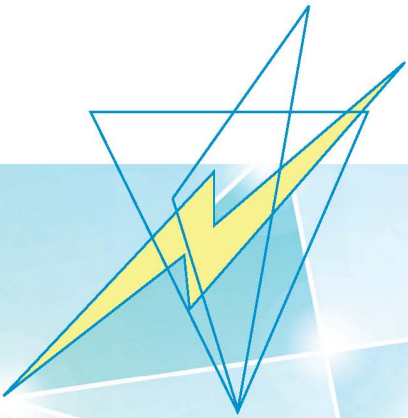
ANTENNA SYSTEMS and Electronic Warfare Applications



RICHARD A. POISEL

Antenna Systems and Electronic Warfare Applications

For a complete listing of titles in the
Artech House Electronic Warfare Library,
turn to the back of this book.

Antenna Systems and Electronic Warfare Applications

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 U.S. Library of Congress.

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

Cover design by Vicki Kane

ISBN 13: 978-1-60807-484-6

685 Canton Street
Norwood, MA 02062

10 9 8 7 6 5 4 3 2 1

To Jim, Teresa, and Sheryl.

Table of Contents

Preface

There are many excellent textbooks available that cover RF antennas. This book is not intended to usurp those books whatsoever. Our primary interest lies in applying this theory to applications in electronic warfare (EW) systems. The necessary background information is included so that this book is reasonably self-contained, however. We discuss the technical parameters of the common types of antennas used in EW systems. The emphasis here is on the EW applications of the common types of these antennas, however.

The material is divided into three principal sections. The first section, Antenna Fundamentals, consists of Chapters 2–4. Chapter 2 discusses the basic principles of electromagnetic (EM) wave propagation. That is followed by a presentation of the fundamental principles that all types of antennas share in Chapter 3. Finally in Chapter 4 we investigate the properties of transmission lines. For those readers familiar with these topics these chapters can be skipped on the first reading, and used for reference in the rest of the book.

Section 2, Antenna Systems, is the main part of the book. We explore the various types of antennas and focus on EW applications. We begin with dipole antennas in Chapter 5, where the major types are investigated. Chapter 6 delves into monopole antennas and their characteristics while Chapter 7 explores loop antennas. Traveling wave antennas are discussed in Chapter 8. The characteristics of antenna arrays, where two or more of the basic antenna types are configured into a single antenna, are examined in Chapter 9. Chapter 10 presents several EW applications of antenna arrays. The ubiquitous Yagi-Uda antenna is explored next in Chapter 11, where dipoles, monopoles, and loops are configured into arrays with particular properties. Chapter 12 presents investigations of some frequency independent antennas. While no antenna is truly frequency independent (such an antenna would be infinitely large), it is generally accepted that an antenna with a 10:1 bandwidth or larger is frequency independent for practical purposes. Aperture EW antennas are presented in Chapter 13. Chapter 14 explores the properties of small antennas. Small size coupled with broad bandwidth are the goal of most EW antenna designs, so size is a particularly important topic. Fundamental characteristics of patch antennas are examined in Chapter 15, followed by discussions of EW applications of patch antennas in Chapter 16. Adaptive antenna

arrays are explored in Chapter 17. Fractal antenna designs are presented next in Chapter 18. The section concludes with a discussion on genetic antenna design techniques in Chapter 19.

The third section, Related Antenna Topics, consisting of Chapters 20 through 22, examines some antenna-related topics. These topics are rarely included in books on antennas. Chapter 20 discusses the basics of antenna matching, with some specific designs presented. Topics related to matching: multicouplers, combiners, and diplexers are examined in Chapter 21. Finally, the section concludes with an exploration of the characteristics of radomes in Chapter 22.

Four appendices are included at the end. These topics were relegated to appendices to keep the technical flow as smooth as possible in the main chapters of the book. Appendix A presents the basic properties of RF amplifiers used in active EW systems. RF switches are surveyed in Appendix B. A popular automated technique for analyzing antenna structures is the method of moments; that topic is examined in Appendix C. Lastly, Appendix D discusses the characteristics of dielectric materials.

We provide a list of acronyms toward the end of the book. Also included in this list are the common definitions of some of the variables that we use. Since there are not enough letters in the English and Greek alphabets to avoid occasional duplication, these definitions apply to the majority of the uses of the variables, but occasionally the symbols may be locally redefined. We hope that it is obvious when this happens so confusion does not arise.

Chapter 1

Introduction to Electronic Warfare Antenna Systems

1.1 Introduction

Antennas are used in systems, including *electronic warfare* (EW) systems, to convert electrical signals into propagating *electromagnetic* (EM) waves, and vice versa, for converting propagating EM waves into electrical signals. Frequently, antennas must be broadband, because the frequencies of the signals of operation or *signals of interest* (SOIs) can be located over broad portions of the frequency spectrum. Limited physical space on these systems normally precludes using several antennas, each designed for a relatively narrow frequency band. Thus, antennas that are effective over a wide frequency range are used, which implies that their Q must be low. However, because of the distances over which many systems need to operate, their gain must be high in order to provide adequate sensitivity that implies high Q requirements,

The general considerations for antennas include those just mentioned, as well as such mundane issues as erectability, affordability, manufacturability, supportability, maintainability, and so forth. We will discuss the first set of requirements here, while we will leave the latter set for engineering trade-off studies.

Antenna design historically has been more of an art than a science. Understanding the complexity of especially the near-field components of EM waves is difficult. The far field is almost always assumed to be a plane wave—an assumption that is reasonable since EM waves propagate in a spherical fashion unless disturbed by an intervening structure.

Antennas are normally treated as reciprocal. That is, they exhibit the same characteristics whether they are used for transmitting or receiving. We provide a proof of this property in Chapter 3. Reciprocity is generally true if the antenna is a passive device. It may not be true, however, if the antenna contains active elements such as amplifiers, which are unidirectional.

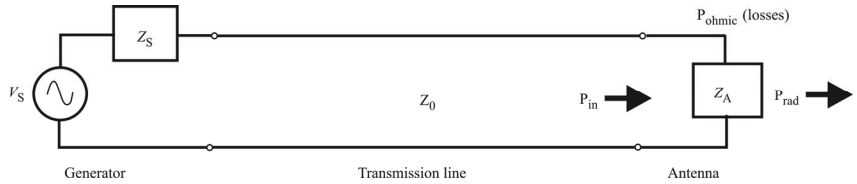


Figure 1.1 Typical transmitting system. Z_S = source impedance, Z_A = antenna impedance, Z_0 = transmission line characteristic impedance, P_{in} = total power delivered to the antenna terminals, P_{ohmic} = antenna ohmic (I^2R) losses comprised of conduction losses and dielectric losses, and P_{rad} = total power radiated by the antenna.

Several antenna patterns are illustrated in the rest of the book. Where possible these patterns were obtained using the *method of moments* (MoM) algorithm [1]. It should be noted, however, that the patterns shown are representative. Particular antenna types are usually employed because they have features that are useful for the intended application. The patterns shown here are intended to show some of those features. However, keep in mind that the characteristics of an antenna are strongly dependent on the dimensions employed, both the size of the antenna and the size of the material (e.g., wire) used to construct the antenna, along with the operating frequency and the environment into which it is deployed.

The two-dimensional (2-D) and three-dimensional (3-D) antenna patterns were obtained using a version of the numerical electric code (NEC) modeling program that incorporates the MoM technique. The program is entitled 4NEC2 Version 5.7.5.[1]

This chapter introduces the types of antennas discussed in the rest of the book.

1.2 Transmitting Systems

A typical transmitter includes a signal generator, transmission line, and transmitting antenna, as shown in Figure 1.1. The generator is modeled by a complex source voltage V_S and a complex source impedance Z_S. The generator could just as well be a current source, as most semiconductor amplifiers [bipolar junction transistor (BJTs) or metal oxide semiconductor field effect transistors (MOSFET)] are. In addition, an antenna radiates energy by changing the momentum of the current not a voltage, so a current source is a more applicable

[1] Newer versions are now available at the time of this wrriting. This program is made available license free by Arie Voors on the Internet free of charge at http://home.ict.nl/~arivoors/.

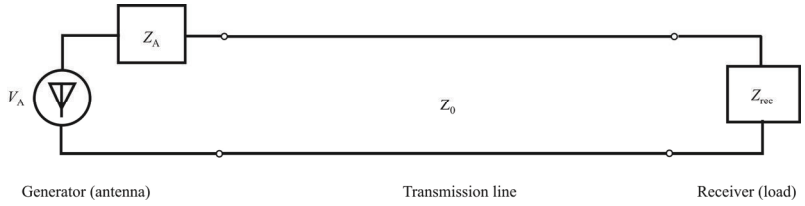


Figure 1.2 Typical receiving system.

model. The antenna is used to convert the electrical signals coming from the transmission line to (normally) free space. Of the components shown in Figure 1.1, we will not be discussing the generator or its characteristics. Books have been written on power amplifiers [2, 3]. We focus only on the antenna, but a chapter is provided that describes the salient characteristics of the transmission line. An appendix on amplifiers is also included for reference.

1.3 Receiving Systems

A typical receiving system can be represented by a generator (receiving antenna), transmission line, and load (receiver), as shown in Figure 1.2. In this case, the antenna is used to convert the EM energy in space into electrical signals for transmission to the receiver down the transmission line. Again, we will be discussing only the antenna and transmission line here. The receiver is discussed only as a load to the transmission line. Characteristics of receivers are a book topic in itself [4].

1.4 Types of Antenna Systems

We will be presenting discussions on the most popular types of antenna systems throughout this book. In this section we give brief introductions to those antenna types.

1.4.1 Wire Antennas

The wire antenna category includes those antenna systems that are composed of one or more wires that take on different shapes. For our purposes, this category includes dipoles, monopoles, loops, and helix.

Figure 1.3 Vertically oriented (polarized) dipole antenna.

These antennas are widely used because of their simple construction and reasonable performance for many applications. Their characteristics have been thoroughly studied and are well understood. Chapters 5–7 present considerably more detail on their character.

Many important characteristics of these antennas are established by the ratio of the diameter of the wire of which the antenna is made to its length.

1.4.1.1 Dipole

A diagram of a vertically oriented dipole antenna is shown in Figure 1.3. The dipole consists of a single wire that is split and fed somewhere along its length. The point where it is split is referred to as its *feed point*. That is the place where the transmission line connecting the antenna to the rest of the system is attached. The dipole shown in Figure 1.3 is split at the middle, and is called a *center fed dipole*.

Considerably more detail is presented on dipole antennas in Chapter 5.

1.4.1.2 Monopoles

A monopole antenna is shown in Figure 1.4. It consists of a single wire that is normally fed at one end. There is a finite ground plane associated with a monopole antenna, although to determine many of the properties of these antennas, it is often assumed that the ground plane is infinite. These ground planes are often considered to be *perfect electrical conductors* (PEC). The characteristics of this ground plane, to a large extent, determine the radiating characteristics of the

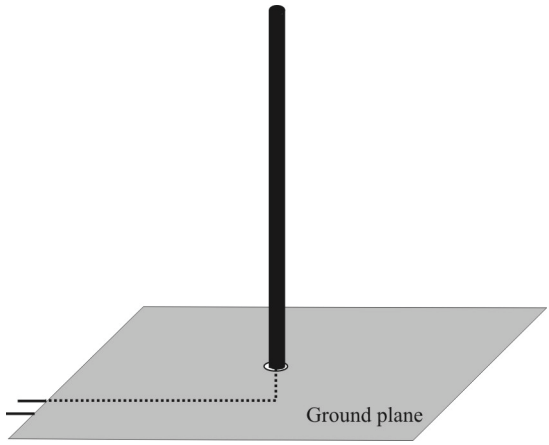


Figure 1.4 Monopole.

monopole. In practice ground planes are typically constructed of wires placed every 3^{o} with length $\lambda/2$. Monopoles are discussed in depth in Chapter 6.

1.4.1.3 Loops

An illustration of a single-turn loop antenna is shown in Figure 1.5, while a depiction of a multiple-turn loop antenna is shown in Figure 1.6. Loop antennas are comprised of a single wire that is bent into the shape of a loop or loops. The actual shape of the loop could be circular, as shown in Figures 1.5 and 1.6, square, rectangular, triangular, and so forth, as will be illustrated in Chapter 7. This shape has little to do with the radiation characteristics of the loop antenna at considerable distances away from the antenna, the case of interest to us.

Loops are notoriously inefficient for transmission systems and are seldom used in that role. When loaded with a ferrite (iron) core, that is, a multiple turn loop antenna is constructed by winding a wire on a ferrite rod, these antennas are very popular as receiving antennas in small applications. Loops are discussed in depth in Chapter 7.

1.4.1.4 Helix

The helix antenna is similar to the multiturn loop discussed above. The helix is a multiturn loop antenna with the multiple loops intentionally separated by more than the minimal amount. They share many of the same characteristics.

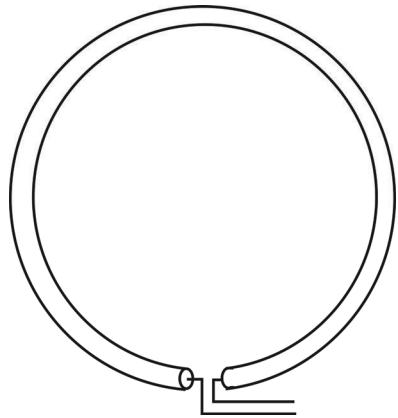


Figure 1.5 Single-turn loop antenna.

Figure 1.6 Multiple-turn loop antenna.

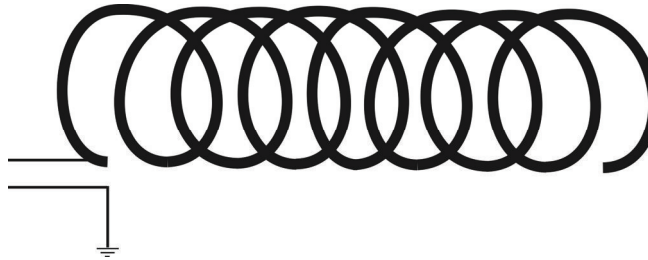


Figure 1.7 Helix antenna. If structured so radiation is off the axis of the loops it is called an axial mode helix. If structured so the radiation is perpendicular to the axis, it is called a radial mode helix.

The helix comes in two flavors determined by the radiating mode. The *axial mode* helix radiates primarily off the end of the axis determined by the center of the antenna windings. The radial mode helix is constructed so that the predominant radiating direction is normal to the axis determined by the antenna windings.

An illustration is provided in Figure 1.7.

1.4.2 Aperture Antennas

Aperture antennas are constructed so that EM fields are created across an opening, an aperture. There are many types of aperture antennas. A diagram of one of them, the pyramidal horn, is shown in Figure 1.8. When the vertical dimension is made substantially smaller than the horizontal one, this antenna is referred to as an H-plane sectoral horn, and if the reverse is true, the horizontal dimension is made substantially smaller than the vertical dimension, it is referred to as an E-plane sectoral horn. E and H refer to the EM waves associated with the antenna. We will discuss E and H fields substantially more in Chapters 2–4.

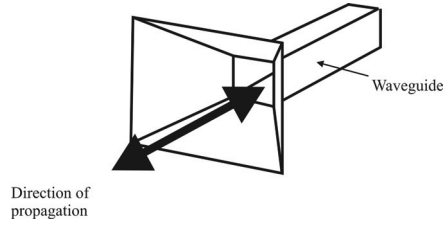


Figure 1.8 Pyramidal horn antenna.

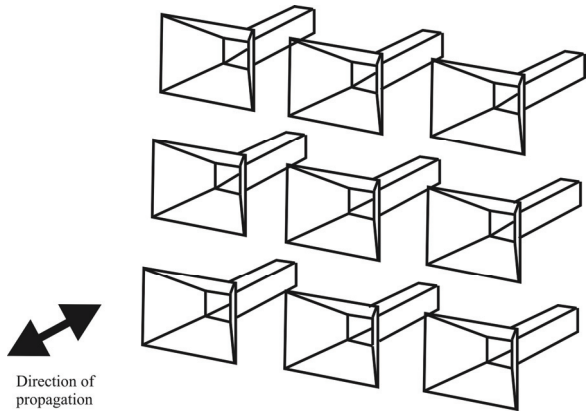


Figure 1.9 Antenna array of pyramidal horn antennas.

Imagine a perfectly opaque plane sheet of absorber with an arbitrarily shaped hole cut in it. This is illuminated from the back side by a plane EM wave at normal incidence (i.e., traveling at right angles to the plane of the absorber). The part of the wave passing through the hole in the absorbing sheet travels on, having been diffracted at the edges of the hole, and as we follow it further on its travels, it spreads out to form a "far field radiation pattern" of a "uniformly illuminated aperture antenna" having the shape of the hole. Aperture antennas are considered in detail in Chapter 12.

1.4.3 Antenna Arrays

Many useful characteristics of antenna systems accrue when the simpler configurations, such as those discussed above, are formed into systems with multiple elements. Such an antenna system is referred to as an *array*.

Arrays can be constructed from any of the simpler types and we will discuss several types of arrays in Chapters 9, 10, 16, and 17.

One particular type of array is referred to as an adaptive array, otherwise referred to generically as a *beamformer*. Such a configuration is of significant importance to EW systems, so we have devoted an entire chapter (Chapter 17) to their characteristics. We focus on the *uniform linear array* (ULA) and the *uniform circular array* (UCA) because of their particular importance, although arrays, whether adaptive or not, can be formed with any arbitrary geometry.

An example of an antenna array comprised of the pyramidal horn is shown in Figure 1.9. Other examples of antenna arrays are the Yagi-Uda and log-periodic, which are both arrays comprised of elementary wire elements (dipoles,

monopoles, loops, or helix). An example of the *log-periodic dipole array* (LPDA) architecture, similar to the familiar TV antenna, is illustrated in Figure 1.10.

1.4.4 Frequency-Independent Antennas

The principal characteristic of frequency-independent antennas is their bandwidth. Other than traveling-wave antennas, all of the antennas discussed herein are resonant antennas. Resonant antennas are dependent on standing-wave creation over the length of the antenna. These standing waves are caused by an impedance mismatch at the end of the antennas. The antenna impedance only matches at a single frequency (with wavelength λ) or other mathematically related frequencies ($\lambda/2$, $\lambda/4$, and so forth). At all other frequencies the antenna is mismatched and standing waves are generated. Therefore, such antennas are referred to as resonant structures.

When antennas operate over a relatively larger bandwidth (by whatever mechanism), say, 10:1, then they are referred to as *frequency-independent antennas*. Over their band of operation, they satisfy a relatively low value of *voltage standing wave ratio* (VSWR); < 2 is such a typical parameter. The LPDA shown in Figure 1.10 is an example of a frequency-independent antenna. We discuss the characteristics of the log-periodic and other types of frequency-independent antennas in Chapter 12.

1.4.5 Microstrip Patch Antennas

A microstrip or patch antenna is a low-profile antenna that has a number of advantages over other antennas (see Figure 1.11). They are lightweight, inexpensive, and easy to integrate with accompanying electronics. While the antenna can be 3-D in structure (wrapped around an object, for example), the elements are usually flat, hence, their other name, planar antennas.

Figure 1.11 shows a patch antenna in its basic form: a flat plate over a ground plane (usually a PC board). The center conductor of a coax serves as the feed probe to couple electromagnetic energy in and/or out of the patch.

The electric field is zero at the center of the patch, maximum (positive) at one side, and minimum (negative) on the opposite side. The minimum and maximum continuously change side according to the instantaneous phase of the applied signal. The electric field between the radiating patch antenna element and the conducting ground plane does not stop abruptly at the patch's periphery as in a cavity; rather, the fields extend past the outer periphery to some degree. This field extension is known as the fringe field and causes the patch to radiate.

We consider the characteristics of patch antennas in Chapter 15 and their applications in EW systems in Chapter 16.

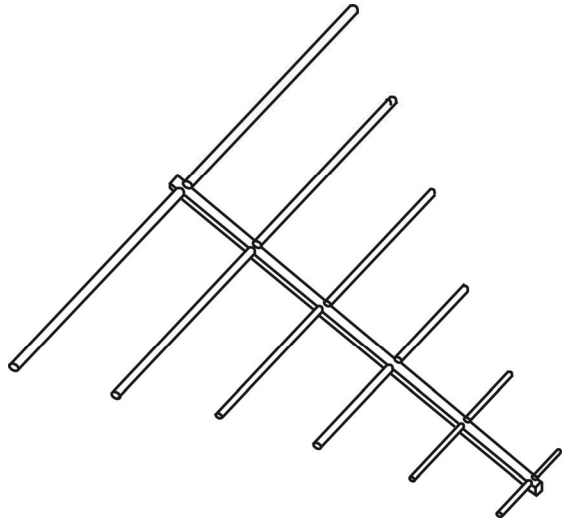


Figure 1.10 Log-periodic is an example of a frequency-independent antenna.

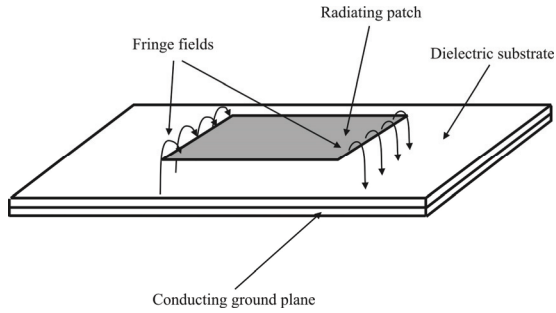


Figure 1.11 Patch antenna.

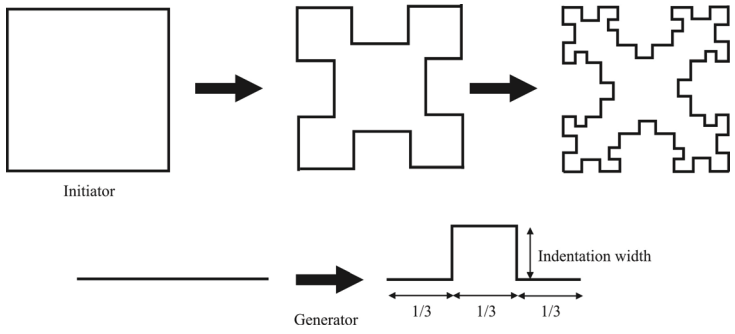


Figure 1.12 The iterative-generation procedure for a Minkowski island fractal. Each straight segment of the geometry is replaced with the generator. The initiator, the square, is shown, along with the first two generating iterations, or prefractals. The indentation length need not be the same as the indentation width in the generator.

1.4.6 Fractal Antennas

As we will discuss in Chapter 14, the limitation of the bandwidth of an antenna is primarily determined by the size of a sphere that can completely enclose the antenna. Efficient use of the space inside the sphere can maximize the bandwidth. Fractal antenna geometries attempt to maximize antenna bandwidth in precisely this way. Beginning with the *initiator*, each straight line segment is replaced with the *generator* as illustrated in Figure 1.12. The resulting antenna length is typically much longer than a monopole of the same "size." The antenna Q is lower and therefore the bandwidth is broader.

There are ample reasons given in published works concerning maximizing bandwidths for us to look at space filling geometries to lower the Q of an antenna. Hansen wrote [5]: "…it is clear that improving bandwidth for an electrically small antenna is only possible by fully utilizing the volume in establishing a TM or TE mode, or by reducing efficiency." Balanis wrote [6]: "…the bandwidth of an antenna (which can be closed within a sphere of radius r) can be improved only if the antenna utilizes efficiently, with its geometrical configuration, the available volume within the sphere." Fractals and other space-filling curves fill spaces more efficiently than one-dimensional antenna structures. We investigate EW applications of fractal antennas more thoroughly in Chapter 18.

1.4.7 Genetically Designed Antennas

Antennas designed using genetic algorithms have been shown to exhibit some interesting characteristics. *Genetic algorithms* (GAs) are search methods based on

the principles and concepts of natural selection and evolution, as espoused by Darwin [7]. These optimization methods operate on a group of trial solutions in parallel, and they operate on the coding of a function's parameters rather than the parameters directly.

In the GA each variable is represented as a binary code called a *gene*. These genes are then arranged and combined to form a *chromosome*. Each chromosome has an associated *fitness value* or *cost* assigning a value of merit to the chromosome. A high fitness value is the characteristic of a good chromosome.

After the starting chromosomes have been created in the GA, a selection strategy determines which chromosomes will take part in the evolution process. These chromosomes mate with one another (according to techniques specified in the specific algorithms) to produce new offspring, which consist of genetic material from two parent chromosomes. The new set of chromosomes produced from the mating process make up the next "generation" of chromosomes, although chromosomes from the previous generation may also be saved and inserted into the new generation. The amount of chromosomes in every generation is kept constant. This process is repeated (selection and mating) until a set amount of generations have been completed.

The primary operation of the GA is that, as the generation number increases, the stronger the population of chromosomes should get, as the weaker chromosomes containing the weaker genes are "weeded out." This, however, is not always the case, as a chromosome with very poor fitness may be one step of mating away from being the best chromosome. In order to prevent this, or to allow for chromosome values that may not have been in the optimizer's search space, the algorithm also makes use of "mutations." Mutations are low-probability random changes, which may occur in a chromosome. We discuss genetically designed antenna systems further in Chapter 19.

1.4.8 Active Antennas

An antenna is active if it contains components that are not passive, such as amplifiers. Active devices essentially perform an impedance transformation and can be used to match the characteristics of, particularly, wire antennas with transmission lines. A depiction of an active receive antenna is presented in Figure 1.13.

Active antennas can be either receive or transmit, although use for transmission is limited because the amplification provided by the antenna can be built into the high power amplifiers prior to the antenna much more economically. One particular type of active antenna we will discuss uses negative impedance converters (NICs) to cancel appropriate antenna impedances.

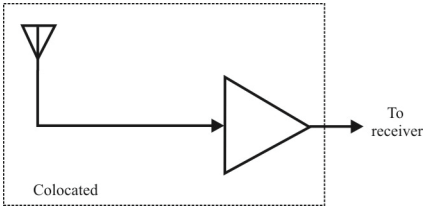


Figure 1.13 An active antenna includes one or more embedded active devices in its structure.

EW applications of active antennas are discussed in Chapter 14 where we discuss antenna miniaturization techniques.

1.5 Concluding Remarks

We have introduced the major categories of antennas that we investigate in the later chapters. We have put together material here to cover a broad waterfront on antennas that is spread over several other sources. What are somewhat unique are the operational requirements that rely on the fundamental physics. In most military and many civilian settings, the requirements on antenna systems are that they must be simultaneously broad-band in frequency coverage while remaining as small as possible. These two parameters are normally at odds with one another.

The antenna system designs we discuss in the remainder of the book have many applications. These include communication systems, radars, and EW applications to mention a few. We will concentrate mostly on the EW applications. Where appropriate, when one or more special features about a particular application or feature of an antenna system is especially important or unique, we point them out.

The book is structured in three parts. In Part I we cover the basic physics involved with antennas, including radiation characteristics, fundamental antenna parameters, and transmission lines. In Part II we delve into the attributes of each of the major antenna types individually. Part III covers some topics related to antenna systems. In particular, antenna matching is discussed in Chapter 20, multicouplers are presented in Chapter 21, and the salient characteristics of antenna radomes are explained in Chapter 22. We have included five appendices that discuss related topics, but not deemed appropriate to include in the main text structure. Appendix A discusses the classes of amplifiers and Appendix B considers the basic characteristics of various types of RF switches used in antennas systems. Appendix C presents some background material on the ubiquitous *method of moments* (MoM) antenna analysis technique. Appendix D presents background

material on dielectric materials that are used for various purposes in antenna systems.

References

[1] Harrington, R. F., "Matrix Methods for Field Problems, *IEEE Proceedings*, Vol. 55, February 1967, pp. 136–149.
[2] Kenington, P. B., *High Linearity RF Amplifier Design*, Norwood, MA: Artech House, 2000.
[3] Gonzalez, G., *Microwave Transistor Amplifiers, Analysis and Design,* 2nd ed., Upper Saddle River, NJ: Prentice Hall, 1997.
[4] Vaccaro, D. D., *Electronic Warfare Receiving Systems*, Norwood, MA: Artech, 1993.
[5] Hansen, R. C., *Proceedings of the IEEE*, Vol. 69 No. 2. February 1981.
[6] Balanis, C. A., *Antenna Theory: Analysis and Design*, 3rd ed., New York: Wiley, 2005.
[7] Darwin, C., *The Origin of Species,* 6th ed., Amherst, New York: Prometheus Books, 1991.

Part I
Antenna Fundamentals

Chapter 2

Principles of Electromagnetic Radiation

2.1 Introduction

We will cover the fundamental principles of EM wave radiation in this chapter as background for later discussions about particular types of EW antenna systems. The purpose of RF antennas is either to convert electrical signals into EM waves (transmitting) or converting EM waves into electrical signals (receiving). An antenna is therefore a sensor of EM waves. Whether transmitting or receiving, the properties of EM radiation are critical for the understanding of antennas. These properties are all based on Maxwell's equations [1].

If an antenna consists of only passive devices, then it is reciprocal. That is, its characteristics for transmitting waves are the same as for receiving EM waves. Therefore, for the most part, analysis of an antenna performance can be approached from either point of view—usually the easiest approach is chosen, and mostly that is when transmitting.

On the other hand, if an antenna contains active components such as an amplifier, then, in general, its transmitting characteristics are different from its receiving characteristics.

We will not cover the principles of EM wave propagation here. For the reader interested in delving into that topic further, [2, 3] are recommended.

We begin this chapter with a discussion of vector and vector fields. EM wave propagation is dictated by what are known as Maxwell's equations, so their characteristics are presented next. After that we delve into the basics of wave propagation in a fairly mathematical way. The chapter ends with a brief discussion of how EM waves are actually formed.

2.2 Coordinate Systems

Although there are many ways to mathematically describe points in 3-D Euclidean space, we will use spherical and rectangular coordinates in this chapter. The

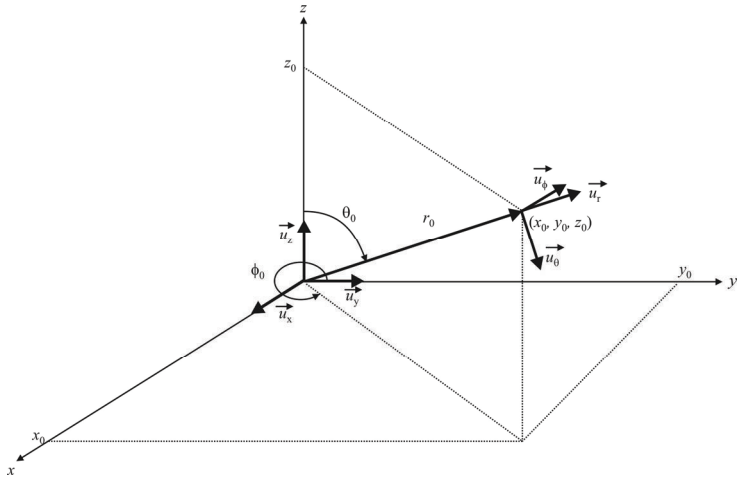


Figure 2.1 Polar and cartesian coordinate definitions.

relationship between spherical coordinates and rectangular coordinates is illustrated in Figure 2.1. A point $\vec{x}$ in 3-D space can be expressed in rectangular coordinates by the 3-tuple (x_0, y_0, z_0) or, equivalently, with unit vectors as[1]

$$\vec{x} = x_0\vec{u}_x + y_0\vec{u}_y + z_0\vec{u}_z \tag{2.1}$$

where the $\vec{u}_i$ are unit vectors[2] in the dimensions defined by $i \in \{x, y, z\}$, as shown in Figure 2.1. The unit vectors in Euclidean space coordinate systems are orthogonal. This same point can be expressed as the 3-tuple (r_0, θ_0, ϕ_0), or equivalently, in spherical coordinates using spherical unit vectors as

$$\vec{x} = r_0\vec{u}_r + \phi_0\vec{u}_\phi + \theta_0\vec{u}_\theta \tag{2.2}$$

Of course, since (2.1) and (2.2) are two different ways of expressing the same notion, there are (simple) transformations between them.

[1] In this book, vectors are represented with an arrow over the vector label, namely, $\vec{x}$, while a matrix is denoted in bold, namely, **X**. Both vectors and matrices can be lower or upper case.
[2] Recall that a unit vector is a vector of unit length oriented in a specified direction.

2.3 Electromagnetic Waves

All waves, including EM waves, come in two basic types: transverse waves and longitudinal waves. The characteristic of the transverse wave is that the direction in which the wave is traveling and the amplitude variation are perpendicular to each other. Water waves, rope waves, phone cord waves, and EM waves are of this type. With a longitudinal wave the wave direction and the amplitude variation direction are parallel to each other. Some systems have both types. The waves in the ocean, for example, exhibit both types. While the wave itself consists of water that is oscillating up and down, the wave travels toward the shore, orthogonal to the amplitude variation.

2.4 Vector Fields

The radiation of power is dictated by the equations of Maxwell. These equations were the culmination of research conducted in 1855 to 1865 by several researchers, including Faraday, Gauss, Coulomb, Ampere, Maxwell, and others. We will discuss them briefly here, as they are the fundamental physical laws that describe how antennas operate.

These equations relate the four fundamental fields involved with radiation: the electric field, $\vec{E}$, the magnetic field density, $\vec{B}$, the magnetic field, $\vec{H}$, and the electric flux density, $\vec{D}$.

Here, as elsewhere in the EM wave literature, we adopt the following conventions with the del operator given by

$$\nabla = \frac{\partial}{\partial x}\vec{u}_{\mathrm{x}} + \frac{\partial}{\partial y}\vec{u}_{\mathrm{y}} + \frac{\partial}{\partial z}\vec{u}_{\mathrm{z}} \tag{2.3}$$

where $\vec{u}_{\mathrm{x}}$, $\vec{u}_{\mathrm{y}}$, and $\vec{u}_{\mathrm{z}}$ are the x, y, and z unit vectors, respectively.[3] The del operator is only defined in the Cartesian coordinate system, however, the related concept of *divergence* of a vector $\vec{v}$ is given by

$$\nabla\vec{v} \triangleq \frac{\partial}{\partial x}\vec{v}_{\mathrm{x}} + \frac{\partial}{\partial y}\vec{v}_{\mathrm{y}} + \frac{\partial}{\partial z}\vec{v}_{\mathrm{z}} \tag{2.4}$$

[3] There will be some minor misuse of the del notation, as it is used both as a vector operator as well as a scalar operator. Hopefully this will not confuse the reader.

where $\vec{v}_x = v_x\vec{u}_x, \vec{v}_y = v_y\vec{u}_y$, and, $\vec{v}_z = v_z\vec{u}_z$ with v_x, v_y, and v_z the magnitudes of the x, y, and z components, respectively. In the spherical coordinate system the divergence is given by

$$\nabla\vec{v} = \frac{1}{r^2}\frac{\partial}{\partial r}r^2\vec{v}_{\mathrm{r}} + \frac{1}{r\sin\theta}\frac{\partial}{\partial\phi}\sin\phi\vec{v}_{\phi} + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}\vec{v}_{\theta} \tag{2.5}$$

The del operator can also be applied to scalars. When it is so applied to scalar s, the result is

$$\nabla s = \frac{\partial s}{\partial x}\vec{u}_{\mathrm{x}} + \frac{\partial s}{\partial y}\vec{u}_{\mathrm{y}} + \frac{\partial s}{\partial z}\vec{u}_{\mathrm{z}} \tag{2.6}$$

For any two vectors $\vec{a}$ and $\vec{b}$:

$\vec{a}\bullet\vec{b}$ is the *dot product*

$$\vec{a}\bullet\vec{b} \triangleq \left|\vec{a}\right|\left|\vec{b}\right|\cos\varphi \tag{2.7}$$

$\vec{a}\times\vec{b}$ is the *cross product*

$$\vec{a}\times\vec{b} \triangleq \left|\vec{a}\right|\left|\vec{b}\right|\sin\varphi\vec{u}_{\mathrm{N}} \tag{2.8}$$

where $\vec{u}_{\mathrm{N}}$ is the normal unit vector (vector of unit magnitude pointing according to the right-hand screw method, that is, in the direction as if $\vec{a}$ were turned into $\vec{b}$ with the motion of a right-hand screw) through an angle φ.

1. $\nabla\psi$ is the *gradient* of scalar field ψ and is a vector field that points in the maximum space rate of change, and whose magnitude is the greatest rate of change [4]. In the spherical coordinate system, the gradient of ψ is

$$\nabla\psi = \frac{\partial\psi}{\partial r}\vec{u}_r + \frac{1}{r}\frac{\partial\psi}{\partial\theta}\vec{u}_{\theta} + \frac{1}{r\sin\theta}\frac{\partial\psi}{\partial\phi}\vec{u}_{\phi} \tag{2.9}$$

2. $\nabla\bullet\vec{V}$ is the *divergence* of vector flux density $\vec{V}$, which is the outflow of flux from a small closed surface per unit volume as the volume shrinks to zero; the

divergence of a vector field is a (signed) scalar [5]. The divergence in the spherical coordinate system is given by

$$\nabla \bullet \vec{V} = \frac{1}{r^2}\frac{\partial r^2 V_{\mathrm{r}}}{\partial r} + \frac{1}{r\sin\theta}\frac{\partial \sin\theta V_{\theta}}{\partial \theta} + \frac{1}{r\sin\theta}\frac{\partial V_{\phi}}{\partial \phi} \tag{2.10}$$

3. $\nabla \times \vec{V}$ is the *curl* of vector field $\vec{V}$ and is the rate of rotation or angular velocity at a point in the field [6]. In spherical coordinates, the curl of $\vec{V}$ is

$$\begin{aligned}\nabla \times \vec{V} = \frac{1}{r\sin\theta}\left[\frac{\partial}{\partial\theta}(\sin\theta V_{\phi}) - \frac{\partial V_{\theta}}{\partial \phi}\right]\vec{u}_r + \frac{1}{r}\left[\frac{1}{\sin\theta}\frac{\partial V_{\mathrm{r}}}{\partial \phi} - \frac{\partial r V_{\phi}}{\partial r}\right]\vec{u}_{\theta} \\ + \frac{1}{r}\left[\frac{\partial r V_{\theta}}{\partial r} - \frac{\partial V_{\mathrm{r}}}{\partial \theta}\right]\vec{u}_{\phi}\end{aligned} \tag{2.11}$$

In determinant form, the curl is given by[4]

$$\nabla \times \vec{V} = \begin{vmatrix} \dfrac{\vec{u}_{\mathrm{r}}}{r^2\sin\theta} & \dfrac{\vec{u}_{\theta}}{r\sin\theta} & \dfrac{\vec{u}_{\phi}}{r} \\ \dfrac{\partial}{\partial r} & \dfrac{\partial}{\partial \theta} & \dfrac{\partial}{\partial \phi} \\ V_{\mathrm{r}} & rV_{\theta} & r\sin\theta V_{\phi} \end{vmatrix} \tag{2.12}$$

2.4.1 Divergence Theorem

Let $\vec{F}$ be a vector field whose components have continuous first-order partial derivatives. Then the divergence theorem states that the outward flux of $\vec{F}$ through S is equal to the triple integral of the divergence on V. Intuitively, it states that the sum of all sources within V minus the sum of all sinks in V gives the net flow out of the region defined by V:

$$\iiint_V \nabla \bullet \vec{F} dv = \iint_S \vec{F} \bullet \vec{n}_{\mathrm{S}} ds \tag{2.13}$$

[4] The determinant form of the curl operation is intended to be a simpler way to remember the equations. The determinant form and equation form lead to identical results.

where $\vec{n}_S$ is the unit normal vector of the surface *S*.

2.4.2 Electric Field and Electric Flux

Electric charges exert forces on each other (like charges repel, dislike charges attract) without physical contact because each charge creates an electric field, denoted by $\vec{E}$, and the electric fields exert forces on other charges. We measure the amount of electric fields acting over a surface *S* by the electric flux, denoted by ϕ.[5] If the electric field is constant, then the electric flux is given by

$$\phi = \vec{E} \bullet \vec{S} \tag{2.14}$$

where the area of the surface is *S* and the direction of $\vec{S}$ is orthogonal to the plane containing *S*. If the electric field is not constant over *S*, then the electric flux is given by

$$\phi = \oiint_S \vec{E} \bullet d\vec{s} \tag{2.15}$$

2.4.3 Constitutive Relationships

We will discuss Maxwell's equations momentarily. However, Maxwell's equations alone are insufficient to solve for the four vector quantities: $\vec{E}, \vec{D}, \vec{H}, \text{and } \vec{B}$ (twelve scalar quantities, one for each dimension per vector). Two additional vector equations are needed that are referred to as the *constitutive relationships*.

The constitutive relationships describe the properties of matter with respect to electric and magnetic forces. These relationships indicate that in a linear, isotropic medium there is a linear relationship between $\vec{E}$ and $\vec{D}$ and between $\vec{H}$ and $\vec{B}$, to wit,

$$\vec{D} = \ddot{\varepsilon} \bullet \vec{E} \tag{2.16}$$

$$\vec{B} = \ddot{\mu} \bullet \vec{H} \tag{2.17}$$

[5] By convention, we use ϕ to represent both the electric flux and the angle in the horizontal plane in antenna patterns. The difference should be clear by the context.

In an anisotropic medium,[6] the dielectric permittivity $\ddot{\varepsilon}$ and the magnetic permeability $\ddot{\mu}$ are *tensors* (matrices) as indicated by the notation $(\ddot{\bullet})$. In a vacuum, which is isotropic, the dielectric permittivity and the magnetic permeability are constants (or tensors whose *diagonal elements only* are nonzero and equal) given by

$$\varepsilon_0 = 8.854\times10^{-12}\,\mathrm{F\,m^{-1}} \tag{2.18}$$

$$\mu_0 = 4\pi\times10^{-7}\,\mathrm{H\,m^{-1}} \tag{2.19}$$

In an isotropic medium, the vectors $\vec{D}$ and $\vec{E}$ are collinear, as are the vectors $\vec{B}$ and $\vec{H}$. In most cases of interest to us, the medium will be assumed to be isotropic, so μ and ε will be single valued constants.

Dielectric properties of materials affect the electric fields. Materials with dielectric properties with relative permittivity (dielectric constant) $\varepsilon_r > 1$ are built from atomic/molecular components that have the properties of electric dipoles. When submerged in an external electric field, the dipoles tend to orient themselves in such a way that their own fields have a cancellation effect on the external field. The electric force $\vec{F}_e = q\vec{E}$ exerted on a point charge q from a source Q_S in such medium is ε_r times weaker than the electric force of the same source in a vacuum. Properties of dielectric materials of interest to us are discussed in Appendix D.

Magnetic materials, on the other hand, with relative permeability (magnetic constant) $\mu_r > 1$ are made of atomic/molecular components that, when submerged in an external magnetic field, tend to orient themselves in such a way that their own magnetic fields align with the external field. The magnetic force $\vec{F}_m = q\vec{v}\times\vec{B}$ exerted on a moving point electric charge q in such a medium will be μ_r times stronger than the force that this same source (e.g., electric currents) would create in a vacuum.

For the most part we are concerned with isotropic media, that is, media where the equations

$$\vec{B} = \mu_0\mu_r\vec{H} \qquad \vec{D} = \varepsilon_0\varepsilon_r\vec{E} \tag{2.20}$$

[6] An anisotropic medium has properties that are directionally dependent, as opposed to isotropy, which means uniformity in all directions.

are valid. The only case that applies for our analysis involving an anisotropic medium is when dielectric material is placed around an antenna to slow the wave velocity in the material, which has the effect of causing a short antenna to embrace characteristics of longer antennas.

2.5 Maxwell's Equations

The greatest single advance in theoretical physics in the 1800s was the formulation of what are now known as Maxwell's equations, named after the nineteenth-century physicist James Clerk Maxwell (1831–1879). The eponymous equations were actually developed by several physicists, but it was Maxwell who determined the importance of these four equations from the many in the field of theoretical physics that completely define the entire field of electromagnetics. Maxwell's unique contribution was adding one term to one of the equations, a displacement current. The equations are named after Maxwell because he showed how they work together to define completely the field of electromagnetics. In 1865 James Clerk Maxwell published the set of equations now bearing his name.

Maxwell's four equations govern all the characteristics of magnetic and electric fields. Individually they are titled: (1) Ampere's law, generalized by Maxwell to include the displacement current; (2) Faraday's law of induction; (3) Gauss' electric law; and (4) Gauss' magnetic law.

As mentioned, Maxwell's unique contribution inclusion of the displacement current, the last term of Ampere's law, the first of Maxwell's equations, described below. This describes what happens to electric fields when they change, such as when the static electricity from your finger arcs to the doorknob (and gives you a shock). The electric field around your finger suddenly drops to zero. This sudden change in the electric field generates a magnetic field. The second equation, Faraday's law, says the reverse is true as well. If you change a magnetic field, (say, by turning an electromagnet on and off), you generate an electric field. Electric and magnetic fields can complement each other, each changing from one into the other, essentially forever. The result in this case is what we know as light, or radio waves, or, in general, EM radiation.

What Maxwell's equations do not explain is why nature behaves this way. They do not explain why, for example, two magnets attract or repel. They only allow us to calculate how strongly and in what direction the magnets are pulled. The question why has yet to be answered.

Maxwell's laws completely describe the characteristics of electromagnetism. First presented by Oliver Heaviside and William Gibbs in 1884, the formal structure of Maxwell's equations is based on the work of Maxwell during the 1860s [1].

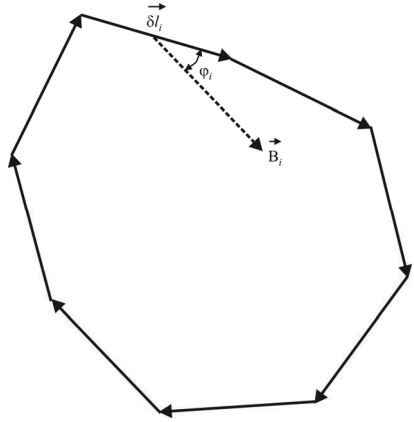


Figure 2.2 Ampere's law.

Each of the four components of the set of equations addresses four different aspects of electromagnetism. As a group, Maxwell's equations articulate the relationships that exist between the electric charge, the magnetic field, the electric field, and the electric current. Only one of the equations was actually developed by Maxwell himself. The other three were preexisting understandings posited by others in the field and interpreted by Maxwell in his 1861 work "On Physical Lines of Force" [1]. However, the explanations that were made by Maxwell, along with the manner in which he connected the fundamentals of Ampere's law, Faraday's law, and Gauss laws were widely adopted. With only some slight changes to allow for developments in the study of electromagnetism, Gibbs and Heaviside extracted these four elements from Maxwell's earlier work and attached the appellation Maxwell's equations to the combination.

2.5.1 Ampere's Law Generalized by Maxwell to Include the Displacement Current

Ampere's law is the magnetic equivalent of Gauss' electric law. The difference is that it refers to a closed loop and the surface enclosed by it (rather than a closed surface and the volume enclosed by it, as is the case with Gauss' electric law).

Consider a closed loop, not necessarily a circle, which is broken into small elements of length $\Delta\vec{l}_i$ with a magnetic field $\vec{B}_i$ at each element as illustrated in Figure 2.2. The sum over elements of the component of the magnetic field along the direction of the element, times the element length, is proportional to the current *I* that passes through the loop as

$$\sum_i B_i \delta l_i \cos\varphi_i = \mu_0 I \tag{2.21}$$

This is Ampere's law. For the case of a wire, the loop can be a circle drawn around the wire, and since the field is always tangent to the circle, cos φ = 1. The circumference of the circle of radius r is $2\pi r$; therefore, Ampere's law becomes

$$2\pi r B(r) = \mu_0 I \tag{2.22}$$

which is the expression for the magnetic field of a wire.

For a static electric field, the law states that the line integral of the magnetic field around a closed loop is proportional to the current flowing in the loop. As modified to include the displacement current, the integral form of the law can be written as

$$\oint_L \vec{B} \bullet d\vec{l} = \mu_0 I + \mu_0 \varepsilon_0 \frac{\partial}{\partial t} \oiint_S \vec{E} \bullet d\vec{s} \tag{2.23}$$

The second term on the right in (2.23) is the displacement current term added by Maxwell.

In differential form, Ampere's law is stated as

$$\nabla \times \vec{B} = \frac{1}{c^2} \frac{\partial \vec{E}}{\partial t} + \frac{1}{\varepsilon_0 c^2} \vec{J} \tag{2.24}$$

where

$\vec{J}$ is the electric current density, A m^{-2}
I is the electric current, A

All other variables are as defined previously.

2.5.2 Faraday's Law of Induction

Faraday's law of induction describes how an electric field can be induced by a changing magnetic flux. It relates the voltage produced in a coil in a magnetic field to the number of loops in the coil, the change in magnetic flux, and the time to carry out the change. Mathematically, it says that the line integral of the electric field around a closed loop L is equal to the negative of the rate of change of the

magnetic flux through the area enclosed by the loop, *S*. The integral form of this law is given by

$$\oint_L \vec{E} \bullet d\vec{l} = -\frac{d\Phi_B}{dt} = -\mu \frac{d}{dt} \iint_S \vec{H} \bullet d\vec{s} \tag{2.25}$$

In differential form this is

$$-\nabla \times \vec{E} = \frac{\partial \vec{B}}{\partial t} + \vec{M} \tag{2.26}$$

where

S is the surface determined by loop *L*, m^2;
$d\vec{s}$ is the infinitesimal segment of the surface *S* normal to *S*, m^2;
Φ_B is the magnetic flux through *S*;
$\vec{M}$ is the magnetic current density,[7] V m^{-2}.

Faraday's law is the theoretical basis for how electric generators work because the line integral given in (2.25) is equal to the voltage generated in the loop.

2.5.3 Gauss' Electric Law

Gauss' electric law states that the total outward electric flux through a closed surface *S* enclosing a charge is proportional to the net charge enclosed by the surface. The integral form of the law is given by

$$\iint_S \vec{E} \bullet d\vec{s} = \frac{q_{\text{enclosed}}}{\varepsilon_0} = 4\pi k q_{\text{enclosed}} = \frac{1}{\varepsilon_0} \iiint_{V_S} \rho dv \tag{2.27}$$

where $k = 1/(4\pi\varepsilon_0)$. The integral form of Gauss electric law has applications in determining the electric fields around charged objects.

In differential form this law gives a measure of the density of the enclosed sources and is given by

[7] $\vec{M}$ represents magnetic currents and is a fictitious quantity since magnetic currents do not exist in nature.

$$\nabla \bullet \vec{D} = \rho \tag{2.28}$$

or

$$\nabla \bullet \vec{E} = \frac{\rho}{\varepsilon_0} = 4\pi k\rho \tag{2.29}$$

where

ρ is the electric charge density, C m^{-3};
V_S is the volume bounded by S, m^3.

Equations (2.28) and (2.29) follow from (2.24) and the continuity relation

$$\nabla \bullet \vec{J} = -\frac{\partial \rho}{\partial t} \tag{2.30}$$

2.5.4 Gauss' Magnetic Law

Gauss' law for magnetism states that the net magnetic flux out of any closed surface is zero. In other words, for any closed surface containing a magnetic dipole, the magnetic flux directed inward toward the South Pole will equal the flux outward from the North Pole. This is stated by

$$\oiint_S \vec{B} \bullet d\vec{s} = 0 \tag{2.31}$$

which is the integral form of the law.

In differential form, this law states that the divergence of a magnetic vector field is proportional to the point source density ρ_m. Mathematically this is stated as[8]

$$\nabla \bullet \vec{B} = \rho_m \tag{2.32}$$

Equation $\nabla \bullet \vec{B} = 0$ follows from (2.32) provided that $\vec{M} = 0$.

---

[8] ρ_m is a fictitious quantity introduced via the continuity relation $\nabla \bullet \vec{M} = -\partial \rho_m / \partial t$. In nature $\nabla \bullet \vec{B} = 0$.

2.6 EM Wave Propagation

In this section we will develop the theory of wave propagation beginning with the foundations provided by Maxwell's equations presented in Section 2.3. This development helps in understanding the subtleties associate with the various types of antennas presented in the later chapters.

2.6.1 Wave Equation

From (2.26)–(2.32), in empty space (no current or charge so that $\rho = \rho_m = J = Q = 0$)

$$\nabla \bullet \vec{E} = 0 \tag{2.33}$$

$$\nabla \bullet \vec{B} = 0 \tag{2.34}$$

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{2.35}$$

$$\nabla \times \vec{B} = \varepsilon_0 \mu_0 \frac{\partial \vec{E}}{\partial t} \tag{2.36}$$

Invoking (2.33), the following curl property results

$$\nabla \times (\nabla \times \vec{E}) = \nabla (\nabla \bullet \vec{E}) - \nabla^2 \vec{E} = -\nabla^2 \vec{E} \tag{2.37}$$

Using (2.35) and (2.36) we get

$$\nabla \times (\nabla \times \vec{E}) = -\frac{\partial}{\partial t}(\nabla \times \vec{B}) = -\varepsilon_0 \mu_0 \frac{\partial^2 \vec{E}}{\partial t^2} \tag{2.38}$$

which, when combined with (2.37), yields the *homogeneous* (no forcing function) *fundamental wave equation*[9]

$$\nabla^2 \vec{E} - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \vec{E} = 0 \tag{2.39}$$

[9] This equation is fundamental because it relates the geometric behavior to the time behavior of the EM wave.

The constant c is generally referred to as the speed of propagation, and for EM waves in free space it follows from the above derivation that

$$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} \approx 3 \times 10^8 \text{ ms}^{-1} \tag{2.40}$$

the speed of light in a vacuum.

We only consider one of the components in (2.39), the field component associated with the radial vector, $\vec{r}$, denoted $E(\vec{r},t)$. Any field of the form $E(\vec{r},t) = f(t - \vec{r}^{\,\mathrm{T}}\vec{\alpha})$ satisfies (2.39), provided $|\vec{\alpha}| = 1/c$. The field can be interpreted as a wave traveling in the direction $\vec{\alpha}$, with the speed of propagation given by $1/|\vec{\alpha}| = c$. For the latter reason, $\vec{\alpha}$ is referred to as the *slowness vector*.

We consider only narrowband signals. Taking the origin as a reference, a narrowband transmitted EM wave can be expressed as

$$E(0,t) = s(t)e^{j\omega t} \tag{2.41}$$

where $s(t)$ is slowly varying with time compared to the carrier $e^{j\omega t}$. For $|\vec{r}| << c/B$, where B is the bandwidth of $s(t)$, we can write

$$E(\vec{r},t) = s(t - \vec{r}^{\,\mathrm{T}}\vec{\alpha})e^{j\omega(t - \vec{r}^{\,\mathrm{T}}\vec{\alpha})} \approx s(t)e^{j(\omega t - \vec{r}^{\,\mathrm{T}}\vec{\beta})} \tag{2.42}$$

The *wave vector* $\vec{\beta} = \vec{\alpha}\omega$ was introduced in (2.42). The magnitude of $\vec{\beta}$,

$$|\vec{\beta}| = \beta = \omega/c = 2\pi/\lambda \tag{2.43}$$

is called the *wave number* or the *propagation phase constant*, where λ is the *wavelength.*[10] Note that $\vec{\beta}$ points in the direction of propagation. For example, in the *x*-*y*-plane we have

[10] It should be noted that β is the propagation phase constant for an unbounded medium, such as the atmosphere, where (2.43) applies. In general (e.g., bounded transmission lines), this expression is not necessarily true and the propagation phase constant is given by k in (2.42). For all conditions we consider herein except for transmission lines in Chapter 5, (2.43) is a valid approximation.

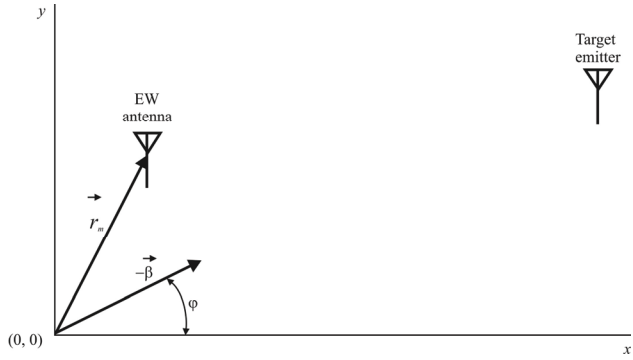


Figure 2.3 Two-dimensional array geometry.

$$\vec{\beta} = \beta\left[\cos\varphi \quad \sin\varphi\right]^{\mathrm{T}} \tag{2.44}$$

where φ is the direction of propagation, defined counterclockwise relative the *x*-axis (see Figure 2.3).

Equation (2.42) assumes far-field conditions, ss an isotropic point source gives rise to a spherical traveling wave (near and far regions are discussed much more extensively in Chapter 3). All points lying on the surface of a sphere of radius *R* will then share a common phase and collectively are referred to as a *wave front.* Therefore, the distance between the target emitter and the receiving antenna array determines whether the curvature of the wave should be taken into account. Far-field receiving conditions imply that the radius of propagation is so large compared to the physical size of the array that a flat plane of constant phase can be considered, thus resulting in a *plane wave* as indicated in (2.42). Except in Chapter 3 where the propagation regions are defined, we will assume such a plane wave.

2.6.2 Production of Electromagnetic Waves by a Dipole Antenna

Electromagnetic waves are produced when electric charges are accelerated. That is, a charge must radiate energy when it undergoes acceleration. Radiation is not produced by stationary charges or steady currents.

A common way of producing EM waves is to apply a sinusoidal voltage source to an antenna, causing the charges to accumulate near the tips of the antenna. The effect is to produce an oscillating electric dipole. This production process is depicted in Figure 2.4.

At $t = 0$, the ends of the rods are charged so that the upper rod has a maximum positive charge and the lower rod has an equal amount of negative charge. At this

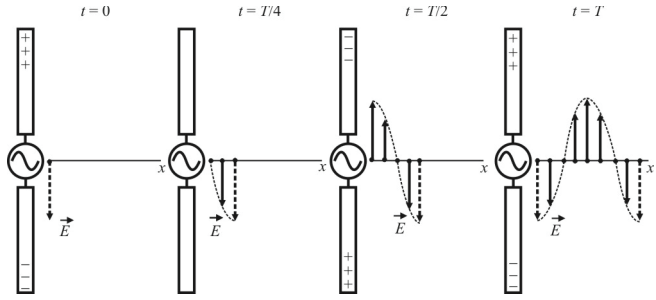


Figure 2.4 Production of an EM wave by a dipole antenna.

instant the electric field near the antenna points downward. The charges then begin to decrease. After $t = T/4$ s, the charges vanish momentarily and the electric field strength is zero. Subsequently, the polarities of the rods are reversed with negative charges continuing to accumulate on the upper rod and positive charges on the lower until $t = T/2$ s, when the maximum is attained again, but with opposite polarity. At this moment, the electric field near the rod points upward. As the charges continue to oscillate between the rods, electric fields are produced and move away at the speed of light. The motion of the charges also produces a current, which, in turn, sets up a magnetic field encircling the rods. However, the behavior of the fields near the antenna is expected to be very different from that far away from the antenna.

2.6.3 Time-Harmonic Fields

In harmonic analysis of EM fields, the *field phasors* are[11]:

$$\begin{aligned}
\mathcal{E}(x,y,z,t) &= \mathrm{Re}\left\{\vec{E}(x,y,z)e^{j\omega t}\right\} \\
\mathcal{H}(x,y,z,t) &= \mathrm{Re}\left\{\vec{H}(x,y,z)e^{j\omega t}\right\} \\
\mathcal{B}(x,y,z,t) &= \mathrm{Re}\left\{\vec{B}(x,y,t)e^{j\omega t}\right\} \\
\mathcal{D}(x,y,z,t) &= \mathrm{Re}\left\{\vec{D}(x,y,z)e^{j\omega t}\right\}
\end{aligned} \tag{2.45}$$

[11] $\mathcal{E}(\cdot)$ should not be confused with the statistical expectation operator denoted by $\mathcal{E}\{\cdot\}$.

In all cases of interest to us, the time variability of the EM waves need not be explicitly indicated because all terms have the same factor, $e^{j\omega t}$, as indicated in (2.45). The usual way to represent fields is with phasors where the time dependence is assumed. From this point on, we will denote time-dependent field vectors with script letters, while their phasors will be denoted with associated upper case letters as

$$\mathcal{F}(x,y,z,t) \Leftrightarrow F(x,y,z), \quad \mathcal{F} \in \{\mathcal{E},\mathcal{H},\mathcal{B},\mathcal{D}\}, F \in \{E,H,B,D\} \tag{2.46}$$

From Fourier theory we know that

$$\frac{\partial \mathcal{F}(x,y,z,t)}{\partial t} \Leftrightarrow j\omega F(x,y,z) \tag{2.47}$$

$$\frac{\partial \mathcal{F}(x,y,z,t)}{\partial v} \Leftrightarrow \frac{\partial F(x,y,z)}{\partial v}, \qquad v \in \{x,y,z\} \tag{2.48}$$

Maxwell's equations in phasor form are given as

$$\nabla \times \vec{H} = j\omega(\varepsilon' - j\varepsilon'')\vec{E} + \sigma\vec{E} + \vec{J} = j\omega\bar{\varepsilon}\vec{E} + \vec{J} \tag{2.49}$$

where

$$\bar{\varepsilon} = \varepsilon' - j\left(\varepsilon'' + \frac{\sigma}{\omega}\right) \tag{2.50}$$

and

$$-\nabla \times \vec{E} = j\omega(\mu' - j\mu'')\vec{H} + \vec{M} = j\omega\bar{\mu}\vec{H} + \vec{M} \tag{2.51}$$

Equations (2.49) and (2.51) are the general form of Maxwell's equations. They include the fictitious magnetic currents $\vec{M}$ (in reality there are no such currents). The dielectric losses (due to alternate field conductivity $\omega\varepsilon''$) and the static field conductivity losses σ are both represented by the imaginary part of the *complex dielectric permittivity* $\bar{\varepsilon}$. Often, the dielectric losses are represented by the *dielectric loss angle* δ_d (or its tangent, known as the *loss tangent*):

$$\bar{\varepsilon} = \varepsilon'\left[1 - j\left(\frac{\varepsilon''}{\varepsilon'} + \frac{\sigma}{\omega\varepsilon'}\right)\right] = \varepsilon'\left[1 - j\left(\tan\delta_d + \frac{\sigma}{\omega\varepsilon'}\right)\right] \tag{2.52}$$

Similarly, the magnetic losses are described by the imaginary part of the *complex magnetic permeability* $\bar{\mu}$ or by the magnetic loss-angle δ_m :

$$\bar{\mu} = \mu' - j\mu'' = \mu'\left(1 - j\frac{\mu''}{\mu'}\right) = \mu'(1 - j\tan\delta_m) \tag{2.53}$$

As mentioned, we are mostly concerned about loss-free, homogeneous, isotropic media but when considering material loaded antennas, the material loading changes these three characteristics of the media immediately adjacent to the antenna. The ground can also disturb the homogeneity in the medium around antennas located close to the ground.

2.6.4 Radiated Power Flow

2.6.4.1 Poynting Vector

The Poynting vector, $\vec{P}$, is a vector representing the density and the direction of the EM wave energy flow [8]. In the time domain, the Poynting vector is defined as

$$\vec{P}(t) = \vec{E}(t) \times \vec{H}(t), \qquad \mathrm{W\,m^{-2}} \tag{2.54}$$

Thus, the total power leaving volume V, denoted here by Π, is found as

$$\Pi(t) = \oiint_{S_V} \vec{P}(t) \bullet d\vec{s}, \quad \mathrm{W} \tag{2.55}$$

Since

$$\mathcal{E}(t) = \mathrm{Re}\left\{\vec{E}e^{j\omega t}\right\} = \frac{1}{2}\left(\vec{E}e^{j\omega t} + \vec{E}^* e^{-j\omega t}\right) \tag{2.56}$$

and

$$\mathcal{H}(t) = \mathrm{Re}\left\{\vec{H}e^{j\omega t}\right\} = \frac{1}{2}\left(\vec{H}e^{j\omega t} + \vec{H}^* e^{-j\omega t}\right) \tag{2.57}$$

then we can express the Poynting vector as

$$\vec{\mathcal{P}}(t) = \frac{1}{2}\mathrm{Re}\left\{\vec{E}\times\vec{H}^{*}\right\} + \frac{1}{2}\mathrm{Re}\left\{\vec{E}\times\vec{H}e^{2j\omega t}\right\} \tag{2.58}$$

The first term in (2.58) is the average value of the power flux density, $\vec{\mathcal{P}}_{\mathrm{avg}}$,

$$\vec{P}_{\mathrm{avg}} = \frac{1}{2}\mathrm{Re}\{\vec{E}\times\vec{H}^{*}\} \tag{2.59}$$

about which the power flux density fluctuates and has no time dependence. It is a vector of unchanging direction showing a constant outflow (positive value) or inflow (negative value) of EM energy. It describes the active energy flow, which is the time-average power flux, as

$$\Pi_{\mathrm{avg}} = \oiint_{S_{\mathrm{V}}} \vec{\mathcal{P}}_{\mathrm{avg}} \bullet d\vec{s} \tag{2.60}$$

The second term in (2.58) is a vector changing its direction with a double frequency (2ω). It describes the reactive power flow (i.e., the power that oscillates in space without contributing to the overall transport of energy in any specific direction).

Definition [8]: The complex *Poynting vector* is

$$\vec{P} \triangleq \frac{1}{2}\vec{E}\times\vec{H}^{*} \tag{2.61}$$

the real part of which is the average power flux density $\vec{\mathcal{P}}_{\mathrm{avg}}$.

From (2.61) and the fact that $H\vec{u}_{\phi} = E\vec{u}_{\theta}/\eta,$ the average power density is

$$\vec{P}_{\mathrm{den}} = \frac{1}{2\eta}E\vec{u}_{\theta}\times E\vec{u}_{\phi} = \frac{1}{2\eta}\left|E\right|^{2}\vec{u}_{\mathrm{r}} \tag{2.62}$$

This flow of power is in the direction of the radial unit vector—that is, in the direction of an expanding sphere.

2.6.4.2 Radiated Power

Definition: *Radiated power* is the average power radiated by the antenna:

$$\Pi_{\text{rad}} \triangleq \oiint_{S_V} \vec{P}_{\text{avg}} d\vec{s} = \oiint_{S_V} \text{Re}\{\vec{P}\} d\vec{s} = \frac{1}{2} \oiint_{S_V} \text{Re}\{\vec{E} \times \vec{H}\} d\vec{s} \tag{2.63}$$

2.6.4.3 Plane Electromagnetic Waves

Consider for simplicity an EM wave propagating in the +*x*-direction, with the electric field $\vec{E}$ pointing in the +*y*-direction and the magnetic field $\vec{B}$ in the +*z*-direction, as depicted in Figure 2.5.

This is an example of a plane wave since at any instant both $\vec{E}$ and $\vec{B}$ are uniform over any plane perpendicular to the direction of propagation. In addition, the wave is transverse because both fields are perpendicular to the direction of propagation, which points in the direction of the cross product $\vec{E} \times \vec{B}$.

Using Maxwell's equations, we can obtain the relationship between the magnitudes of the fields. Consider a rectangular loop that lies in the *xy* plane, with the left side of the loop at *x* and the right at $x + \Delta x$ as illustrated in Figure 2.6. The bottom side of the loop is located at *y*, and the top side of the loop is located at $y + \Delta y$. Let the unit vector norm to the loop be in the positive *z*-direction, $\vec{n} = \vec{u}_z$.

From (2.25), rewrite the left-hand side as

$$\begin{aligned} \oint \vec{E} \bullet d\vec{s} &= E_y(x + \Delta x)\Delta y - E_y(x)\Delta y = [E_y(x + \Delta x) - E_y(x)]\Delta y \\ &= \frac{\partial E_y}{\partial x}(\Delta x \Delta y) \end{aligned} \tag{2.64}$$

where we have used the expansion

$$E_y(x + \Delta x) = E_y(x) + \frac{\partial E_y}{\partial x}\Delta x + \cdots \tag{2.65}$$

Similarly, the rate of change of magnetic flux on the right-hand side is given by

$$-\frac{d}{dt} \iint \vec{B} \bullet d\vec{A} = -\left(\frac{\partial B_z}{\partial t} \right) \Delta x \Delta y \tag{2.66}$$

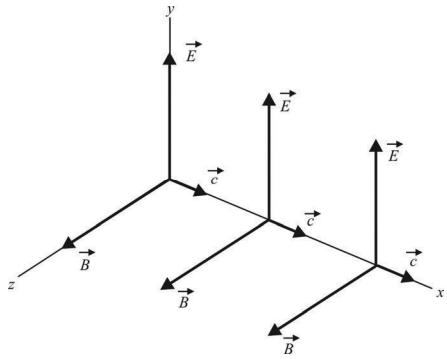


Figure 2.5 Plane EM wave where the electric field and magnetic fields are orthogonal to one another and to the direction of propagation.

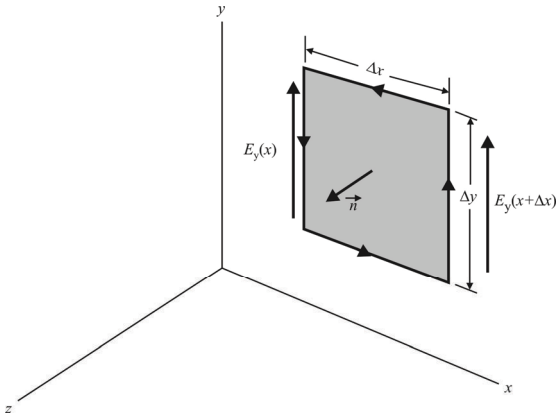


Figure 2.6 Spatial variation of the electric field $\vec{E}$.

Combining (2.64) with (2.66), we get

$$\frac{\partial E_y}{\partial x} = -\frac{\partial B_z}{\partial t} \tag{2.67}$$

A similar derivation using a rectangle in the *xz*-plane and the Ampere-Maxwell equation (2.23) (without sources, so $I = 0$) yields

$$-\frac{\partial B_z}{\partial x} = \mu_0 \varepsilon_0 \frac{\partial E_y}{\partial t} \tag{2.68}$$

As mentioned in the discussion following (2.39), it can be shown that any function of the form $\psi(x \pm vt)$ satisfies the one-dimensional wave equation from (2.39) as (we will assume the *x*-direction)

$$\frac{\partial^2 \psi(x,t)}{\partial x^2} - \frac{1}{v^2}\frac{\partial^2 \psi(x,t)}{\partial t^2} = 0 \tag{2.69}$$

One possible solution to (2.69) is

$$\begin{aligned} \vec{E} &= E_y(x,t)\vec{u}_y = E_0 \cos[\beta(x - vt)]\vec{u}_y = E_0 \cos(\beta x - \omega t)\vec{u}_y \\ \vec{B} &= B_z(x,t)\vec{u}_z = B_0 \cos[\beta(x - vt)]\vec{u}_z = B_0 \cos(\beta x - \omega t)\vec{u}_z \end{aligned} \tag{2.70}$$

where the fields are sinusoidal, with amplitudes E_0 and B_0. As before, the angular wave number β is related to the wavelength by

$$\beta = \frac{2\pi}{\lambda} \tag{2.71}$$

and, of course,

$$\omega = 2\pi f \tag{2.72}$$

When considering free space, $v = c$.

The characteristics of this EM wave are illustrated in Figure 2.7. We see that the $\vec{E}$ and $\vec{B}$ fields are always in phase. The relationship between the amplitudes of

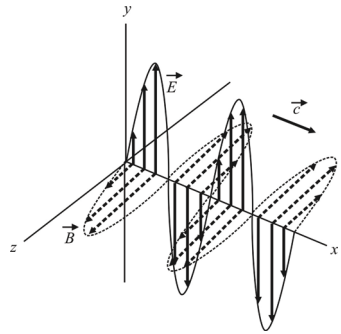


Figure 2.7 Plane EM wave propagating in the +*x* direction.

the wave components are obtained from (2.67). Taking the partial derivatives of (2.70), we get

$$\frac{\partial E_y}{\partial x} = -\beta E_0 \sin(\beta x - \omega t) \tag{2.73}$$

and

$$\frac{\partial B_z}{\partial t} = \omega B_0 \sin(\beta x - \omega t) \tag{2.74}$$

Using (2.67) results in $E_0\beta = \omega B_0$ so that

$$\frac{E_0}{B_0} = \frac{\omega}{\beta} = c \tag{2.75}$$

2.6.4.4 Impedance of Free Space

Using (2.75) with (2.17) and (2.40), we get

$$\frac{E}{\mu_0 H} = c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} \tag{2.76}$$

Thus,

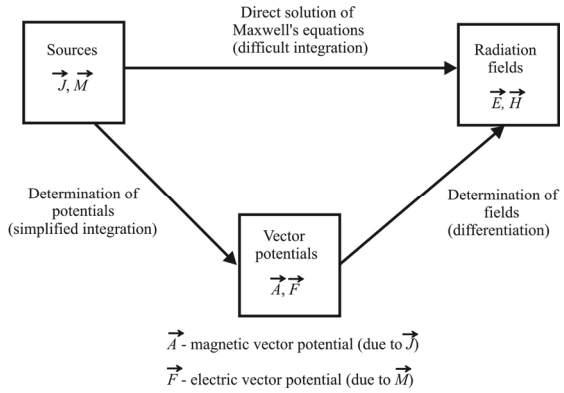


Figure 2.8 Solution of field problems is frequently simplified by using vector potentials rather than Maxwell's equation directly.

$$\frac{E}{H} = \sqrt{\frac{\mu_0}{\varepsilon_0}} \triangleq \eta_0 \tag{2.77}$$

$\eta_0 \approx 120\pi\,\Omega$ is known as the *impedance of free space*.

2.7 Determination of Antenna Radiation Fields Using Vector Potential Functions

The sources of antenna EM radiation fields are: (1) $\vec{J}$, an electric current density (A m^{-2}) vector, and (2) $\vec{M}$, a magnetic current density (V m^{-2}) vector, an artifice devised to facilitate easier field calculation. Some problems involving electric currents can be cast into equivalent forms involving magnetic currents and solved using $\vec{M}$. *Vector potentials* (VPs) were developed to provide an easier solution path to field problems as illustrated in Figure 2.8. Finding the VPs is often much simpler than directly trying to integrate Maxwell's equations. The radiation fields are then relatively easy to calculate with the VPs.

2.7.1 Vector and Scalar Potentials

2.7.1.1 The Magnetic Vector Potential

We first consider electric sources ($\vec{J}$ and $\vec{\rho}$), which are real currents and charges:

$$\nabla \times \vec{E} = -j\omega\mu\vec{H} \tag{2.78}$$

$$\nabla \times \vec{H} = j\omega\varepsilon\vec{E} + \vec{J} \tag{2.79}$$

Since $\nabla \cdot \vec{B} = 0$, we can assume that

$$\vec{B} = \nabla \times \vec{A} \tag{2.80}$$

Substituting (2.80) in (2.78) and (2.79) yields

$$\vec{E} = -j\omega\vec{A} - \nabla\phi \tag{2.81}$$

$$j\omega\varepsilon\vec{E} = \nabla \times \left(\frac{1}{\mu}\nabla \times \vec{A}\right) - \vec{J} \tag{2.82}$$

Substituting (2.81) into (2.82) and rearranging, we get a single equation in the magnetic VP $\vec{A}$. In an isotropic homogeneous region, the result is

$$\nabla \times \nabla \times \vec{A} + j\omega\mu\varepsilon(j\omega\vec{A} + \nabla\phi) = \mu\vec{J} \tag{2.83}$$

Here, ϕ denotes the electric SC. Assuming that

$$\nabla \bullet \vec{A} = -j\omega\mu\varepsilon\phi \tag{2.84}$$

and since $\nabla \times \nabla \times = \nabla \bullet \nabla \bullet (-\nabla^2)$, (2.83) becomes

$$\nabla^2\vec{A} + \omega^2\mu\varepsilon\vec{A} = -\mu\vec{J} \tag{2.85}$$

Equation (2.84) is known as the *Lorenz' gauge condition*.

If the region is lossless, then μ and ε are real numbers and (2.85) can be written as

$$\nabla^2\vec{A} + \beta^2\vec{A} = -\mu\vec{J} \tag{2.86}$$

where $\beta = \omega\sqrt{\mu\varepsilon}$ is the phase constant of the medium. If the region is lossy (which is rarely true in most antenna problems, the most significant exception being

material loading), complex permittivity $\bar{\varepsilon}$ and complex permeability $\bar{\mu}$ are introduced changing (2.85) into

$$\nabla^2 \vec{A} - \gamma^2 \vec{A} = -\bar{\mu}\vec{J} \tag{2.87}$$

$\gamma = \alpha + j\beta = j\omega\sqrt{\bar{\mu}\bar{\varepsilon}}$ is the *propagation constant*, and α is the *attenuation constant*. For example, if the region has losses due to finite conductivity, σ, the complex dielectric permittivity is

$$\gamma^2 = -\omega^2 \mu \underbrace{\left(\varepsilon + \frac{\sigma}{j\omega}\right)}_{\bar{\varepsilon}} \tag{2.88}$$

2.7.1.2 The Electric Vector Potential

Because there are no physically existing magnetic charges, then $\nabla \cdot \vec{B} = 0$, and the magnetic field is therefore a *solenoidal field.*[12] As a consequence, there are no physically existing magnetic currents either. However, the fictitious magnetic currents $\vec{M}$ are a useful tool when applied with the equivalence principle. These currents are introduced in Maxwell's equations in a manner dual to that of the electric currents $\vec{J}$. Now, we consider the field due to magnetic sources *only*; that is, we assume that $\vec{J} = \vec{0}$ and that $\rho = 0$, and therefore $\nabla \bullet \vec{D} = 0$. Then the system of Maxwell's equations is:

$$\nabla \times \vec{E} = -j\omega\mu\vec{H} - \vec{M} \tag{2.89}$$

$$\nabla \times \vec{H} = j\omega\varepsilon\vec{E} \tag{2.90}$$

[12] It can be shown that any vector field $\vec{F}(\vec{r})$ can be decomposed into two terms: a longitudinal (also known as curl-free or irrotational) field $\vec{F}_{\parallel}(\vec{r})$ and a solenoidal (also known as divergence-free, or transverse) field $\vec{F}_{\perp}(\vec{r})$

$$\vec{F}(\vec{r}) = \vec{F}_{\parallel}(\vec{r}) + \vec{F}_{\perp}(\vec{r})$$

with

$$\nabla \times \vec{F}_{\parallel}(\vec{r}) = 0, \qquad \nabla \bullet \vec{F}_{\perp}(\vec{r}) = 0$$

Because $\vec{D}$ is solenoidal, it can be expressed as the curl of a vector, namely the electric VP $\vec{F}$

$$\vec{D} = -\nabla \times \vec{F} \tag{2.91}$$

2.8 Antenna Radiation Fields

2.8.1 Principles of Radiation

EM wave radiation is produced by accelerating or decelerating electric charge (time-varying current element) and that radiation energy is in the form of photons. When an EM wave is incident on a charged particle, the electric and magnetic components of the wave exert a force on the particle, setting it into motion. Since the wave is periodic in time, so is the motion of the particle. Thus, the particle is accelerated and, consequently, emits radiation. More precisely, energy is absorbed from the incident wave by the particle and re-emitted as electromagnetic radiation. Such a process is clearly equivalent to the scattering of the electromagnetic wave by the particle.

The force just referred to carries the appellation the *Lorentz force*. Its effect on a particle is exhibited in Figure 2.10. It is the force on a point charge due to the electromagnetic fields and is given by

$$\vec{F}_{\text{Lorentz}} = q(\vec{E} + \vec{v} \times \vec{B}) \qquad \text{newtons} \tag{2.92}$$

where

$\vec{E}$ is the electric field, V m^{-1}
$\vec{B}$ is the magnetic field, teslas (T)
q is the electric charge of the particle, coulombs (C)
$\vec{v}$ is the velocity of the particle, m s^{-1}

Using the relationships

$$\vec{E} = -\nabla \phi - \frac{\partial \vec{A}}{\partial t} \tag{2.93}$$

$$\vec{B} = \nabla \times \vec{A} \tag{2.94}$$

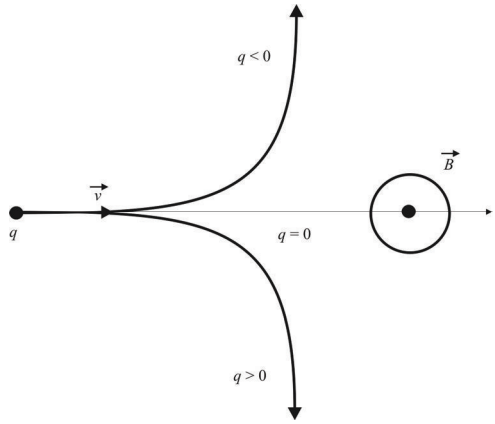


Figure 2.10 Lorentz force. Path of a particle with charge q, under the influence of a magnetic field $\vec{B}$ caused by a current on a wire on the right (directed perpendicularly out of the page), for different values of q. The charge obeys the left-hand rule when greater than zero and the right-hand rule when less than zero.

where $\vec{A}$ and ϕ are the magnetic VP and scalar potential, respectively, the Lorentz force is also given in terms of the vector and scalar potentials as

$$\vec{F}_{\text{Lorentz}} = q\left[-\nabla\phi - \frac{\partial\vec{A}}{\partial t} + \vec{v}\times(\nabla\times\vec{A})\right] \tag{2.95}$$

A useful feature of (2.95) is its separation of the portion of the force due to the irrotational or grad ϕ portion $(\nabla\phi)$ of the force, which is due to electrical charges, and the solenoidal portion of the force or $\vec{A}$- field portion $[-\partial\vec{A}/\partial t + \vec{v}\times(\nabla\times\vec{A})]$, which corresponds to the part that appears as magnetic or as electric force depending upon the relative velocity of the frame of reference.

2.8.2 Current Element

The concept of current element is essential in the theory of EM wave radiation since the time-varying current element is the elementary source of EM radiation. It has the same significance as the concept of a point charge in electrostatics. Due to the principle of superposition, the field radiated by a complex antenna in a linear medium can be determined by summing the fields radiated by these elementary sources (i.e., from these current elements).

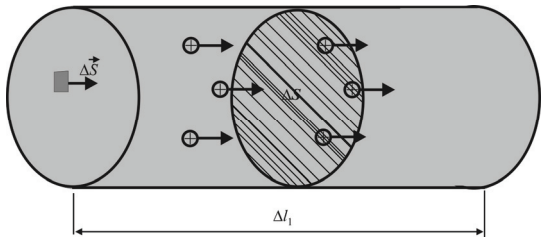


Figure 2.11 A segment of a wire carrying current.

Definition: A *current element* $I\delta l$, (A m), is a wire of length δl and current I.

Assume the existence of a small segment of a very thin wire (see Figure 2.11), where electric currents can be excited (don't worry how the currents are formed for now). The current i flowing through the wire cross-section ΔS is defined as the amount of charge passing through ΔS in 1 second:

$$i = \rho \Delta S \delta l_1 = \rho \Delta S v, \quad \text{A} \tag{2.96}$$

where:

ρ is the electric charge volume density, C m^{-3}
v is the velocity of the charges normal to the cross-section, m s^{-1}
δl_1 is the distance traveled by a charge in 1 second, m

Equation (2.96) can be also written as

$$\vec{j} = \rho \vec{v}, \quad \text{Am}^{-2} \tag{2.97}$$

where $\vec{j}$ is the electric current density. The product $\rho \Delta S$ is the charge per unit length (charge line density) along the wire. Thus, from (2.96) it follows that

$$i = v\rho_l, \quad \text{A} \tag{2.98}$$

and, palpably,

$$\frac{di}{dt} = \rho_l \frac{dv}{dt} = \rho_l a, \qquad \text{A s}^{-1} \tag{2.99}$$

where a (m s^{-2}) is the acceleration of the charge. The time derivative of a current source would then be proportional to the amount of charge q enclosed in the volume of the current element and to its acceleration:

$$\delta l \frac{di}{dt} = \delta l \rho_l a = q \cdot a, \qquad \text{A m s}^{-1} \tag{2.100}$$

It is not immediately obvious from Maxwell's equations that the time varying current is the source of a radiated EM field. However, the system of the two first-order Maxwell's equations in an isotropic medium,

$$-\nabla \times \vec{\mathcal{E}} = \mu \frac{\partial \vec{\mathcal{H}}}{\partial t} \tag{2.101}$$

$$\nabla \times \vec{\mathcal{H}} = \varepsilon \frac{\partial \vec{\mathcal{E}}}{\partial t} + \vec{j} \tag{2.102}$$

can be easily reduced to a single second-order equation either for $\vec{\mathcal{E}}$, or $\vec{\mathcal{H}}$, by taking the curl ($\nabla \times$) of both sides of (2.101). Making use of (2.102), we obtain

$$\nabla \times \nabla \times \vec{\mathcal{E}} + \mu\varepsilon \frac{\partial^2 \vec{\mathcal{E}}}{\partial t^2} = -\mu \frac{\partial \vec{j}}{\partial t} \tag{2.103}$$

From the vector wave equation (2.103), it is obvious that the time derivative of the electric currents is the source for the wave-like propagation of the vector $\vec{\mathcal{E}}$ in homogeneous and isotropic medium. Likewise, we can obtain the wave equation for the magnetic field $\vec{\mathcal{H}}$ and its sources:

$$\nabla \times \nabla \times \vec{\mathcal{H}} + \mu\varepsilon \frac{\partial^2 \vec{\mathcal{H}}}{\partial t^2} = \nabla \times \vec{j} \tag{2.104}$$

To accelerate/decelerate charges, we need sources of electromotive force and/or discontinuities of the medium in which the charges move. Such discontinuities can be bends or open ends of wires (such as in wire antennas discussed in later chapters), change in the electrical properties of the region, and so forth.

2.9 Radiation as Closed Field Lines

Maxwell's four equations describe the electric and magnetic fields arising from varying distributions of electric charges and currents and how those fields change in time. The equations were the mathematical distillation of decades of experimental observations of the electric and magnetic effects of charges and currents. Maxwell's own contribution, which he introduced in 1891, is just the last term of Ampere's law, however realizing the necessity of that term had profound and dramatic consequences. It made evident for the first time that varying electric and magnetic fields could feed off each other and these fields could propagate indefinitely through space, far from the varying charges and currents where they originated. Previously the fields had been envisioned as tethered to the charges and currents giving rise to them. Maxwell's new term (he called it the *displacement current*) freed them to move through space in a self-sustaining fashion, and even predicted their velocity—it was the velocity of light.

There are two types of antenna electric fields: the close *reactive field*, which exists between a source electric charge and a sink electric charge of opposite sign, and an unbound, *radiated electric field*, which does not [9]. An electric field is *bound* when it begins and ends on an electric charge: the positive charge where a field line begins as a "source," and the negative charge where a field line ends as a "sink" (these could be reversed with no consequence; they are defined this way by convention).

The bound fields of antennas exist close to the antenna and are responsible for the impedance (comprised of a real part and an imaginary part) as opposed to the real part only of the antenna impedance.

Radiated fields, on the other hand, form a closed loop with neither source nor sink. Such fields form closed loops within themselves and the energy within these loops closes upon itself. Figure 2.12 illustrates this point for a particular cycle of a harmonic electric dipole. When the E-field intensity is large (positive or negative), the field loops are closer together, while when close to zero they move further apart. (To say this more accurately, the field intensity is larger due to higher concentrations of energy.) Collectively the loops of energy would resemble those shown in Figure 2.13 with these loops speeding away from the antenna at the velocity of light (in a vacuum). The near field lines initiate and terminate at point charges, whereas the farther away fields are self-perpetuating, the E-field inducing the B-field, and vice versa.

Another way to envision the generation of EM waves from a sinusoidal generator and two-wire transmission line is with Figure 2.14, which shows a voltage source connected to a two-conductor transmission line. When a sinusoidal voltage is applied across the transmission line, an electric field is created within the line that is sinusoidal in nature and this results in the creation of electric

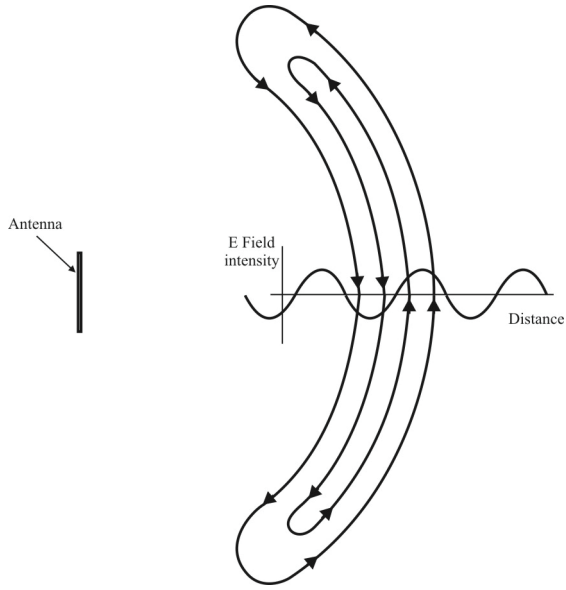


Figure 2.12 An EM wave in free space is comprised of closed loops of propagating energy away from the antenna. Such a wave travels at the speed of light in free space (vacuum), somewhat less in the atmosphere.

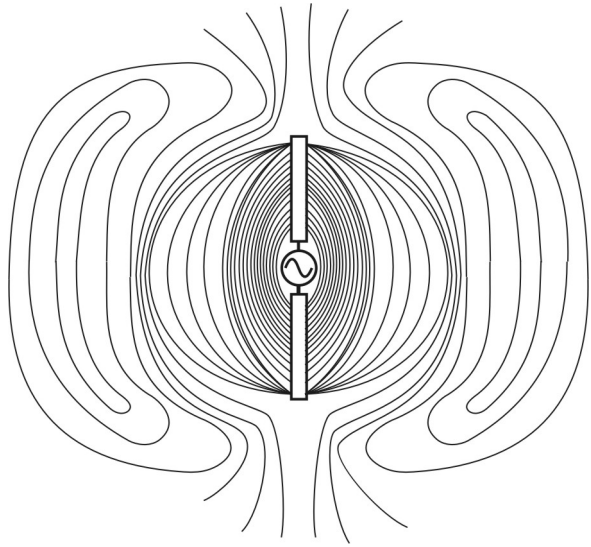


Figure 2.13 A few representative electric field lines of a dipole.

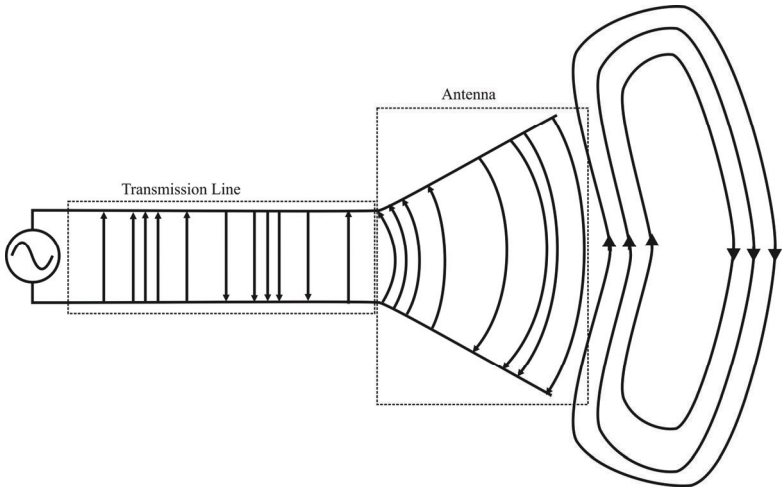


Figure 2.14 Wave generation.

lines of force that are tangential to the electric field. The magnitude of the electric field is indicated by the bunching of the electric lines of force. The free electrons on the conductors are displaced by the electric lines of force, and the movement of these charges causes the flow of current, which in turn leads to the creation of a magnetic field.

Because of the time varying electric and magnetic fields, EM waves are created and travel between the conductors. As these waves approach open space at the end of the line, free space waves are formed by connecting the open ends of the electric lines. Because the sinusoidal source continuously creates the electric disturbance, EM waves are created continuously and travel through the transmission line and through the antenna and are radiated into the free space. Inside the transmission line and the antenna, the electromagnetic waves are sustained due to the charges, but as soon as they enter the free space, they form closed loops and are radiated.

2.10 EM Waves Summary

In summary, we have shown that Maxwell's equations may be manipulated to produce 3-D wave equations for electric and magnetic fields. The solutions to these equations are called EM waves, which have the following properties:

- The waves travel through a vacuum with the speed of light c, where

$$c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}} \approx 3.00 \times 10^8 \qquad \text{m s}^{-1} \tag{2.105}$$

- Electromagnetic waves are transverse waves with $\vec{E} \perp \vec{B} \perp$ direction of propagation. That is, the electric and magnetic fields are perpendicular to each other and perpendicular to the direction of wave propagation. (Hence, electromagnetic waves are transverse waves.)
- $\vec{E}$ and $\vec{B}$ oscillate in phase, with $|E|/|B| = c$ at all times for which the fields are nonzero.
- EM waves carry energy in the direction of propagation. The Poynting vector $\vec{P}$ has magnitude equal to the rate of energy flow through a unit area and direction equal to the direction of propagation. The Poynting vector is given by

$$\vec{P} = \frac{1}{\mu_0} \vec{E} \times \vec{B} = P \vec{u}_{\text{P}} \tag{2.106}$$

- EM waves carry momentum and hence can exert a radiation pressure, P_{r}. If a wave strikes a surface that is perfectly absorbing, $P_{\text{r}} = P/c$, while if the surface is perfectly reflective, $P_{\text{r}} = 2P/c$.
- The electric field and the magnetic field each satisfy a wave equation. These two wave equations, which can be obtained from Maxwell's third and fourth equations, are

$$\frac{\partial^2 \vec{E}}{\partial x^2} = \mu_0 \varepsilon_0 \frac{\partial^2 \vec{E}}{\partial t^2} \tag{2.107}$$

$$\frac{\partial^2 \vec{B}}{\partial x^2} = \mu_0 \varepsilon_0 \frac{\partial^2 \vec{B}}{\partial t^2} \tag{2.108}$$

The electric and magnetic fields of a sinusoidal plane electromagnetic wave propagating in the positive x direction can be written as

$$E = E_{\max} \cos(\beta x - \omega t) \tag{2.109}$$

$$B = B_{\max} \cos(\beta x - \omega t) \tag{2.110}$$

where ω is the angular frequency of the wave and $\beta = 2\pi / \lambda$ is the angular wave number. These equations represent special solutions to the wave equations for E and B. The wavelength and frequency of electromagnetic waves are related by

$$\lambda = \frac{c}{f} \approx \frac{3 \times 10^8}{f} \tag{2.111}$$

The average value of the Poynting vector for a plane electromagnetic wave has a magnitude

$$P_{\text{avg}} = \frac{E_{\max} B_{\max}}{2\mu_0} = \frac{E_{\max}^2}{2\mu_0 c} = \frac{c}{2\mu_0} B_{\max}^2 \tag{2.112}$$

The intensity of a sinusoidal plane electromagnetic wave equals the average value of the Poynting vector taken over one or more cycles.

The electromagnetic spectrum includes waves covering a broad range of wavelengths, from long radio waves at more than 10^4 m to gamma rays at less than 10^{-14} m.

2.11 Concluding Remarks

We presented the fundamentals of EM wave radiation and propagation characteristics in this chapter. We started with Maxwell's equations and then described the flow of power in an EM wave with the Poynting vector. The first principles presented in this chapter form the foundations for all of the antenna properties to be discussed.

References

[1] Maxwell, J. C., "On Physical Lines of Force," *Philosophical Magazine and Journal of Science*, March 1861, pp. 12–24, 85–95, 163–175, 281–291, 338–349.

[2] Poisel, R. A., *Introduction to Communication Electronic Warfare Systems*, 2nd ed., Norwood, MA: Artech House, 2008, Ch. 4.

[3] Poisel, R. A., *Modern Communication Jamming Principles and Techniques*, 2nd ed, Norwood, MA: Artech House, 2011, Ch. 5.

[4] Jordan, E. C., and K. G. Balman, *Electromagnetic Waves and Radiating Systems*, Englewood Cliffs, NJ: Prentice Hall, 1968, p. 10.

[5] Hayt, W. H., Jr., *Engineering Electromagnetic*, New York: McGraw-Hill, 1967, p. 79.
[6] Jordan, E. C., and K. G. Balman, *Electromagnetic Waves and Radiating Systems*, Englewood Cliffs, NJ: Prentice Hall, 1968, p. 12.
[7] http://www.wisegeek.com/what-are-maxwells-equations.htm.
[8] Poynting, J. H., "On the Transfer of Energy in the Electromagnetic Field," *Philosophical Transactions Royal Society of London*, Vol. 175, 1884, pp. 343–361.
[9] Schantz, H., *The Art and Science of Ultrawideband Antennas*, Norwood, MA: Artech House, 2005, pp. 127–135.

Chapter 3

Fundamental Antenna Properties

3.1 Introduction

Before delving into the characteristics of the individual antenna types used in systems, we first present in this chapter some of the properties that all antennas possess. It serves as an introduction to the terminology used in the subsequent chapters.

Frequently, antennas must be broadband, because the frequencies of interest can be located over broad portions of the frequency spectrum. Limited real estate on these systems normally precludes using several antennas, each designed for a relatively narrow frequency band. Thus antennas that are effective over a wide frequency range are desired [1–3].

Antenna design historically has been more of an art than a science. Understanding the complexity of especially the near-field components of electromagnetic waves is difficult. The far field is almost always assumed to be a plane wave—an assumption that is reasonable because EM waves propagate in a spherical fashion unless disturbed by an intervening structure. Because EW systems are rarely located close to the transmitting antenna, the characteristics of the far field are significantly more important for our purposes.

We will show in this chapter that antennas are as reciprocal. That is, they exhibit the same characteristics whether they are used for transmitting or receiving. This is generally true if the antenna is a passive device. It may not be true, however, if the antenna contains active elements such as amplifiers, which are unidirectional.

Several antenna patterns are illustrated in the chapters to follow. It should be noted, however, that the patterns shown are representative. Particular antenna types are usually employed because they have features that are useful for the intended application. The patterns shown here are intended to show some of those features. However, keep in mind that the characteristics of an antenna are strongly dependent on the dimensions employed, both the size of the antenna as well as the

size of the material (e.g., wire) used to construct the antenna, along with the operating frequency and the environment into which the antenna is deployed.

Where possible these patterns shown were obtained using the MoM algorithm [4–6]. The 2-D and 3-D antenna patterns were obtained using a version of the numerical electric code (NEC) modeling program that incorporates the MoM technique.

This chapter begins with a discussion of antenna fundamental principles of operation. Fairly simple equivalent circuits are used to analyze antennas, both transmit and receive. That is followed with consideration of some techniques for broadening the bandwidth of antennas. Then we consider the structure of the radiation patterns in the far field. Last, we consider the effects of elevating antennas above a PEC. An appendix is included that presents a proof of the reciprocity property of antennas.

3.2 Basic Antenna Principles

3.2.1 Antenna Parameter Definitions

Three of the more important characteristics of antennas are frequency response, directionality, and impedance. The first determines the bandwidth over which an antenna will radiate energy effectively. The second describes how an antenna concentrates energy in particular directions while minimizing the propagation of energy in others.

The impedance characteristics of antennas that are practical present significant design issues. Antennas used for communication and radar purposes can be tuned with appropriate circuitry at the output of the PA so that the output impedance of the amplifier conjugate matches that of the antenna. In receiving, it is frequently required that the antennas have a wide instantaneous bandwidth, precluding the ability to tune the amplifier's output. Without matching the amplifier with the cable and antenna, significant *voltage standing waves* (VSW) ensue, which causes energy to propagate back to the amplifier from the antenna, and reduces the amount of energy that is radiated. The level of these waves is indicated by the *voltage standing wave ratio* (VSWR), which is the ratio of the maximum value of the VSW to the minimum value.

Impedance characteristics for small antennas, typical for low-frequency applications, frequently have a relatively small resistive component plus a relatively large reactive component. The reactive component can be either inductive or capacitive depending on the frequency and the type of antenna. Transmission lines that interconnect a *power amplifier* (PA) output with the antenna are typified by a characteristic impedance that remains fairly constant

along its length for high-quality cable—typical values are 50Ω, 75Ω, 100Ω, and 300Ω. If the PA output is not matched to this characteristic impedance then matching components must be placed between the PA and the transmission line to effect this match. Antennas are very rarely (if ever) naturally matched to the transmission line and must be matched with extra components. Even if they are matched, that only happens at a single frequency.

The impedance characteristics are important because maximum power is transferred from a source to a load, including antennas, when the impedance of the load is a conjugate match to that of the source. That is, if the output impedance of the source is $Z_S = R_S + jX_S$, then maximum power transfer occurs when $Z_L = R_S - jX_S$. Matching the impedance at one or two frequencies or matching to within reasonable tolerances across a small band of frequencies is relatively easy to achieve. The difficulty of matching the impedances like this arises when trying to match the impedances across a wide bandwidth.

Antennas, normally being passive devices, do not increase the amount of power in a signal. They do, however, concentrate the power in certain directions, both for transmitting and receiving signals. There are active antennas, but the active components are not normally utilized to increase power levels—the PA just prior to the antenna can best accomplish that function. Active components are designed into antennas for other purposes, such as impedance matching the transmission line to the impedance of the antenna.

It should be noted that although the radiation patterns of antennas are approximately known, there are always holes and bumps in the patterns of any real antenna that must be measured. A signal may be received from a particular azimuth very well, yet rotating the antenna slightly could cause the signal to disappear totally.

The common pattern parameters are described in this section (see Figure 3.1). A *radiation lobe* is a clear peak in the radiation intensity surrounded by regions of weaker radiation intensity. The *main lobe*(*s*) (major lobe, main beam) is the radiation lobe in the direction(s) of maximum radiation, also called boresight. A *minor lobe* is any radiation lobe other than the main lobe. A *side lobe* is a radiation lobe in any direction other than the direction(s) of intended radiation or in the direction to the reverse of the main lobe(s). The *back lobe* is the radiation lobe in the direction opposite to the main lobe(s). The *half-power beamwidth* (HPBW) is the angular width of the main beam at the half-power points (–3 dB). The *first null beamwidth* (FNBW) is the angular width between the first nulls on either side of the main beam(s).

The ability of an antenna to concentrate its energy in a particular direction is referred to as its *directivity*. This directivity is reflected in the gain pattern of an antenna, denoted as *G*, which is a function of the azimuth and elevation angles of

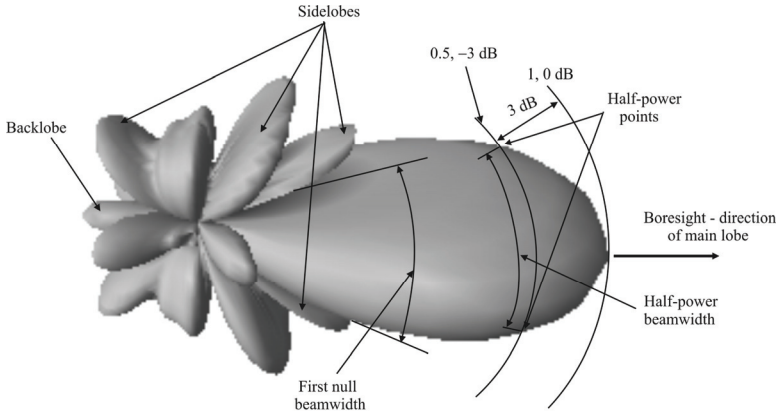


Figure 3.1 Antenna parameters.

interest relative to some physical parameter of the antenna – usually the direction of maximum gain [7].

3.2.2 Radiation Patterns

An antenna radiation pattern is a 2-D or 3-D graphical depiction of the antenna radiation properties as a function of position in Euclidean 3-space. Polar plots are also used to display the radiation pattern. The two most common types of antenna patterns are:

Power Pattern - normalized power versus 2-D or 3-D position.
Field Pattern - normalized $\left|\vec{E}\right|$ or $\left|\vec{H}\right|$ versus 2-D or 3-D position.

These patterns are typically normalized to the maximum value of the pattern so the maximum value of the pattern is equal to 1. Many of the pattern parameters can then be expressed as decibels relative to this value.

3.2.2.1 Antenna Field Types

A (transmit) antenna generates electric (E) and magnetic (H) fields. There are two parts to these fields. The *reactive field* is the component of the antenna fields comprised of stationary, standing waves that represent stored energy. These fields exist close to the antenna. The *radiation field* is the component of the fields that represent transmitted energy.

3.2.2.2 Antenna Patterns

An antenna pattern can be an isotropic pattern, a directional pattern, or an omnidirectional pattern. An *isotropic pattern* is an antenna pattern in which the radiation is uniform in all directions. A *directional pattern* is a pattern characterized by more radiation in some directions than others. All realizable antennas are directional antennas. An *omnidirectional pattern* is a pattern that is uniform in a specified plane.

Alternately, a pattern can be described by its E-plane or H-plane. These principal *plane patterns* are the E-plane and H-plane patterns of a linearly polarized antenna (polarization is discussed shortly). The *E-plane* is the plane containing the electric field vector in the direction of maximum radiation. The *H-plane* is the plane containing the magnetic field vector in the same direction.

Omnidirectional antennas are frequently used in communication EW systems due to the nature of the communication EW mission—the direction to target emitters is not always known and so must be assumed to be anywhere.

3.2.3 Average Power Radiated by an Antenna

The time-average vector power density emitted by an antenna is given by the Poynting vector given by (2.61) repeated here for convenience

$$\vec{P}_{\text{den}} = \frac{1}{2}\text{Re}\{\vec{E}\times\vec{H}^*\} = P_{\text{den}}\vec{u}_{\text{r}}, \quad \text{W m}^{-2} \tag{3.1}$$

The total time-average power radiated by the antenna, denoted by P_{rad}, is found by integrating the time-average power density magnitude over the surface of a closed surface surrounding the antenna denoted by S:

$$P_{\text{rad}} = \oint_S P_{\text{den}}\, d\vec{s} \tag{3.2}$$

$$= \frac{1}{2}\text{Re}\left\{\oint_S \vec{E}\times\vec{H}^*\, d\vec{s}\right\}, \qquad \text{W} \tag{3.3}$$

3.2.4 Directivity

3.2.4.1 Isotropic Antenna

An isotropic antenna is a hypothetical antenna that radiates equally well in all directions. It can be shown that a physical isotropic antenna cannot exist if the far

field radiation is linearly polarized [8, 9]. It is normally thought of as a sphere, radiating outward and expanding with distance.

As we will see shortly, isotropic radiation has no preferred departure point from an isotropic antenna. Most antennas, and all real antennas, do not have this property—they have preferred directions in which they radiate, and, similarly, they have preferred directions in which they radiate less than other directions.

Whereas there is no real manifestation of isotropic antennas, they form a convenient standard against which other antennas can be compared. When the gain (defined below) of an antenna is expressed in decibels relative to an isotropic antenna, it is indicated as dBi.

As opposed to isotropic antennas, isotropic radiation is (theoretically) possible. Many types of antennas, including dipoles and monopoles, exhibit approximately isotropic radiation in one of the fundamental radiation planes. As mentioned, this is referred to as omnidirectional radiation. This will be seen in more detail in Chapters 5–7.

3.2.4.2 Radiation Intensity

The radiation intensity, expressed as $U(\phi, \theta)$, is the power radiated by an antenna per unit solid angle. It is a far-field parameter and therefore is independent of r, the distance from the antenna, but is a function of ϕ and θ. It is defined as

$$U(\phi,\theta) = r^2 P_{\text{den}}(\phi,\theta) \qquad \text{W/unit solid angle} \tag{3.4}$$

when $P_{\text{den}}(\phi,\theta)$ is expressed in W m^{-2}.

3.2.4.3 Directivity

Most antennas have preferred directions for radiating or receiving energy [10–12]. This preference is referred to as *directivity* and is denoted by D. The directivity at coordinates (x_0, y_0, z_0) or (r, ϕ, θ), where these coordinates are defined in Figure 2.1, of an antenna is the ratio of the power density at that coordinate relative to the power density at the same coordinates due to an isotropic antenna radiating the same total power. Thus, using:

$$D(\phi,\theta) = \frac{U(\phi,\theta)}{U_{\text{avg}}} = \frac{4\pi U(\phi,\theta)}{P_{\text{rad}}} \tag{3.5}$$

The maximum directivity is defined as $D_0 = D(\phi,\theta)\big|_{\max}$ and the directivity range for any antenna is $0 \le D(\phi,\theta) \le D_0$. Note that the directivity of an isotropic radiator is $D(\theta,\phi) = 1$.

It is convenient to define the radiation *intensity pattern function* $F(\phi, \theta)$ as

$$U(\phi,\theta) \triangleq B_0 F(\phi,\theta) \tag{3.6}$$

where B_0 is a constant. All of the directivity parameters are included in $F(\phi, \theta)$ with those parameters not affecting the directional characteristics included in B_0. The directivity then becomes

$$D(\phi,\theta) = 4\pi \frac{F(\phi,\theta)}{\int_0^{2\pi}\int_0^{\pi} F(\phi,\theta)\sin\phi d\phi d\theta} \tag{3.7}$$

The maximum directivity is

$$D_0 = D(\phi,\theta)\big|_{\max} = 4\pi \frac{F(\phi,\theta)\big|_{\max}}{\int_0^{2\pi}\int_0^{\pi} F(\phi,\theta)\sin\phi d\phi d\theta} = \frac{4\pi}{\Omega_A} \tag{3.8}$$

where the term Ω_A in (3.8) is the *beam solid angle* and is defined by

$$\Omega_A \triangleq \frac{\int_0^{2\pi}\int_0^{\pi} F(\phi,\theta)\sin\phi d\phi d\theta}{F(\phi,\theta)\big|_{\max}} = \int_0^{2\pi}\int_0^{\pi} F_n(\phi,\theta)\sin\phi d\phi d\theta, \quad \text{rad}^2 \tag{3.9}$$

where the normalized intensity pattern function is given by

$$F_n(\phi,\theta) = \frac{F(\phi,\theta)}{F(\phi,\theta)\big|_{\max}} \tag{3.10}$$

In words, the beam solid angle is the solid angle through which all of the antenna power would flow if the radiation intensity were $U(\phi,\theta)\big|_{\max}$ for all angles in Ω_A.

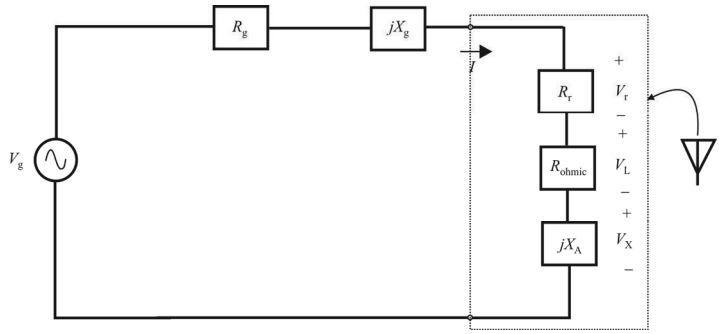


Figure 3.2 Typical transmitting system with explicit representation of the impedances involved.

3.2.5 Antenna Impedance

The complex antenna impedance is comprised of the resistive (real) and reactive (imaginary) components:

$$Z_A = R_A + jX_A, \quad \Omega \tag{3.11}$$

where R_A represents the antenna resistance that is comprised of the dissipation (ohmic) losses and the radiation resistance, and X_A represents the antenna reactance where energy is stored in the near field.

The antenna resistance is the sum of two resistances that separately represent the ohmic losses and the radiation:

$$R_A = R_r + R_{ohmic} \tag{3.12}$$

where R_r represents the radiation resistance and R_{ohmic} represents the ohmic loss.

3.2.6 Transmitting Antenna

The complex power associated with any element in the equivalent circuit shown in Figure 3.2 is given by

$$P = V_{rms} I_{rms}^* = \frac{1}{2} V_{peak} I_{peak}^*, \quad \text{W} \tag{3.13}$$

The peak current for the equivalent circuit shown in Figure 3.2 is

$$I = \frac{V_g}{Z_g + Z_A} = \frac{V_g}{(R_g + R_r + R_{ohmic}) + j(X_g + X_A)} \tag{3.14}$$

The power radiated by the antenna (P_{rad}) may be written as

$$P_{rad} = \frac{1}{2} V_r I^* = \frac{1}{2} (IR_r) I^* = \frac{1}{2} |I|^2 R_r \tag{3.15}$$

but

$$|I| = \frac{|V_g|}{\sqrt{(R_g + R_r + R_{ohmic})^2 + (X_g + X_A)^2}} \tag{3.16}$$

so

$$P_{rad} = \frac{|V_g|^2 R_r}{2[(R_g + R_r + R_{ohmic})^2 + (X_g + X_A)^2]} \tag{3.17}$$

The power dissipated as heat (P_L), which represents a loss, can be written as

$$P_L = \frac{1}{2} V_L I^* = \frac{1}{2} (IR_{ohmic}) I^* = \frac{1}{2} |I|^2 R_{ohmic} \tag{3.18}$$

$$= \frac{|V_g|^2 R_{ohmic}}{2[(R_g + R_r + R_{ohmic})^2 + (X_g + X_A)^2]}, \quad \text{W} \tag{3.19}$$

Similarly, the reactive power (imaginary component of the complex power) stored in the antenna near field (P_X) is

$$P_X = \frac{j|V_g|^2 X_A}{2[(R_g + R_r + R_{ohmic})^2 + (X_g + X_A)^2]} \tag{3.20}$$

Maximum power is transferred from a source to a load when the impedances are conjugate matched. Therefore, for maximal power transfer

$$Z_{\mathrm{A}} = Z_g^* \tag{3.21}$$

$$R_{\mathrm{A}} = R_{\mathrm{r}} + R_{\mathrm{ohmic}} = R_{\mathrm{g}} \tag{3.22}$$

$$X_{\mathrm{A}} = -X_{\mathrm{g}} \tag{3.23}$$

The current in this case is

$$I = \frac{V_{\mathrm{g}}}{Z_{\mathrm{g}} + Z_{\mathrm{A}}} = \frac{V_{\mathrm{g}}}{2(R_{\mathrm{r}} + R_{\mathrm{ohmic}})} \tag{3.24}$$

where the power radiated by the antenna is

$$P_{\mathrm{rad}} = \frac{1}{2}|I|^2 R_{\mathrm{r}} \tag{3.25}$$

$$= \frac{|V_g|^2 R_{\mathrm{r}}}{8(R_{\mathrm{r}} + R_{\mathrm{ohmic}})^2} \tag{3.26}$$

and the power dissipated in heat (and therefore lost) is

$$P_{\mathrm{L}} = \frac{1}{2}|I|^2 R_{\mathrm{ohmic}} \tag{3.27}$$

$$= \frac{|V_g|^2 R_{\mathrm{ohmic}}}{8(R_{\mathrm{r}} + R_{\mathrm{ohmic}})^2} \tag{3.28}$$

The power available from the generator source is

$$P = \frac{1}{2} V_{\mathrm{g}} I^* == \frac{|V_g|^2}{4(R_{\mathrm{r}} + R_{\mathrm{ohmic}})^2} \tag{3.29}$$

while the power dissipated in the generator resistance is

$$P_{\mathrm{g}} = \frac{1}{2}|I|^2 R_{\mathrm{g}} = \frac{|V_{\mathrm{g}}|^2 R_{\mathrm{g}}}{8(R_{\mathrm{r}} + R_{\mathrm{ohmic}})^2} = \frac{|V_{\mathrm{g}}|^2}{8(R_{\mathrm{r}} + R_{\mathrm{ohmic}})^2} = \frac{1}{2}P \tag{3.30}$$

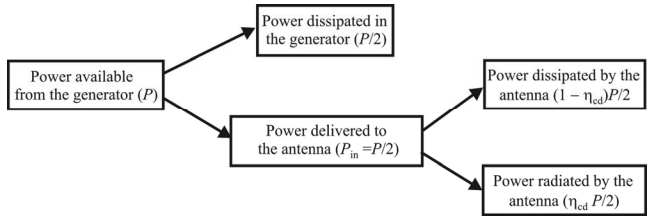


Figure 3.3 Transmit power relationships for maximum power transfer.

A summary of these expressions for maximum power transfer is shown in Figure 3.3. With an ideal transmitting antenna (perfectly efficient) given maximum power transfer, one-half of the power available from the generator is radiated by the antenna.

3.2.7 Receive Antenna

The equivalent circuit with the impedances explicitly separated is shown in Figure 3.4. Again, maximum power transfer occurs when the load is conjugate matched to the generator (antenna) impedance, which means

$$Z_A = Z_{rec}^* \tag{3.31}$$

$$R_A = R_r + R_{ohmic} = R_{rec} \tag{3.32}$$

$$X_A = -X_{rec} \tag{3.33}$$

The resulting current is

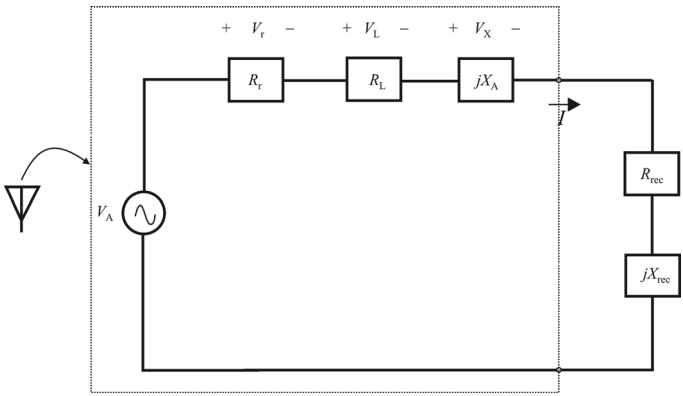


Figure 3.4 Receiving system with impedances explicitly shown.

$$I = \frac{V_{\mathrm{A}}}{Z_{\mathrm{g}} + Z_{\mathrm{A}}} = \frac{V_{\mathrm{A}}}{2(R_r + R_{\mathrm{ohmic}})} \tag{3.34}$$

So the power captured by the receiving antenna is

$$P = \frac{1}{2} V_{\mathrm{A}} I^* = \frac{1}{2} V_{\mathrm{A}} \frac{V_A^*}{2(R_r + R_{\mathrm{ohmic}})} = \frac{|V_{\mathrm{A}}|^2}{4(R_{\mathrm{r}} + R_{\mathrm{ohmic}})} \tag{3.35}$$

Some of the power captured by the receiving antenna is reradiated (scattered). The power scattered by the antenna (P_{scat}) is determined by the voltage across the radiation resistance R_{r} and is given by

$$P_{\mathrm{scat}} = \frac{1}{2} V_{\mathrm{r}} I^* = \frac{|V_{\mathrm{A}}|^2 R_{\mathrm{r}}}{8(R_{\mathrm{r}} + R_{\mathrm{ohmic}})^2} = \eta_{\mathrm{cd}} \frac{P}{2} \tag{3.36}$$

The captured power dissipated by the receiving antenna in the form of heat is caused by the ohmic losses represented by R_{L} and is given by

$$P_{\mathrm{L}} = \frac{1}{2} V_{\mathrm{L}} I^* = \frac{|V_{\mathrm{A}}|^2 R_{\mathrm{ohmic}}}{8(R_{\mathrm{r}} + R_{\mathrm{ohmic}})^2} = (1 - \eta_{\mathrm{cd}}) \frac{P}{2} \tag{3.37}$$

The remaining power is the useful power delivered to the receiver and is given by

$$P_{\mathrm{rec}} = \frac{1}{2} (I R_{\mathrm{rec}}) I^* = \frac{|V_{\mathrm{A}}|^2 R_{\mathrm{rec}}}{8(R_{\mathrm{r}} + R_{\mathrm{ohmic}})^2} = \frac{|V_{\mathrm{A}}|^2}{8(R_{\mathrm{r}} + R_{\mathrm{ohmic}})} = \frac{P}{2} \tag{3.38}$$

since the receiver is conjugate matched to the antenna.

These relationships are summarized in Figure 3.5 for maximum power transfer at the receiving system. With an ideal receiving antenna ($\eta_{\mathrm{cd}} = 1$), given maximum power transfer, one-half of the power captured by the antenna is reradiated and one-half is useful in the receiver.

3.2.8 Antenna Reciprocity

We show in Appendix 3A that passive antennas are reciprocal. This means than the antenna has the all same properties described in this chapter (e.g., radiation

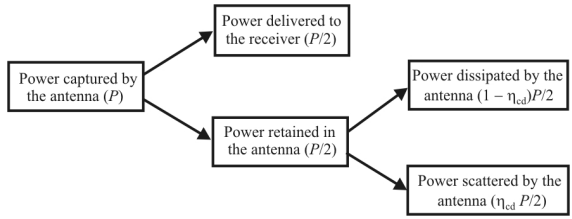


Figure 3.5 Power transfer relationships at the receiving antenna system.

pattern, impedance bandwidth, and so forth) whether it is used for transmitting or receiving.

This property does not extend to active antennas, however. Making an antenna active, that is, incorporating active components, such as amplifiers, into the antenna structure can change its characteristics so that receiving and transmitting characteristics are not equivalent.

3.2.9 Antenna Loss Resistance

The antenna loss resistance (conductor and dielectric losses) for many antennas is typically difficult to calculate. In these cases, the loss resistance is normally measured experimentally. However, the loss resistance of wire antennas can be calculated easily and accurately.

Assuming a conductor of length L and cross-sectional area A that carries a uniform current density, the DC resistance is

$$R_{\mathrm{DC}} = \frac{l}{\sigma A} \tag{3.39}$$

where σ is the *conductivity* of the conductor (Siemens m^{-1} or mhos m^{-1}). Conductivities for some common materials are given in Table 3.1.

Table 3.1 Some Electrical Conductivities

Material	Conductivity (S m^{-1})	Temperature (^{o}C)
Silver	63.0×10^6	20
Copper	59.6×10^6	20
Annealed copper	58.0×10^6	20
Aluminum	$37.8. \times 10^6$	20
Sea water	5	
Drinking water	0.0005 to 0.05	
De-ionized water	5.5×10^{-6}	

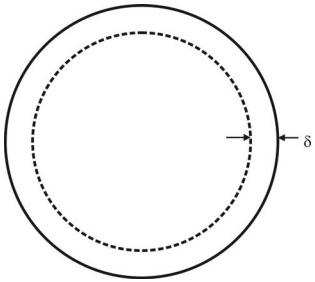


Figure 3.6 Skin depth given by δ.

At high frequencies, the current tends to crowd toward the outer surface of the conductor due to the *skin effect*. The high frequency (HF) resistance can be defined in terms of the *skin depth* δ:

$$\delta = \frac{l}{\sqrt{\pi f \mu \sigma}}, \qquad \text{m} \tag{3.40}$$

where μ is the permeability of the material and f is the frequency in hertz. The skin depth is illustrated in Figure 3.6 for a circular conductor.

If we define the perimeter distance of the conductor as d_{p}, then the HF resistance of the conductor can be written as

$$R_{\text{HF}} = \frac{l}{\sigma d_{\text{p}} \delta} = \frac{l}{d_{\text{p}}} R_{\text{S}} \tag{3.41}$$

where R_{S} is defined as the surface resistance of the material given by

$$R_{\text{S}} = \frac{l}{\sigma \delta} = \sqrt{\frac{\pi f \mu}{\sigma}} \tag{3.42}$$

For (3.41) to be accurate, the skin depth must be a small fraction of the conductor maximum cross-sectional dimension, which, as illustrated in Figure 3.7 for $f < 600$ Hz and Figure 3.8 for 1 kHz $< f <$ 200 kHz, is accurate at frequencies above a few hundred kilohertz. In the case of a cylindrical conductor $(d_{\text{p}} = 2\pi a_0)$, the HF resistance is

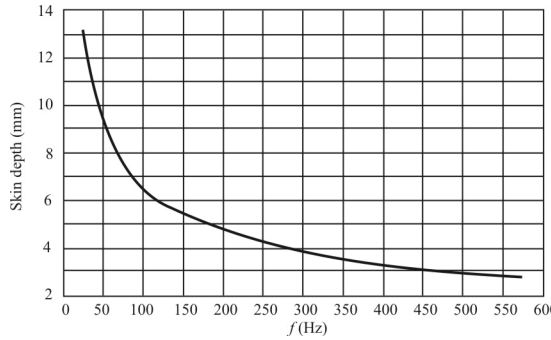


Figure 3.7 Skin depth in copper conductor for $f < 600$ Hz.

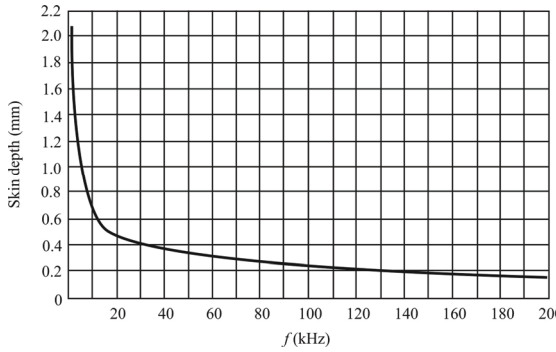


Figure 3.8 Skin depth in copper conductor for 1 kHz $< f <$ 200 kHz. Skin depth is a function of the size of the wire too (see Table A.2 in Appendix A).

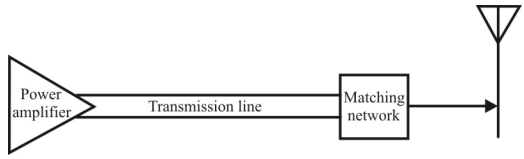


Figure 3.9 System output components.

$$R_{\mathrm{HF}} = \frac{l}{2\pi a_0}\sqrt{\frac{\pi f \mu}{\sigma}} = \frac{l}{2a_0}\sqrt{\frac{f\mu}{\pi\sigma}} \tag{3.43}$$

Equation (3.43) assumes that the current through the conductor is sinusoidal in time and independent of position along the conductor so that

$$I_z(z,t) = I_0 \cos(\omega t) \tag{3.44}$$

Most of the time the current is not independent of position. However, given the actual current distribution on the antenna, an equivalent R_{L} can be calculated.

3.2.10 Antenna Efficiency

When an antenna is driven by a voltage source (generator), the total power radiated by the antenna will not be the total power available from the generator. The loss factors that affect the antenna efficiency can be identified by considering the common example of a generator connected to a transmitting antenna via a transmission line as shown in Figure 3.9. The losses are due to the effects of mismatch at the generator/transmission line connection, losses in the transmission line, the mismatch at the transmission line/antenna connection, and ohmic losses in the antenna structure itself.

3.2.10.1 Antenna Radiation Efficiency

The *antenna radiation efficiency* (η_{cd}) is defined as the ratio of the power radiated from the antenna to the total power delivered to the antenna:

$$\eta_{\mathrm{cd}} = \frac{\mathrm{P}_{\mathrm{rad}}}{P_{\mathrm{in}}} = \frac{P_{\mathrm{rad}}}{P_{\mathrm{rad}} + P_{\mathrm{ohmic}}} \tag{3.45}$$

The total radiated power and the total ohmic losses are determined for the general case of a transmitting antenna using the equivalent circuit. The total radiated power is that "dissipated" in the antenna radiation resistance (R_r) as given by (3.26).

The total ohmic losses for the antenna are those dissipated in the antenna loss resistance (R_{ohmic}) as given by (3.28). Inserting (3.26) and (3.28) into (3.45) yields

$$\eta_{cd} = \frac{R_r}{R_r + R_{ohmic}} \tag{3.46}$$

Thus, the antenna radiation efficiency may be found directly from the antenna equivalent circuit parameters.

The total power delivered to the antenna terminals, P_{in} in Figure 3.3, must equal that lost to I^2R (ohmic) losses plus that radiated by the antenna:

$$P_{in} = P_{rad} + P_{ohmic} \tag{3.47}$$

The antenna radiation efficiency may be written as a product of the *conduction efficiency* (η_c) and the *dielectric efficiency* (η_d):

$$\eta_{cd} = \eta_c \eta_d \tag{3.48}$$

where η_c is the conduction efficiency which includes only conduction losses and η_d is the dielectric efficiency which includes only the dielectric losses.

3.2.10.2 Total Antenna Efficiency

We can define the *total antenna efficiency* (η_T), which includes the losses due to mismatch between the transmission line and the antenna as

$$\eta_T = \eta_r \eta_c \eta_d \tag{3.49}$$

where η_T is the total efficiency, which includes all losses, and η_r is the reflection efficiency, which represents the mismatch losses. The reflection efficiency represents the ratio of power delivered to the antenna terminals to the total power incident on the transmission line/antenna connection. The reflection efficiency is found from transmission line theory in terms of the *reflection coefficient* Γ:

$$\eta_r = 1 - |\Gamma|^2 \tag{3.50}$$

$$\Gamma = \frac{Z_A - Z_0}{Z_A + Z_0} \tag{3.51}$$

The total antenna efficiency then becomes

$$\eta_T = \eta_{cd}(1 - |\Gamma|^2) \tag{3.52}$$

3.2.11 Antenna Gain

The antenna directivity and antenna gain are basically the same concept, except the gain takes the antenna efficiency into account while the directivity does not. The directivity $D(\theta, \phi)$ is the ratio of the antenna radiated power density at a distant point to the *total antenna radiated power* (P_{rad}) radiated isotropically, while the gain of an antenna, $G(\theta, \phi)$, is the ratio of the antenna radiated power density at a distant point to the *total antenna input power* (P_{in}) radiated isotropically. Thus, the antenna gain accounts for the ohmic losses in the antenna, while the antenna directivity, being dependent on the total radiated power, does not.

The directivity is given by (3.5), while the gain is given by

$$G(\phi, \theta) = \frac{U(\phi, \theta)}{P_{in} / 4\pi}, \quad \text{unitless} \tag{3.53}$$

The relationship between the directivity and gain of an antenna may be found using the definition of the radiation efficiency of the antenna, (3.46), yielding

$$G(\phi, \theta) = \eta_{cd} D(\phi, \theta) \tag{3.54}$$

3.2.12 Antenna Polarization

The *polarization* of a plane wave is defined by the figure traced by the instantaneous electric field at a fixed observation point. The most commonly encountered polarizations in communication and radar systems, assuming the wave is approaching out of the page, are illustrated in Figure 3.10.

The polarization of the antenna in a given direction is defined as the polarization of the wave radiated in that direction by the antenna. Note that any of the polarization figures in Figure 3.10 may be rotated by some arbitrary angle.

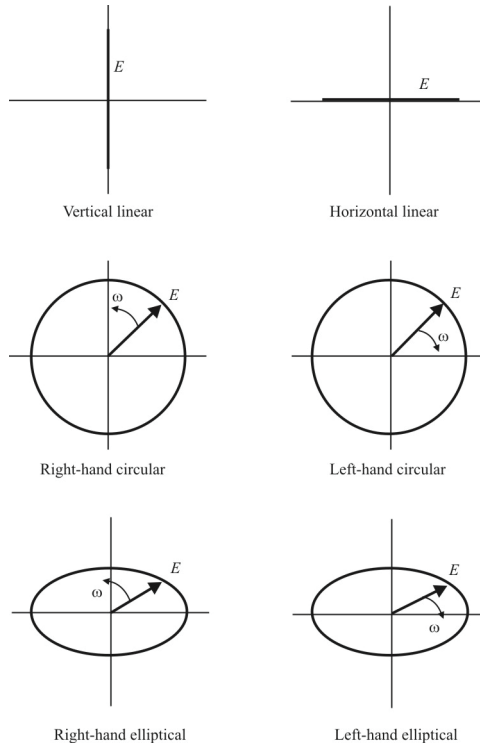


Figure 3.10 Common polarizations. The EM wave is assumed to be coming out of the page.

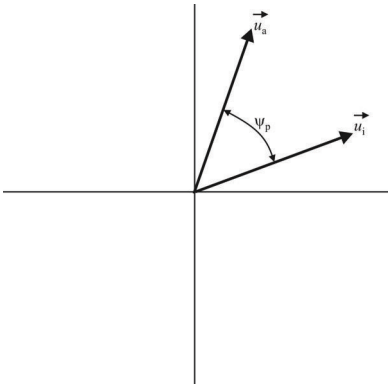


Figure 3.11 Angle between the two polarization unit vectors.

3.2.12.1 Polarization Loss Factor

Denote the incident wave as

$$\vec{E}^i = E^i \vec{u}_i \tag{3.55}$$

with polarization defined by the unit vector $\vec{u}_i$. Further denote the polarization (orientation really) of the antenna as $\vec{u}_a$, with response

$$\vec{E}^a = E^a \vec{u}_a \tag{3.56}$$

The *polarization loss factor* (PLF) is then given by the dot product of these two unit vectors:

$$\text{PLF} = \left|\vec{u}_i \bullet \vec{u}_a\right|^2 = \left|\cos\psi_p\right|^2 \tag{3.57}$$

where ψ_p is the angle between the two polarization unit vectors illustrated in Figure 3.11. We note that if the polarizations are totally mismatched the received signal would be nil.

3.2.12.2 General Polarization Ellipse

The vector electric field associated with a +*z*-directed plane wave can be written in general phasor form as (see Figure 3.12)

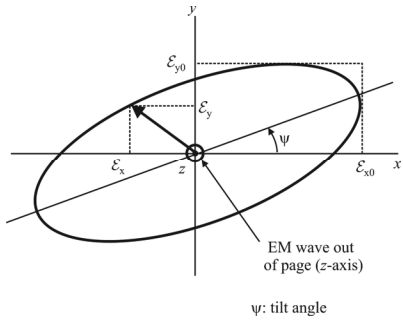


Figure 3.12 The general polarization ellipse.

$$\vec{E} = (E_x \vec{u}_x + E_y \vec{u}_y) e^{-j\beta z} \tag{3.58}$$

where $\beta = 2\pi/\lambda$, and E_x and E_y are complex phasors that may be defined in terms of magnitude and phase as

$$E_x = E_{x0} e^{j\varphi_x} \qquad\qquad E_y = E_{y0} e^{j\varphi_y} \tag{3.59}$$

The instantaneous components of the electric field are found by finding the real part of the product of the phasor components and $e^{j\omega t}$:

$$\mathcal{E}_x(z,t) = \mathrm{Re}\{E_{x0} e^{j\varphi_x} e^{-j\beta z} e^{j\omega t}\} = E_{x0} \cos(\omega t - \beta z + \varphi_x) \tag{3.60}$$

$$\mathcal{E}_y(z,t) = \mathrm{Re}\{E_{y0} e^{j\varphi_y} e^{-j\beta z} e^{j\omega t}\} = E_{y0} \cos(\omega t - \beta z + \varphi_y) \tag{3.61}$$

Linear Polarization

If we define the phase difference between the two electric field components as

$$\Delta\varphi = \varphi_y - \varphi_x \tag{3.62}$$

we find that a phase difference of

$$\Delta\varphi = n\pi, \qquad n = 0, 1, 2, \ldots \tag{3.63}$$

defines a linearly polarized wave:

$$\mathcal{E}_x(z,t) = E_{x0}\cos(\omega t - \beta z + \varphi_x) \tag{3.64}$$

$$\mathcal{E}_y(z,t) = E_{y0}\cos(\omega t - \beta z + \varphi_x + n\pi) = \pm E_{y0}\cos(\omega t - \beta z + \varphi_x)\begin{Bmatrix} n\,\text{even} \\ n\,\text{odd} \end{Bmatrix} \tag{3.65}$$

Some examples of linear polarization are:

- If $E_{y0} = 0$, then we have linear polarization in the *x*-direction ($\psi = 0$).
- If $E_{x0} = 0$, then we have linear polarization in the *y*-direction ($\psi = 90^o$).
- If $E_{x0} = E_{y0}$ and *n* is even, then we have linear polarization with $\psi = 45^o$.
- If $E_{x0} = E_{y0}$ and *n* is odd, then we have linear polarization with $\psi = 135^o$.

Circular Polarization

If $E_{x0} = E_{y0}$ and

$$\Delta\varphi = (2n+1/2)\pi \qquad n = 0,1,2,\ldots \tag{3.66}$$

then

$$\mathcal{E}_x(z,t) = E_{x0}\cos(\omega t - \beta z + \varphi_x) \tag{3.67}$$

$$\mathcal{E}_y(z,t) = E_{x0}\cos(\omega t - \beta z + \varphi_x + (2n+1/2)\pi) = -E_{x0}\sin(\omega t - \beta z + \varphi_x) \tag{3.68}$$

The E-fields and therefore the polarization oscillate at the radiation frequency with some initial phase offsets. This is left-hand circular polarization.

The terms left-hand and right-hand refer to the direction of propagation and the rotation of the field. When your thumb is pointed in the direction of propagation of the field, if the fingers of the right hand point in the direction of rotation, then it is right-hand polarization, likewise for left-hand polarization.

If $E_{x0} = E_{y0}$ and

$$\Delta\varphi = -(2n+1/2)\pi \qquad n = 0,1,2,\ldots \tag{3.69}$$

then

$$\mathcal{E}_x(z,t) = E_{x0}\cos(\omega t - \beta z + \varphi_x) \tag{3.70}$$

$$\mathcal{E}_y(z,t) = E_{x0}\cos(\omega t - \beta z + \varphi_x - (2n+1/2)\pi) = E_{x0}\sin(\omega t - \beta z + \varphi_x) \quad (3.71)$$

This is right-hand circular polarization.

Elliptical Polarization

Elliptical polarization employs the same definitions as circular polarization except that $E_{x0} \neq E_{y0}$:

- If $E_{x0} \neq E_{y0}$ and $\Delta\varphi = (2n+1/2)\pi$, then we have left-hand elliptical polarization.
- If $E_{x0} \neq E_{y0}$ and $\Delta\varphi = -(2n+1/2)\pi$, then we have right-hand elliptical polarization

With elliptical polarization the E-fields rotate in orientation and trace an ellipse in space at any given point.

3.2.13 Antenna Effective Area

The antenna *effective area* (A_{eff}) is defined according to

$$P_{rec} \triangleq P_{den} A_{eff}, \qquad \text{W} \quad (3.72)$$

where P_{den} is the amplitude of the power density vector of the incident wave, $\vec{P}_{den}$ (magnitude of the Poynting vector). According to the equivalent circuit under matched conditions as illustrated in Figure 3.5, the reactive components cancel and

$$R_A \triangleq R_r + R_L = R_{rec} \quad (3.73)$$

then

$$P_{rec} = \frac{1}{2} V_{rec} I^* = \frac{1}{2}\frac{V_A}{2}\left(\frac{V_A}{2R_A}\right)^* = \frac{|V_A|^2}{8R_A} \quad (3.74)$$

From (3.72), the antenna effective area is

$$A_{\text{eff}} = \frac{P_{\text{rec}}}{P_{\text{den}}} = \frac{|V_{\text{A}}|^2}{8P_{\text{den}}R_{\text{A}}} = \frac{|V_{\text{A}}|^2}{8P_{\text{den}}(R_{\text{r}} + R_{\text{L}})}, \qquad \text{m}^2 \tag{3.75}$$

Note that A_{eff} is not the same as the physical area of the antenna—it typically is between 40% and 70% of the physical area.

3.2.13.1 Maximum Directivity and Effective Aperture

Assume the transmitting and receiving antennas are lossless, oriented for maximum response, and that they are separated by distance *r*. The receive antenna effective aperture and maximum directivity are given by $A_{\text{eff,R}}$ and $D_{0,\text{R}}$ respectively; likewise, for the transmitter, $A_{\text{eff,T}}$ and $D_{0,\text{T}}$. If we denote the total power transmitted by the transmit antenna by P_{T}, the power density at the receive antenna ($P_{\text{den,R}}$) is

$$P_{\text{den,R}} = \frac{P_{\text{T}}}{4\pi r^2} D_{0,\text{T}} \tag{3.76}$$

The total power received by the receive antenna (P_{R}) is

$$P_{\text{R}} = P_{\text{den,R}} A_{\text{eff,R}} = \frac{P_{\text{T}} D_{0,\text{T}} A_{\text{eff,R}}}{4\pi r^2} \tag{3.77}$$

which gives

$$D_{0,\text{T}} A_{\text{eff,R}} = \frac{P_{\text{R}}}{P_{\text{T}}} 4\pi r^2 \tag{3.78}$$

Assuming a linear, isotropic medium, as we will show later, if we interchange the transmit and receive antennas, (3.78) still holds true by interchanging the respective transmit and receive quantities, which gives

$$D_{0,\text{R}} A_{\text{eff,T}} = \frac{P_{\text{R}}}{P_{\text{T}}} 4\pi r^2 \tag{3.79}$$

Then (3.78) and (3.79) yield

$$D_{0,\mathrm{T}} A_{\mathrm{eff,R}} = D_{0,\mathrm{R}} A_{\mathrm{eff,T}} \tag{3.80}$$

or

$$\frac{D_{0,\mathrm{T}}}{A_{\mathrm{eff,T}}} = \frac{D_{0,\mathrm{R}}}{A_{\mathrm{eff,R}}} \tag{3.81}$$

If the transmit antenna is an isotropic radiator then

$$D_{0,\mathrm{T}} = 1 \qquad A_{\mathrm{eff,T}} = \frac{\lambda^2}{4\pi} \tag{3.82}$$

which gives

$$\frac{D_{0,\mathrm{R}}}{A_{\mathrm{eff,R}}} = \frac{4\pi}{\lambda^2} \qquad \text{(for any antenna)} \tag{3.83}$$

Therefore, the *equivalent aperture* (area) of a lossless antenna may be defined in terms of the maximum directivity as

$$A_{\mathrm{eff}} = \left(\frac{\lambda^2}{4\pi}\right) D_0 \tag{3.84}$$

The overall antenna efficiency, η_{T}, given in (3.52) may be included to account for the ohmic and mismatch losses in the antenna:

$$A_{\mathrm{eff}} = \eta_{\mathrm{T}} \left(\frac{\lambda^2}{4\pi}\right) D_0 = \eta_{\mathrm{cd}} (1 - |\Gamma|^2) \left(\frac{\lambda^2}{4\pi}\right) D_0 \tag{3.85}$$

The effect of polarization mismatch loss can also be included by incorporating the PLF in (3.57) to yield

$$A_{\mathrm{eff}} = \eta_{\mathrm{cd}} (1 - |\Gamma|^2) \left(\frac{\lambda^2}{4\pi}\right) D_0 \left|\vec{u}_{\mathrm{i}} \bullet \vec{u}_{\mathrm{R}}\right|^2 \tag{3.86}$$

From (3.76), the power density, in watts per square meter, at a distance r from an antenna in free space is given by

$$P_{\text{den}} = \frac{P_{\text{T}} G_{\text{T}}}{4\pi r^2} \tag{3.87}$$

where P_{T} is the transmitter power supplied to the antenna. This power is spread out equally by the isotropic antenna onto the surface of a sphere, the area of which is given by $4\pi r^2$. Thus, the power in any selected square meter in free space is decreasing at a rate proportional to $1 / r^2$.

The gain of an antenna is related to its effective area by

$$G = \frac{4\pi A_{\text{eff}}}{\lambda^2} \tag{3.88}$$

3.2.14 Effective Height of an Antenna

A parameter similar to the effective area of an antenna is its *effective height*. An antenna converts a propagating wave into a current or, vice versa, converts a current into an EM wave. This current is a result of or generates a voltage at the antenna terminals. The effective height, h_{e}, of a wire receive antenna relates the open circuit voltage at the antenna terminals to the incident electric field by

$$V = h_{\text{e}} E \tag{3.89}$$

where

V is the RMS voltage induced at the output of the antenna, V
E is the RMS amplitude of the electric field, V m^{-1}
h_{e} is the effective height, m

Palpably, the antenna/EM wave polarization mismatch can be taken into account by using vectors in (3.89) rather than amplitudes.

3.2.15 Antenna Resonance and Bandwidth

Two major parameters associated with an antenna are the antenna resonant point or center operating frequency and the antenna bandwidth or the frequency range over which some antenna performance parameter is acceptable. In almost all cases, these two factors are very important.

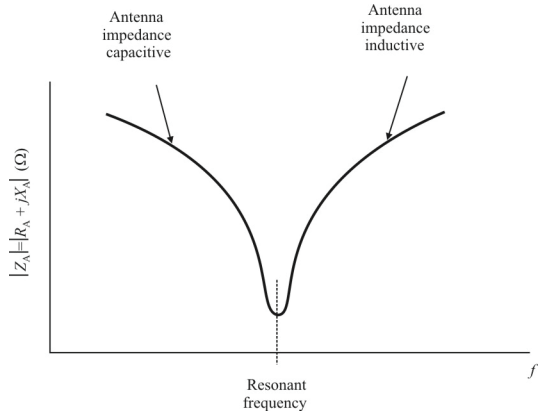


Figure 3.13 Resonant frequency of a tuned antenna.

3.2.15.1 Antenna Resonance

Most RF antennas are a form of tuned circuit consisting of inductance and capacitance, and as a result it has resonant frequencies (see Figure 3.13). This is the frequency where the capacitive and inductive reactances cancel each other. Below this frequency the antenna impedance is capacitive $(X_A < 0)$, and above resonance the antenna impedance is inductive $(X_A > 0)$. At resonance the RF antenna appears purely resistive, the resistance being a combination of the loss resistance and the radiation resistance.

The capacitance and inductance of an RF antenna are determined by its physical properties and the environment in which it is embedded. The major feature of the RF antenna design is its dimensions, both its overall size, as well as the size of the components with which it is made. We can show that the larger the antenna or, more accurately, the antenna elements, the lower the resonant frequency. For example, antennas for the UHF frequency range have relatively small elements, while those for VHF have larger elements indicating a lower frequency. Antennas for short-wave applications are larger still.

3.2.15.2 Antenna Bandwidth

Most RF antennas are operated around resonance. This means that there is only a limited bandwidth over which an RF antenna design can operate efficiently. Outside this the levels of reactance rise to levels that may be too high for

satisfactory operation. Other parameters of the antenna may also be impaired away from the center operating frequency [13, 14].

The antenna bandwidth is particularly important where radio transmitters are concerned as damage may occur to the transmitter due to signals that are reflected from the antenna back to the transmitter output if the antenna is operated outside its operating range and the radio transmitter is not adequately protected. In addition to this, the signal radiated by the RF antenna will be less for a number of reasons.

For receiving purposes the performance of the antenna is less critical in some respects. It can be operated outside its normal bandwidth without any fear of damage to the receiver. Even a random length of wire will pick up signals, and it may be possible to receive several distant stations. However, for the best reception it is necessary to ensure that the performance of the RF antenna design is optimum.

The bandwidth of an antenna is the frequency range over which some parameter is within some acceptable range. This parameter is frequently the VSWR and a measure of less than 2:1 is typical. A VSWR of 2:1 corresponds to a 10% or 9.5 dB loss. There are two generally accepted definitions of the bandwidth of an antenna depending on whether the narrowband or broadband characteristics of the antenna are under consideration.

The definitions of the two bandwidth measurements are:

- Narrowband by percent:

$$B = \frac{f_U - f_L}{f_C} \times 100 \tag{3.90}$$

where f_U is the upper frequency limit f_L is the lower frequency limit, and f_C is the center frequency.

- Broadband by ratio:

$$B = \frac{f_U}{f_L} \tag{3.91}$$

When $f_U / f_L > 2$, the antenna is considered wideband.

Impedance Bandwidth

For a tuned antenna with resonance at ω_0 the impedance at ω is given by

$$Z_0(\omega) = R_0(\omega) + jX_0(\omega) \tag{3.92}$$

and the *fractional matched VSWR bandwidth* [$\text{FBW}_V(\omega_0)$] is given by [13, 15]

$$\text{FBW}_V(\omega_0) = \frac{\omega_U - \omega_L}{\omega_0} \approx \frac{4\sqrt{\gamma}R_0(\omega_0)}{\omega_0 \left|Z_0'(\omega_0)\right|} \tag{3.93}$$

The prime in (3.93) denotes differentiation with respect to ω, and

$$\sqrt{\gamma} = \sqrt{\frac{\alpha}{1-\alpha}} = \frac{s-1}{2\sqrt{s}} \le 1 \tag{3.94}$$

when s represents the constant value of the VSWR on either side of ω_0, and

$$\alpha = \left|\Gamma_0\right|^2 = \frac{(s-1)^2}{(s+1)^2} \tag{3.95}$$

In order to increase the impedance bandwidth of an antenna, there are a number of measures that can be taken. One is the use of thicker conductors. Another is the actual type of antenna used. For example a folded dipole which is described in Chapter 5 has a wider bandwidth than a nonfolded one of similar size.

The impedance characteristics of a 1 m center-fed, untuned straight-wire antenna with diameter 1 mm is illustrated in Figure 3.14. Assuming that the antenna is tuned with a series impedance consisting of a lossless capacitor or inductor, then $R_0(\omega) = R(\omega)$. The frequencies at which $X_0(\omega_0) = 0$ is a *resonant frequency* if $X_0'(\omega_0) > 0$ and an *antiresonant frequency* if $X_0'(\omega_0) < 0$ where the prime indicates differentiation with respect to frequency. At ω_0 where $X_0(\omega_0) = 0$ for an untuned antenna, then the antenna is said to have a *natural resonant frequency* at ω_0 if $X_0'(\omega_0) > 0$ and a *natural antiresonant frequency* at ω_0 if $X_0'(\omega_0) < 0$. Thus, for the 1 m wire antenna characterized by the impedances shown in Figure 3.14, points A and C are natural resonant frequencies while point B is a natural antiresonant frequency. Note that when $X_0(\omega_0) < 0$, the antenna impedance is capacitive while when $X_0(\omega) > 0$ the antenna impedance is inductive.

The characteristic impedances of an unturned circular loop antenna with a radius of 0.348 m and wire diameter 1 mm are shown in Figure 3.15. This antenna

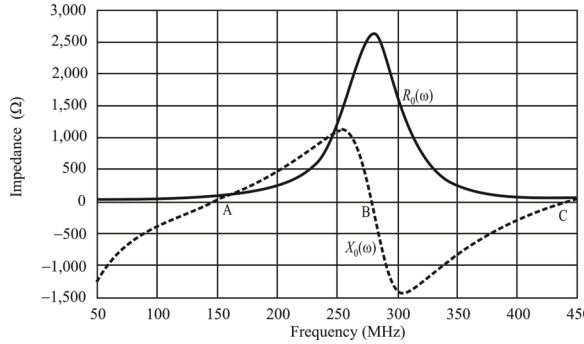


Figure 3.14 Impedance characteristics of a center-fed, untuned, straight-wire antenna with length 1 m and diameter 1 mm.

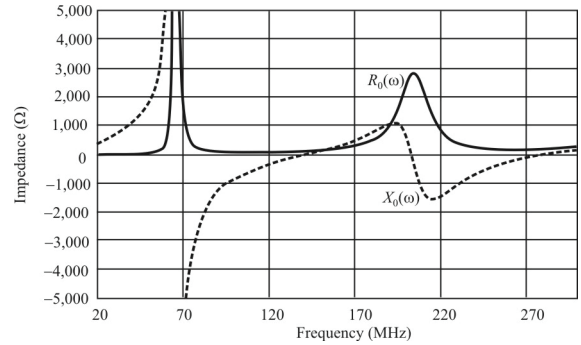


Figure 3.15 Impedance characteristics of an untuned circular loop antenna with a radius of 0.348 m and wire diameter of 1 mm.

has natural antiresonance frequencies of about 66 MHz and 200 MHz and natural resonance frequencies of 144 MHz and 288 MHz.

3.2.16 Radiation Pattern

Another feature of an antenna that changes with frequency is its radiation pattern. In the case of a beam it is particularly noticeable. In particular, the front-to-back ratio will fall off rapidly outside a given bandwidth, and so will the gain. In an antenna such as a Yagi this is caused by a reduction in the currents in the parasitic elements as the frequency of operation is moved away from resonance. For beam antennas such as the Yagi the *radiation pattern bandwidth* is defined as the frequency range over which the gain of the main lobe is within 1 dB of its maximum.

For many beam antennas, especially high gain ones, the impedance bandwidth is wider than the radiation pattern bandwidth, although the two parameters are interrelated in many respects.

3.2.17 Antenna *Q*

Definition. The Q of a linear and passive antenna at resonance is given by the ratio of the magnitude of the energy stored in the antenna, $|W(\omega_0)|$, to the total power accepted by the antenna, $P_A(\omega_0)$, multiplied by the resonant frequency.[1] That is

$$Q(\omega_0) \triangleq \frac{\omega_0 |W(\omega_0)|}{P_A(\omega_0)} \tag{3.96}$$

The approximate expression for antenna Q can be determined by [13]

$$Q(\omega_0) \triangleq \frac{2\sqrt{\gamma}}{\mathrm{FBW}_V(\omega_0)} \triangleq \frac{\omega_0 |Z_0'|}{2R_0(\omega_0)} \tag{3.97}$$

where the parameters are as defined above for (3.93). Equation (3.97) applies when $R_0'(\omega)$ and $X_0'(\omega)$ do not change much over the antenna bandwidth.

Note that the conventional definition of antenna Q is given by

[1] This definition is consistent with the definition of Q for any RLC network.

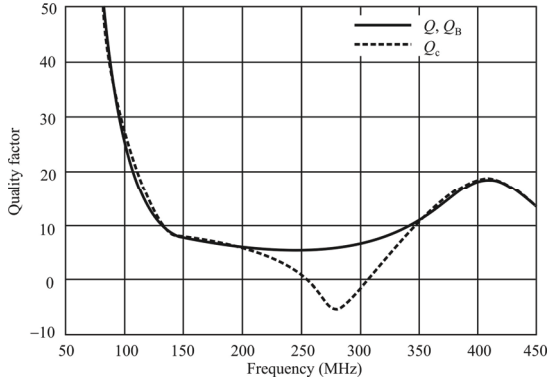


Figure 3.16 Comparison of Q with Q_c for a center-fed, tuned, lossless straight-wire antenna 1 m long with diameter 1 mm. $s = 3.5$.

$$Q_c(\omega_0) \triangleq \frac{\omega_0 X_0'}{2R_0(\omega_0)} \triangleq \frac{1}{B} \tag{3.98}$$

which is different from (3.97) by excluding R_0'. Therefore, $Q_c(\omega_0) = Q(\omega_0)$ only when $R_0' = 0$, which is, of course, when R_0 is a constant.

Defining Q as the inverse of the bandwidth is a reasonable definition because the Q is a measure of the "sharpness" of the antenna response (smaller bandwidths imply a higher Q and vice versa) and the wider B is the smaller Q becomes.

Rhodes defines the half-power bandwidth B of an antenna as [14]

$$B = \frac{2R_0}{\omega_0 \left| X_0' \right|} \tag{3.99}$$

and then defines the Q_B of that antenna to be

$$Q_B = \frac{1}{B} \tag{3.100}$$

Figure 3.16 compares Q, Q_B, and Q_c for a 1 m center-fed, tuned, lossless straight-wire antenna with 1 mm diameter. Note from Figure 3.16 that the conventional definition of the Q factor given by (3.98) does not match well the Q as determined by (3.97) near the natural antiresonance frequencies of the antenna.

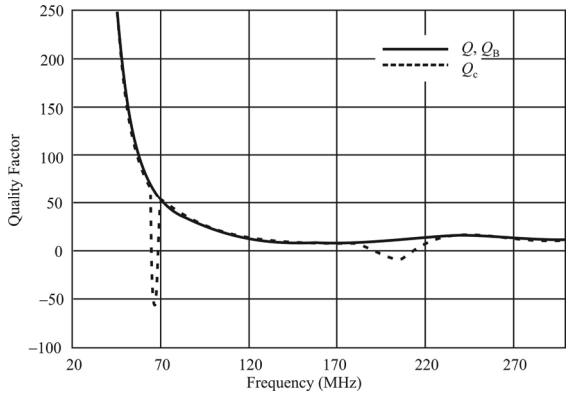


Figure 3.17 Quality factor comparison for a tuned loop antenna with 3.5:1 matched VSWR bandwidth, 0.348 m radius with 1 mm wire diameter. $s = 3.5$.

The quality factors for the circular loop antenna are compared in Figure 3.17. Again, the conventional definition of Q does not yield very accurate results near the antiresonant frequencies.

3.3 Broadband Antennas

In many applications, antennas must cover a sizable span of the frequency spectrum; the plethora of threat environments into which the EW system is to be deployed is not normally known a priori, for example. Simple antennas such as the wire dipole or monopole do not provide the bandwidth themselves, especially when they are small. We present in this section techniques and methodologies for broadening the spectrum coverage range of antennas [15–17].

3.3.1 Resistive Loading

A technique for widening the bandwidth of an electrically small antenna is with resistive loading. The concept follows the logic that reducing the antenna Q while keeping all other factors constant widens the bandwidth (it lowers the gain at the same time, however).

An equivalent circuit of an antenna is shown in Figure 3.18. It consists of two resistances: the loss resistance, R_{loss}, and the radiation resistance, R_{r}, as well as a reactance, X. For most small antennas this reactance is negative (capacitive). Now

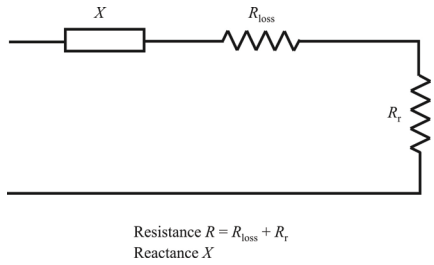


Figure 3.18 Antenna model.

consider Figure 3.19. The Q for the narrower bandwidth characteristic is approximated by

$$Q_1 = \frac{|X|}{R_r + R_{loss}} = \frac{f_C}{f_{U1} - f_{L1}} \tag{3.101}$$

where f_C is the center frequency, X is the reactance of the antenna at f_C, R_r is the radiation resistance at f_C, R_{loss} is the resistive loss at f_C, f_{U1} is the upper 3 dB frequency, and f_{L1} is the lower 3 dB frequency. Increasing R_{loss} obviously lowers the Q due to the factor in the denominator of the second factor in (3.101). From the third factor of (3.101) lower Q implies that $f_U - f_L$ is larger, yielding the frequency characteristic as exhibited by $f_{U2} - f_{U1}$. Therefore, adding resistive loading in the antenna broadens the antenna's bandwidth.

This does not come for free, however. The impedance of the antenna is obviously changed with resistive loading, and the sensitivity is decreased.

The radiation resistance of electrically short antennas is quite low—on the order of 1or 2 ohms when $l << \lambda$. Increasing the loss resistance in such an antenna

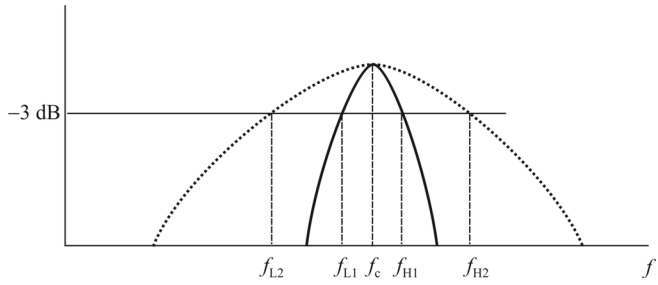


Figure 3.19 Resistive loading.

can easily cause the resistively loaded antenna to become unusable due to inefficiencies.

3.4 Radiation from Infinitesimal Sources

3.4.1 Introduction

The radiation from infinitesimal sources is an important topic because the radiation from more complex antennas can be analyzed by the superposition of the radiation fields generated by dissecting the complex antenna into such small sources.

3.4.2 Infinitesimal Dipole

Definition: The *infinitesimal dipole* is a dipole whose length δl is much smaller than the wavelength λ of the excited EM wave.[1] The infinitesimal dipole is equivalent to a current element $I\delta\vec{l}$, where

$$I\delta\vec{l} = -\frac{dQ}{dt}\delta\vec{l} \tag{3.102}$$

The infinitesimal dipole is also known as a *Hertzian dipole*. A current element is best illustrated by a very short (compared to λ) piece of infinitesimally thin wire with current I. Because the element is so short, the current in the infinitesimal dipole can be assumed to have constant magnitude along $\delta\vec{l}$. The ideal current element is unrealizable, but is used to analyze larger antenna structures.

3.4.2.1 Antenna Field Regions

A point is in a particular *field region* depending on how far the point is from the antenna. The *reactive near-field region* is the region immediately surrounding and possibly within the antenna where the reactive field is dominant. The *radiating near-field* (*Fresnel*) *region* is the region between the reactive near field and the far field where the radiation fields are dominant and the shape of the field distribution is dependent on the distance from the antenna. The *far-field* (*Fraunhofer*) *region* is farthest away from the antenna and is where the field *distribution* (shape, not absolute power levels) is essentially independent of the distance from the antenna. These regions are illustrated in Figure 3.20.

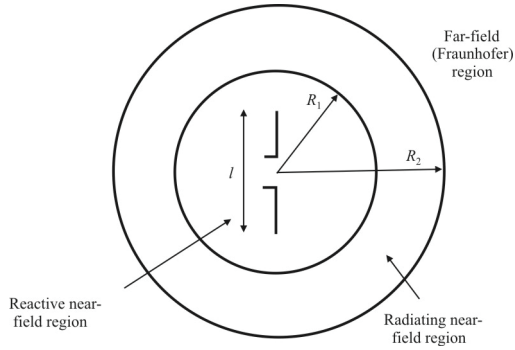


Figure 3.20 Radiation zones.

The space surrounding the antenna is divided into these three regions according to the predominant field behavior in those regions. The boundaries between the regions are not distinct and the field behavior changes very gradually as these boundaries are approached and crossed. For our purposes considering EW antennas, we are mostly concerned with the far-field characteristics. We will briefly discuss the other two regions, however. Understanding near-field behaviors is important when we analyze mutual coupling in antenna arrays in Chapter 9.

Care should be taken when determining the region boundaries. It will be shown that the region boundaries are determined by the geometry of the antenna in question, and, in particular, its size. It is important to be sure that the antenna dimensions are accurately accounted for to correctly determine the boundaries on these regions [19].

3.4.2.2 Far-Field Vectors Due to Current Element Radiation

Consider the two points, p and q, shown in Figure 3.21. The magnetic vector potential and electric vector potentials at observation point p due to electric current density at point q, $\vec{J}(\mathrm{q})$, and magnetic current density at point q, $\vec{M}(\mathrm{q})$, are defined as integrals of the (antenna) electric current density $\vec{J}(\mathrm{q})$ or magnetic current density $\vec{M}(\mathrm{q})$ as, respectively,

$$\vec{A}(\mathrm{p}) = \frac{\mu}{4\pi}\iiint_V \vec{J}(\mathrm{q})\frac{e^{-j\beta r_{\mathrm{pq}}}}{r_{\mathrm{pq}}}d\vec{v}' \tag{3.103}$$

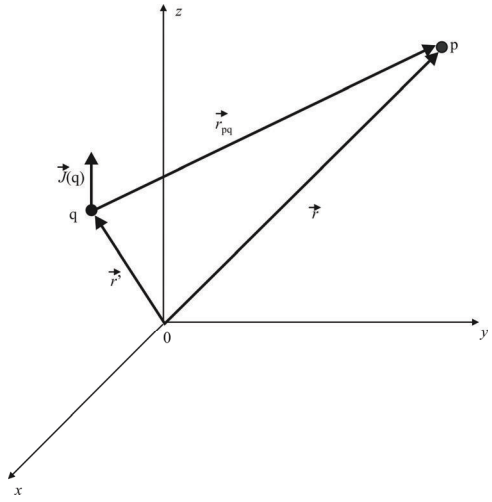


Figure 3.21 Far-field diagram.

$$\vec{F}(\mathrm{p}) = \frac{\varepsilon}{4\pi}\iiint_V \vec{M}(\mathrm{q})\frac{e^{-j\beta r_{pq}}}{r_{pq}}d\vec{v}' \tag{3.104}$$

We make the following approximation for the integrand (due to Kirchoff)

$$\frac{e^{-j\beta r_{pq}}}{r_{pq}} \approx \frac{e^{-j\beta(r-\vec{u}_r\cdot\vec{r}')}}{r} \tag{3.105}$$

where $\vec{r}$ is the position vector of the observation point p, its length is $r = |\vec{r}|$, and its direction is given by the unit vector $\vec{u}_r$ so that $\vec{r} = r\vec{u}_r$; $\vec{r}'$ is the position vector of the integration point q. This is the *far-field approximation*. So in the far field these integrals simplify to

$$\vec{A}(\mathrm{p}) = \frac{\mu e^{-j\beta r}}{4\pi r}\iiint_V \vec{J}(\mathrm{q})e^{-j\beta\vec{u}_r\cdot\vec{r}'}d\vec{v}' \tag{3.106}$$

$$\vec{F}(\mathrm{p}) = \frac{\varepsilon e^{-j\beta r}}{4\pi r}\iiint_V \vec{M}(\mathrm{q})e^{-j\beta\vec{u}_r\cdot\vec{r}'}d\vec{v}' \tag{3.107}$$

In most of cases of interest to us, the closed form solution of the VP integrals given by (3.106) and (3.107) is impossible to obtain. For the evaluation of the far fields (the Fraunhofer region), simplifying approximations are applied, from which approximations to the boundaries of these regions are determined. Consider the VP integral for a linear current source:

$$\vec{A} = \frac{\mu e^{-j\beta r_{pq}}}{4\pi r_{pq}} \int_{L'} I(\vec{l}\,') e^{-j\beta \vec{u}_r \cdot \vec{r}\,'} d\vec{l}\,' \tag{3.108}$$

where the observation point p(x, y, z) is at a distance from the source point q(x', y', z') given by $r_{pq} = \sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2}$ (see Figure 3.22). It is assumed that q(x', y', z') is along the integration contour L'.

For the infinitesimal dipole the current is constant along l, that is $I(\vec{l}\,') = I_0 \delta \vec{l}$. In practical antennas the solution of (3.108) can be very complicated depending on the function $I(l')$. Because of the assumed infinitesimal size of the source, the distance, r_{pq}, between the integration point and the observation point can be considered constant and equal to the distance from the center of the dipole so $r_{pq} = r = \sqrt{x^2 + y^2 + z^2}$. However, as the maximum dimension of the antenna, D_{max}, becomes comparable to the wavelength, the error, especially in the phase term, increases.

Thus, we have

$$\vec{A} = \frac{\mu e^{-j\beta r_{pq}} I_0 \delta l}{4\pi r_{pq}} \int_{L'} e^{-j\beta \vec{u}_r \cdot \vec{r}\,'} d\vec{l}\,' \tag{3.109}$$

Since we are primarily interested in the radiated field at considerable distances from the antenna and since the dipole is very small, $r_{pq} \approx r$ both in the exponent term and in the denominator. Furthermore, $\vec{u}_r$ and $\vec{r}\,'$ are approximately orthogonal at this range ($\vec{u}_r$ points toward p and $\vec{r}\,'$ is the distance vector from the origin to q so it points along the z-axis) so that $e^{-j\beta \vec{u}_r \cdot \vec{r}\,'} = e^{-j\beta \times 0} = 1$. Therefore, the integral in (3.109) becomes

$$\int_{L'} e^{-j\beta \vec{u}_r \cdot \vec{r}\,'} d\vec{l}\,' = \int_{L'} 1 \times d\vec{l}\,' = \vec{u}_z \tag{3.110}$$

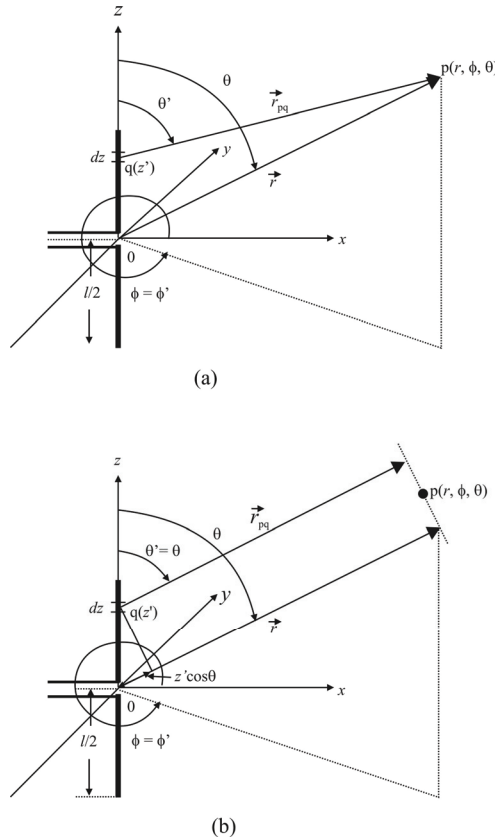


Figure 3.22 Far-field approximation. (a) finite size *z*-oriented dipole, and (b) *z*-oriented dipole far-field approximation.

and (3.109) reduces to

$$\vec{A} = \mu I_0 \delta l \frac{e^{-j\beta r}}{4\pi r} \vec{u}_z \tag{3.111}$$

Equation (3.111) gives the field due to an electric current element (infinitesimal dipole) expressed with the magnetic VP $\vec{A}$.

We can divide the function in (3.111) that depends on $r, e^{-j\beta r_{pq}} / r_{pq}$, into two factors: (1) the amplitude factor $1/r_{pq}$, and (2) the phase factor $e^{-j\beta r_{pq}}$. The amplitude factor is not very sensitive to errors in r_{pq}. In both the Fresnel and the Fraunhofer regions we will assume that $r_{pq} \approx r$. This approximation, however, cannot be used in the phase term as large errors ensue. To keep the phase term error low, the maximum error in βr_{pq} must be kept below $\pi/8$ radians (22.5°). Assuming that the wire is infinitesimally thin, we can ignore the variations along the *x* and *y* axes so that

$$\begin{aligned} r_{pq} &\approx \sqrt{x^2 + y^2 + (z - z')^2} \\ &= \sqrt{x^2 + y^2 + z^2 + (z'^2 - 2zz')} \\ &= \sqrt{r^2 + (z'^2 - 2rz'\cos\theta)} \end{aligned} \tag{3.112}$$

Using the binomial expansion[2] with $r^2 = a$, $z'^2 - 2rz'\cos\theta = b$, and $n = ½$, we get

$$\begin{aligned} r_{pq} &\approx (r^2)^{1/2} + \frac{1}{2}(r^2)^{-1/2}(z'^2 - 2rz'\cos\theta) + \frac{\frac{1}{2}\left(-\frac{1}{2}\right)}{2}(r^2)^{-3/2}(z'^2 - 2rz'\cos\theta)^2 + \cdots \\ &= r - z'\cos\theta + \frac{z'^2}{2r} - \frac{z'^2\cos^2\theta}{2r} + \mathcal{O}\left(\frac{1}{r^2}\right) \end{aligned} \tag{3.113}$$

where the last term in (3.113) refers to terms of the order $1/r^2$ and higher. Equation (3.113) can be further simplified to

---

[2] The binomial expansion is given by

$$(a+b)^n = a^n + na^{n-1}b + \frac{n(n-1)}{2!}a^{n-2}b^2 + \frac{n(n-1)(n-2)}{3!}a^{n-3}b^3 + \cdots$$

$$r_{\text{pq}} = r - z'\cos\theta + \frac{1}{2r} z'^2 \sin^2\theta + \mathcal{O}(1/r^2) \tag{3.114}$$

This expansion is used below to mathematically define the reactive near-field region, the radiating near-field region, and the far-field region.

3.4.2.3 Far-Field Radiation from an Infinitesimal Dipole

Far-Field Approximation

Since the third- and higher-order terms in (3.114) vary as $1/r$ or higher orders, we ignore them in the far field where r is large so that

$$r_{\text{pq}} \approx r - z'\cos\theta \tag{3.115}$$

The most significant error term in r_{pq} is the third one in (3.114)

$$\varepsilon(r) = \frac{1}{2r} z'^2 \sin^2\theta \tag{3.116}$$

which has its first maximum at $\phi = \pi/2,$ which is

$$\varepsilon_{\max} = \frac{z'^2}{2r} \tag{3.117}$$

So to limit the maximum error in βr_{pq} to $\pi/8$ radians, we must have

$$\beta \frac{z'^2_{\max}}{2r} \le \frac{\pi}{8} \tag{3.118}$$

but z' is actually half the largest dimension of the antenna so $z_{\max} = l/2$ and the smallest distance from the antenna this corresponds to is

$$r \ge 2\frac{l^2}{\lambda} \tag{3.119}$$

This is the (lower) limit on the edge of the far-field region that we show in Figure 3.20.

3.4.2.4 Field Vectors in the Far-Field of the Infinitesimal Dipole

The potentials (3.106) and (3.107) have the form of spherical waves as we would expect in the far field of the antenna. Also note that the complete *r*-dependence of the potentials is given outside the integrals. The $\vec{r}\,'$ term in the potential integrands can be expressed in terms of whatever coordinate system best fits the geometry of the source current; however, spherical coordinates are usually best for the field coordinates in the far field because frequently symmetry can be exploited to simplify the mathematics. The results of the far-field potential integrations in (3.106) and (3.107) may be written as

$$\vec{A}(\mathrm{p}) \approx \frac{e^{-j\beta r}}{r}[A_{\mathrm{r}}(\theta,\phi)\vec{u}_{\mathrm{r}} + A_{\phi}(\theta,\phi)\vec{u}_{\phi} + A_{\theta}(\theta,\phi)\vec{u}_{\theta}] \tag{3.120}$$

$$\vec{F}(\mathrm{q}) \approx \frac{e^{-j\beta r}}{r}[F_{\mathrm{r}}(\theta,\phi)\vec{u}_{\mathrm{r}} + F_{\theta}(\theta,\phi)\vec{u}_{\theta} + F_{\phi}(\theta,\phi)\vec{u}_{\phi}] \tag{3.121}$$

In antenna theory, the spherical coordinates are the preferred choices to describe $\vec{A}$ because the far-field radiation is of primary interest and the field is analyzed very far from the source. It is therefore safe to assume that the EM wave propagates only radially away from the source. The transformation from rectangular to spherical coordinates is given by

$$\begin{bmatrix} A_{\mathrm{r}} \\ A_{\theta} \\ A_{\phi} \end{bmatrix} = \begin{bmatrix} \sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta \\ \cos\theta\cos\phi & \cos\theta\sin\phi & -\sin\theta \\ -\sin\phi & \cos\phi & 0 \end{bmatrix} \begin{bmatrix} A_{\mathrm{x}} \\ A_{\mathrm{y}} \\ A_{\mathrm{z}} \end{bmatrix} \tag{3.122}$$

Applying (3.122) to $\vec{A}$ in (3.111) yields

$$\begin{aligned} A_{\mathrm{r}} &= A_{\mathrm{z}}\cos\theta = \mu I\delta l\frac{e^{-j\beta r}}{4\pi r}\cos\theta \\ A_{\theta} &= -A_{\mathrm{z}}\sin\theta = -\mu I\delta l\frac{e^{-j\beta r}}{4\pi r}\sin\theta \\ A_{\phi} &= 0 \end{aligned} \tag{3.123}$$

Take note that $\vec{A}$ does not depend on ϕ (due to the cylindrical symmetry of the dipole), and the dependence on r, which is $e^{-j\beta r}/r$, is separable from the dependence on θ. Also note that we can write

$$\vec{A}^{\inf}(\vec{r}) = \frac{\mu I_0 \delta l}{4\pi r} e^{-j\beta r} \vec{u}_z \tag{3.124}$$

Let us now find the field vectors $\vec{H}$ and $\vec{E}$. We know from (2.17) and (2.80) that

$$\vec{H} = \frac{1}{\mu} \nabla \times \vec{A} \tag{3.125}$$

The curl operator ($\nabla\times$) is expressed in spherical coordinates in (2.11) to get

$$\vec{H} = \frac{1}{\mu r}\left[\frac{\partial r A_\theta}{\partial r} - \frac{\partial A_r}{\partial \theta}\right]\vec{u}_\phi \tag{3.126}$$

The magnetic field $\vec{H}$ has only a $\vec{u}_\phi$ component. Finally,

$$H_\phi = j\beta I_0 \delta l \sin\theta \left(1 + \frac{1}{j\beta r}\right)\frac{e^{-j\beta r}}{4\pi r} \tag{3.127}$$

$$H_\theta = H_r = 0 \tag{3.128}$$

The field, $\vec{E}_A$, due to an electric current source $\vec{J}$ and the magnetic field, $\vec{H}_F$, due to magnetic current source $\vec{M}$ are defined by

$$\vec{E}_A = -j\omega\vec{A} - \frac{j}{\omega\mu\varepsilon}\nabla(\nabla\bullet\vec{A}) \tag{3.129}$$

$$\vec{H}_F = -j\omega\vec{F} - \frac{j}{\omega\mu\varepsilon}\nabla(\nabla\bullet\vec{F}) \tag{3.130}$$

We also know that from (2.90) that

$$\vec{E} = \frac{1}{j\omega\varepsilon}\nabla\times\vec{H} \tag{3.131}$$

and, in spherical coordinates, these become

$$E_{\mathrm{r}} = \eta_0 I_0 \delta l \cos\theta \left(1 + \frac{1}{j\beta r}\right) \frac{e^{-j\beta r}}{2\pi r^2} \tag{3.132}$$

$$E_{\theta} = j\eta_0 \beta I_0 \delta l \sin\theta \left[1 + \frac{1}{j\beta r} - \frac{1}{(\beta r)^2}\right] \frac{e^{-j\beta r}}{4\pi r} \tag{3.133}$$

$$E_{\phi} = 0 \tag{3.134}$$

Note that the higher order terms in $1/r$ can normally be ignored when dealing with long ranges from the antenna. Also note that E_{r} is smaller than E_{θ} because the former decreases as $1/r^2$ while the latter decreases as $1/r$. Therefore, in the far field we have

$$E_{\mathrm{r}} \approx E_{\phi} = H_{\mathrm{r}} = H_{\theta} = 0 \tag{3.135}$$

$$E_{\theta} \approx \frac{j\eta_0 \beta \delta l}{4\pi r} \sin\theta \tag{3.136}$$

$$H_{\phi} \approx j \frac{\beta I_0 \delta l e^{-j\beta r}}{4\pi r} \sin\theta \tag{3.137}$$

Considering the far region and ignoring the $1 / r$ terms in E_{θ}, the ratio of E_{θ} (3.133) to H_{ϕ} (3.127) yields η_0, the characteristic impedance of a vacuum:

$$\frac{E_{\theta}}{H_{\phi}} = \frac{j\eta_0 \, \beta I \delta l \sin\theta \dfrac{e^{-j\beta r}}{4\pi r}}{j\beta \, I \delta l \sin\theta \dfrac{e^{-j\beta r}}{4\pi r}} = \eta_0 \tag{3.138}$$

3.4.2.5 Far-Field Power Density and Radiated Power of the Infinitesimal Dipole

The complex Poynting vector P describes the complex power density flux. It is given by (2.61)

$$\vec{P} = \frac{1}{2}\vec{E} \times \vec{H}^* = \frac{1}{2}\left(E_{\mathrm{r}}\vec{u}_{\mathrm{r}} + E_{\theta}\vec{u}_{\theta}\right) \times \left(H_{\phi}^*\vec{u}_{\phi}\right) = \frac{1}{2}\left(E_{\theta}H_{\phi}^*\vec{u}_{\mathrm{r}} - E_{\mathrm{r}}H_{\phi}^*\vec{u}_{\theta}\right) \tag{3.139}$$

Substituting (3.132), (3.133), and (3.127) into (3.139) yields

$$P_{\mathrm{r}} = \frac{\eta_0}{8}\left|\frac{I\delta l}{\lambda}\right|^2 \frac{\sin^2\theta}{r^2}\left[1 - j\frac{1}{(\beta r)^3}\right]$$
$$P_{\theta} = j\eta_0\beta\frac{\left|I\delta l\right|^2 \cos\theta\sin\theta}{16\pi^2 r^3}\left[1 + j\frac{1}{(\beta r)^2}\right] \tag{3.140}$$

The overall power, Π, is calculated over a sphere, and, therefore, only the radial component of the Poynting vector, P_{r}, contributes

$$\Pi = \oiint_S \vec{P}\bullet d\vec{s} = \oiint_S \left(P_{\mathrm{r}}\vec{u}_{\mathrm{r}} + P_{\theta}\vec{u}_{\theta}\right)\bullet\vec{u}_{\mathrm{r}} r^2 \sin\theta d\theta d\phi \tag{3.141}$$

$$\Pi = \frac{\pi}{3}\eta_0\left(\frac{I\delta l}{\lambda}\right)^2\left[1 - j\frac{1}{(\beta r)^3}\right], \qquad \mathrm{W} \tag{3.142}$$

According to (2.59), the radiated power is equal to the real part of the complex power (the time average of the total power flow). Therefore, from (3.142) the radiated power of an infinitesimal electric dipole is

$$\Pi_{\mathrm{rad}} = \frac{\pi}{3}\eta_0\left(\frac{I\delta l}{\lambda}\right)^2, \qquad \mathrm{W} \tag{3.143}$$

The radiation resistance R_{r} describes the power "loss" due to radiation in an equivalent circuit of the antenna so that

$$\Pi = \frac{1}{2}R_{\mathrm{r}}I^2 \tag{3.144}$$

and so

$$R_{\mathrm{r}} = \frac{2\Pi}{I^2} \tag{3.145}$$

Therefore, for the infinitesimal dipole in free space or a vacuum

$$R_{\mathrm{r}} = \frac{2\pi}{3}\eta_0\left(\frac{\delta l}{\lambda}\right)^2 \tag{3.146}$$

$$= 80\pi^2 \left(\frac{\delta l}{\lambda} \right)^2, \qquad \Omega \tag{3.147}$$

The radiation resistance will be discussed in more detail later.

In summary, the most important features of the far field are:

1. There are no radial components.
2. The angular field distribution is independent of the distance from the source, *r*.
3. The $\vec{E}$-field and $\vec{H}$-field are orthogonal, $\vec{E} \perp \vec{H}$, and both are orthogonal to the direction of propagation.
4. The intrinsic impedance of free space is given by

$$\eta_0 = \frac{E_\theta}{H_\phi}$$

5. The Poynting power vector is given by

$$\vec{P} = \frac{1}{2} \vec{E} \times \vec{H}^* = \frac{1}{2} \frac{1}{\eta_0} \left| E_\theta \right|^2 \vec{u}_r = \frac{1}{2} \eta_0 \left| H_\phi \right|^2 \vec{u}_r$$

3.4.2.6 Fresnel Zone Radiation from an Infinitesimal Dipole

Radiating Near-Field (Fresnel Region) Approximation

The Fresnel region is the middle field region (see Figure 3.20). The outer limit to this region is the same as the inner limit on the Fraunhofer region, so its upper boundary is specified by

$$r \le 2 \frac{l^2}{\lambda} \tag{3.148}$$

When the observation point is in this region, we use one more term in the expansion of *R* as given by (3.114) to sufficiently reduce the phase error. The approximation this time is

$$R = r - z' \cos\theta + \frac{1}{2r} z'^2 \sin^2\theta + \frac{1}{2r^2} z'^3 \cos\phi \sin^2\theta + O(1/r^3) \tag{3.149}$$

The most significant error term is

$$\epsilon(r) = \frac{1}{2}\frac{z'^3}{r^2}\cos\theta\sin^2\theta \tag{3.150}$$

The angles of the extreme values are found by

$$\frac{\partial\epsilon}{\partial\theta} = 0 = \frac{z'^3}{2r^2}\sin\theta(-\sin^2\theta + 2\cos^2\theta)$$

with roots $\theta_1 = 0$ (minimum), $\theta_2 = \tan^{-1}(-\sqrt{2})$ (maximum), and $\theta_3 = \tan^{-1}(\sqrt{2})$ (maximum).

As above,

$$\beta\epsilon_{\max}(r) = \frac{2\pi}{\lambda}\frac{z'^3}{r^2}\cos\theta_2\sin^2\theta_2$$

and

$$\beta\epsilon_{\max}(r) = \frac{\pi}{12\sqrt{3}}\frac{L^3}{\lambda r^2} \le \frac{\pi}{8}$$

so that

$$r \ge 0.62\sqrt{\frac{l^3}{\lambda}} \tag{3.151}$$

This is the inner edge of the Fresnel region and therefore the outer edge of the reactive near-field region.

In this intermediate region the radiation field is more significant but the angular field distribution is still dependent on the distance from the antenna. It is generally accepted that the Fresnel region is enclosed between two spherical surfaces defined by

$$0.62\sqrt{\frac{D_{\max}^3}{\lambda}} \le r \le \frac{2D_{\max}^2}{\lambda} \tag{3.152}$$

where $D_{\max}$ is the largest dimension of the antenna.

Radiation Fields in the Fresnel Zone

The fields of an infinitesimal dipole in the Fresnel region are obtained by neglecting the higher-order $(1/r)^n$ terms in (3.128) and (3.134) yielding

$$H_\phi \approx j\frac{\beta(I\delta l)e^{-j\beta r}}{4\pi r}\sin\theta \tag{3.153}$$

$$E_r \approx \eta\frac{(I\delta l)e^{-j\beta r}}{2\pi r^2}\cos\theta \tag{3.154}$$

$$E_\theta \approx j\eta\frac{\beta(I\delta l)e^{-j\beta r}}{4\pi r}\sin\theta \tag{3.155}$$

$$H_\theta = H_r = E_\phi = 0 \tag{3.156}$$

when $\beta r \geq 1$. The radial component E_r is still not negligible, but the transverse components (E_ϕ and H_θ) are dominant.

3.4.2.7 Reactive Near-Field Radiation from an Infinitesimal Dipole

Reactive Near-Field Region Approximation

This is the region immediately surrounding the antenna where the reactive field dominates. For our purposes we will assume that this region is a sphere with the antenna at its center with a radius of

$$r \approx 0.62\sqrt{\frac{l^3}{\lambda}} \tag{3.157}$$

It is important to note that (3.157) is appropriate for wire and waveguide aperture antennas, while it is not valid for electrically large reflector antennas.

Radiation Fields in the Reactive Near-Field Region

When (3.157) is true, r is so small that $\beta r << 1$. Then the most significant terms in the field expressions (3.128) and (3.134) are

$$H_{\phi} \approx \frac{(I\delta l)e^{-j\beta r}}{4\pi r^2}\sin\theta \tag{3.158}$$

$$E_{r} \approx -j\eta\frac{(I\delta l)e^{-j\beta r}}{2\pi\beta r^3}\cos\theta \qquad \beta r << 1 \tag{3.159}$$

$$E_{\theta} = -j\eta\frac{(I\delta l)e^{-j\beta r}}{4\pi\beta r^3}\sin\theta \tag{3.160}$$

$$H_{r} = H_{\theta} = E_{\phi} = 0 \tag{3.161}$$

This field is essentially purely reactive because $\vec{H}$ and $\vec{E}$ are in phase quadrature. Actually, the $e^{-j\beta r}$ can be neglected, and then we see that the following conditions exist:

- The H_{ϕ} component is exactly the static magnetic field of a current element $(I\delta l)$.
- The E_{θ} and E_{r} components are exactly the static electric field of a dipole.

That the field is almost purely reactive in the near zone is obvious from the power equation (3.140). Its imaginary part is

$$\mathrm{Im}\{\Pi\} = -\frac{\pi}{3}\eta\left(\frac{I\delta l}{\lambda}\right)^2\frac{1}{(\beta r)^3} \tag{3.162}$$

$\mathrm{Im}\{\Pi\}$ obviously dominates over the radiated power (3.143) when $r \to 0$, since Π_{rad} does not depend on r. Since the radial reactive power flow density P_{r} has a negative imaginary value, it is predominantly magnetic and decreases as $(1/r^5)$

$$P_{r}^{\mathrm{near}} = -j\frac{\eta}{8}\left(\frac{I\delta l}{\lambda}\right)^2\frac{\sin^2\theta}{\beta^3 r^5},\ \mathrm{W} \tag{3.163}$$

The P_{θ} power density component has the same order of dependence on r but is predominantly electric in nature

$$P_{\theta}^{\mathrm{near}} = j\eta\beta\frac{(I\delta l)^2}{16\pi^2 r^3}\frac{1}{(\beta r)^2} = j\frac{\eta}{8}\left(\frac{I\delta l}{\lambda}\right)^2\frac{\sin(2\theta)}{\beta^3 r^5} \tag{3.164}$$

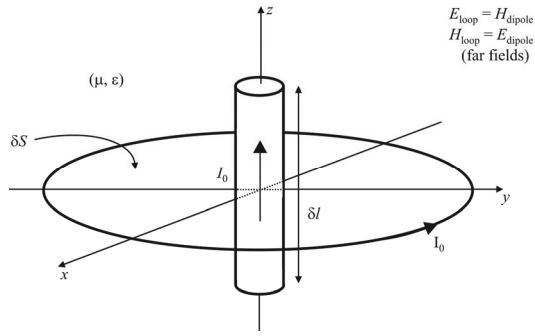


Figure 3.23 Infinitesimal dipole and loop.

3.4.3 Infinitesimal Loop Antennas

Loop antennas have many of the same favorable characteristics as dipoles and monopoles—in particular they are inexpensive and simple to construct [19]. Loop antennas come in a variety of shapes (circular, rectangular, elliptical, and so forth), but the fundamental characteristics of the loop antenna radiation pattern (far field) are, for the most part, independent of the loop shape.

Just as the electrical length of the dipoles and monopoles to a large extent affect the efficiency and other basic characteristics of these antennas, the electrical size of the loop (circumference) determines the basic characteristics of the loop antenna. Loop antennas are usually classified as either electrically small or electrically large based on the circumference of the loop:

- Electrically small loop: circumference $\leq \lambda/10$;
- Electrically large loop: circumference $\approx \lambda$.

The infinitesimal loop antenna is the dual antenna to the electrically short dipole antenna when oriented as shown in Figure 3.23. The current in the dipole and the loop are the same, given by I_0. That is, the far electric field of a small loop antenna is identical to the far magnetic field of the short dipole antenna and the far magnetic field of a small loop antenna is identical to the far electric field of the short dipole antenna.

Given that the radiated fields of the short dipole and small loop antennas are dual quantities, the radiated power for both antennas is the same, and therefore, the radiation patterns are the same. This means that the plane of maximum radiation for the loop is in the plane of the loop. When operated as a receiving antenna, we know that the short dipole must be oriented such that the electric field is parallel to

the wire for maximum response (the polarizations must match). Due to duality, we find that the small loop must be oriented such that the magnetic field is perpendicular to the plane of the loop for maximum response. The loop can therefore be considered a transducer for magnetic waves.

The radiation resistance of the small loop is much smaller than that of the short dipole, however. The loss resistance of the small loop antenna is typically much larger than the radiation resistance, and so, therefore, they are inefficient. Because of this, the small loop antenna is rarely used as a transmit antenna. However, the small loop antenna is suitable as a receive antenna since signal-to-noise ratio is the key factor, especially in EW systems, not antenna efficiency. (It is particularly critical in EW systems because frequently the signals of interest are relatively weak.) The fact that a significant portion of the received signal is lost in the form of heat is not of consequence as long as the antenna provides a large enough signal-to-noise ratio for the given receiver. Small loop antennas are frequently used for communications receiving applications such as pagers, low frequency portable radios, and direction finding.

3.4.3.1 Infinitesimal Loop Antenna

The far fields of an electrically small loop antenna are dependent on the loop area but are essentially independent of the loop shape. Since the magnetic VP integrations required for a circular loop are more complex than those for a square loop, we will use the square loop in the derivation of the far fields of an infinitesimal loop antenna. The square loop, located in the *x*-*y* plane and centered at the coordinate origin shown in Figure 3.24, is assumed to have an area of Δl^2 and carry a uniform current I_0. The distances in Figure 3.24 are approximated with

$$R_1 \approx r + \frac{\delta l}{2}\sin\theta\sin\phi \tag{3.165}$$

$$R_2 \approx r - \frac{\delta l}{2}\sin\theta\cos\phi \tag{3.166}$$

$$R_3 \approx r - \frac{\delta l}{2}\sin\theta\sin\phi \tag{3.167}$$

$$R_4 \approx r + \frac{\delta l}{2}\sin\theta\cos\phi \tag{3.168}$$

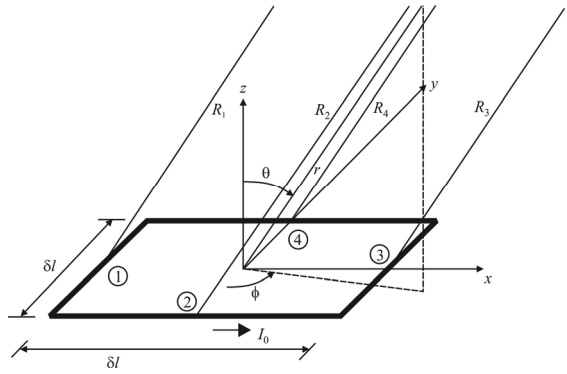


Figure 3.24 The square loop may be viewed as four segments which each represent an infinitesimal dipole carrying current in a different direction. In the far field, the distance vectors from the centers of the four segments become almost parallel.

For the far field, the approximations in (3.165)–(3.168) can be used in the phase terms of the magnetic VP, but we may assume that $R_1 \approx R_2 \approx R_3 \approx R_4 \approx r$ for the magnitude terms. The far-field magnetic VP of a *z*-directed infinitesimal dipole centered at the origin is given by (3.111). The individual far-field magnetic VP contributions due to the four segments of the current loop are

$$\vec{A}_1 \approx \frac{\mu I_0 \delta l}{4\pi r} e^{-j\beta\left(r+\frac{\delta l}{2}\sin\theta\sin\phi\right)}(-\vec{u}_y) \tag{3.169}$$

$$\vec{A}_2 \approx \frac{\mu I_0 \delta l}{4\pi r} e^{-j\beta\left(r-\frac{\delta l}{2}\sin\theta\cos\phi\right)}\vec{u}_x \tag{3.170}$$

$$\vec{A}_3 \approx \frac{\mu I_0 \delta l}{4\pi r} e^{-j\beta\left(r-\frac{\delta l}{2}\sin\theta\sin\phi\right)}\vec{u}_y \tag{3.171}$$

$$\vec{A}_4 \approx \frac{\mu I_0 \delta l}{4\pi r} e^{-j\beta\left(r+\frac{\delta l}{2}\sin\theta\cos\phi\right)}\vec{u}_x \tag{3.172}$$

Combining the (3.169) with (3.171) (the *y*-directed terms) and (3.170) with (3.172) (the *x*-directed terms) yields

$$\vec{A}_1 + \vec{A}_3 \approx \frac{\mu I_0 \delta l}{4\pi r} e^{-j\beta r}\left[e^{j\beta\frac{\delta l}{2}\sin\theta\sin\phi} - e^{-j\beta\frac{\delta l}{2}\sin\theta\sin\phi}\right]\vec{u}_y \tag{3.173}$$

$$= j\frac{\mu I_0 \delta l}{2\pi r} e^{-j\beta r} \sin\left(\frac{\beta \delta l}{2} \sin\theta \sin\phi\right) \vec{u}_y \tag{3.174}$$

$$\vec{A}_2 + \vec{A}_4 \approx \frac{\mu I_0 \delta l}{4\pi r} e^{-j\beta r} \left[e^{-j\beta \frac{\delta l}{2} \sin\theta\cos\phi} - e^{j\beta \frac{\delta l}{2} \sin\theta\cos\phi} \right] \vec{u}_x \tag{3.175}$$

$$= j\frac{\mu I_0 \delta l}{2\pi r} e^{-j\beta r} \sin\left(\frac{\beta \delta l}{2} \sin\theta\cos\phi\right) \vec{u}_x \tag{3.176}$$

When $\delta l << \lambda$, the arguments of the sine functions in (3.174) and (3.176) are very small and may be approximated as

$$\sin x \approx x \qquad x << 1 \tag{3.177}$$

yielding

$$\vec{A}_1 + \vec{A}_3 \approx j\frac{\beta\mu I_0 \delta l^2}{2\pi r} e^{-j\beta r} \sin\theta \sin\phi \vec{u}_y \tag{3.178}$$

$$\vec{A}_2 + \vec{A}_4 \approx j\frac{\beta\mu I_0 \delta l^2}{4\pi r} e^{-j\beta r} \sin\theta \cos\phi \vec{u}_x \tag{3.179}$$

The overall VP is the sum of the four components

$$\begin{aligned} \vec{A} &= \vec{A}_1 + \vec{A}_2 + \vec{A}_3 + \vec{A}_4 \\ &\approx j\frac{\beta\mu I_0 \delta S}{4\pi r} e^{-j\beta r} \sin\theta [\sin\phi \vec{u}_y + \cos\phi \vec{u}_x] \end{aligned} \tag{3.180}$$

where $\delta S = \delta l^2 =$ loop area. The bracketed term in (3.180) is the spherical coordinate unit vector $\vec{u}_\phi$ so the infinitesimal current loop far-field magnetic VP is

$$\vec{A} \approx j\frac{\beta\mu I_0 \delta S}{4\pi r} e^{-j\beta r} \sin\theta \vec{u}_\phi \tag{3.181}$$

The corresponding far fields are thus

$$\vec{E}_\phi \approx j\omega\vec{A} = \frac{\omega\beta\mu I_0 \delta S}{4\pi r} e^{-j\beta r} \sin\theta \vec{u}_\phi \tag{3.182}$$

$$\vec{H}_\theta \approx j\frac{\omega}{\eta}\vec{A} = \frac{\omega\beta\mu I_0 \delta S}{4\pi r\eta} e^{-j\beta r}\sin\theta\vec{u}_\theta \tag{3.183}$$

but

$$\omega\beta\mu = (\omega\mu)\omega\sqrt{\mu\epsilon} = \omega^2\mu\epsilon\sqrt{\frac{\mu}{\epsilon}} = \mu\beta^2 \tag{3.184}$$

so

$$\vec{E}_\phi \approx \frac{\eta_0\beta^2 I_0\delta S}{4\pi r} e^{-j\beta r}\sin\theta\vec{u}_\phi \tag{3.185}$$

$$\vec{H}_\theta \approx \frac{\beta^2 I_0\delta S}{4\pi r} e^{-j\beta r}\sin\theta\vec{u}_\theta \tag{3.186}$$

The fields radiated by an electrically small loop antenna can be increased by adding multiple turns. For the far fields, the added height of multiple turns is immaterial and the resulting far fields for a multiple turn loop antenna can be found by the superposition of the fields due to each of the individual loops, and assuming the loops are identical, this is simply multiplying the single-turn loop antenna fields by the number of turns *N*. So the infinitesimal multiple-turn current loop far fields are

$$\vec{E}_\phi \approx \frac{\eta_0\beta^2 N I_0\delta S}{4\pi r} e^{-j\beta r}\sin\theta\vec{u}_\phi \tag{3.187}$$

$$\vec{H}_\theta \approx \frac{\beta^2 N I_0\delta S}{4\pi r} e^{-j\beta r}\sin\theta\vec{u}_\theta \tag{3.188}$$

Note that this also increases the loss resistance of the multiple loop antenna by the same factor *N*.

3.4.3.2 Loop Antenna Characteristics

The time-average Poynting vector in the far field of the multiple-turn infinitesimal loop, from (2.61), is

$$\vec{P} = \frac{1}{2}\vec{E}\times\vec{H}^* = \frac{1}{2}\left[E_\phi\vec{u}_\phi \times H_\theta^*\vec{u}_\theta\right] = \frac{1}{2}\left[E_\phi\vec{u}_\phi \times \frac{\vec{E}_\phi}{\eta_0}\vec{u}_\theta\right] = \frac{|E_\phi|^2}{2\eta_0}\vec{u}_r$$
$$= \frac{\eta_0}{2}\left[\frac{\beta^2 N\delta S}{4\pi r}\sin\theta\right]^2 \vec{u}_r$$
$$= \frac{\eta_0}{2}\left[\frac{\pi N\delta S}{\lambda^2 r}\right]^2 |I_0|^2 \sin^2\theta\vec{u}_r \tag{3.189}$$

The loop antenna radiation intensity function is then

$$\vec{U}(\theta,\phi) = r^2\vec{P} == \frac{\eta_0}{2}\left[\frac{\pi N\delta S}{\lambda^2}\right]^2 |I_0|^2 \sin^2\theta\vec{u}_r \tag{3.190}$$

The maximum value of (3.190) is

$$U(\theta,\phi)_{\max} = \frac{\eta_0}{2}\left[\frac{\pi N\delta S}{\lambda^2}\right]^2 |I_0|^2 \tag{3.191}$$

The radiated power is

$$\Pi = \int_0^{2\pi}\int_0^{\pi} U(\theta,\phi)\sin\theta d\theta d\phi = \frac{\eta_0}{2}\left[\frac{\pi N\delta S}{\lambda^2}\right]^2 |I_0|^2 2\pi\int_0^{\pi}\sin^3\theta d\theta$$

so the loop antenna radiated power is

$$\Pi = \frac{4}{3}\eta_0\pi^3\left[\frac{N\delta S}{\lambda^2}\right]^2 |I_0|^2 \tag{3.192}$$

The radiation resistance of the loop antenna is found from the radiated power.

$$\Pi = \frac{4}{3}\eta_0\pi^3\left[\frac{N\delta S}{\lambda^2}\right]^2 |I_0|^2 = \frac{1}{2}|I_0|^2 R_r \tag{3.193}$$

so

$$R_{\text{r}} = \frac{8}{3}\eta_0\pi^3\left[\frac{N\delta S}{\lambda^2}\right]^2 \tag{3.194}$$

and the radiation resistance of the multiple turn infinitesimal loop antenna in free space $(\eta_0 = 120\pi)$ is

$$R_{\text{r}} = 320\pi^4\left[\frac{N\delta S}{\lambda^2}\right]^2 \tag{3.195}$$

The radiation resistance of small loops is very small. For example, for $\lambda/100 < b < \lambda/10$, the radiation resistance increases from $\sim 3\times 10^{-3}$ up to $\sim 0.5\Omega$. This is often less than the loss resistance of the loop. That is why small loop antennas are constructed with multiple turns and on a ferromagnetic core. They, of course, have much greater inductive reactance, which is always compensated by a capacitor. This is actually used in narrowband receivers, where the antenna itself is an efficient filter (together with the tuning capacitor), which can be tuned for different frequency bands.

The directivity of the loop antenna is defined by

$$D(\theta,\phi) = 4\pi\frac{U(\theta,\phi)}{\Pi} = 4\pi\frac{\frac{\eta_0}{2}\left[\frac{\pi N\delta S}{\lambda^2}\right]^2 |I_0|^2 \sin^2\theta}{\frac{4}{3}\eta_0\pi^3\left[\frac{N\delta S}{\lambda^2}\right]^2 |I_0|^2}$$

and the loop antenna directivity function is

$$D(\theta,\phi) = 1.5\sin^2\theta \tag{3.196}$$

Given the same directivity function as the infinitesimal dipole, the loop antenna has the same maximum directivity, effective area, and beam solid angle as the infinitesimal dipole. The loop antenna maximum directivity is

$$D_0 = 1.5 \tag{3.197}$$

The loop antenna effective area is

$$A_{\text{eff}} = \frac{\lambda^2}{4\pi} D_0 = \frac{3\lambda^2}{8\pi} \tag{3.198}$$

and the loop antenna beam solid angle is

$$\Omega_{\text{A}} = \frac{4\pi}{D_0} = \frac{8\pi}{3} \tag{3.199}$$

When we compare the radiation resistances of the infinitesimally small dipole and the infinitely small loop (both antennas in free space), we find that the radiation resistance of the small loop decreases much faster than that of the short dipole with decreasing frequency since

$$\begin{aligned} &\text{Dipole:} && R_{\text{r}} = 80\pi^2 \left(\frac{\delta l}{\lambda} \right)^2 && \Rightarrow R_{\text{r}} \propto \lambda^{-2} \\ &\text{Loop:} && R_{\text{r}} = 320\pi^4 \left(\frac{N\delta S}{\lambda^2} \right)^2 && \Rightarrow R_{\text{r}} \propto \lambda^{-4} \end{aligned} \tag{3.200}$$

We see that the radiation resistance of the infinitesimal loop decreases with the inverse fourth power of the wavelength while that of the infinitesimal dipole decreases with the inverse square of the wavelength.

3.4.4 Impedances of Infinitesimal Antennas

The current density was assumed to be uniform on the infinitesimal current loop for our far-field calculations. For a circular loop, the assumption of uniform current is reasonably accurate up to a loop circumference of about 0.2λ. This restriction on the size of the constant current loop in terms of the loop radius is

$$2\pi b \leq 0.2\lambda \tag{3.201}$$

so

$$b \leq 0.032\lambda \tag{3.202}$$

(see Figure 3.25).

The electrically small current loop was found to be a dual source to the infinitesimal dipole. The reactance of the dipole is capacitive while the loop is inductive. The exact reactance of the current loop is dependent on the shape of the

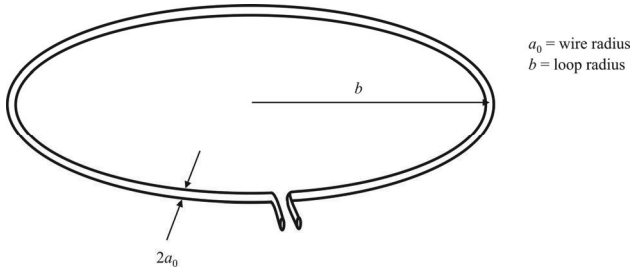


Figure 3.25 Infinitesimal loop.

loop. Approximate formulas for the reactance are given below for a short dipole and an electrically small circular current loop.

Infinitesimal Dipole (length = δl, wire radius = a)

$$C \approx \frac{\pi\epsilon_0 \delta l}{\left[\ln\left(\frac{\delta l}{a}\right) - 1\right]} \qquad X_A = -\frac{1}{\omega C} \approx -\frac{\left[\ln\left(\frac{\delta l}{a}\right) - 1\right]}{\omega\pi\epsilon_0 \delta l} \tag{3.203}$$

Infinitesimal Circular Current Loop (loop radius = b, wire radius = a)

$$L \approx \mu_0 b\left[\ln\left(\frac{8b}{a}\right) - 2\right] \qquad X_A = \omega L \approx \omega\mu_0 a\left[\ln\left(\frac{8b}{a}\right) - 2\right] \tag{3.204}$$

Example: Table 3.2 compares the impedances (R_r and X_A) of the infinitesimal dipole and infinitesimal loop based on the equations in this section. The dipole is 6 cm long and the radius of the loop is 9.5 cm. The wire is #10 AWG copper so a = 4.59 mm and $\sigma = 5.8 \times 10^7$ mhos m^{-1}. The radiation efficiency is given by

Table 3.2 Comparison of Small Dipole and Small Loop

Infinitesimal dipole					
f (MHz)	δl	R_r	R_L	η_{cd}	jX_A
1	0.0002λ	31.6 μΩ	0.962 mΩ	4.18%	–204 kΩ
10	0.002λ	4.16 mμΩ	4.04 mΩ	51.0%	–20.4 kΩ
100	0.02λ	0.316 Ω	9.62 mΩ	97.0%	–4.04 kΩ
Small loop					
f (MHz)	a	R_r	R_L	η_{cd}	jX_A
1	0.00032λ	4.09 nΩ	9.57 mΩ	4.2×10^{-5}%	4.76 Ω
10	0.0032λ	30.9 nΩ	30.3 mΩ	0.102%	27.6 Ω
100	0.032λ	0.309 Ω	95.7 mΩ	76.4%	276 Ω

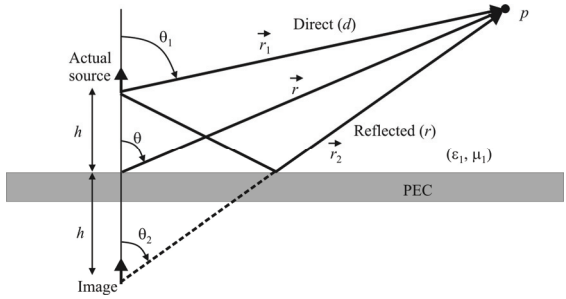


Figure 3.26 Monopole elevated above a PEC.

$$\eta_{cd} = \frac{R_r}{R_r + R_L} \tag{3.205}$$

3.5 Antennas Elevated Above a Perfectly Conducting Ground Plane

Ground-based communications, radar, and EW systems typically use elevated antennas, erected above a mast of some type. At ground level, antennas for the VHF and above spectrum have very short ranges if not elevated. In this section we examine the effects on the pattern of an infinitesimal vertical monopole antenna elevated above a perfectly conducting ground plane (PEC), where $\sigma \to \infty$. A schematic of the antenna positioning is shown in Figure 3.26.

The field at the observation point p is a superposition of the fields of the actual source and the image source as shown in Figure 3.26, both radiating in a homogeneous medium with parameters (ε_1, μ_1). The actual source is a current element $(I_0 \delta l)$ (infinitesimal monopole) with far fields given according to (3.133) as

$$E_\theta^d \approx j\eta \frac{\beta (I\delta l) e^{-j\beta r_1}}{4\pi r_1} \sin\theta_1 \tag{3.206}$$

$$E_\theta^r \approx j\eta \frac{\beta (I\delta l) e^{-j\beta r_2}}{4\pi r_2} \sin\theta_2 \tag{3.207}$$

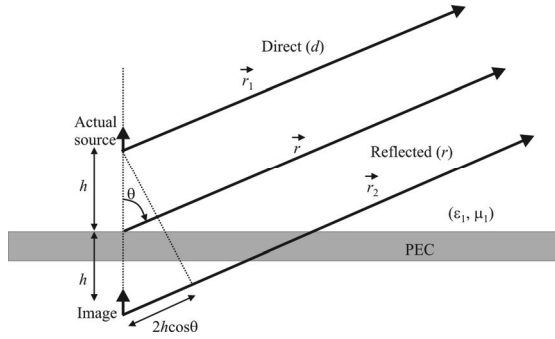


Figure 3.27 Far field of an elevated monopole.

Using the cosine theorem we can express the distances $|r_1|$ and $|r_2|$ in terms of $|r|$ and h, yielding

$$r_1 = \sqrt{r^2 + h^2 - 2rh\cos\theta} \tag{3.208}$$

$$r_2 = \sqrt{r^2 + h^2 - 2rh\cos(\pi - \theta)} \tag{3.209}$$

As above, we approximate the radial distances (but not the phase terms) in (3.206) and (3.207) as being equal in the far field as $r_1 = r_2 = r$. We expand r_1 and r_2 according to the binomial series to obtain approximations of the phase terms, keeping only the first two terms (see Figure 3.27), yielding

$$r_1 \approx r - h\cos\theta \tag{3.210}$$

$$r_2 \approx r + h\cos\theta \tag{3.211}$$

The total far field is

$$\begin{aligned} E_\theta &= E_\theta^{\mathrm{d}} + E_\theta^{\mathrm{r}} \\ E_\theta &= j\eta\beta\frac{I_0\delta l}{4\pi r}\sin\theta\left[e^{-j\beta(r-h\cos\theta)} + e^{-j\beta(r+h\cos\theta)}\right] \\ E_\theta &\approx \underbrace{j\eta\beta I_0\delta l\frac{e^{-j\beta r}}{4\pi r}\sin\theta}_{g(\theta)}\underbrace{\left[2\cos(\beta h\cos\theta)\right]}_{f(\theta)}, \qquad z \geq 0 \\ E_\theta &= 0, \qquad z < 0 \end{aligned} \tag{3.212}$$

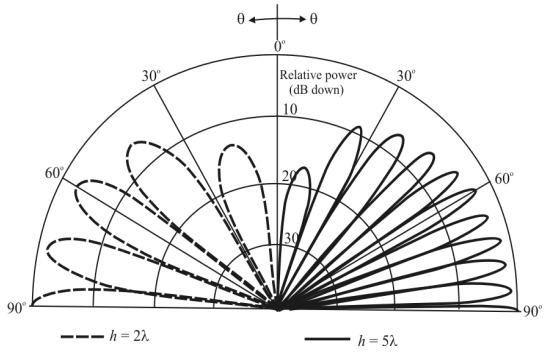


Figure 3.28 Monopole scalloping. As the elevation is increased, the number of lobes increases.

Note that the far-field expression can be decomposed into two factors: the field of the elementary source $g(\theta)$ and the *array factor* $f(\theta)$. Much more will be said about the array factor in Chapter 9 when we discuss antenna arrays.

The normalized power pattern is

$$F(\theta) = \left[\sin\theta\cos(\beta h\cos\theta)\right]^2 \tag{3.213}$$

As the vertical monopole is moved further away from the infinite PEC plane, more and more lobes are introduced in the power pattern. This effect is called *scalloping* of the pattern and is illustrated in Figure 3.28. The number of lobes is given by

$$n = \left\lceil \frac{2h}{\lambda} + 1 \right\rceil \tag{3.214}$$

The total radiated power is given by

$$\begin{aligned}\Pi &= \oiint \vec{P}\overrightarrow{ds} = \frac{1}{2\eta_0}\int_0^{2\pi}\int_0^{\pi/2} |E_\theta|^2 r^2 \sin\theta d\theta d\phi \\ &= \frac{\pi}{\eta_0}\int_0^{\pi/2} |E_\theta|^2 r^2 \sin\theta d\theta \\ &= \eta_0\beta^2 (I_0\delta l)^2 \int_0^{\pi/2} \sin^2\theta\cos^2(\beta h\cos\theta) d\theta\end{aligned}$$

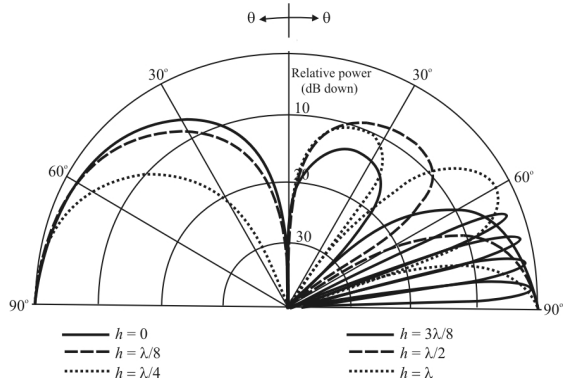


Figure 3.29 Monopole elevation patterns above a PEC when $h < \lambda$.

$$= \pi\eta_0 \left(\frac{I_0 \delta l}{\lambda} \right)^2 \left[\frac{1}{3} - \frac{\cos(2\beta h)}{(2\beta h)^2} + \frac{\sin(2\beta h)}{(2\beta h)^3} \right] \tag{3.215}$$

As $\beta h \to 0$, the radiated power of the vertical monopole approaches twice the value of the radiated power of a monopole of the same length in free space. As $\beta h \to \infty$, the radiated power of both dipoles becomes the same, as would be expected because this is the condition of free space.

In free space, the radiation resistance is given by

$$R_r = \frac{2\Pi}{|I_0|^2} = 2\pi\eta_0 \left(\frac{\delta l}{\lambda} \right)^2 \left[\frac{1}{3} - \frac{\cos(2\beta h)}{(2\beta h)^2} + \frac{\sin(2\beta h)}{(2\beta h)^3} \right] \tag{3.216}$$

As $\beta h \to 0$, the radiation resistance of the vertical monopole approaches twice the value of the radiation resistance of a dipole of the same length in free space

$$R_{in}^{mp} = \frac{1}{2} R_{in}^{dp}, \qquad \beta h = 0 \tag{3.217}$$

As $\beta h \to \infty$, the radiation resistance of both antennas becomes the same, again as expected.

Short monopole radiation patterns when the height above the ground is a wavelength or less are shown in Figure 3.29. Again we see scalloping in the patterns.

The radiation intensity is given by

$$U = r^2 P = r^2 \frac{|E_\theta|^2}{2\eta_0} = \frac{\eta_0}{2}\left(\frac{I_0 \delta l}{\lambda}\right)^2 \sin^2\theta \cos^2(\beta h \cos\theta) \tag{3.218}$$

From (3.218) we can see that the maximum of $U(\theta)$ occurs at $\theta = \pi/2$ (except for $\beta h \to \infty$)

$$U_{\max} = \frac{\eta_0}{2}\left(\frac{I_0 \delta l}{\lambda}\right) \tag{3.219}$$

$U_{\max}$ therefore occurs on the horizon at an equal height (= 0) between the actual antenna and the image. This value is four times greater than $U_{\max}$ of a free-space monopole.

The maximum directivity is given by

$$D_0 = 4\pi \frac{U_{\max}}{\Pi} = \frac{2}{\frac{1}{3} - \frac{\cos(2\beta h)}{(2\beta h)^2} + \frac{\sin(2\beta h)}{(2\beta h)^3}} \tag{3.220}$$

If $\beta h = 0, D_0 = 3,$ which is twice the maximum directivity of a free-space current element $(D_0^{\text{id}} = 1.5)$, the maximum of D_0 occurs when $\beta h = 2.881$ ($h = 0.4585\lambda$). Then $D_0 = 6.566\big|_{\beta h = 2.881}$. The maximum D_0 is illustrated in Figure 3.30.

3.5.1 Horizontal Current Element Above a Perfectly Conducting Plane

A similar analysis can be done for a horizontal current element above a PEC plane yielding similar scalloping and other effects as for the vertical element.

3.5.2 Elevated Mobile Communication System Antennas

Note the effects of elevating the antenna in a cellular or PCS wireless mobile communication system. Even though the frequencies are high and therefore the wavelengths are short, based on the zone discussion above, the mobile terminals are always in the near field of such antennas. Therefore there is always a reactive component to the radiation pattern. As such, approximating an antenna pattern as we are discussing here is only a rough approximation, at best, for such systems.

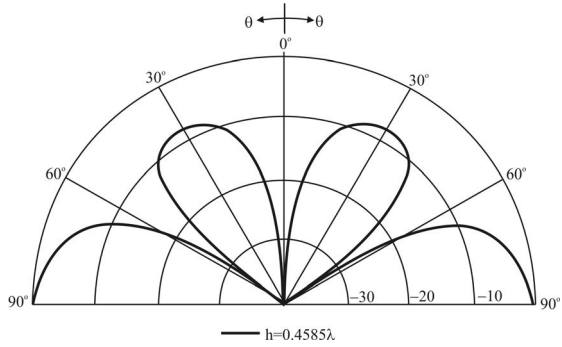


Figure 3.30 Maximum D_0.

3.6 Effects of Masts and Cables on the Patterns of Elevated Antennas

Most ground-based EW systems employ elevated antennas to extend their reach. Antennas located on the ground have very limited range and therefore little utility in general. (There are some exceptions to this; see [18] for an example.) Elevated antennas have considerably greater range, the value of which depends on how high the antenna is elevated.

The far field starts at a distance of $2l^2/\lambda$, with l as the largest dimension of the antenna. If the mast is constructed of a metallic material, however, the proper length to use includes the mast height. Even if the mast is not conductive, the feed cable, a necessity to get the signals between the antenna and whatever it is connected to on the ground, probably is (an exception might be fiber optic cables for receive applications, or when the PA is mounted on the top of the mast with the antenna). These metal components distort the antenna pattern, often considerably.

As an example, consider the patterns that arise when the antenna is a $\lambda/2$ monopole mounted on top of a 20 m metallic mast at a frequency of 1 GHz. Considering just the antenna and its image, the far field begins at 2.4 m. The resulting elevation pattern is illustrated in Figure 3.31. The scallops shown in Figure 3.31 arise because the entire antenna system is several (> 133) wavelengths long. The proper far-field distance in this case is almost 11 km from the antenna tower. (Figure 3.31 is only the first quadrant of four. In practice there will be narrow nulls in the figure 8 pattern as a result.) The horizontal pattern is still omnidirectional assuming that the diameters of the mast and/or cable are small compared to a wavelength.

For a 7 m antenna mast and other parameters the same, the far field begins at about 1.3 km.

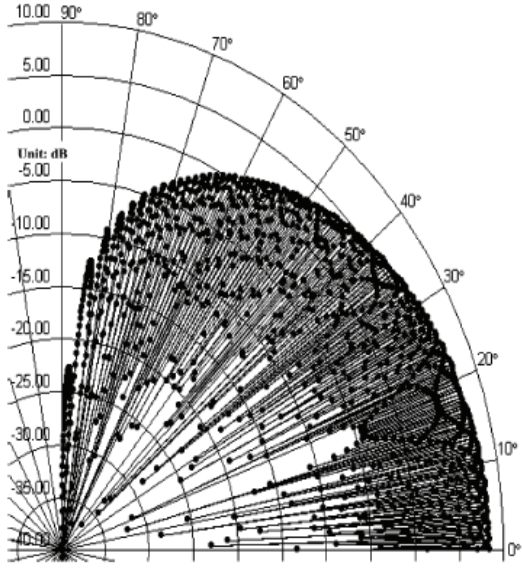


Figure 3.31 Pattern of λ/2 monopole (first quadrant). (Source: [20], © IEEE 2005. Reprinted with permission.)

For airborne EW applications these problems can be somewhat easier to handle, because in that case many of the cables can be in the same plane as the phase center of the antenna. This is the same as mentioned above for mounting the PA on top of the mast with the antenna.

Some adaptive array applications, to include EW applications, mount antennas on the top of antenna masts. When so mounted, it is important to include the mast and any feeding cables into the determination of the resulting antenna pattern. For example, a half-wave dipole at 1 GHz ($\lambda/2 = 0.15$ m) mounted on a 20 m metallic mast (or, equivalently, when there is 20 m of coaxial cable feeding the antenna from the ground level) does not have the familiar doughnut shape once the mast is taken into consideration. The antenna pattern is dramatically affected by a mast or feed cable.

This example illustrates the problem when the improper geometric parameters are used in antenna analyses. Only the far-field can have nulls; therefore, for the antenna in this example, a beam cannot truly be formed inside of about 11 km. Therefore, if we are to try to apply beamforming to cases such as these, some other method must be devised. The technology that comes to the rescue here is the Lorentz Reciprocal Theorem discussed in Appendix 3A. We will apply it in Chapter 9 to show how beams can be formed at a transmitter rather than at the receiver.

3.7 Wire Antennas

A wire antenna consists of a length of one or more segments of wire [21, 22] and are generally assumed to have a length that is much greater than their diameter. If such an antenna consists of a straight wire and is fed somewhere along its length other than at the end, such as at the center, a dipole is formed. If this antenna is fed at one end, a monopole if formed. If the wire is shaped into a circle, rectangle, or triangle, etc., a loop antenna is formed. More complex structures are common as well. The dipole wire antennas are discussed in Chapter 5, monopole antennas are discussed in Chapter 6, and loop antennas are presented in Chapter 7.

The dipole and the monopole are the two most prolific antennas for wireless mobile communication systems. The monopole is perhaps the most common antenna for portable communication equipment, such as cellular telephones, cordless telephones, automobiles, trains, and so forth. It has attractive features such as simple construction, relatively broadband characteristics, and small dimensions at high frequencies. Man-portable equipment frequently employ monopole antennas as well. The monopole has long been popular for tactical, vehicle mounted, VHF communications. An alternative to the monopole antenna for hand-held units is the loop antenna, the microstrip patch antenna, the spiral antennas, and others. We discuss many of these antenna types in the remaining chapters.

The radiation patterns of many antennas are difficult, if not impossible, to compute analytically. The characteristics, such as radiation patterns, of antenna structures for which analytic determination is not possible are determined by computer modeling which is numerically based, or measurement. For computer modeling, the integrals involved are approximated by summations. The popular method of doing this is based on the MoM [23–25].

The thickness of wire antennas can be a significant parameter, especially when the operational bandwidth is a consideration. For example, Figure 3.32 shows the calculated VSWR for two dipole antennas (considered in detail in the next section) of length 0.5 m. Using the definition of bandwidth to be the region of operation where the VSWR is less than 2, as an example, then it can be seen from Figure 5.1 that for the smaller antenna with radius $a_0 = 1\ \text{mm}$, this bandwidth is approximately 24 MHz. For the larger of the two dipoles, $a_0 = 5\ \text{mm}$, then the bandwidth is approximately 48 MHz, twice as wide.

Based on some experimental observations made by Brown and Woodward with results published in 1945 on a monopole mounted over a ground plane screen [26], the two charts shown in Figures 3.33 and 3.34 were obtained. We can see the tremendous variation of the impedance characteristics of this antenna when the ratio of length to diameter is varied.

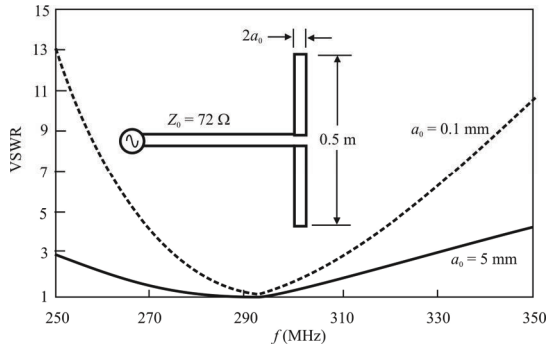


Figure 3.32 The thickness (radius) of an antenna has significant impact on its characteristics as illustrated here.

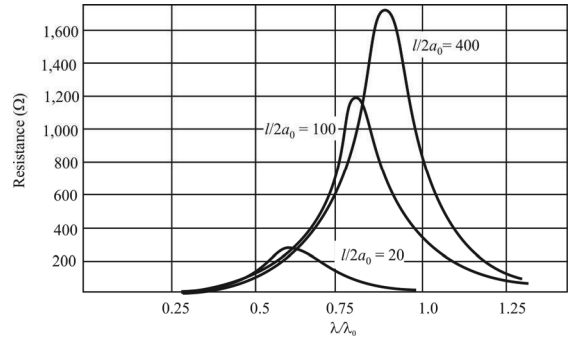


Figure 3.33 Monopole input resistance. (a_0 = radius, l = length).

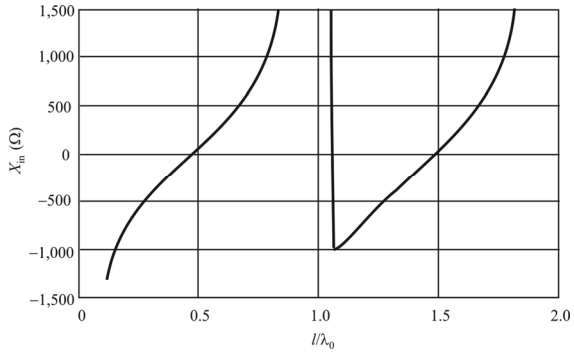


Figure 3.34 Monopole input reactance. l = length, a_0 = radius.

3.8 Concluding Remarks

In this chapter we introduced the fundamental parameters associated with antennas such as the polarization, gain, and so forth. For most applications, some of these parameters are more important than others, but none can be ignored.

These parameters are typically important for any antenna and can be hard to calculate. Sometimes the only way to obtain some of these parameters is by measurement.

References

[1] Jordan, E. C., and K. G. Balmain, *Electromagnetic Waves and Radiating Systems*, 2nd ed., Englewood Cliffs, NJ: Prentice Hall, 1968, Ch. 13.

[2] Andersen, J. B., and R. G. Vaughn, "Transmitting, Receiving, and Scattering Properties of Antennas," *IEEE Antennas and Propagation Magazine*, Vol. 45, No. 4, August 2003, pp. 93–98.

[3] Balanis, C. A., "Antenna Theory: A Review," *Proceedings of the IEEE*, Vol. 80, No. 1, January 1992, pp. 7–23.

[4] Harrington, R. F., "Matrix Methods for Field Problems," *Proceedings of the IEEE*, Vol. 55, No. 2, February 1967, pp. 136–149.

[5] Rawle, W. D., "The Method of Moments: A Numerical Technique for Wire Antenna Design," *High Frequency Electronics*, February 2006, pp. 42–47.

[6] Rogers, S. D., and C. M. Butler, "Efficient Curved-Wire Integral Equation Solution Technique," *IEEE Transactions on Antennas and Propagation*, Vol. 49, No. 1, January 2001, pp. 70–79.

[7] Tai, C.-T., and C. S. Pereira, "Approximate Formula for Calculating the Directivity of an Antenna," *IEEE Transactions on Antennas and Propagation*, March 1976, pp. 235–236.

[8] Mathis, H. F., "A Short Proof That an Isotropic Antenna Is Impossible," *Proceedings of the IRE,* Vol. 39, August 1951, p. 970, http://puhep1.princeton.edu/~mcdonald/examples/EM/mathis_procire_39_970_51.pdf.

[9] Mathis, H. F., "On Isotropic Antennas," *Proceedings of the IRE,* Vol. 42, No. 1810, 1954, http://puhep1.princeton.edu/~mcdonald/examples/EM/mathis_procire_42_1810_53.pdf.

[10] Stutzman, W. L. and G. A. Thiele, *Antenna Theory and Design,* New York: John Wiley & Sons, Inc., 1981, pp. 32–40.

[11] *Reference Data for Radio Engineers,* 6th ed., Indianapolis, IN: Howard W. Sams & Co, 1975, pp. 27–33.

[12] Pender, H. and K. McIlwain, *Electrical Engineers' Handbook Electric Communication and Electronics* 4th ed., New York: John Wiley & Sons, 1963, pp. 6-71–6-73.

[13] Yaghjian, A. D., and S. R. Best, "Impedance, Bandwidth, and Q of Antennas," *Digest IEEE Antennas and Propagation Society International Symposium, 2003*, Vol. 1, Columbus, OH, June 22–27, 2003, pp. 501–503.

[14] Rhodes, D. R., "Observable Stored Energies of Electromagnetic Systems," *Journal Franklin Institute*, Vol. 302, September 1976, pp. 225–237.

[15] Jordan, E. C., and K. G. Balmain, *Electromagnetic Waves and Radiating Systems,* Upper Saddle River, NJ: Prentice Hall, 1968, Ch. 15.

[16] Fujimoto, K., et al., *Small Antennas,* New York: Wiley, 1987.

[17] Fujimoto, K., "Modern Small Antenna Engineering," University of Tsukuba, Japan.

[18] Balanis, C. A., *Antenna Theory,* 3rd Ed., New York: Wiley, 2005, Section 4.2.

[19] Balanis, C. A., *Antenna Theory,* 3rd Ed., New York: Wiley, 2005, Section 5.2.

[20] Hwang, S., A. Medoure, and T. K. Sarkar, "Signal Enhancement in a Near-Field MIMO Environment Through Adaptivity on Transmit," *IEEE Transactions on Antennas and Propagation,* Vol. 33, No. 2, February 2005, pp. 685–694.

[21] Gagliardi, R. M., *Introduction to Communication Engineering,* New York: John Wiley & Sons, 1988, pp. 105–115.

[22] Fujimoto, K., et al., *Small Antennas,* Letchworth, Hertfordshire, UK: Research Studies Press, 1985.

[23] Harrington, R. F., "Matrix Methods for Field Problems," *IEEE Proceedings*, Vol. 55, No. 2, February 1967, pp. 136–149.

[24] Rawle, W. D., "The Method of Moments: A Numerical Technique for Wire Antenna Design," *High Frequency Electronics*, February 2006, pp. 42–45.

[25] Balanis, C. A., *Antenna Theory, Analysis and Design*, 3rd ed., New York: Wiley, 2005, pp. 450–458.

[26] Brown, G. H., and O. M. Woodward, Jr., "Experimentally Determined Impedance Characteristics of Cylindrical Antennas," *Proceedings IRE*, Vol. 33, 1945, pp. 257–262.

Appendix 3A Lorentz Reciprocity Theorem

In this appendix we examine an important theorem due to Lorentz called the *reciprocity theorem* [3A1]. With this theorem, we establish some important antenna properties.

Assume the presence of two sources in a linear, isotropic region:

$$\begin{aligned} \vec{J}^{(1)}, \vec{M}^{(1)} &\rightarrow \vec{E}^{(1)}, \vec{H}^{(1)} \\ \vec{J}^{(2)}, \vec{M}^{(2)} &\rightarrow \vec{E}^{(2)}, \vec{H}^{(2)} \end{aligned} \tag{3A.1}$$

Maxwell's equations (Section 2.5) relate these quantities as (in phasor form)

$$\begin{aligned} \nabla \times \vec{H}^{(1)} &= \vec{J}^{(1)} + j\omega\epsilon\vec{E}^{(1)} & \nabla \times \vec{H}^{(2)} &= \vec{J}^{(2)} + j\omega\epsilon\vec{E}^{(2)} \\ \nabla \times \vec{E}^{(1)} &= -\vec{M}^{(1)} - j\omega\mu\vec{H}^{(1)} & \nabla \times \vec{E}^{(2)} &= -\vec{M}^{(2)} - j\omega\mu\vec{H}^{(2)} \end{aligned} \tag{3A.2}$$

so

$$\begin{aligned} \nabla \bullet (\vec{E}^{(1)} \times \vec{H}^{(2)}) &= \vec{H}^{(2)} \bullet \nabla \times \vec{E}^{(1)} - \vec{E}^{(1)} \bullet \nabla \times \vec{H}^{(2)} \\ &= \vec{H}^{(2)} \bullet (-\vec{M}^{(1)} - j\omega\mu\vec{H}^{(1)}) - \vec{E}^{(1)} \bullet (\vec{J}^{(2)} + j\omega\epsilon\vec{E}^{(2)}) \end{aligned} \tag{3A.3}$$

$$\begin{aligned} \nabla \bullet (\vec{E}^{(2)} \times \vec{H}^{(1)}) &= \vec{H}^{(1)} \bullet \nabla \times \vec{E}^{(2)} - \vec{E}^{(2)} \bullet \nabla \times \vec{H}^{(1)} \\ &= \vec{H}^{(1)} \bullet (-\vec{M}^{(2)} - j\omega\mu\vec{H}^{(2)}) - \vec{E}^{(2)} \bullet (\vec{J}^{(1)} + j\omega\epsilon\vec{E}^{(1)}) \end{aligned} \tag{3A.4}$$

Subtracting (3A.3) from (3A.4) yields

$$\begin{aligned} -\nabla \bullet (\vec{E}^{(1)} \times \vec{H}^{(2)} - \vec{E}^{(2)} \times \vec{H}^{(1)}) \\ = (\vec{E}^{(1)} \bullet \vec{J}^{(2)} - \vec{E}^{(2)} \bullet J^{(1)}) - (\vec{H}^{(1)} \bullet \vec{M}^{(2)} - \vec{H}^{(2)} \bullet \vec{M}^{(1)}) \end{aligned} \tag{3A.5}$$

This the *differential form* of the Lorentz Reciprocity Theorem.

As a special case, suppose there is a region in the medium that is source free. That is, $\vec{J}^{(1)} = \vec{M}^{(1)} = \vec{J}^{(2)} = \vec{M}^{(2)} = \vec{J} = \vec{M} = 0$. Then

$$\nabla \bullet (\vec{E}^{(1)} \times \vec{H}^{(2)} - \vec{E}^{(2)} \times \vec{H}^{(1)}) = 0 \tag{3A.6}$$

Taking the volume integral of both sides of (3A.6) and applying the divergence theorem (2.13):

$$-\oiint_{S_v}(\vec{E}^{(1)}\times\vec{H}^{(2)}-\vec{E}^{(2)}\times\vec{H}^{(1)})\bullet\vec{n}_s\,ds$$
$$=\int\int\int_v[(\vec{E}^{(1)}\bullet\vec{J}^{(2)}-\vec{E}^{(2)}\bullet\vec{J}^{(1)})-(\vec{H}^{(1)}\bullet\vec{M}^{(2)}-\vec{H}^{(2)}\bullet\vec{M}^{(1)})]dxdydz \tag{3A.7}$$

Equation (3A.7) is the *integral form* of the Lorentz Reciprocity Theorem.

Again, in a source-free region, $\vec{J}=\vec{M}=0,$ and

$$-\oiint_{S_v}(\vec{E}^{(1)}\times\vec{H}^{(2)}-\vec{E}^{(2)}\times\vec{H}^{(1)})\bullet\vec{n}_S\,ds=0 \tag{3A.8}$$

The reciprocity theorem's greatest utility is as a means to prove certain general relationships between two antennas or an antenna and itself.

Consider a region containing two antennas, 1 and 2. Take the origin as a point midway between them. Surround both antennas with a mathematical sphere of radius $R_0 >> \lambda, R_0 >> L,$ and $R_0 >> S$ as illustrated in Figure 3A.1.

Integrating the Lorentz Reciprocity Theorem over the volume of the sphere:

$$-\oiint_{S_{v(R_0)}}(\vec{E}^{(1)}\times\vec{H}^{(2)}-\vec{E}^{(2)}\times\vec{H}^{(1)})\bullet\vec{n}_s\,ds$$
$$=\int_{v(R_0)}\int\int\Big[(\vec{E}^{(1)}\bullet\vec{J}^{(2)}-\vec{E}^{(2)}\bullet\vec{J}^{(1)})-(\vec{H}^{(1)}\bullet\vec{M}^{(2)}-\vec{H}^{(2)}\bullet\vec{M}^{(1)})\Big]dxdydz \tag{3A.9}$$

In the far zone, at the sphere's surface,

$$\vec{H}^{(1,2)}=\frac{1}{\eta_0}\vec{a}_{R_{1,2}}\times\vec{E}^{(1,2)} \tag{3A.10}$$

where

$$\vec{a}_{R_{1,2}}=\frac{\vec{R}_0-\vec{r}^{(1,2)}}{\left|\vec{R}_0-\vec{r}^{(1,2)}\right|} \tag{3A.11}$$

Thus, for R_0 in the far zone of both antennas,

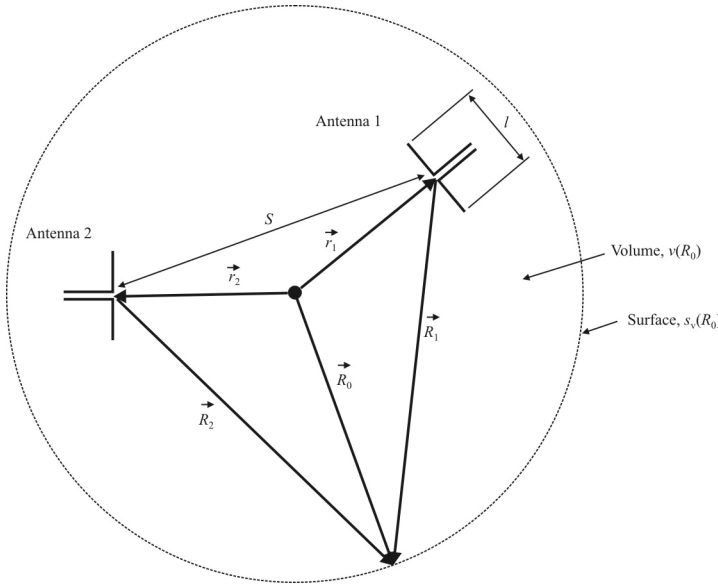


Figure 3A.1 Two antennas.

$$\vec{E}^{(1)} \times \vec{H}^{(2)} = \frac{1}{\eta_0} \vec{E}^{(1)} \times (\vec{a}_{R_2} \times \vec{E}^{(2)})$$

$$= \frac{1}{\eta_0} \left[\vec{E}^{(1)} \bullet \vec{E}^{(2)}) \vec{a}_{R_2} - (\vec{E}^{(1)} \bullet \vec{a}_{R_2}) \vec{E}^{(2)} \right] \tag{3A.12}$$

and

$$\vec{E}^{(2)} \times \vec{H}^{(1)} = \frac{1}{\eta_0} \left[\vec{E}^{(2)} \bullet \vec{E}^{(1)}) \vec{a}_{R_1} - (\vec{E}^{(2)} \bullet \vec{a}_{R_1}) \vec{E}^{(1)} \right] \tag{3A.13}$$

In the limit of a *large* sphere for which

$$\vec{a}_{R_1} \approx \vec{a}_{R_2} \rightarrow \vec{a}_{R_0} = \vec{r}, \qquad \text{as } \vec{R}_0 \rightarrow \infty$$

but in this very far zone,

$$\vec{E}^{(1)} \bullet \vec{a}_r = \vec{E}^{(2)} \bullet \vec{r} = 0$$

Hence, on the surface of this large sphere,

$$\vec{E}^{(1)} \times \vec{H}^{(2)} - \vec{E}^{(2)} \times \vec{H}^{(1)} = \frac{1}{\eta_0}(\vec{E}^{(1)} \bullet \vec{E}^{(2)})\vec{r} - \frac{1}{\eta_0}(\vec{E}^{(2)} \bullet \vec{E}^{(1)})\vec{r} = 0 \qquad \text{(3A.14)}$$

and

$$\begin{aligned} &\oiint_{s_v} (\vec{E}^{(1)} \times \vec{H}^{(2)} - \vec{E}^{(2)} \times \vec{H}^{(1)}) \bullet \vec{n}_s ds \\ &\qquad = 0 = \iiint_v (\vec{E}^{(1)} \bullet \vec{J}^{(2)} - \vec{H}^{(1)} \bullet \vec{M}^{(2)} - \vec{E}^{(2)} \bullet \vec{J}^{(1)} + \vec{H}^{(2)} \bullet \vec{M}^{(1)}) dxdydz \end{aligned} \qquad \text{(3A.15)}$$

or

$$\iiint_v (\vec{E}^{(1)} \bullet \vec{J}^{(2)} - \vec{H}^{(1)} \bullet \vec{M}^{(2)}) dxdydz = \iiint_v (\vec{E}^{(2)} \bullet \vec{J}^{(1)} - \vec{H}^{(2)} \bullet \vec{M}^{(1)}) dxdydz \qquad \text{(3A.16)}$$

Even though (3A.16) was derived for a large spherical volume v and no reflections, it also applies to a region enclosed by a perfect electric or "magnetic" conductors $(\mu \to \infty)$ even if the volume is not large. This can be seen from the fact that

$$(\vec{E}^{(1)} \times \vec{H}^{(2)}) \bullet \vec{n}_s ds = -\vec{H}^{(2)} \bullet (\vec{E}^{(1)} \times \vec{n}_s ds) = -\vec{H}^{(2)} \bullet (\vec{E}^{(1)\text{tang}} \times \vec{n}_s ds) \qquad \text{(3A.17)}$$

and/or

$$(\vec{E}^{(1)} \times \vec{H}^{(2)}) \bullet \vec{n}_s ds = \vec{E}^{(1)} \bullet (\vec{H}^{(2)} \times \vec{n}_s ds) = \vec{E}^{(1)} \bullet (\vec{H}^{(1)\text{tang}} \times \vec{n}_s ds) \qquad \text{(3A.18)}$$

Note that $\vec{E}_t = 0$ at the surface of a PEC and $\vec{H}_t = 0$ at the surface of a PMC.

Thus, we can conclude that, in general,

$$\iiint_v (\vec{E}^{(1)} \bullet \vec{J}^{(2)} - \vec{H}^{(1)} \bullet \vec{M}^{(2)}) dxdydz = \iiint_v (\vec{E}^{(2)} \bullet \vec{J}^{(1)} - \vec{H}^{(2)} \bullet \vec{M}^{(1)}) dxdydz \qquad \text{(3A.19)}$$

Note that although the integral is over volume v (possibly infinite),

$$\vec{J}^{(1)} = \vec{M}^{(1)} = 0, \text{ except over the volume of antenna 1}$$
$$\vec{J}^{(2)} = \vec{M}^{(2)} = 0, \text{ except over the volume of antenna 2}$$

Therefore, we can say that

$$\int\int_{v^{(2)}}\int (\vec{E}^{(1)} \bullet \vec{J}^{(2)} - \vec{H}^{(1)} \bullet \vec{M}^{(2)}) dxdydz = \int\int_{v^{(1)}}\int (\vec{E}^{(2)} \bullet \vec{J}^{(1)} - \vec{H}^{(2)} \bullet \vec{M}^{(1)}) dxdydz \tag{3A.20}$$

These integrals are called *reaction integrals*.

Equation (3A.20), which, as we have just seen, is a consequence of Lorentz reciprocity, basically says that coupling is reciprocal. That is, antenna 1 effect on antenna 2 equals antenna 2 effect on antenna 1, as long as surrounding both antennas is a boundary on which (to a reasonable approximation), $\vec{E}^{\,\text{tang}} = 0, \vec{H}^{\,\text{tang}} = 0$, or there are good absorbing materials. (Different parts of this boundary are allowed to have any one of these three properties.)

Let

$$< i, j > = \int\int_{v^{(j)}}\int (\vec{E}^{(i)} \bullet \vec{J}^{(j)} - \vec{H}^{(i)} \bullet \vec{M}^{(j)}) dxdydz \tag{3A.21}$$

denote the reaction (coupling) of fields $\vec{E}^{(i)}$ and $\vec{H}^{(i)}$ from antenna *i* to the currents $\vec{J}^{(j)}$ and $\vec{M}^{(j)}$ on antenna *j*.

Let us now consider an illustrative application to small linear dipole antennas as represented in Figure 3A.2. From (3A.21)

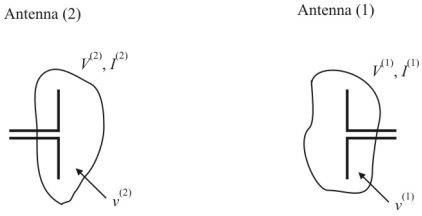


Figure 3A.2 Two dipoles.

$$
\begin{aligned}
<1,2> &= \int \int_{v^{(2)}} \int \vec{E}^{(1)} \bullet \vec{J}^{(2)} dxdydz = \int \vec{E}^{(1)} \bullet \int\int \vec{J}^{(2)} dxdydz \\
&= \int \vec{E}^{(1)} \bullet (I_2 d\vec{l}_2) \\
&= I^{(2)} \int E^{(1)} \bullet d\vec{l}_2
\end{aligned}
\tag{3A.22}
$$

(assuming that $I^{(2)}$ is constant over the length of antenna 2). Thus

$$
<1,2> = -V^{(21)} I^{(2)} \tag{3A.23}
$$

where $V^{(21)}$ is the response at antenna 2 due to current $I^{(1)}$ at antenna 1.

In general, we might describe the coupling between the two antennas as

$$
\begin{aligned}
V^{(1)} &= V^{(11)} + V^{(12)} \triangleq Z_{11} I_1 + Z_{12} I_2 \\
V^{(2)} &= V^{(21)} + V^{(22)} \triangleq Z_{21} I_1 + Z_{22} I_2
\end{aligned}
\tag{3A.24}
$$

when

$$
Z_{ij} = \frac{V^{(ij)}}{I_j} \tag{3A.25}
$$

yet, we also can show that

$$
-V_{ij} I_i = < j,i > \tag{3A.26}
$$

or

$$
V_{ij} = -\frac{< j,i >}{I_i} \tag{3A.27}
$$

Thus, the interantenna coupling or "transimpedance" can be described by

$$
Z_{ij} = -\frac{< j,i >}{I_i I_j} = -\frac{\int \int_{v^{(i)}} \int \vec{E}^{(j)} \vec{J}^{(i)} dxdydz}{I_i I_j} = -\frac{I_i \int \vec{E}^{(j)} \bullet d\vec{l}_i}{I_i I_j} \tag{3A.28}
$$

From the above development we can draw several important conclusions that apply to all antennas even though the transimpedance formulation was for the illustration of small, linear diploes.

- From (3A.20) we can see that $<i, j> = <j, i>$, which is a consequence of Lorentz reciprocity. So then

$$Z_{ij} = -\frac{< j,i >}{I_i I_j} = -\frac{< i,j >}{I_i I_j} = Z_{ji} \tag{3A.29}$$

- A radiating element receives (i.e., $Z_{ij} \neq 0$) only the polarization component that it transmits in the direction being considered (this includes circular and elliptical polarizations). For example, for our illustration with linear dipole antennas,

$$Z_{ij} = 0 \text{ unless } \int \vec{E}^{(j)} \bullet d\vec{l}_i \neq 0$$

  which says that the antenna must have proper polarization to receive or transmit a certain polarization of $\vec{E}$.
- The directional pattern of a radiating element is the same whether transmitting or receiving.

The last property is illustrated with the aid of Figures 3A.3 and 3A.4. The ratio of the time-averaged power received by antenna 2 and dissipated in the receiver load (Z_{receiver}) is

$$\frac{P_{1\to 2}(\theta_A, \phi_A)}{P_{1\to 2}(\theta_B, \phi_B)} = \frac{\left|Z_{21}(\theta_A, \phi_A)\right|^2}{\left|Z_{21}(\theta_B, \phi_B)\right|^2} \tag{3A.30}$$

Now consider the same configuration, but reverse the roles of the transmitter and the receiver (see Figure 3A.4). The ratio of power received and absorbed in the load Z_T for the two positions is

$$\frac{P_{2\to 1}(\theta_A, \phi_A)}{P_{2\to 1}(\theta_B, \phi_B)} = \frac{\left|Z_{21}(\theta_A, \phi_A)\right|^2}{\left|Z_{21}(\theta_B, \phi_B)\right|^2} \tag{3A.31}$$

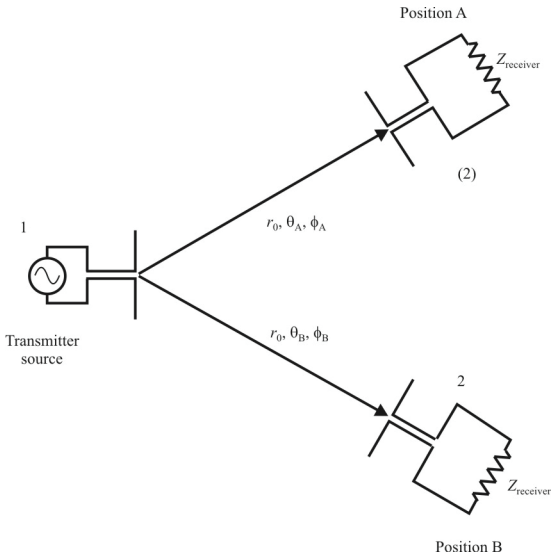


Figure 3A.3 Identical patterns.

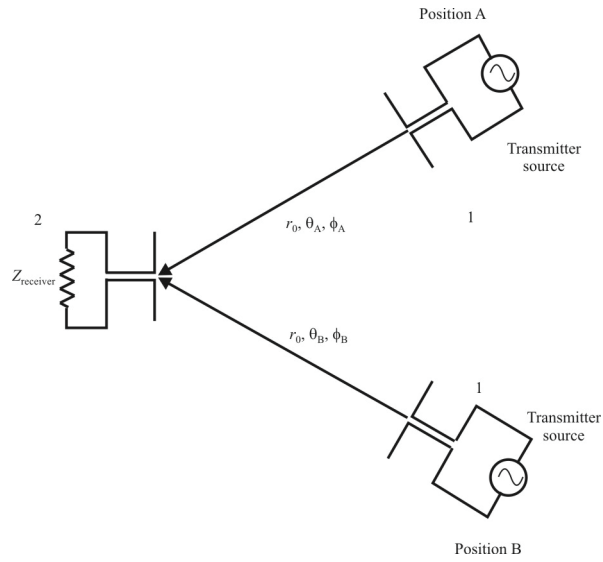


Figure 3A.4 Identical patterns reversed.

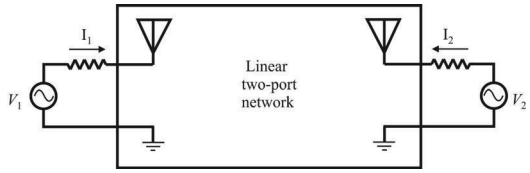


Figure 3A.5 Linear two-port.

but, by reciprocity, $Z_{12}(\theta,\phi) = Z_{21}(\theta,\phi)$. This results in

$$\frac{P_{1\to2}(\theta_A,\phi_A)}{P_{1\to2}(\theta_B,\phi_B)} = \frac{P_{2\to1}(\theta_A,\phi_A)}{P_{2\to1}(\theta_B,\phi_B)} \tag{3A.32}$$

which says that *the transmit and receive patterns of the antennas are identical*.

Now consider a two-antenna system as a two-port linear network. According to circuit theory, for the (properly terminated) linear two-port network shown in Figure 3A.5, $V_1 I_1 = V_2 I_2$, if a voltage, V_1, applied at port 1 produces a current, I_2, at port 2 and if a voltage, V_2, applied at port 2 produces a current, I_1, at port 1.

Reciprocity states that the source and the measurement points can be interchanged without changing the system response. If V_1 is the voltage applied to antenna 1, current I_2 is induced on antenna 2. Similarly, when voltage V_2 is applied to antenna 2, current I_1 is induced on antenna 1. Reciprocity implies that $V_1(\omega)I_1(\omega) = V_2(\omega)I_2(\omega)$. Consequently, the input voltage, $v_2(t)$, required to produce the induced current, $i_1(t)$, is obtained. If we know $v_1(t)$ and the corresponding $i_2(t)$, then

$$v_2(t) = \mathcal{F}^{-1}\left\{\frac{V_1(\omega)I_1(\omega)}{I_2(\omega)}\right\} \tag{3A.33}$$

where $\mathcal{F}^{-1}\{\cdot\}$ denotes the inverse Fourier transform. Thus, the input required for a specified induced current can be synthesized by using two identical antennas. Similarly, to synthesize the input for a specified radiation, we need to have a single antenna and use the radiated fields instead of the induced currents in (3A.33). If the radiation due to input voltage V_1 is R_1 and if R_2 is the required radiated pulse, the requisite time-domain input voltage is calculated from

$$v_2(t) = \mathcal{F}^{-1}\left\{\frac{V_1(\omega)R_2(\omega)}{R_1(\omega)}\right\} \tag{3A.34}$$

by the reciprocity principle. In a similar manner, the principle can be modified for application to the reception problem. The goal here is to illustrate that by proper waveform shaping, any pulse shape can be transmitted or received.

Reference

[3A.1] University of Wisconsin ECE 740 Notes, 2004.

Chapter 4

Transmission Lines

4.1 Introduction

Transmission lines are used to route RF signals to and from the antennas. Their characteristics have such a significant impact on the performance of these systems that inclusion of a discussion of their features is an imperative in any discussion of antenna systems. Note that another use for transmission lines is impedance matching at RF and higher frequencies. We will cover that aspect in detail in Chapter 20.

This chapter is structured as follows. We begin with a description and analysis of the properties of lossless transmission lines. After that, losses that are inevitably present are introduced. Transmission lines can be used as antenna matching elements, and we cover terminating lines next. That is followed by a discussion of the velocity factor. Waveguides are transmission lines for microwave frequencies and they are discussed next. The last topic covered is a presentation of actual cables used for transmission lines.

4.2 Characteristics of Transmission Lines

Transmission lines transfer the RF power from the output of the power amplifiers (PAs) to the antenna(s) and vice versa—they transfer RF energy from the antenna to the receiving subsystem.

In this chapter, we will model a transmission line with distributed circuit elements and examine the voltage and current signals that propagate along that line. The circuit elements include distributed inductors, capacitors, and resistors. The equations that we will derive are called the *telegrapher's equations* because they were originally derived to describe the signals that propagated on the cables

that spanned the United States. Using the telegrapher's equations, we will derive a wave equation that will describe the wave propagation on the transmission line.

In our initial discussion of transmission lines, we will assume that the wave suffers no loss as it propagates and will therefore not include any resistors in the introductory model. Loss terms will be included after we develop the lossless transmission line.

There are three fundamental uses for transmission lines in systems. The first is to convey the RF signals from the antenna(s) to/from the receiving/transmitting components. These generally tend to be relatively long compared to a wavelength, especially for microwave and higher frequencies. The second is as components within microwave circuits. The characteristics of transmission lines vary with the type of loading applied at the output end (as well with other parameters such as frequency and characteristic impedance). They can be short or open circuits at particular frequencies, which makes them useful as frequency-selective signal processing tasks. The third use is as transformers for impedance conversion.

As we discussed in Chapter 2, all EM waves consist of electric and magnetic fields propagating in the same direction of travel, but perpendicular to each other. Along the length of a normal transmission line, both electric and magnetic fields are perpendicular (transverse) to the direction of wave travel. This is known as the *principal mode*, or TEM mode. This mode of wave propagation can exist only where there are two conductors, and it is the dominant mode of wave propagation when the cross-sectional dimensions of the transmission line are small compared to the wavelength of the signal.

The system under consideration in this chapter is shown in Figure 4.1. It consists of a source of signals, called the generator in Figure 4.1, the transmission line, which is characterized mainly by its characteristic impedance Z_0, and a load. The generator is modeled with a lossless signal source in series with source impedance, Z_S.

A depiction of a twin-lead transmission line is given in Figure 4.2 where the magnetic and electric fields are shown. We just show a few of the electric and magnetic flux lines in Figure 4.2, whereas in theory these lines extend to infinity in three dimensions.

The dielectric material between the conductors for most transmission lines, including the one in Figure 4.2, are *double positive* (DP), meaning that $\varepsilon_r > 0$ and $\mu_r > 0$, although man-made dielectric materials have been discovered where this is not true. They are called *double negative* (DN) when both $\varepsilon_r < 0$ and $\mu_r < 0$, or when only one of these parameters is less than zero, they are called *single negative* (SN) dielectrics (see Chapter 14). Interesting and unusual properties result when one or both of these parameters is less than zero. Double negative material coating for wire antennas is discussed in Chapter 14, where their properties for decreasing antenna size are described.

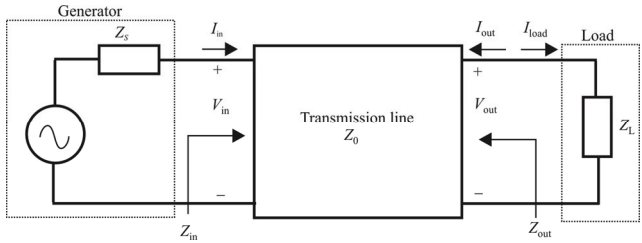


Figure 4.1 Transmission line diagram.

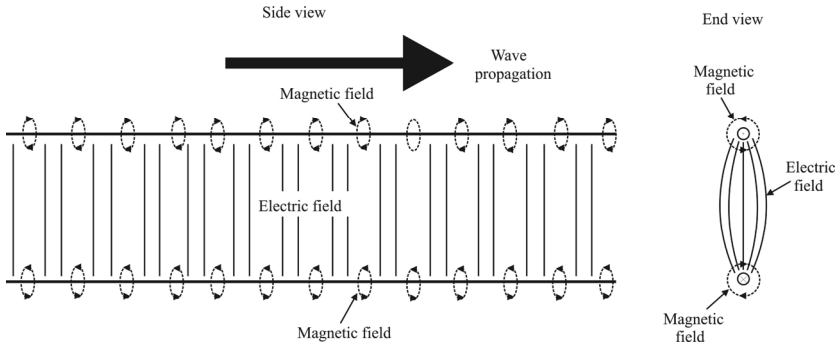


Figure 4.2 Twin-lead transmission line propagation. TEM mode. Both field planes are perpendicular (transverse—the *T* in TEM) to the direction of propagation. In this example, the dielectric is double positive (DP), that is, $\varepsilon_r > 0$ and $\mu_r > 0$, so that the wave propagation behaves according to the right-hand rule.

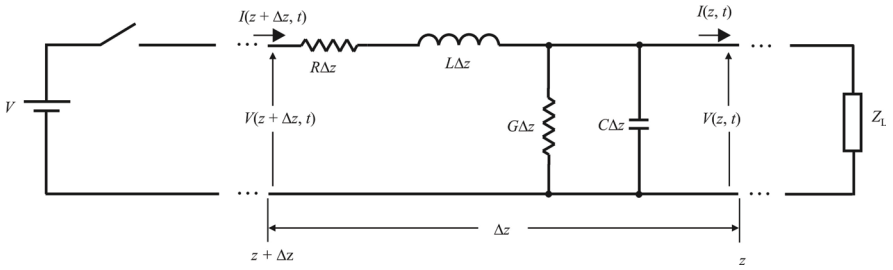


Figure 4.3 Transmission line model.

In fixed installations, to include aircraft installations, the transmission line is usually fixed in place at time of installation. Therefore, the characteristics of the T-line can be established and they will remain more or less the same until something unusual happens, maintenance, for example.

In ground mobile installations, antennas are typically movable and moved. In order to be mobile, an antenna cannot be erected very far above the roof of the vehicle in which it is installed. By international agreement this is typically 4 m. At this height, received signals are likely to be quite weak unless the targets are rather close to the receiving system. It is more common to elevate the ground mobile antenna using an antenna mast. This causes the T-line characteristics to change with each elevation and retraction. Of course, the same comments apply to transmitters as well.

The general lossy transmission line depicted in Figure 4.1 is modeled with an equivalent circuit consisting of an infinite number of distributed inductors, capacitors, and resistors as shown in Figure 4.3 (for a lossless line $R = G = 0$ in Figure 4.3). This model has incorporated some obvious simplifications. In particular, parasitic capacitances exist between the wires that constitute the distributed inductance in Figure 4.3. These capacitances will be assumed to be very small and can be neglected at this stage of our discussion; however, as the frequency of operation is increased, they may come into play and so cannot be ignored.

4.2.1 Velocity Factor

From Maxwell's equations we know that when the line is lossless and if L and C are constant then the speed of propagation is given by

$$v = \frac{1}{\sqrt{LC}} \tag{4.1}$$

Table 4.1 Velocity Factors

Cable	Velocity Factor	Cable	Velocity Factor
RG-8	0.66	RG-214	0.66
LMR-400	0.85	RG-217	0.66
RG-8X	0.84	RG-218	0.66
RG-11	0.75	RG-316	0.79
RG-58	0.66	RG-400	0.70
LMR-195	0.83	LMR-500	0.85
RG-59	0.82	LMR-600	0.86
RG-62	0.84	½ HARD	0.81
RG-174	0.66	7/8 HARD	0.81
RG-213	0.66	LDF all versions	0.88

which is equal to the speed of light c in a vacuum or to a fraction of c in an insulating medium; therefore, we may introduce a speed factor. The *velocity factor* of a transmission line is the rate of signal propagation through the line compared to the rate of propagation through a vacuum. Typical velocity factors vary from about 65% to as high as 97% of free space propagation, depending on the materials used for the conductors and the characteristics of the dielectric material employed. The velocity factor is given by

$$\text{Velocity factor} = \frac{v}{c} = \frac{1}{\sqrt{\varepsilon_r}} \tag{4.2}$$

where v is the velocity of propagation in the medium, c is the velocity of light in a vacuum, and ε_r is the relative dielectric constant of the insulation between the conductors. The velocity factor of flexible coaxial transmission line types vary according to the materials used in their manufacture. Table 4.1 illustrates the velocity factors for a few common types of cables. v depends on the electrical properties of the materials that constitute the line: the conductors and insulators. When the length of the line is l, assuming the switch in Figure 4.3 is closed at $t = 0$; the signal will not appear at the load impedance until $t = l/v$.

4.2.2 Characteristic Impedance

The impedance at any point on the line that loads the source is given by Z_0. This is called the *characteristic impedance* of the line and is given by

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \tag{4.3}$$

For materials commonly used for cable insulation in EW systems, G is frequently small enough (leakage resistance is large) that it can be neglected when compared with ωC.

Depending on the termination impedance and the length of the line, reflection of power can occur. If the length of the line is modeled as being infinite then no reflections will occur. If the length of the line is finite and the load is a conjugate match to the characteristic impedance of the line $(Z_0 = R_0 + jX_0,\ Z_L = R_L + jX_L = R_0 - jX_0)$ then no reflections will occur either. In all other real cases reflections occur, and standing waves are established on the line. These standing waves are due to the interaction of the forward and reflected waves on the transmission line.

When the switch is closed, voltage and current wavefronts begin to propagate down the line from the source to the load. When the wavefront moves a distance dz, the capacitance charged at voltage V is Cdz. Since

$$Q = CV \tag{4.4}$$

then

$$dQ = CVdz \tag{4.5}$$

Also,

$$I = \frac{dQ}{dt} = VC\frac{dz}{dt} \tag{4.6}$$

But $dz/dt = v$ so that using (4.1)

$$I = VCv = VC\frac{1}{\sqrt{LC}} = V\sqrt{\frac{C}{L}} = \frac{V}{Z_0} \tag{4.7}$$

4.2.3 TEM Wave Propagation

The notion of TEM transmission lines has been known for quite some time and is taught in most electrical engineering curriculums. The term TEM mode refers to when both the electric and magnetic fields are parallel to a boundary plane and there are no longitudinal components of either field (this parallelism could be tangential, when, for example, the transmission line is round as in coaxial cables). The propagation of energy is down the transmission line from the source to the

load so is in the plane that is orthogonal to both the electric field and the magnetic field.[1]

Other modes are called the TE and TM modes. This is where the electric field or magnetic field, respectively, of a propagating wave is parallel to a boundary plane, in this case being the surface of the conductors of a transmission line, while at the same time the accompanying magnetic or electric fields, respectively, still have some longitudinal (or axial) components. Both of these terms are normally associated with waveguides.

The TEM mode is the only mode associated with transmission lines as we describe in this chapter. The TE and TM modes require separation between the conductors on the order of a wavelength, while for the TEM mode, separation can be much less than a wavelength. Therefore, waveguides with their TE and TM modes are applicable at higher microwave frequencies while transmission lines are usable essentially down to DC.

4.2.4 Lossless Transmission Lines

4.2.4.1 Reflections

Assuming that the line is lossless, when the switch is closed, the wavefront propagates toward the load. As mentioned, when it gets there, it may or may not be reflected depending on the termination. On the left, the voltage source sees $Z_0 = V_{\text{in}} / I_{\text{in}}$ looking into the line. At the load end of the line, $V_{\text{L}} / I_{\text{L}} = Z_{\text{L}}$. If the ratio $V_{\text{L}} / I_{\text{L}}$ changes because $R_L \neq \text{Re}\{Z_0\}$, we have reflected wavefronts traveling in the reverse direction with respect to the incident one. It follows that the sign of the current is reversed.

$$V_{\text{L}} = V_{\text{incident}} + V_{\text{reflected}} \tag{4.8}$$

$$I_{\text{L}} = I_{\text{incident}} + I_{\text{reflected}} \tag{4.9}$$

$$V_{\text{L}} = V + \Gamma V \tag{4.10}$$

where Γ is the reflection coefficient. The reflected current is

$$\text{Reflected current} = I_{\text{reflected}} = -\Gamma \frac{V_{\text{L}}}{Z_0} \tag{4.11}$$

because $V_{\text{L}} / I_{\text{L}} = Z_0$.

[1] In actuality, the energy is transferred from the source to the load via the electric and magnetic fields associated with the current in the line [1, 2]. For our purposes it is sufficient to envision the current itself as the transfer mechanism.

$$\frac{V+\Gamma V}{V/Z_0-\Gamma V/Z_0}=Z \tag{4.12}$$

so

$$\Gamma=\frac{Z/Z_0-1}{Z/Z_0+1}=\frac{z-1}{z+1} \tag{4.13}$$

or

$$\Gamma=\frac{Z-Z_0}{Z+Z_0} \tag{4.14}$$

This expression can be inverted to give

$$\frac{Z}{Z_0}=\frac{1+\Gamma}{1-\Gamma} \tag{4.15}$$

Admittances can be normalized in the same way.

4.2.4.2 Telegrapher's Equations

Kirchhoff's laws are used to analyze the circuit shown in Figure 4.3. Assuming a lossless line so that that $R = G = 0$, the current entering the node at the location z is I(z). Using Kirchhoff's current law

$$I(z+\Delta z,t)=C\Delta z\frac{\partial V(z,t)}{\partial t}+I(z,t) \tag{4.16}$$

where the first term on the right is the current passing through the capacitor and the second term is the current out of the node into the next node located at z. Equation (4.16) it can be rewritten as

$$\frac{I(z+\Delta z)-I(z,t)}{\Delta z}=C\frac{\partial V(z,t)}{\partial t} \tag{4.17}$$

Finding the limit as $\Delta z \to 0$ yields

$$\lim_{\Delta z \to 0} \frac{I(z+\Delta z) - I(z,t)}{\Delta z} = \frac{\partial I(z,t)}{\partial z} \tag{4.18}$$

Thus

$$\frac{\partial I(z,t)}{\partial z} = C \frac{\partial V(z,t)}{\partial t} \tag{4.19}$$

Similarly, using the same lossless assumptions, Kirchhoff's voltage law around the loop yields

$$V(z+\Delta z) = L\Delta z \frac{\partial I}{\partial t} + V(z,t) \tag{4.20}$$

where the first term on the right is the voltage across the inductor and the second term is the voltage at the node at z. Rewriting (4.20) yields

$$\frac{V(z+\Delta z,t) - V(z)}{\Delta z} = L \frac{\partial I(z,t)}{\partial t} \tag{4.21}$$

Taking the limit as before results in

$$\frac{\partial V(z,t)}{\partial z} = L \frac{\partial I(z,t)}{\partial t} \tag{4.22}$$

The two linear coupled first-order partial differential equations (4.19) and (4.22) are known as the telegrapher's equations. Second-order partial differential equations can be obtained from (4.19) and (4.22) that eliminate one of the dependent variables $V(z, t)$ or $I(z, t)$ from each of them. Differentiating (4.19) with respect to z yields

$$\frac{\partial^2 I(z,t)}{\partial z^2} = C \frac{\partial}{\partial z} \frac{\partial V(z,t)}{\partial t} \tag{4.23}$$

Assuming the differentiations on the right can be reversed, which is normally the case, and using (4.22), then

$$\frac{\partial^2 I(z,t)}{\partial z^2} = LC \frac{\partial^2 I(z,t)}{\partial t^2} \tag{4.24}$$

A similar derivation yields

$$\frac{\partial^2 V(x,t)}{\partial x^2} = LC\frac{\partial^2 V(x,t)}{\partial t^2} \tag{4.25}$$

Assume that the transmission line is infinite in length and is excited with a sinusoidal voltage generator at $z = -\infty$ or $z = +\infty$ and all the transients have settled out. In this case, a sinusoidal voltage and current wave will propagate down the line, and we can separate out the time dependence from the spatial term as

$$V(z,t) = V(z)e^{j\omega t} \quad \text{and} \quad I(z,t) = I(z)e^{j\omega t} \tag{4.26}$$

where $V(z)$ and $I(z)$ are *phasors*. These phasors have an amplitude and may have a nonzero phase and are functions of distance on the line, but they have no time dependence:

$$V(z) = \left|V(z)\right|e^{j\theta_V} \quad \text{and} \quad I(z) = \left|I(z)\right|e^{j\theta_I} \tag{4.27}$$

The voltage and the current both have the same time dependence because they are coupled by the first-order partial differential telegrapher's equations (4.19) and (4.22). Using the velocity in a lossless line given by (4.1), substituting (4.27) into (4.24) and (4.25) yields

$$\frac{\partial^2 V(z)}{\partial z^2} - \frac{(j\omega)^2}{v^2}V(z) = 0 \tag{4.28}$$

and

$$\frac{\partial^2 I(z)}{\partial z^2} - \frac{(j\omega)^2}{v^2}I(z) = 0 \tag{4.29}$$

which can be rewritten as

$$\frac{\partial^2 V(z)}{\partial z^2} + \beta^2 V(z) = 0 \tag{4.30}$$

and

$$\frac{\partial^2 I(z)}{\partial z^2} + \beta^2 V(z) = 0 \tag{4.31}$$

where β is the *wave number*,

$$\beta = \frac{\omega}{v} \tag{4.32}$$

Because $\omega = 2\pi f$ and $f = v/\lambda$, β can also be expressed as

$$\beta = \frac{2\pi}{\lambda} \tag{4.33}$$

Solutions of (4.30) can be written as

$$V(z,t) = A_1 e^{j(\omega t - \beta z)} + A_2 e^{j(\omega t + \beta z)} \tag{4.34}$$

The current wave $I(z, t)$ can be calculated by inserting the voltage wave solution (4.34) into either (4.19) or (4.22). The result is

$$I(z,t) = \frac{A_1}{Z_0} e^{j(\omega t - \beta z)} - \frac{A_2}{Z_0} e^{j(\omega t + \beta z)} \tag{4.35}$$

If we assume that the line is infinitely long and the wave is launched from the point $z = -\infty$ and is propagating to increasing values of the coordinate z, then the second term in (4.34) is zero. Using the real component of (4.34), then from (4.22)

$$\frac{\partial V_0 \cos(\omega t - \beta z)}{\partial z} = -L \frac{\partial I_0 \cos(\omega t - \beta z)}{\partial t} \tag{4.36}$$

$$V_0 \beta = L \omega I_0 \tag{4.37}$$

From (4.37) we see that

$$Z_0 = \frac{V_0}{I_0} = L \frac{\omega}{\beta} \tag{4.38}$$

Notice that this expression for Z_0 is independent of C.

The ratio ω/β is the phase velocity v of the wave. This velocity is given by (4.1), so

$$Z_0 = \frac{V_0}{I_0} = Lv = L\frac{1}{\sqrt{LC}} = \sqrt{\frac{L}{C}} \tag{4.39}$$

4.2.5 Lossy Transmission Lines

In reality there are losses in transmission lines. These losses are due to series resistance of the wires and leakage currents through the insulators, although the latter can frequently be ignored in modern, high-quality cables. When the series resistance and shunt conductance elements in Figure 4.3 are included, the Kirchhoff equations become

$$\frac{\partial I(z,t)}{\partial z} = C\frac{\partial V(z,t)}{\partial t} + GV(z,t) \tag{4.40}$$

$$\frac{\partial V(z,t)}{\partial z} = L\frac{\partial I(z,t)}{\partial t} + RI(z,t) \tag{4.41}$$

where, as before, C, G, L, and R are the values per unit length. Equations (4.40) and (4.41) are also known as the telegrapher's equations.

It is often convenient to assume that there is a sinusoidal excitation of the transmission line. In that case, (4.40) and (4.41) can be rewritten as

$$\frac{\partial I(z)}{\partial z} = [G + j\omega C]V(z) \tag{4.42}$$

$$\frac{\partial V(z)}{\partial z} = [R + j\omega L]I(z) \tag{4.43}$$

where the current and voltage, $I(z)$ and $V(z)$, are in phasor notation. The variation with time is assumed and therefore not explicitly noted; only the variation with distance on the line is specifically included. The terms within the square brackets can be replaced with distributed admittance and distributed impedance quantities, respectively, producing

$$\frac{dI(z)}{dz} = YV(z) \tag{4.44}$$

$$\frac{dV(z)}{dz} = ZI(z) \tag{4.45}$$

The coupled first-order ordinary differential equations can be replaced with a second-order ordinary differential equation for either of the dependent variables. From (4.44)

$$\frac{d^2 I(z)}{dz^2} = Y \frac{dV(z)}{dz} \tag{4.46}$$

and substituting this into (4.45) yields

$$\frac{d^2 I(z)}{dz^2} = ZYI(z) = \gamma^2 I(z) \tag{4.47}$$

where

$$\gamma = \sqrt{ZY} = \alpha + j\beta \tag{4.48}$$

is known as the *propagation constant* with a real part α and an imaginary part β. The real part, α, known as the *attenuation function*, indicates the change in magnitude of the voltage or current, while the imaginary part, β, called the *phase function*, indicates the change in phase of the voltage or current, and, as previously, $\beta = 2\pi / \lambda$. In addition, the attenuation function (constant) is given by

$$\alpha = R\sqrt{\frac{L}{C}} \tag{4.49}$$

A transmission line is *distortionless* if its properties do not vary with frequency. Such a line can be distortionless only over a limited frequency band, as lines are characterized by distributed values of capacitance, inductance, resistance, and conductivity, all of which, in general, are functions of frequency. For a distortionless transmission line,

$$\frac{R}{L} = \frac{G}{C} \tag{4.50}$$

Example: The attenuation on a 50 Ω distortionless transmission line is 0.01 dB m^{-1}. The line has a capacitance of 0.1×10^{-9} F m^{-1}.

The characteristic impedance is

$$Z_0 = \sqrt{\frac{L}{C}} = 50\,\Omega$$

so

$$L = Z_0^2 C = 50^2 \times 0.1 \times 10^{-9} = 2.5 \times 10^{-7}\,\mathrm{H\,m^{-1}}$$

The attenuation constant is

$$\alpha = R\sqrt{\frac{L}{C}} = 0.01\,\mathrm{dB\,m^{-1}}$$

but

$$x\,\mathrm{Np\,m^{-1}} = 20\log_{10} e^x$$

so

$$1\,\mathrm{Np\,m^{-1}} = 20\log_{10}(e) = 8.686\,\mathrm{dB\,m^{-1}}$$

so that

$$\alpha = \frac{0.01}{8.686} = 0.0012\,\mathrm{Np\,m^{-1}}$$

Therefore,

$$R = \alpha\sqrt{\frac{C}{L}} = \alpha Z_0 = 0.0012 \times 50 = 0.0575\,\Omega\,\mathrm{m^{-1}}$$

The distortionless line criterion is

$$\frac{R}{L} = \frac{G}{C} \Rightarrow G = \frac{RC}{L} = \frac{R}{Z_0^2} = \frac{0.0575}{50^2} = 2.3 \times 10^{-5}\,\mathrm{S\,m^{-1}}$$

The phase velocity is

$$v = \frac{1}{\sqrt{LC}} = \frac{1}{\sqrt{2.5 \times 10^{-7} \times 0.1 \times 10^{-9}}} = 2 \times 10^8\,\mathrm{m\,s^{-1}}$$

Likewise,

$$\frac{d^2V(z)}{dz^2} = ZYV(z) = \gamma^2 V(z) \tag{4.51}$$

The solutions to (4.47) and (4.51) are

$$V(z,t) = A_1 e^{j(\omega t+\gamma z)} + A_2 e^{j(\omega t-\gamma z)} \tag{4.52}$$

$$I(z,t) = B_1 e^{j(\omega t+\gamma z)} + B_2 e^{j(\omega t-\gamma z)} \tag{4.53}$$

where

$$\gamma = \alpha + j\beta = \sqrt{ZY} = \sqrt{(R + j\omega L)(G + j\omega C)} \tag{4.54}$$

Substituting (4.52) and (4.53) into (4.45) yields

$$\gamma A_1 e^{j(\omega t+\gamma z)} - \gamma A_2 e^{j(\omega t-\gamma z)} = ZB_1 e^{j(\omega t+\gamma z)} + ZB_2 e^{j(\omega t-\gamma z)} \tag{4.55}$$

therefore

$$ZB_1 = \gamma A_1 \tag{4.56}$$

$$ZB_2 = -\gamma A_2 \tag{4.57}$$

so, using (4.48) in (4.56), we get

$$B_1 = -\frac{A_1}{Z_0} \tag{4.58}$$

And, using (4.48) in (4.57), we get

$$B_2 = \frac{A_2}{Z_0} \tag{4.59}$$

Note that the time dependence has been added back into (4.52) and (4.53).

4.2.6 Standing Waves

Expressions (4.52) and (4.53) are comprised of two constituent components, one a signal propagating in the $+z$ direction and another propagating in the $-z$ direction. The $+z$ is usually assumed to be in the forward direction, from the generator to the

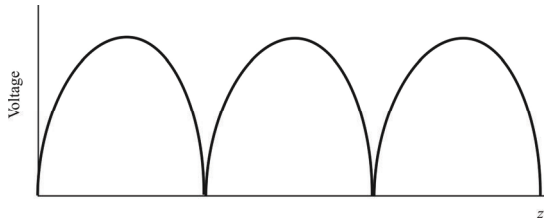


Figure 4.4 Standing wave.

load (incident to the load), and vice versa for the –*z* component (reflected from the load). The generator, of course, is the source of the forward traveling wave. At the load, an impedance mismatch between the line and the load will cause reflections back down the line. If there is no impedance mismatch at the load, so the load is the exact conjugate match to the line, then there will be no reflections. This exact matching, of course, only occurs at a single frequency. At all other frequencies there is some amount of mismatch so reflections are created.

These opposite traveling waves interact as they traverse the line. The net result is standing waves are established, voltage and current signals that appear to be stationary at any given point. The ratio of the peak in the voltage standing wave to the minimum is referred to as the *voltage standing wave ratio* (VSWR) or sometimes just *standing wave ratio* (SWR). The VSWR (*S*) can be shown to be

$$S = \frac{1+|\Gamma|}{1-|\Gamma|} \tag{4.60}$$

Recall that the signals involved here are sinusoids. The voltage standing wave pattern is illustrated in Figure 4.4 in the case of equal forward and reverse voltage levels. Because the levels are the same, at some points the waves will be completely supportive causing the amplitude of the standing wave to be twice the level of the constituent signals, while at other points they totally obliterate each other and the level is zero. In general, these standing waveforms will be sinusoidal in character as illustrated in Figure 4.5.

4.2.7 T-Matrices

Equations (4.52) and (4.53) can be manipulated to yield the T-matrix

$$\mathbf{T} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{4.61}$$

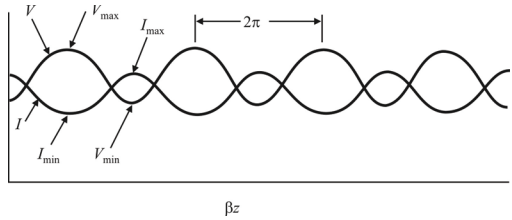


Figure 4.5 Standing waves.

for the line [3]. The equation for a lossy line of length l excited by a sinusoidal source is given by

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} \cosh \gamma l & Z_0 \sinh \gamma l \\ Y_0 \sinh \gamma l & \cosh \gamma l \end{bmatrix} \begin{bmatrix} V_2 \\ -I_2 \end{bmatrix} \tag{4.62}$$

When the line is lossless, $\gamma = j\beta$, and because

$$\cosh j\beta z = \cos \beta z \tag{4.63}$$

$$\sinh j\beta z = j \sin \beta z \tag{4.64}$$

the T-matrix for a lossless line of length l excited by a sinusoidal source is

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} \cos \gamma l & jZ_0 \sin \gamma l \\ jY_0 \sin \gamma l & \cos \gamma l \end{bmatrix} \begin{bmatrix} V_2 \\ -I_2 \end{bmatrix} \tag{4.65}$$

The characteristic impedance of a transmission line with losses is given by

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \tag{4.66}$$

Note that when R and G are small compared to the reactance values (lossless line again), then

$$Z_0 = \sqrt{\frac{L}{C}} \tag{4.67}$$

as before.

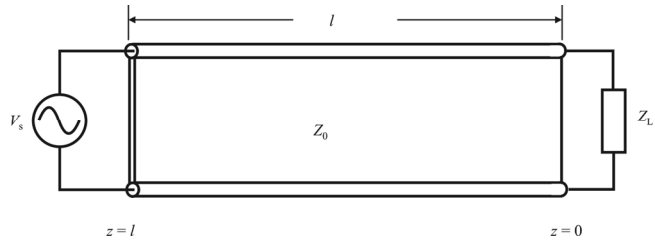


Figure 4.6 Transmission line.

4.3 Terminating Transmission Lines

Real transmission lines are not infinitely long—they are driven by a source at a finite point and are terminated with a load (that may be an open or a short circuit) at a finite point. We therefore must take this into consideration and do so in this section. For simplicity we assume lossless lines; the extension to lines with losses is straightforward.

A finite length of a lossless transmission line is illustrated in Figure 4.6. The transmission line is characterized with impedance Z_0 that is a real quantity. It is convenient to assume that the source of the time-harmonic wave is at $z = l$ and the termination is located at $z = 0$. This termination could be either an impedance or another transmission line with a different characteristic impedance. We will also assume that the signal generator was also turned on at the time $t = -\infty$ so all transient effects have disappeared.

There is also an impedance associated with the energy source at the input. The effects of this impedance are the same as that at the load output. This impedance is typically matched to the impedance of the transmission line.

For many applications Z_L represents a matching network that performs an impedance conversion to conjugate match the transmission line to an antenna. However, as we will see in Chapter 20, Z_L could be a short circuit $(Z_L = 0)$ or an open circuit $(Z_L = \infty)$. hese configurations are used for impedance matching.

The voltage at any point along the line can be expressed from (4.34) as

$$V(z,t) = A_1 e^{j(\omega t - \beta z)} + A_2 e^{j(\omega t + \beta z)} \tag{4.68}$$

The current at any point along the line is calculated from the substitution of (4.68) into the telegrapher's equations (4.19) or (4.22). We find

$$I(z,t) = \frac{A_1}{Z_0} e^{j(\omega t - \beta z)} - \frac{A_2}{Z_0} e^{j(\omega t + \beta z)} \tag{4.69}$$

If we evaluate the ratio of (4.68) to (4.69) at the location of the load impedance Z_L ($z = 0$), this ratio is equal to the load impedance. Because $z = 0$, $\beta z = 0$, the exponential terms will cancel, and we get

$$Z_L = \frac{V(0,t)}{I(0,t)} = \frac{A_1 + A_2}{A_1 - A_2} Z_0 \tag{4.70}$$

We know the amplitude of the signal that is incident upon the load because we can measure it; that is, A_1 is a known value. Knowing the amplitude of the incident wave allows us to compute the amplitude of the reflected wave A_2 from (4.70) to be

$$A_2 = \frac{Z_L - Z_0}{Z_L + Z_0} A_1 = \frac{Z_L / Z_0 - 1}{Z_L / Z_0 + 1} A_1 = \frac{z_L - 1}{z_L + 1} A_1 \tag{4.71}$$

and if $Z_L = Z_0$, so that $z_L = z_0$, the load is matched to the line and $A_2 = 0. A_2 = 0.$

The reflection coefficient, Γ, being a complex number, can be expressed in magnitude/angle form as

$$\Gamma = \rho e^{j\psi} \tag{4.72}$$

In terms of the reflection coefficient Γ, the voltage and the current at any point along the transmission line can be expressed as

$$V(z,t) = A_1 \left[e^{j(\omega t - \beta z)} + \Gamma e^{j(\omega t + \beta z)} \right] \tag{4.73}$$

$$I(z,t) = \frac{A_1}{Z_0} \left[e^{j(\omega t - \beta z)} - \Gamma e^{j(\omega t + \beta z)} \right] \tag{4.74}$$

The ratio of these two quantities is the impedance, $Z(z)$, at a particular location z. If the load impedance were a short circuit ($Z_L = 0$), then the reflection coefficient $\Gamma = -1$. At the load, the total voltage at $z = 0$ that consists of the sum of the incident and the reflected components must equal zero. If this impedance were an open circuit ($Z_L = \infty$), then the reflection coefficient $\Gamma = 1$. In this case, the total voltage can be arbitrary, but the total current must equal zero. If the load impedance is equal to the characteristic impedance of the line $Z_L = Z_0$, the reflection coefficient $\Gamma = 0$, and we say that the line is *matched*. This matching is very important in practice because all of the energy is transported down the line, all is absorbed by the load impedance, and none will be reflected back toward the

source. We will find that there are techniques that can be used in order to achieve this desirable state, even if the load impedance has a value that is different from the value of the characteristic impedance of the transmission line. Such matching of the transmission line to the load is covered in Chapter 20.

4.3.1 Finite Length Lines

A transmission line is either nonresonant or resonant. A nonresonant line is a line that has no standing waves. A resonant line is a line that does.

4.3.1.1 Nonresonant Lines

A nonresonant line is either infinitely long or terminated in its characteristic impedance. Since no reflections occur, all the energy traveling down the line is absorbed by the load that terminates the line. In addition, because the load impedance of such a line is equal to Z_0, no special tuning elements are required to effect a maximum power transfer.

4.3.1.2 Resonant Lines

A resonant line has a finite length and is not terminated in its characteristic impedance. Therefore, reflections of energy do occur. The load impedance is different from the Z_0 of the line; therefore, the input impedance is not purely resistive but has a reactive component. Tuning elements are used to eliminate the reactance and to bring about maximum power transfer from the source to the line. The line also may be used for a resonant or tuned circuit.

A resonant line is said to be resonant at an applied frequency. This means that at one frequency the line acts as a resonant circuit. It may act either as a high-resistive circuit (parallel resonant) or as a low-resistive circuit (series resonant). The line may be made to act in this manner by either open- or short-circuiting it at the output end and cutting it to some multiple of a quarter-wavelength.

4.3.2 Line Terminations

There are a large variety of possible terminations for RF transmission lines. Each type of termination has a characteristic effect on the standing waves on the line. We will cover many of them here.

4.3.2.1 Termination in the Characteristic Impedance

When the line is terminated in Z_0, a constant AC voltage will exist along the length of the line. Provided there are no losses in the line, the voltage trace will be a straight line. See Figure 4.7(a). If there are losses in the line, the amplitude of the

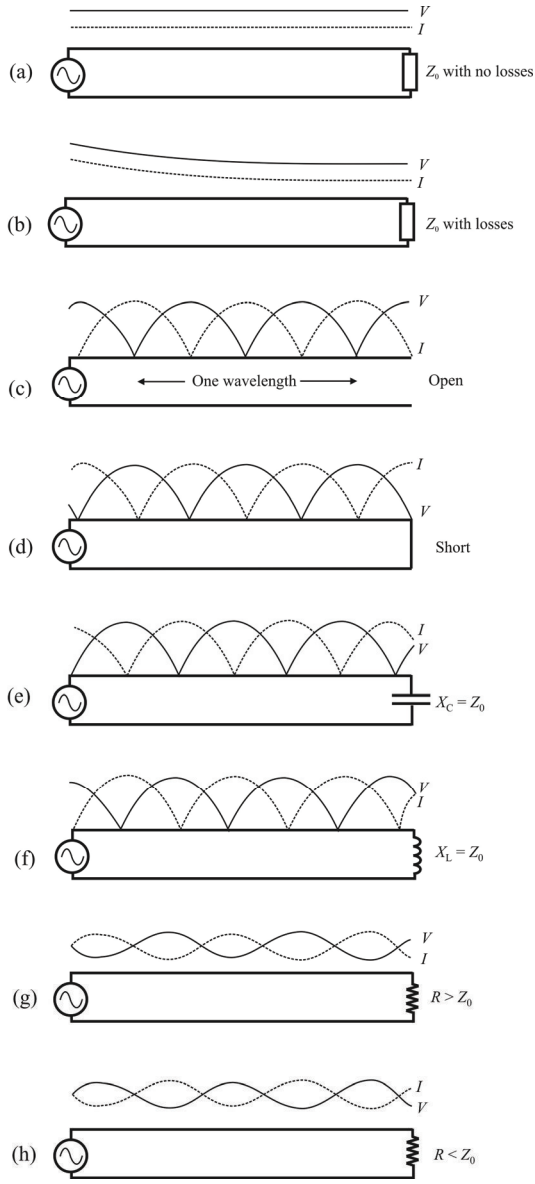


Figure 4.7(a–h) Effects of various terminations on standing waves.

voltage and current will diminish as they move down the line. See Figure 4.7(b). The losses are due to DC resistance in the line itself.

4.3.2.2 Termination in an Open Circuit

When the transmission line is terminated in an open circuit [see Figure 4.7(c)], the voltage is maximum at the load end, but the current is minimum. The distance between two adjacent zero current points is λ/2. The voltage is zero at a distance of λ/4 from the load end of the line, which is true at any frequency, incidentally. A voltage peak occurs at the load end of the line, at λ/2 from the end, and at each λ/2 thereafter.

Due to these peaks and valleys in the voltage and current along the line when it is terminated in an open circuit, the impedance looking into the line at these lengths varies, depending on the length of the line (measured in wavelengths). This property is used to advantage to match the source to the transmission line, which is the load to the source.

4.3.2.3 Termination in a Short Circuit

When the transmission line is terminated in a short circuit, shown in Figure 4.7(d), the voltage is zero at the short circuit and maximum at λ/4 from the end. The current is maximum at the short circuit, zero at λ/4 from the end, and alternately maximum and zero every λ/4 thereafter.

The same comments apply here as those given above for an open circuit. The length of a short-circuited transmission line can be varied to match the source to its load for maximum power transfer.

4.3.2.4 Termination in a Capacitor

When a line is terminated in capacitance, the capacitor does not absorb energy (assuming it is lossless), but returns all of the energy back into the line, resulting in 100% reflection. The current and voltage relationships are somewhat more involved than in previous types of termination. For our purposes here, assume that the capacitive reactance is equal to Z_0. Current and voltage are in phase when they arrive at the end of the line, but in flowing through the capacitor and the characteristic impedance (Z_0) connected in series, they shift their phase relationship. Current and voltage arrive in phase and leave out of phase. This results in the standing-wave configuration shown in Figure 4.7(e). The standing wave voltage is minimum at a distance of exactly λ/8 from the end. If the capacitive reactance is greater than Z_0 (smaller capacitance), the termination looks more like an open circuit; the voltage minimum moves away from the end. If the capacitive reactance is smaller than Z_0, the minimum moves toward the end.

4.3.2.5 Termination in an Inductor

When the line is terminated in an inductance, both the current and voltage shift in phase as they arrive at the end of the line. When X_L is equal to Z_0, the resulting current minimum is located λ/8 from the end of the line. When the inductive reactance is increased, the standing waves appear closer to the end. When the inductive reactance is decreased, the standing waves move away from the end of the line. See Figure 4.7(f).

4.3.2.6 Termination in a Resistance Not Equal to Z_0

Whenever the termination is not equal to Z_0, reflections occur on the line. For example, if the terminating element contains resistance, it absorbs some energy, but if the resistive element does not equal the Z_0 of the line, some of the energy is reflected. The amount of voltage reflected may be found by using

$$V_r = V_i \left(\frac{R_L - Z_0}{R_L + Z_0} \right) \tag{4.75}$$

where:

V_r is the reflected voltage
V_i is the incident voltage
R_L is the terminating resistance
Z_0 is the characteristic impedance of the line

If we try different values of R_L in the preceding equation, we find that the reflected voltage is equal to the incident voltage only when R_L equals 0 or is infinitely large. When R_L equals Z_0, no reflected voltage occurs. When R_L is greater than Z_0, V_r is positive, but less than V_i. As R_L increases and approaches an infinite value, V_r increases and approaches V_i in value. When R_L is smaller than Z_0, V_e has a negative value. This means that the reflected voltage is of opposite polarity to the incident wave at the termination of the line. As R_L approaches zero, V_r approaches V_i in value. The smaller the value of V_r, the smaller is the peak amplitude of the standing waves and the higher the minimum values.

Termination in a Resistance Greater Than Z_0

When R_L is greater than Z_0, the end of the line is somewhat like an open circuit; that is, standing waves appear on the line. The voltage maximum appears at the end of the line and also at half-wave intervals back from the end. The current is

minimum (not zero) at the end of the line and maximum at the odd quarter-wave points. Because part of the power in the incident wave is consumed by the load resistance, the minimum voltage and current are less than for the standing waves on an open-ended line. See Figure 4.7(g).

Termination in a Resistance Less Than Z_0

When R_L is less than Z_0, the termination appears as a short circuit. The line terminates in a current peak and a voltage minimum. The values of the maximum and minimum voltage and current approach those for a shorted line as the value of R_L approaches zero.

A line does not have to be any particular length to produce standing waves; however, it cannot be an infinite line. Voltage and current must be reflected to produce standing waves. For reflection to occur, a line must not be terminated in its characteristic impedance. Reflection occurs on lines terminated in opens, shorts, capacitances, and inductances, because no energy is absorbed by the load. If the line is terminated in a resistance not equal to the characteristic impedance of the line, some energy will be absorbed and the rest will be reflected. See Figure 4.7(h).

4.4 Waveguides

At UHF–microwave signal frequencies (between 300 MHz and 300 GHz), two-conductor transmission lines of any substantial length operating in standard TEM mode become impractical. Lines small enough in cross-sectional dimension to maintain TEM mode signal propagation for these signals tend to have low voltage ratings that make them unsuitable for handling the larger powers associated with EW systems. In addition they suffer from large, parasitic power losses due to conductor "skin" and dielectric effects. Fortunately, though, at these short wavelengths there exist other modes of propagation that are not as "lossy," if a conductive tube is used rather than two parallel conductors. It is at these high frequencies that *waveguides* become practical.

We present some basic properties of waveguides in this chapter but a full discussion of this subject is beyond the scope of this book. For those interested in delving further into the subject, [4] is recommended.

A waveguide is a special form of transmission line consisting of a hollow, metal tube. The tube wall provides distributed inductance, while the space between the tube walls, which could be filled with air or some other dielectric, provides distributed capacitance. Figure 4.8 depicts two common types of waveguides.

When an electromagnetic wave propagates down a hollow tube, only one of the fields—either electric or magnetic—will actually be transverse to the wave's direction of travel. The other field will "loop" longitudinally to the direction of

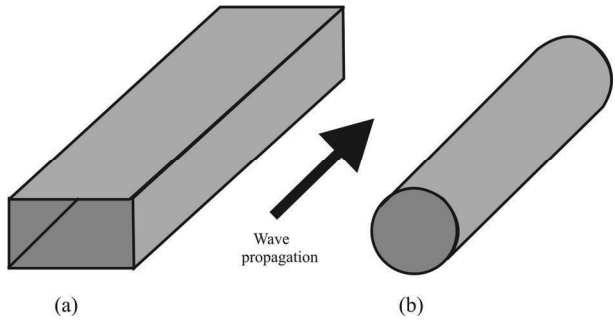


Figure 4.8 Waveguides: (a) rectangular and (b) tubular.

travel, but it still remains perpendicular to the other field. Whichever field remains transverse to the direction of travel determines whether the wave propagates in *TE* (transverse electric) mode or *TM* (transverse magnetic) mode. These fields associated with a signal propagating through a waveguide are depicted in Figure 4.9. We will briefly discuss them in this section.

A comparison of waveguides with transmission lines is provided in Table 4.2 [5]. Because of their modes of propagation, waveguides are practical only for signals of extremely high frequency, where the wavelength approaches the cross-sectional dimensions of the waveguide. Below such frequencies, waveguides cannot be used as electrical transmission lines.

When functioning as transmission lines, however, waveguides are considerably simpler than two-conductor cables, especially coaxial cables, in their manufacture and maintenance. There are few concerns with proper conductor-to-conductor spacing as there are with two-wire transmission lines.

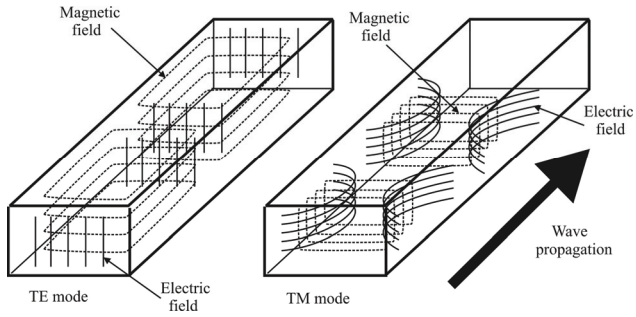


Figure 4.9 Waveguide transverse electric (TE) and transverse magnetic (TM) modes. Magnetic flux lines appear as continuous loops while electric flux lines have beginning and end points.

Table 4.2 Characteristics of Transmission Lines and Waveguides

Transmission Line	Waveguide
Two or more conductors separated by some insulating medium (two-wire, coaxial, microstrip).	Metal waveguides are typically one enclosed conductor filled with an insulating medium (rectangular, circular), while a dielectric waveguide consists of multiple dielectrics.
Normal operating mode is the TEM mode (can support TE and TM modes, but these modes are typically undesirable).	Operating modes are TE or TM modes (cannot support TEM mode).
No cutoff frequency for the TEM mode. Transmission lines can transmit signals from DC up to high frequency.	Must operate the waveguide at a frequency above the respective TE or TM mode cutoff frequency for that mode to propagate.
Significant signal attenuation at high frequencies due to conductor and dielectric losses.	Lower signal attenuation at high frequencies than transmission lines.
Small cross-section transmission lines (like coaxial cables) can only transmit low-power level due to the relatively high fields concentrated at specific location within the device (field levels are limited by dielectric breakdown).	Metal waveguides can transmit high-power levels. The fields of the propagating wave are spread more uniformly over a larger cross-sectional area than the small cross-section transmission line.
Large cross-section transmission lines (like power transmission lines) can transmit high-power levels.	Large cross-section (low-frequency) waveguides are impractical due to large size and high cost.

Source: [5].

When air is used as the dielectric, there is little concern about the consistency of the dielectric material. Moisture is not as severe a problem in waveguides as it is with some other types of transmission lines and so waveguides are often spared the necessity of gas "filling."

Waveguides may be thought of as conduits for electromagnetic energy, the waveguide itself acting as nothing more than a "director" of the energy rather than as a signal conductor in the normal sense of the word. As mentioned previously, all transmission lines function as conduits of electromagnetic energy when transporting high-frequency waves, directing the waves as the banks of a river direct its water. However, because waveguides are single-conductor elements, the propagation of electrical energy down a waveguide is of a very different nature from that down a two-conductor transmission line.

Many variations of each mode, TE and TM, exist for a given waveguide. The particular fields involved in a propagation mode for a given waveguide depend on the frequency of the propagating wave.

Signals are typically introduced to and extracted from waveguides by means of small antenna-like coupling devices inserted into the waveguide. Sometimes these coupling elements take the form of a dipole, which is nothing more than two open-ended stub wires of appropriate length. Other times, the coupler is a single

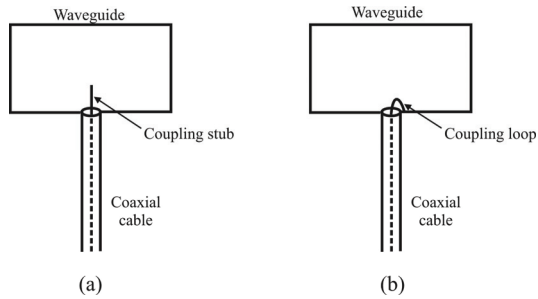


Figure 4.10 Waveguide coupling with (a) a coupling stub and (b) a coupling loop.

monopole 1/4λ stub, or a short loop of wire terminated on the inside surface of the waveguide, as illustrated in Figure 4.10.

Just as transmission lines are able to function as resonant elements in a circuit by termination by a short or open circuit, a dead-ended waveguide may also resonate at particular frequencies. When used as such, the device is called a *cavity resonator*. If a resonant cavity is open on the undriven end, it functions as an antenna. When flared, such a cavity becomes a horn antenna (see Chapter 13).

A cavity's resonant frequency may be altered by changing its physical dimensions. Cavities with movable plates, screws, and other mechanical elements for tuning are manufactured to provide coarse resonant frequency adjustment.

4.4.1 Wave Characteristics

For any time-harmonic source of EM radiation, the phasor electric and magnetic fields associated with the EM waves that propagate away from the source through a medium characterized by (μ, ε) must satisfy the source-free Maxwell's equations given by (2.35) and (2.36), or (2.78) and (2.79), with $\vec{J} = 0$ repeated here for convenience (in phasor form)

$$\vec{\nabla} \times \vec{E} = -j\omega\mu\vec{H} \tag{4.76}$$

$$\vec{\nabla} \times \vec{H} = j\omega\varepsilon\vec{E} \tag{4.77}$$

Equations (4.76) and (4.77) can be manipulated into wave equations (the Helmholtz equations) for the electric and magnetic fields as

$$\vec{\nabla}^2\vec{E} + k^2\vec{E} = 0 \tag{4.78}$$

$$\vec{\nabla}^2\vec{H} + k^2\vec{H} = 0 \tag{4.79}$$

where

$$k = \omega\sqrt{\mu\varepsilon} \tag{4.80}$$

The wavenumber k is real-valued for lossless media and complex valued for lossy media. The electric and magnetic fields of a general wave propagating in the $+z$-direction (either unguided, as in the case of a radiated plane wave or guided, as in the case of a transmission line or waveguide) through an arbitrary medium with a propagation constant of γ are characterized by a z-dependence of $e^{-\gamma z}$. Hence, the electric and magnetic fields of the wave may be written in rectangular coordinates as

$$\vec{E}(x,y,z) = \vec{e}(x,y)e^{-\gamma z} \tag{4.81}$$

$$\vec{H}(x,y,z) = \vec{h}(x,y)e^{-\gamma z} \tag{4.82}$$

where

$$\gamma = \alpha + j\beta \tag{4.83}$$

and where α is the wave attenuation constant and β is the wave phase constant. The propagation constant is purely imaginary ($\alpha = 0$, $\gamma = j\beta$) when there are no losses or complex-valued when losses are present ($\alpha \neq 0$).

Vectors $\vec{e}(x,y)$ and $\vec{h}(x,y)$ in (4.81) and (4.82) may contain both transverse field components and longitudinal field components. The longitudinal field components that emerge from (4.78) and (4.79) can be expressed in terms of the transverse field components as

$$\vec{E}_x = \frac{1}{h^2}\left(-\gamma\frac{\partial\vec{E}_z}{\partial x} - j\omega\mu\frac{\partial\vec{H}_z}{\partial y}\right) \tag{4.84}$$

$$\vec{E}_y = \frac{1}{h^2}\left(-\gamma\frac{\partial\vec{E}_z}{\partial y} + j\omega\mu\frac{\partial\vec{H}_z}{\partial x}\right) \tag{4.85}$$

$$\vec{H}_x = \frac{1}{h^2}\left(j\omega\varepsilon\frac{\partial\vec{E}_z}{\partial y} - \gamma\frac{\partial\vec{H}_z}{\partial x}\right) \tag{4.86}$$

$$\vec{H}_y = \frac{1}{h^2}\left(-j\omega\varepsilon\frac{\partial\vec{E}_z}{\partial x} - \gamma\frac{\partial\vec{H}_z}{\partial y}\right) \tag{4.87}$$

Table 4.3 Waveguide Modes

Type of Wave	Longitudinal Components	Mode
Transverse electromagnetic (TEM mode)	$\vec{E}_z = 0, \vec{H}_z = 0$	Plane waves, transmission line
Transverse electric (TE mode)	$\vec{E}_z = 0, \vec{H}_z \neq 0$	Waveguide
Transverse magnetic (TM mode)	$\vec{E}_z \neq 0, \vec{H}_z = 0$	Waveguide
Hybrid (EH or HE modes)	$\vec{E}_z \neq 0, \vec{H}_z \neq 0$	Waveguide

where the constant h is defined by

$$h^2 = \gamma^2 + \omega^2\mu\varepsilon \tag{4.88}$$

so that

$$\gamma = \sqrt{h^2 - k^2} \tag{4.89}$$

The equations for the transverse fields in terms of the longitudinal fields describe the different types of possible modes for guided and unguided waves, and these are listed in Table 4.3.

Consider the case of guided or unguided waves propagating through an ideal lossless medium where k is real-valued. For TEM modes, the only possible way for the transverse fields to be nonzero with $\vec{E}_z = \vec{H}_z = 0$ is for $h = 0$. This case then yields

$$\gamma = \sqrt{-\beta^2} \tag{4.90}$$

so that

$$\beta = k \tag{4.91}$$

Hence, for unguided TEM waves (radiated plane waves) moving through a lossless medium or guided TEM waves (waves on a transmission line) propagating on an ideal transmission line, we have $\gamma = jk = j\beta$.

For the waveguide modes (TE, TM or hybrid modes), h cannot be zero because this would yield unbounded results for the transverse fields in (4.84)–(4.87). Thus, $\beta \neq k$ for waveguides and the waveguide propagation constant can be written as

$$\gamma = \sqrt{h^2 - k^2} = \sqrt{-k^2\left(1 - \frac{h^2}{k^2}\right)} = jk\sqrt{1 - \frac{h^2}{k^2}} \tag{4.92}$$

The propagation constant (k) of a wave in a waveguide (TE or TM waves) has different fundamental characteristics from the propagation constant for a wave propagating on a transmission line (TEM waves, where $k = \beta$). We can express the ratio h/k in (4.92) in terms of the *cutoff frequency* (f_c) for the given waveguide mode as

$$\frac{h}{k} = \frac{h}{\omega\sqrt{\mu\varepsilon}} = \frac{h}{2\pi f\sqrt{\mu\varepsilon}} = \frac{f_c}{f}$$

so that

$$f_c = \frac{h}{2\pi\sqrt{\mu\varepsilon}} \tag{4.93}$$

The waveguide propagation constant, expressed in terms of the waveguide cutoff frequency is

$$\gamma = jk\sqrt{1 - \left(\frac{f_c}{f}\right)^2} \tag{4.94}$$

From (4.94) we can see that the cutoff frequency behavior of the waveguide modes is

If $f < f_c$, $\gamma = \alpha$ (purely real), $e^{-\gamma z} = e^{-\alpha z}$ waves are attenuated (*evanescent modes*).

If $f > f_c$, $\gamma = j\beta$ (purely imaginary), $e^{-\gamma z} = e^{-j\beta z}$ waves are unattenuated (*propagating modes*).

Therefore, in order to propagate a wave down a waveguide, the source must operate at a frequency higher than the cutoff frequency for that particular mode, and these cutoff frequencies are in general different for different modes. If a waveguide source is operated at a frequency lower than the cutoff frequency of the waveguide mode, then the wave is rapidly attenuated in the vicinity of the source of the wave.

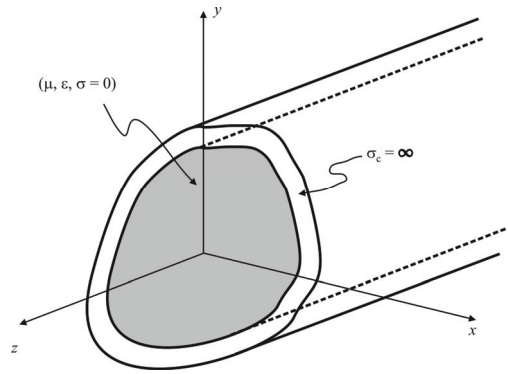


Figure 4.11 Perfect waveguide.

4.4.2 TE and TM Modes in Ideal Waveguides

An ideal waveguide is characterized by a perfect electrical conductor shell, a perfect inside insulator, and perfectly uniform dielectric material (see Figure 4.11). Waves propagate along the waveguide (+*z*-direction) within the waveguide through the lossless dielectric. The electric and magnetic fields of the guided waves must satisfy the source-free Maxwell's equations (4.76) and (4.77).

We will make the following assumptions:

- The waveguide is infinitely long, oriented along the *z*-axis, and uniform along its length.
- The waveguide is constructed from ideal materials [perfectly electric conducting pipe (PEC) is filled with a perfect insulator (lossless dielectric)].
- Fields are time-harmonic.

The cross-sectional size and shape of the waveguide dictate the discrete modes that can propagate along it. That is, there are only discrete electric and magnetic field distributions that will satisfy the appropriate boundary conditions on the inside surface of the waveguide conductor.

If the single nonzero longitudinal field component associated with a given waveguide mode can be determined ($\vec{E}_z$ for a TM mode, $\vec{H}_z$ for a TE mode), the remaining transverse field components can be found using (4.84)–(4.87). That is the methodology used to derive the wave's transverse fields.

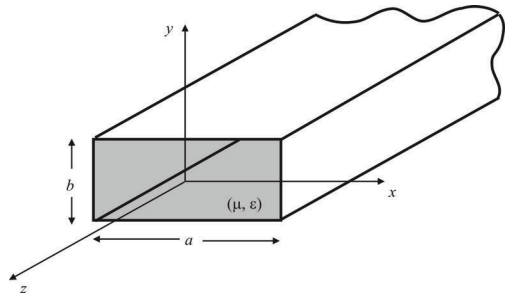


Figure 4.12 Ideal rectangular waveguide.

4.4.3 Ideal Rectangular Waveguides

For the ideal rectangular waveguide shown in Figure 4.12, we solve the wave equations by first finding $\vec{E}_z$ and $\vec{H}_z$ by using the wave equations (4.78) and (4.79) and then using (4.84)–(4.87).

The rectangular waveguide can support either TE or TM modes. The rectangular cross-section ($a > b$) allows for *single-mode operation*. Single-mode operation means that only one mode propagates in the waveguide over a given frequency range. A square waveguide cross-section ($a = b$) does not allow for single-mode operation.

Assuming sinusoidal excitations we proceed from Maxwell's curl equation (in phasor form)

$$\nabla \times \vec{E} = -j\omega\mu\vec{H} \tag{4.95}$$

$$\begin{vmatrix} \vec{u}_x & \vec{u}_y & \vec{u}_z \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ E_x & E_y & E_z \end{vmatrix} = -j\omega\mu\vec{H} \tag{4.96}$$

or

$$\vec{u}_x : \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = -j\omega\mu H_x \tag{4.97}$$

$$\vec{u}_y : -\left(\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z}\right) = -j\omega\mu H_y \tag{4.98}$$

$$\vec{u}_z: \qquad \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{yz} = -j\omega\mu H_z \tag{4.99}$$

However, the spatial variation in z is known so that

$$E_x = E_x^0 e^{-j\beta z} \qquad E_y = E_y^0 e^{-j\beta z} \qquad E_z = E_z^0 e^{-j\beta z} \tag{4.100}$$

and since

$$\frac{\partial\left(e^{-j\beta z}\right)}{\partial z} = -j\beta\left(e^{-j\beta z}\right) \tag{4.101}$$

we see that

$$\frac{\partial E_x}{\partial z} = -j\beta E_x^0 e^{-j\beta z} = -j\beta E_x \tag{4.102}$$

$$\frac{\partial E_y}{\partial z} = -j\beta E_y^0 e^{-j\beta z} = -j\beta E_y \tag{4.103}$$

$$\frac{\partial E_z}{\partial z} = -j\beta E_z^0 e^{-j\beta z} = -j\beta E_z \tag{4.104}$$

Consequently from (4.97)–(4.99), these curl equations simplify to

$$\frac{\partial E_z}{\partial y} + j\beta E_y = -j\omega\mu H_x \tag{4.105}$$

$$-\frac{\partial E_z}{\partial x} - j\beta E_x = -j\omega\mu H_y \tag{4.106}$$

$$\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = -j\omega\mu H_z \tag{4.107}$$

We can perform a similar expansion of Ampere's equation $\nabla\times\vec{H} = j\omega\varepsilon\vec{E}$ (again, phasor form) and subsequent algebraic manipulation to obtain

$$\frac{\partial H_z}{\partial y} + j\beta H_y = j\omega\varepsilon E_x \tag{4.108}$$

$$-\frac{\partial H_z}{\partial x} - j\beta H_x = j\omega\varepsilon E_y \tag{4.109}$$

$$\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = j\omega\varepsilon E_z \tag{4.110}$$

Now (4.108)–(4.110) can be manipulated to produce simple algebraic equations for the transverse (x and y) components of $\vec{E}$ and $\vec{H}$. For example, from (4.105)

$$H_x = \frac{j}{\omega\mu}\left(\frac{\partial E_z}{\partial y} + j\beta E_y\right) \tag{4.111}$$

Substituting for E_y from (4.109), we find

$$\begin{aligned} H_x &= \frac{j}{\omega\mu}\left[\frac{\partial E_z}{\partial y} + j\beta\frac{1}{j\omega\varepsilon}\left(-j\beta H_x - \frac{\partial H_z}{\partial x}\right)\right] \\ &= \frac{j}{\omega\mu}\frac{\partial E_z}{\partial y} + \frac{\beta^2}{\omega^2\mu\varepsilon}H_x - \frac{j\beta}{\omega^2\mu\varepsilon}\frac{\partial H_z}{\partial x} \end{aligned} \tag{4.112}$$

or

$$H_x = \frac{j}{k_c^2}\left(\omega\varepsilon\frac{\partial E_z}{\partial y} - \beta\frac{\partial H_z}{\partial x}\right) \tag{4.113}$$

where $k_c^2 = k^2 - b^2$ and $k^2 = \omega^2\mu\varepsilon$. Similarly, we can show that

$$H_y = -\frac{j}{k_c^2}\left(\omega\varepsilon\frac{\partial E_z}{\partial x} + \beta\frac{\partial H_z}{\partial y}\right) \tag{4.114}$$

$$E_x = -\frac{j}{k_c^2}\left(\beta\frac{\partial E_z}{\partial x} + \omega\mu\frac{\partial H_z}{\partial y}\right) \tag{4.115}$$

$$E_y = \frac{j}{k_c^2}\left(-\beta\frac{\partial E_z}{\partial y} + \omega\mu\frac{\partial H_z}{\partial x}\right) \tag{4.116}$$

From (4.113)–(4.116), we can see that all transverse components of $\vec{E}$ and $\vec{H}$ can be determined from only the axial components E_z and H_z. It is this fact that allows the mode designations TEM, TE, and TM. Furthermore, we can use superposition to reduce the complexity of the solution by considering each of these

mode types separately, then adding the fields together at the end. These are perhaps the most important takeaways from these equations.

4.4.3.1 Rectangular Waveguide TM Modes

The resulting transverse fields of the TM modes within the rectangular waveguide are given by

$$E_{\mathrm{x}}^{\mathrm{TM}_{mn}}(x,y,z) = -\frac{j\beta}{h^2}\left(\frac{m\pi}{a}\right)E_0 \cos\frac{m\pi x}{a}\sin\frac{n\pi y}{b}e^{-j\beta z} \tag{4.117}$$

$$E_{\mathrm{y}}^{\mathrm{TM}_{mn}}(x,y,z) = -\frac{j\beta}{h^2}\left(\frac{n\pi}{b}\right)E_0 \sin\frac{m\pi x}{a}\cos\frac{n\pi y}{b}e^{-j\beta z} \tag{4.118}$$

$$H_{\mathrm{x}}^{\mathrm{TM}_{mn}}(x,y,z) = \frac{j\omega\varepsilon}{h^2}\left(\frac{n\pi}{b}\right)E_0 \sin\frac{m\pi x}{a}\cos\frac{n\pi y}{b}e^{-j\beta z} \tag{4.119}$$

$$H_{\mathrm{y}}^{\mathrm{TM}_{mn}}(x,y,z) = -\frac{j\omega\varepsilon}{h^2}\left(\frac{m\pi}{a}\right)E_0 \cos\frac{m\pi x}{a}\sin\frac{n\pi y}{b}e^{-j\beta z} \tag{4.120}$$

where a and b are defined in Figure 4.12, m = 1, 2, …, n = 1, 2, …, and E_0 is a constant.

In general, the cutoff frequency will increase as the mode indices increase. Thus, in practice, only the lower-order modes are important as the waveguide is operated at frequencies below the cutoff frequencies of the higher-order modes.

4.4.3.2 Rectangular Waveguide TE modes

The resulting transverse fields of the TE modes within the rectangular waveguide are given by

$$E_{\mathrm{x}}^{\mathrm{TE}_{mn}}(x,y,z) = -\frac{j\omega\mu}{h^2}\left(\frac{n\pi}{b}\right)H_0 \cos\frac{m\pi x}{a}\sin\frac{n\pi y}{b}e^{-j\beta z} \tag{4.121}$$

$$E_{\mathrm{y}}^{\mathrm{TE}_{mn}}(x,y,z) = -\frac{j\omega\mu}{h^2}\left(\frac{m\pi}{a}\right)H_0 \sin\frac{m\pi x}{a}\cos\frac{n\pi y}{b}e^{-j\beta z} \tag{4.122}$$

$$H_{\mathrm{x}}^{\mathrm{TE}_{mn}}(x,y,z) = \frac{j\beta}{h^2}\left(\frac{m\pi}{a}\right)H_0 \sin\frac{m\pi x}{a}\cos\frac{n\pi y}{b}e^{-j\beta z} \tag{4.123}$$

$$\vec{H}_{\mathrm{y}}^{\mathrm{TE}_{mn}}(x,y,z) = -\frac{j\beta}{h^2}\left(\frac{n\pi}{b}\right)H_0 \cos\frac{m\pi x}{a}\sin\frac{n\pi y}{b}e^{-j\beta z} \tag{4.124}$$

where m = 0, 1, 2, …, n = 0, 1, 2, … but m and n are not simultaneously zero, and H_0 is a constant.

Thus the number of discrete TM modes is infinite based on the possible values of the indices m and n. An individual TM mode is designated as the TM_{mn} mode.

4.4.4 Rectangular Waveguide TE and TM Mode Parameters

The propagation constant in the rectangular waveguide is the same for both the TE_{mn} and TM_{mn} waveguide modes (γ_{mn}) and is given by

$$\gamma_{mn} = \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 - \omega^2\mu\varepsilon} \tag{4.125}$$

Equation (4.125) can be used to determine the cutoff frequency for the respective mode. Parameter h_{mn} is given by

$$h_{mn} = \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2} \tag{4.126}$$

The propagation characteristics of the wave are defined by the relative sizes of the parameters h_{mn} and k. The propagation constant may be written in terms of the attenuation and phase constants as

$$\gamma_{mn} = \alpha_{mn} + j\beta_{mn} \tag{4.127}$$

so that,

if $h_{mn} = k$, then $\gamma_{mn} = 0$ ($\alpha_{mn} = \beta_{mn} = 0$) (cutoff frequency)

if $h_{mn} > k$, then γ_{mn} is *real*, [$\gamma_{mn} = \alpha_{mn}$] (evanescent modes)

if $h_{mn} < k$, then γ_{mn} is *imaginary*, [$\gamma_{mn} = j\beta_{mn}$] (propagating modes)

Therefore, the cutoff frequencies for the TE and TM modes in the rectangular waveguide are given by

$$f_{c_{mn}} = \frac{1}{2\sqrt{\mu\varepsilon}}\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2} \tag{4.128}$$

The cutoff frequency for a particular rectangular waveguide mode depends on a and b as well as the material inside the waveguide (μ, ε), and the indices of the

Table 4.4 Cutoff Frequencies for the Example

Mode	f_c (GHz)
TE_{10}	0.833
TE_{01}, TE_{20}	1.667
TE_{11}, TM_{11}	1.863
TE_{21}, TM_{21}	2.357
TE_{30}	2.5

mode (m, n). The rectangular waveguide must be operated at a frequency above the cutoff frequency for the respective mode to propagate.

From (4.128), we see that the cutoff frequencies of both the TE_{10} and TE_{01} modes are less than those of the lowest-order TM mode (TM_{11}). Given $a > b$ for the rectangular waveguide, the TE_{10} has the lowest cutoff frequency of any of the rectangular waveguide modes and is thus the dominant mode (the first to propagate). This is the mode that is most frequently used in practice. Note that the TE_{10} and TE_{01} modes are *degenerate modes* (modes with the same cutoff frequency) for a square waveguide. The rectangular waveguide allows operation at a frequency above the cutoff of the dominant TE_{10} mode but below that of the next highest mode to achieve single-mode operation. A waveguide operating at a frequency where more than one mode propagates is said to be *overmoded*.

Example (Ideal rectangular waveguide propagating modes)

A rectangular waveguide (a = 2 cm, b = 1 cm) filled with deionized water (μ =1, ε = 81) operates at 3 GHz. Determine all propagating modes and the corresponding cutoff frequencies.

$$f_{c_{mn}} = \frac{1}{2\sqrt{\mu\varepsilon}}\sqrt{\left(\frac{m}{a}\right)^2+\left(\frac{n}{b}\right)^2} = \frac{c}{2\sqrt{(1)(81)}}\sqrt{\left(\frac{m}{0.02}\right)^2+\left(\frac{n}{0.01}\right)^2} \tag{4.129}$$

The numerical results are displayed in Table 4.4.

4.4.5 TEM Mode

The transverse electric and magnetic (TEM) modes are characterized by $E_z = 0$ and $H_z = 0$. In order for this to occur, it can be shown that it is necessary for $f_c = 0$. In other words, there is no cutoff frequency for waveguides that support TEM waves. Rectangular, circular, elliptical, and all hollow, metallic waveguides cannot support TEM waves. It can be shown that at least two separate conductors are required for TEM waves. Examples of waveguides that will allow TEM modes

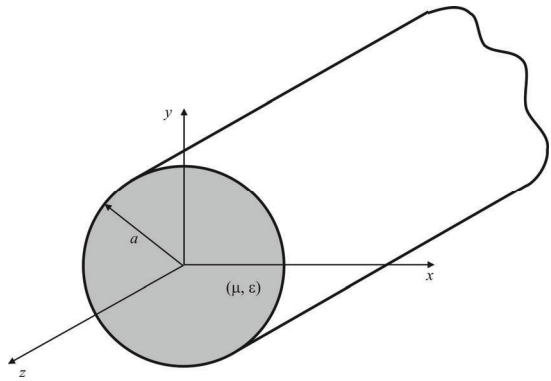


Figure 4.13 Ideal circular waveguide.

include coaxial cable, parallel plate waveguide, stripline, and microstrip. The microstrip "waveguide" will support the "quasi-TEM" mode, which like regular TEM modes has no nonzero cutoff frequency. However, if the frequency is large enough, other modes will begin to propagate on a microstrip. This is usually not a desirable situation.

4.4.6 Circular Waveguide

The same techniques used to analyze the ideal rectangular waveguide may be used to determine the modes that propagate within an ideal circular waveguide shown in Figure 4.13 [radius = a, filled with dielectric (μ, ε)]. Solutions evolve for the TE and TM propagating modes in terms of Bessel functions. The cutoff frequencies for the circular waveguide can be written in terms of the zeros associated with Bessel functions and derivatives of Bessel functions.

The cutoff frequencies of the TE and TM modes in a circular waveguide are given by

$$f_c^{\mathrm{TE}_{mn}} = \frac{J'_{mn}}{2\pi a\sqrt{\mu\varepsilon}} \tag{4.130}$$

and

$$f_c^{\mathrm{TM}_{mn}} = \frac{J_{mn}}{2\pi a\sqrt{\mu\varepsilon}} \tag{4.131}$$

where J_{mn} and J'_{mn} are the nth zero of the mth-order Bessel function and Bessel function derivative, respectively. The values of these zeros are shown in Tables 4.5 and 4.6. Note that the dominant mode in a circular waveguide is the TE_{11} mode, followed in order by the TM_{01} mode, the TE_{21} mode and the TE_{01} mode.

Example (Circular waveguide)

We want to find the size (radius) of an air-filled circular waveguide yielding a frequency separation of 1 GHz between the cutoff frequencies of the dominant mode and the next highest mode.

The cutoff frequencies of the TE_{11} mode (dominant mode) and the TM_{01} mode (next highest mode) for an air-filled circular waveguide are

$$f_{c_{11}}^{\mathrm{TE}} = \frac{J'_{11}}{2\pi a\sqrt{\mu_0\varepsilon_0}} = \frac{1.8412}{2\pi a\sqrt{\mu_0\varepsilon_0}}$$

$$f_{c_{01}}^{\mathrm{TM}} = \frac{J_{01}}{2\pi a\sqrt{\mu_0\varepsilon_0}} = \frac{2.4049}{2\pi a\sqrt{\mu_0\varepsilon_0}}$$

For a difference of 1 GHz between these frequencies, we must satisfy

$$f_{c_{01}}^{\mathrm{TM}} - f_{c_{11}}^{\mathrm{TE}} = \frac{2.4040 - 1.8412}{2\pi a\sqrt{\mu_0\varepsilon_0}} = 10^9$$

Solving this equation gives $a = 2.6896$ cm.

The cutoff frequencies for this waveguide are

$$f_{c_{11}}^{\mathrm{TE}} = \frac{J'_{11}}{2\pi a\sqrt{\mu_0\varepsilon_0}} = \frac{1.8412}{2\pi\times 0.026896\sqrt{\mu_0\varepsilon_0}} = 3.266\,\mathrm{GHz}$$

$$f_{c_{01}}^{\mathrm{TM}} = \frac{J_{01}}{2\pi a\sqrt{\mu_0\varepsilon_0}} = \frac{2.4049}{2\pi\times 0.026896\sqrt{\mu_0\varepsilon_0}} = 4.266\,\mathrm{GHz}$$

One unique feature of the circular waveguide is that some of the higher order modes (TE_{0n}) have particularly low loss. The

Table 4.5 TE Modes

J'_{mn}	$m = 0$	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	…
$n = 1$	3.8318	1.8412	3.0542	4.2012	5.3175	4.4155	4.5013	…
$n = 2$	4.0156	5.3315	4.7062	8.0153	4.2824	10.5199	4.7349	…
$n = 3$	10.1735	8.5363	4.9695	4.3459	12.6819	13.9872	15.2682	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

J'_{mn} are the nth zero of the mth-order Bessel function derivative.

Table 4.6 TM Modes

J_{mn}	$m = 0$	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	…
$n = 1$	2.4049	3.8318	5.1357	4.3802	4.5884	8.7715	4.9361	…
$n = 2$	5.5201	4.1056	8.4173	4.7610	4.0647	12.3386	13.5893	…
$n = 3$	8.6537	10.1735	4.6199	13.0152	14.3726	15.7002	14.0038	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

J_{mn} are the nth zero of the mth-order Bessel function.

magnetic field distribution for these modes generates lower currents on the walls of the waveguide than the other waveguide modes. Therefore, a circular waveguide carrying this mode is commonly used when signals are sent over relatively long distances (microwave antennas on tall towers).

The general equations for the circular waveguide TE_{mn} and TM_{mn} mode attenuation constants due to conductor loss are given by

$$\alpha_{c_{mn}}^{TE} = \frac{R_S}{a\eta'\sqrt{1-\left(\frac{f_{c_{mn}}}{f}\right)^2}}\left[\left(\frac{f_{c_{mn}}}{f}\right)^2 + \frac{m^2}{(J'_{mn})^2 - m^2}\right] \tag{4.132}$$

$$\alpha_{c_{mn}}^{TM} = \frac{R_S}{a\eta'\sqrt{1-\left(\frac{f_{c_{mn}}}{f}\right)^2}} \tag{4.133}$$

where η' is the impedance of the dielectric layer.

Example (Circular waveguide attenuation)

An air-filled copper waveguide (a = 5 mm, $\sigma_c = 5.8 \times 10^7$ mhos/m) is operated at 30 GHz. Determine the loss in dB m^{-1} for the TM_{01} mode.

$$f_{c_{01}}^{TM} = \frac{J_{01}}{2\pi a\sqrt{\mu_0\varepsilon_0}} = \frac{2.4049}{2\pi(0.005)\sqrt{\mu_0\varepsilon_0}} = 22.95\,\text{GHz}$$

$$\delta = \frac{1}{\sqrt{\pi f \mu \sigma_c}} = 0.38154\,\mu\text{m}$$

$$R_S = \frac{1}{\sigma_c \delta} = 0.045189\,\Omega \qquad \eta' = \eta_0 = 377\,\Omega$$

$$\alpha_{c_{mn}}^{TM} = \frac{0.045189}{(0.005)(377)\sqrt{1-\left(\frac{22.95}{30}\right)^2}} = 0.0372\,\text{Np m}^{-1}$$

The attenuation in terms of dB m^{-1} is

$$20\log_{10}(e^{-0.0372}) = -0.3231\,\text{dB m}^{-1}$$

(a loss of 0.3231 dB m^{-1}).

4.4.7 Waveguide Summary

Waveguides are metal tubes functioning as "conduits" for carrying energy in the form of EM waves. They are practical only for signals of extremely high frequency, where the signal wavelength approaches the dimensions of the waveguide.

Wave propagation through a waveguide may be classified into two broad categories: TE (transverse electric) or TM (transverse magnetic), depending on which field (electric or magnetic, respectively) is perpendicular (transverse) to the direction of wave travel. Wave travel along a standard, two-conductor transmission line is via the TEM (transverse electric and magnetic) mode, where both fields are oriented perpendicular to the direction of travel. TEM mode is only possible with two conductors and so is not applicable to a waveguide.

An enclosed-ended waveguide serves as a resonant element in a microwave circuit and is called a cavity resonator. A cavity resonator with an open end functions as an antenna (a horn), sending or receiving RF energy to/from the direction of the open end.

4.5 Transmission Line Cables

There are many possible types of configurations that can be used as transmission lines. For specificity, we discuss three in this section. The characteristics of the three transmission lines are given in Figure 4.14. These lines are very common in EW system design and analysis.

4.5.1 Coaxial Cables

Coaxial cables are probably the most common type of transmission line in use. The constituent parts to coaxial lines are illustrated in Figure 4.15. They consist of an inner conductor, surrounded by a dielectric insulator. That insulator in turn is surrounded by a (typically) braided outer conductor, also called a shield. This structure is enclosed in a nonconducting outer jacket. A coaxial cable is shown in Figure 4.16.

Coaxial cable is an excellent example of a TEM transmission line as the skin effect causes the current of the outer conductor to be concentrated on its inner surface. The magnetic fields generated by the equal and opposite currents of the concentric outer surface of the inner conductor and the inner surface of the outer conductor cancel outside the outer conductor in both the near and far fields, leaving no net magnetic field outside of the outer conductor that would couple to

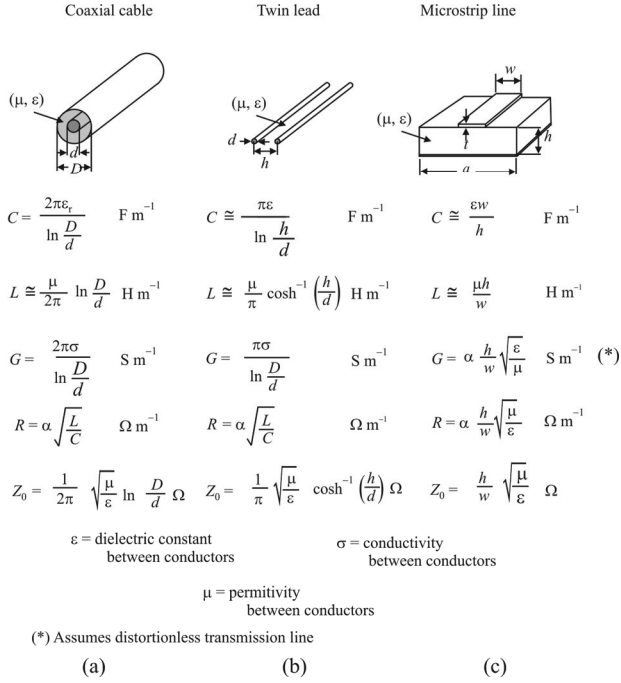


Figure 4.14 Common transmission lines encountered in EW systems: (a) coaxial cable, (b) twin lead, and (c) microstrip lines.

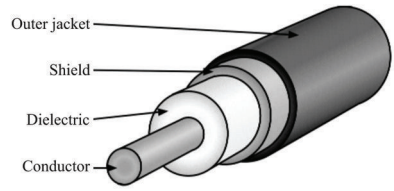


Figure 4.15 Coaxial cable unbalanced transmission line.

Figure 4.16 Coaxial cable.

nearby objects. Likewise, currents impinging on the cable from outside the outer conductor cannot be coupled into the inner conductor because of the equal and opposite currents on the conductors.

There are many types of coaxial cables manufactured by many sources. Some are specialized for particular applications, while others are intended for general applications. We will discuss some of the properties of coaxial cables in detail in the following.

4.5.1.1 Fields in a Coaxial Cable

The fields in a coaxial cable carrying a current I in the z direction (into the page) are illustrated in Figure 4.17. The E-field is radial, originating at the wire and moving toward the shield, while the magnetic field, orthogonal to E, is concentric between the shield and the wire.

4.5.1.2 Capacitance of Coaxial Lines

The capacitance in coaxial cables depends upon the ratio of the two radii; coaxial cables of the same cross-sectional shape but of different sizes have equal capacitance.

Assume the coaxial cable is very long and straight and is made of a conductor wire of radius d in the middle and a conductor shell of radius D ($> d$). The space in between is filled with a medium of dielectric constant ε. Suppose the inner and outer conductor carry uniform charge density λ and $-\lambda$, respectively. As seen in Figure 4.17, the E field is in the radial direction. Consider a cylindrical Gaussian

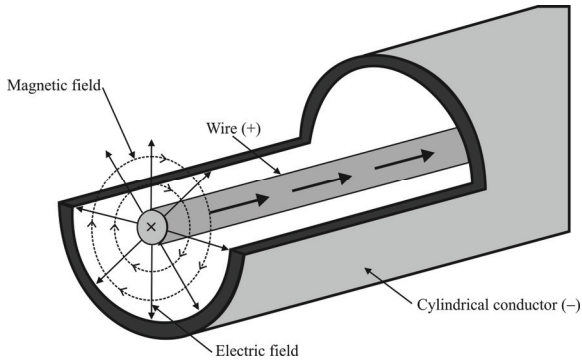


Figure 4.17 Fields in a coaxial cable

surface coaxial with the inner conducting wire, with length l and radius r. By Gauss' law, we have

$$E = \frac{\lambda}{2\pi\varepsilon r} \tag{4.134}$$

Therefore,

$$\begin{aligned} V &= -\int_{d}^{D} E dr \\ &= \frac{\lambda}{2\pi\varepsilon} \ln \frac{D}{d} \end{aligned} \tag{4.135}$$

The capacitance per unit length is hence

$$\begin{aligned} C &= \frac{\lambda}{V} \\ &= \frac{2\pi\varepsilon}{\ln(D/d)} \quad \text{F m}^{-1} \end{aligned} \tag{4.136}$$

4.5.1.3 Series Inductance of Coaxial Cables

To find L, let the inner conductor carry uniform current I in one direction and the shell carries the same current but in the opposite direction. By symmetry and the

fact that $\vec{\nabla} \cdot \vec{B} = 0$, the B-field is tangential (see Figure 4.17) and with equal magnitude for a fixed r. By Ampere's law,

$$B = \frac{\mu I}{2\pi r} \tag{4.137}$$

with direction determined by the right-hand rule. Hence, for a segment with unit length, the magnetic flux is given by

$$\begin{aligned} \Phi &= \int_d^D B dr \\ &= \frac{\mu I}{2\pi} \ln \frac{D}{d} \end{aligned} \tag{4.138}$$

Since $\Phi = LI$, we have

$$L = \frac{\mu}{2\pi} \ln \frac{D}{d} \qquad \text{h m}^{-1} \tag{4.139}$$

We see that $LC = \mu\varepsilon$, which is independent of the physical dimension of the cable.

4.5.1.4 Series Resistance of Coaxial Cable

At low frequencies the current density in a long homogeneous conductor of ordinary size is uniform over the entire cross section. The DC resistance per unit length is

$$R_{\text{DC}} = \frac{1}{\pi\sigma r_i^2} + \frac{1}{\pi\sigma\left[(r_\text{o} + t)^2 - r_\text{o}^2\right]} \qquad \Omega\,\text{m}^{-1} \tag{4.140}$$

where r_o represents the radius of the inner surface of the outer conductor, r_i represents the radius of the inner conductor, t represents the thickness of the outer conductor, σ is the conductivity in the inner and outer conductors, and all the other symbols have their customary meanings. The unit series resistance R is related to the conductivity of the conductors and is frequency dependent due to the skin effect. At high frequencies, all the current is a thin layer next to the surface of the wire and the resistance R_h is (h for high frequency):

$$R_{\mathrm{h}} = \frac{\sqrt{f\mu\rho}}{2a\sqrt{\pi}} = \frac{\sqrt{f\mu}}{2a\sqrt{\pi\sigma}} \qquad \Omega \tag{4.141}$$

where f is frequency in hertz, μ is the wire's absolute permeability in henries per meter, and a is the inner radius of the dielectric between the inner and outer conductors.

4.5.1.5 Shunt Conductance

Shunt conductance of transmission lines may be a result of the combined effects of leakage of current through the insulation or over its surface and dielectric losses.

$$G = \frac{2\pi\sigma_{\mathrm{d}}}{\ln\left(\frac{r_{\mathrm{o}}}{r_{\mathrm{i}}}\right)} \qquad \mho\,\mathrm{m}^{-1} \tag{4.142}$$

where σ_{d} is the conductivity of dielectric between the inner and outer conductors. At high frequencies, typically over 1 MHz, the effective conductivity of a dielectric material at high frequency is given by

$$\sigma_{\mathrm{d}} = 2\pi f \varepsilon_0 \varepsilon_{\mathrm{r}} \tan\delta \tag{4.143}$$

where $\tan\delta$ is the loss tangent usually specified by the cable manufacturer. So the nominal shunt conductance per unit length at frequencies above about 1 MHz is

$$G = \frac{4\pi^2 f \varepsilon_0 \varepsilon_{\mathrm{r}} \tan\delta}{\ln\left(\frac{r_0}{r_{\mathrm{i}}}\right)}, \qquad \mho\,\mathrm{m}^{-1} \tag{4.144}$$

We can derive similar equations for parallel wire transmission lines, twisted wire transmission line, or any other form of transmission line.

Specific losses for some common coaxial transmission lines are listed in Table 4.7. Other characteristics are given in Table 4.8.

4.5.2 Twin Lead

Twin lead is a two-conductor ribbon cable used as a transmission line to carry radio frequency (RF) signals (see Figure 4.18).

Table 4.7 Coaxial Cable Signal Loss (dB per 10 m)

Loss	RG-58	RG-8X	RG-213	RG-6	RG-11
1 MHz	0.1	0.1	0.1	0.1	0.1
10 MHz	0.5	0.3	0.2	0.2	0.1
50 MHz	1.1	0.8	0.5	0.5	0.3
100 MHz	1.6	1.2	0.7	0.7	0.5
200 MHz	4.4	1.8	101	0.9	0.8
400 MHz	4.7	4.6	1.6	1.4	1.2
700 MHz	4.6	4.7	4.2	1.9	1.6
900 MHz	4.7	4.2	4.6	4.0	1.8
1 GHz	4.2	4.5	4.8	4.0	1.9
Ω	50	50	50	75	75

Source: http://www.astrosurf.com/luxorion/qsl-transmission-line4.htm.

Table 4.8 Some Common Coaxial Cables

Type	Impedance (Ω)	Core (mm)	Dielectric			Overall Diameter (mm)	Shields	Notes
			Type*	VF	mm			
RG-6/U	75	1.0	PF	0.75	4.7	4.86	Double	Low loss at high frequency
RG-6/UQ	75	1.0	PF	0.75	4.7	7.57	Quad	Quad-shield RG-6
RG-8/U	50	2.17	PE	0.75	7.2	10.3		
RG-8X	50	1.0	PF	0.75	4.7	4.1	Double	
RG-11/U	75	1.63	PE	0.66	7.2	10.5		Long drops and underground conduit
RG-58/U	50	0.9	PE	0.66	2.9	5.0	Single	Radio communications
RG-213/U	50	.21	PE	0.66	7.2	10.3	Single	Less lossy than RG-58

PE = solid polyethylene, PF = polyethylene foam

Figure 4.18 Twin lead.

4.5.2.1 Twin Lead Characteristics

Twin lead is supplied in several different sizes, with 450Ω, 300Ω, and 75Ω characteristic impedances. The most common, 300Ω twin lead, was once widely used to connect television sets and FM radios to their receiving antennas. Balanced 300Ω twin lead for television installations has been largely replaced with 75Ω unbalanced coaxial cable feed lines because twin lead is susceptible to interactions with its surroundings while coaxial cable is not (or not as much).

Twin lead consists of two multi-stranded copper wires, held apart by a plastic (usually polyethylene) ribbon (see Figure 4.18). The plastic also serves as insulation for the wires. In 300Ω twin lead, the wire is usually 20 or 22 gauge, about 4.5 mm apart. In 450Ω twin lead, which consists of 16 or 18 gauge wire about 20 mm apart, the plastic ribbon usually has rectangular openings between the wires every few centimeters in order to reduce dielectric signal loss.

Twin lead has the advantage that its losses are an order of magnitude smaller than coaxial cable discussed above. Its disadvantages are that it is more vulnerable to interference and must be kept away from metal objects. For this reason, when installed along the outside of buildings and on antenna masts, standoff insulators must be used.

4.5.2.2 Impedance Matching

As a transmission line, transmission efficiency will be maximum when the impedance of the antenna, the characteristic impedance of the twin-lead line, and the impedance of the PA driving the line (transmit) or receiver (receiving) are the same. For this reason, when attaching a twin-lead line to a coaxial cable connection, a balun with a 4:1 ratio is commonly used. We cover baluns more extensively in Chapter 20 when we discuss antenna matching and Chapter 22 in the discussion on multicouplers. Its purpose is twofold: first, it transforms twin lead's 300Ω impedance to match the 75Ω coaxial cable impedance; and second, it

Figure 4.19 Balun. A 300 to 75-ohm balun, showing twin lead on the right-hand side.

transforms the balanced, symmetric transmission line to the unbalanced coaxial input. A typical balun for this purpose is shown in Figure 4.19. In general, twin lead has a higher efficiency than coaxial cable when there is an impedance mismatch between the transmission line and the source (or sink). For receive-only use this means that the system must operate under slightly less than optimum conditions; for transmit use, this can often result in significantly more energy lost as heat in the transmission line as well as higher levels of reflected energy back to the PA from the antenna.

4.5.2.3 Electrical Properties

Some electrical properties of twin lead are given in Table 4.9. Twin lead can be connected directly to a suitably designed antenna:

- A dipole (whose impedance at resonance is ~73 ohms in free space);
- A folded dipole (a better match, since its characteristic impedance in free space is around 300 ohms);
- A Yagi, log-periodic, or similar balanced antenna.

4.5.3 Microstrip Lines

The previous discussions of transmission lines, including coaxial cables and twin lead, refer primarily to cables connecting different elements in an EW system. We now switch topics and discuss transmission lines that are designed on circuit boards and/or integrated circuits, and, in particular, monolithic microwave

Table 4.9 Electrical Properties of 300Ω Twin Lead (20 gauge)

Parameter		300 Ω	75 Ω
Capacitance (pF/10 m)		48	200
Velocity Factor		0.80	0.71
Loss (dB/10 m)	100 MHz	0.36	
	300 MHz	0.72	
	500 MHz	1.02	

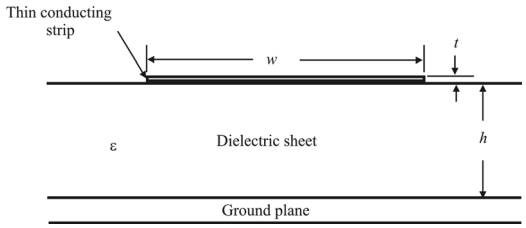


Figure 4.20 Cross-sectional view of a microstrip transmission line.

integrated circuits (MMICs). We will now discuss microstrip lines. This section is patterned largely on the paper by Maloratsky [6].

A number of different transmission lines are generally available for use with microwave ICs (MICs). In addition to the microstrip line shown in Figure 4.14, there are the stripline, suspended stripline, slotline, coplanar waveguide, and finline [6]. We will confine our discussion to the microstrip line as being representative of these technologies; each has its advantages and disadvantages in any particular application.

A cross section of a microstrip transmission line is depicted in Figure 4.20. The microstrip line consists of a single conductor trace on one side of a dielectric substrate and a ground plane on the opposite side (see Figure 4.14). In a microstrip line, the wavelength, Λ, as in any other dielectric material, is lower than the wavelength of the same signal in free space—it is given by

$$\Lambda = \frac{\lambda}{\sqrt{\varepsilon_{\text{eff}}}} \tag{4.145}$$

where ε_{eff} is the effective dielectric constant, which, as we will see, depends on the dielectric constant of the substrate material and the physical dimensions of the microstrip line, and λ is the free-space wavelength.

In a microstrip line, the EM fields exist partly in the air above the dielectric substrate and partly within the substrate. The effective dielectric constant of the line is then some combination of the dielectric constant of air (which is 1) and that of the dielectric substrate. The effective dielectric constant is shown in Figure 4.21 as a function of physical dimensions and relative dielectric constant of the dielectric material [6].

A popular configuration for a microstrip transmission is surface mounting on a printed circuit board (PCB). The transmission line is printed on one side of the PCB while the other side is a ground sheet. An unshielded microstrip line is not a practical structure, however. It is open to the air, and it is desirable to have covers to protect the line from the environment as well as to facilitate EMC and prevent EMI. Covering the basic microstrip configuration with metal top plates on the top

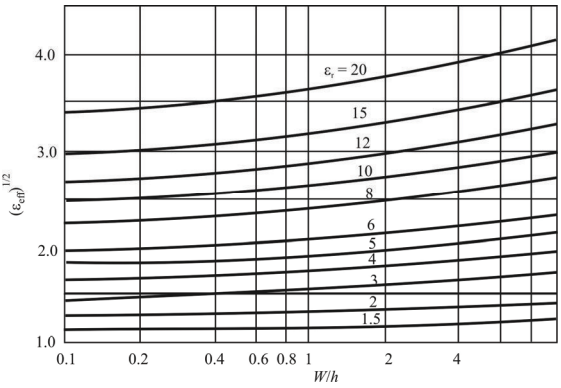


Figure 4.21 Effective dielectric constant.

and on the sides leads to a more realistic circuit configuration, a shielded microstrip line with a housing. The housing, of course, modifies the electrical characteristics somewhat.

If the size of the housing is incorrectly designed, parasitic propagation of EM waves may occur. The parasitic modes will not appear if

$$\lambda_0 < \frac{2}{\sqrt{\left(\frac{m}{l}\right)^2 + \left(\frac{n}{a}\right)^2}} \tag{4.146}$$

where m and n are positive integers. Equation (4.146) is the condition for wave propagation in a waveguide with dimensions $l \times a$. If (4.146) is not satisfied, parasitic modes can arise, and the height H must be chosen to suppress these modes.

Wheeler provided an empirical equation for approximating the characteristic impedance of microstrip lines [7, 8]. The characteristic impedance is a function of the dimensions of the line as well as the dielectric constant of the material between the strip and the ground plane. It is given by (variables are defined in Figure 4.20)

$$Z_0 = \frac{42.4}{\sqrt{\varepsilon+1}} \ln\left\{1 + \left(\frac{4h}{w}\right)\left[\left(\frac{14+8/\varepsilon}{11}\right)\left(\frac{4h}{w}\right) + \sqrt{\left(\frac{14+8/\varepsilon}{11}\right)^2\left(\frac{4h}{w}\right)^2 + \frac{1+1/\varepsilon}{2}\pi^2}\right]\right\} \tag{4.147}$$

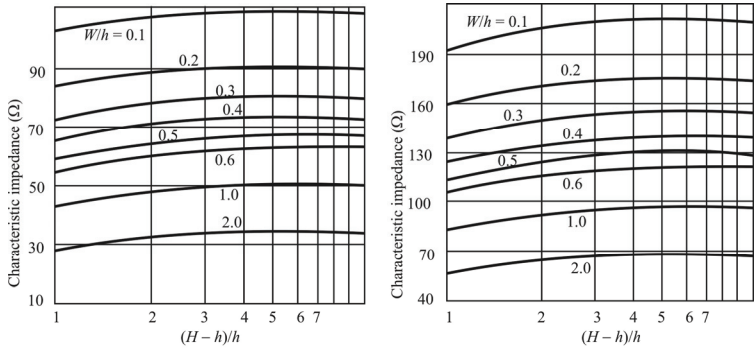


Figure 4.22 (a, b) Microstrip characteristic impedance.

which is claimed to be accurate to within 1% over most of the range of the variables. Clearly it is real-valued. When the dielectric is air ($\varepsilon = 1$) the formula simplifies to

$$Z_0 = 30 \ln \left\{ 1 + \frac{1}{2}\left(\frac{8h}{w}\right)\left[\frac{8h}{w} + \sqrt{\left(\frac{8h}{w}\right)^2 + \pi^2}\right]\right\} \tag{4.148}$$

The relationships between characteristic impedance and the physical dimensions of shielded microstrip lines for two examples: substrates with low (2) and high (4.6) relative dielectric constants are shown in Figure 4.22 [6]. The top cover tends to reduce the impedance. When the ratio of the distance from the top cover to the dielectric substrate and the substrate thickness $[(H - h)/h]$ is greater than 10, the enclosure effects can be considered negligible (see Figure 4.22). The characteristic impedance range of a microstrip line is 20 to 120Ω, and they vary with the W/h ratio as we see in Figure 4.23 [6]. The upper limit is set by production tolerances while the lower limit is set by the appearance of higher-order propagating modes.

There are three types of losses that occur in microstrip lines: conductor (or ohmic) losses, dielectric losses, and radiation losses. An idealized microstrip line, being open to a semi-infinite air space above the conducting strip, acts similar to a patch antenna, which is constructed almost identically to the microstrip line. As such, they have a propensity to radiate energy.

When low cost is a priority, low dielectric constants substrate materials (5 or less) are used. Similar materials provide relief for mechanical tolerances at millimeter-wave frequencies. However, the lower the dielectric constant, the less

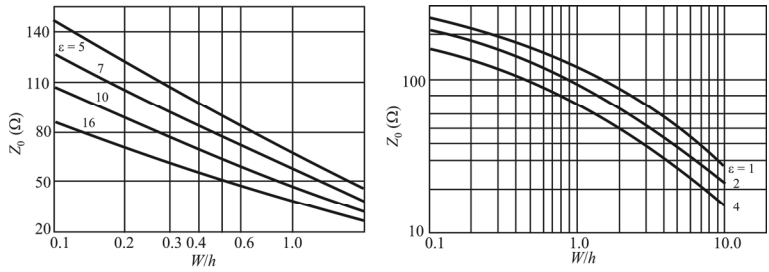


Figure 4.23 Microstrip characteristic impedance. (a) high and (b) low dielectric constants.

the concentration of energy is in the substrate region and, hence, the more radiation occurs.

Most of the EM field is concentrated in the dielectric between the conductive strip and the ground plane when high dielectric constant materials are used, which results in reduced radiation losses. Another significant benefit in using a higher dielectric constant is that the package size decreases by approximately the square root of the dielectric constant.

Normally with high substrate dielectric constant materials, conductor losses in the strip conductor and the ground plane dominate over dielectric and radiation losses. Conductor losses are a result of normal conductor factors related to the metallic material composing the ground plane and walls, including the material's conductivity, skin effects, and surface roughness.

In a microstrip line, conductor losses increase with increasing characteristic impedance due to the greater resistance of narrow strips. Conductor losses follow a characteristic trend that is opposite to radiation loss with respect to W/h.

4.6 Concluding Remarks

We discussed the fundamental properties of transmission lines in this chapter. Transmission lines transport RF signals to and from the antenna(s), power amplifiers, and/or receiver(s). Their properties are critical for optimally performing systems. Incorrect usage generates high VSWRs, causing reflections back from the antenna to the PA output. That energy is not radiated. On receive, poorly performing transmission lines can cause excessive losses, thereby reducing sensitivity and raising the noise figure.

We discussed three specific types of transmission lines as examples of the wide variety of cables available: (1) coaxial cable, (2) twin lead, and (3) microstrip lines. We also discussed at length the properties of waveguides that perform the transmission line function at higher (microwave) frequencies.

References

[1] Jordan, E. C., and K. G. Balmain, *Electromagnetic Waves and Radiating Systems,* 2nd ed., Englewood Cliffs, NJ: Prentice Hall, 1968, Ch. 14.

[2] Shadowitz, A., *The Electromagnetic Field,* New York: Dover Publications, 1975, Ch.12.

[3] Potter, J. L., and S. Fich, *Theory of Networks and Lines*, Englewood Cliffs, NJ: Prentice Hall, 1963, pp. 298–304.

[4] Shadowitz, A., *The Electromagnetic Field,* New York: Dover Publications, 1975, Ch.18.

[5] http://www.ece.msstate.edu/~donohoe/ece3323waveguides.pdf.

[6] Maloratsky, L. G., "Reviewing the Basics of Microstrip Lines," *Microwaves & RF,* March 2000, pp. 79–84.

[7] Wheeler, H. A., "Transmission-Line Properties of a Strip on a Dielectric Sheet on a Plane," *IEEE Transactions on Microwave Theory & Techniques,* Vol. 25, No. 8, August 1977, pp. 631–644.

[8] Wheeler, H. A., "Transmission-Line Properties of Parallel Strips Separated by a Dielectric Sheet," *IEEE Transactions on Microwave Theory & Techniques,* Vol. 3, No. 3, March 1965, pp. 172–184.

Part II
Antenna Systems

Chapter 5

Dipole Antennas

5.1 Introduction

The characteristics of dipole antennas are examined in this chapter. A dipole antenna consists of two wires, usually of equal length, configured colinearly (end to end) and are fed at the middle, although offset feeds are also possible. All the dipoles examined in this chapter are fed at their center. Other cases are considered later in Chapter 20 when we consider matching techniques.

This chapter is organized as follows. We discussed the radiation patterns of the infinitesimal dipole in Chapter 2. As mentioned there, such antennas are not physically realizable, but their analysis is useful for leading to the analysis of larger antennas. The current distribution on such an infinitesimal dipole is assumed to be a constant value. We begin this chapter with an analysis of a "short" dipole. Such an antenna is assumed to be longer than the infinitesimal dipole but short enough so that its current distribution can be approximated as triangular. Their length is on the order of $\lambda/10$ or less. Then we examine dipoles that are longer so that their current distribution is modeled as portions of sine patterns along the length of the antenna. We then focus on the important case of when $l = \lambda/2$. After that we examine the characteristics of the folded dipole. That is followed by exploring the effects of traps along the length of the antenna. The last subject we investigate is the sleeve dipole antenna.

5.2 Dipoles

The dipole is one of the simplest antenna types [1]. It consists of two elements that are aligned in the same direction. Such a structure is illustrated in Figure 5.1 where

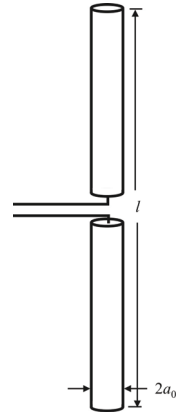


Figure 5.1 Vertically oriented dipole antenna structure.

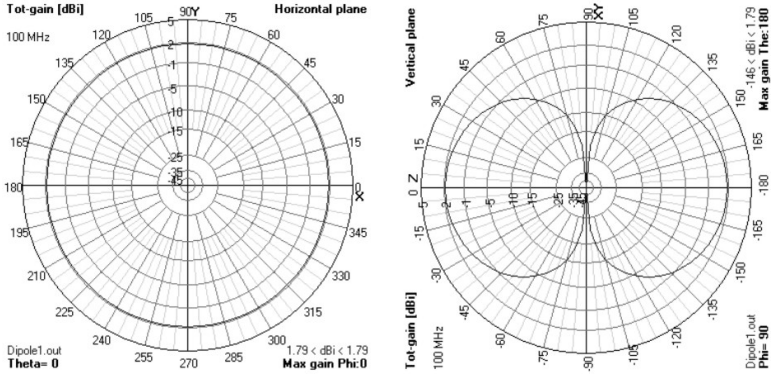


Figure 5.2 Dipole 2-D patterns.

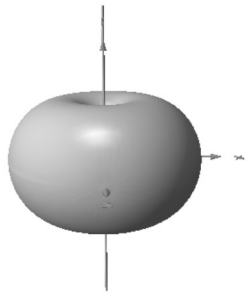


Figure 5.3 Vertically oriented dipole 3-D E-field pattern in free space.

l denotes the length of the dipole and a_0 is its radius. Figure 5.2 shows the 2-D radiation patterns when $l = \lambda/2$. For the dipole, as well as the other antennas discussed here, the radiation pattern depends on the physical size of the antenna. For the dipole, in the far field the most important parameter that determines its radiation pattern is its length. Because the radiation pattern depends on the length of the antenna, then the gain, efficiency, and directivity of these antennas are frequency dependent. This is true for all known types of antennas, although the characteristic variations with frequency are not the same for all types of antennas as will be shown.

Dipole antennas are popular antennas for many applications because of their simplicity in many, perhaps most, nonfixed applications. For example, in EW applications, the geographical relationship between the EW system and the target array cannot be completely a prior known. The dipole antenna provides an element of gain while providing omnidirectional coverage in the azimuth plane. This gain is provided at the expense of gain in the vertical plane, however, so in applications where the vertical plane response is important as well, simple dipoles have limited application. Crossed dipoles can be used in an array to address this situation.

The radiation pattern in three dimensions is shown in Figure 5.3. Again the uniform radiation in azimuth and nulled radiation off the ends of the antenna are evident. The 3-D pattern is compared with the uniform pattern of the isotropic antenna in Figure 5.4 [1].

5.2.1 Short Dipole

A dipole is considered short when $\lambda / 50 < l < \lambda / 10$ and $a_0 << \lambda$ (less than $\lambda / 50$ is considered infinitesimal and was analyzed in Chapter 3). The current distribution on a short dipole is modeled as triangular, as illustrated in Figure 5.5. Recall from Chapter 3 that the current distribution for the infinitesimal dipole (it

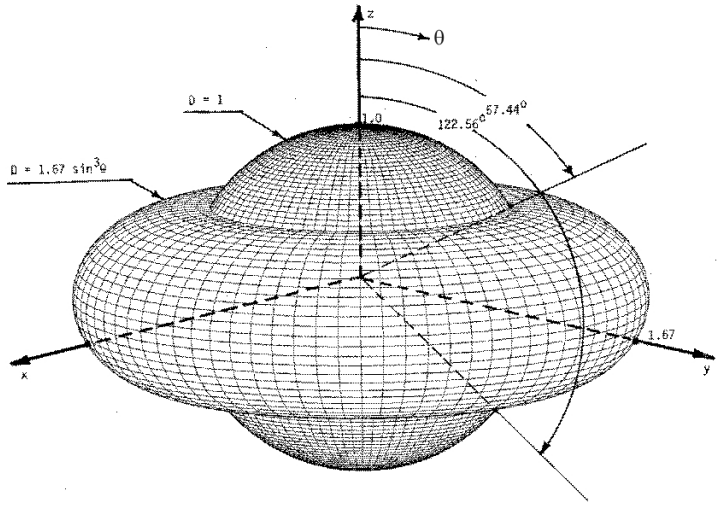


Figure 5.4 Comparison of a dipole with an omnidirectional antenna. (Source: [1]. © IEEE 1992. Reprinted with permission.)

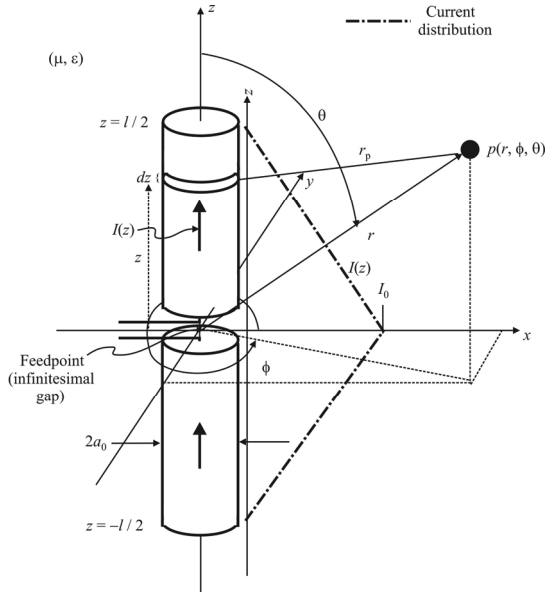


Figure 5.5 Short dipole current distribution is approximated as triangular (dotted lines).

doesn't get any shorter than that) is modeled as uniform or a constant value. In this case, the triangular current distribution is given by

$$\vec{I}(z) = I_0\left(1-\frac{2|z|}{l}\right)\vec{u}_z$$
$$= \begin{cases} I_0\left(1+\dfrac{2z}{l}\right)\vec{u}_z, & \left(-\dfrac{l}{2} \le z \le 0\right) \\ I_0\left(1-\dfrac{2z}{l}\right)\vec{u}_z, & \left(0 \le z \le \dfrac{l}{2}\right) \end{cases} \tag{5.1}$$

The magnetic vector potential is given by

$$\vec{A}(\vec{r}) = \frac{\mu}{4\pi}\iiint_V \vec{J}(\vec{r}\,')\frac{e^{-j\beta r_p}}{r_p}d\vec{v}\,' \tag{5.2}$$

For the far field, making the approximations

$$d\vec{v}\,' \to dl = dz \qquad \left(-\frac{l}{2} \le z \le \frac{l}{2}\right) \tag{5.3}$$

$$r_p \approx r \tag{5.4}$$

yielding

$$\vec{A}^{sd}(\vec{r}) \approx \frac{\mu}{4\pi}\int_{-l/2}^{l/2} I_0\frac{e^{-j\beta r}}{r}dz\vec{u}_z \tag{5.5}$$

$$= \frac{\mu I_0 l}{8\pi r}e^{-j\beta r}\vec{u}_z \tag{5.6}$$

$$= \frac{1}{2}\vec{A}^{\inf}(\vec{r}) \tag{5.7}$$

where $\vec{A}^{\inf}$ is given by (3.124). So the magnitude of the VP $\vec{A}$ for the short dipole is one-half that of the infinitesimal dipole. Therefore the magnitudes of the far-field patterns for the small dipole are one-half those for the infinitesimal dipole and since they both are oriented along the *z*-axis, the angular patterns are identical. Thus, the short dipole far-field radiation patterns are given by

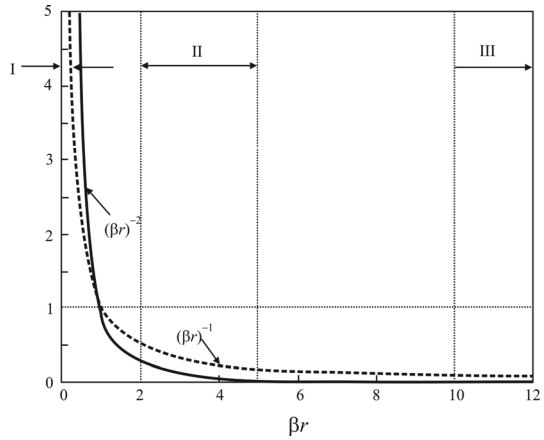


Figure 5.6 Field regions of a short dipole. Three distinct regions can be identified.

$$E_r \approx E_\phi = H_r = H_\theta = 0 \tag{5.8}$$

$$E_\theta = j\eta \frac{\beta I_0 l e^{-j\beta r}}{8\pi r} \sin\theta \tag{5.9}$$

$$H_\phi = \frac{\beta I_0 l e^{-j\beta r}}{8\pi r} \sin\theta \tag{5.10}$$

Regions can be defined based on (5.8)–(5.10), as illustrated in Figure 5.6. Region I is where $\beta r << 1$ ($r << \lambda / 2\pi$) and the $(\beta r)^{-2}$ terms in the brackets in (3.133) dominates. In region II, $\beta r \approx 1$ ($r \approx \lambda / 2\pi$) and the $(\beta r)^{-1}$ in (3.132) and (3.133) term dominates. In region III, $\beta r >> 1$ ($r >> \lambda / 2\pi$) and the constant term in the brackets dominates. This leads to the following approximations for the fields of the short dipole:

Near Fields:

$$\begin{aligned}
\vec{E}_r &\approx -j\eta_0 \frac{\beta I_0 l}{2\pi r^3} \cos\theta e^{-j\beta r} \vec{u}_r \\
\vec{E}_\theta &\approx -j\eta \frac{\beta I_0 l}{4\pi r^3} \sin\theta e^{-j\beta r} \vec{u}_\theta \\
\vec{H}_\phi &\approx \frac{\beta I_0 l}{4\pi r} \sin\theta e^{-j\beta r} \vec{u}_\phi
\end{aligned} \tag{5.11}$$

Near Radiating Fields:

$$\vec{E}_{\mathrm{r}} \approx \eta \frac{\beta I_0 L}{2\pi r^2} \cos\theta e^{-j\beta r} \vec{u}_{\mathrm{r}} \tag{5.12}$$

$$\vec{E}_{\theta} \approx j\eta \frac{\beta I_0 L}{4\pi r} \sin\theta e^{-j\beta r} \vec{u}_{\theta} \tag{5.13}$$

$$\vec{H}_{\phi} \approx j \frac{\beta I_0 L}{4\pi r} \sin\theta e^{-j\beta r} \vec{u}_{\phi} \tag{5.14}$$

The far fields produced by the short dipole, the ones of most interest to us when considering EW (with the exception of mutual coupling in an array) are given by:

Far Fields:

$$\vec{E} \approx j\eta_0 \frac{\beta I_0 l e^{-j\beta r}}{8\pi r} \sin\theta \vec{u}_{\theta} \tag{5.15}$$

$$\vec{H} \approx j \frac{\beta I_0 l e^{-j\beta r}}{8\pi r} \sin\theta \vec{u}_{\phi} \tag{5.16}$$

while the radiated power for the short dipole is

$$\Pi = \eta \frac{\pi}{12} \left(\frac{|I_0| l}{\lambda} \right)^2 = \frac{1}{2} |I_0|^2 R_{\mathrm{r}} \tag{5.17}$$

yielding the radiation resistance

$$R_{\mathrm{r}} = \eta \frac{\pi}{6} \left(\frac{l}{\lambda} \right)^2 = 20\pi^2 \left(\frac{l}{\lambda} \right)^2 \tag{5.18}$$

which is one-fourth the value of the radiation resistance for the infinitesimal dipole considered in Chapter 3.

The directivity function for the short dipole is given by

$$D(\theta, \phi) = 1.5 \sin^2 \theta \tag{5.19}$$

with maximum directivity

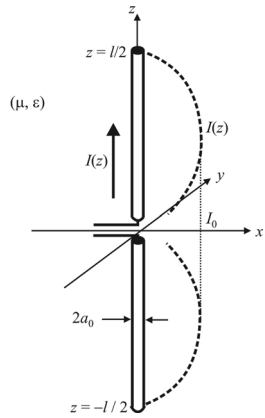


Figure 5.7 Short center-fed dipole. Radiation patterns (dashed lines) are neither constant nor triangular.

$$D_0 = D(\phi,\theta)\big|_{\max} = 1.5 \qquad (1.8 \text{ dB}) \tag{5.20}$$

The effective aperture is given by

$$A_{\text{eff}} = \frac{\lambda^2}{4\pi} D_0 = \frac{\lambda^2}{4\pi}\frac{3}{2} \tag{5.21}$$

with solid beam angle

$$\Omega_A = \frac{8\pi}{3} \text{ radians} \tag{5.22}$$

5.2.2 Finite Length Dipole Antennas

Assume that the dipole antenna is driven at its center and $a_0 << \lambda$ (we will discuss driving a dipole from other than its center in Chapter 20 when we discuss matching techniques). Then we may assume that the current distribution is a sine wave and so is symmetrical along the antenna as shown in Figure 5.7. The current can be expressed as

$$\vec{I}(z) = I_0 \sin\left[\beta\left(\frac{l}{2} - |z|\right)\right]\vec{u}_z, \qquad \left(-\frac{l}{2} \le z \le \frac{l}{2}\right) \tag{5.23}$$

$$= \begin{cases} I_0 \sin\left[\beta\left(\frac{l}{2}+z\right)\right]\vec{u}_z, & \left(-\frac{l}{2} \le z \le 0\right) \\ I_0 \sin\left[\beta\left(\frac{l}{2}-z\right)\right]\vec{u}_z, & \left(0 \le z \le \frac{l}{2}\right) \end{cases} \tag{5.24}$$

5.2.2.1 Radiating Fields

Using the approximation given by (3.115), the far magnetic VP field is found to be

$$\vec{A}(\vec{r}) \approx \mu \frac{e^{-j\beta r}}{2\pi r} I_0 (\cos\theta \vec{u}_r - \sin\theta \vec{u}_\theta) \left[\frac{\cos\left(\frac{\beta l}{2}\cos\theta\right) - \cos\left(\frac{\beta l}{2}\right)}{\beta \sin^2\theta} \right] \tag{5.25}$$

The far fields of the center-fed dipole in terms of the magnetic vector potential are:

$$\begin{aligned} \vec{E}_\theta &\approx -j\omega A_\theta \vec{u}_\theta \\ &= j\eta_0 \frac{e^{-j\beta r}}{2\pi r} I_0 \left[\frac{\cos\left(\frac{\beta l}{2}\cos\theta\right) - \cos\left(\frac{\beta l}{2}\right)}{\sin\theta} \right] \vec{u}_\theta \end{aligned} \tag{5.26}$$

and

$$\begin{aligned} \vec{H}_\phi &= \frac{E_\theta}{\eta_0} \vec{u}_\phi \\ &= j \frac{e^{-j\beta r}}{2\pi r} I_0 \left[\frac{\cos\left(\frac{\beta l}{2}\cos\theta\right) - \cos\left(\frac{\beta l}{2}\right)}{\sin\theta} \right] \vec{u}_\phi \end{aligned} \tag{5.27}$$

The time-average complex Poynting vector in the far field is

$$\vec{P} = \frac{1}{2}\vec{E} \times \vec{H}^* = \frac{1}{2}\left(E_\theta \vec{u}_\theta \times H_\phi^* \vec{u}_\phi\right) = \frac{1}{2}\left(E_\theta \frac{E_\theta^*}{\eta_0} \vec{u}_r\right) = \frac{|E_\theta|^2}{2\eta_0} \vec{u}_r$$

$$= \frac{\eta_0 |I_0|^2}{8\pi^2 r^2} \left[\frac{\cos\left(\frac{\beta l}{2}\cos\theta\right) - \cos\left(\frac{\beta l}{2}\right)}{\sin\theta} \right]^2 \vec{u}_r \tag{5.28}$$

The radiation intensity function is given by (3.4). Thus

$$U(\phi,\theta) = \frac{\eta_0 |I_0|^2}{8\pi^2} \left[\frac{\cos\left(\frac{\beta l}{2}\cos\phi\right) - \cos\left(\frac{\beta l}{2}\right)}{\sin\theta} \right]^2 \tag{5.29}$$

The normalized radiation intensity function, $U(\theta) = B_0 F(\theta)$, displays the effect of the antenna length on its radiation pattern where the power pattern is

$$F(\theta) = \left[\frac{\cos\left(\frac{\beta l}{2}\cos\theta\right) - \cos\left(\frac{\beta l}{2}\right)}{\sin\theta} \right]^2 \tag{5.30}$$

These patterns are shown in Figure 5.8 for some values of l. In general, we see that the directivity of the antenna increases as the length goes from a short dipole (a fraction of a wavelength) to a full wavelength. As the length increases above a wavelength, minor lobes are introduced into the radiation pattern. The current distributions for these lengths are also shown in Figure 5.8. The patterns are compared against each other as seen in the horizontal plane pattern cross sections in Figure 5.9.

The total real power radiated by the center-fed dipole is given by

$$\begin{aligned} \Pi &= \int_0^{2\pi}\int_0^{\pi} U(\theta,\phi)\sin\theta d\theta d\phi \\ &= \frac{\eta_0 |I_0|^2}{4\pi} \int_0^{\pi} \frac{\left[\cos\left(\frac{\beta l}{2}\cos\theta\right) - \cos\left(\frac{\beta l}{2}\right)\right]^2}{\sin\theta} d\theta \end{aligned} \tag{5.31}$$

The θ-dependent integral in the radiated power expression cannot be integrated in

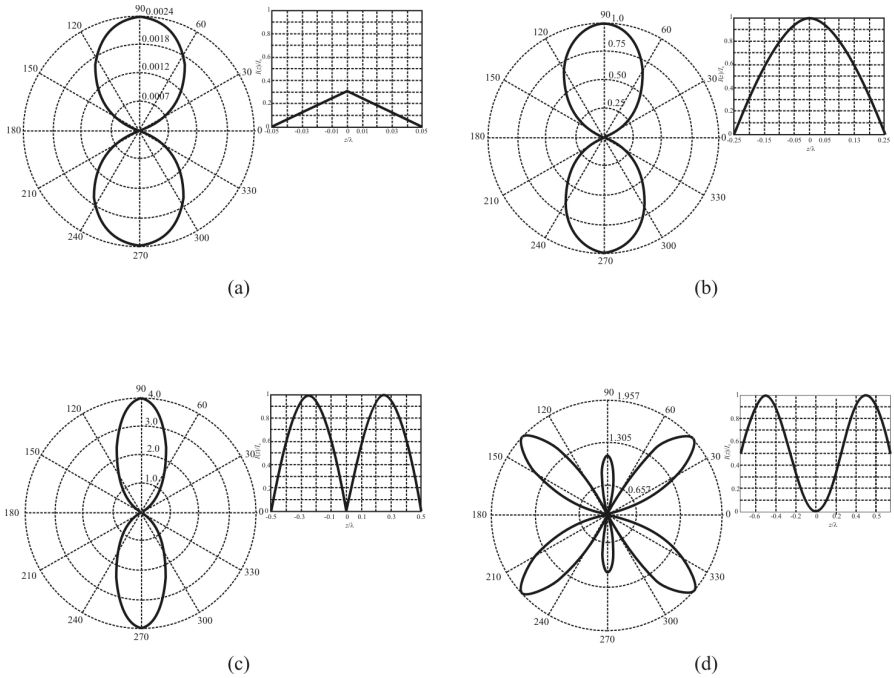


Figure 5.8 Radiation patterns for the center-fed dipole as a function of length. Also shown are the current distributions associated with the respective lengths: (a) $l = \lambda / 10$; (b) $l = \lambda / 2$; (c) $l = \lambda$; and (d) $l = 3\lambda / 2$.

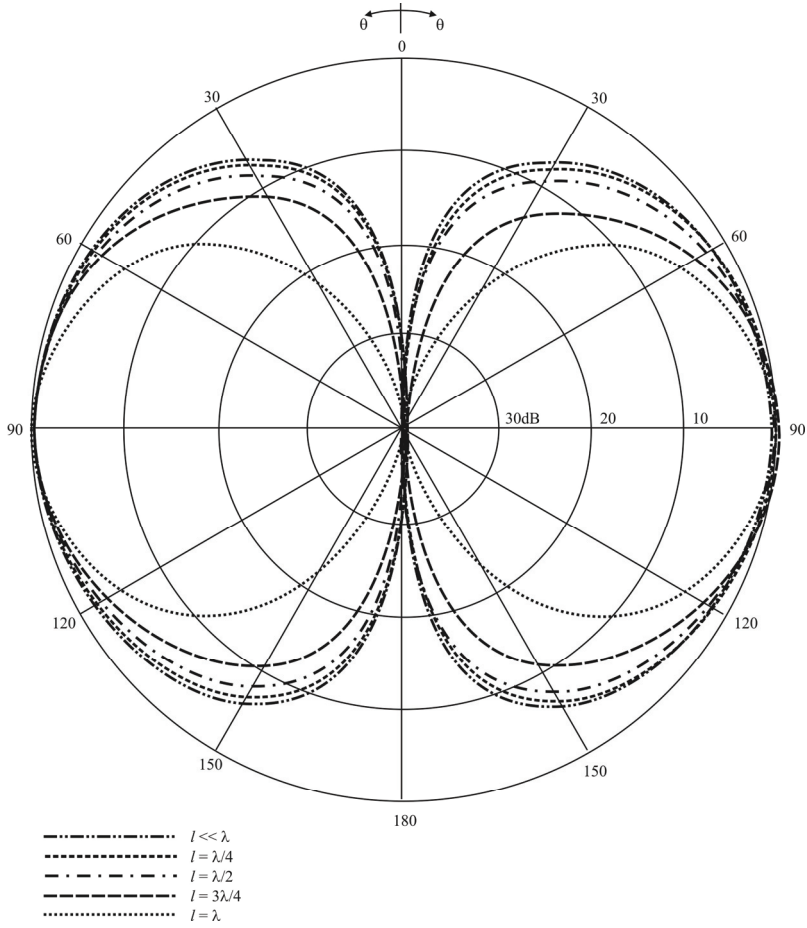


Figure 5.9 Dipole far-field radiation pattern comparison.

closed form. However, the integral may be manipulated into a form using the sine integral and an integral related to the cosine integral [2]:

Sine integral

$$Si(x) = \int_0^x \frac{\sin y}{y} dy \tag{5.32}$$

Cosine integral

$$Ci(x) = \gamma + \ln x - \int_x^\infty \frac{\cos u - 1}{u} du \tag{5.33}$$

Euler's constant in (5.33) is

$$\gamma = 0.57722 \tag{5.34}$$

The radiated power becomes

$$\Pi = \frac{\eta_0 |I_0|^2}{4\pi} \left\{ \begin{array}{r} \gamma + \ln(\beta l) - Ci'(\beta l) + \frac{1}{2}\sin(\beta l)[Si(2\beta l) - 2Si(\beta l)] \\ + \frac{1}{2}\cos(\beta l)[\gamma + \ln(\beta l/2) + Ci(2\beta l) - 2Ci(\beta l] \end{array} \right\} \tag{5.35}$$

Since

$$\Pi = \frac{1}{2}|I_0|^2 R_r \tag{5.36}$$

then[1]

[1] This is the expression for R_r in the induced emf method of determining the input impedance of a wire antenna. It is important to note that this method:

- Assumes sinusoidal current distribution;
- Is accurate for pattern and impedance of dipoles up to half-wavelength and monopoles up to quarter-wavelength;
- Is inaccurate for impedance of dipoles longer than half-wavelength and monopoles longer than quarter-wavelength.

$$R_r = \frac{\eta_0}{2\pi}\left\{\begin{array}{l}\gamma + \ln(\beta l) - Ci(\beta l) + \frac{1}{2}\sin(\beta l)[Si(2\beta l) - 2Si(\beta l)] \\ \qquad + \frac{1}{2}\cos(\beta l)[\gamma + \ln(\beta l/2) + Ci(2\beta l) - 2Ci(\beta l)]\end{array}\right\} \tag{5.37}$$

The intensity function is given by (5.29), which yields

$$U(\theta,\phi) = \frac{\eta_0 |I_0|^2}{8\pi^2}\left[\frac{\cos\left(\frac{\beta l}{2}\cos\theta\right) - \cos\left(\frac{\beta l}{2}\right)}{\sin\theta}\right]^2 \triangleq \frac{\eta_0 |I_0|^2}{8\pi^2} F(\theta,\phi) \tag{5.38}$$

while the directivity is given by (3.6). Denote

$$\begin{array}{l} f(\beta l) = \gamma + \ln(\beta l) - Ci(\beta l) + \frac{1}{2}\sin(\beta l)[Si(2\beta l) - 2Si(\beta l)] \\ \qquad + \frac{1}{2}\cos(\beta l)[\gamma + \ln(\beta l/2) + Ci(2\beta l) - 2Ci(\beta l)] \end{array} \tag{5.39}$$

then the radiated power is

$$\Pi = \frac{\eta_0 |I_0|^2}{8\pi^2} f(\beta l) \tag{5.40}$$

and the directivity is

$$D(\theta,\phi) = 2\frac{F(\theta,\phi)}{f(\beta l)} \tag{5.41}$$

with maximum directivity

$$D_0 = 2\frac{F(\theta,\phi)\big|_{\max}}{f(\beta l)} \tag{5.42}$$

The effective area is

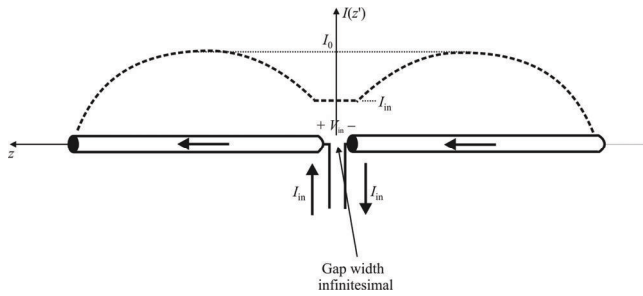


Figure 5.10 Dipole impedance parameters.

$$A_{\text{eff}} = \frac{\lambda^2}{2\pi} \frac{F(\theta,\phi)\big|_{\max}}{f(\beta l)} \tag{5.43}$$

while the solid beam angle is

$$\Omega_{\text{A}} = 2\pi \frac{f(\beta l)}{F(\theta,\phi)\big|_{\max}} \tag{5.44}$$

5.2.2.2 Dipole Input Impedance

The input impedance of the dipole is given by the ratio of voltage to current at the antenna feed point (see Figure 5.10):

$$Z_{\text{in}} = \frac{V_{\text{in}}}{I_{\text{in}}} = R_{\text{in}} + jX_{\text{in}} \tag{5.45}$$

The real time-average power delivered to the terminals of the antenna is

$$\text{Re}\{P_{\text{in}}\} = \frac{1}{2}\text{Re}\{V_{\text{in}} I_{\text{in}}^*\} = \frac{1}{2}\text{Re}\{|I_{\text{in}}|^2 Z_{\text{in}}\} = \frac{1}{2}|I_{\text{in}}|^2 R_{\text{in}} \tag{5.46}$$

and the associated reactive time-average power is

$$\text{Im}\{P_{\text{in}}\} = \frac{1}{2}\text{Im}\{V_{\text{in}} I_{\text{in}}^*\} = \frac{1}{2}\text{Im}\{|I_{\text{in}}|^2 Z_{\text{in}}\} = \frac{1}{2}|I_{\text{in}}|^2 X_{\text{in}} \tag{5.47}$$

When the antenna is lossless ($R_{ohmic} = R_L = 0$), then all the real power delivered to the input terminals is radiated. In that case,

$$\frac{1}{2}\left|I_{in}\right|^2 R_{in} = \frac{1}{2}\left|I_0\right|^2 R_r \tag{5.48}$$

yielding the relationship between the antenna input resistance and the antenna radiation resistance as

$$R_{in} = \frac{\left|I_0\right|^2}{\left|I_{in}\right|^2} R_r \tag{5.49}$$

In a similar fashion, the relationship between the reactive power delivered to the antenna input terminals and that stored in the near field of the antenna can be determined to be

$$X_{in} = \frac{\left|I_0\right|^2}{\left|I_{in}\right|^2} X_A \tag{5.50}$$

The general dipole current is defined by (5.23) so the current I_{in} is the current at $z = 0$ so that

$$I_{in} = I(0) = I_0 \sin\left[\beta\frac{l}{2}\right] \tag{5.51}$$

Based on (5.49)–(5.51), the input resistance and reactance of the antenna are related to the equivalent circuit values of radiation resistance and the antenna reactance by

$$R_{in} = \frac{R_r}{\sin^2\left(\frac{\beta l}{2}\right)} \tag{5.52}$$

and

$$X_{in} = \frac{X_A}{\sin^2\left(\frac{\beta l}{2}\right)} \tag{5.53}$$

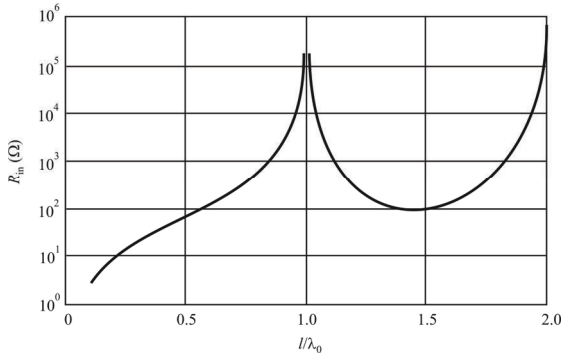


Figure 5.11 Center-fed dipole impedance real-part (resistance) example when $a_0 = 0.0005\lambda$.

Using the induced-emf method [3, see Appendix E], the dipole reactance can be determined to be

$$X_A = \frac{\eta_0}{4\pi}\left\{\begin{array}{l} 2Si(\beta l) + \cos(\beta l)[2Si(\beta l) - Si(2\beta l)] \\ \qquad -\sin(\beta l)\left[2Ci(\beta l) - Ci(2\beta l) - Ci\left(\dfrac{2\beta a^2}{l}\right)\right] \end{array}\right\} \tag{5.54}$$

At small values of l, the center-fed dipole is capacitive. At $l \approx 0.5\lambda$ (0.48λ actually, depending on the wire thickness), the reactance becomes inductive.

The input resistance and reactance are plotted in Figures 5.11–5.13, for a

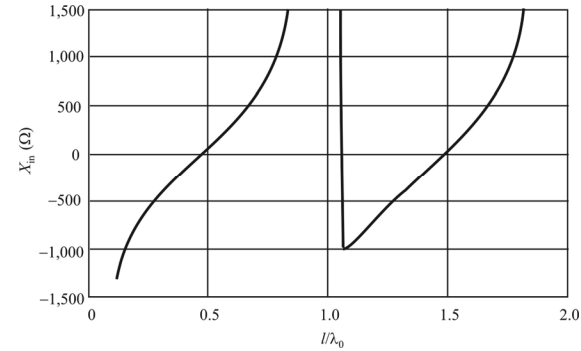


Figure 5.12 Dipole input reactance when $a_0 = 0.0005\lambda$.

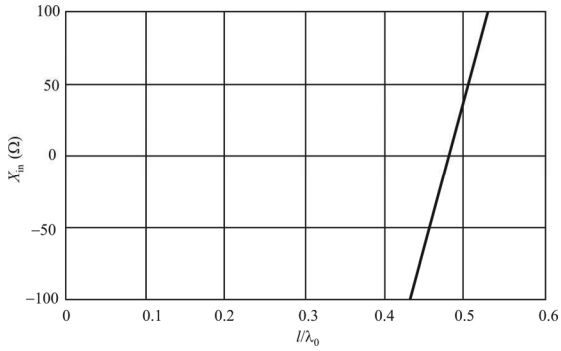


Figure 5.13 Dipole input reactance expanded at $l/\lambda_0 = 0.5$ when $a_0 = 0.0005\lambda$.

dipole of radius $a_0 = 5\times10^{-4}\lambda$. If the dipole is 0.5λ in length, the input impedance is found to be approximately $(73 + j42.5)\Omega$. The first dipole resonance ($X_{in} = 0$) occurs when the dipole length is slightly less than one half-wavelength. The exact resonant length depends on the wire radius, but for wires that are electrically very thin, the resonant length of the dipole is approximately 0.48λ. As the wire radius *increases*, the resonant length *decreases* slightly. Note that R_{in} and X_{in} for a lossless dipole explode (open circuit) at $\beta l = 0, 2\pi, \ldots$, which corresponds to $l = 0, \lambda, 2\lambda, \ldots$, since the denominator goes to zero. This can be seen in Figures 5.11 and 5.12.

Figures 5.14 shows the reactance of a dipole as well as a monopole (the reactance of a monopole is twice that for a dipole, so the ordinate scale labeled end-driven is for a monopole, while that labeled center fed is for a dipole). Likewise, Figure 5.15 shows the resistance of dipoles and monopoles with the same comment applying to the ordinate scales.

5.2.3 Half-Wave Dipole

In this case $l = \lambda/2$ so

$$\beta l = \frac{2\pi}{\lambda}\frac{\lambda}{2} = \pi \tag{5.55}$$

The current distribution is exactly half a wavelength as illustrated in Figure 5.16. The far fields of this antenna are given by

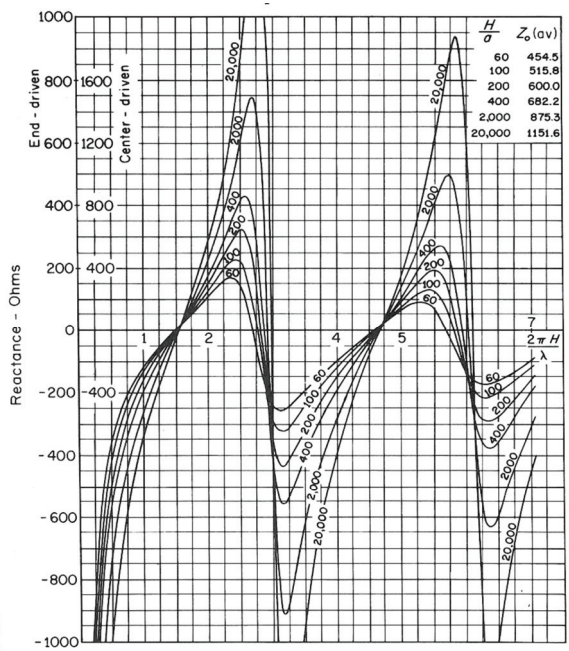


Figure 5.14 Monopole and dipole input reactance as a function of l/λ for several values of l/a_0. Monopole = end-fed, dipole = center-fed. (Source: [4], © Prentice-Hall 1968. Reprinted with permission.)

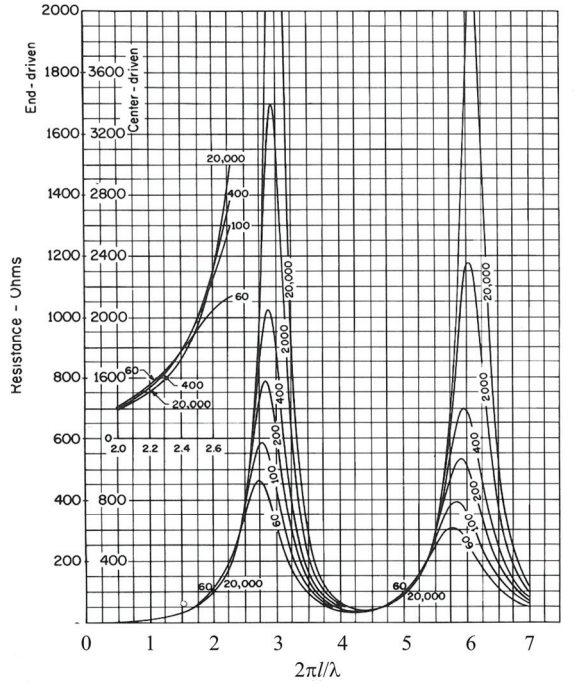


Figure 5.15 Monopole and dipole input resistance as a function of l/λ for several values of l/a_0. Monopole = end-fed, dipole = center-fed. (Source: [4], © Prentice-Hall 1968. Reprinted with permission.)

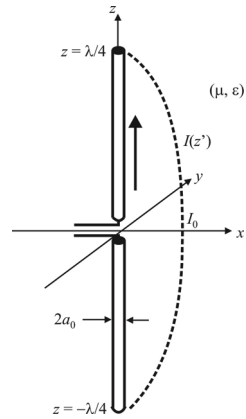


Figure 5.16 Center-fed half-wave dipole.

$$\vec{E}_\theta \approx j\eta_0 \frac{e^{-j\beta r}}{2\pi r} I_0 \frac{\cos\left(\frac{\pi}{2}\cos\theta\right)}{\sin\theta}\vec{u}_\theta \tag{5.56}$$

$$\vec{H}_\phi \approx j \frac{e^{-j\beta r}}{2\pi r} I_0 \frac{\cos\left(\frac{\pi}{2}\cos\theta\right)}{\sin\theta}\vec{u}_\phi \tag{5.57}$$

The radiation intensity function is given by

$$U(\theta,\phi) = \frac{\eta_0 |I_0|^2}{8\pi^2} \frac{\cos^2\left(\frac{\pi}{2}\cos\theta\right)}{\sin^2\theta} \tag{5.58}$$

while the radiated power is

$$\begin{aligned} \Pi &= \frac{\eta_0 |I_0|^2}{4\pi} f(\pi) \\ &= \frac{\eta_0 |I_0|^2}{8\pi} 2.428 \end{aligned} \tag{5.59}$$

Since

$$\Pi = \frac{1}{2}|I_0|^2 R_r \tag{5.60}$$

and $\eta_0 = 120\pi$ in air, then

$$R_r = 73.14\,\Omega \tag{5.61}$$

The half-wave reactance calculated with (5.54) is 42.5Ω yielding the total input impedance of

$$Z_{in} = 73.14 + j42.5\,\Omega \tag{5.62}$$

and we can see that a half wave dipole is inductive ($+j\omega L$).

The directivity is given by

$$D(\theta,\phi) = 1.641\frac{\cos^2\left(\frac{\pi}{2}\cos\theta\right)}{\sin^2\theta} \tag{5.63}$$

Since

$$\left.\frac{\cos^2\left(\frac{\pi}{2}\cos\theta\right)}{\sin^2\theta}\right|_{max} = \left.\frac{\cos^2\left(\frac{\pi}{2}\cos\theta\right)}{\sin^2\theta}\right|_{\theta=\pi/2} = 1 \tag{5.64}$$

the maximum directivity is given by

$$D_0 = 1.641 \qquad (2.15\text{dB}) \tag{5.65}$$

with an effective aperture of

$$A_e = 0.131\lambda^2 \tag{5.66}$$

The polarization is the same as the axis of the antenna—if it is vertically oriented relative to the Earth, then the polarization of the signal will be orthogonal to the Earth's surface as well (at least at the location of the antenna). For $l = \lambda/2$ the elevation 3 dB beamwidth is 78°, while palpably the azimuth beamwidth is

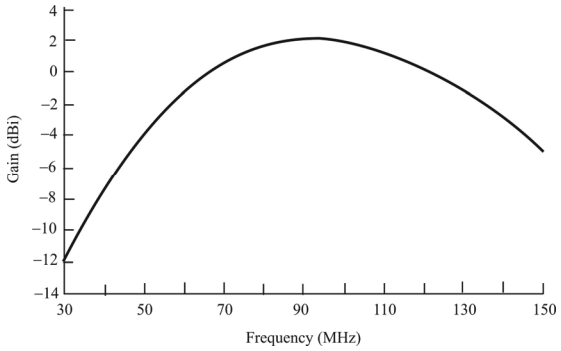


Figure 5.17 Typical gain variation with frequency of a half-wave dipole cut to resonate at $l = \lambda/2$ at 90 MHz.

360°. The 3 dB point is where the radiation pattern is one-half its value at foresight. A half-wave dipole has about a 2 dB gain over an isotropic antenna and can be used economically at virtually any frequency where the radiation patterns are adequate. If additional directionality can be allowed or is required, other antenna types provide better characteristics. The effective area of a half-wave dipole is $A_e = 1.64\lambda^2/4\pi$.

Sometimes broadbanding techniques are used with ES antennas. As discussed in Chapter 3, one such technique places resistive loading along the length of the antenna. Such techniques lower the Q of the antenna, because the Q is given by the ratio of the reactance to the resistance. This method, however, also lowers the gain of the antenna at all frequencies; thus, bandwidth is traded for gain. In some cases this trade-off may be favorable.

Typical gain characteristics of a dipole antenna, cut for $l = \lambda/2$ at 90 MHz (λ = 5.3 m), are shown in Figure 5.17. At 90 MHz, the antenna has a 2 dBi gain while at the low end it exhibits a 12 dBi loss. At 150 MHz, a loss of 5 dBi is typical. A 10 W transmitter using a tuned dipole antenna at a range of 40 km from a UAS ES system at 3,000 meters AGL produces a signal with power P_R out of a receive antenna according to the expression

$$P_R = \frac{P_T G_T G_R}{r^4}\left(h_T h_R\right)^2 \tag{5.67}$$

This target will produce a signal level out of this antenna of –64 dBm at 90 MHz. At 30 MHz it will be 14 dB less, or –78 dBm. At 150 MHz, it will be 7 dB less, or –71 dBm. For most receiving systems, these signal levels are very adequate for intercept purposes.

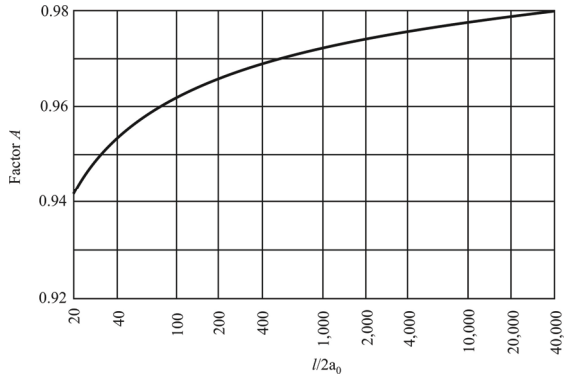


Figure 5.18 Dipole multiplication factor.

In a ground application where h_R = 2 m [an on-the-move (OTM) application), at 10 km this antenna will produce –104 dBm at 90 MHz, –118 dBm at 30 MHz, and –111 dBm at 150 MHz. These levels are approaching the system noise level and, when the system is used in noisy external environments, probably would not be adequate for intercept.

5.2.4 Antenna Length

The length of a dipole is the principal determining factor for the operating frequency of the dipole antenna. Although the antenna may be an electrical half wavelength, or multiple of half wavelengths, it is not exactly the same length as the wavelength for a signal traveling in free space. It means that an antenna will be slightly shorter than the length calculated for a wave traveling in free space.

For a half-wave dipole the length for a wave traveling in free space is calculated with $\lambda / 2 = c / (2f)$, and this is multiplied by a factor A. Typically it is between 0.96 and 0.98 and is mainly dependent upon the ratio of the length of the antenna to the thickness of the wire used as the element. Its value can be approximated from Figure 5.18.

5.3 Folded Dipole

A variation of the dipole, known as a folded dipole, provides a wider bandwidth and a considerable increase in feed impedance. The folded dipole is formed by taking a standard dipole and then taking a second conductor and joining the two at both ends. In this way a complete loop is made as shown in Figure 5.19. The two

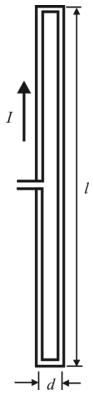


Figure 5.19 Folded dipole.

parallel dipoles of radius a and length l form a narrow loop. The center-to-center separation of the parallel wires is d, which is small relative to the wavelength [5]. If the conductors in the main dipole and the second or "fold" conductor are the same diameter, then we find that there is a fourfold increase in the feed impedance. In free space, this gives a feed impedance of around 300Ω. In addition, the antenna has a wider bandwidth.

5.3.1 Characteristics of the Folded Dipole

In a standard dipole the currents flowing along the conductors are in phase and as a result there is no cancellation of the fields and radiation occurs. When the second conductor is added in the folded dipole, this can be considered as an extension to the standard dipole with the ends folded back to meet each other. As a result the currents in the new section flow in the same direction as those in the original dipole. The currents along both the half-waves are therefore in phase and the antenna will radiate with the same radiation characteristics as a simple half-wave dipole.

The impedance increase can be deduced from the fact that the power supplied to a folded dipole is evenly shared between the two sections that make up the antenna. This means that when compared to a standard dipole, the current in each conductor is reduced in half. As the same power is applied, the impedance has to be raised by a factor of 4 to retain balance.

The 2-D patterns for the folded dipole is shown in Figure 5.20. These patterns are very similar to the dipole. Comparing Figure 5.20 with Figure 5.3, we can see that for the same conditions, the horizontal plane view, being the familiar figure-8

pattern, is flatter than the dipole The dipole half-power beam width is about 90°, while that for the folded dipole is about 50°. This is attributable to the array nature of the folded dipole. The dipole 3-D pattern side and front views are illustrated in Figure 5.21.

Denote the input impedance of a shorted two-wire transmission line of length $l/2$ by Z_T and represent the input impedance of the unfolded dipole by Z_d. Then the input impedance of the folded dipole is given by

$$Z_{\text{folded dipole}} = \frac{4Z_T Z_d}{Z_T + 2Z_d} \tag{5.68}$$

When $l = \lambda/2$, the input impedance of the equivalent transmission line is that of a shorted quarter-wavelength transmission line, which is an open circuit:

$$Z_T = jZ_0 \tan\left(\frac{\beta l}{2}\right) = jZ_0 \tan\left(\frac{2\pi}{\lambda}\frac{\lambda}{4}\right) = jZ_0 \tan\left(\frac{\pi}{2}\right) \to \infty \tag{5.69}$$

The impedance of the half-wave folded dipole at $l = \lambda/2$ becomes

$$Z_{\text{folded dipole}} = \lim_{Z_T \to \infty} \frac{4Z_T Z_d}{Z_T + 2Z_d} = \lim_{Z_T \to \infty} \frac{4Z_d}{1 + 2Z_d / Z_T} = 4Z_d \approx 300 + j170\,\Omega \tag{5.70}$$

since $Z_d \approx 73 + j42.5\Omega$ at resonance. The reactive component can be conjugate matched. Alternately, if the length of the half-wave folded dipole is reduced somewhat, the antenna can be made to resonate where $X_{in} = 0$, so that $Z_{\text{folded dipole}} \approx 300 + j0$, essentially a perfect match and no matching network is needed at all.

The half-wave folded dipole can be made resonant with an impedance of approximately 300Ω, which matches the transmission line impedance of twin lead. Thus, the half-wave folded dipole can be connected directly to a twin-lead transmission line without requiring any matching network (see Chapter 4). Also, the folded dipole has a wider bandwidth than a dipole of the same size. Because of this impedance property and the lack of requirement for a resistive matching network, the folded dipole is often used as the driven element for the Yagi and log-periodic antennas discussed in Chapters 11 and 12.

For some applications it makes sense to use two separate wire sizes for the legs of the folded dipole as shown in Figure 5.22. When the antenna is half wavelength the input impedance is given approximately by

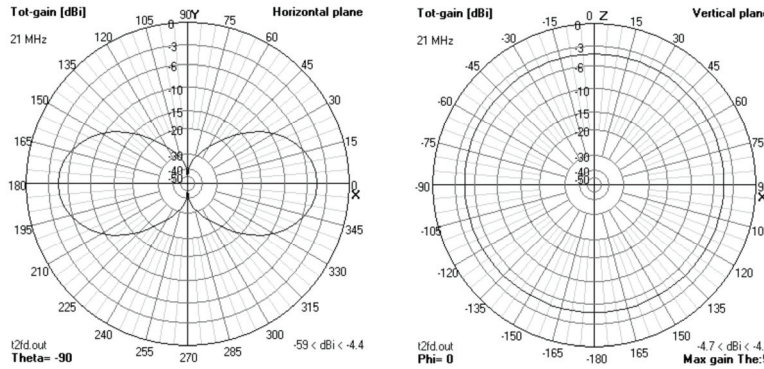


Figure 5.20 Folded dipole 2-D patterns. These patterns are for a horizontally oriented dipole, such as might be used for TV reception with a Yagi or log-periodic antenna.

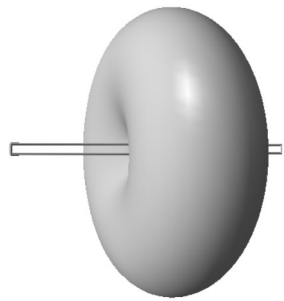


Figure 5.21 Folded dipole 3-D pattern front and side views. In free space. Horizontally oriented folded dipole as shown.

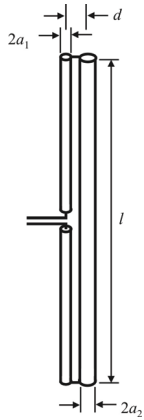


Figure 5.22 Folded dipole different radii. Folded dipole with unequal element radii a_1, and a_2, with separation distance d.

$$Z_{\text{in}} = (1+k)^2 Z_{\text{dipole}} \tag{5.71}$$

where the dimension variables are shown in Figure 5.22. When a_1 and a_2 are considerably less than d, which is typically the case, then k is approximately

$$k \approx \frac{\ln(d/a_1)}{\ln(d/a_2)} \tag{5.72}$$

5.3.2 Applications

Folded dipoles are sometimes used on their own, but they must be fed with a high impedance source, typically 300Ω. However, they find more uses when a dipole is incorporated in another antenna with other elements nearby, forming an array. This has the effect of reducing the dipole impedance. To ensure that it can be fed conveniently, a folded dipole may be used to raise the impedance again to a suitable value. This is the typical driven element of a Yagi antenna array discussed in Chapter 11.

5.4 Dipoles with Traps

In wideband applications where it is necessary to cover only a limited number of frequency bands, it is not necessary to have frequency-independent antennas but

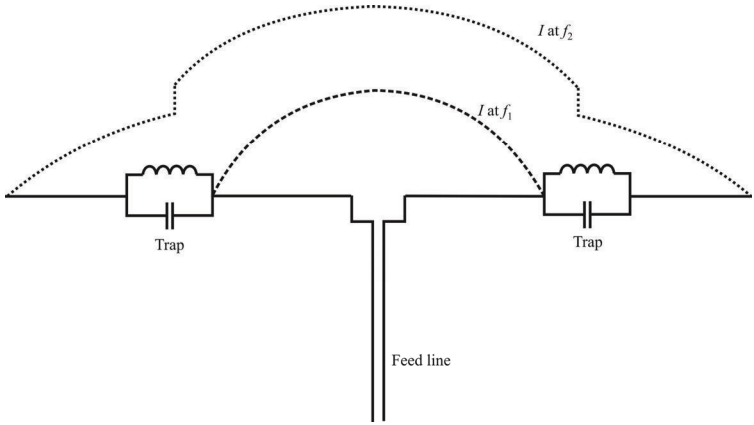


Figure 5.23 Dipole with traps.

rather antennas that can operate at the two (or more) specific bands. A dual-band antenna can be constructed from a single center-fed dipole (or its respective monopole) by means of tuned traps (see Figure 5.23). Each trap consists of a tuned parallel *LC* circuit. At frequency f_1, for which the whole dipole is >>λ/2 long, the trap behaves as an inductor. This reduces slightly the resonant length of the dipole, and has to be taken into account in the antenna design. At another frequency $f_2 > f_1$, the traps become resonant and effectively cut off the outer portions of the dipole, making the dipole much shorter and resonant at this new frequency. If the traps, for example, are in the middle of the dipole's arms, then the antenna can operate equally well at two frequencies separated by an octave. It should be noted that the isolation of the outer portions of the dipole depends not only on the high impedance of the trap but also on the impedance of this outer portion. When the outer portions are about λ/4 long, they have very low impedance compared to the trap's impedance, and their currents are negligible so they are effectively mismatched.

This is not the case if the outer portions are about λ/2 long, however. When the outer portions of the dipole are about λ/2 each, they represent very high impedance themselves in series with the trap. They are no longer isolated. A coil alone can form a trap at certain (very high) frequency because of its own distributed capacitance. It can also act as a 180^o phase shifter (the coil may be viewed as a coiled-up λ/2 element).

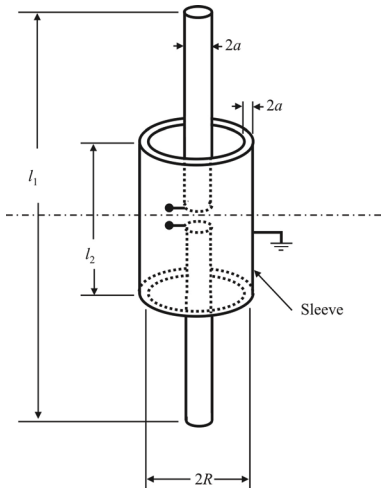


Figure 5.24 Sleeve dipole.

5.5 Sleeve Dipoles

The impedance of a dipole antenna is very frequency sensitive. The addition of a sleeve to a dipole can increase the bandwidth up to more than an octave. A depiction of a sleeve dipole is given in Figure 5.24. A conducting tubular sleeve is placed over the center of the dipole. In practice, this tube is often replaced by two or more conductors close to the antenna that simulate the effects of the sleeve. The resulting antenna, illustrated in Figure 5.25, is referred to as the open-sleeve dipole antenna.

The performance of an open-sleeve dipole where the spacing between the antenna and the reflecting surface shown, S_d, is 0.29λ, and the spacing between conductors (parasitic elements and the antenna proper) is S, with a frequency range of 225–400 MHz, as illustrated in Figure 5.26 [6]. The variables corresponding to this performance are given in Table 5.1. This frequency range corresponds to the low UHF military frequency range.

5.6 Concluding Remarks

We examined the principal characteristics of realistic dipole antennas in this chapter. In all cases it was assumed that the diameter of the dipole was much

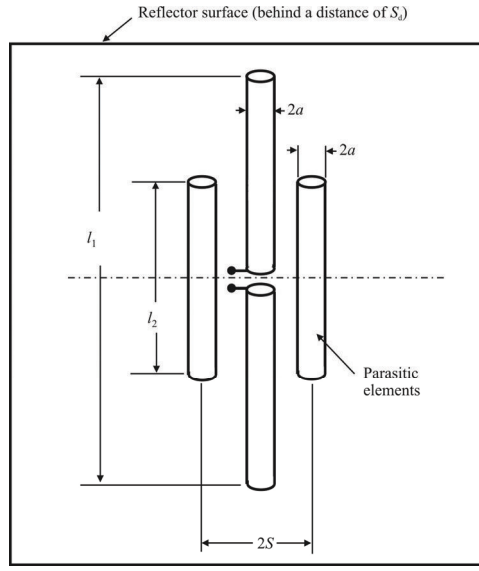


Figure 5.25 Sleeve dipole open. Open-sleeve dipole. The parasitic elements make the sleeve dipole a dipole array.

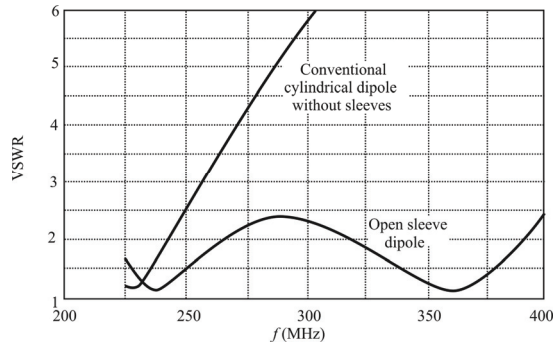


Figure 5.26 Sleeve dipole performance vs conventional..

Table 5.1 Open-Sleeve Dipole Dimensions for Lowest VSWR 225–400 MHz

Parameter	Electrical Dimension at Lowest Frequency (225 MHz)	Electrical Dimension at Highest Frequency (400 MHz)
$2a$	0.026λ	0.047λ
l_1	0.385λ	0.684λ
l_2	0.216λ	0.385λ
S	0.0381λ	0.0677λ
S_d	0.163λ	0.29λ

Source: [6].

smaller than the length. The far field radiation patterns are maximized in the horizontal plane and are theoretically null off the ends of the dipoles. As the length of the dipole gets larger than a wavelength, scalloping starts to form in the radiation patterns so they no longer resemble a figure 8.

We saw that the real part of the driving point impedance of a dipole ranges from on the order of 10Ω for very short lengths to 10^5 or higher at $l = \lambda$, falling to about 100Ω at $l = 3\lambda/2$ and then rising again. The imaginary part of the driving point impedance is capacitive at short lengths, becoming inductive at about $l = \lambda/2$. As the length increases, the impedance gets increasingly more inductive until $l = \lambda$, where it sharply becomes capacitive again until the length approaches $l = 3\lambda/2$, where it becomes inductive.

We investigated the characteristics of a few different dipole structures, in particular, the folded dipole, dipoles with traps, and sleeve dipoles. All of these architectures are attempts to broaden the bandwidth of the basic dipole antenna.

References

[1] Balainis, C. A., "Antenna Theory: A Review," *Proceedings IEEE*, January 1992, pp. 7–23.

[2] Abramowitz, M., and I. A. Stegun, *Handbook of Mathematical Functions*, New York: Dover, 1972, pp. 231–232.

[3] Pender, H., and K. Mcllwain, *Electrical Engineers' Handbook Electric Communication and Electronics* 4theEd., New York: Wiley, 1963, pp. 6-65–6-71.

[4] Jordan, E. C., and K. G. Balmain, *Electromagnetic Waves and Radiating Systems,* 2nd ed., Englewood Cliffs, NJ: Prentice-Hall, 1968, pp. 566–567.

[5] Thiele, G. A., E. P. Ekelman, Jr., and L. W. Henderson, "On the Accuracy of the Transmission Line Model of the Folded Dipole," *IEEE Transactions on Antennas and Propagation*, Vol. AP-28, No. 5, September 1980, pp. 700–703.

[6] Stutzman, W. L., and G. A. Thiele, *Antenna Theory and Design*, New York: Wiley, 1981.

Chapter 6

Monopole Antennas

6.1 Introduction

Monopole antennas are perhaps the simplest kind of antenna. Because of this, they are ubiquitous—there are applications just about anywhere antenna are required. A monopole antenna consists of a single, straight element mounted vertically on an infinite ground plane [1] as illustrated in Figure 6.1. It is a popular configuration for tactical radios in the lower VHF frequency ranges, where, of course, the infinite ground plane is only approximated. When possible, a system of radials is employed to approximate a solid ground plane. We explore the salient properties of monopole antennas in this chapter, and, in particular, those characteristics that make them useful as EW antennas.

All antennas that are situated close to the ground are affected by that ground to some extent. First, the ground forces the antenna's radiation pattern to appear in the half-space above the ground. The monopole has twice the power in the hemisphere above the ground compared with the power in either hemisphere of symmetry for the dipole in free-space. The nature of this reflection of energy above the ground is governed by the polarization of the antenna and the ground conditions. The dielectric constant and conductivity of the ground determine how well the ground acts as a conductor and hence a reflector. In addition, the ground absorbs energy from the antenna; this energy is wasted in the ground's intrinsic resistance. The placement of an artificial ground, called a counterpoise, can decrease the ground losses and enhance the performance of a horizontal antenna.

This chapter is structured as follows. We first examine vertical monopole characteristics and, in particular, the radiating field structures and the impedance characteristics. We then investigate the effects of various ground planes on the vertical monopole characteristics. Next we examine characteristics of the inverted L and F monopole structures. We then investigate the properties of the folded monopole. The sleeve monopole is similar to the sleeve dipole and, just as we examined the properties for sleeve dipoles in the last chapter, we investigate the

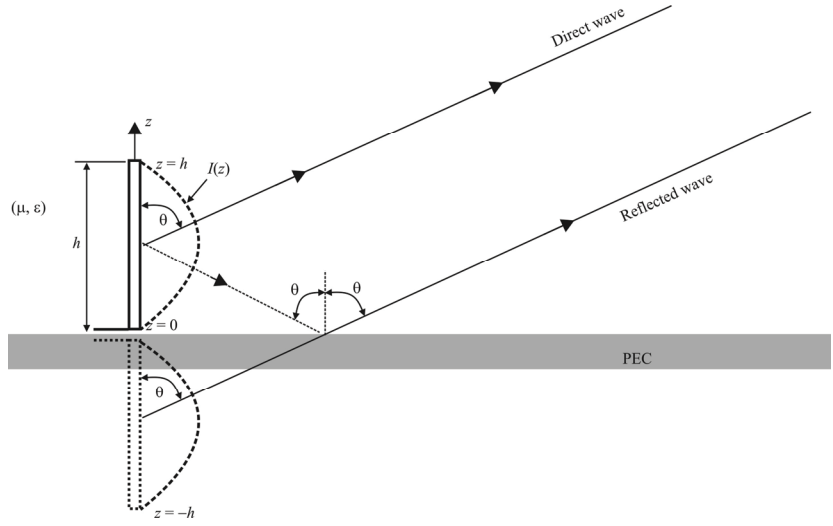


Figure 6.1 Vertical monopole antenna above ground showing how the ground reflects the signal, causing the image of the monopole below ground.

sleeve monopole properties next. Conical and discone antenna structures have characteristics similar to those of the monopole, so we include them next. Lastly, we discuss the properties of helix antennas.

6.2 Vertical Monopoles

A vertical monopole over a PEC can be viewed as half of a dipole and the other half of the dipole is created by a reflecting surface. This gives rise to the antenna image—it appears that the other half of the "dipole" formed by the monopole is under the ground. See Figure 6.1. Whereas the dipole must be fed with balanced feeds, the monopole is fed with asymmetric (unbalanced) feed lines where the signal is carried by one wire and the other is connected to ground. The popular connection possible with coaxial cable is totally compatible with a monopole, whereas for a dipole a balun[1] is required. One advantage of unbalanced lines is that most transmitters/receivers are configured with unbalanced output/input.

[1] Balun = BALanced to UNbalanced converter. A device that converts a balanced feed line to an unbalanced one. See Chapters 20 and 21.

The radiated power of a monopole is half that of its corresponding dipole with the same current. Therefore, the radiation resistance is also half that of the dipole; however, the directivity of the monopole is double that of the dipole. Thus, a $\lambda/4$ monopole has a radiation resistance of $R_r = 36.5\ \Omega$ and a directivity of $D = 3.26$.

The radiation characteristics of the monopole are shown in Figure 6.2 in two dimensions and Figure 6.3 in three dimensions. The azimuth radiation pattern is also 360^o, but the ground plane truncates the elevation pattern. The elevation 3 dB beamwidth is approximately 45^o and the maximum gain of these antennas is 3 dB higher than the associated dipole, or about 5.1 dBi. Thus, the effective area of a monopole is approximately $A_{\text{eff}} = 3.3\lambda^2 / 4\pi$.

6.2.1 Vertical Monopole Radiation Patterns

Some significant conclusions follow from the image theory and the above discussion:

- The field distribution in the upper half-space is the same as that of the respective free-space dipole.
- The total radiated power of a monopole is half the power radiated by its dipole counterpart, since it radiates in half-space (but its field is the same). As a result, the beam solid angle of the monopole is half that of the respective dipole and its directivity is twice the directivity of the dipole.

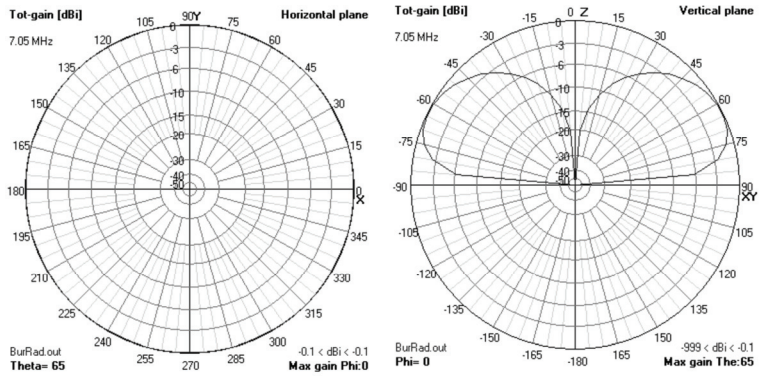
$$D_0^{\text{mp}} = \frac{4\pi}{\Omega_A^{\text{mp}}} = \frac{4\pi}{\frac{1}{2}\Omega_A^{\text{dp}}} = 2D_0^{\text{dp}} \tag{6.1}$$

The current distribution for a vertical monopole antenna of height l when driven by a sinusoidal source is a standing wave of the form

$$I(z) = \frac{I(0)}{\sin(\beta l)} \sin[\beta(l - z)], \qquad 0 \le z \le l \tag{6.2}$$

where $I(0)$ is the current at the feedpoint (the base).

As indicated in Figure 6.1, the PEC ground plane creates an image of the antenna with a current distribution identical to that for the lower arm of a dipole. Together with this image, the monopole antenna appears to be a center-fed dipole for the upper half-space. There is negligible penetration of fields into a high conductivity ground for a monopole antenna, and all that radiation is directed into



(a) (b)

Figure 6.2 λ/2 Vertical monopole. (a) The horizontal plane pattern is omnidirectional, while (b) the vertical plane pattern (b) forms half a doughnut shape.

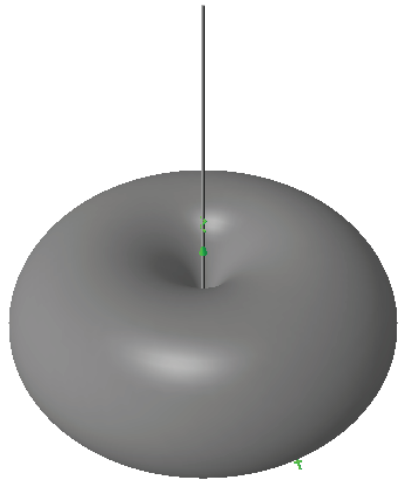


Figure 6.3 3-D monopole pattern.

the upper half-space creating a power density for any elevation angle θ that is twice as high as that for a dipole radiating the same amount of power (see Figure 6.1). This gives a directivity and gain of the monopole antenna that is twice that for the double-length dipole.

The radiation patterns of a few typical monopole antennas as a function of length are given in Figure 6.4 along with the current distribution patterns. Because a monopole antenna is symmetric with respect to angle φ in the azimuthal plane (x-y plane), these patterns are independent of angle ϕ and depend only on zenith angle θ (see Figure 6.1). The power density is given by

$$P_{\mathrm{d}}(\theta) = \frac{30}{\pi r^2}\frac{P}{R_{\mathrm{a}}}\frac{F^2(\theta)}{\sin^2(\beta l)} \tag{6.3}$$

where P is the radiated power, r is the distance to the field point, $\beta = 2\pi/\lambda$ is the propagation constant, and $F^2(\theta)$ is the pattern factor given by

$$F^2(\theta) = \left[\frac{\cos(\beta l\cos\theta) - \cos(\beta l)}{\sin\theta}\right]^2 \tag{6.4}$$

Because monopole antennas help reduce the length of the necessary dipole by a factor of 2 and result in directivities that are twice as large, vertical monopole antennas above ground are extensively used for commercial amplitude modulated (AM) broadcasting in the frequency range 535–1,605 kHz. For this application, the wavelengths are long on the order of 200–600 m, and the monopole antennas are immensely helpful in reducing the required height. It is necessary, of course, to create a good conductivity ground, which for dry or rocky soil conditions is often obtained by burying a conducting screen made of radially spread metal wires with angular separations of 2–3° that extend to a radius at least equal to a quarter-wavelength. This screen, called a counterpoise, is often buried a few inches below the surface of the natural ground but may also be left slightly above ground for rocky or otherwise difficult terrain.

For long distances, there is a precipitous drop in the surface power density at all frequencies due to the increased effect of the curvature of the Earth [2].

Monopole antennas are also the antennas of choice for very low-frequency (VLF) (3 to 30 kHz) and low-frequency (LF) (30 to 300 kHz) communication systems. For these applications, the height l of the antenna is generally a small fraction of the wavelength. From Figure 6.5, we see that for small values of l/λ, the feed-point resistance R_{a} is very small, on the order of a few ohms, and the ohmic losses resulting from the surface resistance of the antenna can be significant by comparison. This results in reduced antenna radiation efficiency η_{r}, given by

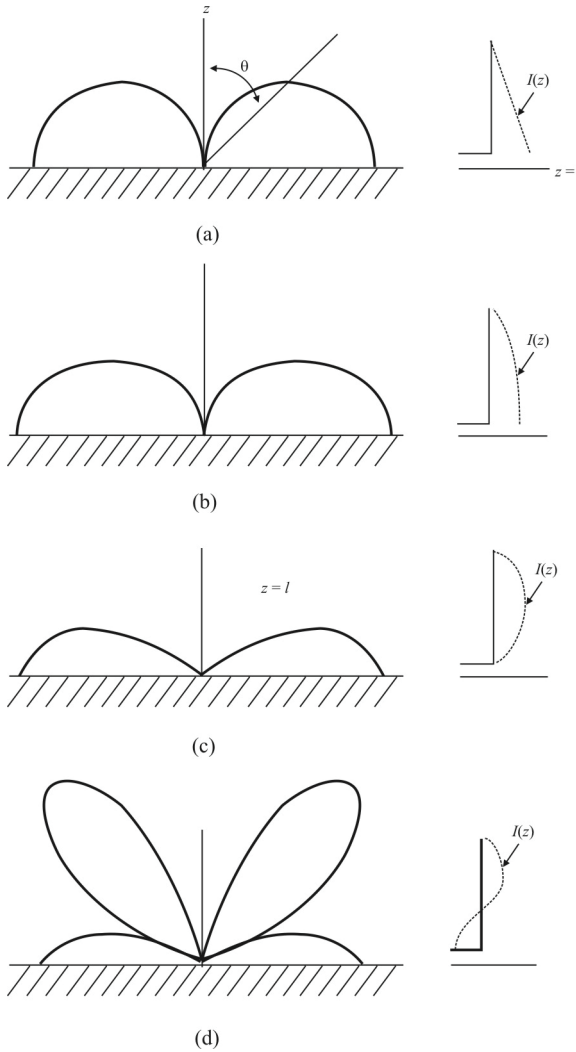


Figure 6.4 Monopole patterns. Radiation patterns and current distributions for some vertical monopoles above an infinite PEC. (a) $h/\lambda = 0.05$ (short monopole, triangular current distribution; (b) $h/\lambda = 0.25$ (quarter-wave monopole); (c) $h/\lambda = 0.5$ (half-wave monopole); and (d) $h/\lambda = 0.75$ (3/4 wave monopole).

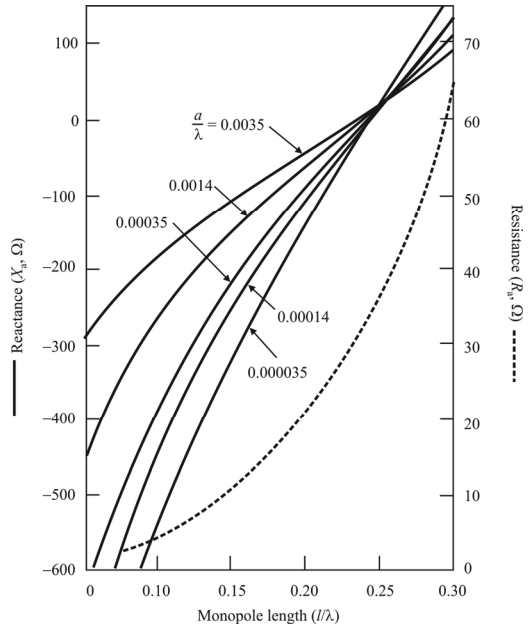


Figure 6.5 Variation in feed point resistance (R_a) and reactance (X_a) for a short end-fed monopole above an infinite PEC ground of height $l\lambda$.

$$\eta_r = \frac{R_a}{R_a + R_0} \tag{6.5}$$

where R_0 is the effective ohmic resistance of the antenna.

6.2.2 λ/4 Vertical Monopole

This is a straight wire of length $l = \lambda/4$ mounted over a ground plane. From the discussion above, it can be expected that the quarter wavelength monopole should be very similar to the half-wavelength dipole in the hemisphere above the ground plane.

- Its radiation pattern is the same as that of a free-space λ/2 dipole, only that it is nonzero only for $0^o < \theta \leq \pi/2$ (above ground).
- The field expressions are the same as those of the λ/2 dipole.

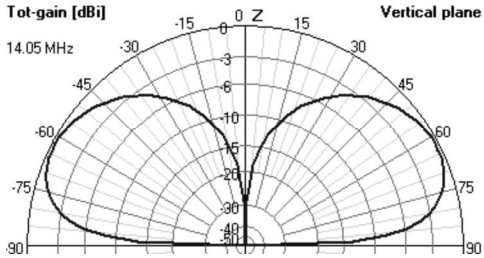


Figure 6.6 λ/4 Vertical monopole. The 0 dB curve corresponds to 5.15 dBi.

- The radiated power of the λ/4 monopole is half that of the λ/2 dipole.
- The directivity of the λ/4 monopole is

$$D_0^{\text{mp}} = 2D_0^{\text{dp}} = 2 \times 1.643 = 3.286 \tag{6.6}$$

The vertical plane radiation pattern of a λ/4 monopole is shown in Figure 6.6. We see that the takeoff angle ~65°. The 0 dB curves in this figure correspond to 5.15 dBi.

6.2.3 λ/2 Vertical Monopole

Likewise, the λ/2 monopole is a vertical wire with length $l \sim 0.48\lambda$ and the self-impedance is ~2 kΩ (real, $X_a = 0$ at $h = \lambda / 2$). The vertical radiation pattern is compared to that of a short vertical monopole in Figure 6.7. The take-off angle is ~ 75°. Ground currents at the base of the antenna are small so radials are less critical.

6.2.4 Short Vertical Monopoles

In the HF range, a λ/4 or λ/2 mobile vertical monopole antenna is generally impractical. The monopole, like the dipole, can be shortened and, since short wire antennas have a capacitive characteristic, can be resonated with a loading coil. The feed point impedance can be quite low (~10Ω, see Figure 6.5) with a good ground system, so an additional matching network is required. Best results are obtained when the loading coil is at the center.

6.2.5 Vertical Monopole Feed Point Impedance

The currents and the charges on a monopole are the same as on the upper half of its dipole counterpart, but the terminal voltage is only half that of the dipole. The

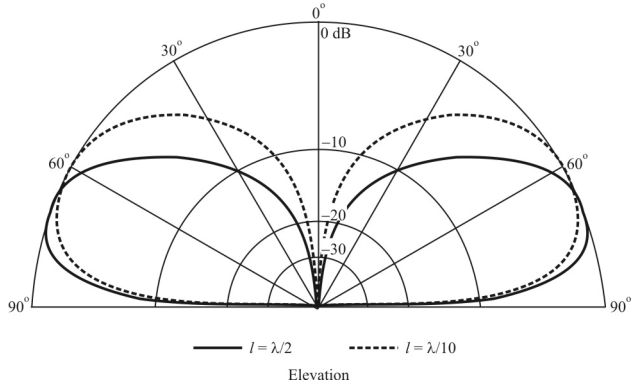


Figure 6.7 Vertical monopole elevation patterns. Comparison of the radiation pattern of a λ/2 monopole with that of a λ /10 monopole.

input impedance of a monopole is therefore only half that of the respective dipole:

$$Z_{\text{in}}^{\text{mp}} = \frac{1}{2} Z_{\text{in}}^{\text{dp}} \tag{6.7}$$

Therefore the input impedance characteristics versus length can be plotted as in Figures 5.17 and 5.18. These charts are repeated here in Figures 6.8 and 6.9. The input reactance is expanded in Figure 6.10 and we see, just as for the dipole, the monopole is resonant at a length just below λ/2 (actually at 0.48λ). Note that these figures do not necessarily correspond to a short dipole, as Figure 6.5 does.

The characteristics of the feed-point resistance and reactance of the short monopole antenna above ground are illustrated in Figure 6.5 as a function of normalized length l/λ, where λ is the free-space wavelength at the radiation frequency. We see from Figure 6.5 that the reactance X_a depends on the conductor radius a, whereas the feed-point resistance R_a is relatively independent of conductor radius a for thin antennas ($a/l << 1$). We assume this thin condition is met everywhere in this chapter.

6.2.6 Vertical Monopole Radiation Resistance

The radiation resistance of a short vertical monopole over a PEC in free space with a height h and at a wavelength λ is

$$R_a = 40\pi^2 \left(\frac{h}{\lambda} \right)^2 \tag{6.8}$$

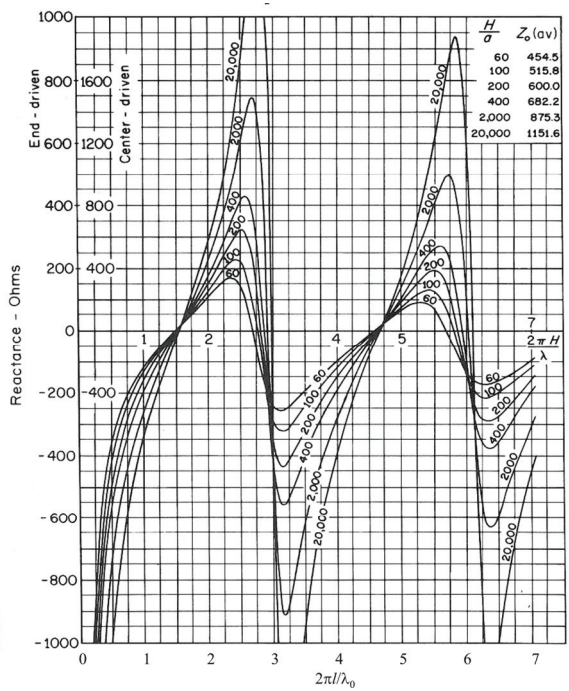


Figure 6.8 Monopole and dipole input reactance as a function of l/λ for several values of l/a_0. Monopole = end-fed, dipole = center-fed. $l = H$, a is the radius. (Source: [4], © Prentice-Hall 1968. Reprinted with permission.)

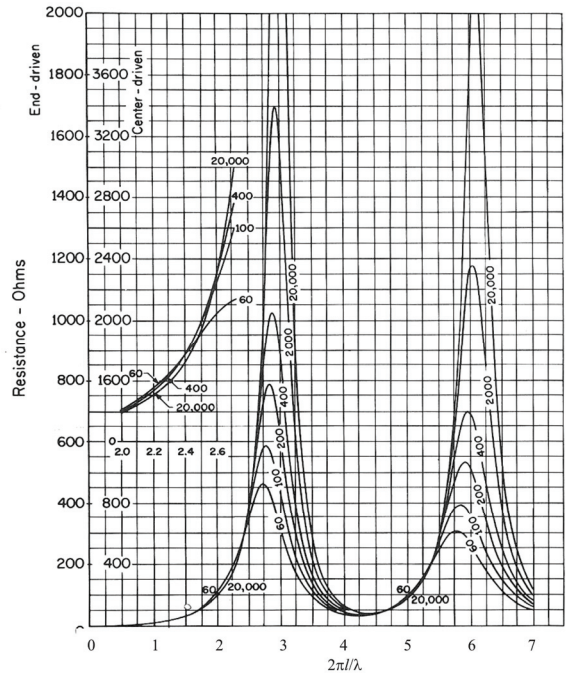


Figure 6.9 Monopole and dipole input resistance as a function of l/λ for several values of l/a_0. Monopole = end-fed, dipole = center-fed. $l = H$. (Source: [4], © Prentice-Hall 1968. Reprinted with permission.)

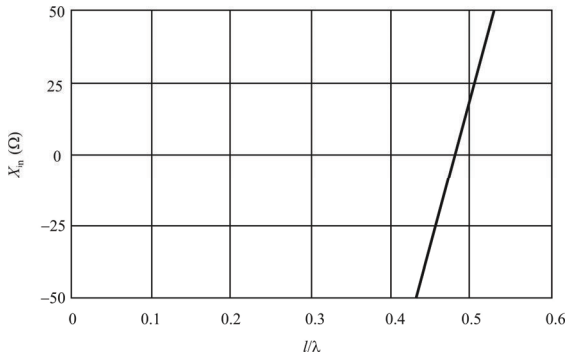


Figure 6.10 Vertical monopole input reactance expanded.

and the input reactance is capacitive up to just below $l/\lambda = 0.5$; then it turns inductive. This is cyclic as seen in Figure 6.8.

The radiation resistance of a dipole of length l is given by (5.18) as

$$R_r^{dp} = 20\pi^2 \left(\frac{l}{\lambda}\right)^2 \tag{6.9}$$

The associated monopole for this dipole is one whose height above ground is $h = l/2$, so the radiation resistance for this monopole is given by

$$\begin{aligned} R_r^{mp} &= 40\pi^2 \left(\frac{l/2}{\lambda}\right)^2 \\ &= 10\pi^2 \left(\frac{l}{\lambda}\right)^2 \end{aligned} \tag{6.10}$$

which is one-half that of the dipole.

6.2.6.1 Monopole Radiation Resistance in Free Space

The characteristics of an arbitrary length monopole antenna in free space are well known. The contribution to the radiation resistance resulting from the power radiated into the angular direction (θ, ϕ) over a ground plane is

$$dR(\theta,\phi) = \frac{\eta_0}{4\pi^2 \sin^2 \beta l} \frac{[\cos(\beta l \cos\theta) - \cos\beta l]^2}{\sin\theta} d\theta d\phi \tag{6.11}$$

where l is antenna length, η_0 is the free-space impedance, and β is the wave number. The total radiation resistance is given by the integral over the half plane of (6.11):

$$R_r = \int_0^{2\pi}\int_0^{\pi} dR(\theta,\phi) \tag{6.12}$$

$$= \frac{\eta_0}{4\pi^2 \sin^2 \beta l}\int_0^{2\pi}\int_0^{\pi} \frac{[\cos(\beta l \cos\theta) - \cos\beta l]^2}{\sin\theta} d\theta d\phi \tag{6.13}$$

The $d\theta$ integral in (6.13) cannot be evaluated in closed form, so it must be evaluated numerically. Note that $\beta l = 2\pi/\lambda \times l = 2\pi l/\lambda$. Equation (6.13) is plotted in Figure 6.11. We note that, comparing Figure 6.11 to Figure 6.9, most of the input resistance consists of the resistive loss in a monopole antenna since the radiation resistance is less than 80Ω.

6.2.7 Elevated Vertical Monopoles

When monopoles are elevated above ground, they typically require a counterpoise, which is somewhat like a ground plane, except it is a resonant structure. Such an antenna configuration is shown in Figure 6.12. The patterns for an elevated monopole with a counterpoise that is elevated less than a wavelength are similar to the patterns for a monopole that is not elevated. When elevated multiple wavelengths, however, scalloping ensues, as illustrated in Figure 6.13. The antenna shown in Figure 6.14 is elevated above the ground plane beneath it by about five wavelengths. The vertical radiation pattern is shown in Figure 6.13 (the horizontal pattern is still omnidirectional in the horizontal plane). The pattern in 3-D is shown in Figure 6.15.

The effects of the method of elevating the antenna above the ground (e.g., antenna mast) and of the technique for feeding the antenna in Figure 6.14 are not included in the patterns shown. If either of these contains conductive materials, such as a metal antenna mast or coaxial cable running between the transmitter on the ground and the antenna, the patterns shown are affected dramatically.

6.2.8 Top Loading

Top loading of the monopole antenna illustrated in Figure 6.16 is often used to improve the feed-point or antenna-equivalent resistance R_a. Because the top set of wires acts as a capacitance to ground rather than the open end of the antenna, the current at the top of the antenna ($z = h$) no longer needs to go to zero as it does for

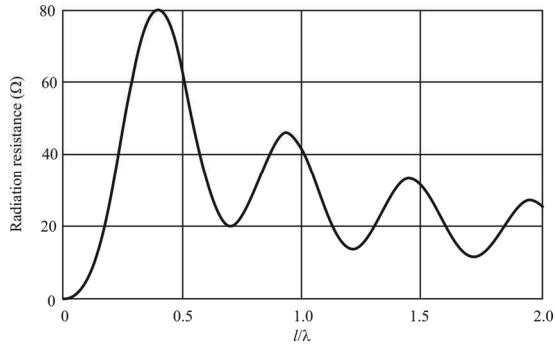


Figure 6.11 Radiation resistance.

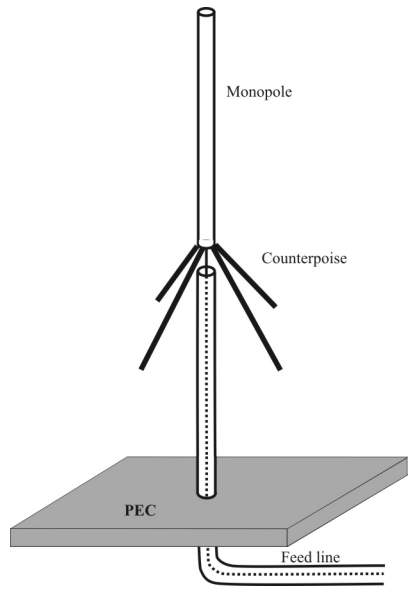


Figure 6.12 Monopole with counterpoise. Elevated monopole with counterpoise.

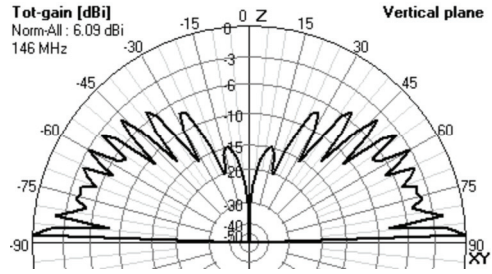


Figure 6.13 Vertical plane monopole pattern elevated with counterpoise.

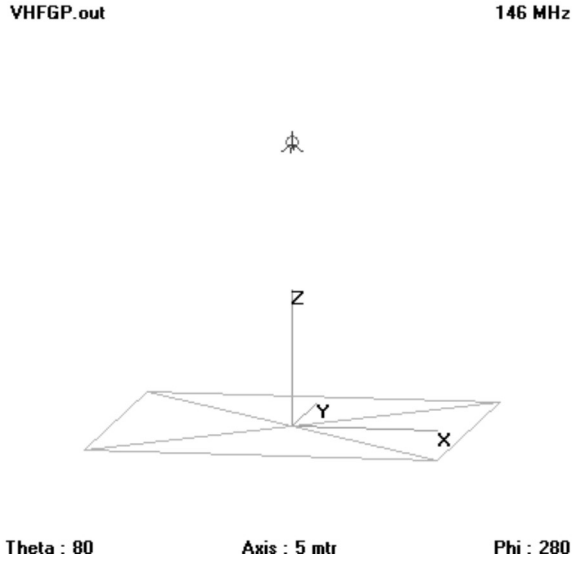


Figure 6.14 Monopole elevated with counterpoise.

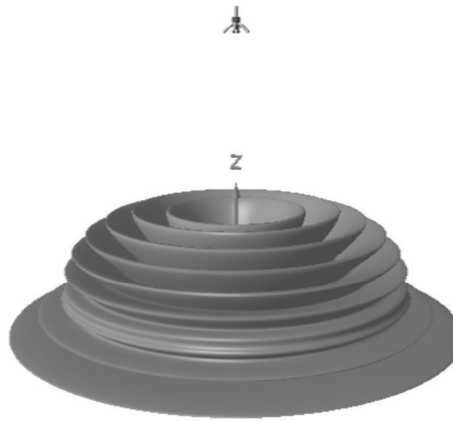


Figure 6.15 Monopole elevated 3-D pattern.

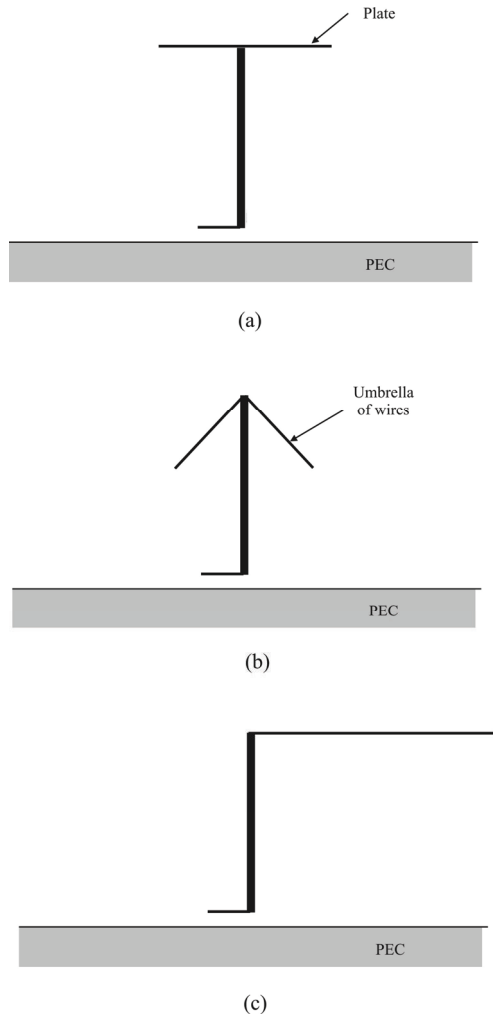


Figure 6.16 Top loading of monopoles: (a) plate, (b) umbrella of wires, and (c) inverted L.

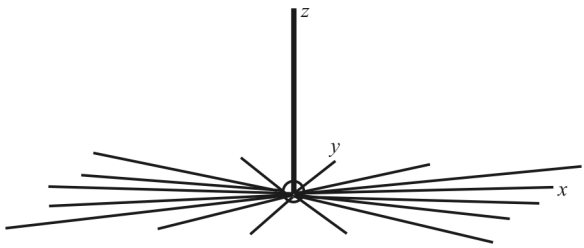


Figure 6.17 Monopole with 16 radials.

an open-ended monopole [from (6.2), $I = 0$ at $z = h$]. The upper region of the antenna can, therefore, support a substantial RF current, resulting in improved radiation efficiency. The antenna-equivalent resistance (6.8) for open-ended short monopoles may be improved by a factor of 2 to 3 by use of top loading, resulting in higher radiation efficiency.

6.3 Ground Planes

To a large extent, the quality of the ground plane determines the monopole antenna pattern close to the horizon. The better the ground plane, the lower to the horizon the antenna pattern falls. Often the ground plane is approximated by radial wires as illustrated in Figure 6.17, where 16 radials are shown. An artificial ground plane is also referred to as a counterpoise, as has already been mentioned.

Equation (6.3) gives the power density radiated from a monopole antenna for a ground of infinite conductivity and for flat earth. For finite conductivity ground $(\sigma < \infty)$, the power density at the surface of the Earth diminishes more rapidly than the square of the distance r from the antenna [2]. Furthermore, the rate at which the power density at the surface of the ground diminishes with distance is faster at higher frequencies. The rate of reduction of the radiated power density as compared to (6.3) is also higher for lower conductivity soil such as sandy or rocky lands, cities, and so forth, rather than propagation over seawater, marshy, or pastoral lands, and so forth.

The effects of imperfect ground conditions are illustrated in Figure 6.18 when $l = \lambda/4$. Two ground conditions are shown: $\sigma = \infty$, corresponding to a PEC ground, and $\sigma = 10^{-2}$ S m^{-1}, corresponding to very good ground conditions. Some ground condition values are given in Table 6.1.

Ground loss resistance is a virtual resistance that represents power lost in the ground system. The reflection losses represent the reduction in signal strength due to reflection of signals from the ground (the ground is a poor reflector for

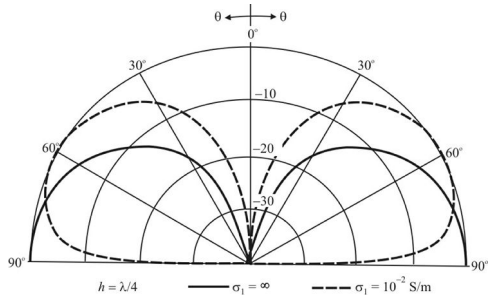


Figure 6.18 Effects of imperfect ground.

Table 6.1 Ground Conductivities and Dielectric Constants

Terrain	Conductivity, σ (mhos m^{-1})	Dielectric Constant, ε (esu)	Quality
Sea water	5	80	Excellent
Fresh water	0.008	80	Poor
Dry, sandy, flat coastal land	0.002	10	Poor
Marshy, forested flat land	0.008	12	Poor
Rich agricultural land, low hills	0.01	15	Poor
Pastoral land, medium hills, and forestation	0.005	13	Very good
Rocky land, steep hills	0.002	10	Poor
Mountainous (hills up to 1,000 m)	0.001	5	Poor
Cities, residential areas	0.002	5	Very poor
Cities, industrial areas	0.0004	3	Extremely poor

Source: [5].

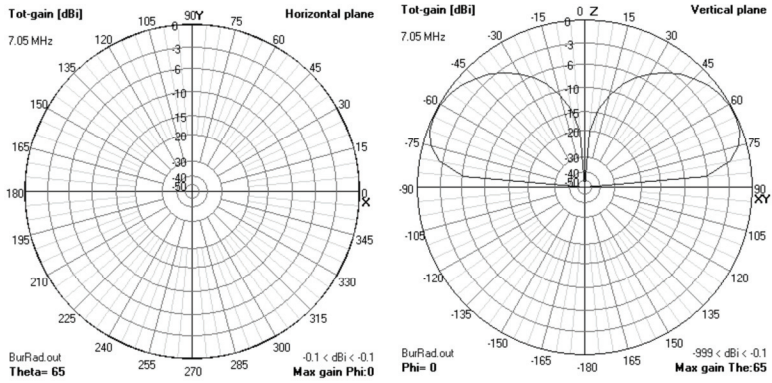
vertically polarized RF). The ground is part of the vertical antenna, not just a reflector of RF, unless the antenna is far removed from the Earth (usually only true in the VHF region). RF currents flow in the ground in the vicinity of a vertical antenna. The region of high current is near the feed point for verticals less than λ/4 long and is ~λ/3 out from the feed point for a λ/2 vertical. To minimize losses, the conductivity of the ground in the high current zones must be very high. Ground conductivity can be improved by using a ground radial system, or by providing a counterpoise. Counterpoises are most practical in the VHF range. At HF, radial systems are generally used. Ground radials can be made of almost any type of wire. The radials do not have to be buried; they may lie on the ground. The radial should extend from the feed point like spokes of a wheel. The length of the radials is not critical because they are not resonant. They should be as long as possible, however. For small radial systems ($N < 16$), the radials need only be λ/8 long. For large ground systems ($N > 64$) the length should be ~ λ/6. Elevated counterpoise wires are usually λ/4 long. Note that the radials used in a counterpoise are not grounded.

6.3.1 Uniform Ground Plane

Ground emplaced radials (as opposed to elevated resonant ground plane systems) can be rather small diameter wire since none of them carries appreciable current. They are effectively in parallel with the ground, which also carries currents. As opposed to elevated ground plane systems, ground radials supplement ground return currents in the earth and do not try to replace them entirely. Elevated ground plane radials need to be a bit longer than quarter-wavelength at the operating frequency since they are resonant structures.

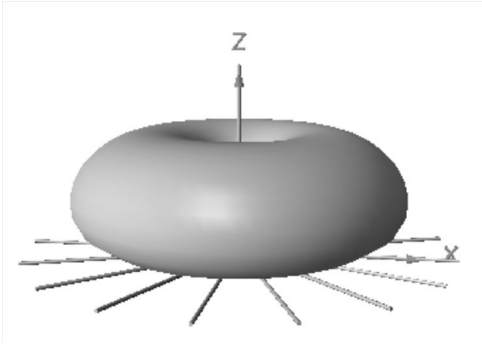
Ground radials seldom need to be longer than 0.28λ regardless of the height of the antenna, even a half-wave vertical radiator. Due to detuning caused by the ground, insulated wires laid on the ground tend to be electrically 0.25λ when the physical length is close to 0.28λ. A shortened monopole with loading coils will have a more compact near field where the majority of the antenna field is located. The ground needs only reach out as far as the near field extends. Close to the antenna, field intensity drops off with the square of the distance from the base of the antenna.

The vertical radiation patterns of the antenna shown in Figure 6.17 are shown in Figure 6.19. We can see the characteristic omnidirectional pattern in the horizontal plane and the lobes in the vertical plane. The effects of the imperfect ground plane are evident in Figure 6.19(b). The three-dimensional pattern is shown in Figure 6.20, which emphasizes the doughnut shape of the radiation pattern. The minimum number of ground radials is probably 8, but more is better.



(a) (b)

Figure 6.19(a, b) Monopole with 16 radials 2-D patterns with $l = \lambda/2$.



Figuer 6.20 Monopole with 16 radials 3-D pattern when $l = \lambda/2$.

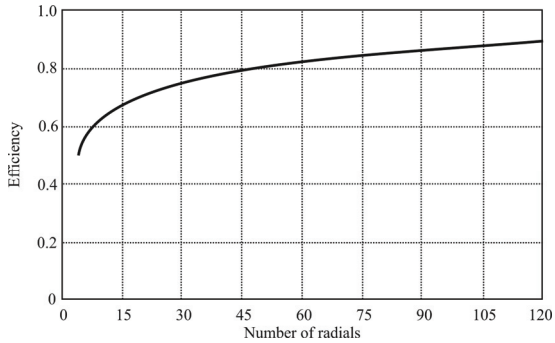


Figure 6.21 Efficiency improvement as a function of the number of radials.

Four ground radials are insufficient. However, as illustrated in Figure 6.21, more than 32 radials produces diminishing returns as the efficiency more or less levels off after that point [6].

The sufficient lengths, beyond which the returns on increased length diminish, are shown in Figure 6.22 [6]. The radial length also affects the gain of the monopole antenna. These effects are shown in Figure 6.23 [6]. There is a clear gain advantage by using longer radials. However, a radial length of a quarter-wavelength should be sufficient in most cases. Relative gains past this point are on the order of 0.5 dB.

The antennas used in handheld phones of cell systems and PCS are of the monopole variety. In those cases there is a very poor ground plane both in terms of consistency and conductivity. The antenna patterns of such antennas are best

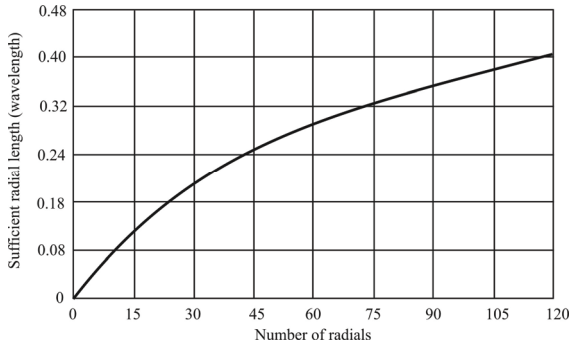


Figure 6.22 Radial length versus the number of radials.

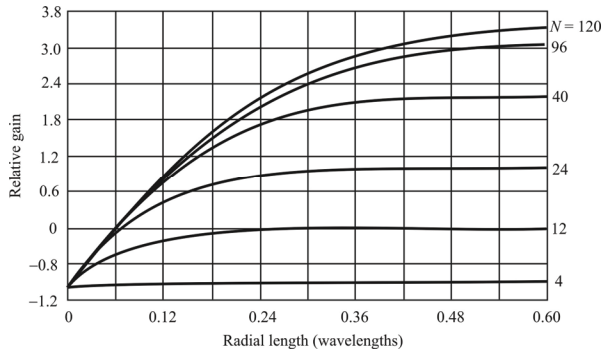


Figure 6.23 Relative gain as a function of radial length.

modeled as randomly varying.

6.3.2 Nonuniform Ground Planes

Monopole antennas are popular for mounting on vehicles to include military tactical vehicles. In cases when the monopole is on a movable object it is difficult to establish a good ground plane, with radials or otherwise. The metal mass of the vehicle is typically substituted for the uniform ground plane oftentimes assumed when analyzing monopole performance.

The effects on the monopole pattern shown in Figure 6.1 when the antenna is mounted on the rear bumper of the vehicle shown in Figure 6.24, that, in turn, is sitting on a ground plane are illustrated in Figures 6.25 and 6.26. While these are modeling results, they vividly demonstrate the complexity of the resultant pattern. Furthermore, each such nonuniform ground plane will have its own perturbations.

The elevation pattern shown in Figure 6.25 barely resembles a monopole pattern at all. Most of the energy is radiated forward, over the vehicle structure, with very little in the reverse direction. Likewise, the horizontal pattern in Figure 6.25 is not even close to omnidirectional, again with most of the energy launched in the direction corresponding to the maximum distance across the vehicle that is possible.

6.4 Inverted-L and Inverted-F Antennas

When a monopole is bent the L-antenna (or inverted-L antenna) ensues as illustrated in Figure 6.27. The L-antenna has a reduced size and low profile

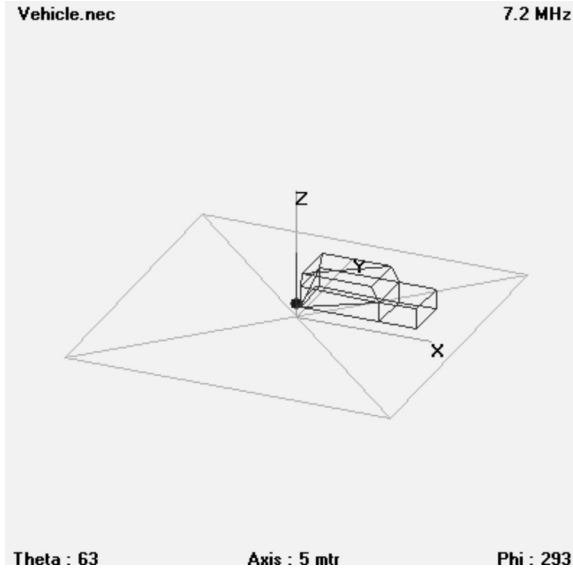
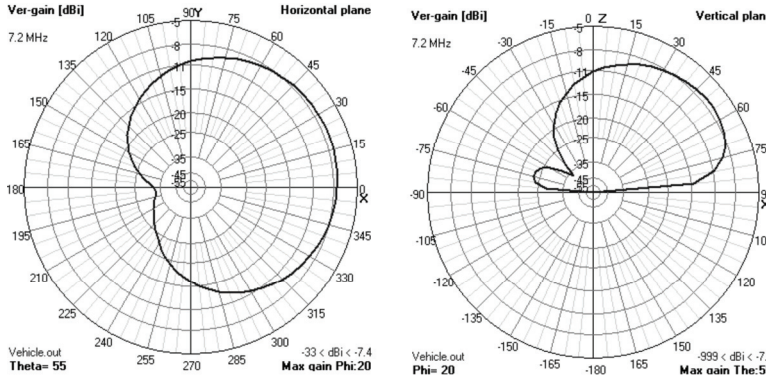
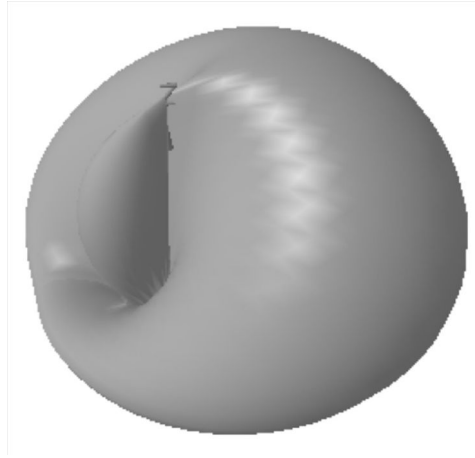


Figure 6.24 Vehicle upon which a monopole is attached to the rear bumper.



(a) (b)

Figure 6.25 Vehicle mounted monopole 2-D patterns: (a) horizontal and (b) vertical.



Figue 6.26 3-D monopole pattern mounted on a car.

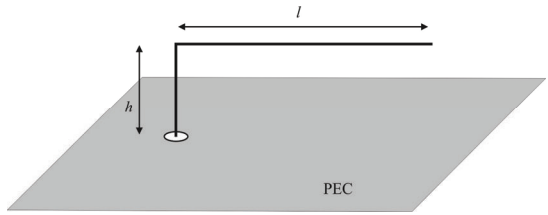


Figure 6.27 Inverted-L monopole.

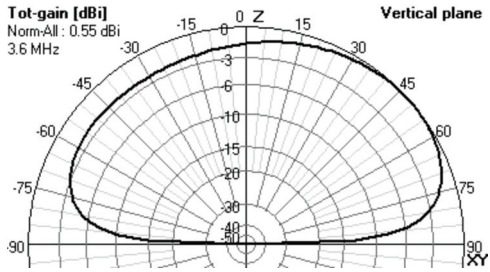


Figure 6.28 2-D vertical plane pattern for an inverted-L antenna.

compared to its monopole counterpart. These features were exploited in the first L-antenna designs, which were made for missile applications [7]. The reduction of the monopole's height results in reduced radiation resistance and the bandwidth is also reduced. In addition, the main angle of radiation is depressed as there is substantial radiation not only from the monopole h but also from the arm l. The null off the end of the vertical disappears as illustrated in the vertical 2-D pattern shown in Figure 6.28 and the 3-D pattern shown in Figure 6.29.

An adaptation of the L-antenna is the tapped inverted-F antenna (see Figure 6.30). It uses a tap along the bent arm segment L, to obtain better impedance matching. Both types of wire antennas can be analyzed using equivalent transmission line models. Their patterns in both principal planes are similar to that of inverted L, while the tapped inverted F has better impedance matching properties.

6.5 Folded Monopole

The folded monopole is a variation on the standard linear monopole. Folding results in an increase in the feedpoint resistance relative to the linear or open-ended monopole. The exact ratio of impedance transformation depends on the relative diameters of the wires. The same equation for the transformation ratio that applies to the folded dipole (5.70) holds for the folded monopole as well. While the ratio of wire diameters provides the key variable in the equation, the spacing between the wires plays a significant role in two ways. As seen in (5.70), the terms of the ratio itself are each ratios of diameter a_i to spacing d. The ratio of the impedance compared to the non-folded antenna is given by

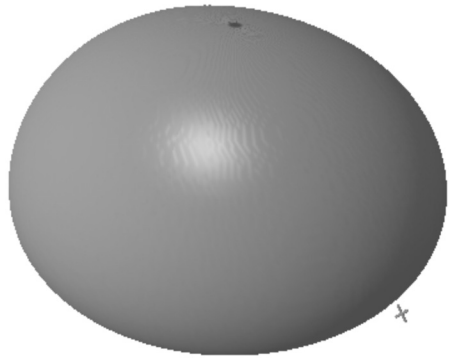


Figure 6.29 Inverted-L 3-D pattern.

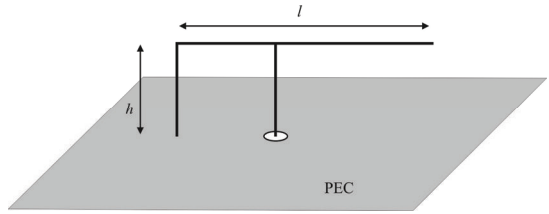


Figure 6.30 Inverted-L tapped monopole.

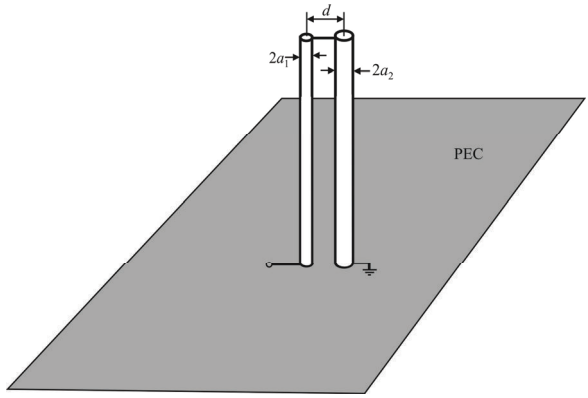


Figure 6.31 Folded monopole.

$$R = \left[1 + \frac{\ln(2d / a_1)}{\ln(2d / a_2)}\right]^2 \tag{6.14}$$

Second, the folded monopole is itself a transmission line. As such, the wires must be close enough to each other to ensure that the pair forms a transmission line and not a half-loop.

A folded monopole antenna is illustrated in Figure 6.31 where the feed point is connected to arm 1, and arm 2 is grounded at the bottom end. As for dipoles, this provides a method of increasing the feed-point impedance given in the following relationship

$$Z_{\mathrm{i}} = (1+c)^2 (R_{\mathrm{a}} + jX_{\mathrm{a}}) \tag{6.15}$$

where R_{a} and X_{a} plotted in Figures 6.8 and 6.9 are the resistance and reactance of the single arm (not folded) monopole antenna, and the factor c is given approximately by

$$c \approx \frac{\ln(d / a_1)}{\ln(d / a_2)} \tag{6.16}$$

In (6.16), a_1 and a_2 are radii of wires or rods used for arms 1 and 2, respectively, and d is the center-to-center spacing between the two arms of the antenna,

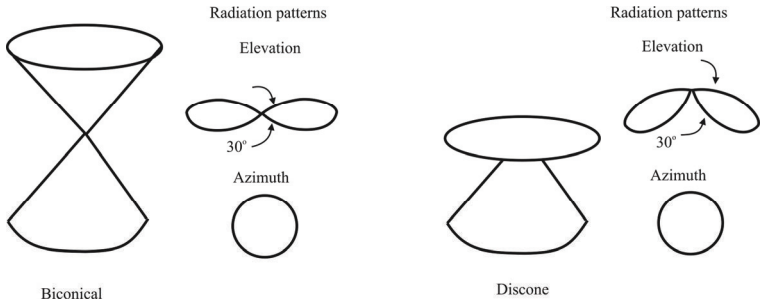


Figure 6.32 Biconical and discone antenna configurations and the radiation patterns.

generally much smaller than the height of each of the arms (see Figure 6.31). For a folded monopole antenna where equal radii arms are used, the feed-point impedance is, therefore, four times higher than that for a monopole antenna that is not folded. This can be truly advantageous, because the feed-point resistance may now be comparable to the characteristic impedance Z_0 of the transmission line, and the reactance of the antenna may easily be compensated by using a lumped element with a reactance that is negative of the reactance $(1 + c)^2 X_a$ at the terminals of the folded monopole antenna.

6.6 Conical/Biconical/Discone Antennas

Conical antennas are so called because they attach some form of cone to a monopole. Biconical antennas derive their name from the dual-cone structure of their configuration. Two cones, one inverted on top of the other as shown in Figure 6.32, form this antenna's configuration. The antenna feed point is at the center, where the two cones meet. When one of the cones is reduced to a plane, then the discone antenna results as shown. When there is only one cone, the antenna is said be conical or, a conical skirt antenna.

These antennas have an omnidirectional pattern in the azimuth plane but are more focused (flatter) in the vertical plane than a dipole or monopole without a cone. The maximum gain of the biconical antenna is approximately 3–4 dBi, while that of the discone is somewhat less. The biconical effective area is given by $A_{\text{eff}} = 2\lambda^2\pi$. Typically the bandwidth of the biconical is somewhat broader than that of the discone and the discone pattern is more focused downward as shown. The beamwidth in the vertical plane is approximately the same. Of course, the discone antenna can have the opposite orientation by putting the plane on the other side of the cone, forcing the pattern lobes in the upward direction.

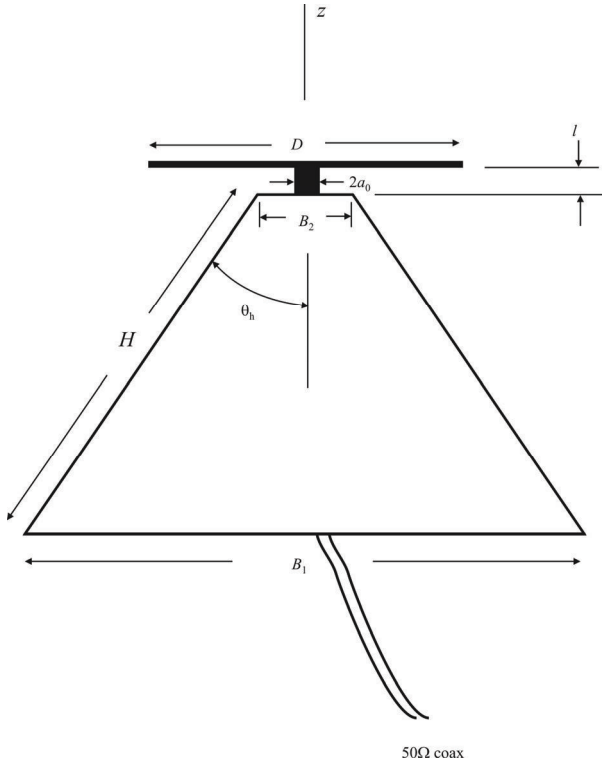


Figure 6.33 Discone antenna.

6.6.1 The Discone Antenna

The discone antenna is widely used where an omnidirectional (horizontal plane) wide bandwidth antenna is needed. One of its primary uses is for all types of radio scanning and monitoring applications for frequencies above 30 MHz.

Although the antenna could be made as a full disc and a cone, it normally is not for weight and wind loading considerations. The antenna consists of a top "disc" formulated from a number of elements arranged in a disc at the top, and further elements pointing downwards in the shape of a cone. The radiating part of the discone antenna shown in Figure 6.33 is the rod (monopole) of diameter $2a_0$ and length l between the cone and the disc at the top.

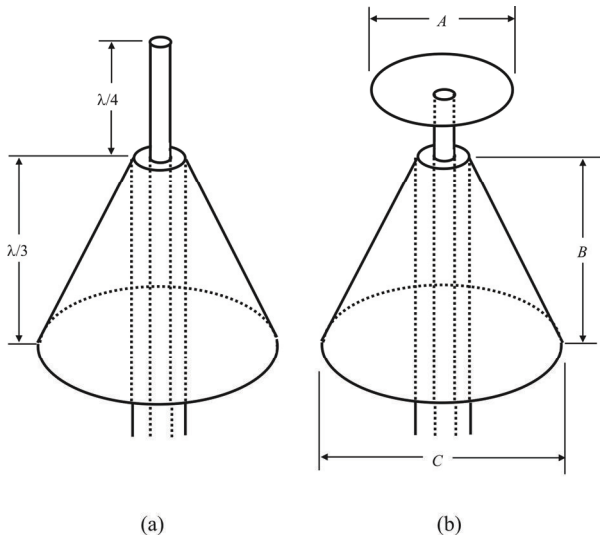


Figure 6.34 Cone antennas: (a) conical skirt monopole and (b) discone.

A conical antenna and a discone are compared in Figure 6.34. The conical skirt monopole in Figure 6.34(a) does not have the top loading plate as the discone does.

The discone can operate over frequency ranges of up to 10:1 depending on the particular design, and it also offers a relatively low angle of radiation (and reception). This makes it ideal for ground-based VHF/UHF applications as its greatest sensitivity is parallel to the Earth.

Although it is widely used for receiving applications, its wideband operation is useful for EW transmitting applications as well. Its wideband operation is both an advantage and nemesis for transmitting applications, however. The wideband width of the antenna means that spurious signals can be radiated more easily and the level of reflected power will vary over the operating range and may rise above acceptable limits.

The disc and cone wire elements simulate an electrically complete disc and cone from which the energy is radiated. As a result the greater the number of elements, the better the simulation, although in reality there is a balance between performance, cost, weight, and wind resistance. Often around six elements are used, but the number is not critical. Energy from the feeder meets the antenna and spreads over the surface of the cone from the apex towards the base until the vertical distance between the point on the cone and the disc is a quarter wavelength. In this way it is possible for the energy to be radiated or received efficiently.

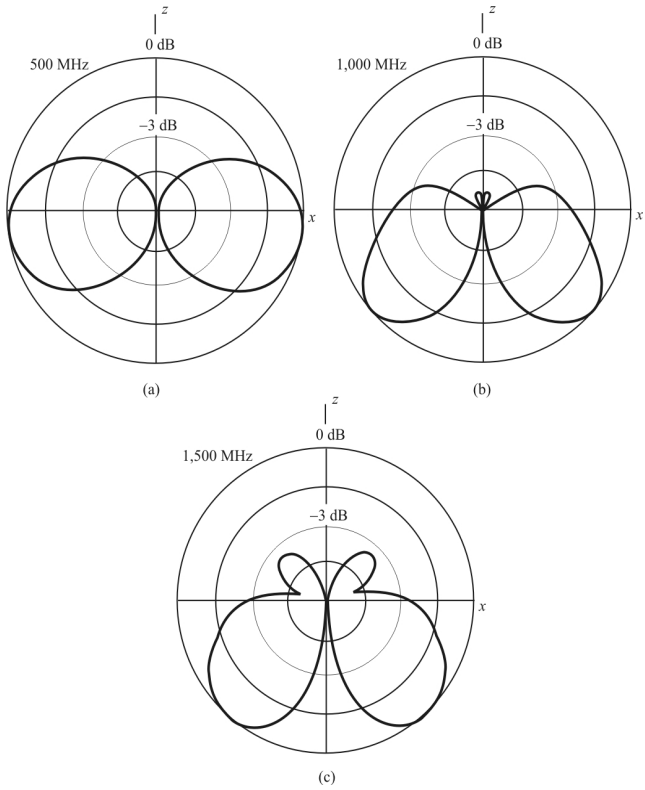


Figure 6.35 Discone patterns. (a) 500 MHz (λ = 60 cm), very similar pattern to a short dipole; (b) 1,000 MHz (λ = 30 cm); and (c) 1,500 MHz (λ = 15 cm), similar to an infinite conical monopole. Measured patterns for a discone with H = 21.3 cm, B_1 = 19.3 cm, $\theta_h = 25^o$.

For a given discone, the radiation pattern slopes further down at the horizon as illustrated in Figure 6.35 for the dimensions indicated. The patterns shown in Figure 6.35(a) closely resemble that of a short dipole. However, as the frequency is raised, side and back lobes begin to appear, much like the scalloping that develops in dipoles and monopoles as the antenna length is increased.

The antenna radiates and receives energy that is vertically polarized, and the radiation pattern is omnidirectional in the horizontal plane. The antenna radiates most of the energy at a low angle which it maintains over most of the operating range. Typically there is little change over a range of 5:1.

With the feed point at the top of the antenna, the current maximum point is also at the top. It is also found that below the minimum frequency the antenna presents a very bad mismatch to the feeder. However, once the frequency rises

above this point then a reasonable match to 50Ω coax is maintained over virtually the whole band of operation. Below the effective cutoff frequency, it has a considerable reactance and produces severe standing waves in the feed line. This happens approximately at the wavelength such that $H \sim \lambda / 4$.

These monopoles have much broader frequency band for their impedance variations (a couple of octaves) than the ordinary quarter-wavelength monopoles. They are a combination of the two basic antennas: the monopole antenna and the biconical antenna.

There are variations of the monopole/cone geometries, which focus on broader bandwidths rather than shaping the radiation pattern. They all provide omnidirectional radiation in the horizontal plane. The discone has the widest bandwidth among these types of antennas, with frequency performance similar to that of a highpass filter.

The radiation patterns for the discone antenna model shown in Figure 6.36 are shown in Figure 6.37 and 6.38 at a frequency of 55 MHz (λ = 5.45 m). The radius of the top-loading plate is approximately 0.5 m. We can see the figure 8 pattern of a dipole in free space. When the frequency is raised to 200 MHz, the radiation patterns in Figures 6.39 and 6.40 ensue. We see that the patterns are directed down much more. Backlobes can be identified as well. The nulls remain at the zenith, however. Typical dimensions of a discone antenna at the center frequency are: $D \sim 0.4\lambda$, $B_1 \sim 0.6\lambda$, $H \sim 0.7\lambda$, $45^o < 2\theta_h < 75^{o,}$ and $\delta << \lambda$. The typical input impedance is designed to be 50Ω. Optimum design formulas are [8]: $B_2 \sim \lambda_U / 75$ at the highest operating frequency, $\delta \sim (0.3 \div 0.5)B^2$, and $D \sim 0.7B_1$. $D_1 \sim 0.7B$.

6.7 Sleeve Monopole Antennas

The impedance of monopole antennas is very frequency sensitive, as we can see in Figuress 6.8 and 6.9. The addition of a sleeve to a monopole can increase the bandwidth up to more than an octave. A typical sleeve monopole is illustrated in Figure 6.41. The monopole is fed with a coaxial transmission line and the inner conductor of the transmission line is connected to the antenna [9, 10].

The exterior of the sleeve acts as the radiating element while the inside wall of the sleeve acts as the outer conductor of the coaxial cable. The length of the sleeve is a design variable and ranges in value from zero (no sleeve) to the full length of the monopole. The antenna dimensions shown affect the impedance more than the radiation pattern.

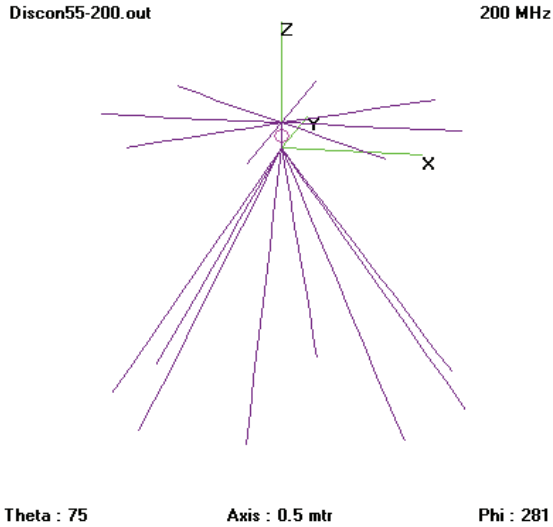


Figure 6.36 Wire model approximation to a discone antenna.

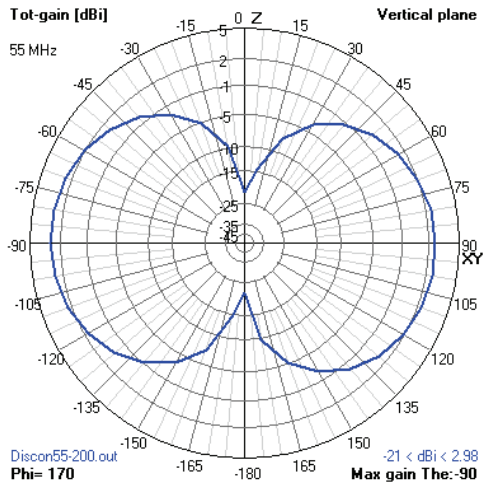


Figure 6.37 Vertical pattern for the discone shown in Figure 6.36 for f = 55 MHz.

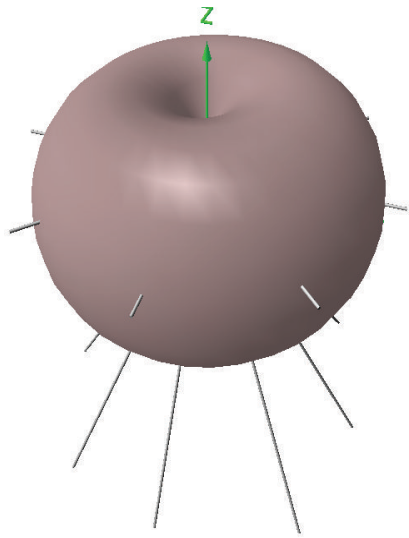


Figure 6.38 3-D pattern for the discone shown in Figure 6.36 at f = 55 MHz.

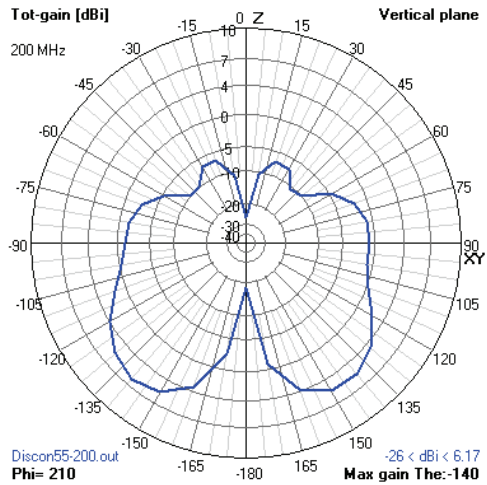


Figure 6.39 2-D vertical plane pattern for the antenna in Figure 6.36 at f = 200 MHz.

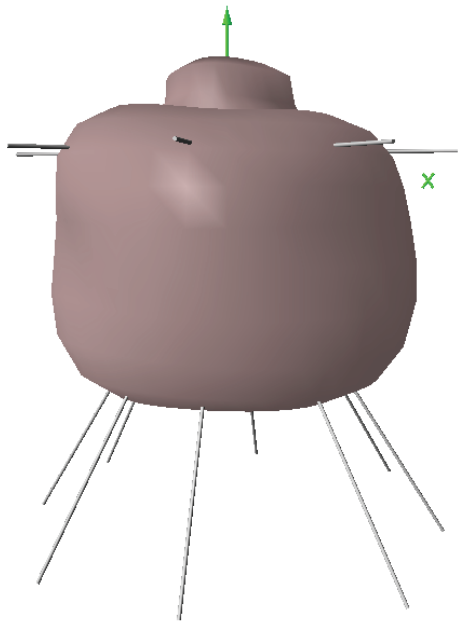


Figure 6.40 3-D pattern for the antenna shown in Figure 6.36 at f = 200 MHz.

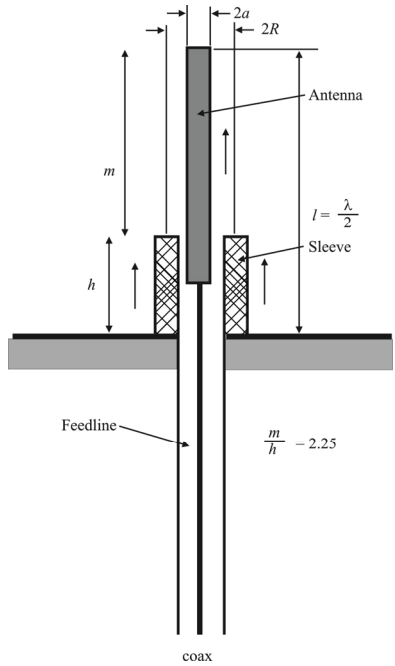


Figure 6.41 Sleeve monopole structure.

The first resonance of this structure occurs at approximately $l = \lambda/4$. It has been determined experimentally that the value of m / h shown, $m / h = 2.25$, yields a nearly optimum radiation pattern over the range of 4:1.[2] The ratio m/h has little impact for $l < \lambda/2$. For longer lengths, however, this ratio has significant impact on the radiation pattern. The specification for an optimum sleeve monopole antenna is given in Table 6.2 [11].

A model of a sleeve antenna using coaxial cable is shown in Figure 6.42. The sleeve is attached to ground at $z = 0$, but the drive point for the antenna extends above this point by a distance h. The resulting 2-D vertical plane pattern is shown in Figure 6.43 (the horizontal pattern is still omnidirectional), while the 3-D pattern is shown in Figure 6.44. We can see the familiar figure 8 vertical pattern in both figures. The advantage of the sleeve is in its bandwidth over the basic monopole.

[2] Optimum in this case means that the frequency response is flat with frequency.

Table 6.2 Specification for Optimum Sleeve Monopole

Pattern bandwidth	4:1
l	$\lambda/4$ at low end of band
m/h	2.25
$2R/2a$	3.0
VSWR	Less than 8:1

Source: [11].

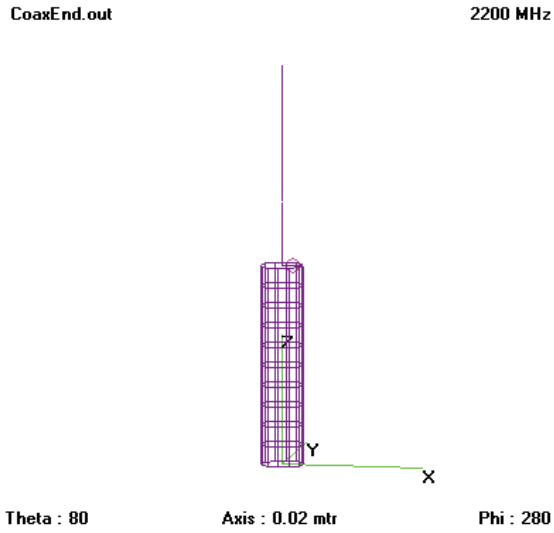


Figure 6.42 Sleeve monopole modeled as a section of coaxial cable. The sheathing on the coaxial cable is attached to ground at $z = 0$, but extends above $z = 0$ at a distance h.

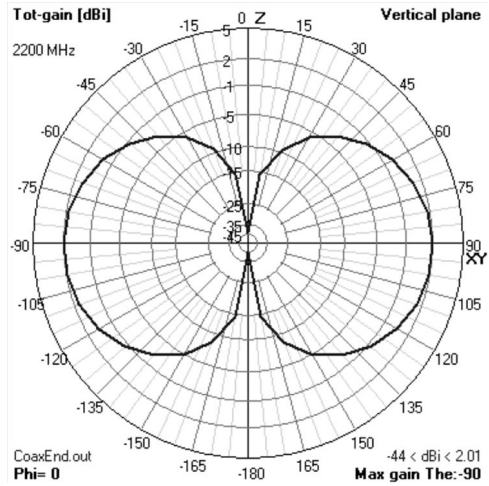


Figure 6.43 Sleeve monopole 2-D vertical plane radiation pattern at f = 2.2 GHz.

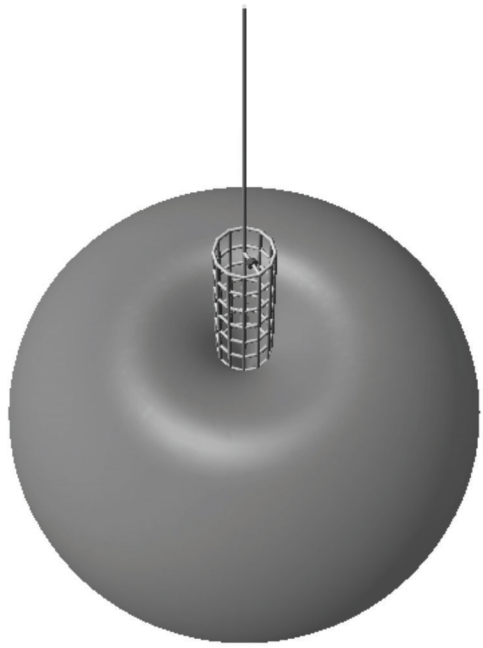


Figure 6.44 Sleeve monopole 3-D radiation pattern at f = 2.2 GHz.

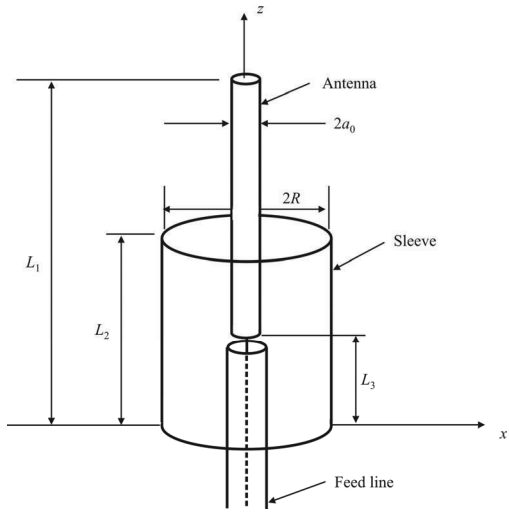


Figure 6.45 Another sleeve monopole.

Li and Gan numerically analyzed the sleeve antenna shown in Figure 6.45 [12]. They evaluated the VSWR performance as some of the physical parameters were varied. The resulting performance is illustrated in Figure 6.46 [12] as the frequency was varied. (The fixed parameters are indicated on Figure 6.46.) Each of the curves corresponds to differing lengths L_2, the height of the sleeve above the ground plane.

From Figure 6.46 we see that each VSWR curve has three frequencies where the VSWR is minimized. When the length of the driven dipole $L_1 = 0.5\lambda$ at f_0, the first frequency, f_1 is approximately equal to f_0, the third frequency, f_3, is slightly above $2.88f_0$ when $L_3 = 0$. As we see, the effects of varying the sleeve length L_2 on

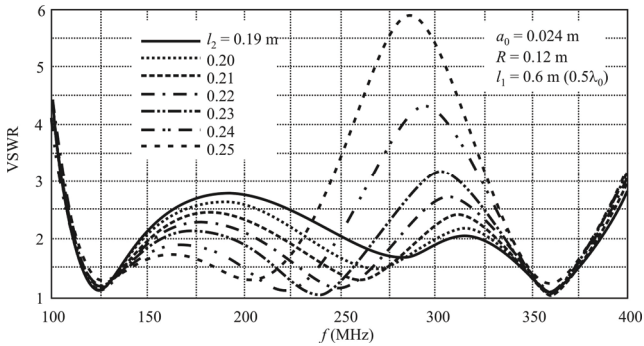


Figure 6.46 Sleeve monopole performance.

f_1 and f_3 are minimal since all the curves have essentially the same value. In addition, the length of the driven dipole has minimal impact on f_2. However, the second null frequency f_2 increases as L_2 gets larger, ranging from about 200 MHz to about 280 MHz for the parameters used in this example. The wavelength at f_2 is related to the length of the sleeve length L_2 and the distance R. Furthermore, in this example, the VSWR remained below 3 for the first four cases between just above 100 MHz to about 400 MHz, a very respectable bandwidth.

When $L_3 \neq 0$, characteristics similar to the case of $L_3 = 0$, that is with three null regions, emerge. The first null region remains near f_0, but the third null region shifts to $2.77f_0$ for L_3 between $0.01\lambda_0$ and $0.02\lambda 0$. The bandwidth is therefore reduced. At the same time, f_2 shifts to lower frequencies.

6.8 Helix Antennas

A helix antenna is constructed by configuring several turns of wire into a coil as shown in Figure 6.47. The normal-mode helix and axial-mode helix antennas are discussed here. Other configurations are possible as well. Each such configuration has its own unique characteristics. For proper operation of a helix antenna, a significant ground plane is required (ideally infinite).

6.8.1 Normal-Mode Helix

In a normal-mode helix, most of the radiation occurs broadside to the axis of the helix. This requires the diameter of the helix to be much smaller than the wavelength. This configuration has similar characteristics to the dipole and monopole. The 2-D pattern for the normal-mode helix shown in Figure 6.48 are shown in Figure 6.49 while the patterns in 3-D are shown in Figure 6.50. The similarity to the monopole is obvious.

6.8.2 Axial-Mode Helix

In an axial-mode helix, the radiation is predominately off the end of the antenna, or along the antenna's axis. The diameter of the antenna must be approximately the same as the wavelength for this to occur. The shape of the antenna is virtually the same as shown in Figure 6.30. This type of antenna is used to create circular polarization in the transmitted wave. This can also be accomplished with crossed dipoles, which are fed in quadrature (90^o out of phase). Circular and elliptical polarizations have favorable characteristics in some circumstances. Elliptical

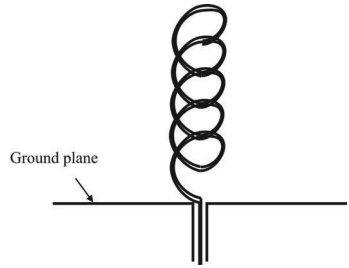


Figure 6.47 Helix antenna.

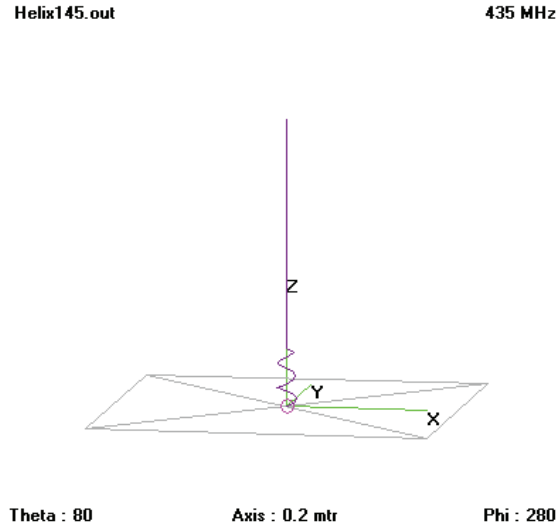


Figure 6.48 Helix 145 normal mode.

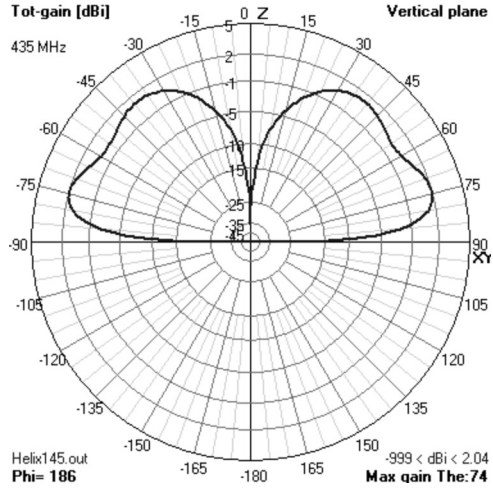


Figure 6.49 Helix normal mode 2D vertical pattern at f = 435 MHz.

Figure 6.50 Helix normal mode 3D radiation pattern at f = 435 MHz.

polarization is produced by utilizing the proper phasing (other than quadrature) between the feed, or by elliptical construction of the antenna.

The gain of the axial-mode helix for lengths of two through seven wavelengths is shown in Figure 6.51 [13], while the helix radius at which this gain is achieved is shown in Figure 6.52 [13]. The gain is relatively insensitive to variations of the other parameters such as wire diameter or the presence of a short stub at the start of the helix. The reference indicates that a 3.5 dB penalty is incurred if the antenna is operated above radials as opposed to an infinite ground plane. Radials are several long (relative to a wavelength) wires spread out radially from the antenna. Also, a half-wavelength square ground plane is as good as an infinite ground plane. These values of gain yield effective areas between $16\lambda^2/4\pi$ and $32\lambda^2/4\pi$.

The radiation patterns for the axial-mode helix shown in Figure 6.53 are shown in Figures 6.54 and 6.55. These characteristics are similar to the patch antenna discussed later in Chapter 15.

6.8.3 Spiral Helix Antennas

A spiral antenna, in particular a log-periodic spiral antenna, is shown in Figure 6.56. As in the wire log-periodic antenna, the spacing between windings is varied as the logarithm of the distance from the end of the antenna. Spiral antennas have wide bandwidths, typically covering 3–4 octaves. The regular (non-log-periodic) spiral antennas are typified by gains of about –6 to +1 dBi where the log periodic shown in Figure 6.56 has a higher gain, on the order of 0 to +6 dBi. These antennas exhibit a circular polarized antenna pattern and can be configured as either right-hand circular or left-hand circular. As shown in Figure 6.56, the beamwidth in either the elevation or azimuth planes is about 80^o.

6.9 Concluding Remarks

We presented the fundamental characteristics of monopole antennas and some of their variants in this chapter. Monopoles, like dipoles, have very useful characteristics in their own right. They are simple to build and therefore inexpensive. They can be made quite rugged for survival in harsh environments. In many applications (in automobiles, for example) the larger power requirements can be accommodated in one or a few transmit antennas and the receive antennas can tolerate the low gain available from a dipole or monopole.

As we will discuss later, dipoles and monopoles are the primary (but not exclusive) building blocks of larger antenna arrays. Such combinations of these

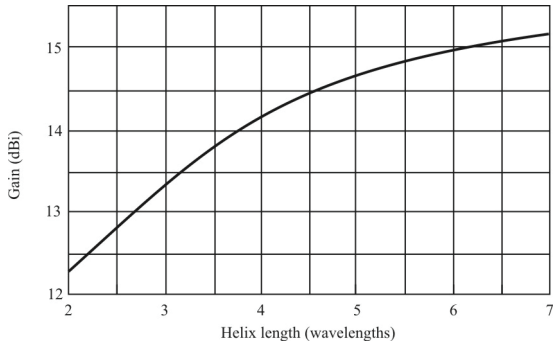


Figure 6.51 Gain variation of an axial-mode helix as a function of the length of the antenna in wavelengths.

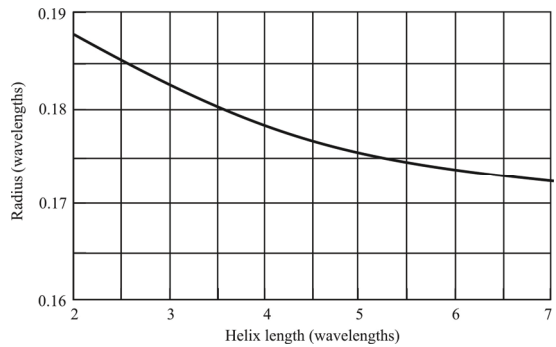


Figure 6.52 Radius of the axial-mode helix at which the gain shown in Figure 6.51 is attained.

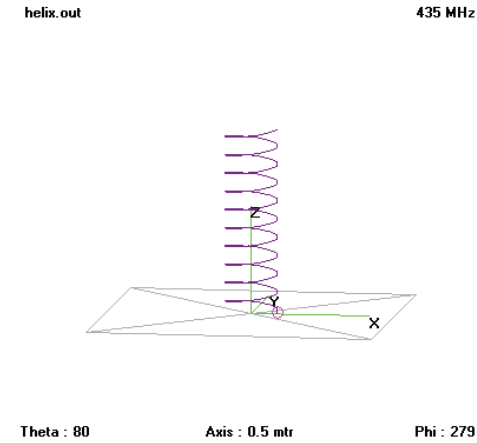


Figure 6.53 Axial mode helix model.

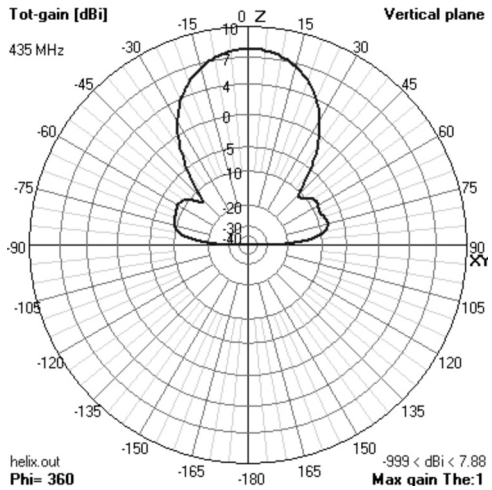


Figure 6.54 Axial mode helix 2-D vertical radiation pattern at f = 435 MHz.

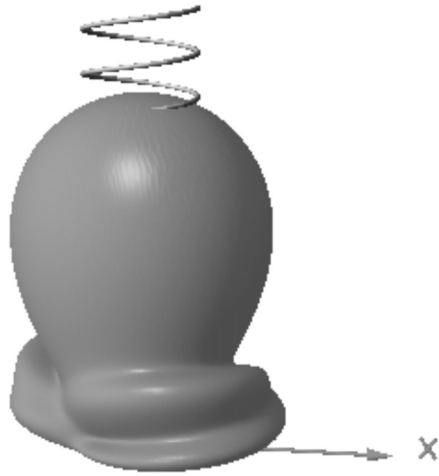


Figure 6.55 Axial mode helix 3-D pattern at f = 435 MHz.

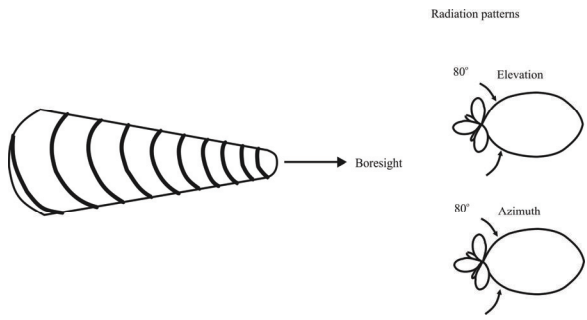


Figure 6.56 Log-periodic helix antenna configuration.

wire antennas have some very desirable characteristics such as high gain. In most EW applications, dipoles and monopoles do not have sufficient gain. In such cases the higher gain available from arrays of these basic elements is highly desirable.

References

[1] Stutzman, W. L., and G. A. Thiele, *Antenna Theory and Design,* New York: John Wiley & Sons, Inc., 1981, pp. 92–96.

[2] Poisel, R. A., *Modern Communications Jamming Principles and Techniques*, 2nd ed., Norwood, MA: Artech House, 2011, Ch. 5.

[3] Jordan, E. C., and K. G. Balman, *Electromagnetic Waves and Radiating Systems*, 2nd ed., Englewood Cliffs, NJ: Prentice-Hall, 1968, p. 548.

[4] Jordan, E. C., and K. G. Balman, *Electromagnetic Waves and Radiating Systems*, 2nd ed., Englewood Cliffs, NJ: Prentice-Hall, 1968, p. 566–567.

[5] *Reference Data for Radio Engineers*, Howard W. Sams, 6th ed., 1976, p. 28-3.

[6] SteppIR. at http://ac6v.com/antprojects.htm.

[7] King, R., et al., "Transmission-Line Missile Antennas," *IRE Transactions on Antennas and Propagation.*, January 1960, pp. 88–90).

[8] Nail, J. J., "Designing Discone Antennas," *Electronics*, Vol. 26, August 1957, pp. 167–169.

[9] Li, J.-Y., and Y.-B. Gan, "Multi-Band Characteristic of Open Sleeve Antenna," *Progress in Electromagnetics Research, PIER* 68, 2006, pp. 135–149.

[10] Poggio, A. J., and P. E. Mayes, "Pattern Bandwidth Optimization of the Sleeve Monopole Antenna," *IEEE Transactions on Antennas and Propagation*, September 1966, pp. 643–645.

[11] Stutzman, W. L., and G. A. Thiele, *Antenna Theory and Design,* New York: Wiley, 1981, p. 279.

[12] Li, J.-Y, and Y.-B. Gan, "The Characteristic of Sleeve Antenna," *Journal Progress in Electromagnetic Research*, Hangzhou, August 23–26, 2005, p. 1.

[13] Emerson, D. T., *ARRL Antenna Compendium*, Vol. 4, 1995, pp. 64–68, http://www.tuc.nrao.edu/~demerson/helixgain/helix.htm.

Chapter 7

Loop Antennas

7.1 Introduction

Loop antennas form an antenna type that is distinct from other wire antennas such as dipoles and monopoles, and that features simplicity, low cost, and versatility. Loop antennas have various shapes: circular, triangular, square, elliptical, and so forth, but the far-field radiation patterns are similar irrespective of the shape of the loop. They are widely used in applications up to the microwave bands (up to ~3 GHz).

A magnetic loop antenna is a winding of copper wire around a frame for air-core loops or around ferromagnetic material for ferrite loops. A loop antenna is actually sensitive to the magnetic field and not the electric field. It outputs a voltage proportional to that field. As we will show, a particularity of this type of antenna is to provide a voltage proportional to the frequency of the signal. The antenna performance is influenced by the number of turns and the area of each loop. For a ferrite antenna, the permeability of the core increases the output voltage. Also, loop antennas have a toroidal reception pattern whose amplitude is maximal in the plane of the loop and zero (in theory) in the plane normal to the loop.

Loop antennas are popular for receiving applications because they have similar desirable construction characteristics as dipoles and monopoles. Just as the electrical length of the dipoles and monopoles affect the efficiency of these antennas, the area of the loop determines the efficiency of the loop antenna.

Loop antennas are usually classified as either electrically small or electrically large based on the circumference of the loop. The terms refer to their size when compared to the wavelength of the frequency in use. A *small* loop is a loop of any configuration and with a constant current whose radius (diagonal if a square) satisfies

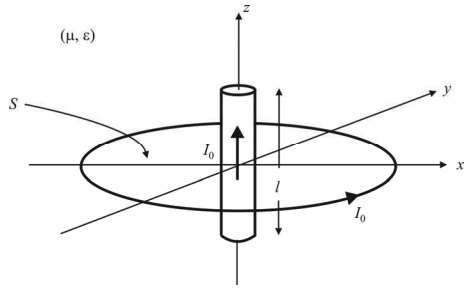


Figure 7.1 Small loop and small dipole. When oriented in the way shown, $E_{loop} = H_{dipole}$ and $H_{loop} = E_{dipole}$.

$$b < \frac{\lambda}{6\pi} \tag{7.1}$$

or whose circumference C satisfies $C < \lambda/3$. A very good approximation of a small loop is the infinitesimal loop (or the infinitesimal magnetic dipole), the characteristics for which were discussed in Chapter 3. An electrically large loop has a circumference that is on the order of a wavelength:

- Electrically large loops–circumference $\approx \lambda$

The electrically small loop antenna is the dual antenna to the electrically short dipole antenna when oriented as shown in Figure 7.1. The far-field electric field of a small loop antenna is identical to the far-field magnetic field of the short dipole antenna and the far-field magnetic field of a small loop antenna is identical to the far-field electric field of the short dipole antenna.

Analysis of loop antennas is divided into two distinct paths. We examined the infinitesimal loop in Chapter 3 and came up with its fundamental parameters. One of the assumptions in that analysis was that the current everywhere on the loop was a constant value. That is the same assumption used for the analysis of the infinitesimal dipole in Chapter 3. The other categorization is when this assumption is not true—that is, for large loop antennas. For dipoles there was an intermediary point that we do not have for loop antennas: a short dipole has a current distribution that has a triangular shape. Both large loop and dipole antennas exhibit non-constant values of current. The analysis of large loops, however, is considerably more difficult than for dipoles. We will concentrate most of our discussion on small loops.

It is because loop antennas assume a constant current at every point around the antenna that loop antennas are rarely used above the low VHF region. The constant current assumption is necessary to yield reasonable analytical results.

This chapter begins with a discussion of the characteristics of small loop antennas, particularly in receiving applications. That is followed by presentation of material on the general characteristics of these antennas. A section of ferrite rod loop antennas is then presented. Adding ferrites to a loop antenna is a good way to increase their radiation resistance and therefore their efficiency. Then we present an equivalent circuit for loop antennas. The last topic covered is a brief discussion of large antennas.

7.2 Small Loop Antennas

Small loop antennas are similar to coils because they have the same current distribution as ordinary "circuit" coils, having the same phase and amplitude through the whole coil. To satisfy this, the total length of the conductor used in the loop antenna must be no more than about 0.1 wavelength long. Any longer than this and the current phase and amplitude will start to vary over the length of the conductor and some of the properties start to change.

Small loop antennas may also be further split into those that use a single turn, and those that have a multiturn loop, as in the case of a coil. Electrically small loops of a single turn have very small radiation resistance comparable to their loss resistance. Their radiation resistance, though, can be substantially improved by adding more turns. Multiturn loops have better radiation resistance but their efficiency is still very poor. Because of this, they are used predominantly as receiving antennas, where losses are not so important. The radiation characteristics of a small loop antenna can be additionally improved by inserting a ferromagnetic core. We discuss the effects of such cores below.

Multiturn loop antennas are not normally used for transmitting because the losses are high and the level of heat dissipated can give rise to rapid temperature increases. Instead, single turn loop antennas may be used if a loop antenna is needed. These antennas have a number of advantages and disadvantages.

The main advantages of loop antennas are their size and directivity. Often a single turn small loop antenna is much smaller than a wavelength. They are also quite directive, and this can be used to direct the radiated power in the desired direction. They find uses for transmitting and receiving, particularly on the MF and HF bands. Here they provide very compact antennas for applications such as logistics (RFID), as well as receiving antennas for MF or medium wave receivers.

There are disadvantages to these antennas as well. The low radiation resistance results in very high levels of current flowing in the antenna. In turn, this means that even small levels of DC resistance can result in significant amounts of power being lost as heat. It is for this reason that single turn small loop antennas are made of very thick wire, or more often made of a tubular conductor due to the

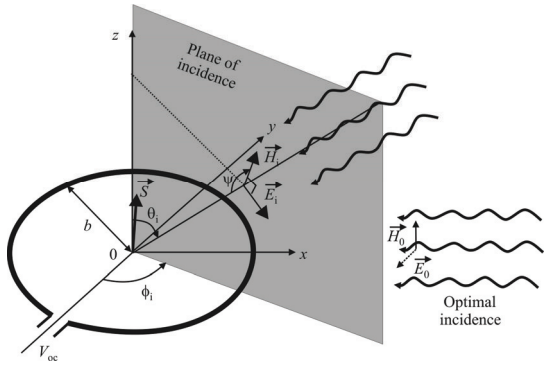


Figure 7.2 Loop receiving antenna.

skin effect. Additionally, this means that they must have an effective form of antenna matching if the energy is to be efficiently transferred to and from the antenna.

A further disadvantage of these antennas is that they typically have a very high Q. Not only does the antenna require tuning to bring it to resonance at the operating frequency, but it also may have such a narrow bandwidth in the MF range or even a little higher, that it may be insufficient to accommodate the carrier and its sidebands.

7.3 General Characteristics of Loop Antennas

Given that the radiated fields of the short dipole and small loop antennas are duals, the radiation patterns are the same. This means that the plane of maximum radiation for the loop is in the plane of the loop (*x*-*y* plane in Figure 7.2). When operated as a receiving antenna, the short dipole must be oriented such that the electric field is parallel to the wire for maximum response. Using duality, the small loop must be oriented such that the magnetic field is perpendicular to the loop for maximum response. This relationship is illustrated in Figure 7.3.

The radiation from a small loop is away from the center of the loop in the direction of the wire making up the loop. See Figure 7.3 where the direction of maximum propagation is shown as well as the directions in which no radiation occurs (theoretically). This can also be seen from a plot of the 2-D radiation patterns shown in Figure 7.4 where the axis definitions correspond to Figure 7.2.

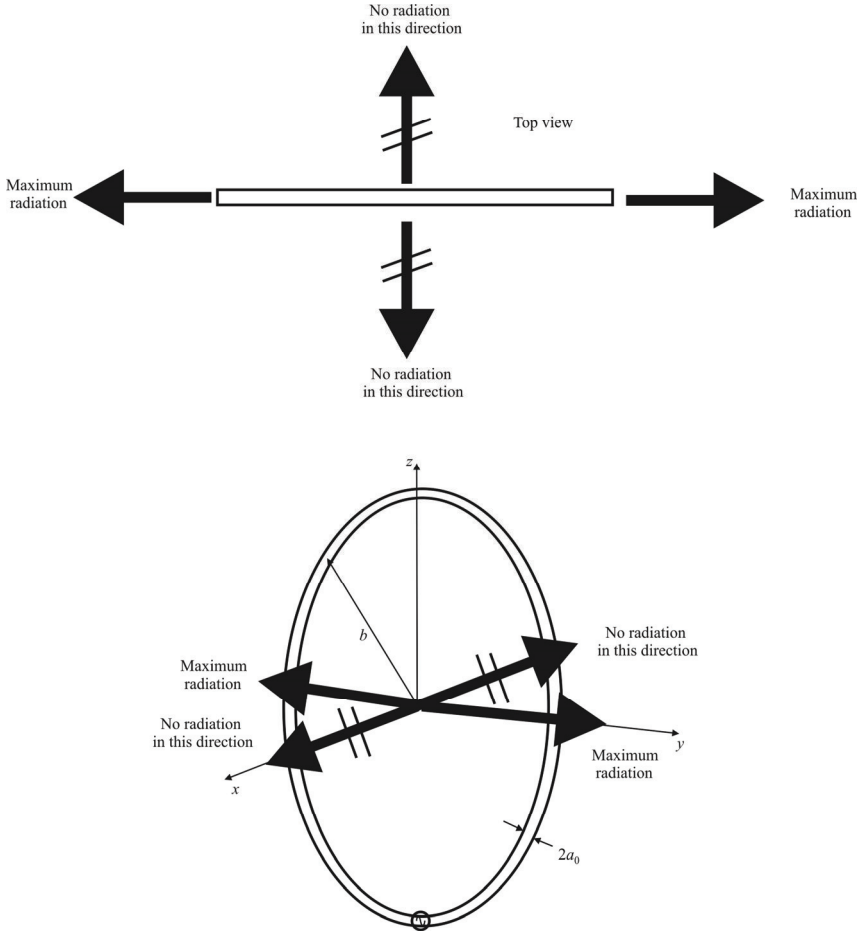


Figure 7.3 Maximum radiation from a small loop antenna is off the sides of the loop, with (theoretically) no radiation in the directions orthogonal to the plane of the loop.

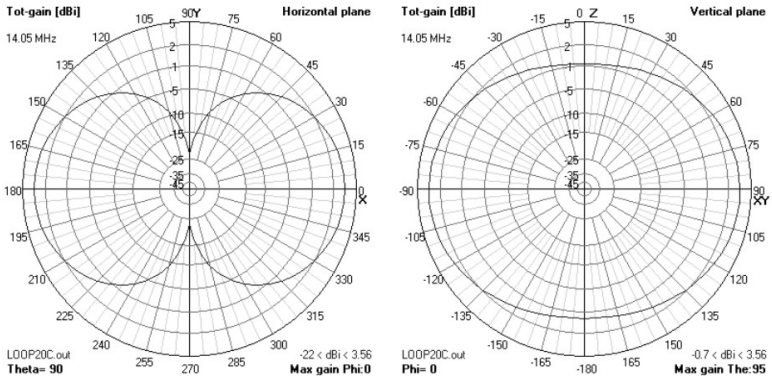


Figure 7.4 Small loop 2-D radiation patterns in free space.

The 3D radiation pattern is shown in Figure 7.5; the antenna for this diagram is in the (*y*-*z*) plane.

Even though it has significant loss resistance compared to its radiation resistance, the small loop antenna will function reasonably well as a receive antenna since the SNR is the driving factor for receiving systems, not antenna efficiency. Even though a considerable portion of the received signal is lost to heat, it is not significant as long as the antenna provides a large enough SNR for the receiver.

7.3.1 The Induced Voltage as a Function of the Magnetic Field

From (2.26), Faraday's law of induction states that the induced electromotive force $e(t)$ in a loop is directly proportional to the time rate of change of magnetic flux $\Psi(t)$ through the loop according to the relation:

$$e(t) = -\frac{d\Psi(t)}{dt} \tag{7.2}$$

where:

$e(t)$ is the induced electromotive force, V
$\Psi(t)$ is the magnetic flux across the circuit, webers (Wb ≡ V·s)

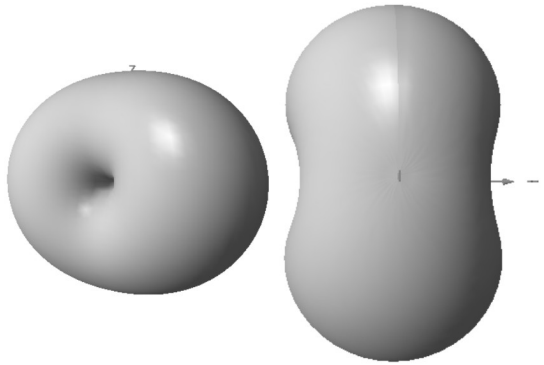


Figure 7.5 Small loop 3-D radiation pattern in free space. The loop is in the *x*-*z* plane. Left: side view, right: top view.

(Recall that $\vec{M}$ does not exist.) Equation (7.2) is valid only for electrically short antennas with respect to the wavelength of interest, which is true for many of our uses.

The inductive magnetic field across the loop depends on the magnetic component, $\vec{H}$, of the EM wave and on the magnetic permeability of the loop core. From (2.20), we know that the inductive magnetic field $\vec{B}$ relates to the magnetic field strength $\vec{H}$ by

$$\vec{B} = \mu_0 \mu_r \vec{H} \tag{7.3}$$

or, in RMS terms,

$$B_{\text{RMS}} = \mu_0 \mu_r H_{\text{RMS}} \tag{7.4}$$

where:

B_{RMS} is the RMS value of the magnetic induction, tesla (T ≡ V s m^{-2}).
μ_0 is the permeability of vacuum, constant of value $4\pi \times 10^{-7}$ h m^{-1}.
μ_r is the relative permeability, dimensionless quantity. μ_r is specific to the medium.
H_{RMS} is the RMS value of the magnetic field strength, A m^{-1}.

The magnetic flux is a measure of quantity of magnetism through the antenna submitted to a given magnetic induction and is given by

$$\Psi(t) = \vec{B}(t) \cdot \vec{S} = \left|\vec{B}(t)\right|\left|\vec{S}\right| \cos\psi \tag{7.5}$$

where

Ψ is the magnetic flux across the circuit, Wb;
$\vec{B}$ is the vector of the magnetic induction;
$\vec{S}$ is the normal vector to the loop surface;
ψ is the angle between the magnetic field lines and the loop normal.

For a sinusoidal magnetic field, the modulus of $\vec{B}$ is

$$\left|\vec{B}\right| = B_0 \cos(\omega t) \tag{7.6}$$

where

B_0 is the strength of the magnetic induction, teslas (T ≡ Wb m^{-1} ≡ V·s m^{-1}).
ω is the angular frequency of the inductive magnetic field, rad·s^{-1} .

For a loop with N turns, each of them of area A, the modulus of $\vec{S}$ is

$$\left|\vec{S}\right| = NA \tag{7.7}$$

Equations (7.2) and (7.5) produce

$$\Psi(t) = NAB_0 \cos(\omega t)\cos\psi \tag{7.8}$$

$$e(t) = NAB_0\omega \sin(\omega t)\cos\psi \tag{7.9}$$

The RMS value of the voltage at the output of the antenna is

$$V_{\text{RMS}} = \omega NAB_{\text{RMS}} \cos\psi \sin\theta_i \tag{7.10}$$

where (θ_i, ϕ_i) are the angles specifying the direction of incidence. For an air-core loop, we have μ_r = 1. For a ferrite loop, the field lines are collected by the ferromagnetic properties of the core, and μ_r can be several orders of magnitude

larger. The antenna will then output an RMS voltage V_{RMS} from a given RMS magnetic field strength H_{RMS} value according to

$$V_{RMS} = \omega\mu_0\mu_r NAH_{RMS}\cos\psi\sin\theta_i \tag{7.11}$$

and we see that the voltage at the antenna terminals is proportional to frequency.

A loop antenna is also resonant at integral multiples of its resonant frequency. The self-impedance of a $\lambda/2$ dipole at these multiples of the resonant frequency is 200–300 Ω. The directivity is lower on harmonic frequencies. Vertically oriented loops will have high angles of radiation on harmonic frequencies. Horizontally oriented loops will have lower angles of radiation on harmonic frequencies.

Thus, the small loop antennas have the following characteristics:

- High radiation resistance provided multiturn ferrite-core constructions are used.
- High losses, therefore, low radiation efficiency.
- Simple construction, small size, and low weight.

7.3.2 The Induced Voltage as a Function of the Electric Field

From (2.75), (7.4), and (7.11) the effective height of a loop antenna is

$$h_e = \frac{2\pi NA\cos\psi\sin\theta_i}{\lambda} \tag{7.12}$$

where:

h_e is the effective height, m
N is the number of turns
A is the area of each winding, m²
λ is the wavelength, m
ψ is the angle between the magnetic field lines and the loop normal

From (3.89), the RMS voltage induced at the antenna terminals is thus

$$V_{RMS} = \frac{2\pi NA\cos\psi}{\lambda} E_{RMS} \tag{7.13}$$

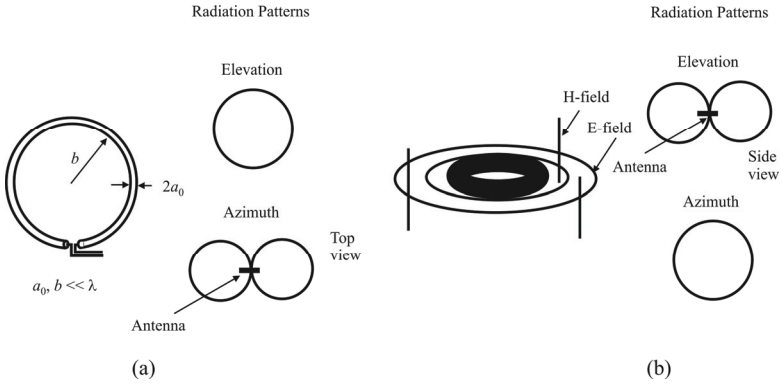


Figure 7.6 Loop antenna patterns: (a) vertical configuration and (b) horizontal configuration.

With a ferrite antenna, the magnetic induction through the antenna is increased, with the relative permeability of the medium, μ_r; then (7.12) becomes

$$h_e = \frac{2\pi NA\mu_r \cos\psi}{\lambda} \tag{7.14}$$

and (7.13) becomes adjusted accordingly.

7.3.3 Radiation Patterns

A loop exhibits similar radiation patterns as the dipole but does not have the gain. In fact a loop exhibits a 2 dB loss relative to an isotropic antenna. Loops need not be circular, but can take on arbitrary shapes. Two orientations of circular loop antennas are shown in Figure 7.6. For the horizontal loop the shape of the E-field forms circles in the horizontal plane, that is, in the same plane as the antenna. The magnetic fields lie in vertical planes for this configuration. The 3-D radiation patterns shown in Figure 7.5 correspond to the orientation shown in Figure 7.6(a).

The effective area of loop antennas is approximately $A_{eff} = 0.63\lambda^2/4\pi$.

7.4 Ferrite Rod Loop Antenna

The radiation resistance of the small loop can be increased significantly by adding multiple turns ($R_r \sim N^2$). However, the addition of more conductor length also increases the antenna loss resistance which reduces the overall antenna efficiency.

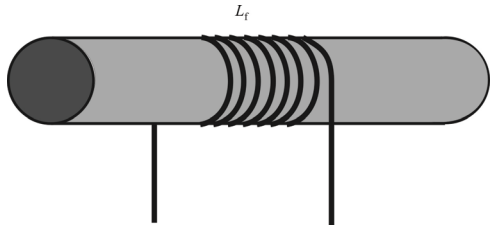


Figure 7.7 Ferrite rod antenna.

To increase the radiation resistance without significantly reducing the antenna efficiency, the number of turns can be decreased when a ferrite material is used as the core of the winding, which has high magnetic permeability in the operating frequency band. Such an antenna is commonly used as a low-frequency receiving antenna.

The ferrite rod antenna consists of a rod made of ferrite, an iron-based magnetic material, with a coil wound around the ferrite rod (see Figure 7.7). This is brought to resonance using a variable tuning capacitor as illustrated in Figure 7.8. As the antenna is tuned, it usually forms the RF tuning circuit for the receiver, enabling both functions to be performed by the same components, thereby reducing the number of components and hence the cost.

Since the antenna senses the magnetic component of the signals, the antenna is directive. It operates best when the magnetic lines of flux are in line with the antenna, as illustrated in Figure 7.9. This occurs when it is at right angles to the direction of the transmitter. Thus, the antenna has a null position where the signal level is at a minimum when the antenna is in line with the direction of the transmitter.

These antennas are very convenient for portable applications, but their efficiency is much less than that of a larger antenna. The performance of the ferrite also limits their frequency response. Normally they are only effective on the long

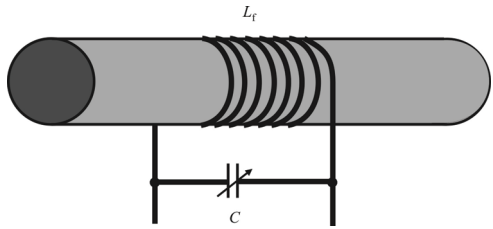


Figure 7.8 Tuned ferrite rod antenna.

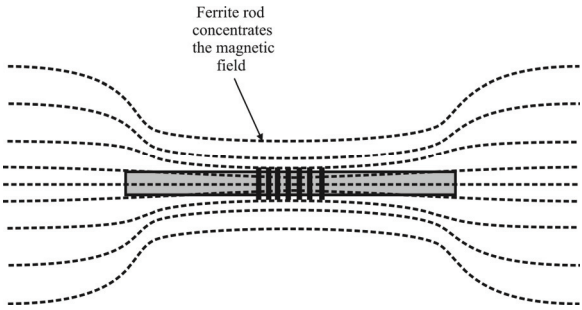


Figure 7.9 Concentration of magnetic flux lines through a ferrite core.

and medium wave bands, but they are sometimes used for lower frequencies in the short wave bands although their performance is significantly degraded, mainly arising from the losses in the ferrite. This limits their operation normally to frequencies of only a few megahertz.

Early HF direction-finding systems used this type of antenna to measure the direction to emitters. The high sensitivity to angle was used to manually (typically) turn the antenna and the direction of the null indicated the direction to the target. In some cases, these antennas were mechanically spun to display the same information on a cathode ray tube display. Other configurations are more popular today using more sensitive dipole and monopole antennas.

7.4.1 Ferrite Rod *Q*

One of the requirements for an efficient ferrite rod antenna is that it should have a high Q over the frequencies where it is expected to operate. At frequencies of a few hundred kilohertz, a medium permeability material would be used, and this would enable a Q on the order of about 10^3 to be obtained. With a Q this large the antenna needs to be tuned if it is to operate over more than a narrow frequency range.

7.4.2 Radiation Resistance

Perhaps the biggest advantages of using a ferrite in the antenna is that it brings the radiation resistance of the overall antenna to a more reasonable level. Because of its small size, the loop is much less than a wavelength in length and without the ferrite it would have a very low radiation resistance. Accordingly the losses due to the resistance of the wire would be respectively high. Placing the ferrite core in the coil has the effect of raising the radiation resistance by a factor of μ^2.

The ferrite rod raises the radiation resistance of the antenna and hence reduces the losses due to the resistance of the wire. However, it introduces other losses because the ferrite itself absorbs power. This arises from the energy required to change the magnetic alignment of the magnetic domains inside the granular structure of the ferrite. The higher the frequency, the greater the number of changes and hence the higher the loss.

A large magnetic permeability $\mu = \mu_0\mu_r$ means large magnetic flux Ψ_m and therefore large induced voltage V_{oc}. The radiation resistance of a small loop was established in Chapter 3 (3.195), and it was shown that it increases as $\sim N^2$. The magnetic properties of a ferrite core depend not only on the relative magnetic permeability μ_r of the material of which it is made, but also on its geometry. The increase in the magnetic flux is then more realistically represented by the *effective relative permeability* μ_{reff}. When we calculated the far fields of a small loop, we made use of the equivalence between an electric current loop and a magnetic current element given by

$$j\omega\mu(IS) = I_m l \tag{7.15}$$

From (7.15), palpably the equivalent magnetic current is proportional to μ. Also the field magnitudes are proportional to I_m, and therefore they are proportional to μ as well. This means that the radiated power, Π_{rad}, is proportional to μ^2, and therefore the radiation resistance increases as $\sim\mu_{eff}^2$.

Finally, we can express the radiation resistance as

$$R_r = \eta_0 \frac{8}{3}\pi^3 \left(N\mu_{eff} \frac{S}{\lambda^2} \right)^2 \tag{7.16}$$

Here $S = \pi b^2$ is the loop area, and η_0 is the intrinsic impedance of a vacuum.

Toroidal cores, one of which is illustrated in Figure 7.10, have the highest μ_{reff}, and ferrite-stick cores, diagrammed in Figure 7.7, have the lowest. The effective magnetic constant is frequency dependent as the magnetic losses of ferromagnetic materials increase with frequency. At microwave frequencies, the magnetic losses are not negligible. They have to be calculated and represented in the equivalent circuit of the antenna as a shunt conductance G_m.

The ferrite has the effect of intensifying the magnetic field inside the loop as illustrated in Figure 7.9. This concentration is produced by the high permeability, μ, of the ferrite material. Usually, it is convenient to use a rod of ferrite material and wind a coil around a central part, as illustrated in Figure 7.7. This increases the loop's radiation resistance by a factor of μ_{eff}^2 to

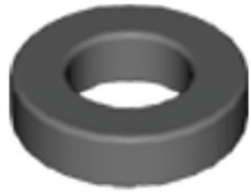


Figure 7.10 Example of a toroidal core.

$$R_r \approx 31,200\left(\frac{\mu_{eff} NS}{\lambda^2}\right)^2 \tag{7.17}$$

For frequencies of a few hundred kilohertz, we can obtain ferrites that provide μ_{eff} values in the range from around 10^2 to around 10^4. So if $\mu_e = 10^3$, we can see that using the ferrite can increase the antenna's radiation resistance by a factor of a 10^6! Hence, the ferrite can have a dramatic effect in improving the antenna's efficiency.

As mentioned, the ferrite itself tends to absorb some of the signal power. For a ferrite rod this loss has an equivalent resistance expressed as

$$R_f \approx 2\pi f \mu_{eff} \frac{\mu''}{\mu'} \mu_0 N^2 \frac{S_f}{l} \tag{7.18}$$

where μ'' is the imaginary (loss) part, and μ' the real part of the ferrite's permeability, S_f is the cross-sectional area of the rod, and l is the length of the rod. We must now add this to the wire's resistive losses to obtain the overall level of loss resistance in the antenna. Fortunately, by choosing a suitable material this increase in loss can be quite small compared to the increase in R_r.

Hence, overall, the ferrite significantly improves the antenna's performance. When viewed from the connecting wires, we find that, even when using a high μ ferrite, the antenna's resistance value is often just a few ohms (or even much less than 1Ω) in series with a significant inductance. This combination of a low resistance with a large inductance can make it difficult to match the antenna as a source or load to the receiver or transmitter electronics. To deal with this problem a capacitor is used to turn the loop into a resonant circuit/antenna as shown in Figure 7.9. The equivalent circuit of the ferrite antenna with this capacitor is shown in Figure 7.11.

The inductance of this antenna is

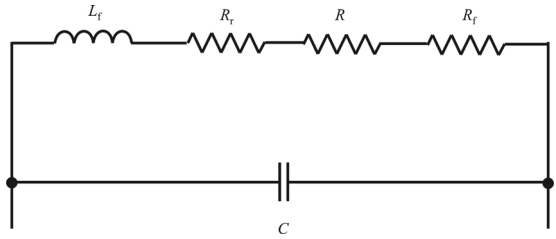


Figure 7.11 Tuned ferrite rod antenna equivalent circuit.

$$L_f = \mu_{eff}\mu_0 N^2 \frac{A_f}{l} \tag{7.19}$$

and by using a suitable parallel capacitance, C, we can convert the antenna's terminal impedance, at the resonant frequency

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{1}{L_f C}} \tag{7.20}$$

into a pure resistance whose magnitude is Q^2 larger than the actual loop resistance, where Q is the circuit's quality factor

$$Q = \frac{2\pi f_0 L_f}{R + R_r + R_f} = \frac{f_0}{\Delta f_{hp}} \tag{7.21}$$

where Δf_{hp} is the half-power half-bandwidth of the resulting resonance. The resulting equivalent circuit at resonance is shown in Figure 7.12.

For signal frequencies up to a few megahertz the tuned ferrite rod antenna can provide antenna efficiencies (and hence gains or effective areas) that can be between 10^3 and 10^6 times better than a Hertzian dipole of similar size. For this reason they are often preferred for small portable devices at the lower-frequency ranges. The high Q characteristic of the antenna can sometimes also help filter out unwanted signals.

The main disadvantages of the antenna are:

- The Q may be so high (i.e., Δf_{hp} so small) that the antenna filters away some of the desired signal, thereby reducing the SNR. (A Q of 10^5 at 10 MHz means that the bandwidth is approximately 100 Hz.)

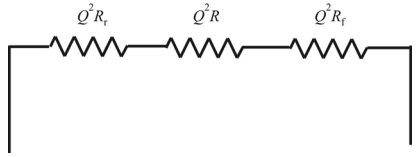


Figure 7.12 Tuned ferrite rod antenna equivalent circuit.

- The dissipation losses in the ferrite make the system unsuitable as a transmit antenna.
- The μ_{eff} value is only larger than unity for small magnetic field levels.

Losses in the ferrite mean that if we try using the antenna in a transmit mode the power dissipated may heat up the material until it decomposes or melts. μ_{eff} tends to fall to unity when a large field is applied so the antenna is unusable for all but the smallest transmit power levels (e.g., it may work for RF identification [RFID] devices). For these reasons the ferrite rod makes an excellent low-frequency receiving antenna, but is not used for signal transmission.

7.5 Loop Antenna Equivalent Circuit

The equivalent circuit of a loop antenna in the transmit mode is illustrated in Figure 7.13. In this figure,

- C_r is the resonance capacitor.
- R_{loss} is the loss resistance of the antenna = $R_{DC} + R_{HF}$.
- R_r is the radiation resistance.
- L_{loop} is the inductance of the loop.
- L_{wire} is the inductance of the loop conductor (wire).
- R_{DC} is the wire resistance.
- R_{HF} is the skin effect and proximity effect resistance = $R_{skin} + R_{proximity}$.
- C_{loop} is the distributed capacitance.

When the antenna is in the receiving mode, there is a lossless voltage generator included in series as shown in Figure 7.13, as the V_{loop}.

7.5.1 Distributed Capacitance

A more complete model would include a distributed capacitance, C_{loop}. This includes the effects of some level of capacitance between the loops. It, however, is

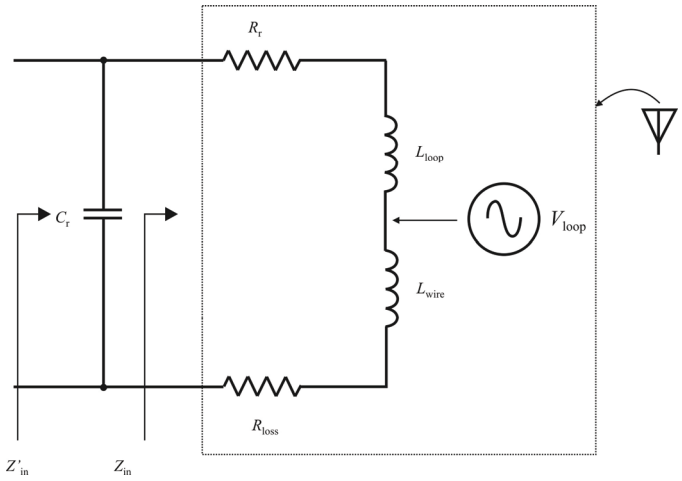


Figure 7.13 Equivalent circuit of a small loop antenna in the transmit mode. In the receive mode the lossless voltage generator is added in series.

normally very small relative to the other parameters and therefore is normally neglected.

7.5.2 Radiation Resistance

The radiation resistance corresponds to the "losses" in the antenna during the transformation of the EM energy into electric energy (if receiving) and electrical energy into EM energy (if transmitting). Unlike an ohmic resistance that converts electrical energy to heat, this resistance is virtual and, consequently, does not generate thermal noise according to the Johnson-Nyquist formula. Its value is given in (3.194), repeated here for convenience

$$R_r = \eta_0 \frac{2}{3}\pi\left(\frac{h_e}{\lambda}\right)^2 = \eta_0 \frac{8}{3}\pi^3\left(\frac{N\mu_r S}{\lambda^2}\right)^2 \tag{7.22}$$

where:

R_r is the radiation resistance, Ω.
η_0 is the impedance of free space (120π Ω, about 377 Ω).
μ_r is the relative permeability of the core.
N is the number of turns.
S is the area of each winding, m².

λ is the wavelength, m.

Since $\eta = \sqrt{\mu_0 / \varepsilon_0} = \mu_0 c = 4\pi c \times 10^{-7}$, (7.22) becomes:

$$R_r = \frac{32}{3c^3}\pi^4 \left(N\mu_r S\right)^2 f^4 \quad \Omega \tag{7.23}$$

7.5.3 Wire Inductance L_{wire}

Square Loop

Consider a square loop of N windings. The wire of the antenna windings itself has an inductance. A wire of length $4Nw$ has an inductance L_{wire} according to

$$L_{wire} \approx \frac{4Nw\mu_0}{2\pi}\left[2.303\log_{10}\left(\frac{16Nw}{2a_0}\right) - 1 + \frac{\mu_r}{4} + \frac{2a_0}{8Nw}\right] \tag{7.24}$$

$$\approx 2\times 10^{-7} \times 4Nw\left[2.303\log\left(\frac{16Nw}{2a_0}\right) - 0.75 + \frac{2a_0}{8Nw}\right] \tag{7.25}$$

where:

L_{wire} is the wire inductance, h.
w is the frame side length, m. Total wire length is $4Nw$.
a_0 is the wire radius, m.

Circular Loop

Now consider a circular loop as in Figure 7.5, with radius b, wire radius a_0, and N windings. The inductance is approximately

$$L \approx N^2 \mu b\left[\ln\left(\frac{8b}{a_0}\right) - 1.75\right] \tag{7.26}$$

7.5.4 Loss Due to Resistance

The actual resistive loss in the antenna is comprised of three parts: (1) wire resistance, R_{DC}; (2) skin depth effect resistance, R_{skin}; and (3) proximity effect

resistance, $R_{\text{proximity}}$. R_{DC} is the DC resistance of the wire and is usually quite small for good conductors. R_{skin} and $R_{\text{proximity}}$ are AC resistances that increase in value as the frequency increases.

Wire Resistance

Electric wire has a resistance that is a function of its length and diameter.

Square Loop Antenna

The resistance of a wire of length $4Nw$ and of radius a_0 is given by:

$$R_{\text{DC}} = \frac{(4Nw)\rho}{\pi(2a_0)^2/4} \tag{7.27}$$

where:

R_{DC} is the wire resistance, Ω.
N is the number of turns.
w is the frame side length, m. Total wire length is $4Nw$.
$2a_0$ is the wire diameter, m. Wire cross section is $\pi(2a_0)^2/7$.
ρ is the conductor resistivity, Ω m. For copper, ρ = 17.78 nΩ m.

Circular Loop Antenna

The wire resistance of a circular loop antenna is approximated by the resistance of a length of wire that has the same total length of the wire in the antenna. Thus, if the radius is a_0 and the total length of the wire in the antenna is l,

$$R_{\text{DC}} = \frac{1}{\sigma}\frac{l}{\pi a_0^2} \tag{7.28}$$

where σ is the conductivity of the metal (for copper at 25°C, $\sigma = 59\times10^6\ \text{S m}^{-1}$).

Nyquist Noise

Both of the DC resistances given in (7.27) and (7.28) are real resistances and therefore generate thermal white noise according to the Johnson-Nyquist formula that expresses the voltage spectral density of the thermal noise given by

$$\overline{v}_{\text{noise}} = \sqrt{4k_B T R_{DC}}, \qquad \text{V Hz}^{-1/2} \tag{7.29}$$

k_B is the Boltzmann constant (= 1.38×10^{-23} J K^{-1}) and T is the absolute temperature in Kelvins. Thermal noise generated in the antenna is generally to be avoided for sensitivity reasons, especially for receiving applications (it raises the noise figure of the system) [1].

Skin Effect Resistance

As we discussed in Chapter 3, in addition to the DC resistance described above, there is an additional high frequency resistance resulting from the fact that the current distribution is no longer homogeneous at high frequencies. This is called the skin effect. The skin effect resistance is given by

$$R_{\text{skin}} = \frac{4Nw}{\pi 2a_0}\sqrt{\pi \mu_0 f \rho} \tag{7.30}$$

where:

R_{skin} is the skin effect resistance, Ω.
N is the number of turns.
w is the frame side length, m. Total wire length $4Nw$.
a_0 is the wire radius, m. Wire perimeter is $2\pi a_0$.
μ_0 is the permeability of free space ($4\pi \times 10^{-7}$ H m^{-1}).
f is the frequency, Hz.
ρ is the copper resistivity, Ω m. ρ = 17.78 nΩ·m.

Proximity Effect Resistance

Another important phenomenon has to be taken into account with multiturn loop antennas: the *proximity effect*. The proximity effect in loop antennas is illustrated in Figure 7.14. When two conductors are spaced close together and carrying current in the same direction, the area available to carry current is reduced, thereby increasing the resistance of the wires (denoted as $R_{\text{proximity}}$). (There is a proximity effect in PCB layout as well, where the two current densities shown in Figure 7.14 are traveling in different directions. In that case, the two conduction regions are attracted to each other. The net effect is the same—a reduction in the conduction cross section and an increase in resistance. We will not concern ourselves further with that problem here.) When the spacing between the turns is very small, the

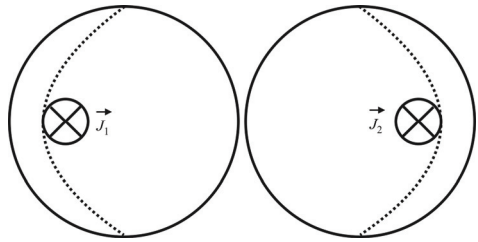


Figure 7.14 Proximity effect decreases the cross section of two parallel wires carrying current in the same direction such as in a multiloop antenna (into the page in this case). The decreased cross section increases the resistance of the wires.

loss resistance due to the proximity effect can be larger than that due to the skin effect.

For a loop antenna with N turns, wire radius b, and loop separation $2c$ (see Figure 7.15), the loss resistance is

$$R_l = \frac{Na_0}{b} R_s \left(\frac{R_p}{R_0} + 1 \right) \tag{7.31}$$

where

R_s is the surface resistance (12.32), Ω.

R_p is the ohmic resistance per unit length due to the proximity effect, Ω m^{-1}.

$R_0 = NR_s / (2\pi b)$ is the ohmic resistance per unit length due to the skin effect, Ω m^{-1}.

The ratio R_p / R_0 has been calculated for different relative spacings c/a_0, for loops with $2 \leq N \leq 8$. The results are shown in Figure 7.16 [2].

Total Resistance

It is normally assumed that the loss resistance of the loop equals the high-frequency loss resistance of a straight wire of the same length as the loop and of the same current distribution. In the case of a uniform current distribution, the high-frequency resistance is

$$R_{HF} = \frac{l}{2\pi a_0} R_s \qquad R_S = \sqrt{\frac{\pi f \mu}{\sigma}} \quad \Omega \tag{7.32}$$

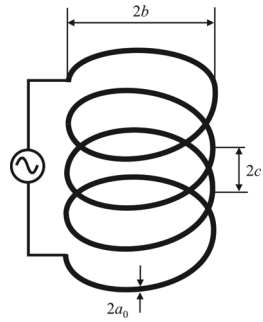


Figure 7.15 The loop dimensions to evaluate the proximity effect.

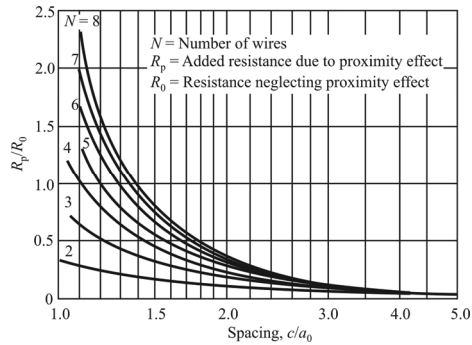


Figure 7.16 Ohmic resistance due to proximity effect.

where l is the length of the wire, a_0 is the radius of the wire and R_S is the surface resistance of the wire (Chapter 3).

7.5.5 Loop Inductance

The inductance of a single circular loop of radius b made of wire of radius a_0 is given by

$$L_{\text{circular}}^{\text{loop}} = \mu b\left(\ln\frac{8b}{a_0} - 2\right) \tag{7.33}$$

The inductance of a square loop with sides b and wire radius a_0 is

$$L_{\text{square}}^{\text{loop}} = 2\mu\frac{b}{\pi}\left[\ln\left(\frac{b}{a_0}\right) - 0.774\right] \tag{7.34}$$

The inductance of a multiturn coil is obtained from the inductance of a single turn loop multiplied by N^2, where N is the number of turns:

$$L_{\text{circular}}^{N\text{loops}} = N^2 L_{\text{circular}}^{1\text{loop}} \tag{7.35}$$

and

$$L_{\text{square}}^{N\text{loops}} = N^2 L_{\text{square}}^{\text{loop}} \tag{7.36}$$

The inductance of the wire itself is very small and is often neglected. It can be shown that the self-inductance of a straight wire of length l is

$$L_{\text{i}} = \frac{\mu_0}{8\pi} l \tag{7.37}$$

For a single loop, $l = 2\pi b$.

7.5.5.1 Air-Core Square Loop

The wire winding around the frame has an inductance L_{loop} that, for a square frame and air core, is approximated with

$$L_{\text{loop}} \approx 8\times10^{-7} N^2 w\left\{\ln\left[\frac{\sqrt{2}wN}{(N+1)l}\right]+0.37942+\frac{(N+1)l}{3wN}\right\} \tag{7.38}$$

where:

L_{loop} is the wiring inductance, h.
N is the number of turns.
w is the frame side length, m.
l is the length of the winding, m.

7.5.5.2 Ferrite Loop

The ferrite loop antenna was discussed above. The wire winding around the ferrite has an inductance L_{loop} that is given by

$$L_{\text{loop}} \approx \frac{\mu_0\mu_r N^2 S}{l} \tag{7.39}$$

where:

L_{loop} is the wiring inductance, h.
μ_0 is the permeability of vacuum, constant of $4\pi \times 10^{-7}$ h m^{-1}.
μ_r is the relative permeability, dimensionless quantity. μ_r is specific to the medium.
N is the number of turns.
S is the area of each winding, m².
l is the length of the winding, m.

7.5.6 Tuning Capacitor

The susceptance of the capacitor B_r must be chosen to eliminate the susceptance of the loop. The equivalent admittance of the loop is

$$Y_{\text{in}} = \frac{1}{Z_{\text{in}}} = \frac{1}{R_{\text{in}} + jX_{\text{in}}} \tag{7.40}$$

where

$$R_{\text{in}} = R_{\text{r}} + R_{\text{loop}}$$
$$X_{\text{in}} = j\omega(L_{\text{A}} + L_{\text{loop}})$$

Then

$$Y_{\text{in}} = G_{\text{in}} + jB_{\text{in}} \tag{7.41}$$

where

$$\begin{aligned} G_{\text{in}} &= \frac{R_{\text{in}}}{R_{\text{in}}^2 + X_{\text{in}}^2} \\ B_{\text{in}} &= -\frac{X_{\text{in}}}{R_{\text{in}}^2 + X_{\text{in}}^2} \end{aligned} \tag{7.42}$$

The susceptance of the capacitor is

$$B_{\text{r}} = \omega C_{\text{r}} \tag{7.43}$$

For resonance to occur at f_0, the condition

$$B_{\text{r}} = -B_{\text{in}} \tag{7.44}$$

must be satisfied, resulting in

$$C_{\text{r}} = \frac{1}{2\pi f_0} \frac{X_{\text{in}}}{R_{\text{in}}^2 + X_{\text{in}}^2} \tag{7.45}$$

At resonance, the input impedance Z'_{in} becomes

$$Z'_{\text{in}} = R_{\text{in}} + \frac{X_{\text{in}}^2}{R_{\text{in}}} \quad \Omega \tag{7.46}$$

7.6 Polarization of Loop Antennas

Horizontally oriented loops (loops in the *x-y* plane) can only transmit in the horizontal plane. Vertical loops, however, can be fed in such a way as to either

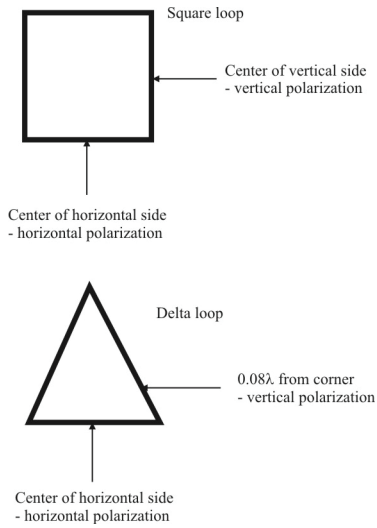


Figure 7.17 Polarizations of loop antennas.

effect horizontal polarization or vertical polarization depending on how (i.e., where) they are fed (see Figure 7.17). If a vertically oriented antenna (loop in the *x*-*z* plane) is fed at either the bottom or the top, the polarization will be horizontal; if fed on the side or bottom corner, it will be vertical. A delta loop should be fed at one of the bottom corners for vertical orientation. If a diamond, then it is fed at mid corner. If a square, then is is fed at the center of one of the vertical sides. Vertical polarization is preferred when the antenna is low because vertically polarized EM waves travel better close to the ground than horizontal ones do.

7.7 Large Loop Antennas

7.7.1 Large Loop Characteristics

7.7.1.1 Circular Loop with Nonuniform Current

When the loop radius becomes larger than about 0.2λ, the constant-current assumption is no longer valid. It is frequently assumed that the distribution is cosine in nature. However, even this one is not satisfactory, especially close to the terminals. That is why the current is often represented by a Fourier series [3]:

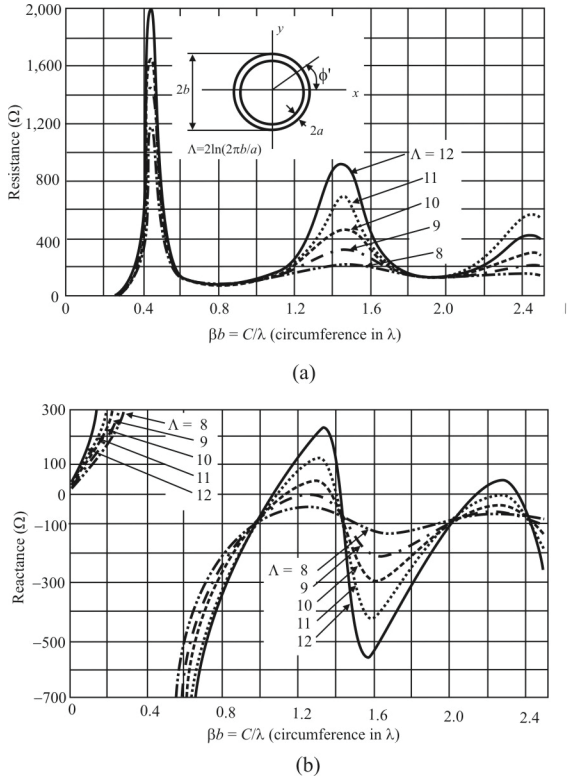


Figure 7.18 Large loop impedance: (a) resistance and (b) reactance.

$$I(\phi') = I_0 + 2\sum_{n=1}^{N} I_n \cos(n\phi') \tag{7.47}$$

for some suitable N.

When the circumference of the loop approaches λ, the maximum of the radiation pattern shifts along the normal to the loop. Then the input resistance of the antenna is also very good (about 50 to 70 mΩ). The maximum directivity occurs when $C = 1.4\lambda$, but then the input impedance is too large.

7.7.1.2 Impedance Characteristics of Large Loop Antennas

The impedance characteristics of large loop antennas are given in terms of the antenna physical parameters and plotted versus the circumference (in wavelengths) in Figure 7.18 [4].

7.8 Concluding Remarks

The basic characteristics of loop antennas were investigated in this chapter. There has been extensive research into the performance of such antennas and a considerable amount is known about their performance. We have but touched the surface here.

Loop antennas are rarely used as transmit antennas due to their inefficiencies. As receive antennas, however, there are configurations, particularly when employing ferrite cores, that make excellent receive antennas.

Large loop antennas are difficult to analyze. A large antenna is a configuration in which it cannot be assumed that the current is constant around the loop. Because of this, almost all applications of loop antennas employ short configurations where the constant current assumption makes sense.

References

[1] Poisel, R. A., *Modern Communications Jamming Principles and Techniques*, 2nd ed., Norwood, MA: Artech House, 2011, Ch. 2.

[2] Smith, G. N., "The Proximity Effect in Systems of Parallel Conductors," *Journal of Applied Physics*, Vol. 43, No. 5, May 1972, pp. 2196–2207.

[3] Wolff, E. A., *Antenna Analysis*, New York: Wiley, 1967.

[4] Storer, J. E., "Impedance of Thin-Wire Loop Antennas," *AIEE Transactions*, Vol. 75, November 1956.

Chapter 8

Traveling Wave Antennas

8.1 Introduction

The ends of resonant antennas (dipoles, monopoles, and so forth) are open circuits and the current of necessity must go to zero there. These antennas can be characterized as *standing wave antennas* since these open circuits cause current to be reflected back toward the source. Therefore, the current on these antennas can be written as a sum of waves traveling in opposite directions (waves that travel toward the end of the wire and are reflected in the opposite direction).

As discussed in Chapter 5, the current on a dipole of length l is given by

$$\begin{aligned}\vec{I}(z) &= I_0 \sin\left[k\left(\frac{l}{2}-|z|\right)\right]\vec{u}_z \qquad -\frac{l}{2}\le z\le\frac{l}{2}\\ &= \begin{cases} I_0 \sin\left[k\left(\frac{l}{2}+z\right)\right]\vec{u}_z & -\frac{l}{2}\le z\le 0\\ I_0 \sin\left[k\left(\frac{l}{2}-z\right)\right]\vec{u}_z & 0\le z\le\frac{l}{2}\end{cases}\end{aligned} \tag{8.1}$$

The current on the upper arm of the dipole can be written as

$$\begin{aligned}I(z) &= I_0 \sin\left[k\left(\frac{l}{2}-z\right)\right]\\ &= \frac{I_0}{2j}\left[e^{jk\left(\frac{l}{2}-z\right)}-e^{-jk\left(\frac{l}{2}-z\right)}\right]\\ &= \underbrace{\frac{I_0}{2j}}_{\text{constant}}\left[\underbrace{e^{jk\frac{l}{2}}e^{-jkz}}_{\substack{+z\text{ directed}\\ \text{wave}}}-\underbrace{e^{-jk\frac{l}{2}}e^{jkz}}_{\substack{-z\text{ directed}\\ \text{wave}}}\right]\end{aligned} \tag{8.2}$$

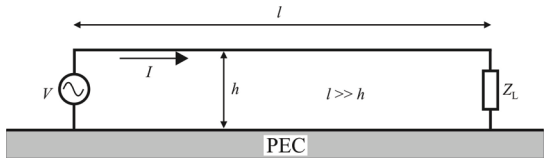


Figure 8.1 Beverage antenna architecture.

which is composed of a wave traveling forward down the antenna (+*z* direction) and a reverse wave traveling in the other direction (–*z* direction). This is a characteristic of resonant antennas.

Traveling wave antennas, on the other hand, have terminations that are matched to the characteristic impedance of the antenna at their ends, not open circuits, so that the current is defined in terms of waves traveling in only one direction, which is a complex exponential as opposed to a sine or cosine. A traveling wave antenna can be formed by a single wire over ground or a single wire transmission line, which is terminated with a matched load resulting in no reflection. Typically, the length of the transmission line is several wavelengths [1–3].

8.2 Beverage Antennas

The antenna shown in Figure 8.1, called a *Beverage* antenna, named after Harold H. Beverage who invented it in 1918. The effects of an imperfect ground may be significant and must be included. The contribution to the far fields due to the vertical conductors is typically neglected since it is small if $l >> h$. Note that the antenna does not radiate efficiently if the height h is small relative to wavelength. We will see shortly that this antenna has a low take-off angle and therefore is popular to use for long-distance HF communication systems; it is also popular for long-distance EW applications.

Traveling wave antennas are commonly formed using wire segments of varying geometries. Therefore, the antenna far field can be obtained by superposition using the far fields of the individual segments.

Consider a segment of a traveling wave antenna (an electrically long wire of length l lying along the z-axis) in a medium with dielectric constant ε and magnetic permeability μ as shown in Figure 8.2. A traveling wave current, $I(z)$, flows in the +z-direction. This current is given by

$$\vec{I}(z) = I_0 e^{-\gamma z}\vec{u}_z$$

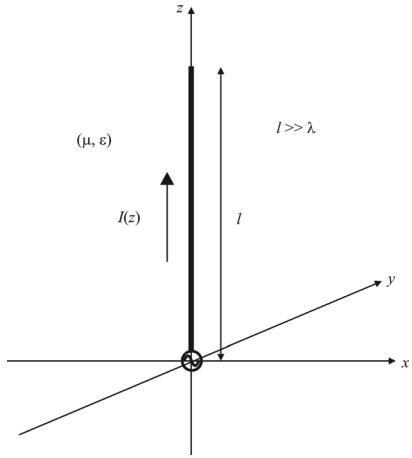


Figure 8.2 Vertical wire.

$$= I_0 e^{-(\alpha + j\beta)z} \vec{u}_z$$
$$= I_0 e^{-\alpha z} e^{-j\beta z} \vec{u}_z \tag{8.3}$$

where α is the attenuation constant and β is the phase constant. If the losses for the antenna, as represented by α, are negligible (ohmic loss in the conductors, loss due to imperfect ground, etc.), then the current can be written as

$$\vec{I}(z) = I_0 e^{-j\beta z} \vec{u}_z \tag{8.4}$$

From (3.103) we know that the far field vector potential (generalized to account for the possibly bounded nature of propagation from the traveling wave antenna so that β is not necessarily equal to k_0) due to a constant current I_0 is given by

$$\vec{A} \approx \mu \frac{e^{-jk_0 r}}{4\pi r} \int_0^l \vec{I}(z) e^{jk_0 z \cos\theta} dz$$
$$= \mu I_0 \frac{e^{-jk_0 r}}{4\pi r} \int_0^l e^{j(k_0 \cos\theta - \beta)z} dz \vec{u}_z$$
$$= -j\mu I_0 \frac{e^{-jk_0 r}}{4\pi r} \left[\frac{e^{j(k_0 \cos\theta - \beta)z}}{k_0 \cos\theta - \beta} \right]_0^l \vec{u}_z$$

$$
\begin{aligned}
&= -j\mu I_0 \frac{e^{-jk_0 r}}{4\pi r}\left[\frac{e^{j(k_0\cos\theta-\beta)l}-1}{k_0\cos\theta-\beta}\right]\vec{u}_z \\
&= -j\mu I_0 \frac{e^{-jk_0 r}}{4\pi r}\left[\frac{e^{j\frac{1}{2}(k_0\cos\theta-\beta)l}}{k\cos\theta-\beta}\right]\left[e^{j\frac{1}{2}(k_0\cos\theta-\beta)l}-e^{-j\frac{1}{2}(k_0\cos\theta-\beta)l}\right]\vec{u}_z \\
&= -j\mu I_0 \frac{e^{-jk_0 r}}{4\pi r}\left[\frac{e^{j\frac{1}{2}(k_0\cos\theta-\beta)l}}{k_0\cos\theta-\beta}\right]\left\{2j\sin\left[\frac{l}{2}(k_0\cos\theta-\beta\right]\right\}\vec{u}_z \\
&= \mu I_0 \frac{e^{-jk_0 r}}{4\pi r} e^{j\frac{1}{2}(k_0\cos\theta-\beta)l}\left[\frac{\sin\left[\frac{l}{2}(k_0\cos\theta-\beta\right]}{\frac{l}{2}(k_0\cos\theta-\beta)}\right]\vec{u}_z
\end{aligned}
\tag{8.5}
$$

If we let

$$
u = \frac{l}{2}(k_0\cos\theta-\beta) \tag{8.6}
$$

then

$$
\begin{aligned}
\vec{A} &= \mu I_0 l \frac{e^{-jk_0 r}}{4\pi r} e^{ju}\frac{\sin u}{u}\vec{u}_z \\
&= \mu I_0 l \frac{e^{-jk_0 r}}{4\pi r}(\cos\theta\vec{u}_r - \sin\theta\vec{u}_\theta)e^{ju}\operatorname{sinc} u
\end{aligned}
\tag{8.7}
$$

The far fields are therefore given by

$$
E_\theta \approx -j\omega A_\theta = j\omega\mu I_0 l\frac{e^{-jk_0 r}}{4\pi r}\sin\theta e^{ju}\frac{\sin u}{u} \qquad H_\phi = \frac{E_\theta}{\eta_0} \tag{8.8}
$$

The 2-D radiation patterns of the 300 m beverage antenna over a PEC ground shown in Figure 8.3 at 1.825 MHz are shown in Figure 8.4 while the 3-D patterns are shown in Figure 8.5. We can see that there are substantial sidelobes in both the horizontal and vertical planes. The first sidelobe in the horizontal plane is about 15 dB below the main lobe while the first sidelobe in the vertical plane is about 9 dB down.

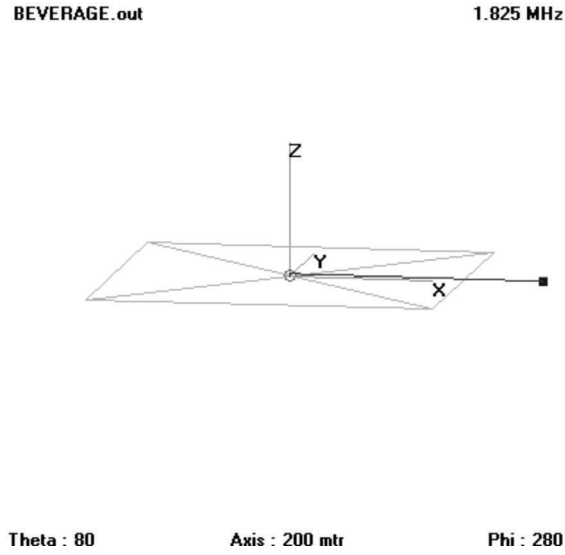


Figure 8.3 Beverage antenna model over a PEC Earth. Antenna length in this case is 320 m. f = 1.825 MHz, λ = 164 m.

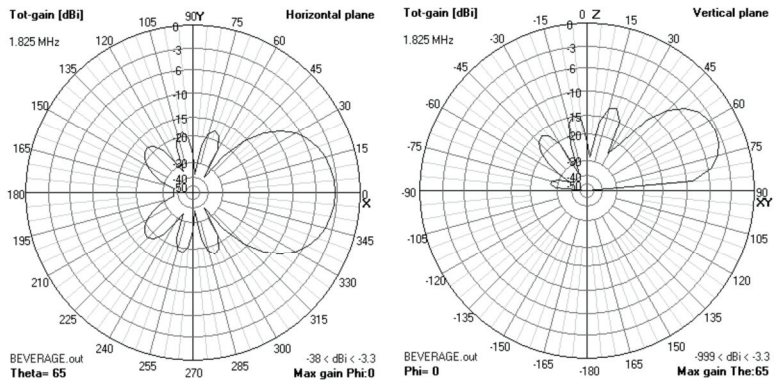


Figure 8.4 Beverage antenna in Figure 8.3 2-D patterns. f = 1.825 MHz, λ = 164 m.

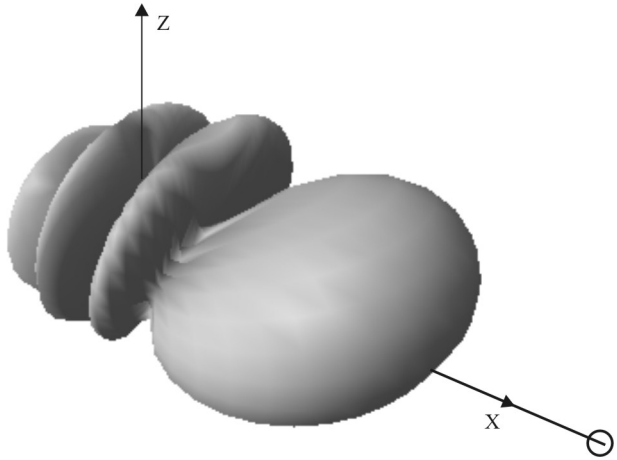


Figure 8.5 Beverage antenna in Figure 8.3 3-D pattern. $f = 1.825$ MHz.

In a lossless transmission line, the phase constant is linearly related to frequency by $\beta = 2\pi/\lambda$. However, primarily because of finite losses, the phase constant of a lossy transmission line wave (a guided wave) can be very different from that of an unbounded medium (an unguided wave, a lossless situation).

However, for a traveling wave antenna, the electrical height of the conductor above ground is typically large (at least $\lambda/2$) and the phase constant approaches that of an unbounded medium (k_0). (Note that $l >> \lambda$ still.) If we assume that the phase constant of the traveling wave antenna is the same as an unbounded medium ($\beta = k_0$) and recalling that $\omega\mu = \eta_0\beta$, then

$$u = \frac{\beta l}{2}(\cos\theta - 1) \tag{8.9}$$

$$E_\theta = j\eta_0\beta I_0 \frac{e^{-j\beta r}}{4\pi r} e^{j\frac{\beta l}{2}(\cos\theta-1)} \sin\theta \frac{\sin\left[\frac{\beta l}{2}(\cos\theta-1)\right]}{\frac{\beta l}{2}(\cos\theta-1)}$$

$$= j\eta_0 I_0 \frac{e^{-j\beta r}}{2\pi r} e^{j\frac{\beta l}{2}(\cos\theta-1)} \sin\theta \frac{\sin\left[\frac{\beta l}{2}(\cos\theta-1)\right]}{\cos\theta - 1} \tag{8.10}$$

Given the far field of the traveling wave segment (8.8), we can determine the time-average radiated power density with the Poynting vector (2.61):

$$\begin{aligned}\vec{P} &= \frac{1}{2}\vec{E}\times\vec{H}^{*} = \frac{|E_\theta|^2}{2\eta_0}\vec{u}_{\mathrm{r}} \\ &= \eta_0 \frac{|I_0|^2}{8\pi^2 r^2}\frac{\sin^2\theta}{(\cos\theta-1)^2}\sin^2\left[\frac{\beta l}{2}(\cos\theta-1)\right]\vec{u}_{\mathrm{r}}\end{aligned} \tag{8.11}$$

The total power radiated by the traveling wave segment is found by integrating the Poynting vector.

$$\begin{aligned}P_{\mathrm{rad}} &= \iint \vec{P}\bullet d\vec{s} \\ &= \frac{\eta_0}{4\pi}|I_0|^2\left[1.415+\ln\left(\frac{\beta l}{\pi}\right)-Ci(2\beta l)+\frac{\sin(2\beta l)}{2\beta l}\right]\end{aligned} \tag{8.12}$$

and the radiation resistance is

$$R_{\mathrm{r}} = \frac{2P_{\mathrm{rad}}}{|I_0|^2} == \frac{\eta_0}{2\pi}\left[1.415+\ln\left(\frac{\beta l}{\pi}\right)-Ci(2\beta l)+\frac{\sin(2\beta l)}{2\beta l}\right] \tag{8.13}$$

where *Ci* is the cosine integral given by [4]

$$Ci(y) = C+\ln y+\int_0^y \frac{\cos u-1}{u}du \tag{8.14}$$

and C is Euler's constant: $C \approx 0.57722$.

The radiation resistance of the ideal traveling wave antenna (where VSWR = 1) is purely real. Figure 8.6 shows a plot of the radiation resistance of the traveling wave segment as a function of segment length. We see that the radiation resistance of the traveling wave antenna is much more uniform than that seen in resonant antennas. Thus, traveling wave antennas are classified as broadband antennas.

The pattern function of the traveling wave antenna is given by

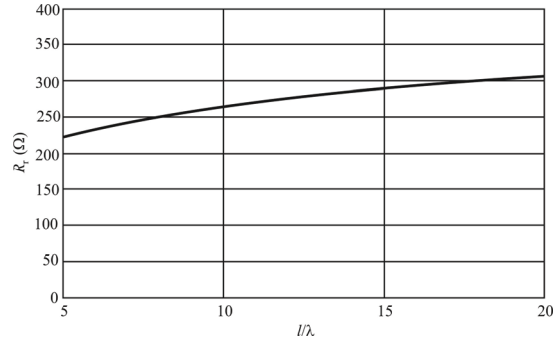


Figure 8.6 Radiation resistance of a Beverage antenna as a function of length in wavelengths.

$$U(\theta,\phi) = r^2 P = \underbrace{\eta_0 \frac{|I_0|^2}{8\pi^2}}_{\text{constant}} \cot^2\frac{\theta}{2}\sin^2\left[\frac{\beta l}{2}(\cos\theta - 1)\right] \tag{8.15}$$

so the normalized pattern function is

$$\begin{aligned} F(\theta,\phi) &= \cot^2\frac{\theta}{2}\sin^2\left[\frac{\beta l}{2}(\cos\theta - 1)\right] \\ &= \cot^2\frac{\theta}{2}\sin^2\left[\frac{\pi l}{\lambda}(\cos\theta - 1)\right] \end{aligned} \tag{8.16}$$

The normalized pattern function of the traveling wave segment shown in Figure 8.7 that is oriented along the *x*-axis is shown in Figure 8.8 for antenna lengths of 5λ, 10λ, 15λ, and 20λ (obtained by varying the frequency). It can be argued that $l = 5\lambda$ violates the assumption that $l >> \lambda$, however, so this particular pattern is at best an approximation.

We see from Figure 8.8 that as the electrical length of the traveling wave segment increases, the main beam becomes somewhat narrower while the angle of the main beam moves slightly toward the axis of the antenna. Also note that the pattern function of the traveling wave segment always has a null at $\theta = 0^o$ (at zenith). In addition, note that with $l >> \lambda$, the sine function in the normalized pattern function varies much more rapidly (more peaks and nulls) than the cotangent function.

The approximate angle of the main lobe for the traveling wave segment is found by determining the first peak of the sine function in the normalized pattern function.

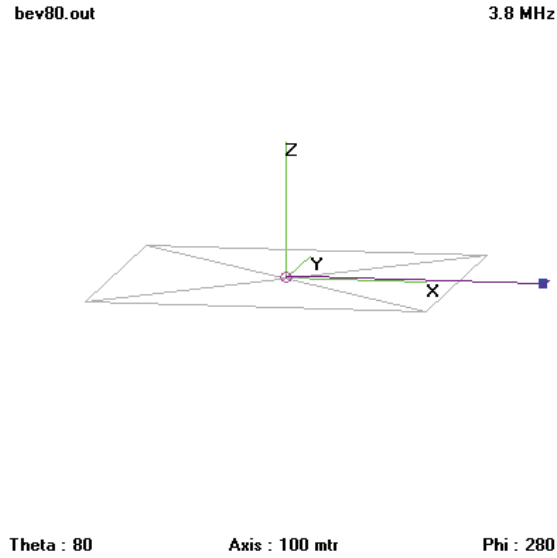
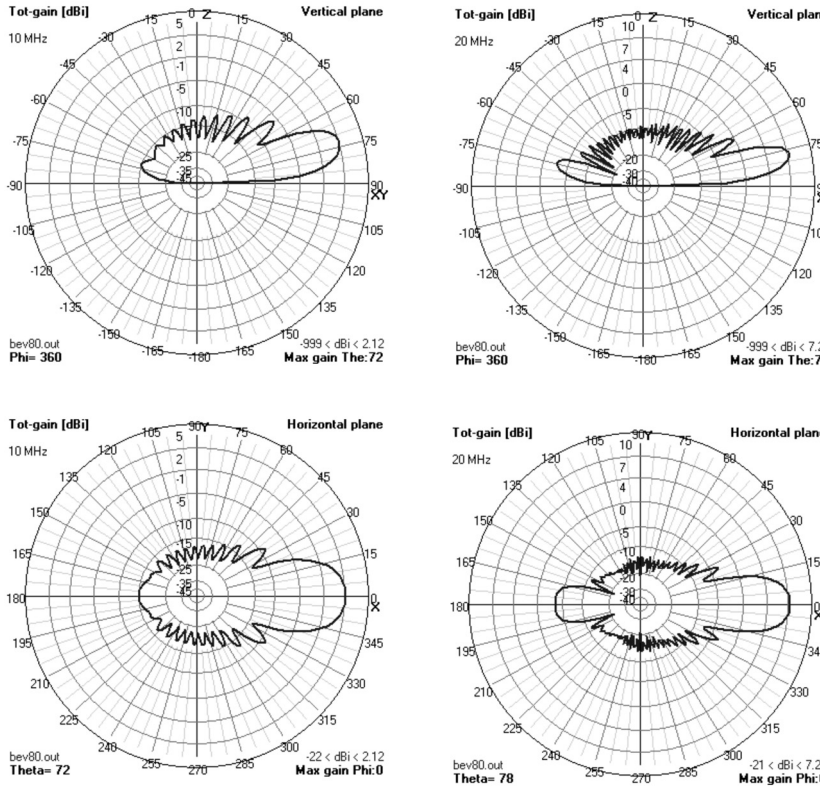
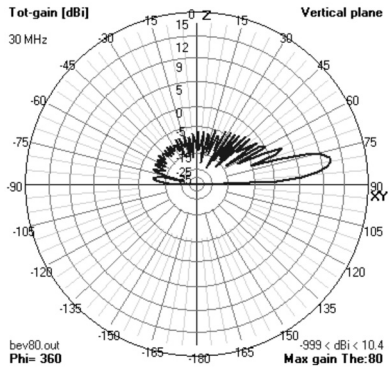


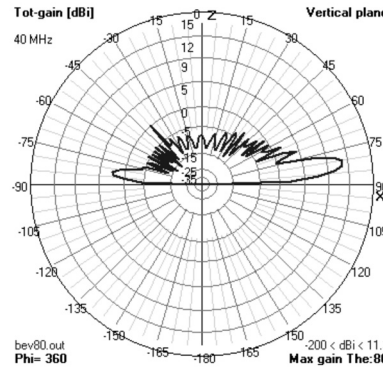
Figure 8.7 Beverage antenna where length is 160 at f = 3.8 MHz, λ = 80 m.



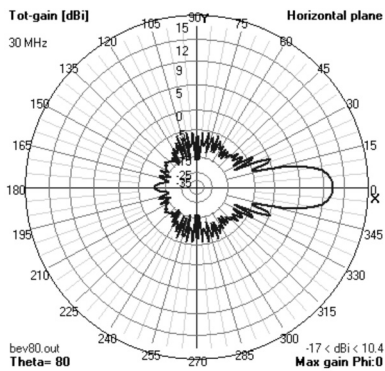
(a) (b)

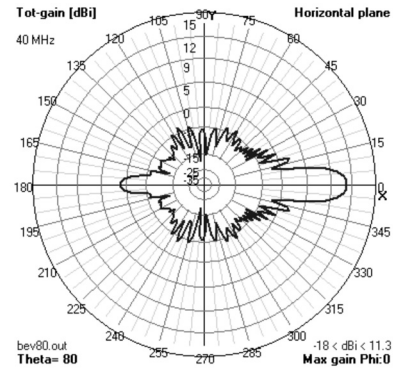
Figure 8.8 (a) 2-D radiation patterns for the antenna in Figure 8.7 for $l = 5\lambda$, (b) $l = 10\lambda$, (c) $l = 15\lambda$, and (d) $l = 20\lambda$.





(c)





(d)

Figure 8.8 (continued)

$$\sin\left[\frac{\pi l}{\lambda}(\cos\theta-1)\right]_{\theta=\theta_m}=\pm 1$$

$$\frac{\pi l}{\lambda}(\cos\theta_m-1)=(2m+1)\frac{\pi}{2}$$

$$\cos\theta_m=\frac{\lambda}{2l}(2m+1)+1$$

so that

$$\theta_m=\cos^{-1}\left[1+\frac{\lambda}{2l}(2m+1)\right] \qquad m=\ldots,-2,-1,0,1,2,\ldots \tag{8.17}$$

The values of m which yield $0^o \le \theta_m \le 180^o$ are negative. The smallest value of θ_m in this region defines the location of the main beam $(m=\pm 1)$

$$\theta_{\max}=\cos^{-1}\left(1-\frac{\lambda}{2l}\right) \tag{8.18}$$

If we also account for the cotangent function in the determination of the main beam angle, we find

$$\theta_{\max}=\cos^{-1}\left(1-\frac{0.371\lambda}{l}\right) \tag{8.19}$$

$l=5\lambda$: $\theta_{\max}=\cos^{-1}\left(1-\frac{0.371}{5}\right)=22.2^o$

$l=10\lambda$: $\theta_{\max}=\cos^{-1}\left(1-\frac{0.371}{10}\right)=15.7^o$

$l=15\lambda$: $\theta_{\max}=\cos^{-1}\left(1-\frac{0.371}{15}\right)=12.8^o$

$l=20\lambda$: $\theta_{\max}=\cos^{-1}\left(1-\frac{0.371}{20}\right)=11.1^o$

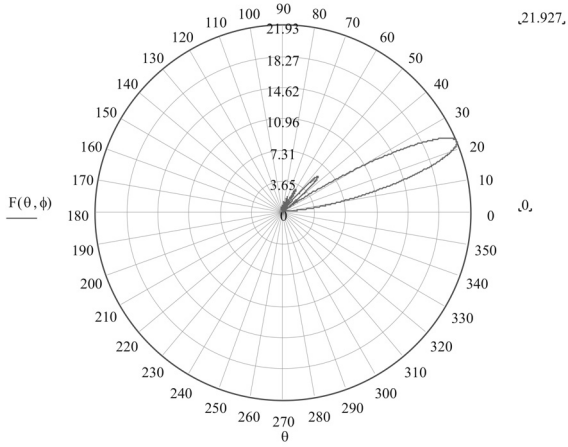


Figure 8.9 Calculated elevation patterh for the Beverage antenna in Figure 8.7 when $l = 5\lambda$. There is a difference from NEC modeling 18° versus 22°.

The radiation pattern for 5λ is shown expanded in Figure 8.9, from NEC modeling, and we see that the modeling results are reasonably close to the calculated take-off angle.

From (3.5) the directivity of the traveling wave segment is

$$D(\theta,\phi) = 4\pi \frac{U(\theta,\phi)}{P_{\text{rad}}} = \frac{2\cot^2\frac{\theta}{2}\sin^2\left[\frac{\beta l}{2}(\cos\theta - 1)\right]}{1.415 + \ln\left(\frac{\beta l}{\pi}\right) - Ci(2\beta l) + \frac{\sin(2\beta l)}{2\beta l}} \tag{8.20}$$

where the maximum directivity can be approximated by

$$D_{\text{o}} \approx \frac{2\cot^2\left[\frac{1}{2}\cos^{-1}\left(1 - \frac{0.371\lambda}{l}\right)\right]}{1.415 + \ln\frac{\beta l}{\pi} - Ci(2\beta l) + \frac{\sin(2\beta l)}{2\beta l}} \tag{8.21}$$

where the sine term in the numerator of the directivity function is assumed to be unity at the main beam.

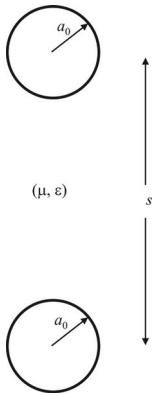


Figure 8.10 Two-wire transmission line end view.

8.3 Traveling Wave Antenna Terminations

8.3.1 Two-Wire Transmission Line

For the two-wire transmission line shown in Figure 8.10 the characteristic impedance is given by

$$Z_0^{\text{two-wire}} = \frac{\eta_0}{\pi} \ln\left[\frac{s}{2a_0} + \sqrt{\left(\frac{s}{2a_0}\right)^2 - 1}\right] \tag{8.22}$$

If $s >> a_0$, then

$$Z_0^{\text{two-wire}} \approx \frac{\eta_0}{\pi} \ln\frac{s}{a_0} \tag{8.23}$$

and in air, where $\eta_0 = 120\pi\ \Omega$,

$$Z_0^{\text{two-wire}} \approx 120 \ln\frac{s}{a_0} \tag{8.24}$$

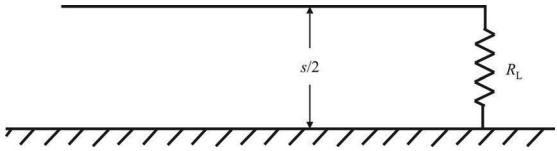


Figure 8.11 One-wire uniform transmission line side view.

8.3.2 One-Wire Transmission Line

Given a traveling wave antenna segment located horizontally above a ground plane, the termination R_L required to match the *uniform transmission line* formed by the cylindrical conductor over ground shown in Figure 8.11 (radius = a_0, height over ground = $s/2$; see Figure 8.12) is the characteristic impedance of the corresponding one-wire transmission line. If the conductor height above the ground plane varies with position, the conductor and the ground plane form a *non-uniform transmission line*. The characteristic impedance of a nonuniform transmission line is a function of position.

For the one-line transmission line shown in Figure 8.11,

$$Z_0^{\text{one-wire}} = \frac{1}{2} Z_0^{\text{two-wire}}$$

$$Z_0^{\text{one-wire}} = \frac{\eta_0}{2\pi} \ln\left[\frac{s}{2a_0} + \sqrt{\left(\frac{s}{2a_0}\right)^2 - 1} \right] \tag{8.25}$$

and if $s >> a_0$

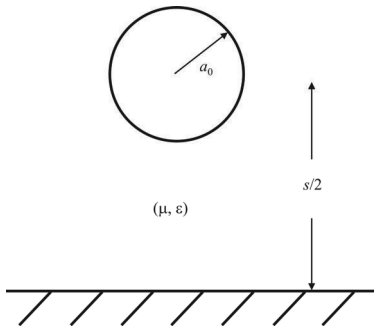


Figure 8.12 One-wire transmission line end view.

$$Z_0^{\text{one-wire}} \approx \frac{\eta_0}{2\pi} \ln \frac{s}{a_0} \tag{8.26}$$

In air,

$$Z_0^{\text{one-wire}} \approx 60 \ln \frac{s}{a_0} \tag{8.27}$$

8.4 Vee Traveling Wave Antenna

As we have seen, the main beam of a single electrically long wire guiding waves in one direction, which we have been calling a traveling wave segment, is inclined at an angle relative to the axis of the wire. Traveling wave antennas are typically formed by implementing several traveling wave segments. These traveling wave segments can be oriented such that the main beams of the constituent wire elements combine to enhance the directivity of the overall antenna. A *vee traveling wave antenna* is formed by connecting two matched traveling wave segments to the end of a transmission line feed at an angle of $2\theta_0$ relative to each other as illustrated in Figure 8.13.

The beam angle of a traveling wave segment relative to the axis of the wire (θ_{max}) depends on the length of the wire. Given the length of the wires in the vee traveling wave antenna, the angle $2\theta_0$ may be chosen such that the main beams of the two tilted wires combine to form an antenna with increased directivity over that of a single wire as we see in Figure 8.13.

We can show that a complete analysis that takes into account the spatial separation effects of the antenna arms (the two wires are not colocated) reveals that by choosing $\theta_0 \approx 0.8\theta_{max}$, the total directivity of the vee traveling wave antenna is approximately twice that of a single conductor. Note that the overall pattern of the vee antenna shown in Figure 8.14 is essentially unidirectional given matched conductors.

If, on the other hand, the conductors of the vee traveling wave antenna are resonant conductors (*vee dipole antenna*), there are reflected waves that produce significant beams in the opposite direction. Thus, traveling wave antennas, in general, have the advantage of essentially unidirectional patterns when compared to the patterns of corresponding resonant antennas.

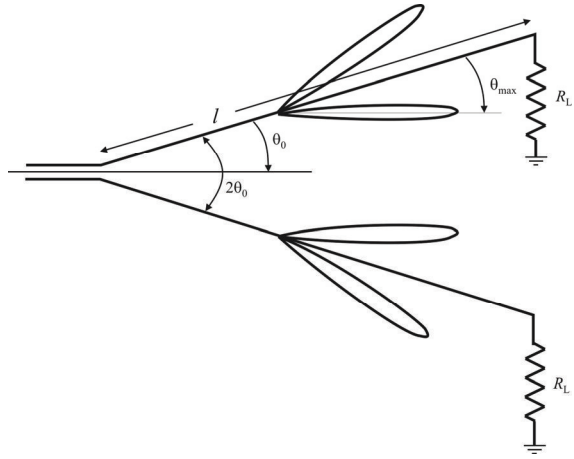


Figure 8.13 Vee traveling wave antenna with associated horizontal plane patterns.

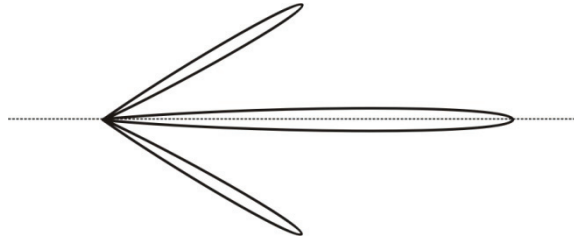


Figure 8.14 Vee antenna horizontal pattern.

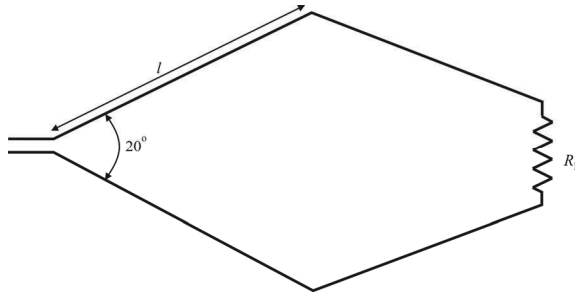


Figure 8.15 Rhombic antenna top view (horizontal polarization).

8.5 Rhombic Antennas

The *rhombic antenna* is formed by connecting two vee traveling wave antennas at their open ends, as in Figure 8.15. It was invented by Edmond Bruce in 1931 and carries its appellation because of its diamond shape. The antenna feed is located at one end of the rhombus and a matched termination is located at the opposite end. Typically, all four conductors of the rhombic antenna have the same length. Note that the rhombic antenna is an example of a nonuniform transmission line. As with all traveling wave antennas, we assume that the reflections from the load are negligible.

To produce a single antenna main lobe along the axis of the rhombic antenna, the individual conductors of the rhombic antenna should be aligned such that the component lobes numbered 2, 3, 5, and 8 in Figure 8.16 are aligned (accounting for spatial separation effects). Beam pairs (1, 7) and (4, 6) combine to form significant sidelobes but at a level smaller than the main lobe.

Modeling results for the rhombic antenna illustrated in Figure 8.15 that is 40 m long at 30 MHz are shown in Figures 8.17 and 8.18. Figure 8.17 shows the 2-D patterns while Figure 8.18 shows the 3-D pattern, illustrating the sidelobes generated.

8.5.1 Semi-Rhombic

A rhombic antenna can also be constructed using an inverted vee antenna over a ground plane as shown in Figure 8.19 (called a semi-rhombic). The termination resistance is one-half that required for the isolated rhombic antenna. The patterns of a semi-rhombic shown in Figure 8.19 are shown in Figures 8.20 and 8.21. The frequency in this case is 3.8 MHz while the antenna is 150 m long. We can see that again there are significant sidelobes, with the first sidelobe in the horizontal plane

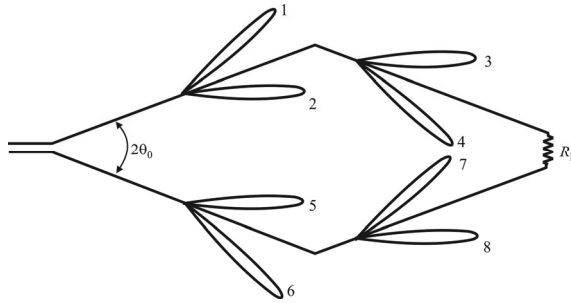


Figure 8.16 Top view of four-leg rhombic with horizontal plane patterns.

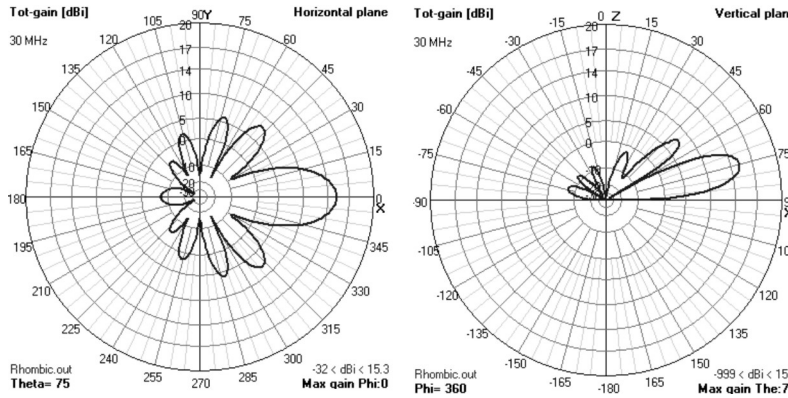


Figure 8.17 2-D radiation patterns of four-leg rhombic.

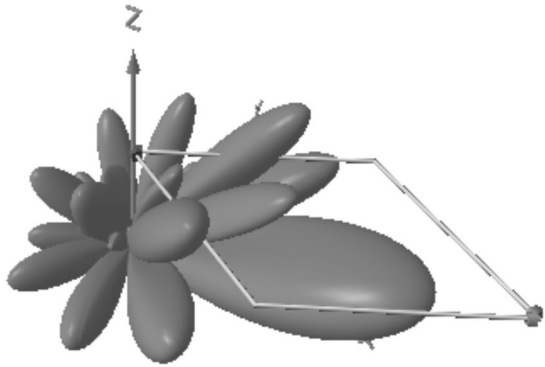


Figure 8.18 3-D patterns of four-leg rhombic antenna.

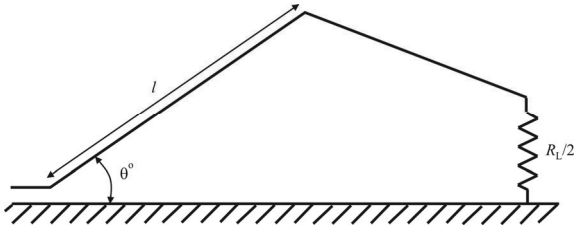


Figure 8.19 Half of a rhombic antenna side view. Also called a semi-rhombic.

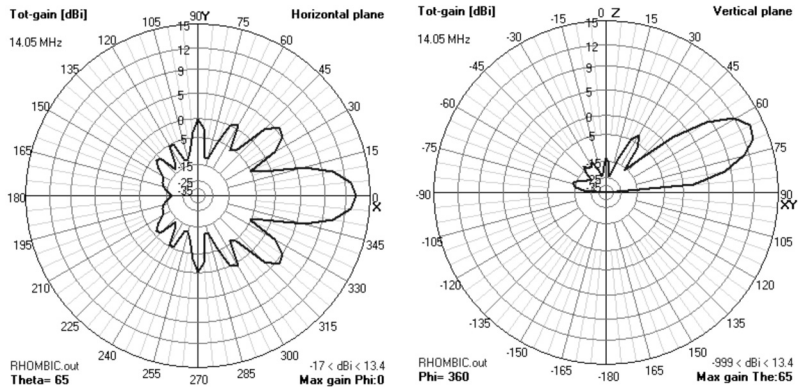


Figure 8.20 Semi-rhombic 2-D patterns at $f = 3.8$ MHz.

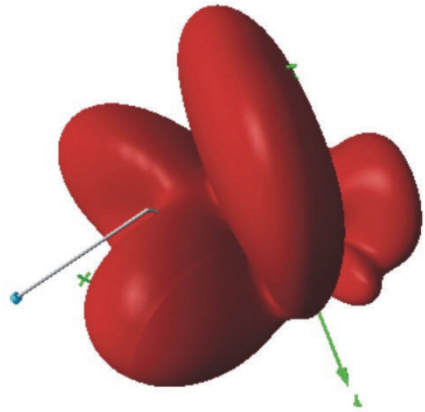


Figure 8.21 Semi-rhombic 3-D pattern.

only 7 dB lower than the main beam. In the vertical plane the first sidelobe is about 15 dB down and the take-off angle is about 67^{o}.

8.6 Concluding Remarks

We have examined the basic principles of operation of traveling wave antennas in this chapter. Traveling wave antennas are nonresonant in that there are no standing waves created. Instead, the antennas are terminated in their characteristic impedance and, just as for transmission lines, when this is the case, all of the energy is transferred to the load.

The Beverage and rhombic antennas are popular for long-distance communications and intercept in the HF frequency range, as well as EA in this range, primarily because of the low take-off angle of the main lobe in the radiation pattern. This facilitates small incidence angles of the HF signals at the ionospheric layers that refract the signal back to the Earth. Intercept of HF signals from several 1,000s of kilometers is possible with these antennas, as is conducting effective EA.

Microstrip antennas in the higher frequency ranges can also be configured as traveling wave antennas. We examine microstrip antennas in Chapter 17.

References

[1] Stutzman, W. L., and G. A. Thiele, *Antenna Theory and Design*, New York: Wiley, 1981, pp. 239–244.

[2] Balanis, C. A., *Antenna Theory*, 3rd ed., New York: Wiley, 2005, Ch. 8.

[3] Jordan, E. C., and K. G. Balmain, *Electromagnetic Waves and Radiating Systems*, 2nd ed., Englewood Cliffs, NJ: Prentice-Hall, 1968, pp. 356–359.

[4] Abramowitz, M., and I. A. Stegun (Eds.), *Handbook of Mathematical Functions*, New York: Dover, 1965, pp. 231–233.

Chapter 9

Antenna Arrays

9.1 Introduction

In this chapter we discuss antenna arrays, which are arrangements of multiple, simpler antenna types (wire dipoles or monopoles, for example). Such structures exhibit some characteristics not available in the simpler configurations, such as very narrow boresight beams and broader bandwidth [1–4].

Antenna arrays can be thought of as sampled aperture antennas. The elements in the array sample the EM wave as it passes by. Therefore, much of the theory of sampled data systems can be and has been applied to the analysis of antenna arrays. The array elements can be of any variety and each type brings its own characteristics. The effects of the array can be analyzed separately, as we will show.

Frequently antennas must be broadband, since the frequencies of interest can be located over wide portions of the frequency spectrum. Limited space on some systems normally precludes using several antennas, each designed for a relatively narrow frequency band. Thus, antennas that are effective over a wide frequency range are desirable.

An antenna array consists of many small antennas configured in a particular way [5]. Adjusting the phase characteristics of the signal path to each of these antennas from the signal source controls the beam shape of this antenna. By adjusting the phase of the signal slightly, a beam can be steered in particular directions. A *phased array* is an array of identical elements that achieves a given pattern through the control of the element excitation phasing. Phased arrays can be used to steer the main beam of the antenna without physically moving the antenna.

Antennas with a given radiation pattern may be arranged in a pattern (line, circle, plane) to yield a different radiation pattern. In a *linear array* the antenna elements are arranged along a straight line. In a *circular array* the antenna elements are arranged around a circular ring. In a *planar array* the antenna elements are arranged over some planar surface, for example, a rectangular array.

In a *conformal array* the antenna elements are arranged to conform to some nonplanar surface, such as the skin of an aircraft or missile.

There are several array design variables that can be changed to achieve the overall array pattern design. These include the general array shape as just described (linear, circular, planar), the element spacing, the element excitation amplitude, and the element excitation phase.

The important subject of array signal processing, or adaptive arrays, and in particular beamforming, is a legitimate discussion in a chapter on antenna arrays. While we introduce the subject in this chapter, that topic is so significant that an entire chapter is devoted to it. That material is presented in Chapter 17.

Phased arrays form a class of antennas that offer many important features such as electronic beam scanning. They are, however, in general, more complex and therefore more expensive than many other types of antennas. This flexibility has to be balanced against their complexity and cost. In this chapter the fundamental principles of phased arrays are described.

This chapter is structured as follows. We begin with a discussion of general pattern characteristics of arrays, including a discussion of the pattern multiplication theorem. That is followed with a section on linear antenna arrays, including the uniform linear array and uniform circular array. We include derivations of their array factors. The next topic is a presentation of the characteristics of planar arrays, followed by a discussion on active arrays. Every array has certain amounts of mutual coupling and impedance between elements, so the next section covers that topic. We conclude with a presentation on a novel way of determining the input impedance of an antenna array.

9.2 Pattern of a Generalized Array

A 3-D array with an arbitrary geometry is depicted in Figure 9.1. In spherical coordinates, the vector from the origin to the *m*th element of the array is given by $\vec{r}_m = (r_m, \theta_m, \phi_m)$ and $\vec{u}_\beta = (1, \theta, \phi)$ is the unit-length vector in the direction of the source of an incident wave. Throughout this discussion it is assumed that the source of the wave is in the far field of the array and the incident wave can be treated as a plane wave.

Pattern Multiplication Theorem: The pattern of an array of identical antennas is proportional to the product of the pattern of a single element and a factor called the *array factor*. The pattern of a single element is a function of the element geometry and element excitation while the array factor is a function of the physical configuration of the array and the excitation of the elements. That is,

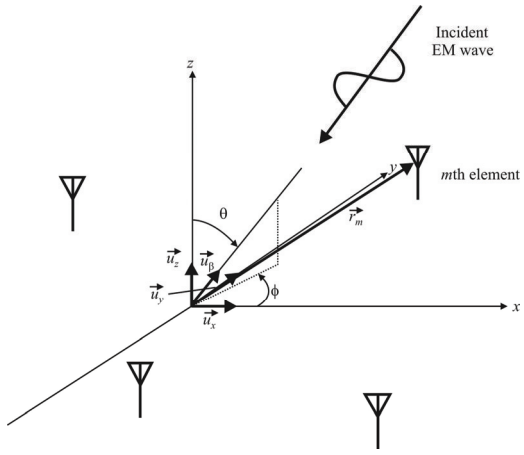


Figure 9.1 An arbitrary 3-D array.

$$\text{Array Pattern} \propto \text{Element Pattern} \times \text{Array Factor} \tag{9.1}$$

Example: Consider the image of the linear element shown in Figure 9.2(a) over a perfectly electrical conductor (PEC, for which $\sigma \to \infty$). The far field of the two-element array shown in Figure 9.2(b) at point *P*, a distance *r* from the array, is given by

$$E_\theta \approx \underbrace{j\eta_0 \frac{\beta I_0 \delta l e^{-j\beta r}}{4\pi r}}_{\text{constant}} \underbrace{\sin\theta}_{\text{element pattern}} \underbrace{[2\cos(\beta h \cos\theta)]}_{\text{array factor}} \tag{9.2}$$

To find the array factor, it is necessary to find the relative phase of the received plane wave at each element. The phase is referenced to the phase of the plane wave at the origin, or some other convenient point, such as the first antenna. Thus, the phase of the received plane wave at the *m*th element is the phase constant $\beta = 2\pi/\lambda$ multiplied by the projection of the element position $\vec{r}_m$ onto the plane wave arrival vector $-\vec{u}_\beta$. This is given by $-\beta\vec{u}_\beta \bullet \vec{r}_m$ with the dot product taken in rectangular coordinates.

In rectangular coordinates,

$$-\vec{u}_\beta = \sin\phi\cos\theta\vec{u}_x + \sin\phi\sin\theta\vec{u}_y + \cos\phi\vec{u}_z \tag{9.3}$$

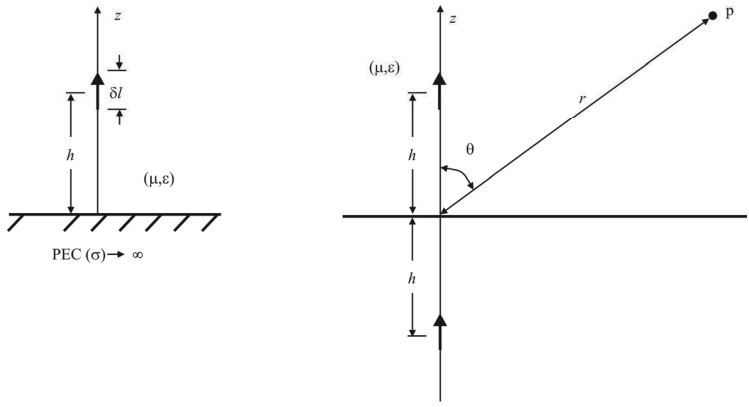


Figure 9.2 An elevated antenna. We assume that the far field of an antenna is all that is of interest for examination of EW antenna system performance; (a) segment of linear element and (b) far-field point.

$$\vec{r}_{\mathrm{m}} = \rho_{\mathrm{m}} \sin\phi_{\mathrm{m}} \cos\theta_{\mathrm{m}} \vec{u}_{\mathrm{x}} + \rho_{\mathrm{m}} \sin\phi_{\mathrm{m}} \sin\theta_{\mathrm{m}} \vec{u}_{\mathrm{y}} + \rho_{\mathrm{m}} \cos\phi_{\mathrm{m}} \vec{u}_{\mathrm{z}} \tag{9.4}$$

and the relative phase of the incident wave at the *m*th element is

$$\begin{aligned} \zeta_m &= -\beta \vec{u}_{\beta} \cdot \vec{r}_{\mathrm{m}} \\ &= \beta r_{\mathrm{m}} (\sin\phi\cos\phi\sin\phi_{\mathrm{m}} \cos\theta_{\mathrm{m}} + \sin\phi\sin\theta\sin\phi_{\mathrm{m}} \sin\theta_{\mathrm{m}} + \cos\phi\cos\phi_{\mathrm{m}}) \\ &= \beta(x_{\mathrm{m}} \sin\phi\cos\theta + y_{\mathrm{m}} \sin\phi\sin\theta + z_{\mathrm{m}} \cos\phi) \end{aligned} \tag{9.5}$$

9.2.1 Array Factor

For an array of *M* elements, the array factor is given by

$$AF(\theta,\phi) = \sum_{m=0}^{M-1} I_m e^{j(\zeta_m + \varphi_m)} \tag{9.6}$$

where I_m is the magnitude and φ_m is the phase of the weighting of the *m*th element.

The normalized array factor is given by

$$f(\phi,\theta)=\frac{\mathrm{AF}(\phi,\theta)}{\max\left|\mathrm{AF}(\phi,\theta)\right|} \tag{9.7}$$

This would be the same as the array pattern if the array consisted of ideal isotropic elements.

9.2.2 Array Pattern

If each element has a pattern $p_m(\theta, \phi)$, which may be different for each element, the normalized array pattern is given by

$$f(\theta,\phi)=\frac{\sum_{m=0}^{M-1} I_m p_m(\theta,\phi)e^{j(\zeta_m+\varphi_m)}}{\max\left|\sum_{m=0}^{M-1} I_m p_m(\theta,\phi)e^{j(\zeta_m+\varphi_m)}\right|} \tag{9.8}$$

In (9.8), the element patterns must be represented such that the pattern maxima are equal to the element gains relative to a common reference.

9.3 Linear Antenna Arrays

9.3.1 *M*-Element Uniform Linear Array

A ULA is defined by uniformly-spaced identical elements of equal magnitude with a linearly progressive phase from element to element. An *M*-element ULA oriented along the *z*-axis is shown in Figure 9.3.

9.3.1.1 Array Factor

For a linear array of *M* elements, the array factor for the ULA shown in Figure 9.3 is

$$\mathrm{AF}(\phi,\theta)=\sum_{m=0}^{M-1} I_m e^{j\beta md\cos\theta} \tag{9.9}$$

where I_m is the element current for the *m*th element. The array factor for an 8-element time scanned ULA is illustrated in Figure 9.4.

The normalized array factor is given by

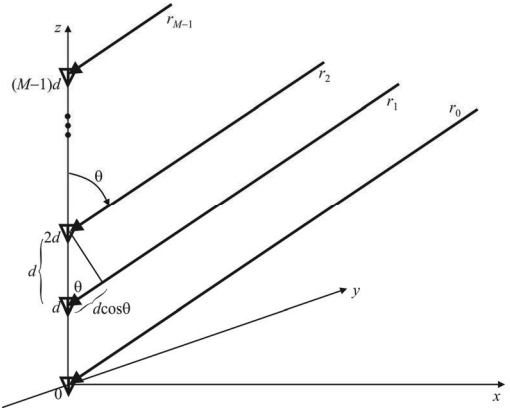


Figure 9.3 Uniform linear array stretching in the *z*-direction.

$$f(\phi,\theta)=\frac{\mathrm{AF}(\phi,\theta)}{\max\left|\mathrm{AF}(\phi,\theta)\right|} \tag{9.10}$$

This would be the same as the array pattern if the array consisted of ideal isotropic elements.

The array factor is independent of the antenna type assuming all of the elements are identical. Thus, to simplify the algebra, isotropic radiators may be utilized in its derivation. The field of an isotropic radiator located at the origin is

$$E_{\theta 0}=I_0\frac{e^{-j\beta r}}{4\pi r} \tag{9.11}$$

Assume that the elements of the array are uniformly-spaced with a separation distance *d* as illustrated in Figure 9.3. In the far field of the array:

$$\begin{aligned} r_0 &= r \\ r_1 &\approx r-d\cos\theta \\ r_2 &\approx r-2d\cos\theta \\ &\vdots \\ r_{M-1} &\approx r-(M-1)d\cos\theta \end{aligned}$$

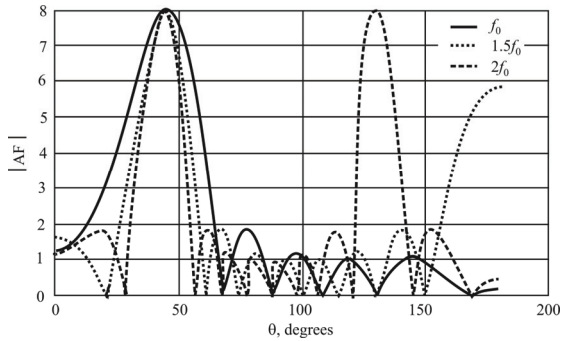


Figure 9.4 Array factor of eight-element time scanned ULA. $d = 0.37\lambda$ at f_0, designed to steer the beam to $\theta_0 = 45^o$ at f_0.

The current magnitudes on the array elements are assumed to be equal and the current on the array element located at the origin is used as the phase reference (zero phase).

$$I_0 = I_0, I_1 = I_0 e^{j\varphi_1}, I_2 = I_0 e^{j\varphi_2}, \cdots, I_{M-1} = I_0 e^{j\varphi_{M-1}} \tag{9.12}$$

The far fields of the individual array elements are, given that $r_0 = r_1 = r_2 \approx r$,

$$\begin{aligned}
E_{\theta 0} &\approx I_0 \frac{e^{-j\beta r}}{4\pi r} = E_0 \\
E_{\theta 1} &\approx I_0 e^{j\varphi_1} \frac{e^{-j\beta(r - d\cos\theta)}}{4\pi r} = E_0 e^{j(\varphi_1 + \beta d\cos\varphi)} \\
E_{\theta 2} &\approx I_0 e^{j\varphi_2} \frac{e^{-j\beta(r - 2d\cos\varphi)}}{4\pi r} = E_0 e^{j(\varphi_2 + 2\beta d\cos\varphi)} \\
&\vdots \\
E_{\theta(M-1)} &\approx I_0 e^{j\varphi_{M-1}} \frac{e^{-j\beta[r-(M-1)d\cos\varphi]}}{4\pi r} = E_0 e^{j[\varphi_{M-1} + (M-1)\beta d\cos\varphi]}
\end{aligned}$$

The overall array far field is found using superposition, since we are assuming a linear medium, as

$$\begin{aligned}
E_\theta &= E_{\theta 0} + E_{\theta 1} + E_{\theta 2} + \ldots + E_{\theta(M-1)} \\
&= E_0[1 + e^{j(\varphi_1 + \beta d\cos\theta)} + e^{j(\varphi_2 + 2\beta d\cos\theta)} + \cdots + e^{j[(\varphi_{M-1} + (M-1)\beta d\cos\varphi)]}] \\
&= E_0[\mathrm{AF}]
\end{aligned} \tag{9.13}$$

where the array factor is

$$\mathrm{AF} = [1 + e^{j(\varphi_1 + \beta d\cos\varphi)} + e^{j(\varphi_2 + 2\beta d\cos\theta)} + \cdots + e^{j[(\varphi_{M-1} + (M-1)\beta d\cos\theta)]}] \tag{9.14}$$

For the ULA the phase progression over the array is linear (increases by a constant value, α, from element to element) and given by

$$\varphi_0 = 0, \quad \varphi_1 = \alpha, \quad \varphi_2 = 2\alpha, \cdots, \varphi_{M-1} = (M-1)\alpha \tag{9.15}$$

Inserting this linear phase progression into (9.14) gives

$$\begin{aligned}\mathrm{AF} &= 1 + e^{j(\alpha+\beta d\cos\theta)} + e^{j2(\alpha+\beta d\cos\theta)} + \cdots + e^{j(M-1)(\alpha+\beta d\cos\theta)]} \\ &= 1 + e^{j\psi} + e^{j2\psi} + \cdots + e^{j(M-1)\psi}\end{aligned} \tag{9.16}$$

where

$$\psi = \alpha + \beta d\cos\varphi \tag{9.17}$$

so that

$$\mathrm{AF} = \sum_{m=0}^{M-1} e^{jm\psi} \tag{9.18}$$

The function ψ is defined as the *array phase function* and is a function of only the element spacing, phase shift, frequency, and elevation angle. Note that

$$(\mathrm{AF})e^{j\psi} = e^{j\psi} + e^{j2\psi} + e^{j3\psi} + \cdots + e^{jM\psi} \tag{9.19}$$

and subtracting (9.16) from (9.19) yields

$$(\mathrm{AF})(e^{j\psi} - 1) = e^{jM\psi} - 1 \tag{9.20}$$

so that

$$\mathrm{AF} = \frac{e^{jM\psi} - 1}{e^{j\psi} - 1} = \frac{e^{jM\frac{\psi}{2}}}{e^{j\frac{\psi}{2}}} \frac{e^{jM\frac{\psi}{2}} - e^{-jM\frac{\psi}{2}}}{e^{j\frac{\psi}{2}} - e^{-j\frac{\psi}{2}}} = e^{j(M-1)\psi/2} \frac{\sin\left(\frac{M\psi}{2}\right)}{\sin\left(\frac{\psi}{2}\right)} \tag{9.21}$$

The complex exponential term in (9.21) appears because the array phase center is shifted relative to the origin. If the position of the array is shifted so that the center of the array is located at the origin, this phase term goes away and the array factor becomes

$$\mathrm{AF} = \frac{\sin\left(\frac{M\psi}{2}\right)}{\sin\left(\frac{\psi}{2}\right)} \tag{9.22}$$

Plots of the absolute value of the array factor, $|\mathrm{AF}|$, versus the array phase function ψ as the number of elements in the array is varied are illustrated in rectangular form in Figure 9.5 and in polar form in Figure 9.6. *Note that these are plots of AF versus* ψ*, not* θ.

Examining Figure 9.5 we can identify some general characteristics of the array factor for a ULA:

- $\mathrm{AF}_{\max} = M$ at $\psi = 0$, which represents the main lobe;
- The total number of nodes is $M - 1$, one main lobe and $M - 2$ sidelobes;
- The width of the main lobe is $4\pi/M$ and the minor lobe width is $2\pi/M$.

The array factor is frequently normalized so that the maximum value for any value of M is unity, yielding

$$\mathrm{AF}_{\mathrm{norm}} = \frac{1}{M}\frac{\sin\left(\frac{M\psi}{2}\right)}{\sin\left(\frac{\psi}{2}\right)} \tag{9.23}$$

The nulls of the array function are found by determining the zeros of the numerator term where the denominator is not simultaneously zero.

$$\sin\left(\frac{M\psi}{2}\right) = 0 \tag{9.24}$$

$$\frac{M\psi}{2} = \pm m\pi, \qquad m = 1, 2, 3, \ldots, \qquad m \neq 0, M, 2M, 3M, \cdots \tag{9.25}$$

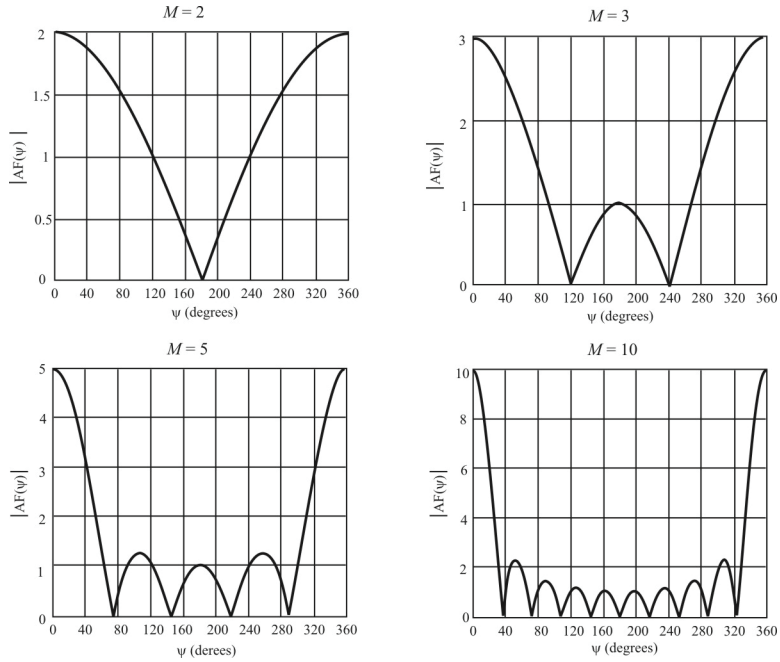


Figure 9.5 Vertical plane array factor for a uniform *M*-element linear array in rectangular coordinates. Horizontal plane array factor is omnidirectional. Note that these are plots versus ψ not θ.

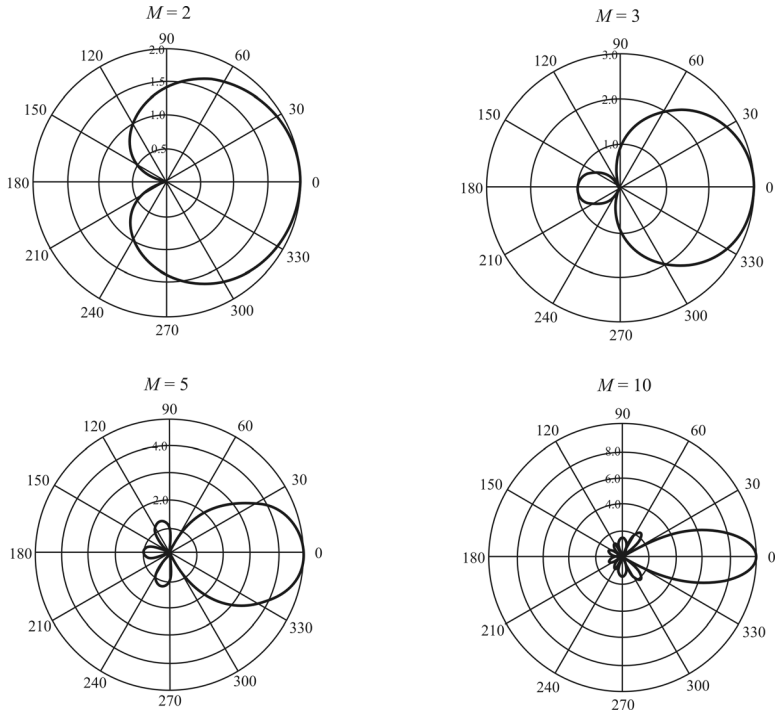


Figure 9.6 Vertical plane array factor for *M*-element ULA in polar form in free space. Horizontal patterns are similar in shape.

$$\varphi_m = \cos^{-1}\left[\frac{\lambda}{2\pi d}\left(-\alpha \pm \frac{2m\pi}{M}\right)\right], \qquad m = 1,2,3,\ldots, \qquad m \neq 0, M, 2M, 3M, \cdots \quad (9.26)$$

The peaks of the array function are found by determining the zeros of the numerator term where the denominator is simultaneously zero.

$$\varphi_m = \cos^{-1}\left[\frac{\lambda}{2\pi d}\left(-\alpha \pm 2m\pi\right)\right], \qquad m = 0,1,2,\ldots \quad (9.27)$$

The $m = 0$ term

$$\theta_m = \cos^{-1}\left[\frac{\lambda\alpha}{2\pi d}\right] \quad (9.28)$$

represents the angle that makes $\psi = 0$ (main lobe).

9.3.1.2 Beam Steering

Consider a general ULA excitation a_m (m = 0, 1, ..., M–1). The corresponding radiation pattern is

$$E_a(z) = \sum_{m=0}^{M-1} a_m z^m \quad (9.29)$$

where

$$z = \exp(j\beta d \cos\theta) \quad (9.30)$$

Now apply a linear phase gradient, ξ, to the array—that is, let the array excitations become a_0, $a_1 \exp(j\xi)$, $a_2 \exp(j2\xi)$, and so forth. The new radiation pattern is

$$\begin{aligned} E_b(z) &= \sum_{m=0}^{M-1} a_m \exp(jm\xi) z^m \\ &= \sum_{m=0}^{M-1} a_m z'^m \end{aligned} \quad (9.31)$$

where

$$z' = \exp[j(\beta d \cos\theta + \xi)] \tag{9.32}$$

Then, if

$$\beta d \cos\theta_b + \xi = \beta d \cos\theta_a \tag{9.33}$$

$$E_b(\cos\varphi_b) = E_a(\cos\varphi_a) \tag{9.34}$$

the radiation pattern keeps the same form but with an angular shift determined by the phase shift ξ. From (9.34):

$$\cos\varphi_b - \cos\varphi_a = -\frac{\xi}{\beta d} \tag{9.35}$$

The right-hand side of (9.35) is proportional to the phase gradient ($d\xi/dx$) along the array.

If the main beam peak is initially at $\theta = 90^o$, broadside to the array, then $\cos\theta_a = 0$, and $\cos\theta_b = \sin\alpha$, where α is the beam deflection angle. In that case

$$\sin\alpha = -\frac{\xi}{\beta d} \tag{9.36}$$

When $d = \lambda/2$, and α is small so that $\sin\alpha \approx \alpha$, then

$$\alpha(\text{radians}) \approx -\frac{\xi(\text{radians})}{\pi}$$

$$\alpha(\text{degrees}) \approx -\frac{\xi(\text{degrees})}{\pi} \tag{9.37}$$

Thus, for a phase shift of 30^o, the deflection angle is $\approx 10^o$. Using (9.36), a phase shift of 90^o for $\lambda/2$ spaced elements gives a deflection angle of 30^o.

The use of electronically controlled phase shifters on each element of the "phased" array allows the beam-pointing direction of an array antenna to be varied rapidly and arbitrarily, without any mechanical movement of the antenna. This is in contrast to a fixed-beam antenna, such as a simple horn-fed parabolic reflector, where the antenna is rotated mechanically to "scan" the beam. Such an antenna has difficulty moving to arbitrary points as well. However, the latter is in general much less expensive.

These properties are extremely valuable in some EW applications. Radars normally depend on moving the antenna either up and down (height finding radars) or around in a circle (azimuth scanning radars). When a phased array is used instead, these "movements" are not physical at all, but electronic, making them much faster. ELINT ES systems can take advantage of these properties as well for much the same reasons.

9.3.1.3 Grating Lobes

We showed in the last section that beams can be steered with the ULA in desired direction by applying weights to antennas to produce a phase gradient. However, if the spacing between the elements are too large (greater than $\lambda/2$) *grating lobes* develop. A grating lobe is an antenna pattern segment pointing in an undesired direction.

The radiation pattern of a ULA with M elements is

$$\left|E_{\mathrm{a}}(z)\right| = \frac{\sin\left(\frac{1}{2}M\beta d\cos\theta_{\mathrm{a}}\right)}{\sin\left(\frac{1}{2}\beta d\cos\theta_{\mathrm{a}}\right)}$$

Now suppose there is a phase gradient, ξ, superimposed. Using (9.33):

$$\left|E_{\mathrm{b}}(z)\right| = \frac{\sin\left[\frac{1}{2}M\left(\beta d\cos\theta_{\mathrm{b}} + \xi\right)\right]}{\sin\left[\frac{1}{2}\left(\beta d\cos\theta_{\mathrm{b}} + \xi\right)\right]} \tag{9.38}$$

The peak of this radiation pattern is at $\theta_{\mathrm{b}} = \theta_{\mathrm{m}}$, where

$$\cos\theta_{\mathrm{m}} = -\frac{\xi}{\beta d} \tag{9.39}$$

Equation (9.38) can therefore be rewritten as

$$\left|E_{\mathrm{b}}(z)\right| = \frac{\sin\left[\frac{1}{2}M\beta d\left(\cos\varphi - \cos\varphi_{\mathrm{m}}\right)\right]}{\sin\left[\frac{1}{2}\beta d\left(\cos\varphi - \cos\varphi_{\mathrm{m}}\right)\right]} \tag{9.40}$$

Grating lobes appear if the denominator of (9.40) becomes zero for some $\theta \neq \theta_{\mathrm{m}}$. This will occur if

$$\cos\varphi - \cos\varphi_{\mathrm{m}} = \pm\frac{\lambda}{d} \tag{9.41}$$

Let the deflection angle from broadside (where $\theta = 90^{\mathrm{o}}$) be denoted by α, so that

$$\theta = 90^{\mathrm{o}} + \alpha \tag{9.42}$$

Equation (9.41) then becomes

$$\sin\alpha_{\mathrm{m}} - \sin\alpha = \pm\frac{\lambda}{d} \tag{9.43}$$

Since the extreme values of $\sin\alpha$ are ± 1, if $\alpha_{\mathrm{m}} > 0$, the magnitude of the left-hand side of (9.43) has a maximum value of

$$\sin\alpha_{\mathrm{m}} + 1 \tag{9.44}$$

Thus, the condition for no grating lobes is

$$\begin{aligned} \frac{\lambda}{d} &> 1 + \sin\alpha_{\mathrm{m}} \\ \frac{d}{\lambda} &< \frac{1}{1+\sin\alpha_{\mathrm{m}}} \end{aligned} \tag{9.45}$$

For example, if the maximum beam steering angle is 30^{o} from broadside, (9.45) requires $d < 2\lambda/3$. If the beam can be steered by 90^{o}, to "end fire," $d < \lambda/2$ is needed to avoid grating lobes.

Nulls

A null of the radiation pattern need not be, but is here taken to mean, zero—no transmission or reception at a particular angle. A null can also mean a minimum of the pattern. Consider the complex variable:

$$z = \exp(j\beta d \cos\varphi) \tag{9.46}$$

As φ varies from 0 to π, z moves along a locus in the Argand diagram, as shown in Figure 9.7. The modulus of z is 1, while its argument is

$$\frac{2\pi d}{\lambda}\cos\varphi$$

Therefore, z varies between $+2\pi d / \lambda$ ($\varphi = 0$) to $-2\pi d / \lambda$ ($\varphi = \pi$) and the locus of z is an arc on the unit circle in the Argand diagram. For a large value of d, z may traverse the unit circle several times as φ varies from 0 to π. If $d = \lambda/2$, the locus of z as φ varies from 0 to π form a single complete circle. If $d << \lambda/2$ the locus of z is a small arc on the unit circle.

9.4 Circular Antenna Arrays

9.4.1 *M*-Element Uniform Circular Array

The M-element UCA is shown in Figure 9.8 ($M = 8$). Define the element signal column vector $\vec{E}$ in which the kth element component E_k consists of quiescent receiver channel noise voltage n_k plus a summation of voltages associated with I external, narrowband interference sources:

$$\vec{E} = \begin{bmatrix} E_1 & E_2 & \cdots & E_K \end{bmatrix}^{\mathrm{T}} \tag{9.47}$$

where

$$E_k = n_k + \sum_{i=1}^{I} A_i e^{ju_i(2k-K-1)} \tag{9.48}$$

$$u_i = \left(\frac{\pi d}{\lambda}\right)\sin\theta_i \tag{9.49}$$

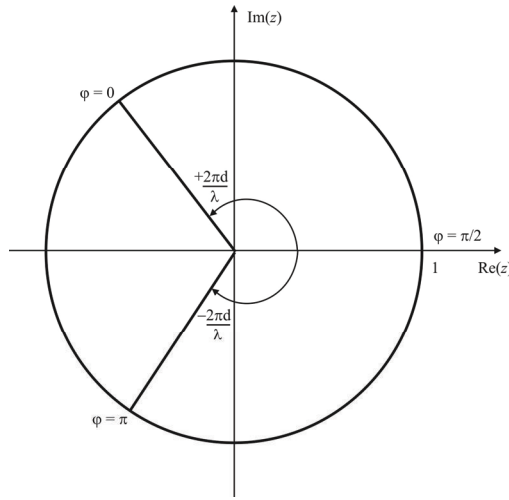


Figure 9.7 Argand diagram for $z = \exp(j\beta d \cos \varphi)$.

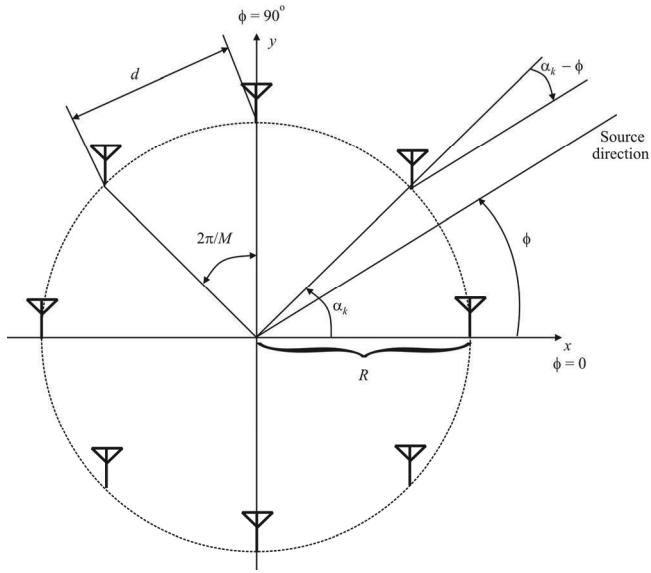


Figure 9.8 $M = 8$ uniform circular array.

We assume that the interference sources are statistically independent; A_i is the element channel voltage amplitude due to the ith source, and θ_i is its azimuth angle from array boresight. It is assumed that a given source will induce equal voltage amplitudes at all of the array elements. Element phase is referenced to the geometric center of the array.

The beam-steering signals shape an antenna beam that is steered in some desired azimuth direction θ_0. For quiescent conditions wherein only receiver noise is present, the adaptive weights will settle to values denoted by the quiescent weight column vector $\vec{W}_{\mathrm{q}}$. The components of this vector are equal to the array weights needed to generate the desired quiescent shaped-beam pattern $G_{\mathrm{q}}(\theta)$. Thus, define $\vec{W}$, as this desired vector, so that

$$\vec{W}_{\mathrm{q}} = \begin{bmatrix} W_{\mathrm{q}1} & W_{\mathrm{q}2} & \cdots & W_{\mathrm{q}K} \end{bmatrix}^{\mathrm{T}} \tag{9.50}$$

where

$$W_{\mathrm{q}k} = a_k e^{-ju_0(2k-K-1)} \tag{9.51}$$

$$u_0 = \left(\frac{\pi d}{\lambda}\right)\sin\theta_0 \tag{9.52}$$

The values of element coefficients a_k are chosen to achieve the desired beam shape and sidelobe levels. The quiescent beam pattern can then be expressed as

$$G_{\mathrm{q}}(\theta) = \vec{S}^{\mathrm{T}}\vec{W}_{\mathrm{q}} = \sum_{k=1}^{K} a_k e^{j(u-u_0)(2k-K-1)} \tag{9.53}$$

where $\vec{S}$ is a column vector representing element signals of unit amplitude. Phase factor u is associated with far-field angle variable θ:

$$\vec{S} = \begin{bmatrix} S_1 & S_2 & \cdots & S_K \end{bmatrix}^{\mathrm{T}} \tag{9.54}$$

where

$$S_k = e^{ju(2k-K-1)} \tag{9.55}$$

$$u = \frac{\pi d}{\lambda}\sin\theta \tag{9.56}$$

Consider the circular antenna configuration in Figure 9.8. For a circular array, it is more convenient to define an element pattern and a phase factor referenced to the center of the circle as opposed to the origin. Assuming the eight-element configuration shown in Figure 9.8, these may be written as

$$\text{phase factor } u_k = \pi \frac{D}{\lambda} \cos \varphi_k \tag{9.57}$$

where

$$\varphi_k = \theta - \psi_k \tag{9.58}$$

and

$$\psi_k = \left(\frac{\pi}{K} \right) (2k - K - 1) \tag{9.59}$$

Here D is the diameter of the circle, and to achieve an element spacing of $\lambda/2$, $D = 2.5629$. The pattern associated with A_k is based upon a measured element pattern. These A_k and u_k must be utilized in the new element signals E_k and S_k [see (9.48) and (9.49)].

A desired quiescent shaped-beam pattern, $G_q(\theta)$, must also be chosen and, for a circular array, it is of interest to select a simple omnipattern. This choice is naturally suited to EW systems because it is normally not known a priori from which direction the SOI will arrive. Therefore, in (9.51) for the W_{qk}, select $a_k = 1$ and $u_0 = 0$ for all elements to set up the omnipattern.

9.5 Planar Arrays

Just as a linear array can be considered as a sampled aperture in one dimension, a planar array can be thought of as a sampled aperture in two dimensions. Much of the previous analysis of planar apertures is therefore appropriate to derive many of the properties of planar arrays, provided they are at least several wavelengths across and are composed of elements that radiate predominantly into one half-space, such as open-ended waveguides, or dipoles in front of a planar reflecting sheet. Typical aperture shapes are rectangular, elliptical, or circular. Figure 9.9 show a rectangular grid lattice.

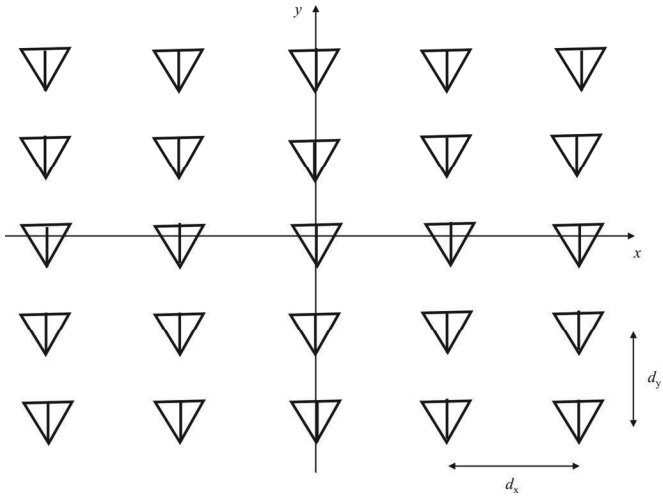


Figure 9.9 A rectangular lattice. Signal propagation direction is along the *z*-axis, that is, out of the page.

Beam steering to a direction (S_{10}, S_{20}) can be accomplished using the appropriate phase distribution. The direction cosines S_1 and S_2 are

$$\begin{aligned} S_1 &= \sin\theta\cos\phi \\ S_2 &= \sin\theta\sin\phi \end{aligned} \tag{9.60}$$

where θ and ϕ are the usual spherical polar coordinates (see Chapter 2) with respect to *z* as the normal to the array aperture (the surface). Let the element lattice spacings in the *x* and *y* directions be d_x and d_y. Beamwidths and first sidelobe levels can generally be found using continuous aperture theory. The interelement phase shifts in the *x* and *y* directions corresponding to these directional cosines are

$$\begin{aligned} \phi_x &= -\beta d_x S_{10} \\ \phi_y &= -\beta d_y S_{20} \end{aligned} \tag{9.61}$$

in order to steer the beam to (θ_0, ϕ_0).

As the beam is scanned, its beamwidth changes as the inverse of the projected aperture. For a rectangular aperture, with an aperture distribution separable in *x* and *y*, and scanned in one of the two principal planes (*x*-*z* or *y*-*z*), the beamwidth varies approximately as $\cos^{-1}\alpha$. For scanning in planes other than these, or for other aperture shapes, the relationship is more complicated.

The conditions for the appearance of grating lobes were analyzed in Section 9.3 for a ULA. Consider a rectangular lattice with interelement spacings d_x and d_y as illustrated in Figure 9.9. The array factor for an unsteered beam is

$$P(S_1, S_2) = \sum_l \sum_m a_{lm} \exp[j\beta(ld_x S_1 + md_y S_2)] \tag{9.62}$$

If the beam is now steered to (S_{10}, S_{20}):

$$P(S_1, S_2) = \sum_l \sum_m a_{lm} \exp(j\beta[ld_x (S_1 - S_{10}) + md_y (S_2 - S_{20})]) \tag{9.63}$$

the main lobe is where $S_1 = S_{10}$, $S_2 = S_{20}$. A grating lobe will appear if

$$\beta d_x (S_1 - S_{10}) = 2\pi n \tag{9.64}$$

or

$$\beta d_y (S_2 - S_{20}) = 2\pi p \tag{9.65}$$

where n and p are negative or positive integers, and are such that

$$S_1 - S_{10} = \frac{\lambda}{d_x} n \tag{9.66}$$

or

$$S_2 - S_{20} = \frac{\lambda}{d_y} p \tag{9.67}$$

As would be expected, a rectangular lattice scanned in a principal plane behaves similarly to a linear array. For arbitrary scan angles Figure 9.10 can be used, which shows the "inverse lattice" for the rectangular grid array, with lattice points spaced by λ/d_x and λ/d_y. The inverse lattice coordinates are $S_1 - S_{10}$ and $S_2 - S_{20}$. The inverse lattice points correspond to the main beam and grating lobes.

For a broadside beam, $S_{10} = S_{20} = 0$, and the coordinates become S_1, S_2. For arbitrary observation angles θ and ϕ, $S_1^2 + S_2^2 < 1$, and a circle of unit radius shown in Figure 9.10 encompasses all possible cases. If this circle does not include any inverse lattice points (other than 0, 0), grating lobes will not appear. Then to make

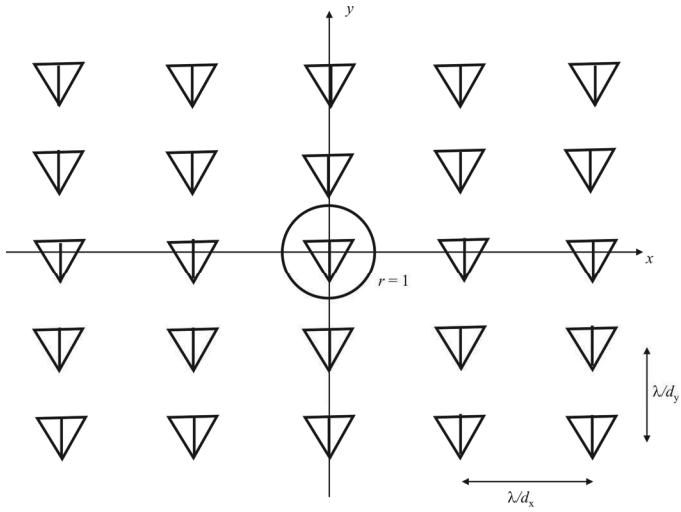


Figure 9.10 Inverse rectangular lattice.

sure the circle does not contain any other elements, we must have $\lambda/d_x > 1$ and $\lambda/d_y > 1$ so the condition for no grating lobes becomes

$$\begin{aligned} d_x &< \lambda \\ d_y &< \lambda \end{aligned} \tag{9.68}$$

for an unsteered beam.

Now let the main beam be scanned to a direction (S_{10}, S_{20}). The unit circle is centered where $S_1 = 0$, $S_2 = 0$, which is now at coordinates ($-S_{10}$, $-S_{20}$) in Figure 9.10. If the circle now includes inverse lattice points other than (0, 0), grating lobes will occur. If the scan angle from broadside is θ_0, the center of the circle is displaced from (0, 0) by a radial distance $\sin\theta$ $(= \sqrt{S_{10}^2 + S_{20}^2})$. Now if the beam can be steered anywhere within a cone about broadside with maximum scan angle θ_0, a circle centered at (0, 0) with radius $1 + \sin\theta_0$, encompasses all the possible cases. The condition for no grating lobes is then

$$d_x < \frac{\lambda}{1 + \sin\theta_0} \tag{9.69}$$

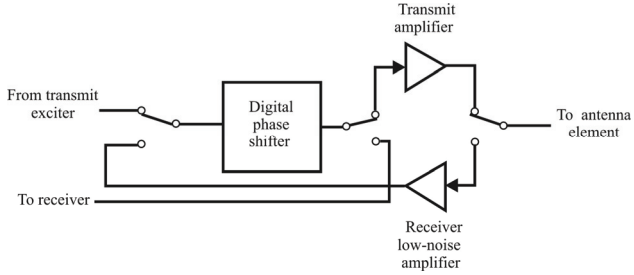


Figure 9.11 Transmit/receive module.

$$d_y < \frac{\lambda}{1+\sin\theta_0} \tag{9.70}$$

A square lattice with spacing d has an area per element, A_{el}, of d^2. If we consider the conditions for no grating lobes, without scanning

$$A_{el}(\text{square lattice}) < \lambda^2 \tag{9.71}$$

For scanning out to θ_0, the area is reduced by the factor $(1+\sin\theta_0)^2$.

9.6 Active Arrays

9.6.1 Transmit/Receive Modules

A single scanning beam for a planar array can be produced using RF phase shifters at each array element. These are usually digitally controlled and can scan the beam in two planes. The array amplitude distribution is usually fixed, with a network distributing the power from a single source. Alternatively, individual elements can have complete integrated transmit and receive modules. The main problem with these approaches is high cost, as the number of phase shifters or modules is equal to the number of array elements. Figure 9.11 shows one possible architecture of a transmit/receive module.

9.6.2 Feed Networks

The feed network of a planar phased array is where the adjustable elements are located. A planar array can be fed in various ways. Figure 9.12 shows three

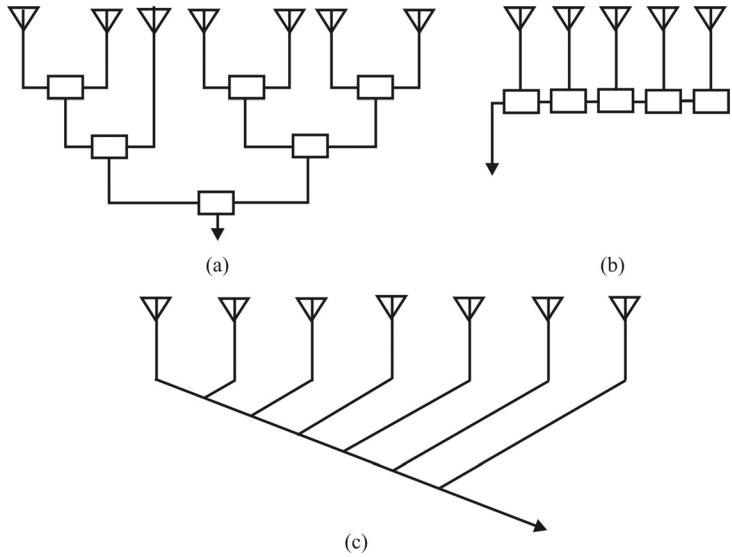


Figure 9.12 Some array feed networks: (a) corporate (parallel) feed network, (b) series feed network, and (c) equal path length, series feed network.

distribution networks for a planar array. The amplitude and phase shifts of the feed lines driving the antennas can be adjusted to control the angle off boresight to which the main beam is steered. We will delve much more deeply into beamforming in Chapter 14, but for now we discuss some of the beamforming aspects of phased arrays.

9.6.3 IF Beamforming

In order to use an IF beamforming scheme, it is necessary to equip each array element with a complete receiver front end (mixer, IF amplifier, and LO drive). Ideally, these front ends will be identical, especially as regards tracking in phase and amplitude.

The fractional bandwidth at the IF is, of course, much higher than the fractional RF bandwidth. This can lead to significant difficulties when wide bandwidths are necessary, which they usually are for EW applications.

9.6.4 Baseband Beamforming

A fully digital baseband beamformer has separate I and Q digital processing for each element in the phased array aperture, thus doubling the resources required to

form the beam. Analog baseband beamforming is also possible. In a system where the number of elements is high, this has obvious problems of cost and complexity. In practice, the full advantages offered by an optimum performance system may not be required, and a hybrid approach with baseband beamforming confined to one dimension may be advantageous.

The biggest advantage of digital baseband beamforming is the flexibility it offers in beam pattern control.

9.6.5 RF Beamforming

The desired amplitude and phase distribution can be formed using a fixed or variable feed network. A standard phased array consists of a number of antenna elements distributed in space, for example, equi-spaced along a straight line, a linear array as discussed above [6–13]. A single beam is formed by feeding the array elements with the amplitudes and phases needed to create the desired beam shape [14]. The beamforming techniques described in this section operate directly at RF.

The configuration in Figure 9.12(a) is referred to as a *corporate feed*. This could be used with phase shifters at each element to provide a steerable single beam; however, this approach has several drawbacks. Two of the more serious are the high cost of phase shifters, and the interconnection problem. In addition to the connections to the array elements themselves, the phase shifters need control lines, and a thin planar structure for the complete antenna assembly is tenuous. A phase shifter is also required for each antenna. The power divisions at the junctions may be fixed or variable—the former is usual as the resource implication is less. In that case, a fixed amplitude distribution results, with corresponding fixed main beam width and sidelobe levels. The phase shifters can then be used to vary the array phase distribution—for example, to steer the beam electronically. The phase shifters are normally composed of a series of discrete valued switched elements [for example, using PIN diodes (see Appendix B)], although continuously variable units are also used.

Figure 9.12(b) shows a series distribution network, where power is progressively coupled out of a feed line into the array elements to give the amplitude distribution. The RF path length between successive elements is one wavelength, λ_0, at a particular frequency, f_0 (or $\lambda_0/2$ plus a phase-reversed connection). At that frequency, the array elements are cophased and a broadside beam is produced. At another frequency f, wavelength λ, there is a phase difference ϕ between elements, given by

$$\phi = 2\pi\left(\frac{\lambda_0}{\lambda} - 1\right) \tag{9.72}$$

For a given array spacing d, the beam is then deflected from broadside by an angle α; using (9.36):

$$\sin\alpha = -\frac{\phi\lambda}{2\pi d} = \frac{\lambda - \lambda_0}{d} \tag{9.73}$$

Depending on the application, this effect can either be a benefit or a drawback. A numerical example can be used to illustrate the effects. Consider an antenna array that has to provide a specific narrow beamwidth. Say that at f_0 (λ_0), the beam points correctly, and this is broadside to the array. Then at $f = f_0 \pm \Delta f / 2$, the beam is deflected from broadside by a small angle α, where

$$|\alpha| = \left(\frac{\lambda}{\lambda_0} - 1\right)\frac{\lambda_0}{d} = \left(-\frac{f_0}{f} - 1\right)\frac{\lambda_0}{d} \approx \frac{\Delta f}{f_0}\frac{\lambda_0}{2d} \tag{9.74}$$

Let d be $\lambda_0/2$ and the beamwidth be 2^o. Then, if $\alpha = \pm 1^o$, there will be a 3 dB reduction in signal at the band edges. Then

$$\frac{\Delta f}{f_0} \approx \frac{\pi}{180} \approx 1.8\%$$

The equal path length, series feed is another form of amplitude distribution network and is shown in Figure 9.12(c). This does not exhibit frequency scanning, but is less compact than a standard series feed as in Figure 9.12(b).

The ease with which both sum and difference beams for tracking applications can be produced in a phased array depends on the beamforming technique that is used. In the case of digital or IF beamforming (Sections 9.6.2 and 9.6.3, as well as later in Section 17.7) good performance can be obtained for both the sum and difference beams since the SNR is established before beamforming takes place. However, in the case of RF beamforming, the most common and simplest way of achieving the sum and difference patterns is to split the array in half (in the plane of the difference pattern) and feed the two halves through a hybrid junction. Unfortunately the resulting difference pattern has high sidelobes due to the large discontinuity in the aperture distribution at the center of the array. A number of techniques exist for improving the difference pattern sidelobes and these involve some increase in design complexity together with a marginal increase in cost.

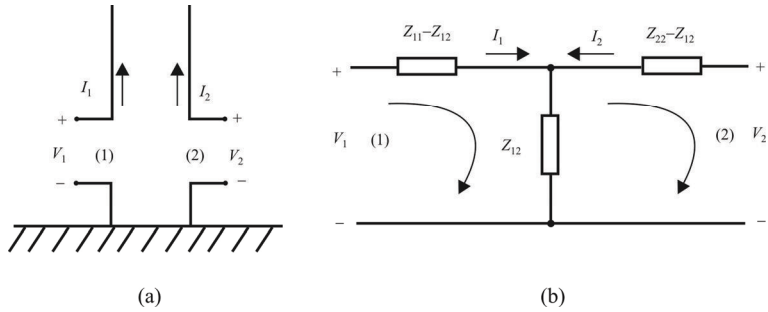


Figure 9.13 Mutual coupling model: (a) two antennas making up a two-port network and (b) equivalent network model.

9.7 Mutual Coupling and Mutual Impedance

Whenever two or more antenna elements are in close proximity to one another, they interact. This interaction carries the appellation *mutual coupling*. The element currents are not the same as they would be for an element in isolation (free space). A model for two elements in Figure 9.13(a) interacting in this way is shown in Figure 9.13(b).

The impedances in Figure 9.13(b) are the port parameters by assuming that the two antennas form a two-port network. The two-port model is given by

$$\begin{aligned} V_1 &= Z_{11}I_1 + Z_{12}I_2 \\ V_2 &= Z_{21}I_1 + Z_{22}I_2 \end{aligned} \tag{9.75}$$

So Z_{11} is found by measuring the voltage and current at port 1 with port 2 open circuited (or, since it is an antenna element, by removing it) since

$$Z_{11} = \left.\frac{V_1}{I_1}\right|_{I_2=0} \tag{9.76}$$

Likewise, Z_{22} is determined by measuring the voltage and current at port 2 with the antenna at port 1 open circuited (or not there) since

$$Z_{22} = \left.\frac{V_2}{I_2}\right|_{I_1=0} \tag{9.77}$$

All of these measurements are, of course, frequency dependent and, in general, complex values.

Assuming that the network is symmetrical (this is certainly true for an array of dipoles, for example), then $Z_{21} = Z_{12}$. Either of these is found in the same way: Z_{12} is found by measuring the voltage at port 1 when it is an open circuit ($I_1 = 0$) caused by a known current source at port 2 because

$$Z_{12} = \left.\frac{V_1}{I_2}\right|_{I_1=0} \tag{9.78}$$

Notice with port 2 open circuited, the impedance seen looking into port 1 is $Z_{11} - Z_{12} + Z_{12} = Z_{11;}$ similarly for port 2.

If the frequency of operation is high enough (say, above 300 MHz or so), measuring the Z-parameters as indicated here is difficult and it may be necessary to resort to the S-parameters where the ports are not opened or shorted but are terminated with their characteristic impedances. Conversion to the Z-parameters is then possible if required, using

$$Z_{11} = Z_0 \frac{(1+S_{11})(1-S_{22})+S_{12}S_{21}}{(1-S_{11})(1-S_{22})-S_{12}S_{21}} \tag{9.79}$$

$$Z_{12} = Z_0 \frac{2S_{12}}{(1-S_{11})(1-S_{22})-S_{12}S_{21}} \tag{9.80}$$

$$Z_{21} = Z_0 \frac{2S_{21}}{(1-S_{11})(1-S_{22})-S_{12}S_{21}} \tag{9.81}$$

$$Z_{22} = Z_0 \frac{(1-S_{11})(1+S_{22})+S_{12}S_{21}}{(1-S_{11})(1-S_{22})-S_{12}S_{21}} \tag{9.82}$$

Creating opens and shorts in circuits at high frequencies is difficult to accomplish, which is why the S-parameters are an attractive alternative.

For an array with M elements, all (or most) or the antenna elements will have an effect on all the others. Hence, the two-port analysis needs to be extended to M ports. This is accomplished in a straightforward way and many textbooks are available that indicate how this is done. Also computer programs are available to assist in this process.

The process described here applies to both transmitting as well as receiving antennas as long as they are constructed with passive elements. Active elements cause nonsymmetrical circuits, and the model shown in Figure 9.13 would not apply.

9.7.1 Mutual Impedance

When designing the feed system of such an antenna array, this mutual interaction among the elements must be considered. The value of the impedance at the input of each antenna element has to be determined taking this interaction into account. Knowing these impedances allows us to determine the input current of each element and currents at the feed ports of the array. Moreover, the feed system can be designed to establish a suitable value of the input impedance of the entire antenna array [15].

Neglecting antenna losses, the radiation impedance Z_S is equal to the input impedance of the antenna Z_S given as a ratio of input voltage V_S and input current I_S as

$$Z_S = \frac{V_S}{I_S} \tag{9.83}$$

The above-defined values of input impedance Z_S are usually provided for basic types of antennas and they typically are valid for antennas placed in *free space*, that is, at satisfactorily long distances from other antennas or objects.

In an antenna array, single antenna elements mutually influence each other and the impedance of each element depends on the types and placement of elements in the surrounding region. The current I is assumed to be excited in the antenna in some way. The current I is characterized by the current I_S and by a function of the current distribution. The antenna radiates and creates a given electric field E in its surroundings, which is proportional to the current on the antenna. A certain field intensity E_t appears even on the surface of the radiating antenna. At the same time, the antenna acts as a receiving antenna, and the received signal results in a voltage at the input port of the antenna. Voltage V_S in (9.83) can be therefore considered as a voltage produced on the antenna by the receipt of its own radiation. This radiation is proportional to the current magnitude I_S at its input and the quantity Z_S (self-impedance) plays the role of the proportionality constant.

This can be simply applied to the whole antenna array. Even in the array, there is a certain field intensity E_t on the surface of each antenna element. This field is created not only by the self-radiation of the antenna element but also by the radiation of the other elements.

In Figure 9.14, an antenna array is depicted consisting of M elements (dipoles), which are fed by currents I_{Si} on their input terminals. In analogy to (9.83), input voltages to the antenna elements are given by the set of equations

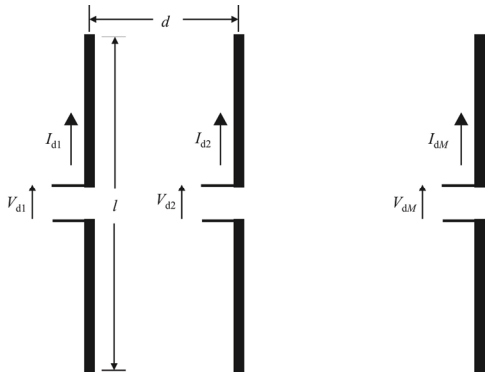


Figure 9.14 Array of *M* dipoles.

$$\begin{aligned} V_1 &= Z_{11}I_1 + Z_{12}I_2 + \ldots Z_{1M}I_M \\ V_2 &= Z_{21}I_1 + Z_{22}I_2 + \ldots Z_{2M}I_M \\ &\vdots \\ V_M &= Z_{M1}I_1 + Z_{M2}I_2 + \ldots Z_{MM}I_M \end{aligned} \tag{9.84}$$

Impedance coefficients Z_{jk} in (9.84) express the mutual coupling between *j*th and *k*th element and are called *mutual impedances* (when $j \neq k$). A mutual impedance is a complex quantity, the real part of which represents a resistance and the imaginary part represents a reactance. A mutual impedance satisfies $Z_{jk} = Z_{kj}$. Their magnitudes depend on the shape, dimensions, and the mutual positions of antenna elements, as well as their current distributions.

Coefficient Z_{jj} is called the *self-impedance*, which is the ratio of the voltage to the current at the input of *j*th antenna element in the array, and is equal to the radiation impedance the *j*th element in free space.

The magnitude of the mutual impedance between two elements depends on the antenna dimensions and on the distance between antenna elements. The real and imaginary components of the mutual impedance can be both positive and negative, and their maximum values decrease as the distance between elements, *d*, increases. The influence of very distant elements is negligible. Typical characteristics of the components of the mutual impedance $Z_{jk} = R_{jk} + jX_{jk}$ on the distance *d* (multiplied by wave-number $\beta = 2\pi/\lambda$) between two parallel $\lambda/2$ dipoles are depicted in Figure 9.15.

The magnetic field intensity $H_\phi(z)$ on the surface of a relatively thin wire is primarily determined by the current $I(\zeta)$ in the immediate vicinity of the element *dz*. Applying Ampere's law (2.21), we get

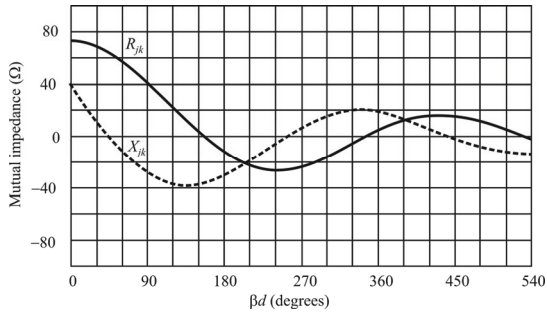


Figure 9.15 Mutual impedance between two λ/2 dipoles at distance d (βd in degrees).

$$H_{\phi}(z) = \frac{I(\zeta)}{2\pi a_0} \tag{9.85}$$

Electric field intensity $E_z(z)$ on the surface of the antenna wire can be expressed in term of vector potential $A_z(z)$ given a known current distribution $I(z)$ from (5.5) as

$$\vec{A}(z) = A_z(z)\vec{u}_z \tag{9.86}$$

with magnitude

$$A_z(z) = \frac{\mu}{4\pi}\int_{-l/2}^{l/2} I(u)\frac{e^{-j\beta r}}{r}du \tag{9.87}$$

where

$$r = \sqrt{(z-\xi)^2 + a_0^2} \tag{9.88}$$

Electric field intensity $E_z(z)$ is given by

$$E_z(z) = -j\omega\left[A_z + \frac{1}{\beta^2}\nabla\nabla\cdot\vec{A}_z\right] = -j\omega\left(A_z + \frac{1}{\beta^2}\frac{\partial^2 A_z}{\partial z^2}\right) \tag{9.89}$$

If the radius of the antenna wire a_0 is significantly smaller than its length l, and if radiation at the ends of the wire is neglected, the Poynting vector has only a radial component given by

$$\Pi = -E_z(z)H_\phi^*(z) \tag{9.90}$$

An element of the wire surface $dS = 2\pi a_0 dz$ then radiates the power

$$dP_S = \Pi 2\pi a_0 dz = -E_z(z)I_z^*(z)dz \tag{9.91}$$

The complex power radiated by the whole antenna P_S is found by integrating (9.91) along the complete length

$$P_S = \int_{-l/2}^{l/2} dP_S = -\int_{-l/2}^{l/2} E_z(z)I_z^*(z)dz \tag{9.92}$$

Components of the radiation impedance Z_S at the input port of the antenna can be found by dividing the power P_S by the squared input current I_S

$$Z_S = -\frac{1}{|I_S|^2}\int_{-l/2}^{l/2} E_z(z)I_z^*(z)dz \tag{9.93}$$

The approach described above is conceptually simple, but in solving any given problem, evaluation of the integrals can be problematic. The main advantage of this approach is that is allows us to compute impedances of elements in an antenna array.

Consider the set of parallel dipoles shown in Figure 9.14. The total electric field intensity $E_{z1}(z)$ on the surface of the first element is equal to the sum of contributions of all the elements of the array

$$E_{z1}(z) = E_{11}(z) + E_{12}(z) + \ldots + E_{1M}(z) \tag{9.94}$$

Power P_{S1} radiated by the first element is found by substituting (9.94) into (9.92). This yields

$$P_{S1} = -\int_{-l/2}^{l/2} \left[E_{11}(z)I_1^*(z) + E_{12}(z)I_1^*(z) + \ldots + E_{1M}(z)I_1^*(z)\right]dz$$

$$= I_{\mathrm{d1}}^{*}\left[\begin{array}{c} -I_{\mathrm{d1}}\int\limits_{-l/2}^{l/2}\frac{E_{11}(z)}{I_{\mathrm{d1}}}\frac{I_1^{*}(z)}{I_{1\mathrm{d}}^{*}}dz - I_{\mathrm{d2}}\int\limits_{-l/2}^{l/2}\frac{E_{12}(z)}{I_{\mathrm{d2}}}\frac{I_1^{*}(z)}{I_{1\mathrm{d}}^{*}}dz - \ldots \\ -I_{\mathrm{dM}}\int\limits_{-l/2}^{l/2}\frac{E_{1M}(z)}{I_{\mathrm{dM}}}\frac{I_1^{*}(z)}{I_{1\mathrm{d}}^{*}}dz \end{array}\right] \tag{9.95}$$

The integrals

$$Z_{jk} = \int\limits_{-l/2}^{l/2}\frac{E_{jk}(z)}{I_{\mathrm{d}k}}\frac{I_j^{*}(z)}{I_{\mathrm{d}j}^{*}}dz \tag{9.96}$$

are the mutual impedances with units of ohms. The integrals (9.96) do not depend on the magnitudes of currents in the antenna elements because the field intensity E_{jk} excited on the surface of the jth element by the radiation of the kth element is proportional to the current $I_{\mathrm{d}k}$ in the kth element of the antenna array.

For parallel dipoles, results of the integration in (9.96) are expressed in the form of integral sine and integral cosine. In order to get numerical values of mutual impedance components Z_{jk}, numerical computation is usually required.

Dependence of the magnitude of the components of mutual impedance on the distance between dipoles was discussed above and shown in Figure 9.15 for the couple of parallel symmetric dipoles with a standing current. Negative values of the mutual impedances R_{jk} and X_{jk} express the fact that for a given spatial arrangement, radiation of kth element reduces the magnitude of real or imaginary part of the power $P_{\mathrm{S}j}$ which is radiated by jth element of the antenna array. Decreasing magnitude of mutual impedance for long distances corresponds to the decrease of the electric field intensity E_{jk} for increasing distance.

Exploiting (9.96), self-impedance of a stand-alone antenna element Z_{jj} can be computed if the distance between elements d is set to the value of the radius of the antenna wire a_0. This approach corresponds to the situation described when deriving (9.93): we considered electric field intensity on the surface of jth antenna wire E_{jj} excited by a current flowing through the axis of the same wire.

The method of induced electromotive forces enables us to compute the radiated power, and consequently the values of mutual impedances even in the case of nonparallel antenna wires. Then we have to take into account a different spatial orientation of vectors of electric field intensity E_{jk} with respect to the surface of the jth antenna wire. Decomposing vector E_{jk} to respective components and exploiting the more general expression for the power contribution dP_{S}, we obtain more complicated relations. That way, coupling between crossed wires of a turnstile antenna and other situations can be solved.

For a z-directed dipole the induced-EMF method leads to the integral

$$v_{A,2} \approx -\frac{1}{I_0}\int_{-l/2}^{l/2} E_{z,21}(z) I_2(z)\,dz \tag{9.97}$$

where $v_{A,2}$ is the voltage induced at the terminals of antenna 2 due to antenna 1, with antenna 1 excited by an input current of $i_{A,1}$ [see Figure 9.13(a)]. $E_{z,21}$ is the field radiated by antenna 1 at the location of antenna 2 if antenna 2 were absent and $I_2(z)$ is an equivalent current model for antenna 2 with input current I_0. The mutual impedance is $Z_{A,21} = v_{A,2}/ i_{A,1}$. Using the sinusoidal current model, $E_{z,21}$ can be approximated and the integral for the coupling evaluated. For more accurate results, the MoM with a thin-wire integral equation can be used.

Characterizing mutual coupling for complex array antennas requires numerical models based on surface integral equations, volume integral equations, the finite-difference time-domain method, the finite element method, or other numerical algorithms. The same simulation can be used to obtain the embedded element patterns. Modeling an array with many elements can be computationally challenging, due to the large number of required field or current sample points, mesh elements, or degrees of freedom in the simulation.

For very large arrays, a periodic boundary condition can be implemented, so that only one unit cell in the array must be modeled numerically. This approach ignores array edge effects.

The general expressions for the mutual impedance between two monopoles shown in Figure 9.16 [16] were given by Jordan and Bailman based on an earlier paper by Brown and King [17]. Those expressions are

$$R_{21} = 30\left\{\begin{array}{l}\sin\beta l\cos\beta l[\mathrm{Si}(u_2)-\mathrm{Si}(v_2)-2\mathrm{Si}(v_1)+2\mathrm{Si}(u_1)] \\ -\dfrac{\cos 2\beta l}{2}[2\mathrm{Ci}(u_1)-2\mathrm{Ci}(u_0)+2\mathrm{Ci}(v_1)-\mathrm{Ci}(u_2)-\mathrm{Ci}(v_2)] \\ -[\mathrm{Ci}(u_1)-2\mathrm{Ci}(u_0)+\mathrm{Ci}(v_1)]\end{array}\right\} \tag{9.98}$$

$$X_{21} = -30\left\{\begin{array}{l}\sin\beta l\cos\beta l[2\mathrm{Ci}(v_1)-2\mathrm{Ci}(u_1)+\mathrm{Ci}(v_2)-\mathrm{Ci}(u_2)] \\ -\dfrac{\cos 2\beta l}{2}[2\mathrm{Si}(u_1)-2\mathrm{Si}(u_0)+2\mathrm{Si}(v_1)-\mathrm{Si}(u_2)-\mathrm{Si}(v_2)] \\ -[\mathrm{Si}(u_1)-2\mathrm{Si}(u_0)+\mathrm{Si}(v_1)]\end{array}\right\} \tag{9.99}$$

where

$$u_0 = \beta d$$

$$u_1 = \beta(\sqrt{d^2+h^2}-h) \qquad u_2 = \beta(\sqrt{d^2+(2h)^2}+2h)$$

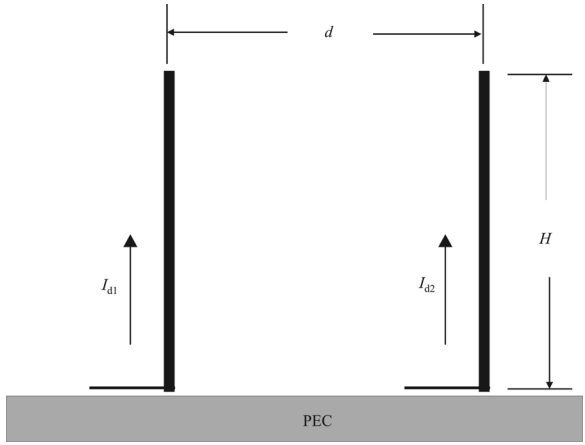


Figure 9.16 Two monopoles mounted on a PEC separated by distance *d* and length *l*.

$$v_1 = \beta(\sqrt{d^2 + h^2} + h) \qquad v_2 = \beta(\sqrt{d^2 + (2h)^2} - 2h)$$

Unfortunately, these equations do not lead to much intuition about antenna behavior; these formulas can only be solved numerically. Doing so, the distance *d* was varied between λ/8 to 3λ/4. The results are shown in Figure 9.17 [18], where the magnitude of the mutual impedance is illustrated and in Figure 9.18 [18] where the phase is shown. We can see that there is reasonable agreement between the *l* = λ/2 curves on these charts and Figure 9.15 since Figure 9.15 is the magnitude of the mutual impedance ($Z^{dp} = 2Z^{mp}$).

9.8 Alternate Technique to Determine the Input Impedance of Antennas and Arrays

Raines developed some useful tools for evaluating antenna impedance parameters without the necessity of calculating them from first principles [19]. This section is largely based on that presentation. This approach uses novel formulas for determining the input conductance or resistance without knowing a priori any self- and mutual impedances. By determining the conductance, the susceptance is also determined because they are Hilbert transform pairs. Similarly, resistance and reactance are also Hilbert transform pairs.

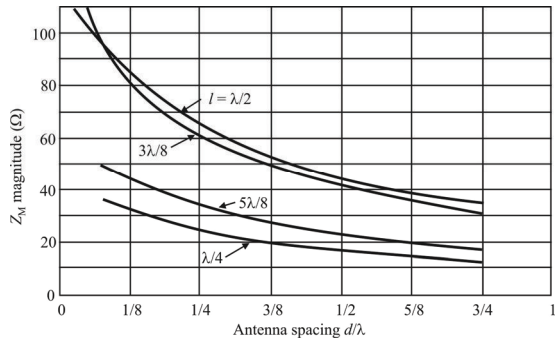


Figure 9.17 Z_M Magnitude. Monopoles have equal length l.

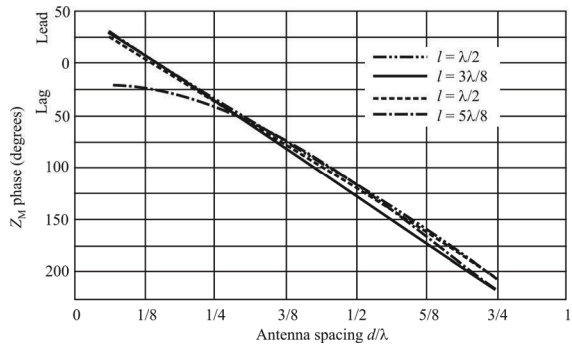


Figure 9.18 Z_M phase.

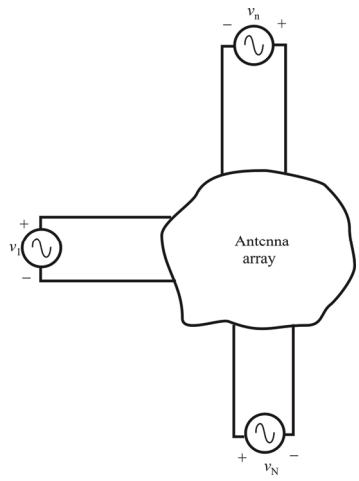


Figure 9.19 Antenna array with N voltage source elements.

9.8.1 Antenna Array

Figure 9.19 shows an antenna array with N elements. Neglecting losses, the total radiated power P and the input voltages V_n are related to the input conductances G_n according to

$$P = \sum_{n=1}^{N} G_n \left| V_n \right|^2 \tag{9.100}$$

In general, the voltages are complex, with nonzero amplitude and phase

$$V_n = A_n e^{j\alpha_n} \tag{9.101}$$

Using (9.101) in (9.100), the latter becomes

$$P = \sum_{n=1}^{N} G_n A_n^2 \tag{9.102}$$

Now, the first derivative of P with respect to the voltage amplitude at the input of the mth element is found from (9.102) as

$$\frac{\partial P}{\partial A_m} = \sum_{n=1}^{N} \left(\frac{\partial G_n}{\partial A_m} A_n^2 + 2A_n \frac{\partial A_n}{\partial A_m} G_n \right) \tag{9.103}$$

The first term on the right in (9.103) is simply the mutual conductance between the *m*th and *n*th elements

$$\frac{\partial G_n}{\partial A_m} = G_{nm} = G_{mn} \tag{9.104}$$

The voltage amplitudes are determined independently by the sources, and therefore are not functions of the other amplitudes so that

$$\frac{\partial A_n}{\partial A_m} = \begin{cases} 0, & n \neq m \\ 1, & n = m \end{cases} \tag{9.105}$$

From (9.104) and (9.105), (9.103) becomes

$$\frac{\partial P}{\partial A_m} = \sum_{\substack{n=1 \\ n \neq m}}^{N} G_{nm} A_n^2 + 2A_m G_{nm} \tag{9.106}$$

Using (9.106), the second derivative of P with respect to A_m is

$$\frac{\partial^2 P}{\partial A_m^2} = \sum_{\substack{n=1, \\ n \neq m}}^{N} \left(\frac{\partial G_{nm}}{\partial A_m} A_n^2 + 2A_n \frac{\partial A_n}{\partial A_m} G_{nm} \right) + 2G_m \tag{9.107}$$

The mutual conductance G_{nm} depends only on the geometry of the array and the wavelength. It is independent of voltage. That is

$$\frac{\partial G_{nm}}{\partial A_m} = 0 \tag{9.108}$$

Using (9.105) and (9.108), (9.107) simplifies to

$$\frac{\partial^2 P}{\partial A_m^2} = 2G_m \tag{9.109}$$

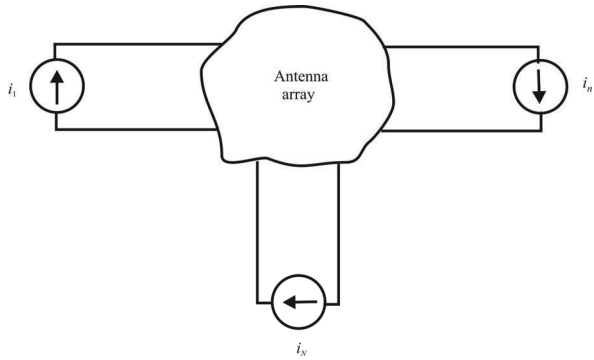


Figure 9.20 Antenna array with current sources.

so

$$G_m = \frac{1}{2}\frac{\partial^2 P}{\partial A_m^2} \tag{9.110}$$

Equation (9.110) indicates that if the amplitude A_m of the mth voltage source is perturbed, then the more the total radiated power P is affected by this perturbation, the greater the input conductance G_m must be.

9.8.2 Formula for Antenna Array with Current Sources

Similar equations for the input resistance may be derived if the excitations are current sources instead of voltage sources, as shown in Figure 9.20.

The total radiated power P is related to the current sources I_n by the input resistance R_n as

$$P = \sum_{n=1}^{N} R_n \left| I_n \right|^2 \tag{9.111}$$

The currents are complex

$$I_n = B_n e^{j\beta_n} \tag{9.112}$$

Following steps analogous to those in the previous section, the input resistance to the mth antenna in the array is

$$R_m = \frac{1}{2}\frac{\partial^2 P}{\partial B_m^2} \tag{9.113}$$

9.8.3 Antenna Susceptance and Reactance

It has long been known that the real and imaginary parts of impedance, admittance, and all realizable physical phenomena are not independent. One uniquely determines the other. For example, susceptance and conductance determine each other, according to

$$B(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{G(u)}{\omega - u}du \tag{9.114}$$

$$G(\omega) = -\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{B(u)}{\omega - u}du \tag{9.115}$$

Equations (9.114) and (9.115) are called Hilbert transform pairs. To make practical use of those equations, B or G must be known over a wide frequency range. Reactance and resistance are similarly related by Hilbert transforms

$$X(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{R(u)}{\omega - u}du \tag{9.116}$$

$$R(\omega) = -\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{X(u)}{\omega - u}du \tag{9.117}$$

In principle, (9.110) and (9.114) could be used to determine the antenna input susceptance B. Similarly, (9.113) and (9.116) could be used to determine the antenna input reactance X. This mathematical formality is usually not necessary, however. Antennas can often be regarded as lengths of waveguide or transmission line, and this point of view provides a good approximation to the input susceptance or reactance. The radiation conductance and radiation resistance are the real challenge.

9.8.4 Summary

This section has derived general and powerful new formulas for determining the input conductance (or resistance) for both single antennas and elements in an arbitrarily large antenna array. The formulas avoid the intermediate, and often

tedious or impractical, steps of determining self- and mutual admittances or impedances. It is necessary only to know the total radiated power and the voltage sources (or current sources).

9.9 Concluding Remarks

We discussed the basic properties of antenna arrays that are relevant to the discussion of EW and other systems in this chapter. In the next chapter we will cover applications of many of these principles to some particular problems. While the concept of beamforming was introduced for antenna arrays in general, we will delve into much more detail in the next chapter on the uniform linear array and uniform circular array for beamforming. We also discuss adaptive arrays at length in Chapter 17.

References

[1] Stutzman, W. L., and G. A. Thiele, *Antenna Theory and Design,* New York: Wiley, 1981, Sect. 7.10.

[2] Johnson, D. H., and D. E. Dudgeon, *Array Signal Processing,* Upper Saddle River, NJ: Prentice-Hall, 1993.

[3] Torrieri, D. J., *Principles of Secure Communication Systems,* Norwood, MA: Artech House, 1992, pp. 367–462.

[4] *Reference Data for Radio Engineers,* 6th ed., Indianapolis: Howard W. Sams & Co, 1975, pp. 27-28–27-49.

[5] Stutzman, W. L. and G. A. Thiele, *Antenna Theory and Design,* New York: John Wiley & Sons, Inc., 1981, pp. 160–161.

[6] Godara, L. C., "Application of Antenna Arrays to Mobile Communications, Part II: Beam-Forming and Direction-of-Arrival Considerations," *Proceedings of the IEEE,* Vol. 85, No. 8, August 1997, pp. 1195–1245.

[7] Applebaum, S. P., "Adaptive Arrays," *IEEE Transactions on Antennas and Propagation*, Vol. AP-24, No. 5, September 1976, pp. 585–599.

[8] Applebaum, S. P., and D. J. Chapman, "Adaptive Arrays with Main Beam Constraints," *IEEE Transactions on Antennas and Propagation,* Vol. AP-24, No. 5, September 1976, pp. 650–662.

[9] Paulraj, A. J., and C. B. Papadias, "Space-Time Processing for Wireless Communications," *IEEE Signal Processing Magazine,* November 1997, pp. 49–89.

[10] Papadias, C. B., and A. J. Paulraj, "Space-Time Signal Processing for Wireless Communications: A Survey," *First IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications*, April 16–18, 1997, pp. 285–289.

[11] Gabriel, W. F., "Adaptive Arrays—An Introduction," *Proceedings of the IEEE,* Vol. 64, No. 2, February 1976, pp. 239–272.

[12] Mailloux, R. J., "Phased Array Theory and Technology, *Proceedings of the IEEE*, Vol. 79, No. 3, March 1982, pp. 246–291.

[13] Sarkar, T. K., J. Koh, R. Adve, R. A. Schneible, M. C. Wicks, S. Choi, and M. S. Palma, "A Pragmatic Approach to Adaptive Antennas," *IEEE Antennas and Propagation Magazine*, Vol. 42, No. 2, April 2000, pp. 39–55.

[14] Van Veen, B. D., and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine,* April 1988, pp. 4–24.

[15] Malherbe, J. A., "Analysis of a Linear Antenna Array Including the Effects of Mutual Coupling," *IEEE Transactions on Education*, Vol. 12, No. 1, February 1989, pp. 29–34.

[16] Jordan, E. C., and K. G. Balmain, *Electromagnetic Waves and Radiating Systems,* Upper Saddle River, NJ: Prentice-Hall, 1968, Ch. 14.

[17] Brown, G. H., and R. King, "High Frequency Models in Antenna Investigations," *Proceedings IRE*, Vol. 22, No. 457, 1934.

[18] Jordan, E. C., and K. G. Balmain, *Electromagnetic Waves and Radiating Systems,* Upper Saddle River, NJ: Prentice-Hall, 1968, p. 541.

[19] Raines, J. K., "Powerful New Formulas for Input Impedance of Antennas and Arrays," *Microwave Journal*, January 2008, pp. 134–139.

Chapter 10

EW Applications of Antenna Arrays

10.1 Introduction

There are many EW applications of antenna arrays—the list is virtually endless. We will consider several such applications in this chapter, but the reader should remember that these are but examples—with a few exceptions, some of the more commonly encountered uses.

We begin with a discussion of circular arrays and their many configurations. For the most part, these arrays are used for localization of the source of emitted radiation. Next we consider beamforming briefly (briefly here because Chapter 19 is devoted to this subject). Lastly, we cover at some length the application of antenna arrays to the problem of intercepting radar signals using a monopulse receiver with its associated antenna array.

10.2 Circular Arrays

10.2.1 Circularly Disposed Antenna Array

The *circularly disposed antenna array* (CDAA) is a circular arrangement of wire antennas, typically monopoles or dipoles, or one of their derivatives (such as a sleeve monopole for broader bandwidth coverage). As discussed in Chapter 6, sleeved monopoles have considerably more bandwidth than regular, unsleeved monopoles. In a typical CDAA architecture, there is more than one ring to broaden the frequency coverage.

One application of the CDAA is for direction finding, particularly in the HF frequency range. One such antenna is the Wullenweber array illustrated in Figure 10.1. This antenna existed for several years at the University of Illinois [1] for ionospheric research purposes and it covered 2–20 MHz. A narrow beam was formed by a network of fixed phase shifters that were mechanically scanned.

Figure 10.1 Wullenweber array. (Source: [1], © University of Illinois. Reprinted with permission.)

A picture of such a mechanical goniometer is shown in Figure 10.2 [2]. The associated phase network is sketched in Figure 10.3 [2]. As the goniometer is commuted, a set of antennas is formed with the delays as shown. This produces a radiation pattern as shown in Figure 10.3 that has a null in the direction of the axis of symmetry. If a signal appears from that direction there is a noticeable null between two substantial energy lobes. The null points in the direction of the transmitter, or at least in the direction of arrival of the signal at the receiver.

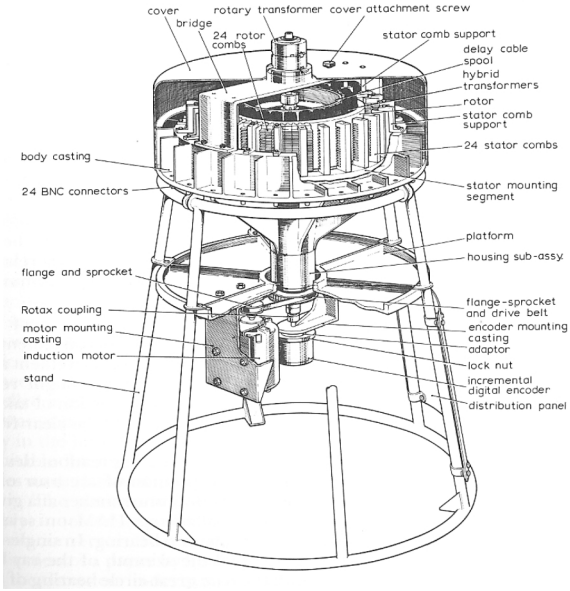


Figure 10.2 Goniometer. (Source: [2], © Peregrinus, 1991. Reprinted with permission.]

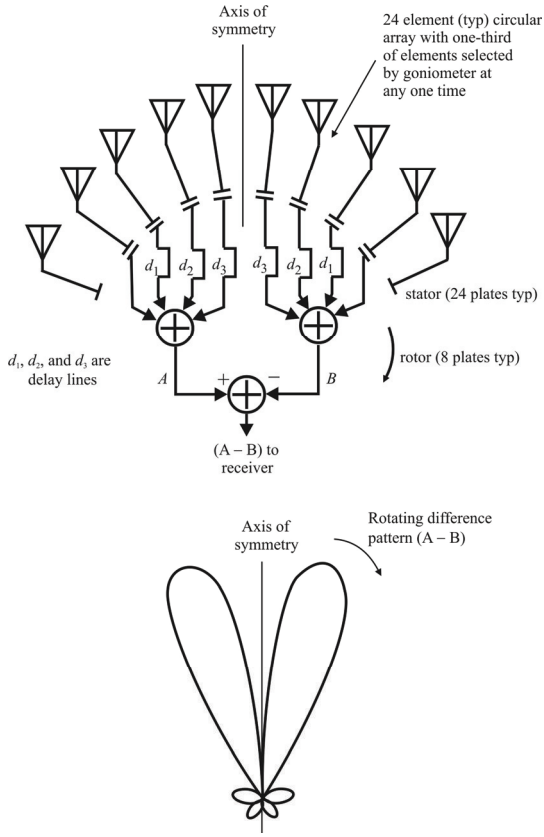


Figure 10.3 Goniometer phase network.

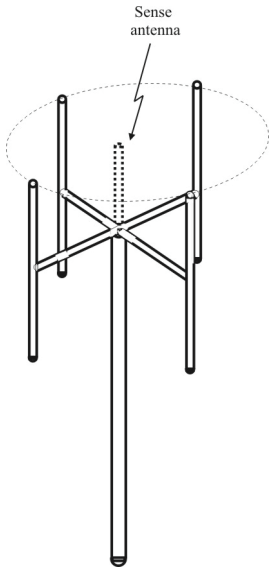


Figure 10.4 Circular Adcock array. (Source: [3], © Artech House, 2008. Reprinted with permission.)

10.2.2 Adcock Antenna Array

The Adcock antenna array consists of a pair of monopoles or dipoles, or one of their variations, that are equidistant from each other in the form of a circular array (see Figure 10.4 [3]). Figure 10.5 shows one of these sets of pairs [4]. The pattern formed by this pair of antennas is illustrated in Figure 10.6 [5]. Two signals impinging on this array at two different angles, φ_1 and φ_2, will cause responses with different amplitudes at the output of the antennas.

The amplitude of the output of this pair of antennas is given by

$$A_v = 2\sin\left(\frac{2\pi R}{\lambda}\sin\theta\cos\phi\right) \tag{10.1}$$

where θ is the elevation angle of the signal (measured from zenith), ϕ is the azimuth angle of arrival, R is the radius of the circle formed by the two pairs of antennas, and λ is the wavelength of the signal.

The antennas connected in this way find the vector difference of the received signal at each antenna so that there is only one output from the pair of antennas.

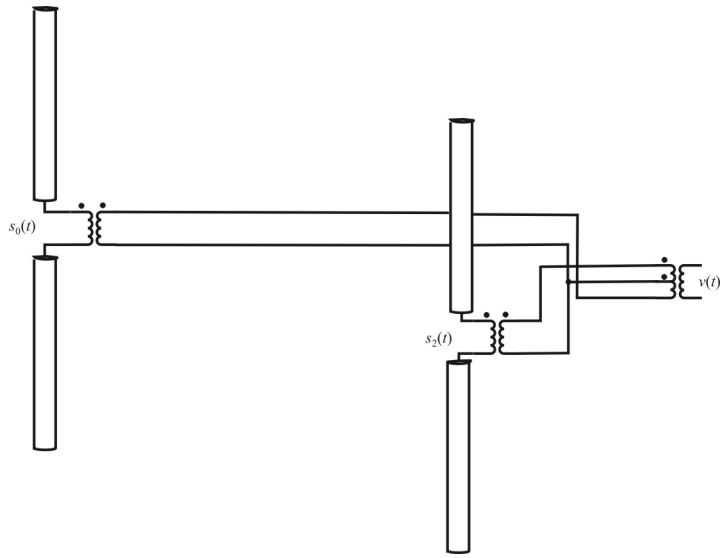


Figure 10.5 Adcock antenna array configuration (Source: [4] © Artech House 2008. Reprinted with permission.)

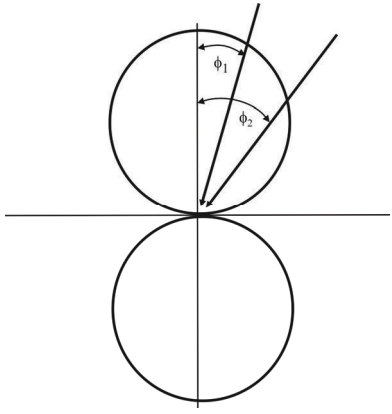


Figure 10.6 Amplitude response pattern for the Adcock antenna configuration shown in Figure 10.5. (Source: [5] © Artech House 2008. Reprinted with permission.)

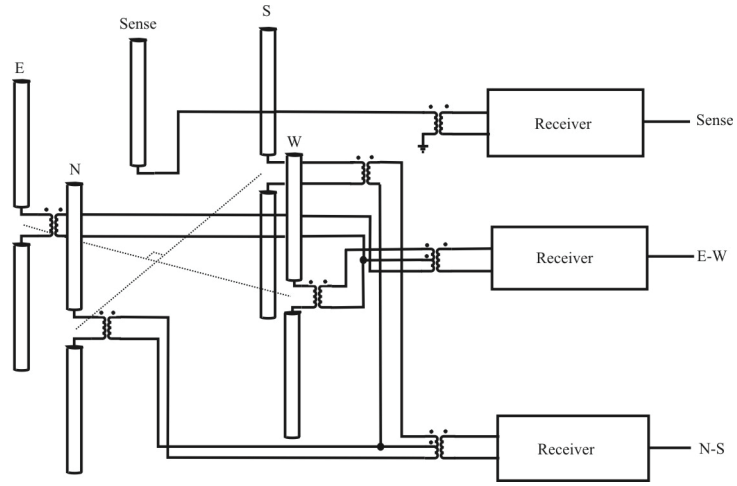


Figure 10.7 Watson-Watt amplitude DF antenna array. (Source: [6] © Artech House 2008. Reprinted with permission.)

Two of these pairs are colocated but perpendicularly oriented to produce what can be referred to as the N-S (North-South) and E-W (East-West) signals that will then be passed to the receiver. In the receiver, the bearing angle can then be computed by taking the arctangent of the ratio of the N-S to E-W signal. This is the Watson-Watt *direction finding* (DF) configuration shown in Figure 10.7 [6]. A sense antenna is included as shown to resolve the natural ambiguity that arises. (The sense antenna in actuality need not be a separate antenna as shown, but can be virtually simulated by summing the output of all four antennas. This would save some complexity in the antenna, but does not eliminate the requirement for the third receiver.)

Perhaps the most prolific use of the Adcock antenna array is in the Watson-Watt amplitude comparison direction finder, depicted in Figure 10.7.

10.2.3 Interferometer Direction Finding

Phase comparison DF systems determine the *direction of arrival* (DOA) of signals impinging on the antenna array by direct phase comparison of the subject signal received by separate, physically displaced antennas [7]. Figure 10.8 shows two antennas so structured for phase interferometry with baseline D and an incident signal at an angle ϕ relative to the boresight z-y plane [8]. Although not explicitly depicted in Figure 10.8, the elevation angle of arrival, measured from the zenith

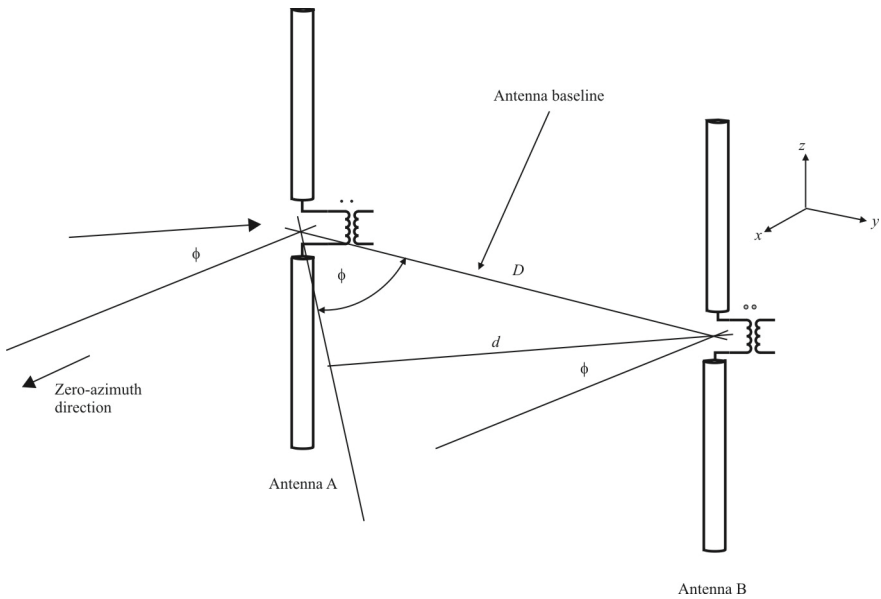


Figure 10.8 Geometry of interferometer baseline. (Source: [8] © Artech House 2008. Reprinted with permission.)

(+*z*-direction) is θ. The propagating wavefront is received at antenna A first and travels an additional distance $d = D \sin \phi \sin \theta$ to antenna B. During the time required to travel the distance between these antennas, the phase of the signal received at antenna A changes by an amount δ given by

$$\delta = \frac{2\pi D}{\lambda} \sin \phi \sin \theta \tag{10.2}$$

The azimuthal AOA[1] is given by

$$\phi = \sin^{-1}\left[\frac{\delta\lambda}{2\pi D \sin \theta}\right] \tag{10.3}$$

where δ is measured and λ, *D*, and θ are known. If the elevation *angle of arrival* (AOA) θ is unknown, a value must be assumed if a single baseline is being used (such is frequently the case for ground-based EW systems or airborne EW systems where the altitude is known and the Earth is assumed to be approximately round). However, if a dual (or more) baseline system is used as illustrated in Figure 10.9 [9] (which depicts a triple baseline system), both azimuth and elevation AOAs may be independently determined. For a two-orthogonal baseline configuration, the total phase delays across the two baselines provide complete DOA information through the expressions

$$\phi = \tan^{-1}\frac{\delta_1}{\delta_2} \tag{10.4}$$

for the azimuth AOA and

$$\theta = 90 - \sin^{-1}\frac{\sqrt{\delta_1^2 + \delta_2^2}}{2\pi D} \tag{10.5}$$

for the elevation AOA where

δ_1 is the phase delay across baseline 1.
δ_2 is the phase delay across baseline 2.

[1] AOA is the angle of arrival, LOB is the line of bearing, and LOP is the line of position. While not strictly equivalent, for our purposes they all represent the same concept and we will use them interchangeably.

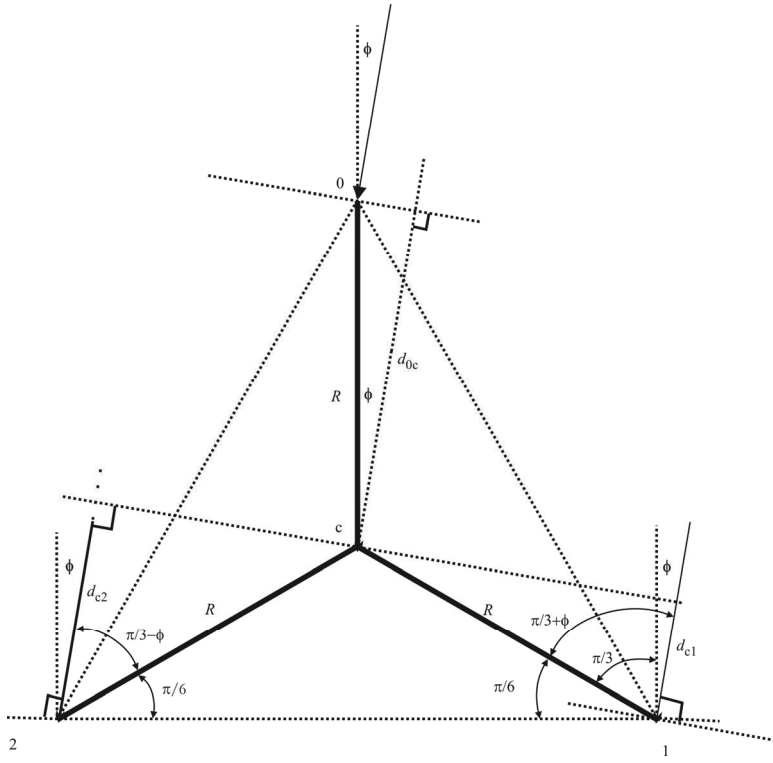


Figure 10.9 Three-antenna interferometer. (Source: [9] © Artech House 2008. Reprinted with permission.)

D is the baseline spacing, assumed the same for each baseline.
λ is the signal wavelength.

and the azimuth angle ϕ is referenced to baseline 2. (For a more in-depth treatment of the four-channel interferometer see [10].)

For the triple-channel interferometer shown in Figure 10.9 [11], the distance traveled by a signal impinging on the array from (ϕ, θ), between antenna 0 and the center of the array, given by d_{0c}, is given by

$$d_{0c} = R\sin\theta\cos\phi \tag{10.6}$$

The sin θ term is required to reflect the signal perpendicular to the x-y plane of the antenna array (θ is the angle from zenith). The phase difference corresponding to this distance is given by

$$\begin{aligned}\varphi_{0c} &= 2\pi f \Delta t_{0c} \\ &= 2\pi \frac{c}{\lambda}\Delta t_{0c}\end{aligned} \tag{10.7}$$

where Δt_{0c} is the time it takes to traverse this distance. $d_{0c} = c\Delta t_{0c}$ however, so

$$\begin{aligned}\varphi_{0c} &= 2\pi \frac{d_{0c}}{\lambda} \\ &= \frac{2\pi R}{\lambda}\sin\theta\cos\phi\end{aligned} \tag{10.8}$$

Likewise, the phase differences from the other two antennas to the center of the array can be determined as

$$\varphi_{c1} = \frac{2\pi R}{\lambda}\sin\theta\cos\left(\phi + \frac{\pi}{3}\right) \tag{10.9}$$

$$\varphi_{c2} = \frac{2\pi R}{\lambda}\sin\theta\cos\left(\frac{\pi}{3} - \phi\right) \tag{10.10}$$

The signals at each of the antennas can thus be represented in terms of the azimuth and elevation AOAs as

$$s_0(t) = s(t)\sin\theta e^{j\frac{2\pi R}{\lambda}\cos\theta\cos\phi} \tag{10.11}$$

$$s_1(t) = s(t)\sin\theta e^{-j\frac{2\pi R}{\lambda}\cos\theta\cos\left(\frac{\pi}{3}+\phi\right)} \tag{10.12}$$

$$s_2(t) = s(t)\sin\theta e^{-j\frac{2\pi R}{\lambda}\cos\theta\cos\left(\frac{\pi}{3}-\phi\right)} \tag{10.13}$$

We determine the azimuth and elevation AOAs by first defining

$$\Delta_1 = \varphi_{c2} - \varphi_{c1} \tag{10.14}$$

so that

$$\begin{aligned}\Delta_1 &= -\frac{2\pi}{\lambda}R\sin\theta\left[\cos\left(\frac{\pi}{3}-\phi\right)+\cos\left(\frac{\pi}{3}+\phi\right)\right] \\ &= \frac{2\pi}{\lambda}R\sin\theta\cos\phi\end{aligned} \tag{10.15}$$

Likewise, by defining

$$\Delta_2 = \phi_{0c} + \phi_{c1} \tag{10.16}$$

we get

$$\Delta_2 = \frac{2\pi}{\lambda}R\sin\theta\left(\frac{1}{2}\cos\phi+\frac{\sqrt{3}}{2}\sin\phi\right) \tag{10.17}$$

and the azimuth AOA is given by

$$\phi = \tan^{-1}\frac{\varphi_{0c}}{\Delta_1} \tag{10.18}$$

The elevation AOA from zenith is determined by expanding

$$\left(\frac{\sqrt{3}}{2}\Delta_1\right)^2 + \left(\Delta_2 - \frac{1}{2}\Delta_1\right)^2$$

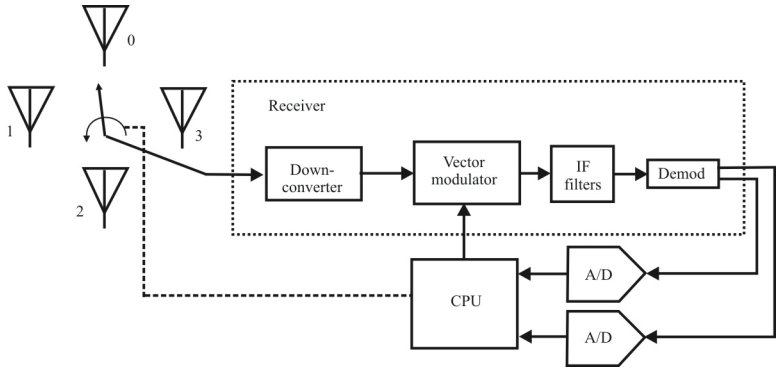


Figure 10.10 Block diagram of a single-channel interferometer DF system based on the pseudo-Doppler technique.

yielding

$$\theta = 90 - \cos^{-1}\sqrt{\frac{4}{3}\left(\frac{\lambda}{2\pi R}\right)^2\left[\left(\frac{\sqrt{3}}{2}\Delta_1\right)^2 + \left(\Delta_2 - \frac{1}{2}\Delta_1\right)^2\right]} \tag{10.19}$$

10.2.4 Pseudo-Doppler Direction Finder

In this section we describe an application of a circular array of vertical antennas for direction finding using a method called *pseudo-Doppler* [12]. A circular array of three (or more) omni-directional (in the horizontal plane) antennas (i.e., dipoles or monopoles) is employed. For specificity, we will describe an array of four elements. A rotating (electrically) commutation switch samples the antennas sequentially. With GAs FET RF switches (see Appendix B), phase matching over temperature and frequency is facilitated, intercept points are high, and the power consumption is extremely low (400 microwatts). The system determines the phase differences of signals at the several antennas and thus is a form of phase interferometer.

10.2.4.1 Block Diagram

A simplified block diagram of the pseudo-Doppler system is shown in Figure 10.10 [12]. The antenna with an SP4T RF switch is shown on the left. The output of the antenna array is sent to the receiver with the IF routed to a *vector modulator*. The FM detected outputs are used for LOB measurements. The

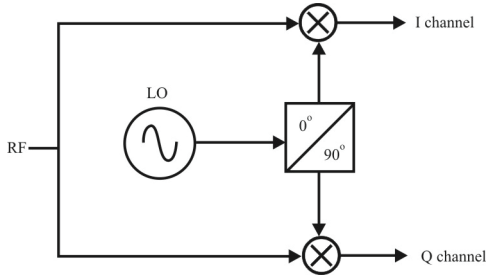


Figure 10.11 Vector modulator.

detected outputs are digitized and passed to a central processing unit (CPU) that controls the entire system and makes the LOB calculations.

A vector modulator, also known as an IQ modulator, divides an RF or IF signal into in phase (I) and quadrature phase (Q) components for processing. See Figure 10.11. Modulating a received signal this way sometimes makes subsequent processing steps easier. The phase of a signal, for example, is given by $\varphi(t) = \tan^{-1}[Q(t)/I(t)]$, FM can be demodulated as $FM(t) = d\varphi(t)/dt \approx \Delta\varphi/\Delta t$, and AM is demodulated as $AM(t) = \sqrt{I^2(t) + Q^2(t)}$.

Consider the system with the vector modulator fixed at unity gain and no phase shift [i.e., $h(t) = 1 + j0$]. The FM response to the quad antenna array as the array is commutated around the elements is illustrated in Figure 10.12 [12]. In this illustration, an unmodulated (CW) signal is traversing across the array with an angle of arrival given by ϕ, measured from some reference direction. The antennas are assumed to be independent ideal spatial field sensors and the maximum spacing between adjacent elements is somewhat shorter than the smallest $\lambda/2$. As the signal traverses across the array, the received signal at each element i is phase shifted (θ_0, θ_1, θ_2, θ_3) by an amount corresponding to the element position in the array with respect to the arriving wavefront. At each element *i,* the received signal is given by:

$$x(t,i) = s(t)m_i \tag{10.20}$$

where the ith element response is

$$m_i = e^{j\frac{2\pi}{\lambda}\cos\left(-\frac{2\pi i}{N_a}+\phi_{AOA}\right)} = e^{j\theta_i(\phi_{AOA})} \tag{10.21}$$

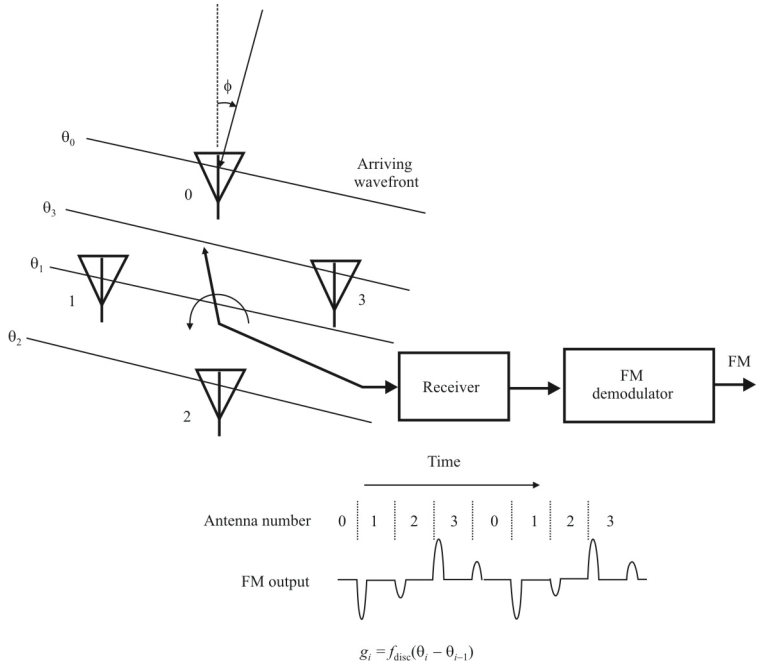


Figure 10.12 FM response to antenna commutation. Impulse outputs are approximately proportional to the phase step between elements.

where $s(t)$ is the signal at the center of the array, R is the array radius in meters, λ is the wavelength of the signal in meters, N_a, is the number of elements in the array, ϕ_{AOA} is the angle of arrival referenced clockwise from element 0, and i is the element index.

As the elements are commutated (CCW in this case) from element i to element $i + 1$, the FM demodulator generates impulses the amplitudes of which are proportional to the phase step between each pair of elements as shown in Figure 10.11. These impulses are sent through a narrow bandpass filter with center frequency equal to the commutation rate. The phase of the sine wave output of this filter with respect to the commutation index is proportional to the azimuth AOA to within a constant. This filter, being narrowband, is an effective way to minimize the thermal and other noise sources, and whose bandwidth can be quite narrow.

10.2.4.2 Adaptive Algorithm

The impulses are sampled with an A/D converter. Each rotation around the array, estimates of the phase steps $\theta_i - \theta_{i-1}$ between each sequential pair of antenna

elements is obtained. These phase steps are removed by applying the negative phase steps at the IF.[2] The inverse modulation is performed by the vector modulator synchronously with the antenna commutation. That is, as the antenna is switched from element i–1 to i, the antenna applies a phase step $e^{j(\theta_i - \theta_{i-1})}$ and the algorithm apples a simultaneous phase step of $e^{j(\hat{\theta}_i - \hat{\theta}_{i-1})}$. The impulses out of the FM detector then become error signals for each phase step. This is shown diagrammatically in the lower right of Figure 10.11. In the figure f_{disc} is the FM discriminator's response to a phase step and is represented by

$$\begin{aligned} e_i(n) &= f_{\text{disc}}\left\{\left(\theta_i - \hat{\theta}_i\right) - \left(\theta_{i-1} - \hat{\theta}_{i-1}\right)\right\} \\ &= f_{\text{disc}}\left\{\left(\theta_i - \theta_{i-1}\right) - \left(\hat{\theta}_i - \hat{\theta}_{i-1}\right)\right\} \end{aligned} \tag{10.22}$$

These error signals are used to refine the estimates of θ_i each pass around the array. Once the θ_is have been estimated, the line of bearing can be calculated with

$$\hat{\phi}_{\text{AOA}} = \tan^{-1}\frac{\hat{\theta}_3 - \hat{\theta}_1}{\hat{\theta}_0 - \hat{\theta}_2} \tag{10.23}$$

as illustrated in Figure 10.13 [12].

Thus, we see that a circular array of dipoles or monopoles can be used to determine the AOA of a signal. This is a fairly common use for antenna arrays in EW systems. The AOAs thus obtained from several systems are typically combined at a central location for the purpose of calculating the geoposition of target systems. This is commonly accomplished by measuring where the AOAs from two or more EW systems intersect [14]. This is known as *triangulation*.

10.3 Butler Matrix

One example of a phased array is the Butler matrix [15 16]. Actually, this is a term that sometimes refers to the entire antenna as well as the phase adjustment

[2] Since phase is relative, only the phase difference between elements is important. Hence, we do not measure θ_i directly but the phase *difference* between each pair of elements. We arbitrarily assign 0 phase to the first element and compute the remainder of the phases based on that assignment. We then remove the average phase from the set of phases. This effectively puts the 0 degree phase reference at the center of the array.

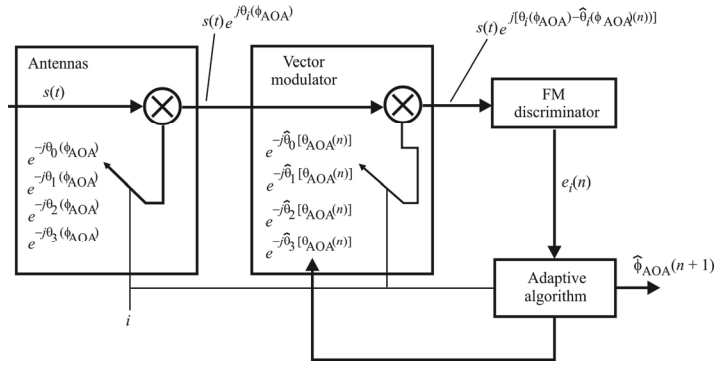


Figure 10.13 Signal flow graph of a four-element single-channel interferometer.

components but it literally refers to the phase adjustment components only. The antenna is normally a CDAA, the outputs of the antennas of which form the input to an arrangement of phase shifters as illustrated in Figure 10.14. (This is but one example of an arrangement of phase shifters that can be structured to form a Butler matrix. See, for example, [17].)

For the DF application, the amplitudes of the outputs of the fixed phase shifters are used to estimate the direction of arrival of signals impinging on the antennas.

A Butler matrix can also be used to form a smart antenna that can steer beams and nulls in specific directions. In this case, one or more of the phase shifters in the matrix are adjustable as in any other phased array.

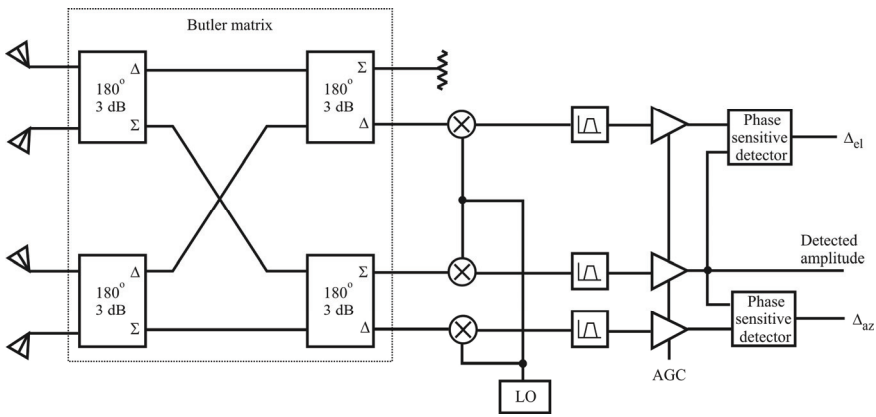


Figure 10.14 Butler matrix.

10.4 Beamforming

We will consider beamforming in considerable depth in Chapter 19. Here we will just mention that beamforming is one of the more successful applications of antenna arrays.

10.4.1 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 compute 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 3).

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 bearing 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" [2]. There are three criteria that are normally associated with these algorithms. The first criterion is their ability to resolve two signals that arrive at closely spaced azimuths. This is from where the term "superresolution" is derived. The closer in azimuth at which two signals arrive the more difficult it is to separate them—to calculate the two distinct angles of arrival. 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 angles of arrival 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.

Let P denote the angular response of an antenna array. A typical plot of P might look like that in Figure 10.15. In this example there are two signals impinging on the antenna array, one at about $\pi/2$ and another at about $3\pi/2$. The abscissa on a chart like this corresponds to the angle around the antenna array, while the ordinate corresponds to the amplitude response of the array.

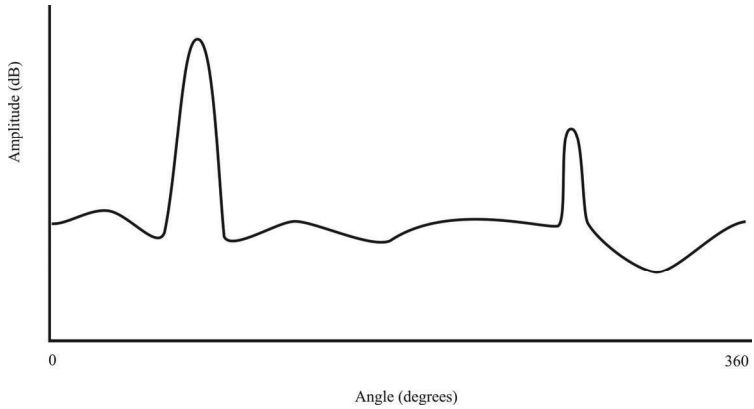


Figure 10.15 Azimuthal response of an array.

There are various ways to compute *P*, 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 discusses 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 that could be arriving at a nonhorizontal elevation angle.

An estimate of the covariance matrix of signals impinging on an antenna array is given by

$$\hat{\mathbf{R}} = \mathbf{MSM}^{\mathrm{H}} + \mathbf{N}$$

This matrix represents the measurable response of the array. The matrix **M** is the matrix of steering vectors of the array, which represents the response of the array if a signal impinges from a particular direction. **S** contains the amplitudes of the signals, and **N** is a noise covariance matrix. The angle of arrival information is contained in the matrix **M**. If the noise at the antennas is uncorrelated, as it will tend to be as more data is averaged, then $\mathbf{N} \rightarrow \sigma^2\mathbf{I}$, where **I** is the identity matrix of appropriate size (equal to the number of antennas, *M*), and where it is assumed that the noise power at each antenna is the same.

The steering vector associated with the *i*th antenna is given by

$$\vec{m}_i = e^{j2\pi\vec{k}\cdot\vec{r}_i}$$

This steering vector is the response of the *i*th antenna if a signal were coming from direction $\vec{k}$.

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 beamforming and MUSIC.

10.4.2 Array Beamforming

One of the oldest techniques for dealing with the cochannel 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 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 a desired direction, enhancing the reception in that direction while suppressing the reception in others. The structure is shown in Figure 10.16.

If $x_j(t)$ represents the output of the *j*th antenna element and $s(t)$ is the signal that is transmitted, then

$$x_j(t) = s(t + t_j) + n_j(t)$$

where $n_j(t)$ represents the additive noise at antenna *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

$$y(t) = \sum_j w_j x_j(t - \tau_j) \tag{10.24}$$

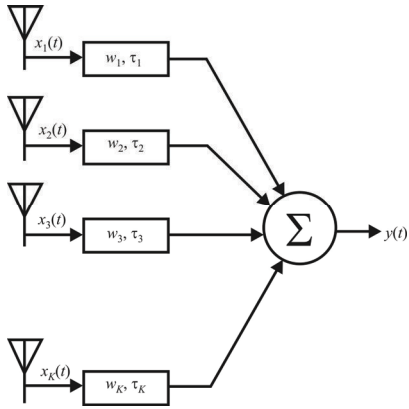


Figure 10.16 One of many types of beamforming network.

If the time delays are adjusted so that $t_j = \tau_j$, and assuming the weights are set to one, then

$$y(t) = Ks(t) + \sum_j n_j(t - \tau_j) \tag{10.25}$$

$$= \sum_j w_j s(t + t_j - \tau_j) + w_j n_j(t - \tau_j) \tag{10.26}$$

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 directions of arrival of signals.

10.4.3 Subspace Method (MUSIC)

A method of superresolution DF based on eigendecomposition of the signal correlation matrix was developed originally by Schmidt called multiple signal classification (MUSIC) [18]. This is but one of many subspace-based techniques for dealing with multiple signals impinging on an antenna array. In this technique the eigenvalues and eigenvectors of the signal correlation matrix are determined. The eigenvectors corresponding to the largest eigenvalues span the signal subspace and the eigenvectors associated with the smallest eigenvalues span the noise subspace. These sets of eigenvectors are complete and orthogonal so their (noise free) inner products are zero. The antenna response pattern of interest here is given by

$$P_{\text{MUSIC}} = \frac{1}{\vec{m}^{H}\mathbf{R}_{\text{N}}\mathbf{R}_{\text{N}}^{\text{H}}\vec{m}} \tag{10.27}$$

where $\mathbf{R}_{\text{N}}$ is the matrix of eigenvectors that spans the noise space [this forces the denominator in (10.27) to equal zero]. The column vector $\vec{m}$ is the steering vector of the array assuming that a signal is coming from a given direction. Like the other methods discussed above, calculation of this spectrum is required for every increment of azimuth because $\vec{m}$ will be different for different azimuths.

The advantages of MUSIC are that it can accommodate a nonuniform antenna array and no a-priori knowledge of the number of signals is required. Its biggest single downfall is it cannot deal with fully correlated signals, such as those caused by multipath reflections.

A comparison of the MUSIC performance versus some other popular types of array processing algorithms is shown in Figure 10.17 [19]. Compared to conventional beamforming, MUSIC has much better capability to resolve two signals. Maximum likelihood can produce extraneous peaks in the array spectrum. Maximum entropy processing can result in offset bias in the peaks. MUSIC has none of these limitations—it does have the correlated signal shortcoming, however.

10.4.4 Electromagnetically Complete Antenna Array

An electromagnetically complete antenna is illustrated in Figure 10.18. It is composed of three electrically small dipoles and three electrically small loops, all orthogonal to each other [20–23].

This antenna array was designed to perform DF in the HF band. In the HF band, antennas with a length of 1 m are electrically small in the far field. This antenna drives a six-channel receiving system, one channel for each antenna that can measure the azimuth as well as the elevation.

A simulation experiment of this antenna connected to a processor that modeled the MUSIC spectrum analysis program was conducted. The MUSIC algorithm computes the spectrum, parameterized by the spatial variables θ and ϕ, given by

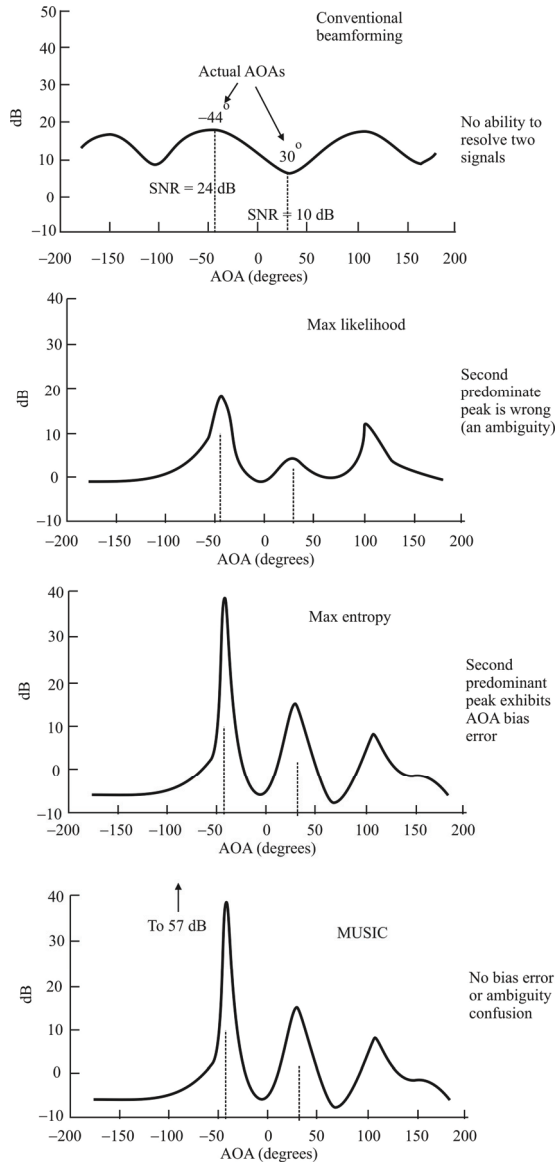


Figure 10.17 Comparison of MUSIC to other array processing algorithms.

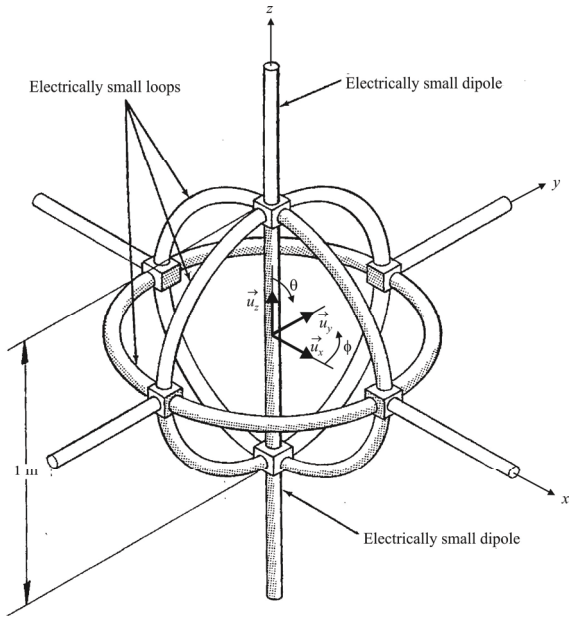


Figure 10.18 Electromagnetically complete antenna array consists of three orthogonal dipoles and three orthogonal loops.

$$S(\theta,\phi) = \left[\frac{\sum_{i=1}^{p} \left| \vec{a}^{\mathrm{H}}(\theta,\phi) e_i \right|^2}{\vec{a}^{\mathrm{H}}(\theta,\phi)\vec{a}(\theta,\phi)} \right]^{-1} \tag{10.28}$$

where $\vec{a}(\theta,\phi)$ is the array steering vector, the response of the array to a signal from (θ, ϕ).

The simulated results are illustrated in Figure 10.19. In this case four signals with an SNR of 40 dB with the signal parameters as given in Table 10.1.

10.5 Monopulse Receivers for Direction Finding

Monopulse receivers are used in antiradar EW systems (ELINT and platform self-protection suites) and for the detection and processing of target radar signals. They

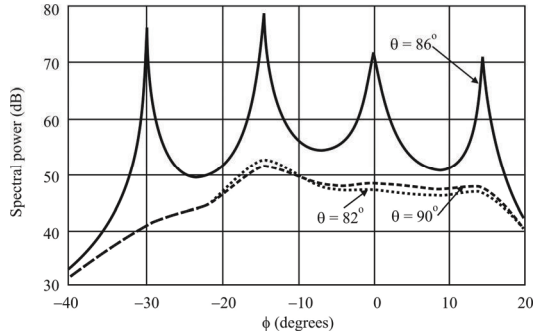


Figure 10.19 CART response. There are four incident signals in this example corresponding to the parameters in Table 10.1.

Table 10.1 Signal Parameters Corresponding to Figure 10.19

Signal Number	Incident Angle		Polarization	
	θ	ϕ	γ	η
1	86	–30	0	90 (horizontal)
2	86	–15	45	90 (left circular)
3	86	0	89	0 (vertical)
4	86	15	45	–90 (right circular)

Source: [20].

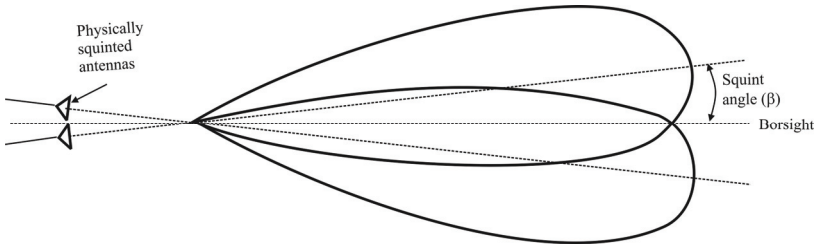


Figure 10.20 Amplitude monopulse patterns.

typically provide an indication of the presence of target signals impinging on the EW platform as well as other related parameters such as frequency, AOA, and target type. We discuss applications of these monopulse receivers, and in particular, their associated antenna arrays in this section [24–28].

Monopulse receivers used in *radar warning receivers* (RWRs, part of platform self-protection suites) and ELINT systems come in two flavors: amplitude comparison and phase comparison. In reality, according to Sherman [29], most modern monopulse receivers combine amplitude processing with phase processing forming a hybrid receiver. It has been shown that at high SNRs, the characteristics of these two receiver types are similar except at boresight [30], so we will analyze the performance characteristics of only one here: amplitude comparison. At the end we provide a brief performance comparison of the two techniques [31].

The antenna patterns of the amplitude monopulse antennas are illustrated in Figure 10.20 [32]. Two antennas are squinted so their amplitude patterns cross as shown. The amplitudes of the incoming signal are compared to get an estimate of the AOA. On the other hand, the antenna patterns for the phase monopulse antennas are parallel with each other. The phase angle of arrival is estimated by measuring the phase difference of the signals from the two antennas, as shown in Figure 10.21 [32].

The term "monopulse" is really a misnomer. The technique, whether amplitude or phase, applies to many more cases than just pulsed signals. The most relevant feature is the simultaneous measurement of parameters. In other types of receivers for radar signals, sequential measurements are made. If a target is moving, or a myriad of other possible conditions occur, the characteristics of the target that are being measured can change from one measurement to the next. That is the greatest advantage of monopulse processing in that the parameter measurements are all made at the same instant. The downside of a monopulse receiver is that more hardware is required.

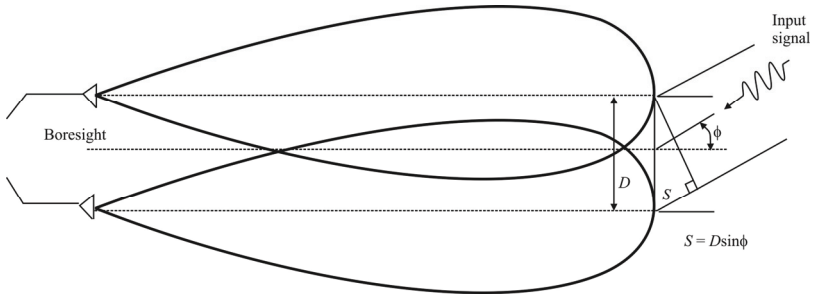


Figure 10.21 Phase monopulse patterns.

The vast majority of the theory about monopulse receivers in RWRs or ELINT systems has its origins in the theory of radars. Radars have been examined in excruciating detail since before World War II, so a considerable amount is known about them. In addition, they come in many forms, to include monopulse radars.

We first consider a monopulse receiving system that incorporates antennas with voltage responses that are approximated by cosine patterns (over a segment of the pattern anyway). The output voltages of the antennas are used directly in that formulation. That is followed by an analysis of a monopulse system whose antenna responses are modeled as Gaussian; that is, their far field voltage patterns follow a Gaussian response characteristic (again, over a portion of the pattern). In the latter case, the angle measurements (azimuth and elevation) are indirectly computed by first forming the sum and difference patterns of the voltage responses of the antennas and forming a ratio of these two patterns.

In both cases we explicitly provide formulas for the angles as well as provide an error analyses assuming the presence of AWGN combined with the signal received at all antennas.

10.5.1 Amplitude Monopulse Receiver

This section follows [33] closely. Consider two channels (there are typically more than two) of the amplitude monopulse DF RF receiver shown in Figure 10.22. We assume the receiver uses a square-law detector (as depicted by the diodes) with the resultant video signals being amplified by a pair of logarithmic amplifiers. The amplified video signals are then coupled to a difference amplifier whose output signal is proportional to the azimuth angle of the RF emitter. The RF signals at the

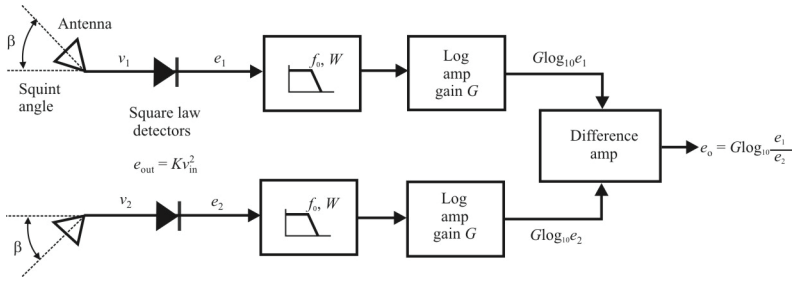


Figure 10.22 Monopulse receiver.

antennas are represented by $v_1(t)$ and $v_2(t)$ and the video detector outputs are $e_1(t)$ and $e_2(t)$. The voltages $v_1(t)$ and $v_2(t)$ are given by

$$\begin{aligned} v_i(t) &= A_{si}\cos(\omega_0 t + \theta_i) + v_{ni}(t), \qquad i = 1,2 \\ &= \text{signal} + \text{noise} \end{aligned} \tag{10.29}$$

where $A_{si}\cos(\omega_0 t + \theta_i)$ is the RF signal voltage and A_{si} is the antenna amplitude pattern function [the amplitudes of the output of the antennas, which are functions of the amplitudes of the input signals and the antenna responses—see (10.39) below]. v_{ni}(t) is narrowband white noise of zero mean and variance σ_{ni}^2, that is, $\mathcal{E}\{v_{ni}(t)\} = 0$ and $\mathcal{E}\{v_{ni}^2(t)\} = \sigma_{ni}^2$. We assume that the noises $v_{n1}(t)$ and $v_{n2}(t)$, are uncorrelated.

Using Rice's representation [34] of random noise, $v_{ni}(t)$

$$v_{ni}(t) = a_i \cos\omega_0 t + b_i \sin\omega_0 t$$

where

$$a_i = \frac{1}{T}\int_0^T v_{ni}(t)\cos\omega_i t dt$$

$$b_i = \frac{1}{T}\int_0^T v_{ni}(t)\sin\omega_i t dt$$

Rice's representation assumes that

$$\mathcal{E}\{a_i b_i\} = 0$$
$$\mathcal{E}\{a_i a_j\} = \mathcal{E}\{b_i b_j\} = \chi(0)\delta_{ij}$$

where

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$
$$\chi(\tau) = \text{autocorrelation function}$$
$$= \mathcal{E}\{x(t)x^*(t-\tau)\}$$

This implies that a_i and b_i, as well as a_i and a_j (b_i and b_j), are independent and uncorrelated. We further assume that the pdfs of the coefficients a_i and b_i are Gaussian.

Equation (10.29) can now be written as

$$v_i(t) = \left[(A_{si}\cos\theta_i + a_i)^2 + (A_{si}\sin\theta_i - b_i)^2 \right]^{1/2} \cos(\omega_0 t + \phi_i)$$

where

$$\phi_i = \tan^{-1} \frac{A_{si}\sin\theta_i - b_i}{A_{si}\cos\theta_i + a_i}$$

The portion of the video detector outputs passed by the lowpass filters is

$$e_i(t) = \frac{1}{2} K \left[(A_{si}\cos\theta_i + a_i)^2 + (A_{si}\sin\theta_i - b_i)^2 \right]$$
$$= S_i(t) + n_i(t)$$
$$= \text{signal} + \text{noise}$$

where

$$S_i(t) = \frac{1}{2} K A_{si}^2$$
$$n_i(t) = \frac{1}{2} K \left[a_i^2 + b_i^2 + 2A_{si}(a_i\cos\theta_i - b_i\sin\theta_i) \right] \qquad (10.30)$$

The output of the difference amplifier is then

$$\begin{aligned} e_{\mathrm{o}} &= G\log_{10}(S_1+n_1)-G\log_{10}(S_2+n_2) \\ &= G\log_{10}\frac{S_1}{S_2}+G\log_{10}\left(1+\frac{n_1}{S_1}\right)-G\log_{10}\left(1+\frac{n_2}{S_2}\right) \end{aligned}$$

The noise portion of the output voltage is

$$e_{\mathrm{N}} = G\log_{10}\left(1+\frac{n_1}{S_1}\right)-G\log_{10}\left(1+\frac{n_2}{S_2}\right) \tag{10.31}$$

Using the series expansion [35],

$$\ln(1+x) = x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\cdots, \qquad -1<x<1$$

and

$$\log_{10} x = 0.4343\ln x$$

neglecting higher-order terms, (10.31) can be rewritten as

$$e_{\mathrm{N}} \approx 0.4343G\left(\frac{n_1}{S_1}-\frac{n_2}{S_2}\right) \tag{10.32}$$

$$e_{\mathrm{N}}^2 = \left(\frac{0.4343G}{S_1S_2}\right)^2\left(S_2^2n_1^2-2S_1S_2n_1n_2+S_1^2n_2^2\right) \tag{10.33}$$

Since the S_i are deterministic, from (10.32) we get the average value of the output noise voltage as

$$\mathcal{E}\{e_{\mathrm{N}}\} = \frac{0.4343G}{S_1S_2}\left(S_2\mathcal{E}\{n_1\}-S_1\mathcal{E}\{n_2\}\right)$$

From (10.30)

$$\mathcal{E}\{n_i\} = \frac{K}{2}\left(\mathcal{E}\{a_i^2\} + \mathcal{E}\{b_i^2\}\right)$$

Therefore,

$$\mathcal{E}\{e_N\} = \frac{0.4343G}{S_1 S_2}\left[S_2 \frac{K}{2}\left(\mathcal{E}\{a_1^2\} + \mathcal{E}\{b_1^2\}\right) - S_1 \frac{K}{2}\left(\mathcal{E}\{a_2^2\} + \mathcal{E}\{b_2^2\}\right)\right] \quad (10.34)$$

The noise variance σ_{ni}^2 is

$$\sigma_{ni}^2 = \mathcal{E}\{a_i^2\} = \mathcal{E}\{b_i^2\}$$

By definition, the variance of a *random variable* (r.v.) x is given by

$$\sigma_x^2 = \mathcal{E}\{(x - \overline{x})^2\}$$

where

$$\overline{x} = \mathcal{E}\{x\}$$

From (10.34) the expected value of the noise voltage is

$$\mathcal{E}\{e_N\} = \frac{0.4343G}{S_1 S_2}\frac{K}{2}\left[2S_2\sigma_{n1}^2 - 2S_1\sigma_{n2}^2\right]$$

Since the antennas are relatively close, we make the reasonable assumption that the noise variances are equal (but the noises are still uncorrelated)

$$\sigma_{n1}^2 = \sigma_{n2}^2 = \sigma_n^2$$

so that

$$\mathcal{E}\{e_N\} = \frac{0.4343GK}{S_1 S_2}\sigma_n^2\left(S_2 - S_1\right)$$

The expected value of the output noise voltage squared (10.33) becomes

$$\mathcal{E}\{e_{\mathrm{N}}^2\} = \left(\frac{0.4343G}{S_1S_2}\right)^2 \left(S_2^2\mathcal{E}\{n_1^2\} - 2S_1S_2\mathcal{E}\{n_1n_2\} + S_1^2\mathcal{E}\{n_2^2\}\right) \tag{10.35}$$

The expected value of n_i^2 from (10.30) is given by:

$$\mathcal{E}\{n_i^2\} = \frac{K^2}{4}\begin{bmatrix} \mathcal{E}\{a_i^4\} + \mathcal{E}\{b_i^4\} + 4\mathcal{E}\{a_i^2\}A_{si}^2\cos^2\theta_i - 4\mathcal{E}\{a_i\}\mathcal{E}\{b_i\}A_{si}^2\sin\theta_i\cos\theta_i \\ +4\mathcal{E}\{b_i^2\}A_{si}^2\sin^2\theta_i + 2\mathcal{E}\{a_i^2\}\mathcal{E}\{b_i^2\} + 4\mathcal{E}\{a_i^3\}A_{si}\cos\theta_i \\ -4\mathcal{E}\{a_i^2\}\mathcal{E}\{b_i\}A_{si}\sin\theta_i + 4\mathcal{E}\{a_i\}\mathcal{E}\{b_i^2\}A_{si}\cos\theta_i - 4\mathcal{E}\{b_i^3\}A_{si}\sin\theta_i \end{bmatrix} \tag{10.36}$$

It can be shown that for a Gaussian distribution, if x is a zero mean r.v. that is Gaussian distributed, then

$$\mathcal{E}\{(x-\mu)^q\} = \begin{cases} (q-1)!!\sigma^q & \text{for } q \geq 2 \text{ and } q \text{ even} \\ 0, & \text{for } q \text{ odd} \end{cases} \tag{10.37}$$

where μ is the mean, σ is the standard deviation, and x!! denotes the *double factorial.*[3] Recall that it was assumed that the coefficients a_i and b_i are normally distributed with zero mean. Therefore, in our case $\mu = 0$. Applying (10.37) to the factors in (10.36), we get

$$\begin{aligned} \mathcal{E}\{a_i^4\} &= \mathcal{E}\{b_i^4\} = 3\sigma_{\mathrm{n}}^4 \\ \mathcal{E}\{a_i^3\} &= \mathcal{E}\{b_i^3\} = 0 \\ \mathcal{E}\{a_i^2\} &= \mathcal{E}\{b_i^2\} = \sigma_{\mathrm{n}}^2 \\ \mathcal{E}\{a_i\} &= \mathcal{E}\{b_i\} = 0 \end{aligned}$$

Therefore, (10.36) becomes

$$\mathcal{E}\{n_i^2\} = K^2(A_{si}^2\sigma_{\mathrm{n}}^2 + 2\sigma_{\mathrm{n}}^4)$$

n_1 and n_2 are independent of each other since the coefficients a_i and b_i are independent. Therefore, in (10.35)

[3] Double factorial, x!!, denotes the product of every odd number between 1 and x.

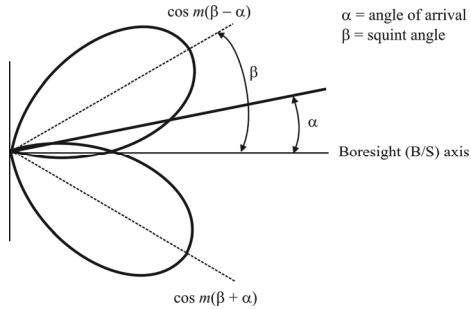


Figure 10.23 Cosine antenna patterns. The antenna response close to the main lobes of the pattern are approximated as cosine functions as shown, perhaps out to ±30° off the main axis of the beams.

$$\begin{aligned}\mathcal{E}\{n_1 n_2\} &= \mathcal{E}\{n_1\}\mathcal{E}\{n_2\} \\ &= K^2\sigma_n^4\end{aligned}$$

Equation (10.35) becomes

$$\mathcal{E}\{e_N^2\} = \left(\frac{0.4343GK\sigma_n}{S_1S_2}\right)^2\left[S_2^2(A_{s1}^2+2\sigma_n^2) - 2S_1S_2\sigma_n^2 + S_1^2(A_{s2}^2+2\sigma_n^2)\right]$$

The variance of the output signal is defined as

$$\begin{aligned}\sigma_{e_o}^2 &= \mathcal{E}\{(e_o-\bar{e}_o)^2\} = \mathcal{E}\{e_N^2\} - \mathcal{E}^2\{e_N\} \\ &= \left(\frac{0.4343GK\sigma_n}{S_1S_2}\right)^2\left[S_2^2(\sigma_n^2+A_{si}^2) + S_1^2(\sigma_n^2+A_{s2}^2)\right] \qquad (10.38)\end{aligned}$$

10.5.1.1 Cosine Antenna Patterns

Assume antenna patterns of the form shown in Figure 10.23 [33]. The receiver input voltage is then

$$\begin{aligned}v_i(t) &= \cos m(\beta\pm\alpha)\left[V_0\cos(\omega_0 t+\theta_i)\right] + v_{ni}(t) \\ &= A_{si}\cos(\omega_0 t+\theta_i) + v_{ni}(t)\end{aligned}$$

Therefore,

$$\begin{aligned} A_{s1} &= V_0 \cos m(\beta - \alpha) \\ A_{s2} &= V_0 \cos m(\beta + \alpha) \end{aligned} \tag{10.39}$$

then

$$\begin{aligned} S_1 &= \frac{1}{2} K A_{s1}^2 = \frac{1}{2} K V_0^2 \cos^2 m(\beta - \alpha) \\ S_2 &= \frac{1}{2} K A_{s2}^2 = \frac{1}{2} K V_0^2 \cos^2 m(\beta + \alpha) \end{aligned}$$

Using (10.38), the standard deviation of the receiver output voltage due to noise is

$$\sigma_{e_0} = \frac{0.4343 G K \sigma_n}{S_1 S_2} \sqrt{S_2^2 (\sigma_n^2 + A_{s1}^2) + S_1^2 (\sigma_n^2 + A_{s2}^2)} \tag{10.40}$$

Substituting in the above values for S_1, S_2, A_{s1}, and A_{s2}, we get

$$\begin{aligned} \sigma_{e_0} = {} & \frac{2(0.4343 G \sigma_n)}{V_0^2 \cos^2 m(\beta - \alpha) \cos^2 m(\beta + \alpha)} \\ & \times \left[\begin{array}{l} \cos^4 m(\beta + \alpha) \left\{ \sigma_n^2 + V_0^2 \cos^2 m(\beta - \alpha) \right\} \\ + \cos^4 m(\beta - \alpha) \left\{ \sigma_n^2 + V_0^2 \cos^2 m(\beta + \alpha) \right\} \end{array} \right]^{1/2} \end{aligned} \tag{10.41}$$

Equation (10.41) is shown in Figure 10.24 [33] as a plot of the standard deviation of the receiver output voltage versus the angle off boresight (α) as a function of various values of the noise variance (σ_n^2) or the noise standard deviation (σ_n). The log amplifier gain G, and the free-space signal voltage amplitude V_o were set equal to unity. The squint angle $\beta = 20^o$, and the antenna pattern factor $m = 2.25$.

We are interested in the relationship between σ_{e_0} and the SNR, denoted here by γ. Starting with (10.40) and substituting for S_1 and S_2, we get

$$\sigma_{e_0} = \frac{0.4343 G \sqrt{2}}{A_{s1}^2 A_{s2}^2 / 4\sigma_n^4} \sqrt{\left(\frac{A_{s2}^2}{2\sigma_n^2} \right)^2 \left(\frac{1}{2} + \frac{A_{s1}^2}{2\sigma_n^2} \right) + \left(\frac{A_{s1}^2}{2\sigma_n^2} \right)^2 \left(\frac{1}{2} + \frac{A_{s2}^2}{2\sigma_n^2} \right)} \tag{10.42}$$

The voltage SNR in channel i is

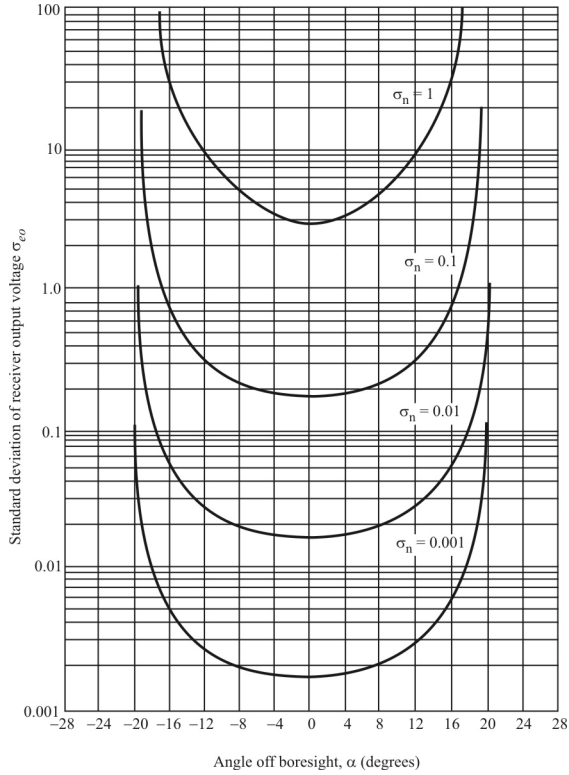


Figure 10.24 Output standard deviation of square law detector. $\beta = 20^{o}$, $m = 2.25$, $V_0 = 1$, $K = 1$.

$$\gamma_{\text{voltage},i} = \frac{A_{si}\cos(\omega_0 t + \theta_i)}{v_{ni}(t)}$$

The power SNR in channel i is (assuming the impedance is 1Ω to remove it from the discussion; it scales both the signals and noise equivalently)

$$\begin{aligned}\gamma_{\text{power},i} &= \frac{A_{si}^2\cos^2(\omega_0 t + \theta_i)}{v_{ni}^2(t)} \\ &= \frac{1}{2}\frac{A_{si}^2\left[1+\cos\{2(\omega_0 t+\theta_i)\}\right]}{v_{ni}^2}\end{aligned} \tag{10.43}$$

but

$$\sigma_n^2 = \mathcal{E}\{v_{ni}^2\}$$

Therefore, the SNR is approximated as

$$\gamma_i = \frac{A_{si}^2}{2\sigma_n^2}$$

which is the power SNR at the input to the video detector. Substituting this result into (10.42) yields

$$\sigma_{e_0} = \frac{0.4343G\sqrt{2}}{\gamma_1\gamma_2}\sqrt{\gamma_2^2\left[\frac{1}{2}+\gamma_1\right]+\gamma_1^2\left[\frac{1}{2}+\gamma_2\right]} \tag{10.44}$$

where

$\gamma_1 = SNR_1$ = power SNR in channel 1

$\gamma_2 = SNR_2$ = power SNR in channel 2

Note that (10.44) appears to be independent of the antenna radiation patterns. However the A_{si} are not, which means that SNR_i are not either, which means that (10.44) is not.

Equation (10.44) is shown in Figure 10.25 [33] as a plot of the standard deviation of the receiver output voltage versus the channel 1 power SNR as a function of the channel 2 power SNR. The gain G = 1. For omnidirectional

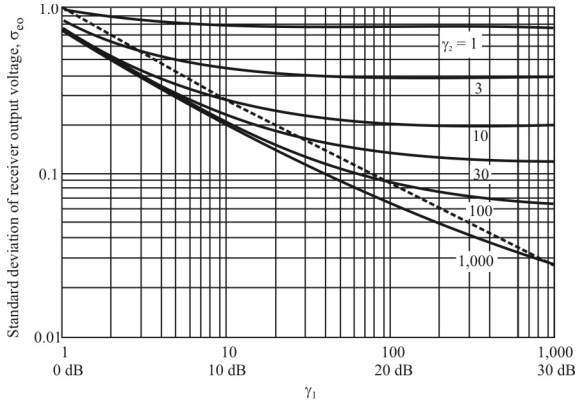


Figure 10.25 Output standard deviation versus channel 1 power SNR for a square law detector. $G = 1$. The dashed line corresponds to the case where $SNR_1 = SNR_2$.

antennas, normally the two SNRs are close in value, so this chart can be misleading. The case when $SNR_1 = SNR_2$ corresponds to the dashed line in Figure 10.25.

In many monopulse applications, however, the antennas indeed do have directional gain so the SNRs may be considerably different in that case. It may be assumed that the thermal noise is the same in both channels since that is determined primarily by the temperature of the antennas. Nonthermal noise, however, may be substantially different at the two antennas since they are looking in different directions.

The angular error in degrees is obtained from the receiver transfer curve once the output voltage error is determined. A typical receiver transfer characteristic is shown in Figure 10.26. Close to boresight, between, say –0.5 to +0.5 beamwidth, where $\alpha \approx 0$, this curve can be approximated as linear so that

$$e_o \approx 2 \times \alpha \tag{10.45}$$

and

$$\sigma_\theta \approx 0.5\sigma_{e_o} \tag{10.46}$$

This yields for a single pulse

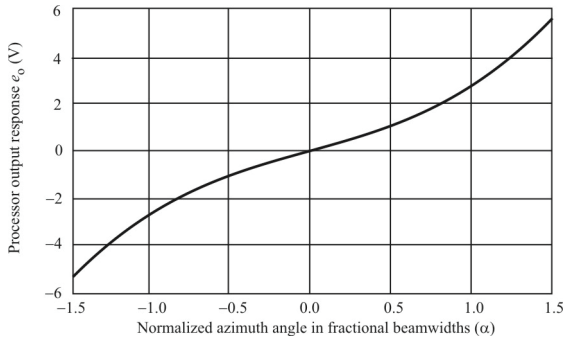


Figure 10.26 Receiver transfer characteristic.

$$\sigma_{e_0} = \frac{0.22G\sqrt{2}}{\gamma_1\gamma_2}\sqrt{\gamma_2^2\left[\frac{1}{2}+\gamma_1\right]+\gamma_1^2\left[\frac{1}{2}+\gamma_2\right]} \tag{10.47}$$

Equation (10.47) is for a single pulse. If there is more than one pulse, say, N_P, integrated before the standard deviation is calculated, then the net standard deviation is given by

$$\sigma_{e_o,t} = \frac{\sigma_{e_o}}{\sqrt{N_P}} \tag{10.48}$$

Using (10.48), plots of the AOA standard deviation for N_P = 1, 10, and 30 are shown in Figure 10.27 for representative values of the SNR.

This conclusion, as mentioned, is performance close to the boresight. In many RWR and ELINT applications, however, operating the receiver near boresight may not be possible much of the time.[4] In those cases the actual receiver transfer characteristic must be used to establish the relationship between σ_{e_o} and σ_θ.

Lipman showed that more accurate approximations allow more precise values of these standard deviations [36]. There it is shown that the signal of interest (SOI) is the logarithm of the square of the envelope of

$$v(t) = A\cos(\omega t + \psi) + n(t) \tag{10.49}$$

[4] This is as opposed to using a monopulse receiver in a tracking radar with a closed-loop servo. The error signals generated by the receiver are used to align the radar's boresight with the target, thereby ensuring operation close to boresight.

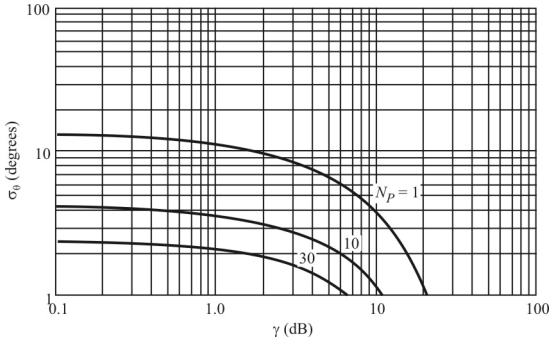


Figure 10.27 AOA standard deviation for the cosine antenna patterns. $SNR_1 = SNR_2$.

where A is the signal amplitude (and $A^2/2$ its power), ω is the radian frequency, and ψ is a uniformly distributed signal phase angle. $n(t)$ is narrowband zero mean Gaussian r.v. with variance σ^2. The noise can be represented as

$$n(t) = x(t)\cos\omega t - y(t)\sin\omega t \tag{10.50}$$

where $x(t)$ and $y(t)$ are independent zero-mean Gaussian r.v.s with variance σ^2. Note that the output SNR is $A^2/2\sigma^2$. So the signal of interest, s, is given by (dropping the notation of dependence on time)

$$\begin{aligned} s &= \ln\left[(A\cos\varphi + x)^2 + (A\sin\varphi - y)^2\right] \\ &= \ln a^2 \end{aligned} \tag{10.51}$$

where a is the envelope of the input signal. The goal is to determine the mean of s, which is the variance of the output error.

Specifically, Lipman showed that the exact expression for this mean is given by

$$\mathcal{E}\{s\} = \ln A^2 + E_1\left(A^2/2\sigma^2\right) \tag{10.52}$$

where $E_1(.)$ is the exponential integral function [37]

$$E_1(u) = \int_u^\infty \frac{e^{-v}}{v}\,dv$$

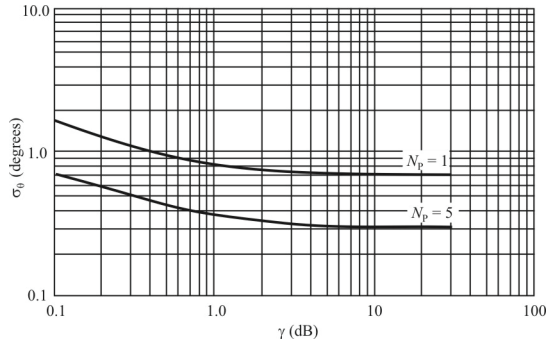


Figure 10.28 AOA standard deviation for $A = 2$ and $N_P = 1$ and 5. This chart assumes that the receiver transfer characteristic shown in Figure 10.26 applies.

Expression (10.52) is charted in Figure 10.28 [33] assuming that the same receiver transfer characteristic as shown in Figure 10.26 applies.

Summary

As we see from Figure 10.24, the variance of the receiver output voltage increases as the signal arrival angle approaches the squint angle of the antennas. The tolerable angle noise voltage would determine the useful field of view of the receiver, and this could conceivably be much less than the maximum field of view. This would depend on the boresight SNR conditions. Increasing the SNR by a factor of 10 reduces the deviation by a factor of $\sqrt{10}$ on boresight. In an amplitude monopulse receiver the SNR in one log amplifier is improving at the same time as the ratio in the other amplifier is deteriorating. We see from Figures 10.24 and 10.25 that the noisy channel becomes the deciding factor in the amount of angular noise incurred by the receiver.

10.5.1.2 Gaussian Antenna Patterns

A 2-D amplitude comparison monopulse DF system generates four beams simultaneously, the maxima of which are squinted off boresight to generate non-coincident patterns in the principal planes. One of the principal planes is shown in Figure 10.29. The angle of arrival of a signal pulse can be obtained by determining the ratio of the signal strengths from the two beams in each of the principal planes. We consider antennas with Gaussian response patterns in this section.

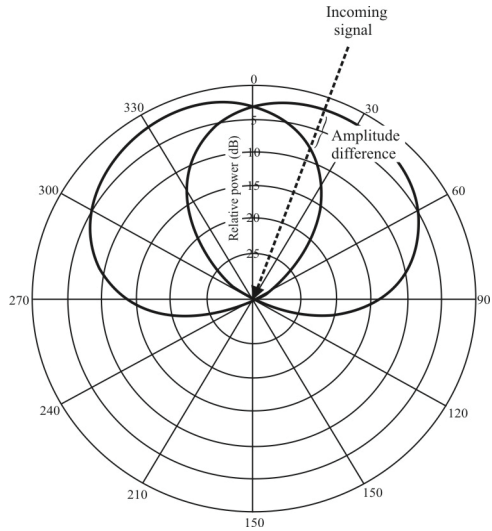


Figure 10.29 Measured principal plane. The angle of arrival of the target signal is determined by the difference in the received amplitudes from the antennas.

The simultaneous comparison of the four beams requires four receiver channels, which can be of crystal video or superheterodyne type, but must be identical in amplitude characteristics if the comparison of the four signals is to produce an accurate bearing. In the block diagram shown in Figure 10.22, the actual bearing is determined by comparing the amplitudes of the signals amplified in the log-IF strips. The logged output voltage difference is proportional to the ratio of the signal strengths received by the two squinted radiation patterns. The value of this ratio can be equated to a particular bearing angle; furthermore, the greater the rate of change of this ratio, the more accurately the bearing angle determination can be made.

In an amplitude comparison system, the offset or squint angle (as shown in Figure 10.20) is chosen to produce a beam crossover level between –1.5 and –3 dB for a practical system. Ideally, for wide frequency ranges, an antenna whose beamwidth is independent of frequency will maintain this crossover level. In selecting an antenna for a practical frequency-independent amplitude comparison monopole, it is necessary to consider several requirements. The antenna must have a pattern that is a figure of revolution about the antenna axis, is circularly polarized, and exhibits a good axial ratio at large angle from boresight. There are several antennas that satisfy these antenna requirements—the cavity-backed planar equiangular spiral, the conical equiangular spiral, and the crossed log-periodic dipole array are a few. Conical equiangular spiral antennas are illustrated in

Figure 10.30. These antennas have a pattern as illustrated in Figure 10.31. See Chapter 12 for a discussion of frequency-independent antennas.

We wish to examine the accuracy that we can expect using these frequency-independent antennas in an amplitude comparison system. Antenna response patterns of the frequency-independent antennas shown in Figure 10.28 can be approximated with a Gaussian function over a substantial portion of its beamwidth around the boresight of each individual antenna. Such a response pattern is given by

$$G(\theta) = A\exp\left[-k\frac{(\theta \pm \beta)^2}{\theta_{3\text{dB}}^2}\right] \tag{10.53}$$

where

A = relative amplitude of antenna peak response.
$k = \ln 4 = 1.388$.
β = squint angle (see Figure 10.31).
$\theta_{3\text{dB}}$ = 3 dB beamwidth of the antennas.

Thus, for the two beams shown in Figure 10.31,

$$G_1(\theta) = A_1\exp\left[-k\frac{(\theta - \beta)^2}{\theta_{3\text{dB}}^2}\right] \tag{10.54}$$

$$G_2(\theta) = A_2\exp\left[-k\frac{(\theta + \beta)^2}{\theta_{3\text{dB}}^2}\right] \tag{10.55}$$

and so

$$R = \frac{G_1(\theta)}{G_2(\theta)} = \frac{A_1}{A_2}\exp\left[4k\frac{\beta\theta}{\theta_{3\text{dB}}^2}\right] \tag{10.56}$$

With this definition of R, we can see that the units of R are volt/volt. Expressed in decibels

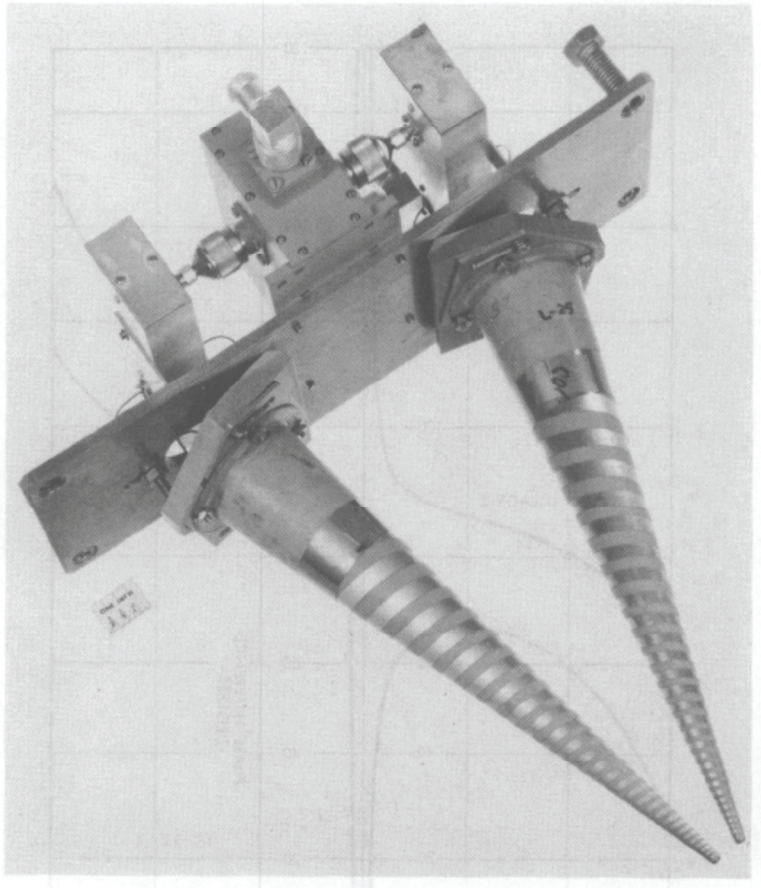


Figure 10.30 Two-element monopulse array. (Source: [38] © IEEE 1971. Reprinted with permission.)

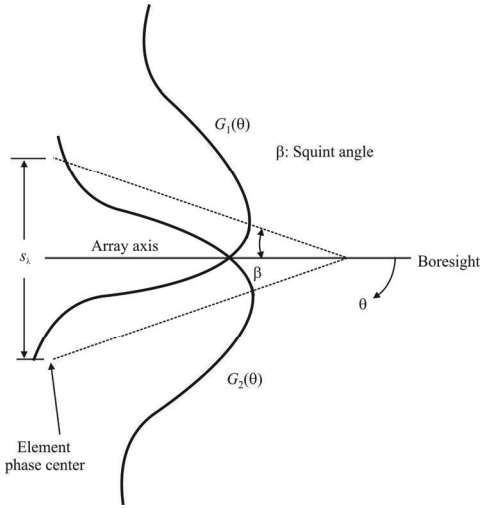


Figure 10.31 Gaussian antenna patterns for two beams. Only the "tops" of the patterns actually approximate a Gaussian distribution.

$$R_{\text{dB}} = 10\log_{10}\frac{A_1}{A_2} + 17.372 \times \frac{1.388}{\theta_{3\text{dB}}^2}\beta\theta = 10\log_{10}\frac{A_1}{A_2} + \frac{24.1}{\theta_{3\text{dB}}^2}\beta\theta \qquad (10.57)$$

The monopulse error slope is given by the rate of change of R with respect to θ. Based on (10.56), the monopole slope is given by

$$k_{\text{m,dB}} = \frac{\partial R_{\text{dB}}}{\partial\theta} = \frac{24.1}{\theta_{3\text{dB}}^2}\beta \qquad \text{dB/beamwidth} \qquad (10.58)$$

or if the non-dB form of R is being considered, the units on the error slope $k_{\text{m}} = \partial R / \partial\theta$ are volts/volt/beamwidth. This error slope value is fairly common in most monopulse receivers and is usually close to 1.6 V/V/beamwidth. The receiver produces the monopulse ratio Δ/Σ and divides it by the slope k_{m} to obtain the *indicated angle*, denoted by θ_{i}. That is,

$$\theta_{\text{i}} = \frac{\Delta / \Sigma}{k_{\text{m}}} \qquad (10.59)$$

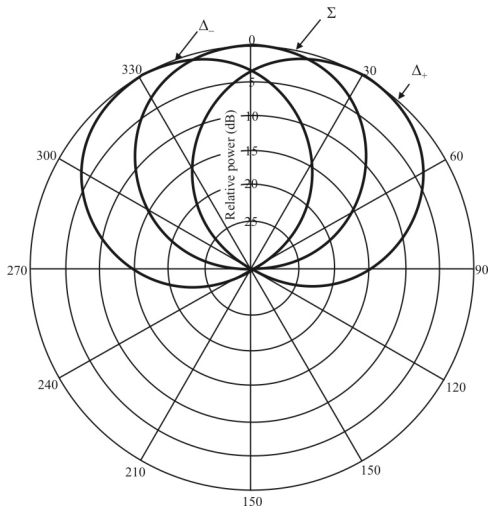


Figure 10.32 Sum and difference patterns. $\beta = 30^{o}$.

Thus, amplitude imbalance ΔR_{dB} affects the angular error approximately according to

$$\Delta\theta \approx \frac{\Delta R_{\text{dB}}\theta_{3\text{dB}}^2}{24.1\beta} \quad \text{degrees} \tag{10.60}$$

We see that the error slope is a function of the antenna beamwidth and squint angle only and is independent of the angle off boresight. In other words, the error slope is constant for all angles of arrival for which the antenna pattern may be approximated by the exponential form (10.53).

The above analysis used the voltage levels from the antennas directly to compute the azimuth angle of arrival. An alternative scheme is to indirectly measure this parameter by manipulating the signals from the antennas in a different way. In particular, the sum and difference antenna patterns can be found and the azimuth angle determined from these patterns [30].

The sum and difference patterns of the two antennas are illustrated in Figure 10.32. These patterns are found using hybrids, for example, as illustrated in Figure 10.33, where both elevation and azimuth antenna arrays are shown. Note that there

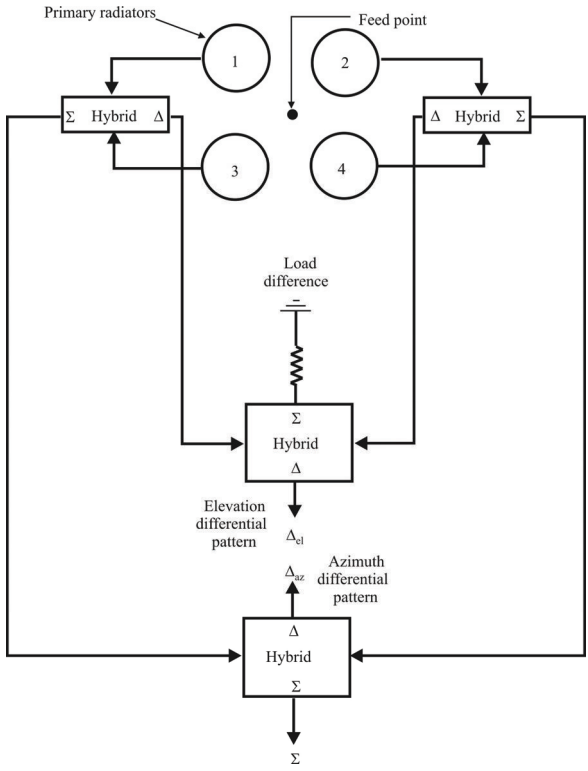


Figure 10.33 Monopulse system using quadrature hybrids. This configuration of quadrature hybrids is sometimes referred to as a Butler matrix.

are many other ways to develop these patterns. We will concentrate on only two antenna channels, specifically the azimuth, as analysis of the other angle is identical with appropriate substitution of antenna identification. Note that in Figure 10.30, the two difference patterns are identical (theoretically), but their squint angles are complementary.

With the Gaussian antenna patterns associated with (10.53), the sum and difference equations for the patterns shown in Figure 10.32 are given by

$$\begin{aligned} P_\Sigma &= \exp(-k\theta^2) \\ P_\Delta &= \exp[-k(\theta-\beta)^2] - \exp[-k(\theta+\beta)^2] \end{aligned} \tag{10.61}$$

where

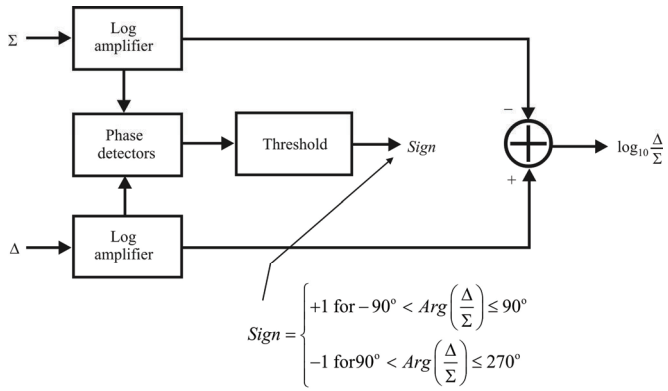


Figure 10.34 Amplitude processor.

P_Σ is the sum beam far-field voltage pattern.
P_Δ is the difference far-field voltage pattern.
θ is the normalized angle off boresight = angle off borresight/θ_{3dB}.
β is the squint angle = 0.5 fractional beamwidths.
$k = 2 \ln 2$.

The amplitude processor analytical model shown in Figure 10.34 [30] evaluates the magnitude of the ratio of the difference amplitude to the sum amplitude ($|\Delta/\Sigma|$). The sign of this ratio is determined by $\arg(\Delta/\Sigma)$. If this quantity is between 90^o and 270^o, the sign is negative; otherwise, it is positive, as shown in Figure 10.34. A plot of a typical antilog at the output of the receiver is shown by the receiver transfer characteristic in Figure 10.31.

Noise in both channels affects the accuracy according to [38]

$$\sigma_\theta = \frac{8.686}{\partial R/\partial\theta}\left[\frac{N_\Sigma/S_\Sigma + N_\Delta/S_\Delta}{2N_P}\right]^{1/2} = \frac{8.686}{\partial R/\partial\theta}\left[\frac{1/\gamma_\Sigma + 1/\gamma_\Delta}{2N_P}\right]^{1/2} \quad (10.62)$$

where

N_Σ is the noise power in sum channel.
N_Δ is the noise power in difference channel.
S_Σ is the signal power in sum channel.
S_Δ is the signal power in difference channel.
N_P = number of pulses integrated.

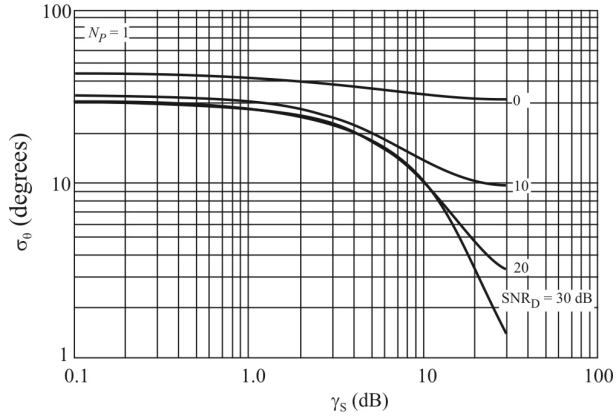


Figure 10.35 AOA standard deviation when $N_P = 1$ (no integration), $\beta = 30^o$, and $\theta_{3dB} = 60^o$.

The monopulse error slope $\partial R / \partial\theta$ is derived above yielding (10.58)

$$\frac{\partial R}{\partial \theta} = \frac{24.1}{\theta_{3dB}^2}\beta \tag{10.63}$$

So that the standard deviation of the measured azimuth (or elevation) error from (10.62), using (10.63), is given by

$$\sigma_\theta = \frac{8.686\theta_{3dB}^2}{24.1\beta}\left[\frac{1/\gamma_\Sigma + 1/\gamma_\Delta}{2N_P}\right]^{1/2} \tag{10.64}$$

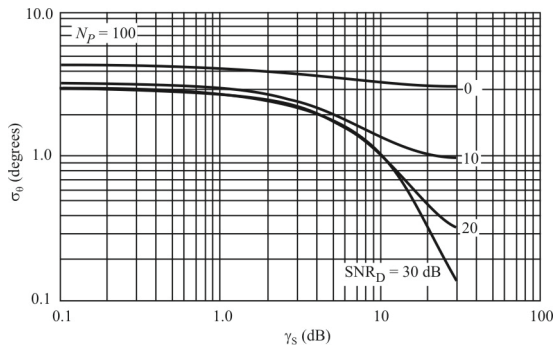


Figure 10.36 AOA standard deviation when $N_P = 100$, $\beta = 30^o$, and $\theta_{3dB} = 60^o$. Note the change of scale on the ordinate from Figure 10.34.

Equation (10.64) is charted in Figures 10.35 and 10.36 for $N_P = 1$ and $N_P = 100$, respectively. Using a single pulse produces relatively inaccurate angle measurements (Figure 10.35) while if integration is possible, the $N_P = 100$ results are respectable for SNR > 10 dB or so (Figure 10.36).

Summary

It is fairly easy to design antennas whose response patterns are approximately Gaussian. As we showed in this section, Gaussian response patterns are particularly easy to work with to determine system performance. Therefore, it is common practice to use such antennas.

10.5.2 Approximate Relationships Between the AOA Standard Deviation and SNR

Sherman [29] wrote the definitive and complete manuscript on monopulse techniques. He devoted an entire chapter (Chapter 11) to analyzing the effects of noise on azimuth and elevation measurements. In particular, he pointed out that under most circumstances (assuming adequate SNR, say SNR > 10 dB) a simple formula for the standard deviation of the measurements in a monopulse receiving system is given by

$$\sigma_\theta \approx \frac{\theta_{3dB}}{R\sqrt{2N_P \times \gamma}}\left[1+\left(\frac{\Delta}{\Sigma}\right)^2\right]^{1/2} \tag{10.65}$$

where

θ_{3dB} is the 3 dB beamwidth of the sum beam.
R is the monopulse error slope given by (10.58) at the target angle in V/V/beamwidth.
N_P is the number of pulses whose measurements are averaged to obtain the angle estimate.
γ is the single pulse SNR.

For (10.65) to apply it is necessary to assume that the noise in the two channels are uncorrelated. Some results are supplied in the reference to account for the possibility of correlated noises, however.

He pointed out that Δ and Σ in (10.65) are complex quantities, in general, and must be taken into account that way. When the computations are complete, however, the real part of the results is the useful indication of the standard

deviation. Equation (10.65) applies to each angle coordinate separately. The errors in the two coordinates may be different because of different beamwidths or different error slopes *R*.

Equation (10.65) is based on a truncated series expansion of Δ/Σ, so it is an approximation. If the SNR is lower, say, 3 dB, then he pointed out that more than the first term in the series expansion should be included to avoid significant errors. It was pointed out that [29]:

> As an example of the contribution of the higher-order terms, consider a target approximately on axis, equal noise powers in both channels, and no correlation between the noises. Then
>
> $$\sigma_{\Delta/\Sigma}^2 = \frac{\sigma_n}{2S}\left(1+\frac{\sigma_n}{S}\right) = \frac{1}{2\lambda}\left(1+\frac{1}{\gamma}\right)$$
>
> $$\sigma_{\Delta/\Sigma} = \sqrt{\frac{1+1/\gamma}{2\gamma}}$$
>
> If γ = 10 dB ($1/\gamma = 0.1$), the higher-order term increases the variance by 10% and the standard deviation by about 5%, which is not usually significant in an error analysis. However, if the SNR is only 3 dB, the higher-order term increases the variance by 50% and the standard deviation by 22%.

The value of *R* in (10.65) must be the error slope at the target angle. If the target is off-axis, this error slope may differ considerably from the slope at boresight. In the AN/FPS-16 radar, for example, Δ/Σ, is

$$\frac{\Delta}{\Sigma} = \sqrt{2}\tan\left(1.14\frac{\alpha}{\theta_{3db}}\right)$$

while the error slope is

$$\frac{\partial R}{\partial \theta} = 1.14\sec^2\left(1.14\frac{\alpha}{\theta_{3dB}}\right)$$

So on the boresight axis ($\alpha = 0$), the slope is 1.6 V/V/beamwidth. At one-half beamwidth off axis, it is 2.3 V/V/beamwidth.

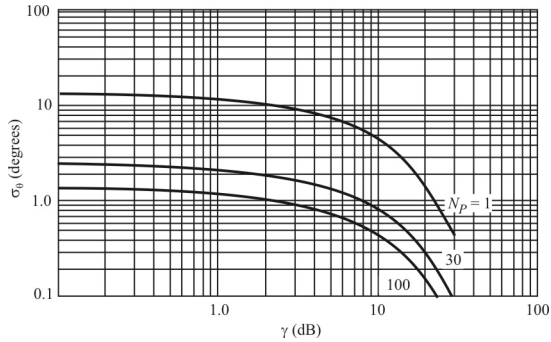


Figure 10.37 AOA standard deviation when $\sigma_\theta = 60^o$, $R = 1.6$, close to boresight, and $N_P = 1, 30, 100$. These curves assume the receiver transfer characteristic is as shown in Figure 10.26.

Equation (10.65) can be simplified if the AOA is close to boresight. In that case, the second term in the brackets is close to zero and the equation becomes

$$\sigma_\theta \approx \frac{\theta_{3dB}}{R\sqrt{2N_p\gamma}} \tag{10.66}$$

where all the variables are as above. This expression is plotted in Figure 10.37 [40] for some representative values of the parameters. Because of the assumption made to derive this simplification, the points below about 8 dB are dubious in Figure 10.37.

10.5.3 Comparison of Amplitude Monopulse to Phase Monopulse Receivers

Sinsky and Lew [30] conducted a Monte Carlo experiment comparing the performance of an amplitude monopulse receiver to a phase monopulse receiver, both in AWGN. The amplitude processor is shown in Figure 10.34. The phase processor is shown in Figure 10.38 [30]. The results of the experiment are illustrated in Figures 10.39 [30] and 10.40 [30]. Monopulse bias errors are plotted in Figure 10.39 for both configurations as a function of SNR. In both cases, there

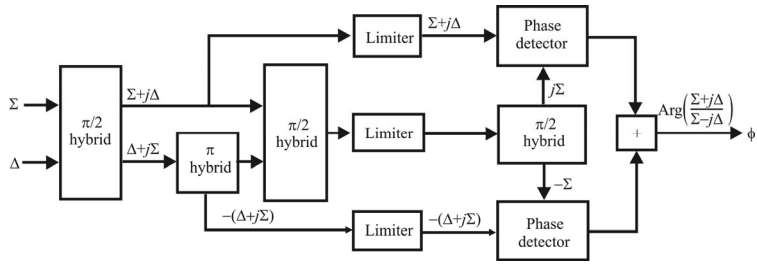


Figure 10.38 Phase processor.

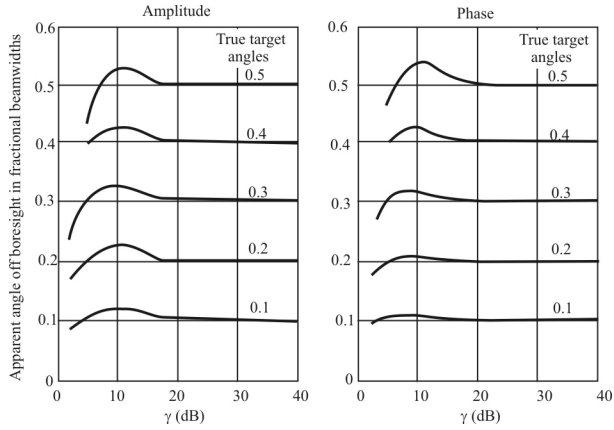


Figure 10.39 Monopulse bias errors due to noise.

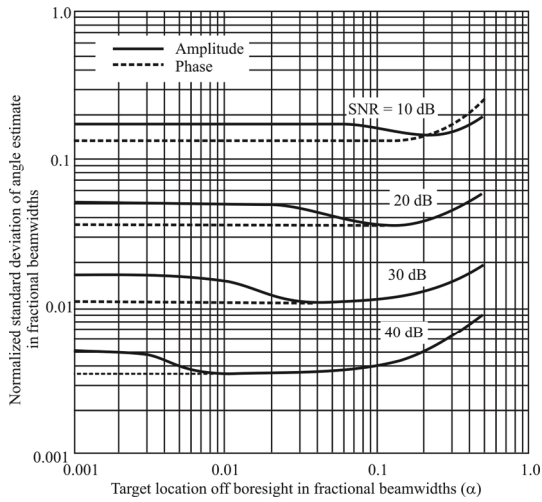


Figure 10.40 Standard deviation of angle errors.

are significant bias errors until the SNR reaches about 20 dB. For the amplitude processor, the maximum bias error is relatively constant as the target moves away from boresight, while the phase processor maximum bias error increases as the target moves away from boresight. The bias errors are due to the nonlinear relationship between processor output voltage and target angle from boresight.

A comparison of the standard deviation about the mean for the two processors is shown in Figure 10.40. At target angles very close to boresight, we see that the amplitude processor has a $\sqrt{2}$ larger standard deviation than the phase processor. This translates to a requirement for 3 dB greater SNR at the amplitude processor's input to achieve the same angle estimate errors as the phase processor. We see that this advantage of the phase processor exists at angles close to boresight. At target angles greater than 0.08 beamwidth, the two processors have comparable errors for 20 dB input SNR. At the 40 dB SNR levels, the processors have essentially equal errors beyond 0.01 beamwidth.

10.5.4 Unresolved Targets

A *radar warning receiver* (RWR) and/or ELINT system typically has many radar signals in its passband at any one time. This produces the necessity of deinterleaving the pulses coming from multiple sources. One of the most useful parameters for sorting pulses arriving at an RWR is the AOA since a radar target cannot rapidly change it from one pulse to the next. Determining the AOA is one of the first measurements made in order to cull and bin as best as possible the arriving pulses prior to calculating and analyzing the *pulse repetition interval* (PRI). The latter takes considerably more computation power and time than for most of the other parameters but arguably produces the best sorting information.

However, when pulses from two or more radars arrive within the bandwidth of the RWR, measuring the AOA of any of them becomes more difficult. This is the so-called *unresolved target issue* associated with a monopulse receiver. A second signal within the passband of the receiver causes the SOI to fade. What the receiver actually measures is some combination of the two or more signals arriving at the antenna. There is no guarantee that the AOA determined by the monopulse receiver correctly corresponds to any of the impinging signals.

When pulses from two signals are present then there are processes that can be used to separate them [39]. When three signals are being received, the computations become much more complicated. With more than three, it is virtually impossible to separate them.

Note that multipath reflections from the same target cause the same issues as if the second signal were coming from another SOI.

Table 10.2 Characteristics of DF Techniques

	Amplitude Comparison	**Phase Interferometer**
Sensor configuration	Typically four to six equi-spaced antenna elements for 360° coverage	Two or more RHC or LHC spirals in fixed array
Theoretical DF accuracy	$\frac{\theta_{3dB}^2 \Delta R}{24\beta}$ (Gaussian antenna pattern)	$\frac{\lambda}{2\pi d \cos\alpha}\Delta\alpha$
Typical DF accuracy	3 to 10 degrees RMS	0.1 to 3 degrees RMS
DF accuracy improvement	Decrease antenna BW; decrease amplitude mistrack; increase squint angle	Increase spacing of outer antennas; decrease phase mistrack
Sensitivity to multipath/reflections	Highly sensitive; mistrack of several decibels can cause large DF errors	Relatively insensitive; interferometer can be made to tolerate large phase errors
Platform constraints	Locate in reflection free area	Reflection free area; real estate for array; prefers flat radome
Applicable receivers	Crystal video; channelizer; acousto-optic; compressive; superheterodyne	Superheterodyne

Source: [28].

Two or more "targets" are unresolvable if the angle measurement cannot be made on each target without significant error due to the presence of the others [40]. Note that to be unresolvable, targets must be closely spaced in all parameter coordinates, not just the AOA.

The conical spiral antennas described in this section have several applications. We discussed one: a monopulse direction finder application. Such devices are commonly used in RWRs in aircraft *self-protect equipment* (ASE). They are also sometimes used in ELINT systems, although these systems are fewer in number than aircraft with the requirement to know that a radar is looking at them so that defensive maneuvers can be attempted. As such, ELINT systems usually opt for the better accuracy that is obtainable from interferometric DF equipment. A summary of the two types of DF approaches is shown in Table 10.2.

10.5.5 Summary

We discussed an amplitude-based monopulse receiver in this section. Error analyses were provided assuming AWGN is added to the SOI. Such receivers are commonly used in RWRs and ELINT systems for the intercept and analysis of radar signals. Specifically they are used to measure signal parameters for the determination of the AOA.

Two types of antenna patterns were considered. First cosine patterns were assumed and the sensitivity was determined. Then, the more common Gaussian antenna pattern was assumed and performance evaluated.

It was determined that at adequate SNR (>10 dB) accurate AOAs (3^o–10^o) can be determined with the amplitude monopulse receiver. For the RWR application, this accuracy is generally adequate. For the ELINT application, however, it generally is not. In addition, at low SNRs (~3 dB or less) there is a bias in the AOA calculation. This bias is caused by the nonlinear transformation from antenna pattern voltage measurements to AOA calculations.

Finally, a comparison of the amplitude monopulse receiver to the less common phase monopulse receiver is provided. It is shown that the phase interferometer approach is considerably more accurate (0.1^o–3^o) at the expense of substantially more hardware and therefore cost of implementation.

10.6 Concluding Remarks

We discussed some applications of antenna arrays in this chapter. Many of these applications provided an estimate of the direction of arrival of a signal at the receiving system. This is a fairly ubiquitous use of antenna arrays in EW systems. Another fairly common application of antenna arrays is beam steering, where the main lobe of an antenna array is steered in particular directions. There are sundry reasons for doing this in EW systems, including nulling out interfering signals, sending C2 signals in a particular direction to minimize opponent ES (this is an example of EP), and sending an EA signal in a particular direction to minimize fratricide.

We discussed monopulse receiving systems at length, giving examples of architectures that take advantage of the phase difference of a signal received at two or more antennas, as well as the amplitude response of directional antennas pointed in slightly different directions. Monopulse receivers are quite popular as receivers in aircraft self-defense suites for detecting the presence of air defense radar signals lighting up aircraft.

References

[1] http://www.ece.uiuc.edu/about/history/wullenweber.html.

[2] Gething, P. J. D., *Radio Direction Finding and Superresolution,* London, U.K.: Peregrinus, 2nd ed., 1991.

[3] Poisel, R. A., *Introduction to Communications Electronic Warfare Systems*, 2nd ed., Norwood, MA: Artech House, 2008, p. 336.

[4] Poisel, R. A., *Introduction to Communications Electronic Warfare Systems*, 2nd ed., Norwood, MA: Artech House, 2008, p. 359.
[5] Poisel, R. A., *Introduction to Communications Electronic Warfare Systems*, 2nd ed., Norwood, MA: Artech House, 2008, p. 360.
[6] Poisel, R. A., *Introduction to Communications Electronic Warfare Systems*, 2nd ed., Norwood, MA: Artech House, 2008, p. 361.
[7] Jenkins, H. H., *Small-Aperture Radio Direction Finding*, Norwood, MA: Artech House, 1991, pp. 120–122.
[8] Poisel, R. A., *Introduction to Communications Electronic Warfare Systems*, 2nd ed., Norwood, MA: Artech House, 2008, p. 339.
[9] Poisel, R. A., *Introduction to Communications Electronic Warfare Systems*, 2nd ed., Norwood, MA: Artech House, 2008, p. 344.
[10] Poisel, R. A., *Introduction to Communications Electronic Warfare Systems*, 2nd ed., Norwood, MA: Artech House, 2008, pp. 347–353.
[11] Poisel, R. A., *Introduction to Communications Electronic Warfare Systems*, 2nd ed., Norwood, MA: Artech House, 2008, pp. 342–347.
[12] Peavey, D., "The Single Channel Interferometer Using a Pseudo-Doppler Direction Finding System," *Proceedings ICASSP* 1997, pp. 4129–4132.
[13] Peavey, D., et al. "Method and Apparatus for Adaptively Determining the Bearing Angle of a Radio Frequency Signal," U.S. Patent #5,426,438, June 1995.
[14] Poisel, R. A., *Electronic Warfare Target Location Methods*, Norwood, MA: Artech House, 2005.
[15] Cuckson, H., "Multi-Octave Mixers and Hybrid Couplers for Wideband Monopulse Direction Finding Receivers," *Proceedings IEE Colloquium on Multi-Octave Active and Passive Components and Antennas*, May 10, 1989, pp. 311–312.
[16] Moody, H. J., "The Systematic Design of the Butler Matrix," *IEEE Transactions on Antennas and Propagation,* November 1964, pp. 786–788.
[17] Poisel, R. A., *Introduction to Communications Electronic Warfare Systems*, 2nd ed., Norwood, MA: Artech House, 2008, p. 353.
[18] Schmidt, R. O., "Multiple Emitter Location and Signal Parameter Estimation," *Proceedings of the RADC Spectrum Estimation Workshop*, Griffiths Air Force Base, Rome NY, pp. 243–258.
[19] Schmidt, R. O., "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Transactions on Antennas and Propagation*, Vol. AP-34, No. 3, March 1986, pp. 276–280.
[20] Bull, J. F., and L. R. Burgess, "A Compact Antenna Array for Direction Finding in the HF Band," *Proceedings IEEE Tactical Communications Conference*, 1990, Vol.1, Tactical Communications. Challenges of the 1990s, April 24–26, 1990, pp.651–657.
[21] Bull, J. F., L. R. Burgess, R. P. Flam, and R. A. Peloso, "A Broadband Electrically Small Adaptive Antenna," *Proceedings IEEE Antennas and Propagation Society International Symposium*, Vol. 24, June 1986, pp. 993–996.
[22] Bull, J. F., and L. R. Burgess, "An Adaptive Antenna Array for Digital HF Communications," *Proceedings IEEE Long Island Section Adaptive Antenna Systems Symposium*, November 1992, pp. 75–80.
[23] Bull, J. F., M. A. Arnao, and L. R. Burgess, "Hypersensitivity Effects in Adaptive Antenna Arrays," *Digest Antenna and Propagation Society International Symposium,* 1990, Vol. 1, May 7–11, 1990, pp. 396–399.
[24] Wiley, R. G., *ELINT–The Interception and Analysis of Radar Signals,* Norwood, MA: Artech House, 2006.
[25] Schleher, D. C., *Introduction to Electronic Warfare*, Norwood, MA: Artech House, 1986.
[26] Wiley, R. G., *Electronic Intelligence: the Analysis of Radar Signals*, Dedham, MA; Artech House, 1982.

[27] Wiley, R. G., *Electronic Intelligence: the Interception of Radar Signals*, Norwood, MA: Artech House, 1985.
[28] Schleher, D. C., *Electronic Warfare in the Information Age*, Norwood, MA: Artech House, 1999.
[29] Sherman, S. M. *Monopulse Principles and Techniques*, Dedham, MA: Artech House, 1984.
[30] Sinsky, A. I., and E. A. Lew, "Comparative Analysis of a Phase and an Amplitude Processor for Amplitude Monopulse Systems," *IEEE Transactions on Antennas and Propagation*, Vol. AP-31, No. 3, May 1983, pp. 519–522.
[31] Vaccaro, D. D., *Electronic Warfare Receiving Systems*, Norwood, MA: Artech House, 1993.
[32] Lipsky, S. E., *Microwave Passive Direction Finding*, New York: Wiley, 1987.
[33] Kehr, D. E., "Variance Analysis of the Angle Output for an Amplitude Monopulse Receiver," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-4, No. 4, July 1968, pp. 535–540.
[34] Rice, S. O., "Mathematical Analysis of Random Noise," *BSTJ*, July 1944.
[35] Abramowitz, M., and I. A. Stegun (eds.), *Handbook of Mathematical Functions*, New York: Dover, 1972, p. 68.
[36] Lipman, M. A., Comment on "Variance Analysis of the Angle Output Voltage for an Amplitude Monopulse Receiver," *IEEE Transactions on Aerospace and Electronic Systems,* September 1971, pp. 1009–1011.
[37] Abramowitz, M., and I. A. Stegan (eds.), *Handbook of Mathematical Functions,* Washington, D.C.: National Bureau of Standards, 1964, Ch. 5.
[38] Bullock, L. G., G. R. Oeh, and J. J. Sparagna, "An Analysis of Wide-Band Microwave Monopulse Direction-Finding Technique," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-7, No. 1, January 1971, pp. 188–203.
[39] Wang, Z., A. Sinha, P. Willett, and Y. Bar-Shalom, "Angle Estimation for Two Unresolved Targets with Monopulse Radar," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 40, No. 3, July 2004, pp. 998–1019.
[40] Sherman, S. M., *Monopulse Principles and Techniques*, Dedham, MA: Artech House, 1984, Ch. 8.

Chapter 11

Yagi-Uda Antennas

11.1 Introduction

The Yagi or Yagi-Uda antenna is used in a wide variety of applications where gain and directivity are required. The full name for the antenna is the Yagi-Uda antenna. It derives it name from its two Japanese inventors, Yagi and his student Uda.[1] They discovered that by adding "elements" of various lengths and spacings in front of and behind a dipole antenna that the performance and effectiveness of the dipole could be greatly increased and the pattern of the dipole RF energy could be "beamed" or focused in one direction. They also discovered that by changing the space between the elements, and the element lengths, they could "tune" it to get various results depending on what they wanted it to do. They found that they could change the forward "gain" of it and also that they could change the way it performed in other aspects [1].

Because of their limited bandwidth, Yagi antennas are only useful when the range of operating frequencies is very limited. In EW systems they are useful as intercept or jamming antennas when the target frequency is known or only varies over a limited frequency range. They are useful as high-gain, directional antennas in a C2 datalink role, however.

There are several advantages to Yagi-Uda arrays. Some of them are:

- Light weight;
- Low cost;
- Simple construction;
- Unidirectional beam (front-to-back ratio);
- Increased directivity over other simple wire antennas;
- Practical for use at HF (3–30 MHz), VHF (30–300 MHz), and UHF (300 MHz–3 GHz).

[1] The actual discovery of the antenna was by Uda, published in Japanese in 1926. Yagi published a similar paper on the research in English in 1928 and the short name for the antenna bears his name.

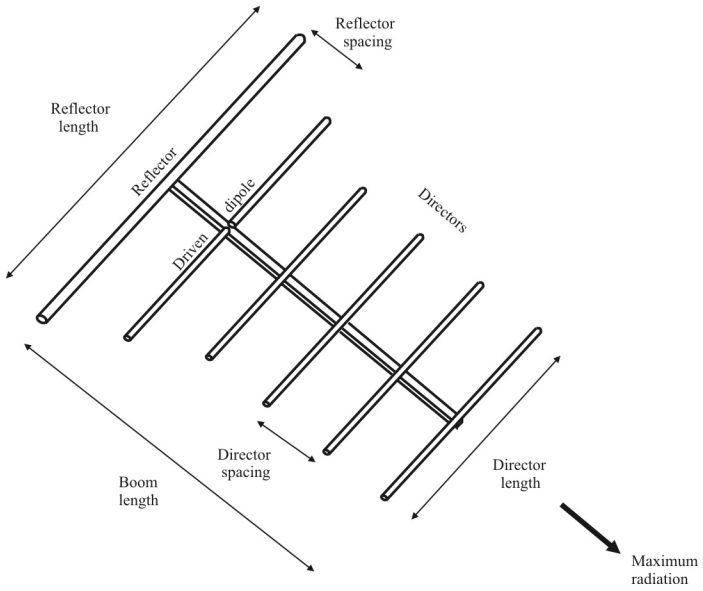


Figure 11.1 Yagi antenna.

This chapter begins with a discussion of the general characteristics and principles of operation of the Yagi-Uda dipole antenna. Included are some details on the design of these antennas. That is followed by discussions of the properties of both circular and square loop arrays. Lastly we cover the characteristics of the Yagi-Uda monopole array.

11.2 Yagi-Uda Dipole Antenna

A Yagi-Uda dipole antenna (Yagi for short), shown in Figure 11.1, consists of typically a single driven dipole along with several (two or more) wire elements arranged to be parallel to each other. An element placed behind the driven dipole is referred to as a reflector, while one or more directors are in front of the driven element(s). All but the driven elements are passive and are driven "parasitically." A parasitically driven element absorbs energy from the electromagnetic wave that is passing by and reradiates this energy. The spacing of the elements is such that the phases of the driven element and that of the reradiated waves add properly off the end of the array.

The main driven element of the Yagi is normally a (typically folded) dipole. Recall from Chapter 5 that the impedance of a dipole at resonance is $Z = 73 + j0$. We showed there that the impedance of a folded dipole is $4Z$, or approximately 300Ω at resonance. Thus the folded dipole in a Yagi is a natural match for 300Ω twin-lead transmission lines (Chapter 4).

Typical 2-D radiation patterns are shown in Figure 11.2 and 3-D in Figure 11.3.

11.2.1 Parasitic Elements

The parasitic elements reradiate signals that are impinging on them in a slightly different phase to that of the driven element. In this way the signal is reinforced in some directions and cancelled out in others. The amplitude and phase of the current that is induced in the parasitic elements are dependent upon their length and the spacing between them and the driven element.

It is difficult to have complete control over both the amplitude and phase of the currents in all the parasitic elements. Thus, it is not possible to obtain complete cancellation in some directions. Nevertheless, it is still possible to obtain a high degree of reinforcement in one direction that produces high gain and also have a high degree of cancellation in others to provide a good front-to-back ratio.

To obtain the required phase shift, an element is made either inductive or capacitive. If the parasitic element is inductive, the induced currents are in such a phase that they reflect the power away from the parasitic element. This causes the antenna to radiate more power away from it. An element that does this is called a reflector. It can be made inductive by tuning it below resonance. This can be done by physically adding some inductance to the element in the form of a coil or, more commonly, by making it somewhat longer than the resonant length. Generally, it is made about 5% longer than the driven element, making it somewhat inductive (see Figure 5.12).

If a parasitic element is capacitive, the phase of the induced currents are such that they direct the power radiated by the whole antenna in the direction of the parasitic element. An element that does this is called a director. It can be made capacitive by tuning it above resonance, which can be done by physically adding some capacitance to the element in the form of a capacitor, or more commonly by making it about 5% shorter than the driven element.

The addition of further directors increases the directivity of the antenna, increasing the gain, and reducing the beamwidth in the process. The addition of further reflectors, however, makes no noticeable difference, so typically there is only one included (see Figure 11.1).

The antenna exhibits a directional pattern consisting of a main forward lobe and a number of spurious sidelobes. The main one of these is the reverse lobe

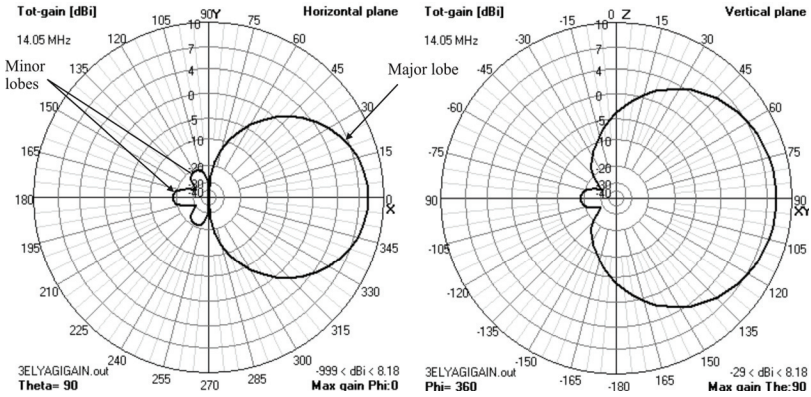


Figure 11.2 Three-element Yagi 2-D patterns.

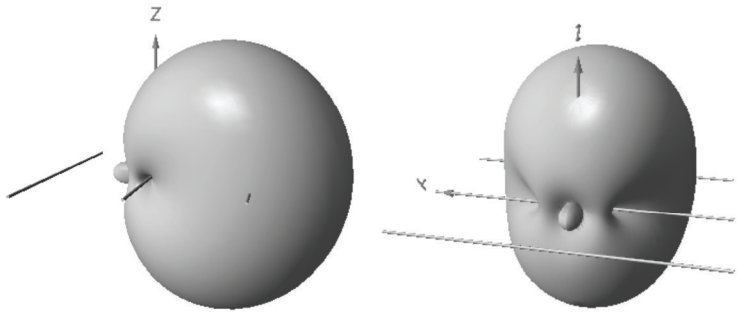


Figure 11.3 Three-element Yagi 3-D patterns. Left: front view, right: rear view.

caused by radiation in the direction of the reflector. The antenna can be optimized to either reduce this or produce the maximum level of forward gain. Unfortunately, the two do not coincide exactly and a compromise on the performance has to be made depending upon the application.

11.2.2 Yagi-Uda Characteristics

11.2.2.1 The Driven Element

The driven element of a Yagi is the feed point where the feed line is attached from the transmitter to the Yagi to perform the transfer of power from the transmitter to the antenna. As mentioned, this element is more often than not a folded dipole in a dipole Yagi antenna.

A dipole driven element will be resonant when its electrical length is (slightly less than) half of the wavelength of the frequency applied to its feed point. The feed point Figure 11.1 is at the center of the driven element.

11.2.2.2 The Director

The director is the shortest of the parasitic elements, and this end of the Yagi is aimed at the receiving station. It is resonant slightly higher in frequency than the driven element due to its length being about 5% shorter than the driven element. The directors' lengths can vary, depending upon the director spacing, the number of directors used in the antenna, the desired pattern, the pattern bandwidth, and the element diameter, all of which are design variables. The number of directors that can be used are determined by the physical size (length) of the supporting boom needed. The directors are used to provide the antenna with directional pattern and gain. The amount of gain is directly proportional to the length of the antenna array and not by the number of directors used. The spacing of the directors can range from 0.1λ to 0.5λ or more and will depend largely upon the design specifications of the antenna.

11.2.2.3 The Reflector

The reflector is the element that is placed at the rear of the driven element (the dipole). Its resonant frequency is lower due to its length being approximately 5% longer than the driven element. Its length will vary depending on the spacing and the element diameter. The spacing of the reflector is between 0.1λ and 0.25λ from the driven element. Its spacing will depend upon the gain, bandwidth, front-to-back ratio, and sidelobe pattern requirements of the final antenna design.

11.2.3 Radiation Patterns

For each element in the Yagi array each with N propagation modes, located at (x_p, y_p, z_p), of length l_p, the element pattern $F_p(\phi, \theta)$ is given by [2]

$$F_p(\phi,\theta) = l_p \sin\phi \exp[j\beta(x_p \sin\phi\cos\theta + y_p \sin\phi\sin\theta + z_p \cos\phi)] \times \sum_{n=1}^{N} (-1)^n \frac{(2n-1) I_{n_p} \cos(\pi l_p' \cos\phi)}{(2n-1)^2 - (2l_p' \cos\phi)^2} \tag{11.1}$$

where $l_p' = l_p / \lambda$. With Λ elements each with N propagating modes, the total AF is given by

$$F(\phi,\theta) = \sum_{p=1}^{\Lambda} F_p(\phi,\theta), \qquad \Lambda = D + 2 \tag{11.2}$$

where D is the number of directors in the array.

The antenna's radiation pattern plays a major role in the overall performance of the Yagi antenna. Example radiation patterns are shown in Figure 11.2. The directional gain, front-to-back ratio, beamwidth, and unwanted (or wanted) sidelobes combine to form the overall radiation pattern. The antenna's radiation pattern bandwidth is the range of frequencies above and below the design frequency in which the pattern remains reasonably consistent. The amount of variation from the antenna's design specification goals that can be tolerated is subjective, and limits put into the design are mainly a matter of choice. Equal spaced, equal length directors may give higher gain at a particular frequency, but the bandwidth is narrower and larger sidelobe levels are created. Wide spacing will increase the bandwidth, but the sidelobes become larger. By varying both the spacing and director lengths, the pattern and the pattern bandwidth may be more controlled.

More directors within a given boom length will not increase the gain by any great amount, but gives better control of the antenna's pattern over a wider range of frequencies in the band of design.

If the length of each succeeding director is reduced by a set factor (%), and the spacing of each succeeding director is increased by another factor, a very clean pattern with good pattern bandwidth can be obtained. The trade-off is a small loss in the optimum forward gain (10% to 15%).

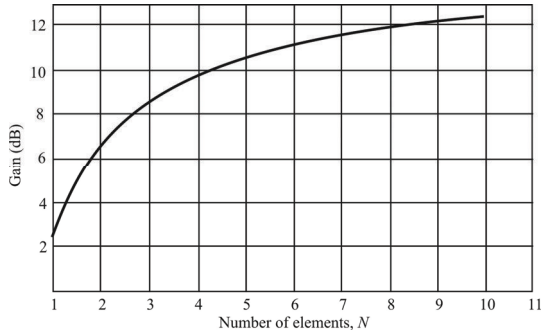


Figure 11.4 Gain versus the number of elements for a Yagi antenna.

11.2.4 Gain

The gain of a Yagi antenna is determined primarily by the number of elements in the array. However, the spacing between the elements also has an effect. The gain as a function of the number of elements is shown in Figure 11.4 [3]. The farther away from the driven element the director is, the lower the current it handles. Therefore, there is a point of diminishing returns on the number of directors. We can see that after five elements or so there is little benefit to adding more.

For highest forward gain, the main lobe becomes narrower in both the elevation and azimuth planes, and a back lobe is always present. When the back lobe is designed out, the pattern gets wider and the forward gain goes down. In some cases, the sidelobes become quite large.

11.2.5 Feed Network

It is possible to vary the feed impedance of a Yagi over a wide range. Although the impedance of the dipole itself would be 73Ω in free space, this is altered considerably by the proximity of the parasitic elements. The spacing, their length, and a variety of other factors all affect the feed impedance presented by the dipole to the feed line. In fact, altering the element spacing has a greater effect on the impedance than it does on the gain, and accordingly setting the required spacing can be used as one design technique to fine-tune the required feed impedance. Nevertheless the proximity of the parasitic elements usually reduces the impedance below the 50Ω level normally required. It is found that for element spacing distances less than 0.2 wavelength the impedance falls away precipitously. As an example, the input impedance of a 15-element Yagi dipole array is given in Table 11.1 [4].

Table 11.1 Input Impedance of 15-Element Yagi Array (Reflector Length = 0.5λ; Director Spacing = 0.34λ; Director Length = 0.406λ

Reflector Spacing	Input Impedance (Ω)
0.25λ	62
0.18λ	50
0.15λ	32
0.13λ	22
0.10λ	12

Source: [4].

To overcome this, a variety of techniques can be pursued. As mentioned, one is to use a folded dipole for the driven element. Other techniques involve using gamma matches, delta matches, baluns and the like (see Chapter 4). Delta matches can be very convenient. They involve "fanning out" the connection to the driven element. This method has the advantage that the driven element does not need to be broken to apply the feed as shown.

A gamma match is another alternative that is often used (see Chapter 4). The outer or braid of the coax feeder is connected directly to the center of the driven element. This can be done because the RF voltage at the center is zero at this point. The inner conductor of the feeder carrying the RF current is taken out along the driven element. The inductance of the arm is then tuned out by the variable capacitor. When adjusting the antenna design, both the variable capacitor and the point at which the arm contacts the driven element are adjusted. Once a value has been determined for the variable capacitor, its value can be measured and a fixed component inserted if required.

There are many ways to feed the Yagi, but, just as for other antenna systems, they all can be condensed into two main categories: the balanced feed and the unbalanced feed.

11.2.5.1 The Balanced Feed System

This may give a broader impedance bandwidth, but the main problem is that the driven element must in most cases be split in the center and insulated from the boom. Construction considerations aside, it is the better of the feed systems. Meeting the requirements of a balanced matching system is usually the main problem, but there are many methods available. One method is to not split the driven element and use a "T" match, which can be described as two gamma matches on each side of the center of the element, fed with a 1:1 balun at the center. The main drawback is that it is difficult to adjust.

11.2.5.2 The Unbalanced Feed System

Another method (for low-impedance feed points) uses a split element insulated

from the boom, and is fed with a step-down 4:1 balun made by combining two quarter wavelength sections of coaxial feed line in parallel, attaching an equal length of insulated wire to the outside of these sections, and connecting it to the center conductors at the feed point end and to the shields at the feed line end. The impedance of this type of balun should be at or near the mid-point value between the feed point impedance and the feed line impedance. For example, two 75 Ω sections in parallel will equal 37.5Ω and will match a 25Ω feed point to a 50Ω feedline with a 1:1 VSWR. The most common method in use today is the gamma match. It will provide an easy and sure method of matching to the feed point without any loss of bandwidth.

11.2.6 Bandwidth and Impedance

Of primary importance for the Yagi antenna is the impedance of the driven element, the point on the antenna where the transfer of RF from the feed line takes place. As discussed many times, maximum energy transfer of RF at the design frequency occurs when the impedance of the feed point is equal to the conjugate impedance of the feed line (ideally there is no reactance so the conjugate is simply the real part). In most antenna designs, the feed line impedance will be 50Ω, but the feed point impedance of the Yagi is rarely 50Ω. In most cases it can vary from approximately 40Ω to around 10Ω, depending upon the number of elements, their spacing, and the antenna's pattern bandwidth. If the feed line impedance does not equal the feed point impedance, the driven element cannot transfer the RF energy effectively from the transmitter, thus reflecting it back to the feed line resulting in a standing wave.

The impedance bandwidth of the driven element is the range of frequencies above and below the center design frequency of the antenna that the driven element's feed point will absorb acceptable RF power from the feed line. The design goal is to have the impedance at the center design frequency of the Yagi, $Z_Y = R_0 + j0\Omega$.

The single driven element in the Yagi resonates at a single frequency, so we would expect the Yagi to exhibit a fairly narrow bandwidth. In fact, the Yagi bandwidth is typically 5–15% of its design frequency [5]. Their beamwidth in the elevation plane is approximately 80^o–90^o while that in the azimuth plane is about 50^o–60^o or so. Their gain is between 5 and 15 dBi, yielding an effective area of between $3\lambda^2/4\pi$ and $13\lambda^2/4\pi$. There are Yagi antenna designs, however, that contain more than one driven element yet still exploit the parasitic nature of the reflector and directors [6]. Wide element spacing, large element diameter, wide pattern bandwidth, and low Q matching systems all add to a wider impedance bandwidth.

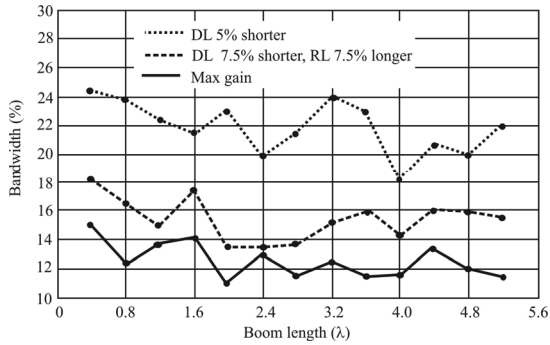


Figure 11.5 Yagi bandwidth.

The bandwidth as a function of boom length (in wavelengths) is shown in Figure 11.5 [7]. We see that maximum gain dictates that the minimum bandwidth is available. Also the bandwidth is a fairly sensitive function of the lengths of the reflector and director relative to the driven element.

11.2.7 Boom Length

Design charts for Yagi antennas are shown in Figure 11.6 [8]. The boom length is the single most important parameter, and these charts plot various parameters versus that parameter. Shown are the number of elements vs. boom length in Figure 11.6(a), gain (dBi) versus boom length in Figure 11.6(b), reflector length versus boom length in Figure 11.6(c), the R_L/D_L ratio in Figure 11.6(d), and bandwidth vs. boom length in Figure 11.6(e). Bear in mind that these charts display approximations and should only be used as a starting point. The actual parameter values should be determined by simulation and/or measurement.

Example (Yagi-Uda Array)

> The simple three-element Yagi-Uda array (one reflector: length = 0.5λ, one director: length = 0.45λ, driven element: length = 0.475λ) is illustrated in Figure 11.7 where all the elements are the same radius (a = 0.005λ) [9]. For $d_R = d_D$ = 0.1λ, 0.2λ, and 0.3λ, we want to determine the E-plane and H-plane patterns, the 3 dB beamwidths in the E- and H-planes, the front-to-back ratios (dB) in the E and *H*-planes, and the maximum directivity (dB). Also, we will plot the currents along the elements in each case.

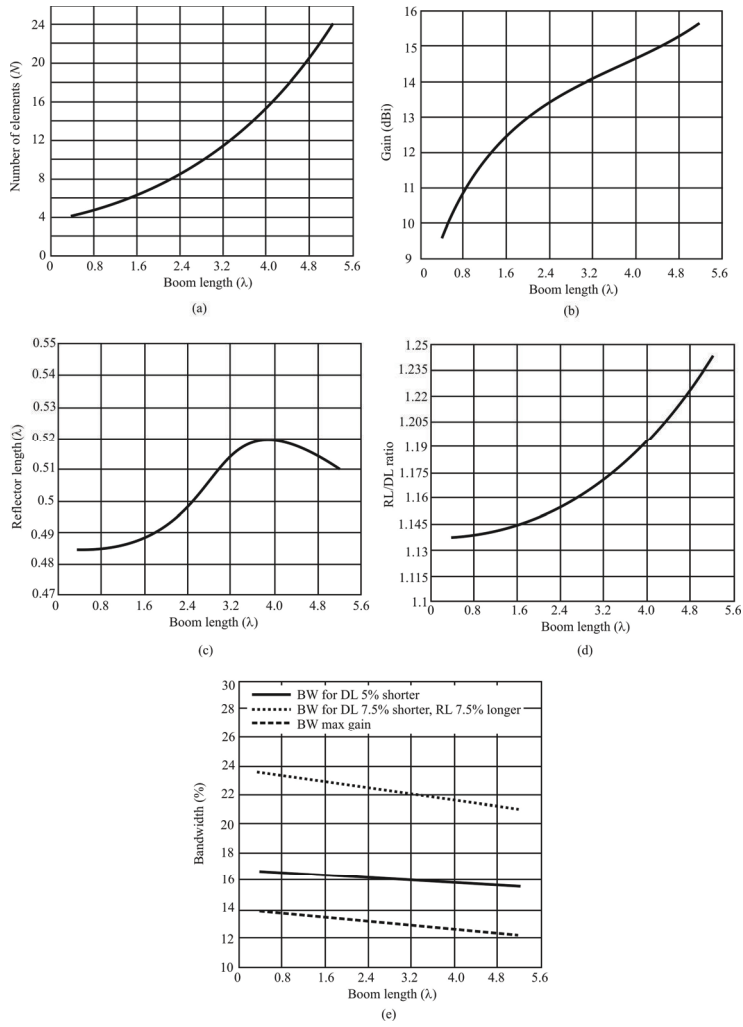


Figure 11.6(a–e) Yagi design charts: (a) number of elements (*N*) versus boom length (λ); (b) gain (dBi) versus boom length (λ); (c) reflector length (λ) versus boom length (λ); (d) *RL*/*DL* ratio; and (e) bandwidth (%) versus boom length (λ). These charts provide a starting point. Final parameters should be determined by measurement and/or simulation.

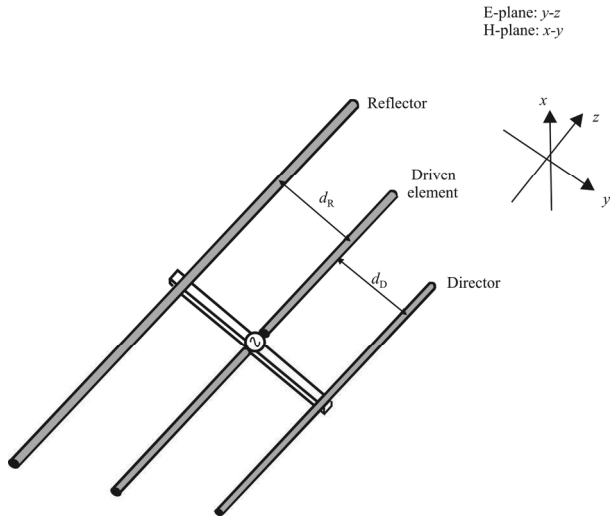


Figure 11.7 Three-element Yagi.

The current distributions are shown in Figures 11.8, 11.9, and 11.10. The E-plane patterns are illustrated in Figures 11.11, 11.12, and 11.13. The H-plane patterns are shown in Figures 11.14, 11.15, and 11.16. We can see that as the spacing increases the beamwidths decrease and the backlobe increases in magnitude [9].

The beamwidths for the three antennas as well as the front-to-back ratios and maximum directivity are given in Table 11.2. Again, as the spacing increases, the front-to-back ratio decreases and the maximum directivity increases.

11.3 Yagi Loop Array

The Yagi can be implemented with other than dipoles. Appel-Hansen [10] showed that a Yagi made of loop antennas as the feeding loop, reflector, and directors would function just as well as the dipole Yagi as a directional antenna. The construction of the antenna is as illustrated in Figure 11.17 [10]. As flat plate loops, the loops are 1.25 mm thick and the width is 1 cm. The radii were varied to examine the performance trade-offs. A single feeding loop and reflector were

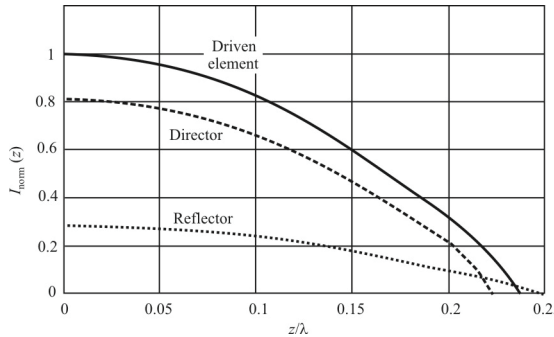


Figure 11.8 Currents 0.1λ. $d_R = d_D = 0.1\lambda$.

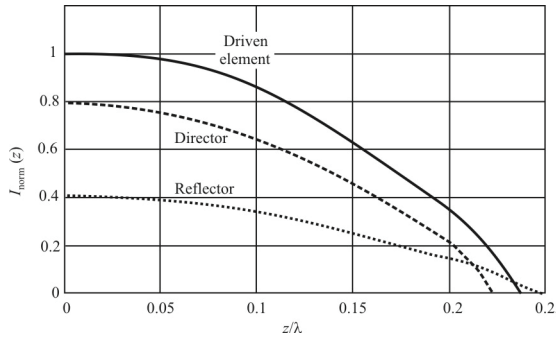


Figure 11.9 Currents 0.2λ. $d_R = d_D = 0.2\lambda$.

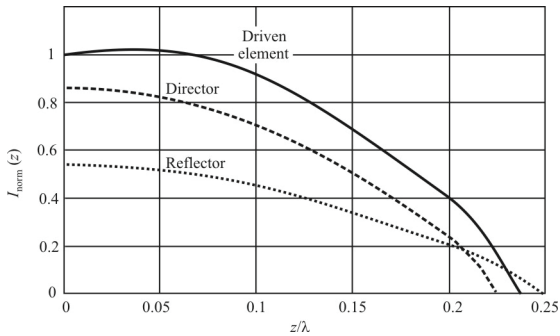


Figure 11.10 Currents 0.3λ. $d_R = d_D = 0.3\lambda$.

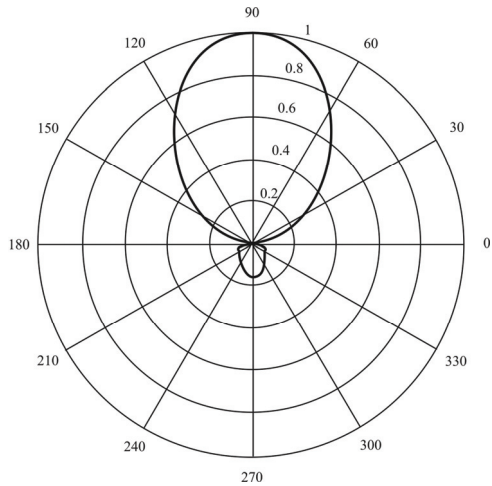


Figure 11.11 E-plane $d_R = d_D = 0.1\lambda$.

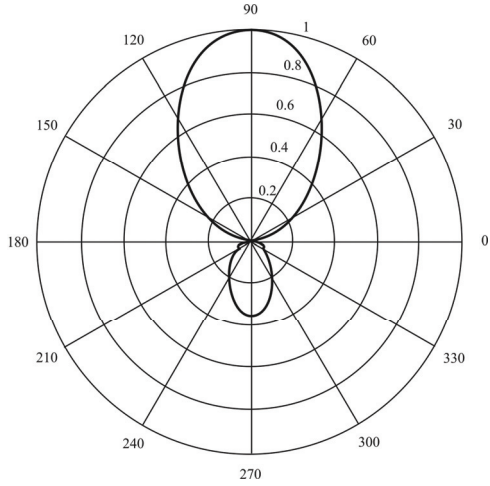


Figure 11.12 E-plane $d_R = d_D = 0.2\lambda$.

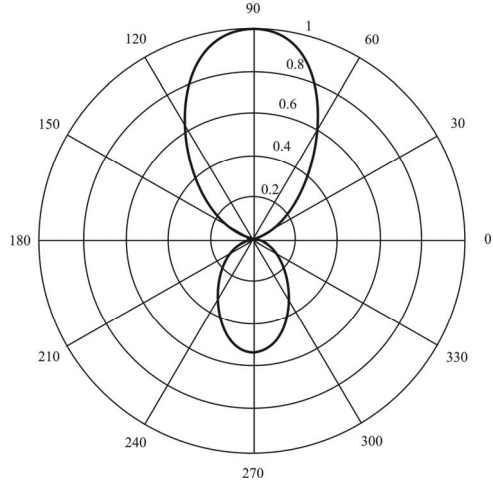


Figure 11.13 E-plane $d_R = d_D = 0.3\lambda$.

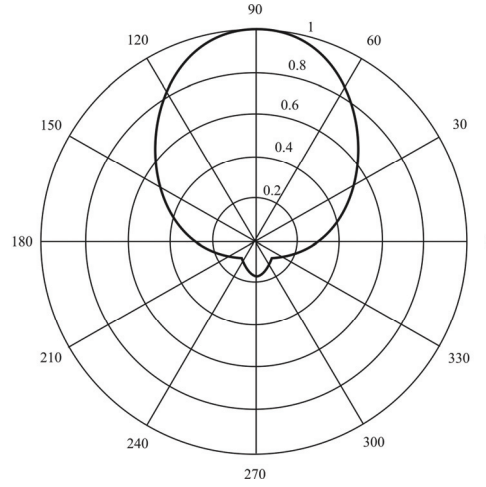


Figure 11.14 H-plane $d_R = d_D = 0.1\lambda$.

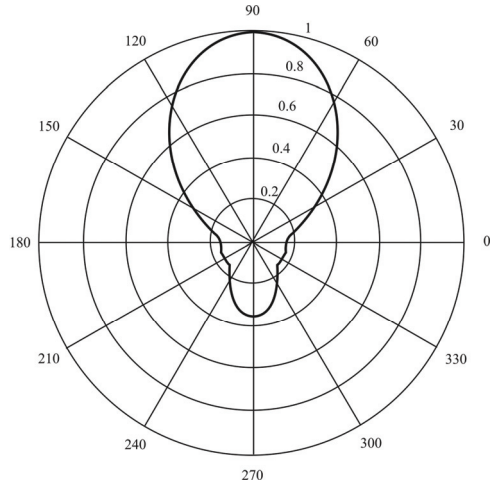


Figure 11.15 H-plane $d_R = d_D = 0.2\lambda$.

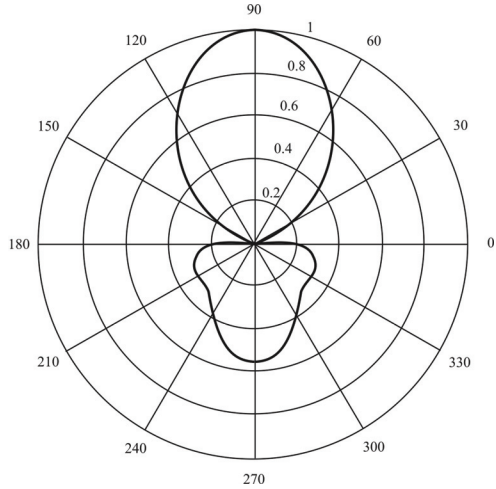


Figure 11.16 H-plane $d_R = d_D = 0.3\lambda$.

Table 11.2 Three-Element Yagi Parameters

d_R, d_D	$\theta_{E,3dB}$ (degrees)	$\theta_{H,\theta 3dB}$ (degrees)	$E_{F/B}$ (dB)	$H_{F/B}$ (dB)	D_{max} (dB)
0.1λ	62.71	86.15	15.86	15.86	7.78
0.2λ	55.84	69.50	9.20	9.20	9.09
0.3λ	51.89	61.71	5.49	5.49	8.97

Source: [9].

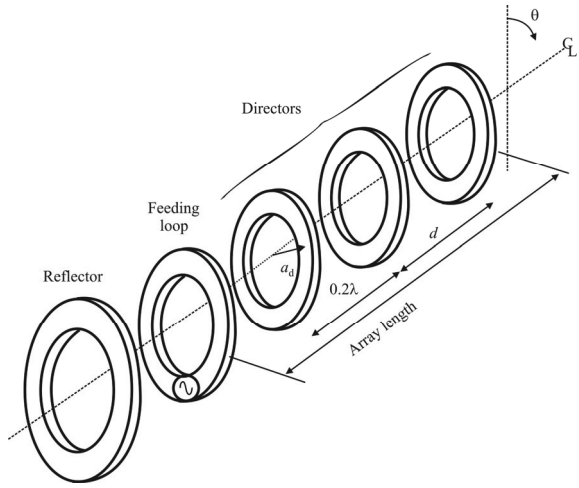


Figure 11.17 Directive loop array.

used, while the number, array length, and director radius were varied. The results of varying the director radius at an array length of 2λ are shown in Figure 11.18 [10]. At this length, the optimum value of βa_d is about $\beta a_d = 0.90$, where $\beta = 2\pi/\lambda$ is the wavenumber. Varying the array length for three types of directors produced the performance illustrated in Figure 11.19 [10]. The maximum gain achieved was approximately 15 dBi and that was with using flat plate loops as well as rods as directors, although the variation of the type of director did not affect the gain by more than 1 dB. Rods simulate the dipole directors in the dipole Yagi. The gain values shown in Figure 11.19 are certainly comparable to the gain achievable with a dipole Yagi of similar size shown in Figure 11.4. For comparison purposes between Figure 11.19 and Figure 11.4, the loop Yagi director spacing is 0.2λ, so at 1λ there are 5 elements, while at 4λ there are 20 elements. For five elements in each configuration, for example, the gains are virtually identical [11–14].

The free-space 2-D radiation patterns for the Yagi loop antenna shown in Figure 11.17 are illustrated in Figure 11.20.

The design of the Yagi loop array begins with optimizing the size and spacing of the feeding loop and reflector. Once the maximum gain is realized in the $\theta = 90^o$ direction with these two, directors are added of varying size and number until the desired parameters, such as directivity, are reached. The maximum end-fire directivity is determined by the electrical length of the array l/λ. The larger the number of directors within the length l, the smaller the electrical size of the directors will be for maximum directivity. Typically $a_d \leq \beta a_d \leq 1.0$.

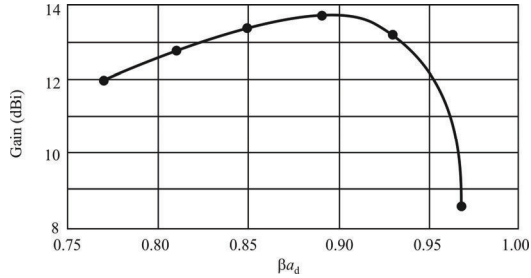


Figure 11.18 Yagi loop array gain as a function of βa_d, with an array length of 2λ.

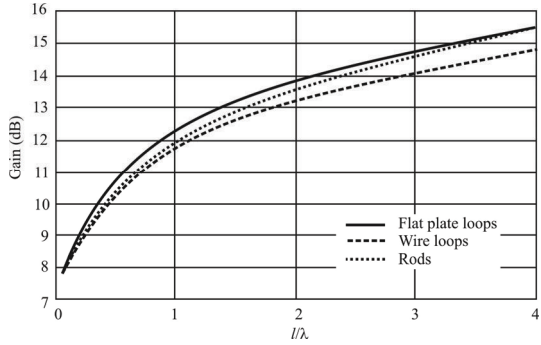


Figure 11.19 Yagi loop array gain.

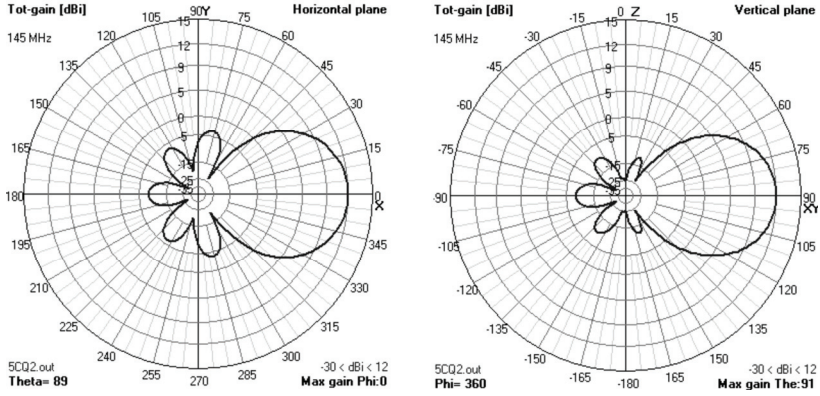


Figure 11.20 Loop Yagi with one reflector and three directors as in Figure 11.17. This is the free-space pattern. The effects of a ground plane are illustrated in Figure 11.23.

The directivity of a Yagi loop array with spacing $d/\lambda = 0.2$ is shown as a function of the number of directors or the length of the array l/λ is shown in Figure 11.21 [10]. We see that the number of directors substantially impacts on the directivity of the Yagi loop antenna.

The patterns for the square Yagi loop array shown in Figure 11.22 are shown in Figures 11.23 and 11.24. Rather than a PEC ground plane, a finite ground plane was incorporated for these charts—in particular, $\varepsilon_r = 13$ and $\sigma = 5\times10^{-3}$ $\text{℧}/\text{m}$. We see that the main lobe has a low take-off angle and there are substantial sidelobes. The antenna is approximately 4 m above the ground plane, and with the relatively poor response on the horizon, would not be an antenna for reliable intercept from a ground platform. Considerable energy loss occurs in the sidelobes as well. It is important to remember, however, that the Yagi loop antenna, as with any other loop antenna, still does not make a very good transmitting antenna due to its inherently high loss resistance and therefore low efficiency.

11.4 Yagi-Uda Monopole Array

Much like the Yagi-Uda arrays of dipoles, parasitic monopoles of grounded elements can be used to help direct the beams away from the reflector monopoles in the direction of director monopoles. A side view of such an array of monopoles along with its image is shown in Figure 11.25.

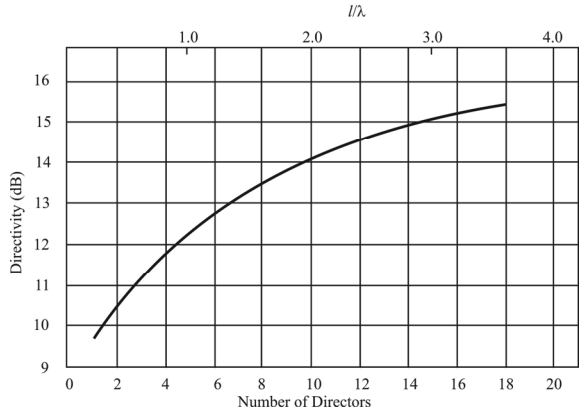


Figure 11.21 The directivity of the Yagi loop array versus the number of directors. Director spacing $d/\lambda = 0.2$.

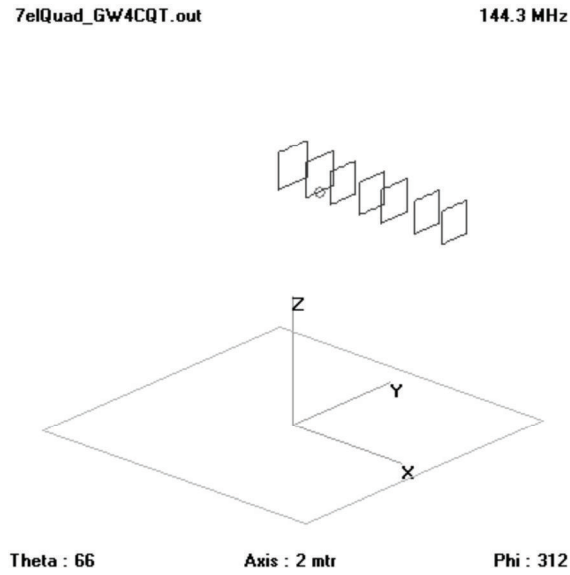


Figure 11.22 4NEC2 model of seven-element (square) loop Yagi. The ground plane characteristics in this case are $\varepsilon_r = 13$ and $\sigma = 5\times10^{-3}$ ℧/m.

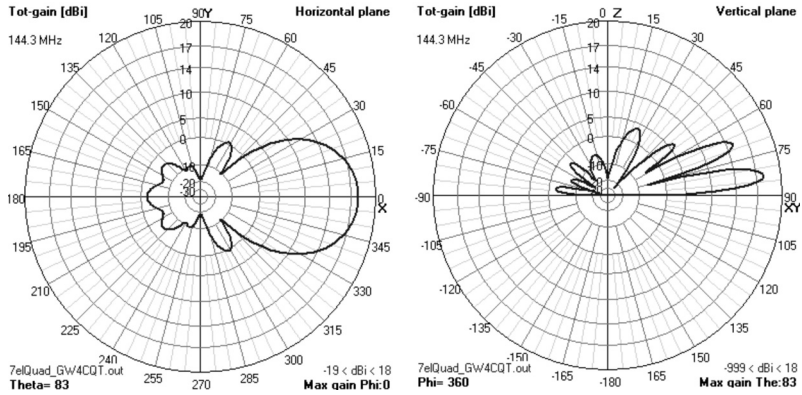


Figure 11.23 2-D patterns of seven-element Yagi array.

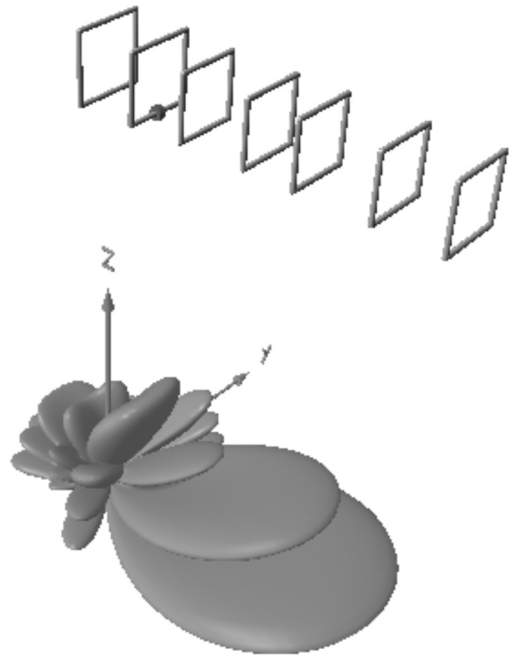


Figure 11.24 3-D pattern of seven-element Yagi array.

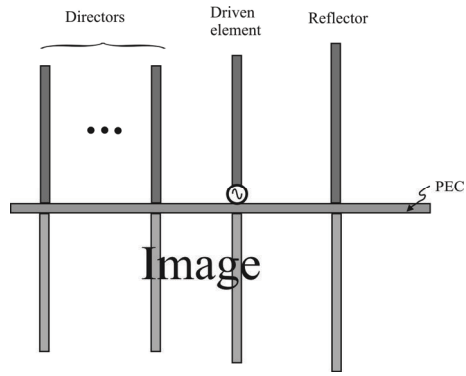


Figure 11.25 Side view of a Yagi monopole array.

An illustration of a five-element Yagi-Uda array of monopoles is shown as Figure 11.26. Forward gains on the order of 10 to 12 dB with front-to-back ratios of 5 to 8 dB may be obtained using such arrangements. One or more of the parasitic monopoles have also been occasionally used for beam shaping.

11.5 Concluding Remarks

We covered the principles of operation of Yagi-Uda antennas in this chapter. This antenna provides relatively high gain with a relatively narrow bandwidth. Its utility in EW systems is primarily as a C2 datalink antenna where the gain can be exploited. It is of use as an intercept or EA antenna as well when the target frequency is unchanging so the antenna can be optimized against that particular target.

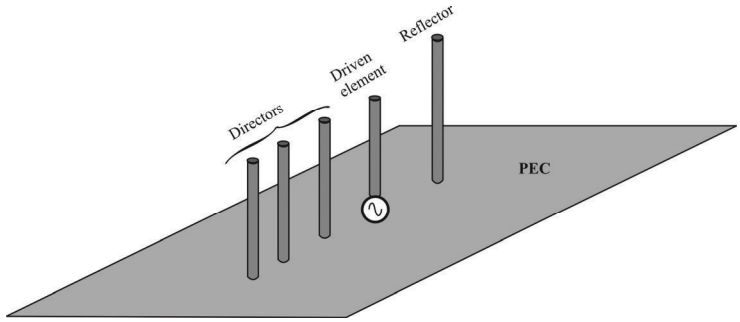


Figure 11.26 Yagi monopole array.

References

[1] Yagi, H., "Bean Transmission of Ultra Short Waves," *Proceedings of the IRE*, Vol. 16, No. 6, June 1928, pp. 715–741.

[2] Thiele, G. A., "Analysis of Yagi-Uda-Type Antennas," *IEEE Transactions on Antennas and Propagation*, Vol. AP-17, No. 1, January 1969, pp. 24–31.

[3] Stutzman, W. L., and G. A. Thiele, *Antenna Theory and Design,* New York: Wiley, 1981, p. 225.

[4] Balanis, C. A., *Antenna Theory*, 3rd ed., New York: Wiley, 2005, p. 593.

[5] Stutzman, W. L., and G. A. Thiele, *Antenna Theory and Design,* New York: Wiley, 1981, pp. 220–222.

[6] Barbano, N., "Log-Periodic Yagi-Uda Array," *IEEE Transactions on Antennas and Propagation*, Vol. AP-14, No. 2, March 1966, pp. 235–238.

[7] http://www.supernec.com/yagi.htm.

[8] http://www.poynting.co.za/tech_training/simpleyagi.shtml.

[9] http://www.ece.msstate.edu/~donohoe/ece4990notes10.pdf.

[10] Appel-Hansen, J., "The Loop Antenna with Director Arrays of Loops and Rods," *IEEE Transactions on Antennas and Propagation*, July 1972, pp. 516–517.

[11] Lindsay, J. E., "A Parasitic End-Fire Array of Circular Loop Elements," *IEEE Transactions on Antennas and Propagation,* Vol. AP-15, September 1967, pp. 697–698.

[12] Shen, L. C., and G. W. Raffoul, "Optimum Design of Yagi Array of Loops," *IEEE Transactions on Antennas and Propagation,* Vol. AP-22, November 1974, pp. 829–830.

[13] Shoamanesh, A., and L. Shafai, "Properties of Coaxial Yagi Loop Arrays," *IEEE Transactions on Antennas and Propagation,* Vol. AP-26, July 1978, pp. 547–550.

[14] Shoamanesh, A., and L. Shafai, "Design Data for Coaxial Yagi Array of Circular Loops," *IEEE Transactions on Antennas and Propagation,* Vol. AP-27, September 1979, pp. 711–713.

Chapter 12

Frequency-Independent EW Antennas

12.1 Introduction

One of the major drawbacks with many antennas is that they have a relatively narrow frequency bandwidth. This is particularly true of the Yagi-Uda antenna discussed in Chapter 11. One design, called the *log-periodic array* (*LPA*), is able to provide directivity and gain while being able to operate over a wide bandwidth. There are other antenna types that are referred to as frequency independent antennas as well, because of their ability to maintain reasonable performance over relatively wide bandwidths. We will present the basic principles of this antenna type in this chapter by examining some examples, including the log-periodic.

An ideal antenna would be frequency independent. That is to say, there are neither low nor high frequency limits. Such an antenna does not exist, however, so there are limits imposed. However, there is no precise definition of what makes an antenna "frequency independent." Clearly, it implies acceptable operation over wide bandwidths. Prior to the 1950s antennas were limited to a bandwidth of about 2:1. Then broad bandwidth antennas were discovered that extended operation over much wider bandwidths. Current definitions make the limit at least 10:1 [1] or 40:1 [2] (these are the ratios of highest to lowest operating frequency).

Antennas that are frequency-independent have geometries that are defined entirely by angles. This is not truly possible in reality, because antennas have to be finite in length, but most of the time, the current in the antenna dies off quite rapidly with distance from the feed, so that the current at the point of this truncation is practically nil. This translates to a very small *end effect,* or the effect that finite truncation of the structure has on its impedance and radiation characteristics.

This chapter begins with a discussion of the general properties of log-periodic antennas. That is followed by a presentation on the *log-periodic dipole array* (LPDA), a widely used broadband antenna. That is followed by a section on the *log-periodic monopole array* (LPMA), which, in turn, is followed by a discussion of the Archimedean spiral antenna. Then the characteristics of the log-periodic spiral antenna are investigated. Next, the planar equiangular spiral antenna is

introduced. Lastly, we include a discussion on the two-arm wire conical spiral antenna.

12.2 Log-Periodic Antennas

The definition of a log-periodic antenna is "an antenna having a structural geometry such that its impedance and radiation characteristics repeat periodically as the logarithm of frequency" [1]. Typical characteristics that exhibit this repetition are input impedance, beamwidth, directivity and sidelobe levels. This periodicity implies variation, but in most cases this variation is small enough that the antennas are considered frequency-independent. In such a case, the number of periodic cycles that the truncation allows determines the overall bandwidth of the structure.

Log-periodic antennas have been around for many years and have been valued for their very broadband behavior. They were developed as part of the same revolutionary research at the University of Illinois in the 1950s and 1960s that produced the first frequency-independent antennas. Much of this research was funded by the United States Army and Air Force—the former to reduce the number of antennas mounted on tactical vehicles, the latter to reduce the overall number of antennas needed for each aircraft.

The most widely known log-periodic antenna is the LPDA array, its most popular historical application being the VHF to UHF frequency spectrum for television use [1]. The toothed planar log-periodic antenna published by R. H. DuHamel and D. E. Isbell in 1957 was a direct predecessor of this design. It was one of the earliest log-periodic antennas in existence [2].

12.3 Log-Periodic Dipole Array

The LPDA is used in a number of applications where wide bandwidth is required along with directivity and a modest level of gain (but less than the Yagi). It is used at VHF and UHF for a variety of applications. It has been used extensively as an EW antenna in the VHF and UHF frequency ranges where tactical mobility is required. EW HF log-periodic antennas have been deployed when mobility is not an issue, as these antennas in that frequency range are quite large. Some EW targets satisfy this criterion, as long-distance HF communication systems require large antennas as well. The LPDA has frequently been used as a high gain, fixed site communication antenna. A diagram of an LPDA along with representative radiation patters are displayed in Figure 12.1 [3].

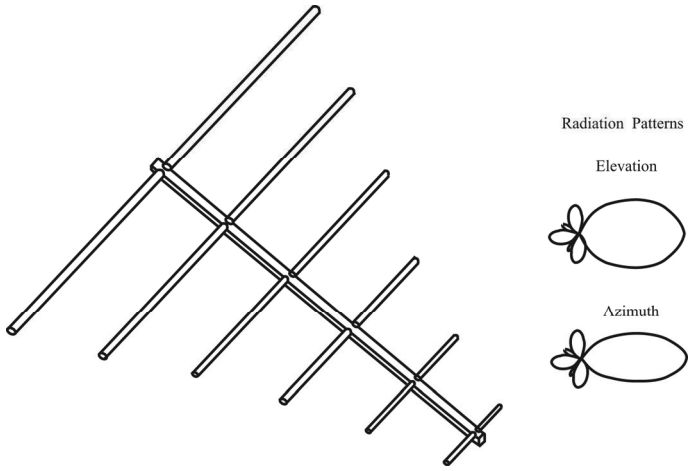


Figure 12.1 Log periodic dipole array. This type of antenna is an example of a frequency-independent antenna. Frequency ranges of 2:1 or more are possible with a single antenna. (Source: [3], © Artech 2002. Reprinted with permission.)

The LPDA antenna is directional and is normally capable of operating over a frequency range of about 2:1 or more. It has many similarities to the Yagi, to include its looks, because it exhibits forward gain and has a significant front to back ratio. In addition to this, the radiation pattern stays roughly the same over the frequency bandwidth, as do parameters like the radiation resistance and the VSWR. However, it offers less gain for its size than the Yagi.

There are several types of log-periodic antennas that can be discussed. They are simple and easily constructed, exhibit moderate gain, and have significant bandwidth [4]. The LPDA consists of a number of dipole elements. These diminish in size from the back towards the front; the main beam of the antenna coming from the smaller front. The element at the back of the array where the elements are the largest is a half-wavelength long at the lowest frequency of operation. The element spacings also decrease towards the front of the array where the shortest elements are located. In operation, as the frequency changes there is a smooth transition along the array of the elements that form the active region (defined below). To ensure that the phasing of the different elements is correct, the feed phase is reversed from one element to the next.

The LPDA exhibits a relatively low VSWR (usually not greater than 2 to 1) over a wide band of frequencies. An LPDA can yield a 1.3-to-l VSWR over a 1.8-to-1 frequency range with a typical directivity of 12.5 dB. The LPDA beamwidth in the elevation plane is seen in Figure 12.2 to be approximately 180^{o}, while that in the horizontal planes is approximately 60^{o}. They exhibit maximum gains of

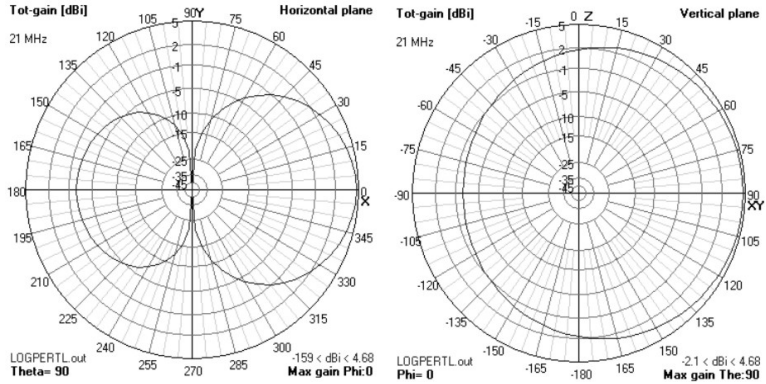


Figure 12.2 Five-element LPDA 2-D patterns.

approximately 6–10 dBi. The effective area of an LPDA is therefore approximately $A_{\text{eff}} = 4\lambda^2/4\pi$. The radiation pattern in 3-D is shown in Figure 12.3.

12.3.1 Operation of the LPDA

The log-periodic dipole antenna basically behaves like a Yagi-Uda array over a wide frequency range. As the frequency varies, the active set of elements for the

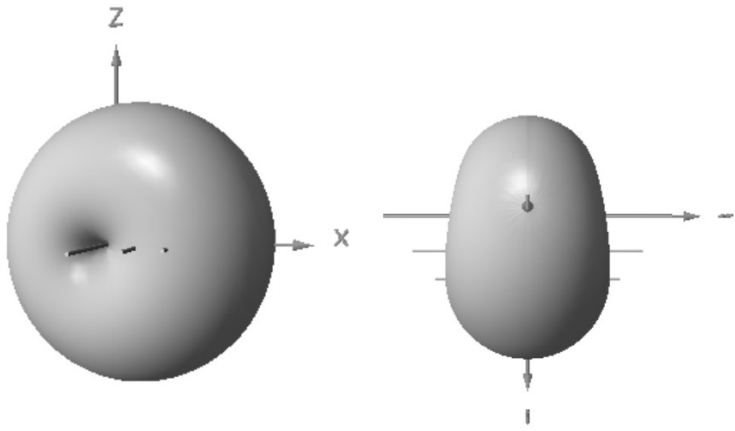


Figure 12.3 Five-element LPDA 3-D patterns: left, side view, right, top view. Antenna is horizontally polarized in this case.

log-periodic antenna (those elements which carry the significant current) moves from the long-element end at low frequency to the short-element end at high frequency. Because they are somewhat shorter than the driven element, the director element current in the Yagi array lags that of the driven element, while the reflector element current leads that of the driven element because it is somewhat longer. This current distribution in the Yagi array points the main beam in the direction of the director.

The principles of the LPDA can perhaps be best understood by comparing it to the Yagi. Since the dipoles are different lengths they each will resonate at a different frequency. Say element n is at resonance. Then element $n - 1$ is slightly longer (about 5%) than element n and serves as the reflection element for element n. Similarly, element $n + 1$ is a little shorter (5%) than element n and serves as its director. Thus, the entire array acts like several Yagi antennas pointed in the same direction. This is how the broadband operation is obtained; at any given (small) range of frequencies, three of the elements are acting like a three-element Yagi antenna and the others are not of the correct length to contribute significantly to the radiation pattern in that range of frequencies.

The LPDA consists of a system of driven elements, but not all elements in the system are active on a single frequency of operation. As mentioned, depending upon its design parameters, the LPDA can be operated over a range of frequencies having a ratio of 2:1 or higher, and over this range its electrical characteristics—gain, feed-point impedance, front-to-back ratio—will remain more or less constant. This is not true of any multi-element directive array antenna, for either the gain factor or the front-to-back ratio or both deteriorate rapidly as the frequency of operation departs from the design frequency of the array. And with resonant antenna structures, off-resonance operation introduces reactance which causes the VSWR in the feeder system to increase.

Since the LPDA acts essentially like the three-element Yagi once the frequency is specified, we would expect the basic characteristics such as radiation patterns and input impedance would be similar between the two. Therefore the element pattern and array factor would be similar to (11.1) and (11.2), respectively.

The geometry of an LPDA is shown in Figure 12.4, with relevant parameters identified. We will discuss all of these parameters subsequently.

12.3.1.1 Active Region

The feeder polarity is reversed between successive elements (see Figure 12.4). Consider the condition when the antenna is approximately in the middle of its operating range. When the signal meets the first few elements, since they are spaced quite close together in terms of the operating wavelength, the fields from

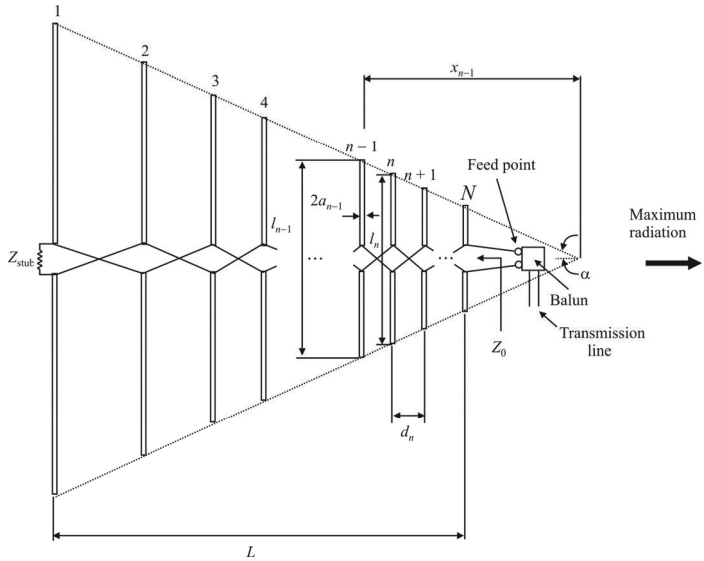


Figure 12.4 Log-periodic antenna. With the Z_{stub} at the end, the LPDA is a form of traveling wave antenna since there are no reflections. Not all log-periodic antennas are so terminated.

these elements will cancel one another out as the feeder sense is reversed between the elements. Then as the signal progresses down the antenna a point is reached where the feeder reversal and the distance between the elements gives a total phase shift of about 360°. At this point the effect that is seen is that of two phased dipoles. The region in which this occurs is called the *active region* of the antenna. Although the example of only two dipoles is given, in reality the active region can consist of more elements. The actual number depends upon the angle α and a design constant.

The elements outside the active region receive little direct power. Despite this, we find that the larger elements are resonant below the operational frequency and appear inductive. Those in front resonate above the operational frequency and are capacitive. These are exactly the same criteria that are found in the Yagi.

The LPDA is frequency-independent in that the electrical properties such as the mean resistance level, R_0, characteristic impedance of the feed-line Z_0, and driving-point admittance Y_0, vary periodically with the logarithm of the frequency. As the frequency f_1 is shifted to another frequency f_2 within the passband of the antenna, the relationship is $f_2 = f_1/\tau$, (τ is a design parameter specified below), where

$$\begin{aligned} f_3 &= f_1 / \tau^2 \\ f_4 &= f_1 / \tau^3 \\ &\vdots \\ f_n &= f_1 / \tau^{n-1} \end{aligned} \tag{12.1}$$

The phase relationship that exists in the set of dipoles in the active region must include an active region of dipoles for the highest and lowest design frequency of the antenna. It has a bandwidth which we shall denote by W_{ar} (bandwidth of the active region).

In order to obtain the same phasing in the log-periodic antenna with all of the elements in parallel, the source would have to be located on the long-element end of the array. However, at frequencies where the smallest elements are resonant at $\lambda/2$, there may be longer elements which are also resonant at lengths of $m\lambda/2$. Thus, as the power flows from the long-element end of the array, it would be radiated by these long resonant elements before it arrives at the short end of the antenna. For this reason, the log-periodic dipole array must be driven from the short-element end, but this arrangement gives the exact opposite phasing required to point the beam in the direction of the shorter elements. This problem is solved by alternating the connections from element to element. Such phasing of the log-periodic dipole elements points the beam in the proper direction.

Sometimes, the log-periodic antenna is terminated on the long-element end of the antenna with a transmission line and load. This is done to prevent any energy that reaches the long-element end of the antenna from being reflected back toward the short-element end. This makes such an arrangement a traveling wave antenna (Chapter 8). For the ideal log-periodic array, not only should the element lengths and positions follow the scale factor τ, but the element feed gaps and radii should also follow the scale factor. In practice, the feed gaps are typically kept at a constant spacing. If different radii elements are used, two or three different radii are used over portions of the antenna.

12.2.3 LPDA Fundamental Parameters

A fundamental parameter of the LPDA is the *shape factor* and is denoted by τ. It is related to the dimensions of the antenna as shown in Figure 12.4 by

$$\tau = \frac{l_{n+1}}{l_n} = \frac{a_{n+1}}{a_n} = \frac{x_{n+1}}{x_n} \tag{12.2}$$

so that

$$l_{n+1} = \tau l_n \tag{12.3}$$

τ depends on the bandwidth of the antenna and the number of elements. It can be calculated as

$$\tau = \left(\frac{l_N}{l_1} \right)^{1/N-1} \tag{12.4}$$

Truncation coefficients are used to establish the minimum value for l_1 (the longest dipole element) and the maximum value for l_N (shortest dipole element) as

$$l_1 = k_1 \lambda_{\max} \tag{12.5}$$

$$l_N = k_2 \lambda_{\min} \tag{12.6}$$

where $k_1 = 0.5$ and $k_2 = 0.25$ are typical values. The performance of such antennas is best determined, and these coefficients adjusted accordingly, by modeling.

Once l_1 and l_N are determined, the number of dipoles can be determined with

$$N = 1 + \frac{\log_{10}(l_1 / l_N)}{\log_{10}(1/\tau)} \tag{12.7}$$

The *spacing factor* of an LPDA is given by

$$\sigma = \frac{d_n}{2l_n} = \frac{1-\tau}{4\tan\alpha} \tag{12.8}$$

Thus,

$$\alpha = \tan^{-1} \frac{1-\tau}{4\sigma} \tag{12.9}$$

The basic design chart relating gain (directivity), τ, and scale factor, σ, is given in Figure 12.5 [5]. To design an LPDA, first the optimum values of σ and τ are found based on the required gain from Figure 12.5. (Initially, it is assumed that the efficiency is unity, so that $D = G$.) Then the longest element is determined from (12.5) then the shortest element is determined from (12.6). The lengths of the elements are computed one at a time using (12.3) until the shortest length is reached. Since the antenna bandwidth is determined by the longest and shortest elements, the bandwidth is only limited by the number of elements included. Of

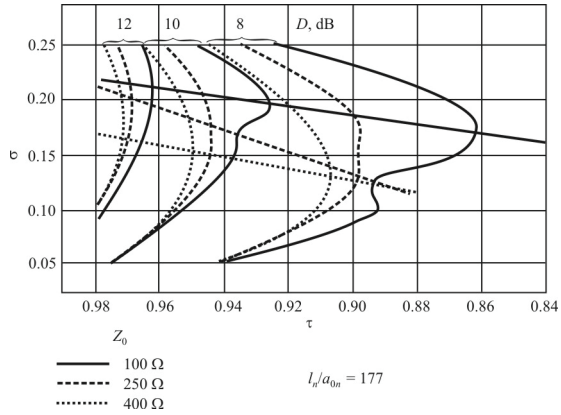


Figure 12.5 Optimal values of σ and τ for LPDAs based on the corrected chart of Peixerio.

course, structural limitations (space, strength) will ultimately put an end to the number of elements that can be included.

12.3.3 Feed Arrangements

As we see in Figure 12.4, the log-periodic array consists of several dipole elements each of which is of different lengths and different relative spacings (and potentially different radii). A distributive type of feeder system is used to excite the individual elements. The element lengths and relative spacings, beginning from the feed point for the array, are seen to increase smoothly in dimension, being greater for each element than for the previous element in the array. It is this feature upon which the design of the LPDA is based and which permits changes in frequency to be made without greatly affecting the electrical operation. With changes in operating frequency, there is a smooth transition along the array of the elements that comprise the active region.

The log-periodic dipole antenna presents a number of difficulties if it is to be fed properly. The feed impedance is dependent upon a number of factors. However, it is possible to control this by altering the spacing and hence the impedance for the feeder that connects each of the dipole elements together. Despite this the impedance varies with frequency, but this can be overcome to a large extent by making the longer elements out of a larger diameter rod. Even so the final feed impedance does not normally match to 50Ω on its own. It is normal for a further form of impedance matching to be required. This may be in the form of a stub or a transformer (balun, see Appendix A). The actual method employed

will depend to a large degree on the application of the antenna and its frequency range.

12.3.4 Active Region Revisited

The operation of the LPDA can be understood with an example. Assume for the moment that the LPDA has 12 elements. Currents flowing in the elements are both real and imaginary, the real current flowing in the resistive component of the impedance of a particular dipole, and the imaginary flowing in the reactive component. Suppose that the frequency is such that element number 6 is near to being half-wave resonant. The imaginary parts of the currents in shorter elements 7 to 12 are capacitive, while those in longer elements 1 to 6 are inductive. The capacitive current components in shorter elements 9 and 10 exceed the conductive components so these elements receive little power from the feeder and act as parasitic directors. The inductive current components in longer elements 4 and 5 are dominant and they act like parasitic reflectors. Elements 7 and 8 receive most of their power from the feeder (element 6) and act like driven elements. The amplitudes of the currents in the remaining elements are small and may be ignored as primary contributors to the far radiation field. The result is that we have a generalized Yagi array with seven elements comprising the active region. It should be noted that this active region is for a specific set of design parameters ($\tau = 0.93$, $\sigma = 0.175$). The number of elements making up the active region will vary with τ and σ. Adding additional elements on either side of the active region cannot significantly modify the circuit or field properties of the array.

This active region determines the basic design parameters for the array, and sets the bandwidth for the structure, W_s. That is, for a design frequency coverage of bandwidth W, there exists an associated bandwidth of the active region such that

$$W_s = W \times W_{ar} \tag{12.10}$$

where W is the operating bandwidth defined as

$$W = \frac{f_H}{f_L} \tag{12.11}$$

where

f_H is the highest frequency.
f_L is the lowest frequency.

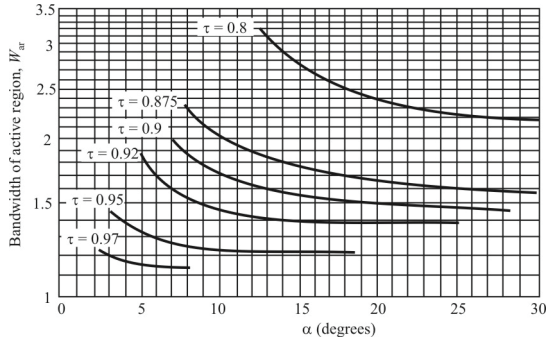


Figure 12.6 Bandwidth of active region.

W_{ar} varies with τ and σ as shown in Figure 12.6 [6]. Elements that fall outside W_{ar} play an insignificant role in the operation of the active region subarray. The designed bandwidth W_s is given by the following empirical equation:

$$W_s = \frac{f_H}{f_L}\left[1.1 + 7.7(1-\tau)^2 \cot\alpha\right] \tag{12.12}$$

The overall length of the array from the shortest element to the longest element (L) is given by

$$L = \frac{l_1}{2}\left(1 - \frac{1}{W_s}\right)\cot\alpha \tag{12.13}$$

where

$$l_1 = \frac{\lambda_{max}}{2} = \frac{c}{2f_L} \tag{12.14}$$

There exists an optimum value for σ, σ_{opt}, for each τ in the range $0.8 \le \tau < 1.0$, for which the gain is maximum; however, the increase in gain achieved by using σ_{opt} and τ near 1.0 (i.e., $\tau = 0.98$) is only 3 dBi when compared with the minimum σ ($\sigma_{min} = 0.05$) and $\tau = 0.9$, shown in Figure 12.7 [6].

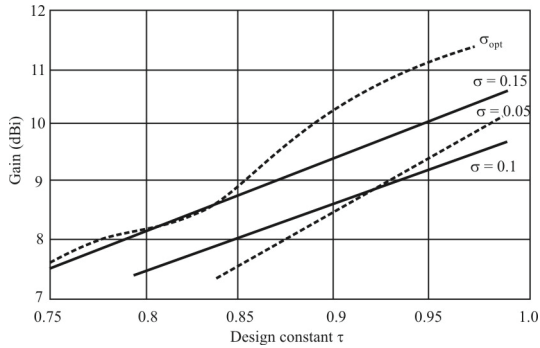


Figure 12.7 Gain as a function of τ for an LPDA.

An increase in τ means more elements and optimum σ means a long boom. A high-gain (12.5 dBi) LPDA can be designed in the HF region with $\tau = 0.9$ and $\sigma = 0.05$. The design variables τ, σ, and α are related as follows:

$$\sigma = \frac{1}{4}(1-\tau)\cot\alpha \tag{12.15}$$

where α is the apex angle shown in Figure 12.4, τ is the design constant, and σ is the relative spacing constant. In addition,

$$\sigma = \frac{d_{n,n-1}}{2/n-1} \tag{12.16}$$

$$\sigma_{\text{opt}} = 0.258\tau - 0.066 \tag{12.17}$$

12.3.5 Dipole Feed Design

The method of feeding the antenna is rather simple. As shown in Figure 12.4, a balanced feeder is required for each element, and all adjacent elements are fed with a 180° phase shift by alternating element connections. In this section the term antenna feeder is defined as the line that connects each adjacent element. The feed-line is that line between antenna and transmitter. The characteristic impedance of the antenna feeder, Z_0, must be determined so that the feed-line impedance and type of balun can be determined. The antenna-feeder impedance Z_0 depends on the mean radiation resistance level R_0 (required input impedance of the active region elements—see Figure 12.8 [6]) and average characteristic impedance

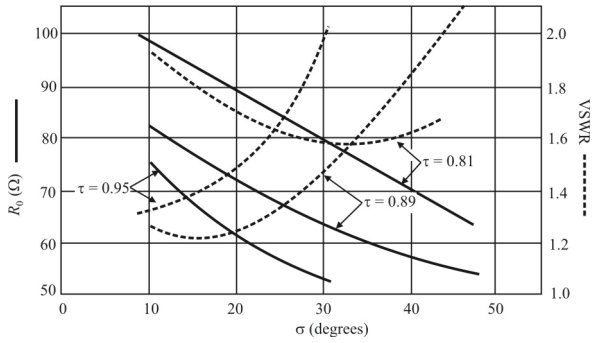


Figure 12.8 R_0 and VSWR versus σ for an LPDA.

of a dipole, Z_d. (Z_d is a function of element radius a and the resonant element half length, h, is $h = \lambda / 4$. See Figure 12.4.) The relationship is

$$Z_0 = \frac{R_0^2}{8\sigma' Z_d} + R_0 \sqrt{\left(\frac{R_0}{8\sigma' Z_d}\right)^2 + 1} \tag{12.18}$$

with

$$Z_d = 120\left(\ln\frac{h}{a} - 2.55\right) \tag{12.19}$$

and the mean spacing factor, σ', is given by

$$\sigma' = \frac{\sigma}{\sqrt{\tau}} \tag{12.20}$$

Z_d is shown in Figure 12.9 [6] as a function of the thickness ratio h/a.

From Figure 12.8 we see that R_0 decreases with increasing τ and increasing σ. Also the VSWR with respect to R_0 has a minimum value of about 1.1 to 1 at σ_{opt}, and a value of 1.8 to 1 at $\sigma = 0.05$. These SWR values are acceptable when using standard RG8/U 50Ω and RG-11/U 75Ω coax for the feed-line.

Once the value of Z_0 has been determined for each band within the array passband, the balun and feed-line may be chosen. That is, if $Z_0 = 100\Omega$, a good choice for the balun would be 1-to-1 balanced to unbalanced, and 75Ω coax feed-

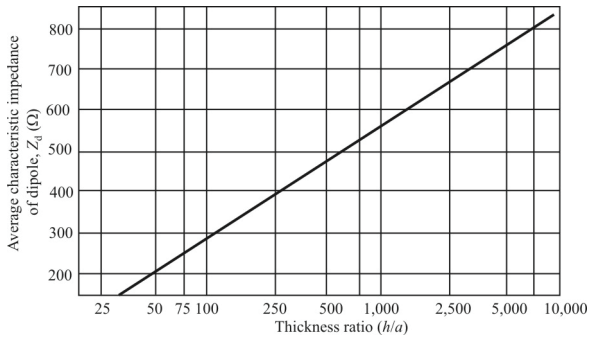


Figure 12.9 Average characteristic impedance, Z_d, for an LPDA.

line. If $Z_0 = 220\Omega$, choose a 4-to-1 balun, and 50Ω coax feed-line, and so on. The balun may be omitted if the array is to be fed with an open-wire feed-line (such as twin-lead connections on TV antennas).

The terminating impedance, Z_{stub}, may be omitted. If it is omitted, the LPDA is a (multi-) resonant antenna while if it is used, as mentioned, the LPDA is a form of traveling wave antenna as discussed in Chapter 8 since there are no standing waves produced. However, if the terminating impedance is used, it should have a length no longer than $\lambda_{max}/8$. The terminating impedance tends to increase the front-to-back ratio for the lowest frequency used. For HF-band operation a 6-inch shorting jumper wire may be used for Z_{stub}. When Z_{stub} is simply a short-circuit jumper, the longest element behaves as a passive reflector. It also might be noted that we could increase the front-to-back ratio on the lowest frequency by moving the passive reflector (element # 1) a distance of 0.15 to 0.25λ behind element 2, as would be done in the case of an ordinary Yagi parasitic reflector. This, of course, would necessitate lengthening the boom. The front-to-back ratio increases somewhat as the frequency increases. This is because more of the shorter inside elements form the active region, and the longer elements become additional reflectors.

12.3.6 Design Procedure

Based on the preceding discussion, a design procedure for an LPDA is given in Figure 12.10 [7].

A photograph of an LPDA is shown in Figure 12.11 [7]. Its overall chraceteristics are given in Table 12.3. We can see that this antenna has acceptable performance over a frquency span of over 20:1. Its average gain over this range is

about 6 dB. The 3 dB beamwidth in the E-field (vertical) plane is 35° while its 3 dB beamwidth in the horizontal plane is 100°. The feed-line impedance is 50Ω.

The measured performance is indicated in Figures 12.12–12.14. Figure 12.12 shows the array factor as a function of frequency. The half-power beamwidth in the E-field plane is shown in Figure 12.13, while the gain versus frequency performance is seen in Figure 12.14.

12.4 Log-Periodic Monopole Array

An LPMA over a PEC ground plane is shown in Figure 12.15. The side view schematic of the LPMA, shown in Figure 12.16, shows how the image generated by the ground plane causes the LPMA to appear much like an LPDA with the exception of its feed structure. τ and σ are the standard defining parameters for log-periodic antennas and are defined by the equations given in the figure. These are the same τ and σ that are given for the LPDA (they are based on $\lambda/4$ lengths as opposed to $\lambda/2$ lengths) [8].

The primary difference between an LPDA and an LPMA is in the feed network. Unlike the LPDA, there is not a single standard feed that is used. A convenient feed for high frequencies is the modulated feed-line structure. It consists of cells that have two different impedances separated by a step discontinuity. The characteristic impedance (Z_0) section is generally the same as the source and load (Z_S and Z_L). The modulated impedance (Z_m) is normally higher and is connected to the elements. By varying:

1. The percentage of each feed cell that is modulated (k);
2. The modulated impedance (Z_m);
3. The length of the total feed (l_{feed});

we can attempt to match the LPMA and the source for a given LPMA physical structure. After matching, the characteristic impedance sections can be varied in length by the phase factor to change the pattern itself with little impact on the VSWR. This feed network is illustrated in Figure 12.17.

A different feed technique using transformers at each monopole is illustrated in Figure 12.18(a) [9]. This antenna operates over 4–9 MHz. Figure 12.18(a) shows an antenna feed line using transformers and coaxial cable. Only transformers 1, 2, 10, and 11 are shown where 1 is located with the tallest monopole at the rear of the antenna. The LPMA antenna is 77 m long and consists

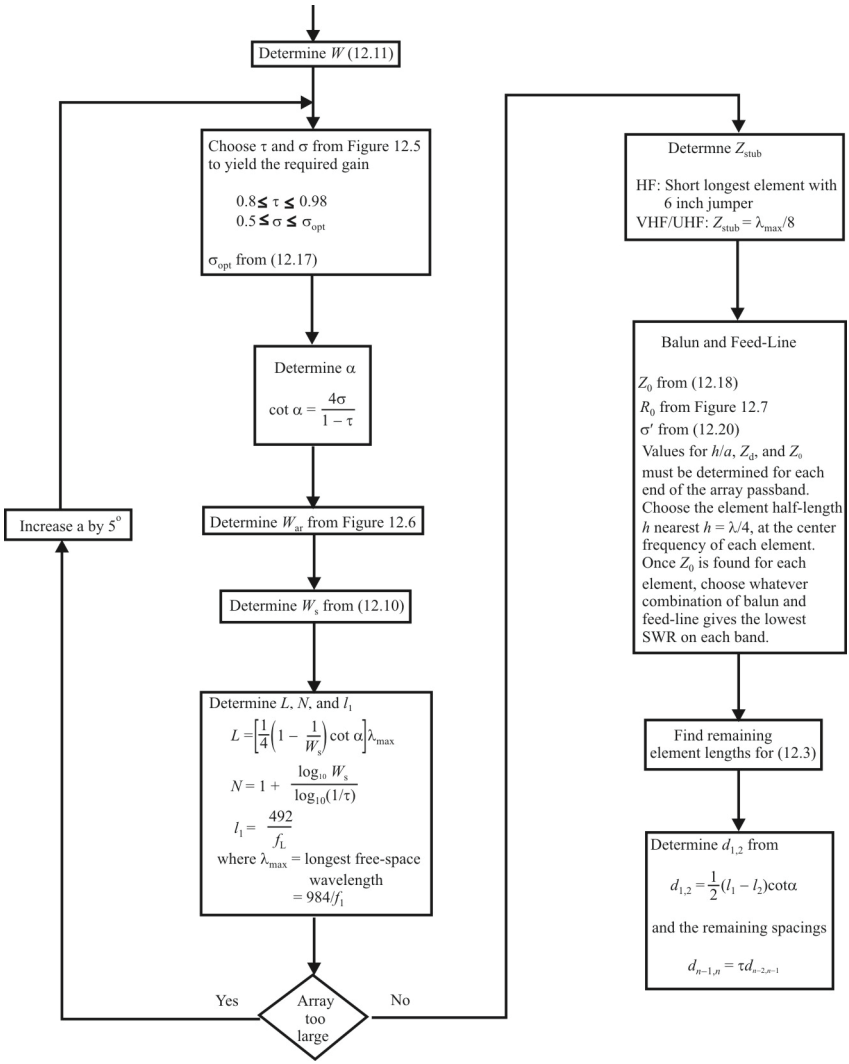


Figure 12.10 LPDA design procedure.

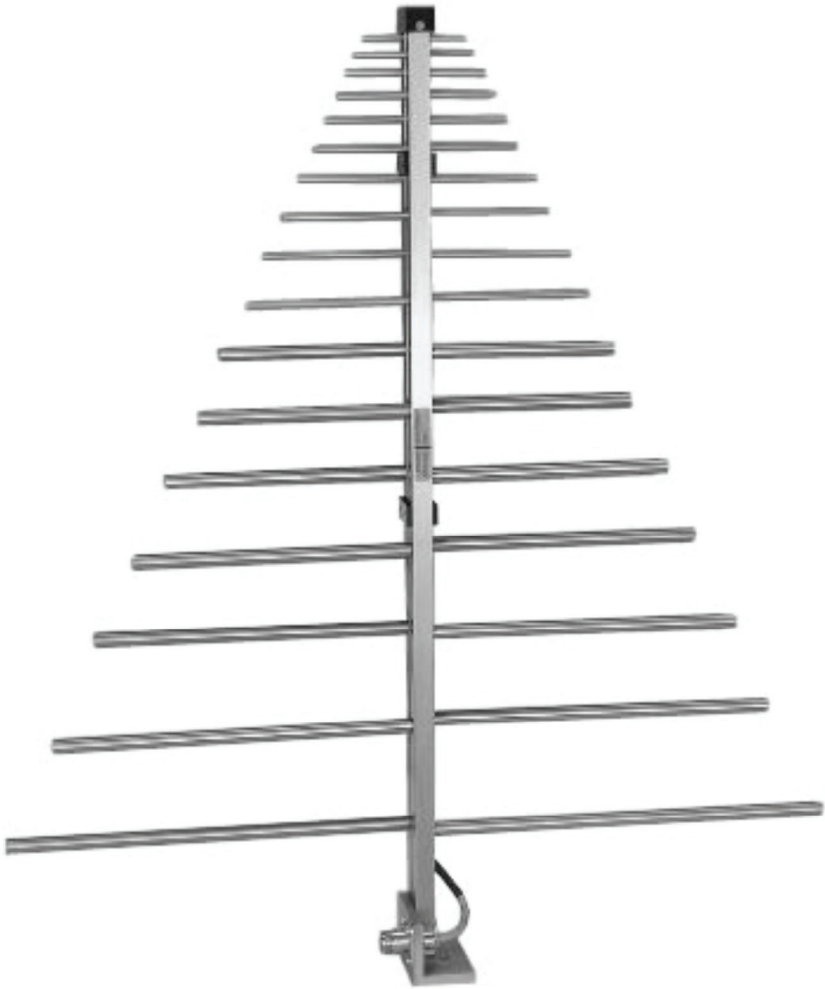


Figure 12.11 Picture of the AH Systems SAS-510-7 LPDA. (Courtesy of AH Systems [7].)

Table 12.3 Overall Specifications of the AH Systems SAS-510-7 LPDA

Parameter	Value
Frequency Range	290–7,000 MHz
Antenna Factor	15–47 dB
Gain	5.8 dBi
Maximum Continuous Power	1,000 W
Maximum Radiated Field	200 V m^{-1}
3 dB Beamwidth (E-Field)	35°
3 dB Beamwidth (H-Field)	100°
Impedance	50 Ω
VSWR	1.65: 1 typ, 2.5:1 max.
Length	63.2 cm
Width	51.1 cm
Weight	0.68 kg

Source: [8].

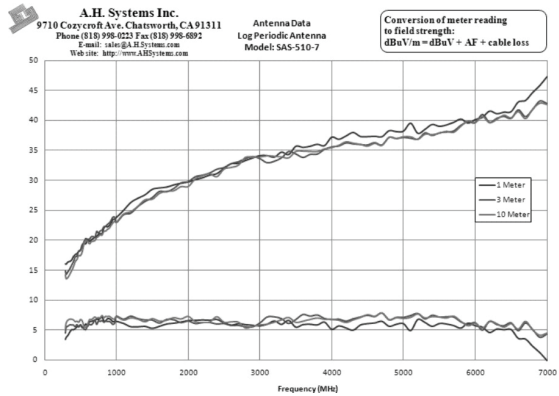


Figure 12.12 LPDA array factor. (Courtesy of AH Systems [7].)

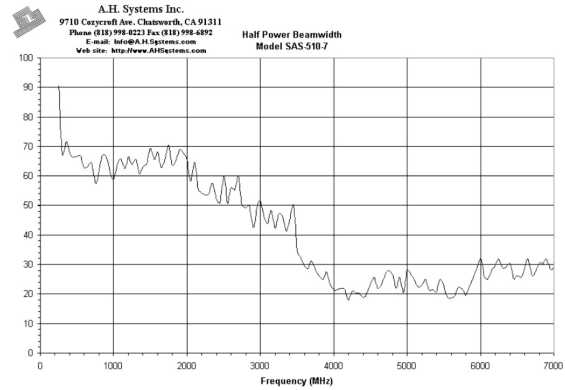


Figure 12.13 LPDA half-power beamwidth (E-field). (Courtesy of AH Systems [7].)

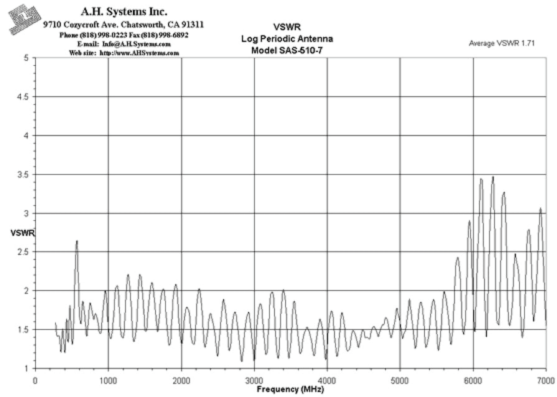


Figure 12.14 LPDA VSWR. (Courtesy of AH Systems [7].)

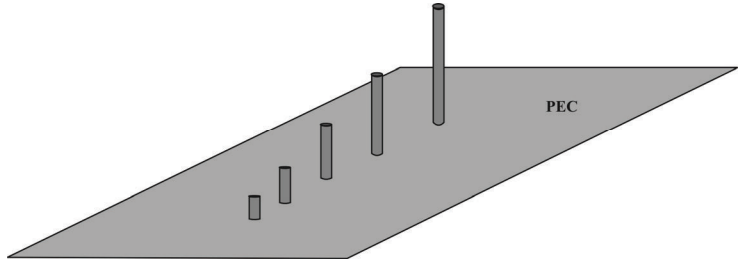


Figure 12.15 Log-periodic monopole array (LPMA).

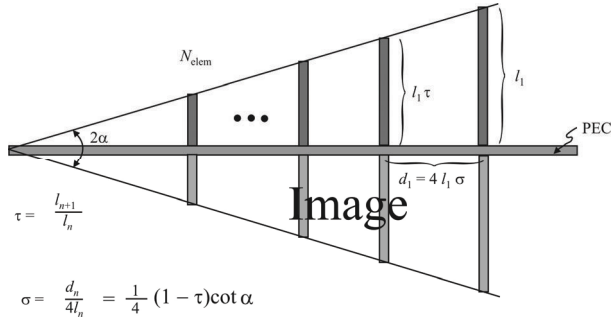


Figure 12.16 Side view of LPMA.

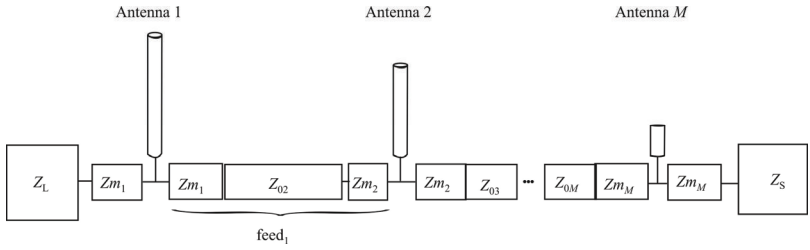


Figure 12.17 LPMA feed network.

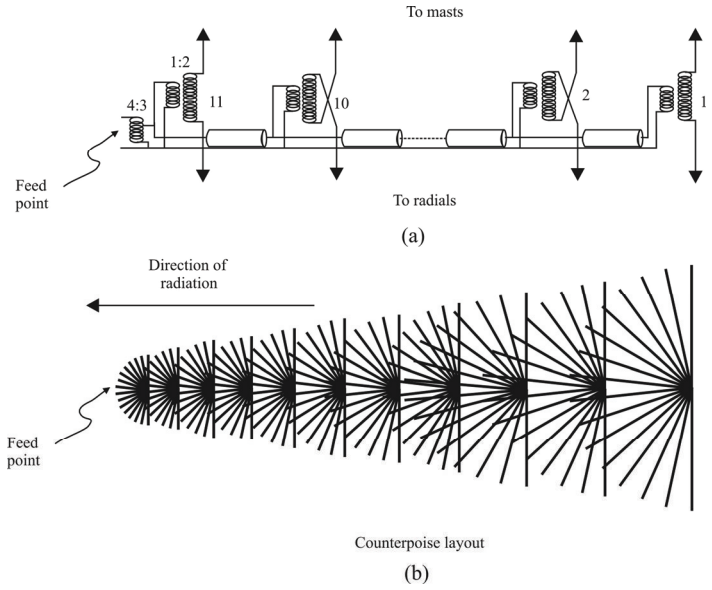


Figure 12.18(a, b) LPMA transformer feed.

of 11 monopoles whose heights range from 5.1 to 17.4 m. Each monopole has a matching counterpoise consisting of 16 wires as shown in Figure 12.18(b). The lengths of the counterpoise wires equal each pole's height.

NEC modeling indicates that input VSWR is less than 1.8:1 over 97% of all operating frequencies and the maximum directivity is between 2 and 8 dBi depending on ground type.

12.4.1 Design Procedure

The design procedure for the LPMA is essentially the same as the one above for the LPDA. The major difference is in the design of the feed networks, which would be customized depending on the application.

12.5 Archimedean Spiral Antenna

The Archimedean spiral antenna is a popular form of frequency independent antenna for higher frequency ranges. Its utility at lower frequencies is limited due to size and weight constraints. The Archimedean spiral is typically backed by a lossy cavity to achieve frequency bandwidths of 9:1 or greater. Specifically we will examine the planar structure of these antennas. A popular variation is to implement spiral antennas on a cone that provides similar but somewhat different properties. Since the results are similar, we will not examine those structures.

12.5.1 Principles of Operation

A two-arm self-complementary Archimedean spiral antenna [10] is shown in Figure 12.19. The antenna is fed in the middle and radiation is out of the page.

Definition: A spiral antenna is *self-complementary* if the metal and air regions of the antenna are equal.

The Archimedean spiral antenna is an example of a self-complementary structure.

The input impedance of a self-complementary antenna is

$$Z_{\text{metal}} Z_{\text{air}} = \frac{\eta^2}{4} \tag{12.21}$$

where η is the characteristic impedance of the medium surrounding the antenna. For a self-complementary Archimedean spiral antenna in free space the input

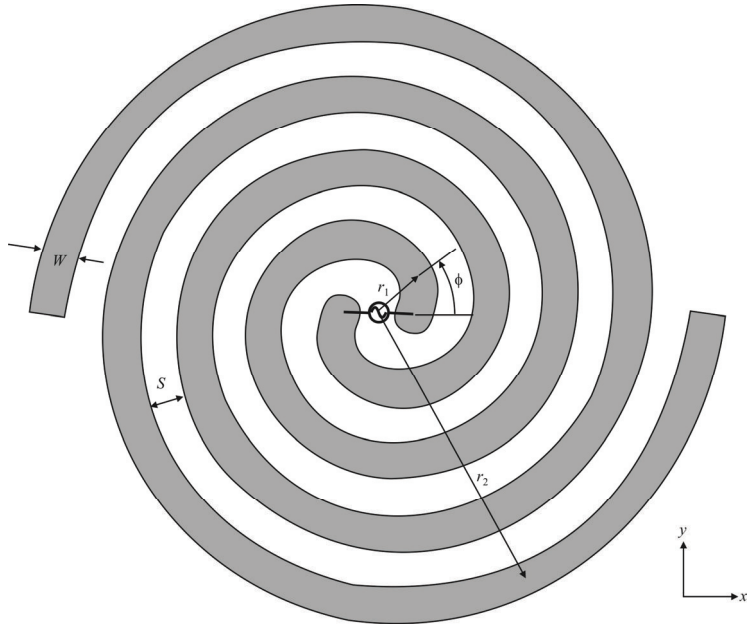


Figure 12.19 Archimedean spiral antenna. Propagation is out of the page with circular or elliptical polarization.

impedance is

$$Z_{\text{in}} = \frac{\eta_0}{2} = 188.5\Omega \tag{12.22}$$

The polarization of this antenna is elliptical/circular.

Each arm of an Archimedean spiral is linearly proportional to the angle, ϕ, shown in Figure 12.19, and is given by

$$r = r_0\phi + r_1 \qquad \text{and} \qquad r = r_0(\phi - \pi) + r_1 \tag{12.23}$$

where r_1 is the inner radius of the spiral shown in Figure 12.19. The proportionality constant, r_0, for a self-complementary spiral is given by

$$r_0 = \frac{S + W}{\pi} = \frac{2W}{\pi} \tag{12.24}$$

while the strip width of each arm is

$$S = \frac{r_2 - r_1}{2N} - W = W \tag{12.25}$$

The spacing and width can be determined from

$$S = W = \frac{r_2 - r_1}{4N} \tag{12.26}$$

where r_2 is the outer radius of the spiral and N is the number of turns.

The above equations apply to a two-arm Archimedean spiral. When there are four arms, the arm width is given by

$$W^{\text{4-arm}} = \frac{r_2 - r_1}{8N} \tag{12.27}$$

and the proportionality constant is

$$r_0^{\text{4-arm}} = \frac{4W}{\pi} \tag{12.28}$$

The Archimedean spiral antenna radiates from a region where the circumference of the spiral equals one wavelength. Similar to the LPDA structure described above, this is called the active region of the spiral. The arms of the spiral are fed 180^{o} out of phase, so when the circumference of the spiral is one wavelength the currents at complementary or opposite points on each arm of the spiral add in phase in the far field.

Based on this view, palpably the low-frequency operating point of the spiral is determined theoretically by the outer radius and is given by

$$f_{\text{low}} = \frac{c}{2\pi r_2} \tag{12.29}$$

where c is the speed of light. In practice the low-frequency point will be higher than predicted by (12.29) due to reflections from the end of the spiral. The reflections can be minimized by using resistive loading at the end of each arm or by adding conductivity loss to some part of the outer turn of each arm, turning the antenna into a type of traveling wave antenna.

Likewise, the high-frequency operating point is determined by the inner radius

$$f_{\text{high}} = \frac{c}{2\pi r_1} \tag{12.30}$$

The high-frequency limit may be less than found from (12.30) due to feed region effects.

A few examples of cavity-backed spiral antennas are shown in Figure 12.20 [11]. A specific antenna is shown in Figure 12.21, while the radiation pattern of this antenna is illustrated in Figure 12.22. This antenna covers the frequency range 0.5–40 GHz in one antenna with a beamwidth of $\pm 50^{o}$ at the frequency corresponding to Figure 12.22.

12.6 Log-Periodic Spiral Antennas

A variant on the Archimedean spiral configuration that provides even greater bandwidth performance is the log-periodic spiral antenna. The configuration for this architecture is illustrated in Figure 12.23.

Efficient spiral log-periodic structures can be effective with very few arms. The antenna shown in Figure 12.23 is such a configuration with only 1.5 spirals in each of two arms. The frequency range of this antenna is 0.4–3.8 GHz.

Figure 12.20 Cavity-backed spiral antennas. (Source: [11] © Cobham Sensor Systems. Reprinted with permission.)

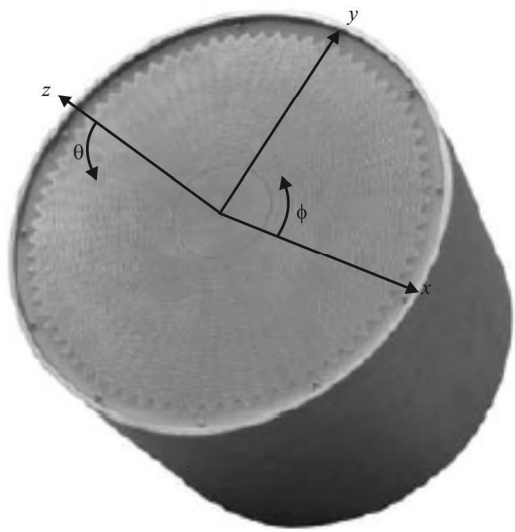


Figure 12.21 Cavity-backed spiral antenna.

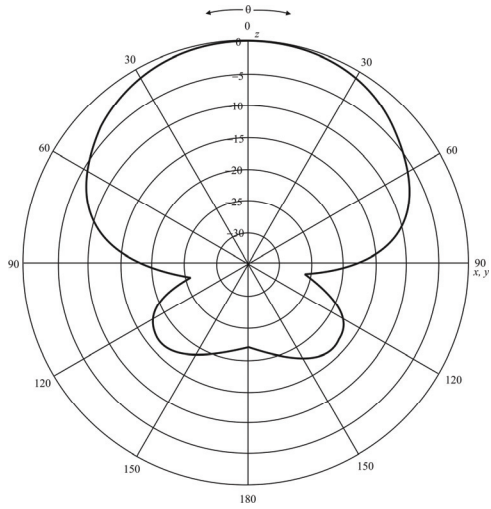


Figure 12.22 Spiral antenna radiation pattern in the *x*-*z* and *y*-*z* plane.

Figure 12.23 Spiral small log-periodic. (Source: [12] © Applied Microwaves and Wireless. Reprinted with permission.)

12.6.1 Principles of Operation

The logarithmic spiral antenna [12] is designed using the equations

$$r_1 = r_0 e^{a\phi} \tag{12.31}$$

and

$$r_2 = r_0 e^{a(\phi-\phi_0)} \tag{12.32}$$

where r_1 and r_2 are the outer and inner radii of the spirals, respectively; r_0 and $r_0 e^{a(-\phi_0)}$ are the initial outer and inner radii; a is the growth rate; and ϕ is the angular position. To obtain the most frequency independent radiation pattern and the most constant input impedance, the dimensions are $r_0 = 2.1$ mm, $a = 0.5$ rad^{-1}, and $\phi_0 = 1.3$ rad $= 75^o$.

The width of the arms are the same as the spacing between the arms making the structure self-complementary, which gives the most frequency-independent parameters. Good radiation patterns can usually be obtained with as few as 1 to 1.5 turns of the spiral. For this antenna, the ends of the two spiral arms are truncated to produce the smallest physical antenna for a given lower resonant frequency. Alternatively, the end of the spiral arms can be tapered, which will result in a more constant input impedance of the antenna.

When the antenna arms are very short compared to one wavelength, the antenna approximates a dipole and the polarization is linear. As the frequency increases, the axial ratio decreases and the polarization becomes elliptical. For frequencies at which the arm lengths of the spiral antenna are greater than one wavelength, or slightly less, the polarization is circular and the input impedance remains almost constant as frequency varies. When this antenna is used for circular polarization, the minimum frequency is where the electrical length of the arm is one wavelength.

12.6.2 Self-Complementary Log-Periodic Slot Antenna

A self-complementary log-periodic slot antenna is shown in Figure 12.24 [13]. This toothed planar log-periodic antenna was designed to make the current flowing from the feedpoint (at the center) to the outer edges take a longer path. Regions were cut out from the edge of a bow-tie shape to make teeth, and these teeth were intended to break up the path of the current so that it has farther to travel, reducing its magnitude at the ends of the antenna and thereby reducing the end effect (see Figure 12.25 [14]). Adding appropriately spaced teeth proved to be a successful

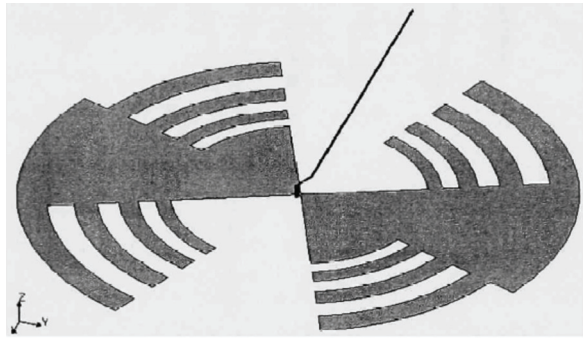


Figure 12.24 Slot log-periodic array. (Source: [13] © B. L. Nance 2004.)

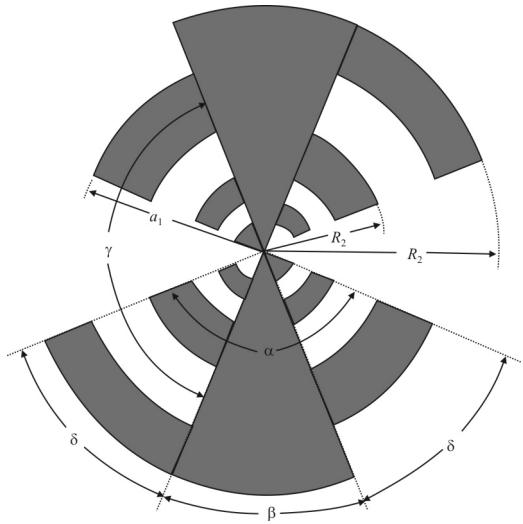


Figure 12.25 Dimensions of a self-complementary log-spiral antenna array.

method for making the antenna much more broadband, but it also rotated the angle of polarization by 90°.

If a straight line were drawn from the origin of one of these antennas, extending outward in any direction, the metal spirals that make up the antenna would intersect this line at ever-increasing increments. The spacing between one metal intersection and the next is in a constant ratio. Similarly, the spacing between a region intersected by air and the next are in another constant ratio. The constants that define these ratios are denoted by τ and σ, and they are used to define the widths of the teeth on the toothed planar log-periodic antenna.

$$\tau = \frac{R_{n+1}}{R_n} \tag{12.33}$$

and

$$\sigma = \frac{f_{n+1}}{a_n} \tag{12.34}$$

where variables a and R refer to radii defined in Figure 12.25. Each of these teeth acts as its own resonator, so that τ also defines a ratio between frequencies radiated by the teeth:

$$\tau = \frac{f_n}{f_{n+1}}, \qquad f_{n+1} > f_n \tag{12.35}$$

where f_n corresponds to the resonant frequency of the tooth bounded by R_n and a_n. The difference between $\log(f_{n+1})$ and $\log(f_n)$ defines one period of operation.

The drawback of adding these teeth, however, is that distances are now a factor in defining the antenna's structure. In other words, the geometry is no longer solely defined by angles. Because of this, it may no longer be frequency independent. This is why the toothed planar log-periodic antenna published by DuHamel and Isbell in [3] was given the added benefit of self-complementarity.

The simulated return loss (S_{11}) for this antenna is shown in Figure 12.26 [13]. This return loss is quite good, with the intended band and more all below the standard –10 dB cutoff point. The region from 4 to 6 GHz appears to be one wide resonance.

This corresponds to the innermost region of the antenna that begins just inside the smallest tooth.

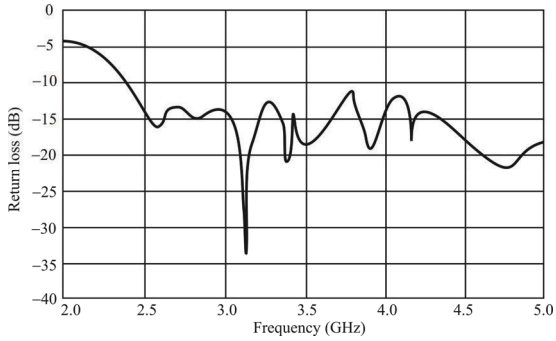


Figure 12.26 Return loss self-complementary log-periodic spiral array. $S_{11} \leq -10$ dB when VSWR ≤ 2.

Figures 12.27 and 12.28 [13] show the calculated real and imaginary components of the self-complementary antenna's impedance. As you can see, 140 Ω seems to be an appropriate median value for the microstrip feed-line impedance. Over most of the operating range we see that the antenna exhibits a capacitive characteristic.

12.7 Two-Arm Wire Conical Archimedes Spiral Antenna

A two-arm spiral conical antenna was analyzed using the method of moments by Zainud-Deen et al. [15]. The structure of this antenna is shown in Figure 12.29. It consists of two wires forming a conical spiral and is fed at the bottom as shown.

12.7.1 Operating Principles

The arm of conical spiral curve is given by

$$r = a\phi \tag{12.36}$$

where a is spiral constant and ϕ is winding angle.

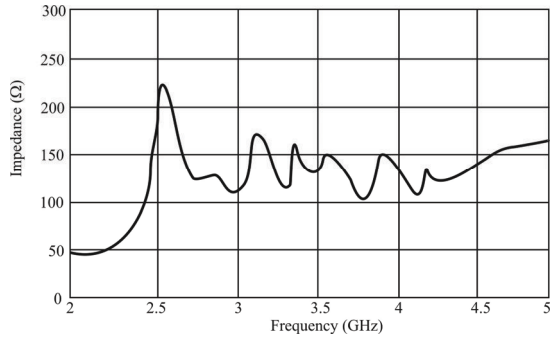


Figure 12.27 $\mathrm{Re}\{Z_L\}$ for the self-complementary log-periodic antenna in Figure 12.23.

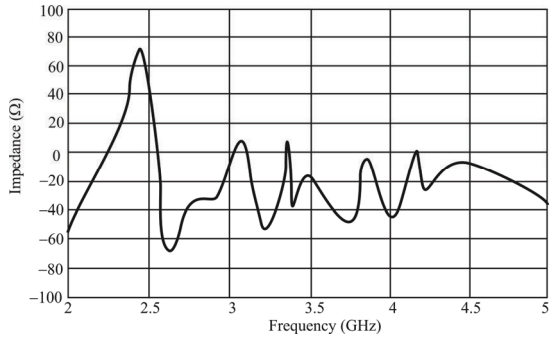


Figure 12.28 $\mathrm{Im}\{Z_L\}$ for the self-complementary log-periodic antenna in Figure 12.24.

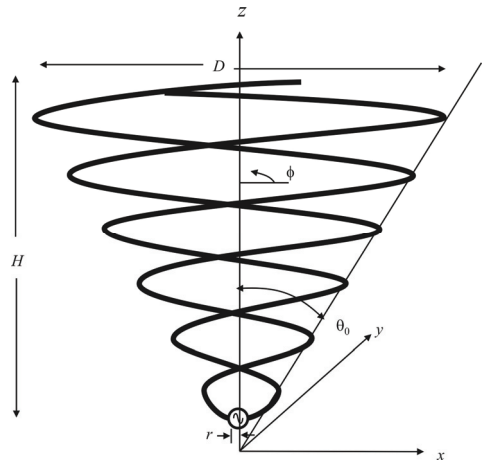


Figure 12.29 Spiral conical log-periodic antenna array.

Antenna dimensions are:

Height of the conical spiral, H= 1 λ;
Spiral constant a = 0.01λ;
Start radius (apex radius) r, = 0.02λ ;
Base diameter D = C /π.

The distance along the z-axis at any point on the conical Archimedes spiral curve given by (12.36) is defined as starting from the tip of the cone (z = 0). Thus,

$$z = \frac{a}{\tan\theta_0}\phi \tag{12.37}$$

where, 0^0**,** is the cone angle. Other important parameters shown in Figure 12.29 are the base diameter, *D,* the apex radius *r,* and the height of conical spiral *H.*

The method of moments was then used to determine the unknown current distribution along the antenna arms, as well as the input impedance, gain, and radiation patterns of the antenna.

12.7.2 Results

Figure 12.30 [15] shows the current distribution along the antenna for a 1 volt delta source. The symmetrical pattern is obtained due to the symmetrical structure of the antenna. The dimensions of the antenna in this case are:

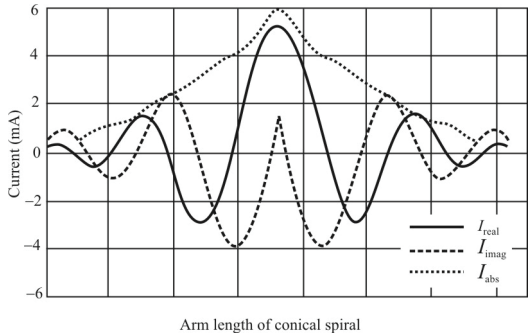


Figure 12.30 Current distribution of a conical log-periodic antenna array. $h = 1.2\lambda$.

Antenna wire radius, R,= 0.0096λ;
Antenna circumference, $C = 1.2\lambda$.

Such an antenna produces a power radiation pattern that is depicted in Figure 12.31. The input impedance of this antenna is nominally $Z_{in} = 183.7 - j54\Omega$ while the maximum power gain is 5.l dB. The axial ratio in the direction of the maximum gain is 0.73 dB. Such low value of the axial ratio indicates that the radiation produced by the conical Archimedes spiral antenna of the above geometrical parameters has very little cross-polarization. We also see that the 3 dB beamwidth at the frequency corresponding to Figure 12.31 is about $\pm 70^{o}$.

As another example, consider a conical spiral antenna with $R_o = 0.096$ cm, $C = 12$ cm, $a = 0.1$ cm, $r_S = 0.2$ cm, and $D = 3.8$ cm over a range of frequencies from 2 GHz to 5 GHz. Figure 12.32 [15] shows the variation of the input impedance of the antenna with frequency, and Figure 12.33 [15] shows the variation of VSWR (referred to a 50Ω coaxial cable) and the reflection coefficient. From these figures, we notice nearly constant values of the input impedance and the VSWR are maintained over this frequency band.

12.7.3 Summary

The MoM was used to calculate the characteristics of the conical spiral antenna. It was shown that these parameters are independent of frequency over a very wide frequency band. It turns out that the height of the antenna (H) and the spiral constant (a) are the most effective geometrical parameters in shaping the radiation patterns of the antenna [15].

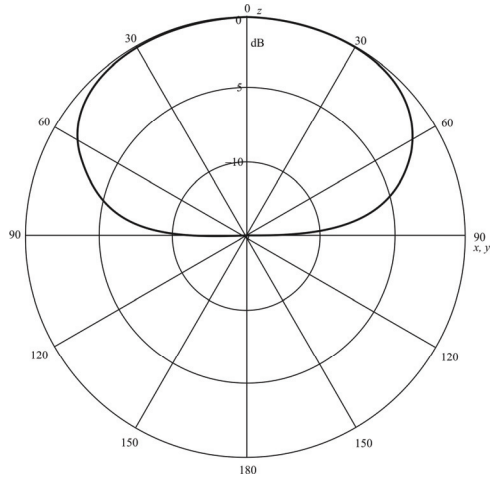


Figure 12.31 Spiral conical power pattern. Radiation pattern in the *x*-*z* and *y*-*z* planes.

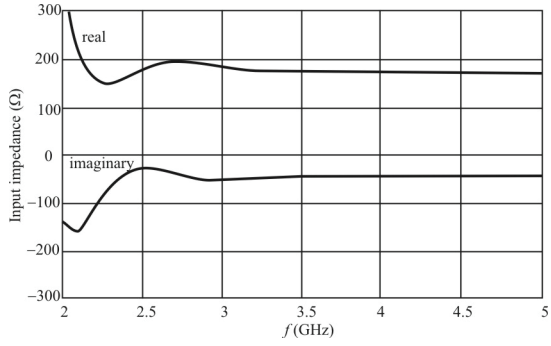


Figure 12.32 Spiral conical input impedance.

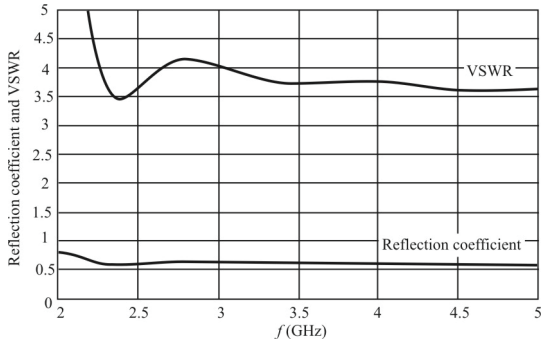


Figure 12.33 VSWR and reflection coefficient of spiral conical array frequency (referenced to 50Ω).

12.8 Concluding Remarks

Some of the properties of frequency-independent antennas were considered in this chapter. They all, one way or another, rely on the log-periodic principle where one or more of the antenna parameters varies according to the logarithm of the frequency.

The LPDA is an adaption of the Yagi-Uda antenna we discussed in Chapter 11 for wideband applications. The operating principles are identical to that of a three-element Yagi, where the active region changes depending on the operating frequency. Significant gain can be obtained from the log-periodic, although not as high as the Yagi. The principal advantage of the log-periodic is its moderate gain coupled with very wide bandwidths. The 10:1 bandwidths are not uncommon for the LPDA.

We also discussed some other types of frequency-independent antennas. While we did not cover all of the types that are available, those that were discussed are representative of the types of antennas used for ELINT, radar, and ASE purposes. These antennas generally are too large to be used below the microwave frequencies, however.

References

[1] Stutzman, W. L., *Antenna Theory*, New York: Wiley, 2005, p. 611.

[2] DuHamel R. H., and D. E. Isbell, "Broadband Logarithmically Periodic Antenna Structures," *IRE National Convention Record* 1957, pp. 119–128.

[3] Poisel, R. A., *Introduction to Communication Electronic Warfare Systems*, Norwood, MA: Artech House, 2002, p. 217.

[4] Isbell, D. E., "Log-Periodic Dipole Arrays," *IRE Transactions on Antennas and Propagation,* May 1960, pp. 260–261.
[5] Peixerio, C., "Design of Log-Periodic Dipole Antennas," *IEE Proceedings,* Vol. 135, Pt. H., No. 2, April 1988, pp. 98–102.
[6] http://www.salsburg.com/Log-Periodic.pdf.
[7] http://www.ahsystems.com/catalog/SAS-510-7.php.
[8] http://www.ewh.ieee.org/soc/es/Nov1998/13/LPARRAY/LPMA.HT.
[9] http://www.ict.csiro.au/asa/asa07/ASA07_paper_030_final.pdf.
[10] http://scholar.lib.vt.edu/theses/available/etd-01082002-073223/unrestricted/Caswell_etd_Ch2.pdf.
[11] http://www.Cobham.com.
[12] Thaysen, J., K. B. Jakobsen, and J. Appel-Hansen, "A Logarithmic Spiral Antenna for 0.4 to 3.8 GHz," *Applied Microwaves and Wireless*, 2002, pp. 32–45.
[13] http://etd.lib.ttu.edu/theses/available/etd-06272008-31295019381325/unrestricted/31295019381325.pdf.
[14] Stutzman, W. L., *Antenna Theory and Design*, New York: Wiley, 1981, p. 289.
[15] Zainud-Deen, S. H., N. F. M. Soliman, A. A. M. Shaalan, and K. F. A. Hussein, "Characteristics of Two-Arm Wire Conical Spiral Antennas," *Proceedings 21st National Radio Science Conference (NRSC2004) (NTI)*, March 16–18, 2004, pp. B20-1–B20-7.

Chapter 13

Aperture Antennas

13.1 Introduction

An aperture antenna contains an opening (the aperture) through which EM waves are transmitted or received. Examples of aperture antennas include slots, waveguides, horns, reflectors, and lenses. The analysis of aperture antennas is typically accomplished using the fields within the aperture to determine the antenna radiation patterns [1–3].

Aperture antennas are commonly used in aircraft or spacecraft applications. The aperture can be mounted flush with the surface of the vehicle, and the opening can be covered with a dielectric material called a Radome, which allows electromagnetic energy to pass through. We will discuss radomes in Chapter 22.

While not strictly limited in frequency, as will be seen, the assumptions on the dimensions of aperture antennas dictates that they are usually applied in the higher-frequency regions (UHF+). The physical dimensions are typically several wavelengths in order to make approximations that make mathematical analysis tractable.

The structure of this chapter is as follows. We first examine linear apertures with uniform, half-cosine, and general aperture distributions. We then consider planar rectangular apertures with uniform and cosine distributions. That is followed by planar circular apertures with uniform distributions. Horn antennas, which are formed from open-ended waveguides, are considered next. Lastly, we have a section on parabolic dishes.

13.2 Linear Apertures

The radiation pattern of a linear aperture, $P(\sin\alpha)$, is found from the inverse Fourier transform of the aperture distribution $E_x(x,0) = E(x,y)\big|_{y=0}$ with the frequency variable replaced with a dimensional one. Mathematically this is

$$P(S=\sin\alpha)=\frac{1}{\lambda}\int_{-\infty}^{\infty}E_x(x,0)\exp\left(j\beta xS\right)dx \tag{13.1}$$

We will calculate the radiation patterns for some standard aperture distributions to both illustrate the methodology and derive several antenna properties that are important for EW systems.

13.2.1 Uniform Distribution

A uniform distribution is constant over a finite, contiguous span and zero elsewhere. That is

$$E_x(x,0)=\begin{cases}E_0, & -a/2\le x\le x/2\\ 0, & |x|>a/2\end{cases} \tag{13.2}$$

Then the radiation pattern is

$$P(S)=\frac{1}{\lambda}\int_{-a/2}^{a/2}E_0\exp(j\beta xS)dx \tag{13.3}$$

$$=E_0\frac{a}{\lambda}\frac{\sin(\beta aS/2)}{\beta aS/2} \tag{13.4}$$

This is the familiar $\operatorname{sinc} x=\sin x/x$ pattern and is illustrated in Figure 13.l(a). The radiation pattern peak is at $S = 0$, normal to the aperture, and the main beam extends from $S=-\lambda/a$ to $S=+\lambda/a.$ The first sidelobe peak is at $S=\pm 3\lambda/2a,$ and zeros of the pattern occur at $S=\pm n\lambda/a.$ Negative values for $P(S)$ indicate a phase difference of 180° relative to the main beam.

Figure 13.1(b) shows the same pattern, normalized to P_{max}, expressed on a decibel scale, as $20\log_{10}\left[|P(S)|/|P_{max}|\right].$ The first (and largest) sidelobe level is at –13.2 dB relative to the main beam peak. The two main features of a directive radiation pattern are the beamwidth [usually indicated at the –3 dB (half-power) level] and the amplitude sidelobe levels. Here the zero-to-zero beamwidth is $2\lambda/a$ on a sin α scale, while the 3 dB beamwidth is ~ $0.9\lambda/a$.

The above case illustrates one very important antenna property, namely the inverse relationship between aperture size in wavelengths and the beamwidth—narrow beams require large apertures.

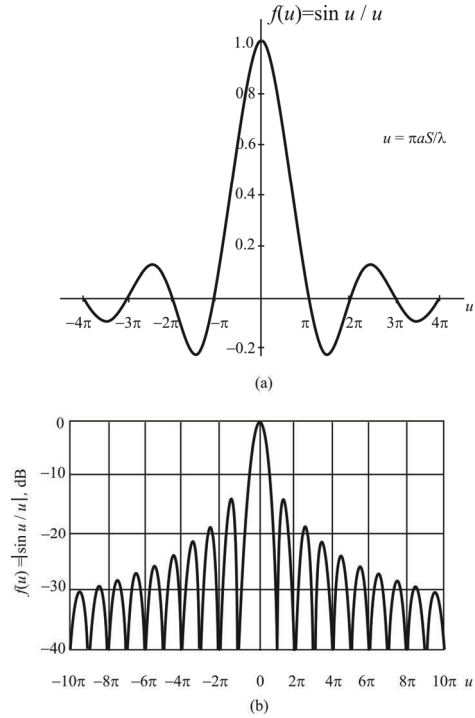


Figure 13.1(a, b) Uniform distribution radiation pattern. Note the difference in abscissa scales.

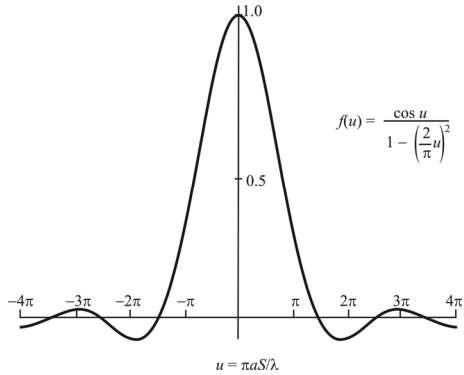


Figure 13.2 Half-cosine aperture radiation pattern.

13.2.2 Half-Cosine Distribution

The half-cosine distribution is given by:

$$E_x(x,0) = \begin{cases} E_0 \cos\left(\dfrac{\pi x}{a}\right), & -a/2 \le x \le a/2 \\ 0, & |x| > a/2 \end{cases} \tag{13.5}$$

and the radiation pattern is

$$P(S) = \frac{E_0}{\lambda} \int_{-a/2}^{a/2} \exp(j\beta xS) \frac{\exp(j\pi x/a) + \exp(-j\pi x/a)}{2} dx \tag{13.6}$$

After some elementary calculus and algebra, we get

$$P(S) = \frac{E_0 a}{\lambda} \frac{\dfrac{\pi}{2}\cos(\beta aS/2)}{\left(\dfrac{\pi}{2}\right)^2 - \left(\dfrac{\beta aS}{2}\right)^2} \tag{13.7}$$

Table 13.1 Four Common Aperture Distributions

Aperture Distribution	Normalized Radiation Pattern	First Sidelobe Level
$f(x)=\begin{cases}1, & \lvert x\rvert \le a/2\\ 0, & \text{otherwise}\end{cases}$	$\dfrac{\sin(\beta aS/2)}{\beta aS/2}$	−13 dB
$f(x)=\begin{cases}\cos(\pi x/a), & \lvert x\rvert \le a/2\\ 0, & \text{otherwise}\end{cases}$	$\dfrac{\cos(\beta aS/2)}{1-(\beta aS/\pi)^2}$	−23 dB
$f(x)=\begin{cases}\cos 2(\pi x/a), & \lvert x\rvert \le a/2\\ 0, & \text{otherwise}\end{cases}$	$\dfrac{\sin(\beta aS/2)}{\beta aS/2}\dfrac{1}{1-(\beta aS/2\pi)^2}$	−32 dB
$f(x)=\begin{cases}1-2\lvert x\rvert/a, & \lvert x\rvert \le a/2\\ 0, & \text{otherwise}\end{cases}$	$\left[\dfrac{\sin(\beta aS/4)}{\beta aS/4}\right]^2$	−26 dB

The characteristics of this radiation pattern are plotted in Figure 13.2 (linear field scale). The first zero of the pattern is not where $\beta aS/2=\pi/2,$ as both numerator and denominator become zero simultaneously at that point. The zero-to-zero beamwidth is now $3\lambda/a$ on a sin α scale, 50% larger than for a uniform aperture distribution. The 3 dB beamwidth is also correspondingly increased. The sidelobe levels have decreased significantly—the first and largest is now –23 dB.

This second case illustrates another important aperture antenna property—using a tapered aperture distribution (an "amplitude taper"), such as the cos function, tends to reduce sidelobe levels at the expense of an increase in main beamwidth. Low sidelobes are often a very desirable property in many EW applications. For example, for threat signal detection, the targets of interest are frequently considerably weaker than the possible interfering, insipid signals. If one of the more powerful signals occurs at boresight, a target signal that is more than 23 dB down at the point of the peaked sidelobe will be masked. Hence low sidelobes can be very important.

Table 13.1 lists four standard aperture distributions and the corresponding radiation patterns.

13.2.3 General Amplitude Distribution

Consider a general amplitude distribution $A(x)$. The corresponding radiation pattern is

$$P(S) = \frac{1}{\lambda} \int_{-\infty}^{\infty} A(x) \exp(j\beta x S) dx \tag{13.8}$$

Now include a linear phase variation across the aperture—that is, the aperture distribution becomes $A(x)\exp(j\psi x)$. Then the resulting radiation pattern is

$$P'(S) = \frac{1}{\lambda} \int_{-\infty}^{\infty} A(x) \exp[j(\beta S + \psi)x] dx \tag{13.9}$$

$$= P(S + \psi/\beta) \tag{13.10}$$

The new pattern has an identical shape to the original one (on a sin α scale) but is shifted by $\delta(\sin\alpha) = -\psi/\beta$. If the first pattern has its main beam peak at $S = 0$, the second pattern has its peak where $S = -\psi/\beta$. As an example, if $\psi = \pi/\lambda$, the beam maximum is changed from $\alpha = 0$ to $\alpha = -30^{\circ}$. This is an example of steering an array in a particular direction. It is accomplished by changing the phase of the radiation field.

The pattern's beamwidth is constant on a sin α scale because the pattern has the same shape on a sin α scale. This is not constant on an α scale, however. Let the beamwidth be 2δ on the sin α scale, and let the beam steering angle be α_0. Then the beamwidth on an α scale is

$$\alpha_1 - \alpha_2$$

where

$$\sin\alpha_1 = \sin\alpha_0 + \delta$$
$$\sin\alpha_2 = \sin\alpha_0 - \delta$$

Assume that δ is small (a narrow beam), and let $\alpha_1 = \alpha_0 + \varepsilon$. Then, using the approximations $\cos\varepsilon \approx 1$ and $\sin\varepsilon \approx \varepsilon$ we get

$$\sin\alpha_1 = \sin(\alpha_0 + \varepsilon) = \sin\alpha_0 \cos\varepsilon + \cos\alpha_0 \sin\varepsilon$$
$$\approx \sin\alpha_0 + \varepsilon\cos\alpha_0$$

Then $\delta \approx \varepsilon\cos\alpha_0$ and the beamwidth on an α scale becomes

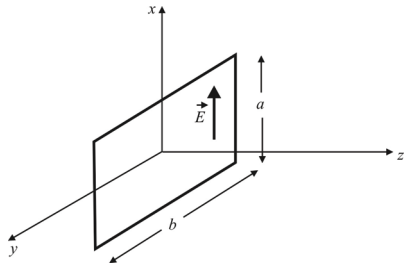


Figure 13.3 Planar aperture.

$$2\delta / \cos\alpha_0$$

The beamwidth therefore increases as the steering angle increases.

13.3 Planar Apertures

Now consider the 3-D generalizations of the formulae for two dimensions. There the fields had no *y* dependence, and an aperture of width *a* was infinitely long in the *y* direction. We now have an aperture such as shown in Figure 13.3. As in the two-dimensional case, there are two independent polarizations, and we shall give results for the E_x case ($E_y = 0$ in the aperture).

The 2-D radiation pattern for a linear aperture is given by (13.8) and the 3-D radiation pattern for a planar aperture is given by the two-dimensional extension of this:

$$P(S_1, S_2) = \frac{1}{\lambda^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} E_x(x, y, 0) \exp[j\beta(xS_1 + yS_2)] dxdy \tag{13.11}$$

where $S_1 = \sin\theta\cos\phi$, $S_2 = \sin\theta\sin\phi$ using conventional spherical polar coordinates θ and ϕ, as discussed in Chapter 2.

In the far field, the E-fields in two dimensions and three dimensions are

$$\vec{E} = \sqrt{\frac{\lambda}{r}} \exp[-j(\beta r - \pi/4)] \vec{u}_\theta P(\sin\theta) \tag{13.12}$$

$$\vec{E}(r, \theta, \phi) = \frac{\lambda}{r} \exp[-j(\beta r - \pi/2)][\cos\phi \vec{u}_\theta - \sin\phi \vec{u}_\phi] P(S_1, S_2) \tag{13.13}$$

respectively. The dependence on r in the3-D case has now changed to that of a spherical, rather than cylindrical, wave. The $\pi/2$ phase shift in (13.13) is half the π shift in (13.12) through a spherical focus. The factor $\cos\phi\vec{u}_\theta - \sin\phi\vec{u}_\phi$ defines the polarization of the wave.

We will now determine radiation patterns for some important cases.

13.3.1 Uniform Distribution over a Rectangular Aperture

Consider a uniformly illuminated rectangular aperture, as in Figure 13.3, where

$$E(x,y,0)=\begin{cases}E_0, & |x|\le a/2, |y|\le b/2\\ 0, & \text{elsewhere}\end{cases} \tag{13.14}$$

The radiation pattern is

$$\begin{aligned}P(S_1,S_2)&=\frac{E_0}{\lambda^2}\int_{-b/2}^{b/2}\int_{-a/2}^{a/2}\exp[j\beta(xS_1+yS_2)]dxdy\\&=\frac{E_0}{\lambda^2}\int_{-a/2}^{a/2}\exp(j\beta xS_1)dx\int_{-b/2}^{b/2}\exp(j\beta yS_2)dy\end{aligned} \tag{13.15}$$

In this case the simple aperture distributions, (13.15), are separable—that is, $E(x, y, 0) = E_1(x)\,E_2(y)$, which allows the x and y integrations to be separated. Then

$$P(S_1,S_2)=\frac{abE_0}{\lambda^2}\frac{\sin\left(\frac{\beta a}{2}S_1\right)}{\frac{\beta a}{2}S_1}\frac{\sin\left(\frac{\beta b}{2}S_2\right)}{\frac{\beta b}{2}S_2} \tag{13.16}$$

which is the product of two sinc functions. If $\phi = 0$, $S_1 = \sin\theta$, $S_2 = 0$, and the pattern in the y-z plane is

$$P(S_1,0)=\frac{abE_0}{\lambda^2}\frac{\sin\left(\frac{\beta a}{2}\sin\theta\right)}{\frac{\beta a}{2}\sin\theta} \tag{13.17}$$

Also if $\phi = \pi/2$, $S_1 = 0$, and $S_2 = \sin\theta$, and the pattern in the y-z plane is

$$P(0,S_2)=\frac{abE_0}{\lambda^2}\frac{\sin\left(\frac{\beta b}{2}\sin\theta\right)}{\frac{\beta b}{2}\sin\theta} \tag{13.18}$$

the only difference being the dimensions of the aperture, a and b. These patterns have the same form as a uniform aperture in two dimensions.

Note that the first zero in the x-z plane is where $\sin\theta=\pm\lambda/a,$ and the first zero in the y-z plane is where $\sin\theta=\pm\lambda/b.$ So we see that the beamwidths in the two principal planes are separately controlled by the aperture dimensions in each plane.

13.3.2 Cosine Distribution over Rectangular Aperture

Now consider a cosine distribution over a rectangular aperture that is characteristic of a horn antenna. Let

$$E_x(x,y,0)=\begin{cases}E_0\cos(\pi y/b), & |x|\le a/2,|y|\le b/2\\ 0, & \text{otherwise}\end{cases} \tag{13.19}$$

The radiation pattern is

$$P(S_1,S_2)=\frac{E_0}{\lambda^2}\int_{-b/2}^{b/2}\int_{-a/2}^{a/2}\cos\left(\frac{\pi y}{b}\right)\exp[j\beta(xS_1+yS_2)]dxdy \tag{13.20}$$

As above, this is again separable, with the two integrals similar to the 2-D examples given earlier. Thus,

$$P(S_1,S_2)=\frac{abE_0}{\lambda^2}\frac{\sin\left(\frac{\beta a}{2}S_1\right)}{\frac{\beta a}{2}S_1}\frac{\frac{\pi}{2}\cos\left(\frac{\beta b}{2}S_2\right)}{\frac{\pi^2}{4}-\frac{\beta^2b^2S_2^2}{4}} \tag{13.21}$$

In the x-z plane (S_2 = 0). The first zero of the pattern is where $\sin\theta=\lambda/a,$ whereas in the y-z plane (S_1 = 0), the first zero is where $\sin\theta=3\lambda/2b.$ In addition, the sidelobes in the y-z plane are lower than in the x-z plane.

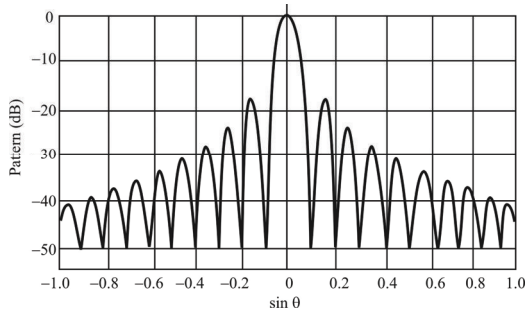


Figure 13.4 Radiation pattern for a circular aperture with uniform distribution.

13.3.3 Uniform Distribution with a Circular Aperture

Now consider a uniform illumination over a circular aperture with diameter D. The radiation pattern in any plane is a Bessel function, shown in Figure 13.4. The first sidelobe level is –13.6 dB, and the 3 dB beamwidth is $1.02\lambda/D$. The (normalized) pattern is

$$\frac{2J_1(\beta a \sin\theta)}{\beta a \sin\theta} \tag{13.22}$$

(where $a = D/2$). If an amplitude taper is superimposed on this field distribution, the beamwidth increases and the sidelobes are (generally) reduced, just as for a linear aperture.

13.3.4 Gain

For an antenna with gain, the power P incident on the effective area of the antenna, A_{eff} at a distance r in the direction of most intense radiation is based on (3.77) and is given by

$$P_R = \frac{GP_T A_{\text{eff}}}{4\pi r^2} \tag{13.23}$$

where P_T is the transmitter power, and G is the gain, the maximum value of $G(\theta, \phi)$. Directivity, D, is defined by (3.5) as the ratio of the maximum radiation intensity to the average radiation intensity (averaged over all angles). For an antenna that is 100% efficient and has no losses, directivity and gain are the same.

As discussed in Chapter 3, for an antenna with losses, G will be less than D by a factor corresponding to the efficiency (3.54). Now (13.23) can be rearranged as

$$G = \frac{4\pi r^2 (P / A_{\text{eff}})}{P_{\text{T}}} \tag{13.24}$$

13.3.4.1 Uniform Distribution over a Rectangular Aperture

Consider the uniformly illuminated rectangular aperture defined by (13.14). Its radiation pattern $P(u_1, u_2)$, is given by (13.15). The far field, $E(r, \theta, \phi)$, is found from (13.13). The maximum value of E for a given r is found where $S_1 = S_2 = 0$; this value is

$$|E| = \frac{abE_0}{r\lambda} \tag{13.25}$$

From (2.62) and (13.25) the power flux density

$$P_{\text{den}} = P / A = \frac{1}{2}\frac{|E|^2}{\eta_0} = \frac{1}{2\eta_0}\left(\frac{abE_0}{r\lambda}\right)^2 \tag{13.26}$$

The power transmitted, P_{T}, is found by assuming that $E/H = \eta_0$ in the aperture—a fair assumption if $a, b >> \lambda$. Then

$$P_{\text{T}} = \frac{1}{2}\frac{E_0^2}{\eta_0} ab \tag{13.27}$$

Substituting (13.26) and (13.28) into (13.24), the gain of a uniformly illuminated rectangular aperture is

$$G = \frac{4\pi A}{\lambda^2} \tag{13.28}$$

where A is the area of the aperture. This result is in fact true for any shape of aperture, provided the illumination is uniform. This is the same result achieved with (3.83) for any lossless antenna when $A_{\text{eff}} = A$.

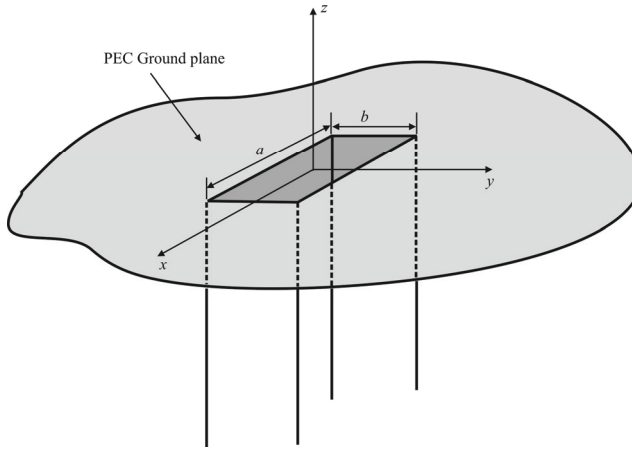


Figure 13.5 Open-ended waveguide.

13.4 Open-Ended Rectangular Waveguide

An open-ended waveguide forms an aperture antenna at the opening. Consider an open-ended rectangular waveguide that connects to a conducting ground plane, which covers the *x*-*y* plane as illustrated in Figure 13.5. Assuming that the waveguide carries only the dominant TE_{10} mode, the field distributions in the aperture of the waveguide are

$$\vec{E} = E_0 \cos\left(\frac{\pi x}{a}\right)\vec{u}_y$$

$$\vec{H} = -\frac{\gamma_{10}}{j\omega\mu} E_0 \cos\left(\frac{\pi x}{a}\right)\vec{u}_x \qquad -a/2 \le x \le a/2,\ -b/2 \le y \le b/2 \qquad (13.29)$$

where

$$\gamma_{10} = j\beta\sqrt{1-\left(\frac{f_{c_{10}}}{f}\right)^2}$$

$$f_{c_{10}} = \frac{1}{2a\sqrt{\mu\epsilon}}$$

The resulting radiated far fields are

$$E_r = H_r = 0 \tag{13.30}$$

$$E_\theta = -j\frac{ab\beta E_0}{4r}e^{-j\beta r}\sin\phi\frac{\cos\left(\frac{\beta a}{2}\sin\theta\cos\phi\right)}{\left(\frac{\beta a}{2}\sin\theta\cos\phi\right)^2-\left(\frac{\pi}{2}\right)^2}\frac{\sin\left(\frac{\beta b}{2}\sin\theta\sin\phi\right)}{\frac{\beta b}{2}\sin\theta\sin\phi} \tag{13.31}$$

$$E_\phi = -j\frac{ab\beta E_0}{4r}e^{-j\beta r}\cos\theta\cos\phi\frac{\cos\left(\frac{\beta a}{2}\sin\theta\cos\phi\right)}{\left(\frac{\beta a}{2}\sin\theta\cos\phi\right)^2-\left(\frac{\pi}{2}\right)^2}\frac{\sin\left(\frac{\beta b}{2}\sin\theta\sin\phi\right)}{\frac{\beta b}{2}\sin\theta\sin\phi} \tag{13.32}$$

$$H_\theta = -\frac{E_\phi}{\eta_0} \tag{13.33}$$

$$H_\phi = \frac{E_\theta}{\eta_0} \tag{13.34}$$

The fields in the E-plane ($\phi = 90^o$) and H-plane ($\phi = 0^o$) reduce to

$$E_\theta = -j\frac{aE_0}{\pi^2 r}e^{-j\beta r}\frac{\sin\left(\frac{\beta b}{2}\sin\theta\right)}{\sin\theta} \qquad \phi = 90^o \tag{13.35}$$

$$H_\phi = -j\frac{abE_0}{r}e^{-j\beta r}\cos\theta\frac{\cos\left(\frac{\beta a}{2}\sin\theta\right)}{(\beta a\sin\theta)^2-\pi^2} \qquad \phi = 0^o \tag{13.36}$$

These two fields are illustrated in Figures 13.6 and 13.7 for two sets of values for (a, b). We can see that with a wider aperture the beamwidth shrinks; for the set of values we used, the half-power bandwidth collapsed from 24^o to 8^o. As mentioned, however, there is a point of diminishing returns after which further beamwidth reduction becomes impractical.

13.5 Horn Antenna

The horn antenna is one type of aperture antenna. It is used in the transmission and reception of microwave signals, and the antenna is normally used in conjunction

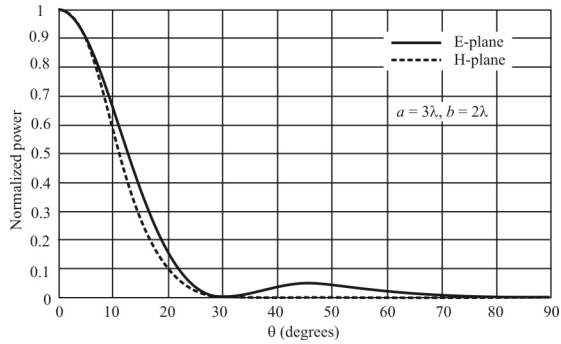


Figure 13.6 Open-ended waveguide radiated power when $a = 3\lambda$ and $b = 2\lambda$.

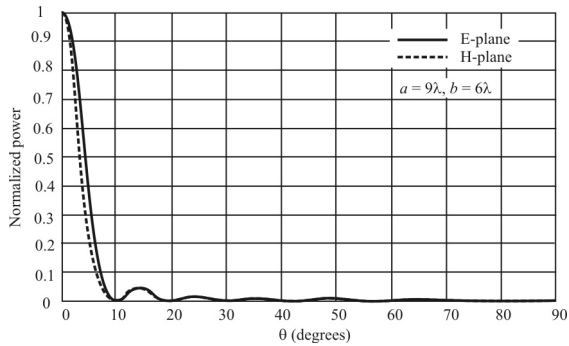


Figure 13.7 Open-ended waveguide radiated power when $a = 9\lambda$ and $b = 6\lambda$.

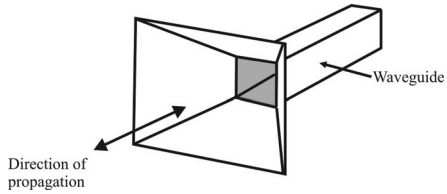


Figure 13.8 Typical horn antenna.

with rigid waveguide feeds. The horn antenna gains its name from its appearance (see Figure 13.8). The waveguide can be considered to open out or to be flared, launching the signal towards the receiving antenna.

Horn antennas are often used as feeds for parabolic "dish" antennas, as well as being used as antennas in their own right. When used as part of a parabolic reflector, the horn is orientated towards the reflector surface and is able to give a reasonably even illumination of the surface while minimizing the signal leakage that does not hit the reflector. In this way it is able to maximize the efficiency of the overall antenna. The use of the horn antenna also minimizes the spurious responses of the parabolic reflector antenna to signals that are in the sidelobes.

The horn antenna may be considered as an RF transformer or impedance match between the waveguide feeder and free space, which has an impedance of $\eta_0 = 120\pi \approx 377\Omega$. As a matching section, the horn antenna reduces reflections and leads to a lower VSWR. It can be viewed as part of the waveguide that is flared out. By having a flared end to the waveguide the horn antenna is formed and this enables the impedance to be matched. Although the waveguide will radiate without a horn antenna, the flaring provides a far more efficient match. In addition to the improved match provided by the horn antenna, it also helps suppress signals traveling via unwanted modes in the waveguide from being radiated (see Chapter 4).

The main advantage of the horn antenna, however, is that it provides a significant amount of directivity and therefore gain. To achieve the maximum gain for a given aperture size, the taper should be long so that the phase of the wavefront is nearly constant across the aperture. However, there comes a point where to provide even small increases in gain, the increase in length becomes too large to be practical. Thus, gain levels are a balance between aperture size and length. However, gain for a horn antenna may be 20 dB or more.

13.5.1 Horn Antenna Types

There are two basic types of horn antenna: pyramid and conical. The pyramid ones, one of which is illustrated in Figure 13.8, as the name suggests are

rectangular, whereas the conical ones are usually circular. The conical horn provides a pattern that is nearly symmetrical, with the E- and H-plane beamwidths being nearly the same. Additionally, it is possible to control the sidelobes better with a conical or corrugated horn antenna.

Pyramidal horn antennas can be further categorized into three types: (a) the E-plane sectoral horn, flared in the direction of the E-plane only (see Figure 13.9), (b) the H-plane sectoral horn, flared in the direction of the H-plane only (see Figure 13.10), and (c) the pyramidal horn antenna, flared in both the E-plane and H-plane (Figure 13.8). The flare of the horns is usually, but not always, linear. Some horn antennas are formed by other flare types such as an exponential flare.

The horn antenna is mounted on a waveguide that is almost always excited in single-mode operation. That is, the waveguide is operated at a frequency above the cutoff frequency of the TE_{10} mode but below the cutoff frequency of the next higher mode (see Chapter 4).

13.5.2 Horn Antenna Gain

The gain of a horn antenna is determined in this section. The maximum value of $|E|$ is found from (13.21) and (13.13):

$$|E| = \frac{abE_0}{r\lambda}\frac{2}{\pi} \tag{13.37}$$

The power flux density is now

$$P/A = \frac{1}{2\eta_0}\frac{abE_0}{r\lambda}\frac{4}{\pi^2} \tag{13.38}$$

The power transmitted, again using the assumption $E/H = \eta_0$, is

$$\begin{aligned} P_T &= \frac{1}{2}\frac{E_0^2}{\eta_0}a\int_{-b/2}^{b/2}\cos^2(\pi y/b)dy \\ &= \frac{1}{2}\frac{E_0^2}{\eta_0}\frac{ab}{2} \end{aligned} \tag{13.39}$$

The gain now becomes

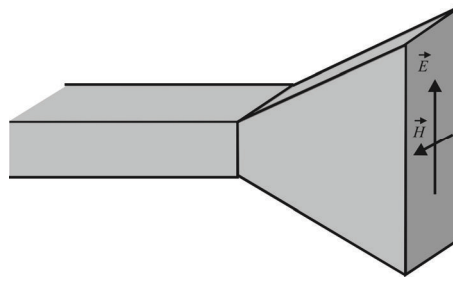


Figure 13.9 E-plane sectoral horn antenna.

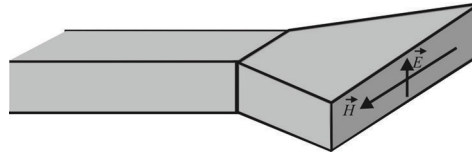


Figure 13.10 H-plane sectoral horn antenna.

$$G = \frac{4\pi ab}{\lambda^2}\frac{8}{\pi^2} \tag{13.40}$$

Compared with the uniformly illuminated aperture case, the maximum power flux density is multiplied by $4/\pi^2$; however the power transmitted is multiplied by 1/2 so the net gain is $8/\pi^2$ times that for the uniform aperture. The effective area is therefore ~80% of the geometrical area for a horn antenna. The *aperture efficiency* of aperture type antennas is defined as the ratio of the effective area to the geometrical area, and for the horn antenna considered here is about 80%.

13.5.3 Horn Antenna Radiation Fields

The field expressions for the horn antennas are similar to the fields of a TE_{10} mode for a rectangular waveguide with the aperture dimensions of *a* and *b*. The difference is in the complex exponential term, parabolic phase error, δ due to the plane of the aperture not corresponding to cylindrical wavefront. For the E plane sectoral horn the fields are given by

$$E_y(x,y) \approx E_1 \cos\left(\frac{\pi}{a}x\right) e^{-j\frac{\beta y^2}{2\rho_1}} \tag{13.41}$$

$$H_x(x,y) \approx \frac{E_1}{\eta}\cos\left(\frac{\pi}{a}x\right) e^{-j\frac{\beta y^2}{2\rho_1}} \tag{13.42}$$

$$H_z(x,y) \approx jE_1\left(\frac{\pi}{\beta a\eta}\right)\sin\left(\frac{\pi}{a}x\right) e^{-j\frac{\beta y^2}{2\rho_1}} \tag{13.43}$$

The E- and H-radiation patterns of an E-plane sectoral horn antenna is illustrated in Figure 13.11. The E- and H-radiation patterns of an H plane sectoral horn antenna is illustrated in Figure 13.12.

The 3-D radiation pattern for the typical horn antenna shown in Figure 13.13 is shown in Figure 13.14 [4]. This pattern is for radiation at 3 GHz. Notice the rounded main beam shape. As mentioned, the cosine function is a good approximation for this shape over a considerable portion of the main beam.

13.5.4 Horn Antenna Feed Techniques

There are several ways to couple the energy into a horn antenna, or more accurately, into the waveguide feeding the antenna. Two of the more common are the coupling stub and coupling loop illustrated in Figure 4.10.

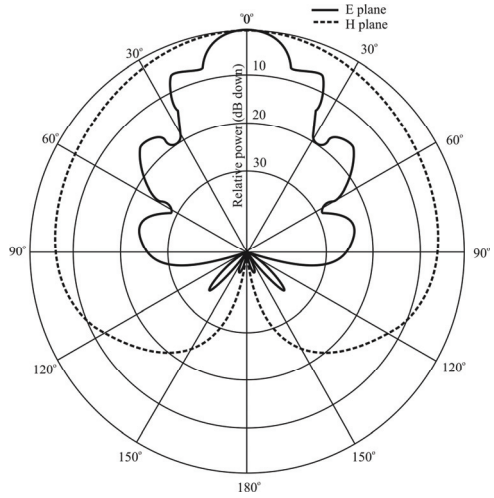


Figure 13.11 E-plane sectoral horn antenna radiation patterns.

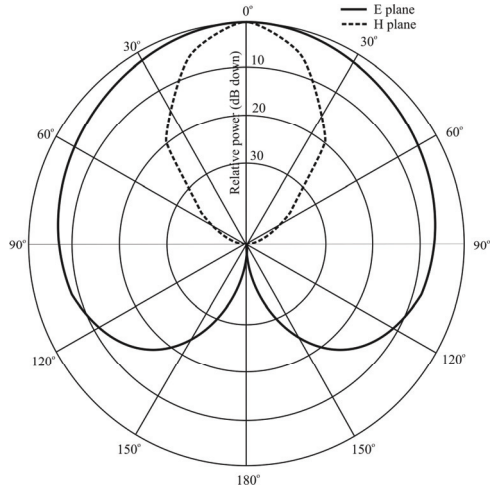


Figure 13.12 H-plane sectoral horn antenna radiation patterns.

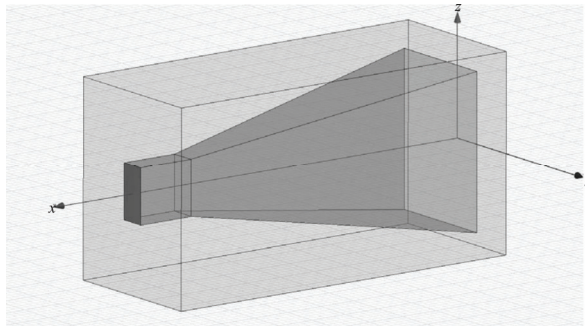


Figure 13.13 Typical E-plane horn antenna. The front is in the –*x*-direction. (Source: [4].)

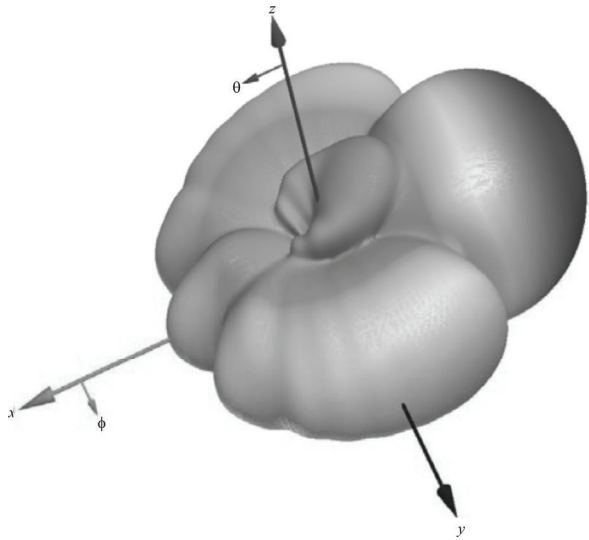


Figure 13.14 Horn pattern for the antenna in Figure 13.11 at 3 GHz. Forward radiation is in the –*x*-direction. (Source: [4].)

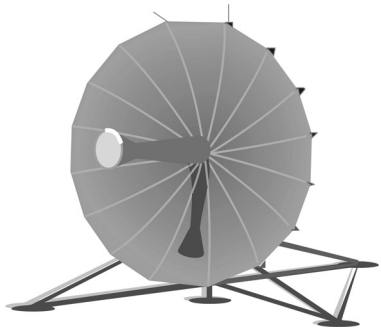


Figure 13.15 Parabolic dish.

13.6 Parabolic Dish Antenna

The parabolic dish can be considered a form of aperture antenna for the higher-frequency ranges, an example of which is shown in Figure 13.15 [5, 6]. The dish antenna finds uses in many radio applications at frequencies usually above about 1 GHz where very high levels of RF antenna gain are required along with narrow beamwidths. The relatively short wavelength of EM energy at these frequencies allows reasonably sized reflectors to exhibit the very desirable highly directional response for both receiving and transmitting. These parabolic reflectors or dish antennas are used for satellite as well as microwave links, employed as radio relays, and mobile phone antennas (infrastructure, not between the mobile handsets and base stations). In all these applications very high levels of gain are required to receive the incoming signals that are often at a very low level. When transmitting, this type of RF antenna is able to concentrate the available radiated power into a narrow beamwidth, ensuring most of the available power is radiated in the desired direction.

Parabolic dish antennas are also used in EW systems as the frequency range dictates. The above mentioned communication uses for these antennas are often times EW targets as well, and they generally are fixed in place. As such, antennas used for EW can be fixed as well.

A typical parabolic antenna consists of a parabolic reflector illuminated by a small feed antenna. The reflector is a metallic surface formed into a paraboloid of revolution and (usually) truncated in a circular rim that forms the diameter of the antenna. The mathematical definition of a parabola is the locus of points that are equidistant from both the focus and a straight line, both lying in a plane. The focal

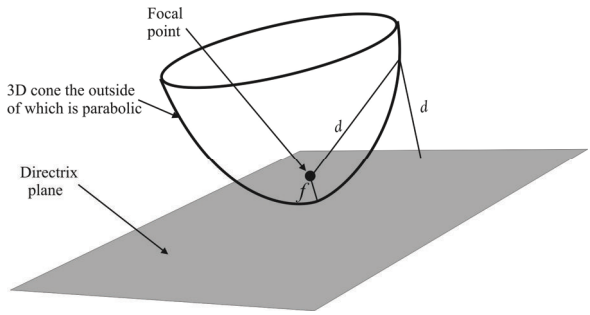


Figure 13.16 3-D parabola.

point has the unique property that, in optics, if the reflective surface reflects light, then the light beams reflected off the surface with a light at the focus are parallel and aligned with the axis of revolution (see Figure 13.16). The same principle applies to such antennas with radio waves.

13.6.1 Parabolic Reflector Antenna Principles of Operation

As mentioned, the parabolic reflector antenna consists of a radiating system that is used to illuminate a reflector that is formed in the shape of a paraboloid. This shape enables a very accurate beam to be obtained. Thus, the feed system forms the actual radiating section of the antenna (for transmitting) and the receiving section of the antenna (for receiving), and the reflecting parabolic surface is purely passive.

There are a number of parameters and terms that are of importance when analyzing parabolic reflector antenna systems:

- ***Focus:*** The focus or focal point of the parabolic reflector is the point at which the incoming signals are concentrated. When radiating, the feed is located at this point and the signals will be reflected by the reflecting surface and travel away in a parallel beam.
- ***Vertex:*** This is the innermost point at the center of the reflecting surface.
- ***Focal length:*** The focal length of a parabolic antenna is the distance from its focus to its vertex.
- ***Aperture:*** The aperture of a parabolic reflector is the size of the surface of the reflector when projected onto a 2-D plane. For a circular reflector, this is quantified by its diameter.

13.6.1.1 Feed Systems

A parabolic antenna is designed around its feed system. The design of the feed system is central to the design of the overall parabolic reflector antenna system. As mentioned, the actual "antenna" in a parabolic antenna, that is, the device that interfaces the transmission line or waveguide containing the RF energy to free space, is the feed element. The reflector surface is strictly passive. In its simplest form, this feed element is at the center of the reflector at the focal point of that dish. The radiation from the feed element induces a current flow in the conductive reflector surface which, in turn, reradiates in the desired direction, perpendicular to the directrix plane of the paraboloid (see Figure 13.16).

13.6.1.2 Parabolic Antenna Focal Length

The focal point is the point at which all reflected waves are concentrated. To ensure that the antenna operates correctly, it is necessary to place the radiating element at the focal point. The focal length f is calculated with

$$f = \frac{D^2}{16d} \tag{13.44}$$

where:

D is the diameter of the reflecting surface in the same units as the wavelength.
d is the depth of the reflector.

The feed element can be any one of a multitude of antenna types. Whichever type is used, it must exhibit a directivity that efficiently illuminates the reflector and must have the correct polarization for the application—the polarization of the feed determines the polarization of the entire antenna system. The simplest feed is a half-wave dipole, which is commonly used at lower frequencies, sometimes in conjunction with a closely coupled parasitic reflector placed behind it. At higher frequencies a horn becomes more feasible and efficient. To adapt the horn to a coaxial antenna cable, a length of waveguide is used to effect the transition.

13.6.1.3 Feed Types

There are three basic forms of feed system that can be used for a parabolic reflector antenna (others are possible as well):

- Focal point feed system;

- Offset feed system;
- Cassegrainian reflector system.

The radiation from the feed element induces a current flow in the conductive reflector surface, which, in turn, reradiates in the desired direction, perpendicular to the directrix plane of the paraboloid.

Focal Feed System

In the focal feed system, the feed is placed at the focal point of the parabolic reflecting surface. The energy from the radiating element is arranged so that it illuminates the entire reflecting surface. Once the energy is reflected it leaves the antenna system in a narrow beam [see Figure 13.17(a)]. As a result considerable levels of gain can be achieved.

Offset Feed system

In the offset feed, the feed element is still located at the focal point, which because of the angles utilized, is usually located below the reflector so that the feed element and support do not interfere with the main beam. This also allows for easier maintenance of the feed, but is usually only found in smaller antennas (such as VSAT systems). This feed system is shown in Figure 13.17(b).

Cassegranian Feed System

The Cassegranian feed system, although requiring a second reflecting surface, has the advantage that the overall length of the dish antenna between the two reflectors is shorter than the length between the radiating element and the parabolic reflector. The Gregorian and Cassegranian types, utilizing a secondary reflector, allows for better control over the collimated beam as well as allowing the antenna feed system to be more compact. The feed element is usually located in a feed horn, which protrudes out from the main reflector. This arrangement is used when the feed element is bulky or heavy such as when it contains a preamplifier or even the actual receiver or transmitter. A Gregorian antenna can be identified by the fact that it uses a concave subreflector, while a Cassegranian antenna uses a convex subreflector as seen in Figure 13.17(c).

13.6.2 Parabolic Dish Antenna Gain

The gain of the parabolic antennas is of paramount importance. Parabolic antennas or parabolic reflector antennas are often used solely for their gain and directivity

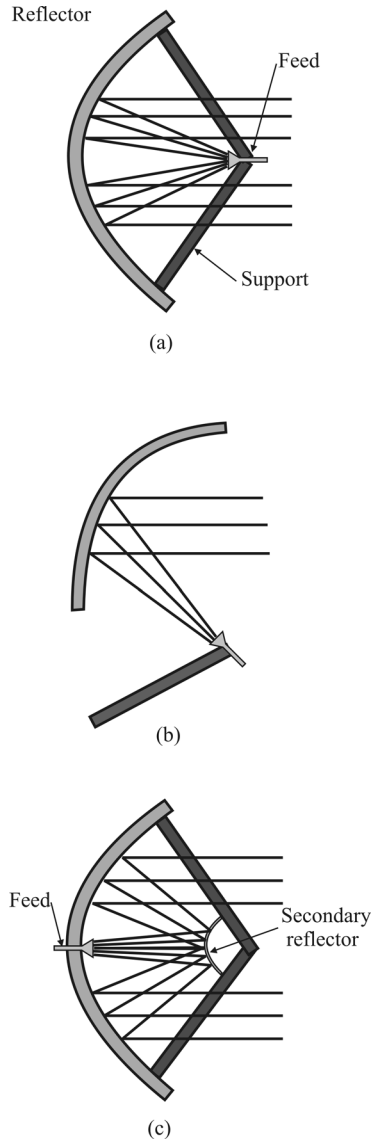


Figure 13.17(a–c) Types of dishes: (a) parabolic, (b) offset, and (c) Cassegranian (as shown with the convex secondary reflector, Gregorian when a concave secondary reflector is used).

[7, 8]. Considering the parabolic antenna as a circular aperture gives the following approximation for the maximum gain:

$$G \approx \eta_{ap} \left(\frac{\pi D}{\lambda} \right)^2 \tag{13.45}$$

where

G is the gain over an isotropic source, linear, not decibels.
η_{ap} is the aperture efficiency factor, a simple way to incorporate the differences between the geometric aperture and the effective aperture, and is generally around 35% to 55%, that is, 0.35 to 0.55 for dish reflector antennas.
D is the diameter of the parabolic reflector, consistent units.
λ is the wavelength of the signal, consistent units.

From this we can see that very large gains can be achieved if sufficiently large reflectors are used (large D). However, when the antenna has a very large gain, the beamwidth is also very small and the antenna requires very careful control over its pointing angle.

There are a number of factors that affect the parabolic antenna gain. These factors include the following:

- Diameter for the parabolic reflector reflecting surface [D in (13.45)];
- Surface accuracy;.
- Quality of illumination of the reflecting surface;
- Frequency or wavelength of the signal being received or transmitted.

Note that most of the gain in parabolic reflector antennas comes from the effects of the concentrating effects of the parabolic surface, not in the feed, at least at the lower frequencies. At higher frequencies, the feed horn can provide more gain than the low-frequency dipoles.

13.6.2.1 Optimizing Parabolic Antenna Gain

To provide the optimum illumination of the reflecting surface, the level of illumination should be greater in the center than at the sides. It can be shown that the optimum situation occurs when the center is around 10 to 11 dB greater than the illumination at the edge. Higher levels of center illumination/lower levels of edge illumination result in lower sidelobe levels.

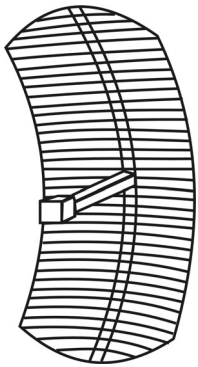


Figure 13.18 Mesh reflector.

The reflecting surface is a major component of the antenna system; however, it is not all that critical—often a wire mesh may be used. Provided that the pitch of the mesh is small compared to a wavelength ($<\lambda/10$) it will be seen as a continuous surface by the radio signals. If a mesh is used then the wind resistance and weight will both be reduced, and these provide significant practical mounting advantages. Such a mesh reflector antenna is shown in Figure 13.18.

13.6.3 Structure

The reflector dish can be solid, mesh, or wire in construction and it can be either fully circular or somewhat rectangular depending on the radiation pattern of the feeding element, which is the controlling element. The performance characteristics of solid antennas are more ideal but are troublesome because of weight and high wind load. Mesh and wire types weigh less, are easier to construct, and have nearly ideal characteristics if the holes or gaps are kept smaller than $\lambda/10$, as mentioned.

In the wire-mesh parabolic antenna illustrated in Figure 13.18, the antenna is oriented to provide horizontal polarization—the reflector wires and the feed element are both horizontal. This antenna has a greater extent in the vertical plane and therefore a narrower beam in that plane. The feed element would have a wider beam in the vertical direction than the horizontal and thereby match the reflector by illuminating it fully.

For most low-rate digital communication systems a small reflector combined with a focal point feed are used, providing the simplest and most economical form of construction. These antennas may not always look exactly like the traditional full dish antenna. For mechanical and production reasons the feed is often offset

from the center and a portion of the paraboloid used, again offset from the center. This provides mechanical advantage. Nevertheless, the principles are exactly the same.

13.6.3.1 Shield

The basic microwave antenna consists of an open dish and a feed system. Usually lacking a radome, standard performance antennas are economical solutions for specific applications. The main drawback of these antennas is the diffraction of microwave energy around the rim of the dish, which results in significant sidelobes at azimuth angles of $\pm 100^{o}$, which can interfere with adjacent point-to-point links.

To block these sideways rim reflections, a shield is placed around the circumference of the antenna, to which a planar radome is attached as seen in Figure 13.19 [9]. These "high-performance" microwave antennas may be further enhanced by the application of RF energy absorbing foam to the inside of the shield, resulting in "ultra-high-performance" antennas. The foam absorbs spurious reflections within the antenna and improves performance through limiting the sidelobes. A shield is shown in Figure 13.19.

Radiation pattern envelopes for standard, high, and ultra-high-performance antennas are compared in Figure 13.20 [9]. The improvement in sidelobe reflection control of the ultra-high-performance antenna over both other antennas is evident. Interestingly though, the high-performance antenna exhibits poorer performance than the standard performance antenna between 20^{o} and 60^{o}—the result of additional reflections off the shield. It nevertheless proves significantly lower back lobes.

13.6.3.2 Mounting

The mounting of reflecting antennas must also be considered for installations seeking to minimize interference. Typical multi-antenna tower installations, exemplified in Figure 13.19, exhibit interference, and even the mounting structure may directly impact the performance of the microwave link. This is particularly true when antennas are mounted on the face of buildings and metallic towers such as the one in Figure 13.19. Figure 13.21 [9] shows significant reflections that can arise when an antenna is mounted too close to a solid structure to the side. Such structure-generated reflections are likely sources of interference for other antennas on the same tower.

The performance of a microwave network can deteriorate due to the impacts of the environment. While a radome might protect the sensitive feed system and reflecting surface from the elements, only a stable tower and mount can protect the

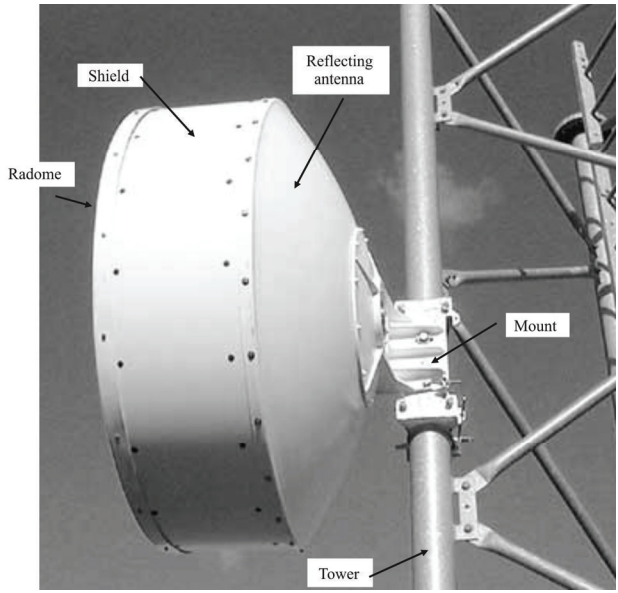


Figure 13.19 Mounted dish.

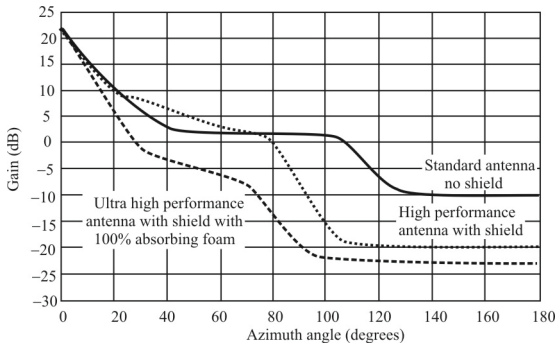


Figure 13.20 Shield performance. Performance levels of standard, high-performance, and ultra-high performance antennas.

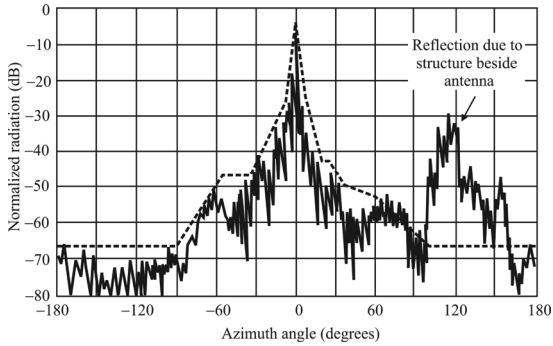


Figure 13.21 Significant reflections may arise when an antenna is mounted too close to a solid structure.

dish from the deleterious effects of wind. Because the beams are so narrow, mechanical stability of an installed antenna is critical to maintain a point-to-point link. In addition, mechanical stability is required to minimize the effects of interference for other links.

While the above comments are written for using the reflecting antenna in a communication link, all of the considerations apply to using these antennas for EW purposes—both EA and ES. The feed techniques and the nature of the reflecting surface have the same limitations.

13.6.4 Antenna Power Gain Patterns

In this section we present some gain patterns for parabolic dish reflector antennas. The model used in the NEC antenna analysis software for parabolic antennas is illustrated in Figure 13.22. For this antenna, $D = 200$ cm.

Figure 13.23 illustrates the 2-D gain pattern for the vertical plane (the horizontal plane pattern is exactly the same since these antennas are symmetric) when $D/\lambda = 0.66$ ($f = 1$ GHz). The 3 dB beamwidth is approximately 80^o while the front-to-back ratio is approximately 17 dB. The maximum gain is about 7 dB at this frequency. The back and sidelobe structure for this antenna is certainly more complicated than that shown in Figure 13.21. The modeling resolution is not fine enough to pull this structure out, however.

The 3-D structure is illustrated in Figure 13.24. The symmetric characteristic is evident in this figure.

When the D/λ ratio is increased to $D/\lambda = 1.67$ by increasing the frequency to 5 GHz, the 2-D effects illustrated in Figure 13.25 ensue. In this case the 3 dB beamwidth is approximately 20^o with a maximum gain of about 17 dB. The front

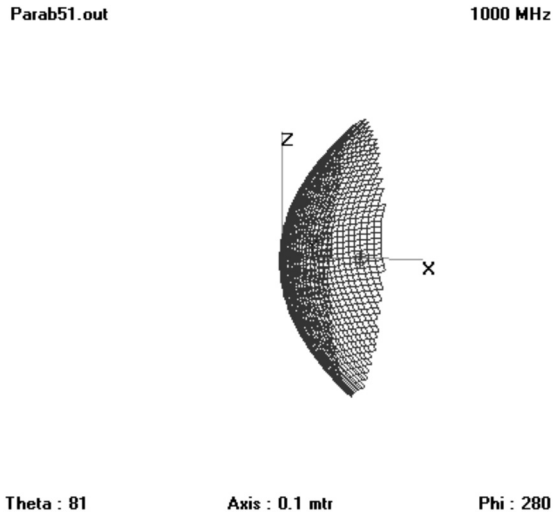


Figure 13.22 4NEC2 model of parabolic dish. f = 1 GHz.

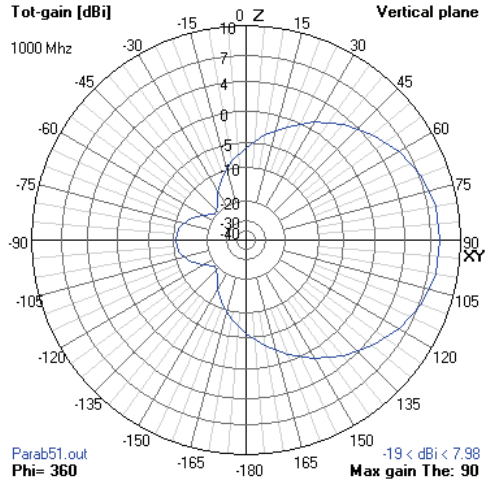


Figure 13.23 2-D vertical plane pattern for the antenna shown in Figure 13.22. f = 1 GHz, R = 100 cm. Horizontal plane pattern is identical due to symmetry.

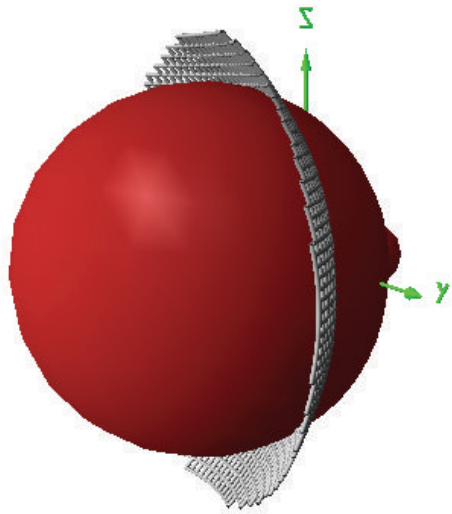


Figure 13.24 3-D radiation pattern for parabolic dish. Parabola51 p33 3D. f = 1 GHz, R = 100 cm.

to back ratio is approximately 37 dB. In addition, some of the rich back-lobe characteristics begin to be identified in Figure 13.21.

The 3-D structure for this antenna is shown in Figure 13.26. The focusing effects of the antenna at higher frequencies are evident when comparing Figure 13.24 with Figure 13.26.

13.7 Concluding Remarks

We presented some fundamental information on aperture antennas in this chapter. It certainly is not a complete expose on this interesting and diverse topic, but hopefully the reader gets a glimpse of some of the primary attributes of aperture antennas.

We also present some information on the open-ended waveguide, where the open end forms an aperture antenna, as well as the horn antenna, one of the most popular aperture antennas. The horn antenna provided a better interface to the propagating medium from the waveguide than just using the open end of the waveguide. In that capacity, the horn is an impedance transformation device.

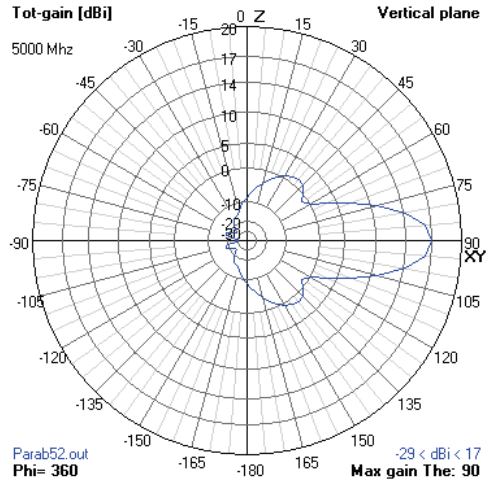


Figure 13.25 2-D vertical pattern for parabolic dish at f = 5 GHz, R = 100 cm.

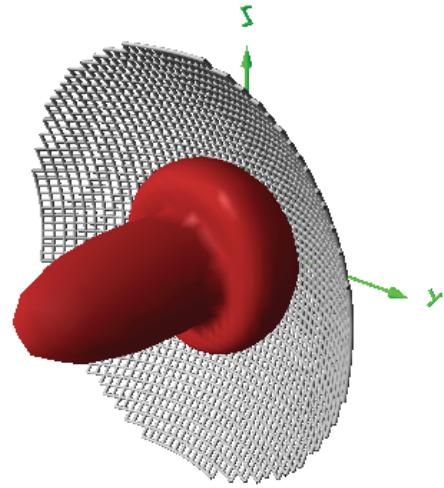


Figure 13.26 3-D radiation pattern for parabolic dish when f = 5 GHz, R = 100 cm.

Some of the more salient characteristics of the parabolic reflector dish antenna were presented. This antenna is used in the microwave frequency band and higher because of its high gain and narrow beam. Like the HF LPDA, as an EW antenna, it is principally useful in installations where the target azimuth is fixed and well known.

References

[1] Johnson, D. H., and D. E. Dudgeon, *Array Signal Processing*, Upper Saddle River, NJ: Prentice Hall, 1993, Ch. 3.
[2] Stutzman, W. L., and G. A. Thiele, *Antenna Theory and Design*, New York: Wiley, 1981, Ch. 13.
[3] Jordan, E. C., and K. G. Balman, *Electromagnetic Waves and Radiating Systems,* Upper Saddle River, NJ: Prentice-Hall, 1968, Ch. 13.
[4] http://www.arl.army.mil/arlreports/2009/ARL-TN-0371.pdf.
[5] http://www.analyzemath.com/parabola/parabola_work.html.
[6] http://www.radio-electronics.com/info/antennas/parabolic/parabolic_reflector.php.
[7] http://www.radartutorial.eu/06.antennas/an13.en.html.
[8] http://en.wikipedia.org/wiki/Parabolic_antenna.
[9] http://www.mwrf.com/Articles/Print.cfm?ArticleID=8086.

Chapter 14

Electrically Small EW Antennas

14.1 Introduction

In most ground, mobile, tactical communication, radar, or EW system applications there is a requirement for small antennas. Fixed and rotary wing aircraft cannot have multiple large antenna arrays protruding from the skin of the aircraft either. With the proliferation of UASs, they demand small antennas as well. Urban warfare also influences the size of antennas that can be employed. We discuss the fundamental properties of *electrically small antennas* (ESAs) in this chapter [1].

We should keep it in mind that, ceteris paribus, a short antenna cannot perform as well as a larger antenna. That is just a law of physics. Utilizing some techniques, however, a short antenna can be made to perform better than otherwise as we will explain in this chapter.

There is a thrust in modern warfare to facilitate operations *on the move* (OTM). This was particularly evident in the first Gulf War in 1990, where entire corps were simultaneously moving into and through Kuwait. The ground EW systems at the time could not operate OTM—they had to stop and erect antennas to operate. Small antennas are a requirement for OTM applications.

In addition to small size, small antennas are usually characterized by low gain and limited bandwidth. In fact, for wire antennas such as dipoles and monopoles there is a minimum Q that the antennas can have. (Recall that the smaller the Q the broader the bandwidth.) ESAs are characterized by:

- Poor radiation performance;
- Very small bandwidth;
- Poor radiation and antenna efficiencies;
- Difficulty in matching to standard transmission lines;
- Poor polarization performance.

The size of a small antenna is defined relative to the wavelength. The most common definition of electrically small antennas is one for which $l < 0.1\lambda_H$, where λ_H is the wavelength at the highest frequency of operation, and that is the definition we use here. Such antennas are referred to as electrically small.

We also discuss methods for foreshortening antennas. This is accomplished predominantly by surrounding an antenna with a dielectric material. As discussed in Appendix D, this effectively slows the wavefront and allows a small antenna to exhibit some attributes of longer antennas. While several forms of dielectric materials exist, perhaps the most prolific come from the ceramic family.

We provide a brief introduction to ESAs in this chapter. An in-depth discourse on the characteristics of small antennas was presented by Fujimoto et al. [2]. That reference is recommended for those who wish to delve deeper into the subject.

This chapter is structured as follows. We first consider techniques of loading a small antenna by immersing it in a volume of dielectric material. Next we discuss active antennas. An active antenna is an antenna with some form of amplifier incorporated. These antennas are normally used as receive antennas; for transmitting the amplifier can be incorporated as part of the power amplifier. Next we discuss active antennas that incorporate the active components not as amplifiers but as *negative impedance converters* (NICs). The bandwidth-limiting components of the antenna impedance are cancelled by adding the negative of the antenna impedance. These antennas can be configured to be both receiving and transmitting antennas—we will discuss both. The last topic we cover is a discussion on a particular technique for using short antennas in the HF range by adding an amplifier. Adding an amplifier to a short antenna does indeed produce a higher signal at the antenna output, but any noise in the antenna or external to the antenna also gets amplified. In fact, the amplifier has a nonzero noise factor that adds noise to whatever noise is already there. We discuss a way of suppressing that noise.

14.2 Miniaturization Issues for Antennas

One set of targets for EW systems worldwide is cellular phone systems and *personal communication systems* (PCSs). These systems operate in designated regions of the 825–2,500 MHz frequency range. It would be highly desirable to provide EW functionality over this entire range with one antenna. In addition, when other commercial services such as WLAN and UMTS are included, the band requirements increase to 5 GHz. Other frequency bands must be covered as well, since communication targets are spread throughout the spectrum. These needs require ultra-wideband (>10:1) antennas, such as the spiral. In this section we

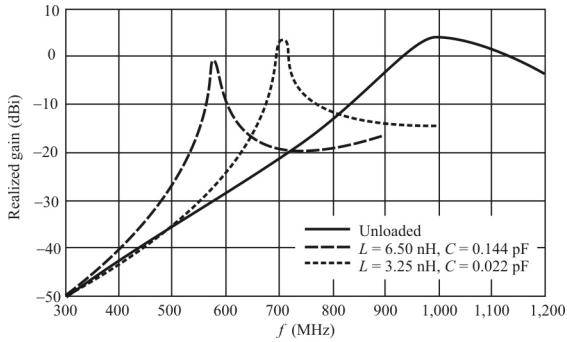


Figure 14.1 Realized gain of resonant loop antenna. Resonance point is moved to a lower frequency by loading the antenna with lumped elements.

consider novel methods for miniaturizing spirals and other antennas using loading techniques including dielectric loading, artificial lumped loads, textured dielectrics, and other approaches within the context of metamaterials. The goal is to achieve miniaturization without compromising gain and by retaining as much bandwidth as possible [3].

An antenna is resonant when at least one of its dimensions is of the size of half a wavelength. The wavelength at a given frequency depends on the dielectric and magnetic properties of the material surrounding the antenna according to

$$\lambda = \frac{\lambda_0}{\sqrt{\varepsilon_r \mu_r}} \tag{14.1}$$

Thus, the size of a resonant antenna can be decreased by increasing the dielectric or magnetic constant around the antenna.

Traditional antenna miniaturization techniques involve use of materials with dielectric constant values ranging over $\varepsilon_r = 9$ to $\varepsilon_r > 100$. Although size reduction using these materials has been demonstrated for narrowband antennas [2], such as patches, loading of broadband antennas is not as easy. The major issues are presented next.

14.2.1 Bandwidth Reduction

The gains of a resonant loop antenna with lumped loads that shift the resonance region to lower frequencies are illustrated in Figure 14.1 [3]. When lumped elements are used to shift the loop's resonance to lower frequencies, the bandwidth is significantly reduced as illustrated in Figure 14.1. The same effect occurs when

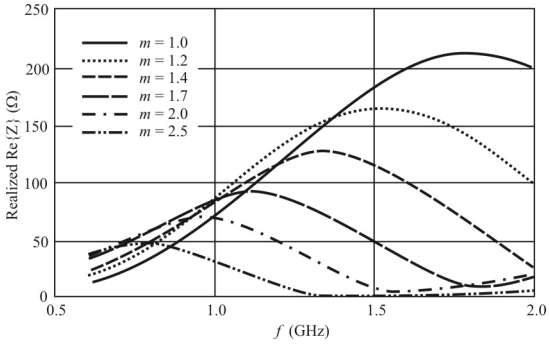


Figure 14.2 Re Z for resonant dipole. The impedance of a resonant dipole decreases as its size is reduced by factor *m*.

dielectric loading is applied to the antenna. [Ferromagnetic loading produces the same effect—see (14.1).] This is because the concentration of electric (magnetic) fields in the high permittivity (permeability) regions produces larger near fields, which, in turn, produce higher *Q* and therefore lower bandwidth.

14.2.2 Impedance Mismatch

The input impedance of a dipole decreases as its size is reduced by a factor *m* as illustrated in Figure 14.2 [3]. In general, we can say that antenna miniaturization below 0.25λ is extremely challenging even when dielectric loading or other wave slow-down techniques are used. A major reason for this relates to impedance mismatches. To address this issue, approaches based on dielectric texturing, tapering, and distribution of lumped loads within the antenna aperture can be used.

14.2.3 Gain Reduction

Gain reduction is virtually inevitable as an aperture size is reduced due to beamwidth broadening when dealing with apertures that are on the order of a wavelength. However, as the antenna aperture is reduced to less than 0.25λ, of greater importance is the impedance bandwidth and its associated impact on gain.

14.2.4 Effects of the Radius of Wire Antennas

The wire radius of wire antennas has a significant influence on the antenna's characteristics. Small changes in the radius can significantly change the fundamental parameters. An example of the effects on the VSWR for a dipole is illustrated in Figure 14.3 [4].

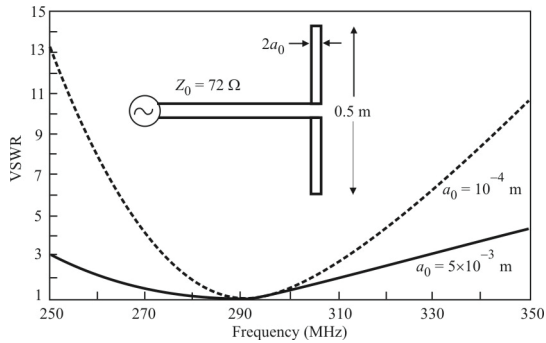


Figure 14.3 Calculated VSWR for two dipoles of different radii.

14.3 *Q* and Efficiency of Electrically Small Antennas

With few exceptions, EW systems are wideband. That is, they cover a significant portion of the frequency spectrum. Antennas typically do not share this property. It is often the case that the antennas are relatively small compared to a wavelength. In other words, they are electrically small. ESAs tend to have a high *Q*—that is, they have narrow bandwidths [5–7]. Furthermore, there is a fundamental theoretical limit to the bandwidth and radiation efficiency of ESAs [8–10]. The bandwidth characteristics of small antennas are investigated in this section.

The *electrical size* of an antenna is the physical dimensions of the antenna defined relative to wavelength. There are generally two categories of electrical size that are analyzed:

- *Electrically small antenna* where the dimensions of the antenna are small relative to the wavelength;
- *Electrically large antenna* where the dimensions of the antenna are large relative to the wavelength.

For example, consider the dipole antenna of length $l = 1$ m as shown in Figure 14.4. The electrical lengths of the dipole at $f = 3$ MHz and $f = 30$ GHz are also shown. At 3 MHz, a 1 m antenna is electrically small. However, at 30 GHz, a 1 m antenna is electrically large.

It has been established that for an ESA, contained within a given volume, the antenna has an inherent minimum value of *Q*. This places a limit on the attainable bandwidth of an ESA [2, 11, 12].

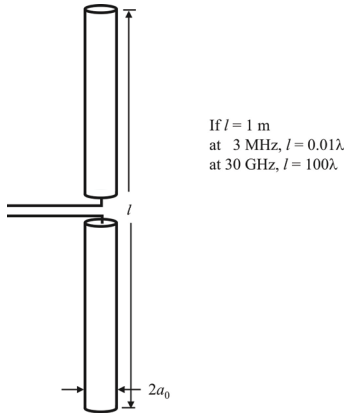


Figure 14.4 A dipole of given dimensions can be either short or long depending on frequency.

The first published work to address the fundamental limits of ESA was written by Wheeler in 1947 [13]. Wheeler defined an ESA as one whose maximum physical dimension is less than $\lambda / 2\pi$. This relation is often expressed as $\beta a < 1$ where $\beta = 2\pi / \lambda$ rad m^{-1} is the wavenumber, λ = free space wavelength in meters, and a = radius of sphere enclosing the maximum dimension of the antenna in meters. A sphere of radius $a = 1 / \beta = \lambda / 2\pi$ is defined as the *radiansphere* [14].

The radiansphere is regarded as the boundary between the near field and far field, the former being where the energy is stored and the latter being where signals are intercepted. Frequently the boundary between the near field and far field is determined from the frequency at which $\beta a = 1$, in other words at one radiansphere [15, 16].

The efficiency of an ESA is determined by the amount of losses in the antenna compared with the radiation resistance:

$$\eta_a = \frac{R_r}{R_r + R_L} \tag{14.2}$$

where

η_a is the efficiency of ESA.
R_r is the radiation resistance (Ω).
R_L is the material loss resistance (Ω).

The input impedance of an ESA is capacitive [that is, the reactance is negative, $jX = 1/(j\omega C) = -j/(\omega C)$], and for maximum power transfer the antenna must be conjugate matched to the transmission line. Thus, a matching network may be required. The efficiency of the antenna with its matching network is expressed as

$$\eta_s = \eta_a \eta_m \tag{14.3}$$

η_s is the efficiency of system (i.e., antenna with matching network).
η_m is the efficiency of matching network.

It can be shown that the efficiency of the matching network is approximately:

$$\eta_m \approx \frac{\eta_a}{1 + \dfrac{Q_a}{Q_m}} \tag{14.4}$$

where

Q_a is the Q of ESA.
Q_m is the Q of matching network.

The approximate bandwidth for an RLC type circuit in terms of Q is:

$$W = \frac{S-1}{Q\sqrt{S}} \tag{14.5}$$

where

S is given by S : 1 VSWR (≥1);
W is the normalized bandwidth.

It is difficult to discuss the Q of small antennas since, in general, they are too small to resonate at any but high frequencies. However, if the matching circuit is included in the analysis then the reactive component of the antenna can be neutralized and it is then reasonable to define the Q of a small antenna. Also the matching network should match the real part of the driving impedance to the real part of the antenna load for maximum power transfer. This is true whether the antenna under consideration is for receiving or transmitting [17].

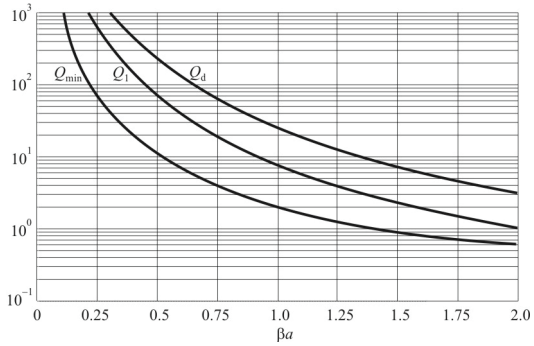


Figure 14.5 Minimum Q for dipole (Q_d), loop (Q_l), and the minimum limit (Q_{min}) given by (14.7). For the dipole and loop $l = 1$ and $a_0 = 0.01$, 100% efficiency.

The most generally accepted definition for the Q of an antenna is

$$Q = \begin{cases} \dfrac{2\omega U_e}{P_{rad}}, & U_e > U_m \\ \dfrac{2\omega U_m}{P_{rad}}, & U_m > U_e \end{cases} \tag{14.6}$$

where U_e is the time-average, nonpropagating, stored electric energy, U_m is the time-average, nonpropagating, stored magnetic energy, ω is the radian frequency, and P_{rad} is the power of the signal propagated [18]. Assuming a perfect, lossless matching network, (14.6) leads to the expression for the minimum Q for a small, linear electric-current antenna as derived by McLean [18] to be

$$Q_{min} = \frac{1}{\beta^3 a^3} + \frac{1}{\beta a} \tag{14.7}$$

where a is the radius of an enclosing sphere. Equation (14.7) is plotted in Figure 14.5. The maximum bandwidth of an antenna, relative to the center frequency, based on (14.7) is given by [19]

$$W = \frac{\beta^3 a^3}{1+\beta^2 a^2} \frac{S_{max}^2 - 1}{2S_{max}} \tag{14.8}$$

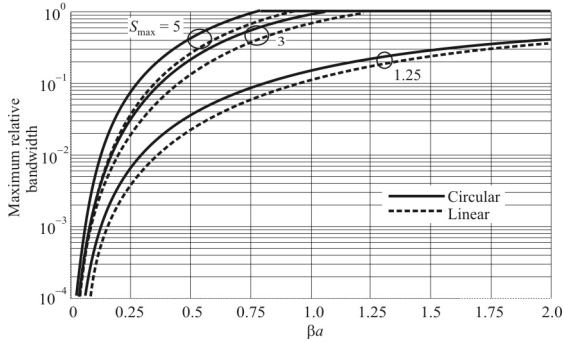


Figure 14.6 Maximum relative bandwidth for circular and linear polarization.

where S_{max} is the maximum VSWR. Equation (14.8) is plotted in Figure 14.6 for representative values of VSWR. Clearly, to obtain any significant bandwidth for VSWR < 2 or so, fairly large antennas are required, $\beta a > 0.57$ or so (βa is about 0.5 at 25 MHz with a 2 m dipole). Also it is evident from Figure 14.6 that the assumptions break down for larger values of VSWR and large values of βa, since a fractional bandwidth greater than one is not possible.

Equation (14.7) assumes a uniform current along the length of the short dipole (approximated for an infinitesimal dipole). In practice, the Q obtained by using that expression is difficult to achieve, and generally higher Qs and therefore narrower bandwidths are obtained. In an attempt to determine a better lower limit on the Q of small antennas, Thiele, Detweiler, and Penno [9] took a different approach. They determined the lower bound on the Q of small antennas by considering the far field and assumed a sinusoidal distribution of the current on the antenna. At the lengths considered, the sinusoidal pattern is approximately triangular. The Q they determined is given by

$$Q = \frac{\displaystyle\int_{-\infty}^{-\pi l/\lambda} \left|E_{\mathrm{n}}(u)\right|^2 du + \int_{\pi l/\lambda}^{\infty} \left|E_{\mathrm{n}}(u)\right|^2 du}{\displaystyle\int_{-\pi l/\lambda}^{\pi l/\lambda} \left|E_{\mathrm{n}}(u)\right|^2 du} \tag{14.9}$$

where $u = [(\beta l)/2]\cos\theta$, l is the length of the dipole, and $E_{\mathrm{n}}(u)$ is the normalized field pattern given by

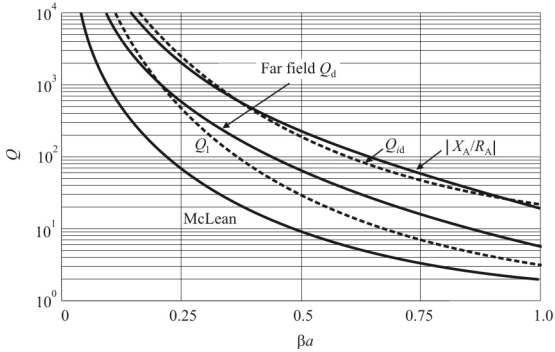


Figure 14.7 Radiation Q for infinitesimal dipole with radius a_0 = 0.005λ, 100% efficiency (Q_{id}), a small dipole (Q_d), as well as the radiation Q for a small loop antenna (Q_i).

$$E_n(u) = \frac{\cos u - \cos\left(\frac{\beta l}{2}\right)}{\left[1-\cos\left(\frac{\beta l}{2}\right)\right]\sqrt{1-\left(\frac{2u}{\beta l}\right)^2}} \tag{14.10}$$

The Q calculated with (14.7) is compared with that computed with (14.10) and that for a typical wire dipole in Figure 14.7 [9]. The "far-field Q_d" is the Q computed with (14.9). We see that the Q computed with (14.9) is lower than that for the uniform distribution assumption by approximately a factor of 2 over this range of βa. The dipole Q, Q_d, is seen to be well approximated with the ratio $Q = X_A / R_A$, when $Q > 10$ anyway.

The minimum Q for a circularly polarized antenna was derived by McLean [18] and is given by

$$Q = \frac{1}{2}\left(\frac{1}{\beta^3 a^3} + \frac{2}{\beta a}\right) \tag{14.11}$$

Following the same procedure as in [20] for the short linear antenna, the maximum bandwidth for such antennas, using (14.11) is

$$W = \frac{2(\beta a)^3}{1+2(\beta a)^2}\frac{S_{max}^2 - 1}{\sqrt{4}S_{max}} \tag{14.12}$$

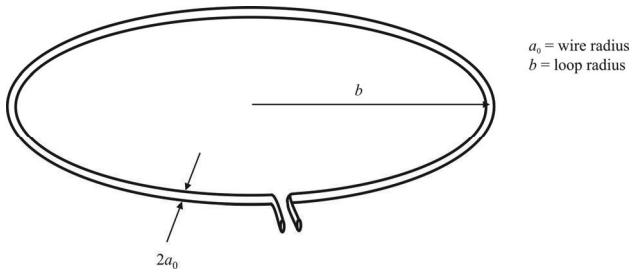


Figure 14.8 Infinitesimal loop.

This function is also plotted in Figure 14.6. The relative bandwidth for this polarization is higher than that for linear polarization for the same value of βa and VSWR. The circular polarization arises with a dipole, monopole, or loop. by correctly phasing and weighting the TM_{01} and TE_{01} modes[1] [18].

The Q of a small dipole is given by approximately [20]

$$Q_{\mathrm{d}} \approx \frac{48\left[\ln\left(\dfrac{l}{2a_0}\right)-1\right]}{(\beta l)^3} \tag{14.13}$$

where l is the total length of the dipole and a_0 is the radius of the wire, and it is assumed that $a_0 << l$. The antenna is illustrated in Figure 14.4. It is assumed that the width of the gap between the elements is infinitesimal.

The Q of a small loop antenna is approximately [20]

$$Q_{\mathrm{l}} \approx \frac{6\ln\left(\dfrac{b}{a_0}\right)}{\pi(\beta b)^3} \tag{14.14}$$

where b is the radius of the loop and a_0 is the radius of the wire and it has been assumed that $2b >> a_0$. Equations (14.13) and (14.14) are plotted in Figure 14.7 as well. The antenna is illustrated in Figure 14.8. We see that the loop Q, Q_{l}, is between the McLean minimum and the Qs for the dipole.

[1] The TE_{01} wave is the mode 1 electric field and the TM_{01} is the mode 1 magnetic field. (See Chapter 4.)

For large Q, the Q is approximately equal to the reciprocal of the fractional bandwidth. That is,

$$Q \approx \frac{f_c}{W} \tag{14.15}$$

where $W = f_H - f_L$ and $f_c = \sqrt{f_H f_L}$, f_c being the center frequency.

14.4 Lumped Element Loaded Antennas

14.4.1 Inductive Loading

Efficiency and input resistance of short monopole and dipole antennas can be significantly improved if a loading coil is used in series with the feed line.

Moving the tuning coil away from the feed point out into the antenna proper is also possible. Short monopoles and dipoles have approximately a triangular current distribution as illustrated in Figure 5.8 for the dipole and in Figure 6.4(a) for the monopole that is maximum at the feed-point and zero at the end (monopole)/ends (dipole). This current distribution brings with it high reactance and low radiation resistance. A loading coil acts to keep the current distribution essentially constant from the feed point to the loading coil with a linear decrease from the loading coil to the end. The radiation resistance can be increased by a factor of more than 4 this way.

14.4.2 Resistive Loading

Resistance added in series with the wire elements increase, the ohmic resistance of an antenna. This will tend to decrease the Q, thereby widening the bandwidth, while at the same time decreasing the efficiency of the antenna due to (3.40) as well as decreasing the overall system sensitivity.

14.4.3 Capacitive Loading

A top loaded monopole was shown in Figure 6.16. A plate mounted to the end is sometimes referred to as a "top hat." These top hats create a capacitance. Short wire antennas inherently exhibit a capacitive reactance characteristic in the first place, so loading such an antenna with further capacitive reactance is usually counterproductive. Other types of small antennas, however, can benefit from

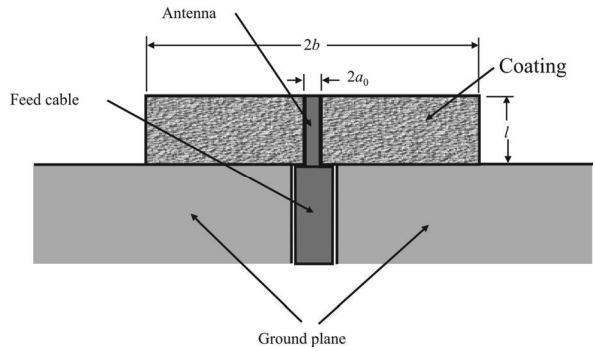


Figure 14.9 Coated monopole.

capacitive loading, for example, the patch antenna and inverted-F monopole. Such capacitive loading can be applied with lumped elements or in a distributed fashion.

14.5 Material Loaded Antennas

The characteristics of wire antennas can be altered by application of material coatings or otherwise modifying the near-field region of the antenna. When an antenna is surrounded (coated) with a material of infinite extent with relative dielectric constant (permittivity) ε_r and permeability μ_r, the wavelength is reduced by $F = \sqrt{\varepsilon_r \mu_r}$ [see (14.1)]. Therefore, for the same response, the antenna length can be reduced by this factor. This reduction does not come without other issues, however. The antenna impedance at resonance is reduced by a factor of $1/F$ and the bandwidth is reduced by $1/F$ as well. Since the coating does not extend to infinity, the effects will only be approximately given by F.

Either ε_r or μ_r or both can be increased to accomplish this foreshortening, although materials that accomplish one rarely accomplish the other. For the coaxial driven monopole shown in Figure 14.9 [2] with the dielectric coating of dimensions shown, the characteristics of some coatings are given in Table 14.1.

As we discussed in Chapter 6, adding ferrite inside a loop antenna significantly improves the performance of such "stick" antennas. Adding such ferrite material increases μ_r. This approach has been used for years in receivers for the lower-frequency range (HF).

14.5.1 Material Loading

Shown in Figure 14.10 [21] are the impedance characteristics of a 1m long, center-fed, untuned, lossless straight wire antenna of diameter 1 mm embedded in a lossy

Table 14.1 Characteristics of Some Antenna Coatings

Case	A (uncoated)	B	C	D
a_0 (cm)	0.3	0.3	0.3	0.3
l (cm)	7.6	14.85	2.15	1.65
Dielectric material	None	Titanium Dioxide	Varium Titanate	Barium Titanate
ε_r	–	18	90	90
b (cm)	–	14.1	14.1	10
Resonant frequency (MHz)	900	900	910	900
Bandwidth, W (MHz)	236 (26%)	145 (16%)	104 (11%)	144 (16%)
Resistance at resonance (Ω)	38.8	14.5	14.5	14.4
F	1.0	1.98	14.53	14.61

Source: [2].

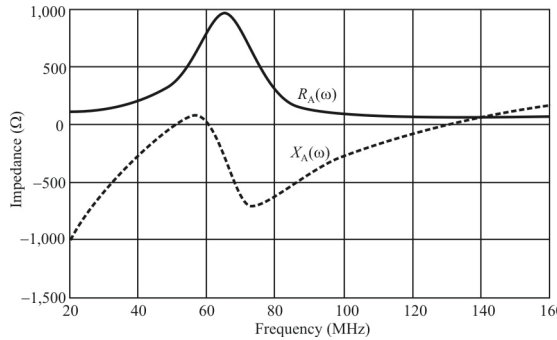


Figure 14.10 Characteristic impedances of a 1 m long, untuned, lossless straight wire antenna with 1 mm diameter, embedded in a lossy dielectric material.

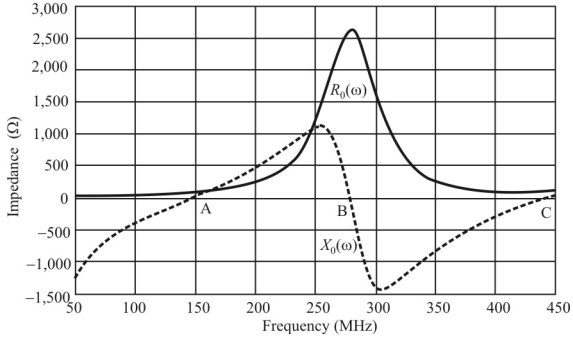


Figure 14.11 Impedance of 1 m long, center-fed, untuned, lossless straight wire antenna with diameter 1 mm. This is Figure 3.14, repeated here for convenience.

dispersive dielectric. The dielectric constant is given by [21]

$$\varepsilon(\omega) = \varepsilon_r(\omega) - j\varepsilon_i(\omega) \tag{14.16}$$

$$= \varepsilon_0 \left[1.5 + \frac{1.2}{1 - \left(\frac{\omega}{2\pi \times 50 \times 10^6} \right)^2 + 0.4j \left(\frac{\omega}{2\pi \times 50 \times 10^6} \right)^2} \right] \tag{14.17}$$

Compare this with the same antenna without the dielectric whose characteristics were first presented in Figure 3.15, repeated here as Figure 14.11 [21] for convenience [21]. Note the change in scales on both axes. The impedances of the loaded antenna peak around 60 MHz while the impedances of the unloaded antenna peak much higher, between 250 and 300 MHz. The bandwidth of the resistive component of the unloaded antenna is much wider than that for the loaded antenna as well. So we can say that the loading lowers the resonant and antiresonant frequencies and lowers the bandwidth of the antenna. In addition, the actual impedance levels are lower in the resonant and antiresonant regions as well.

14.5.2 Double-Negative Materials

A relatively recent approach for loading of electrically small antennas is to use *double negative* (DNG) materials. That is, materials for which $\varepsilon_r < 0$ and $\mu_r < 0$. As far as we know, these materials do not exist naturally in nature and must be

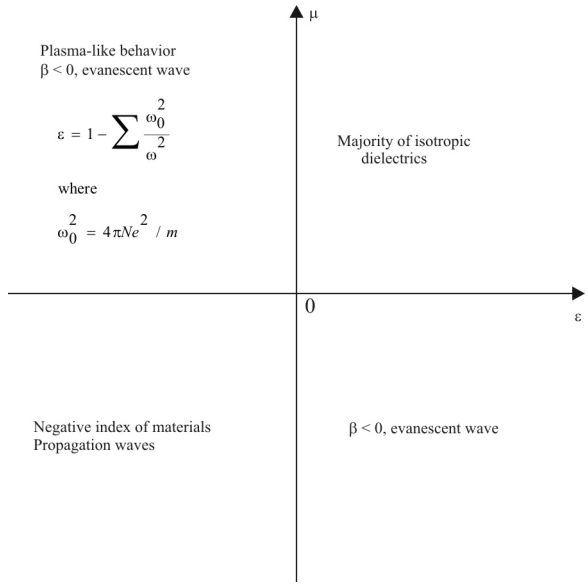


Figure 14.12 Metamaterials structures and their location in ε, μ phase space.

artificially fabricated. Figure 14.12 illustrates where such materials stand in the ε-μ phase space.

When one (but not both) ε_r and μ_r is negative, plane waves decay exponentially, like modes below cutoff in a waveguide (evanescent waves, see Chapter 4). However, when both ε_r and μ_r are negative, waves can still propagate in such a medium since the product $\varepsilon_r\mu_r$ remains positive. In this case ("backward wave"), the phase of the wave moves in the opposite direction from that of the energy flow. Strange properties ensue from this phenomenon, such as the Doppler shift reverses—a train coming toward us has a lower frequency than after it passes by, when they are higher. Another is the creation of a negative index of refraction (see Figure 14.13).

The normalized reactance of *double-positive* (DPS) materials, where $\varepsilon_r > 0$ and $\mu_r > 0$, for an electrically small dipole is approximated with

$$X_R^{DPS}(a) \approx -\frac{1}{(\beta a)^3} \tag{14.18}$$

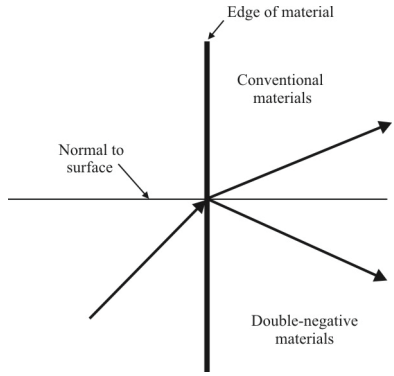


Figure 14.13 Characteristics of negative index of refraction.

where again, a is the radius of the smallest sphere completely containing the dipole. Since this reactance is less than zero, it is capacitive. Meanwhile, for a DNG material it can be shown that

$$X_{\mathrm{R}}^{\mathrm{DNG}} = -X_{\mathrm{R}}^{\mathrm{DPS}} \tag{14.19}$$

That is, it is inductive in nature. The question then naturally arises: Can the small dipole be matched with DNG material loading? The answer is affirmative, and we show a few charts here to show the significance of this.

The antenna configuration is illustrated in Figure 14.14 [22]. The infinitesimal dipole current element is given by $I_0 l$ and the sphere completely enclosing the antenna is shown. Out to a radius of r_1, the material loading is DPS. From r_1 to r_2, the material loading is DNG. Beyond r_2, the material is DPS again. As an example of the effects of this configuration, consider the simulation results shown in Figure 14.15 [22]. This chart shows the radiated power gain as a function of r_2, of the infinitesimal dipole with the DNG shell where the inner radius r_1 = 100 μm. The maximum gain occurs at r_2 = 748.8 μm. We see that a gain over the case of not using the coating is over 70 dB in this case. DNG materials for antenna coatings are still in the research stage, however. A search for naturally occurring DNG materials is ongoing as well.

14.6 Active Receive Antenna Systems

Passive antennas are reciprocal—they can be used with equal performance in both transmit and receive roles. That is not necessarily true with active antennas.

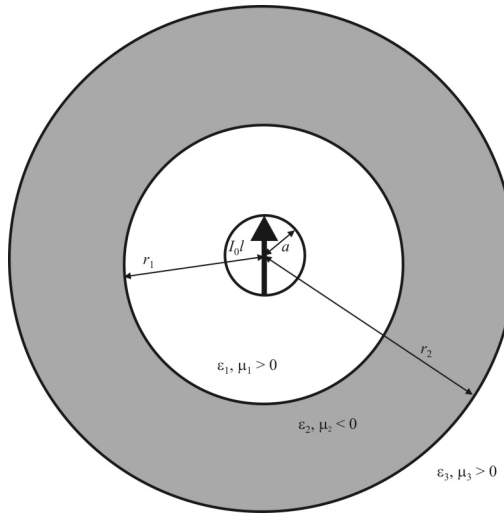


Figure 14.14 Infinitesimal electric dipole surrounded by DNG material.

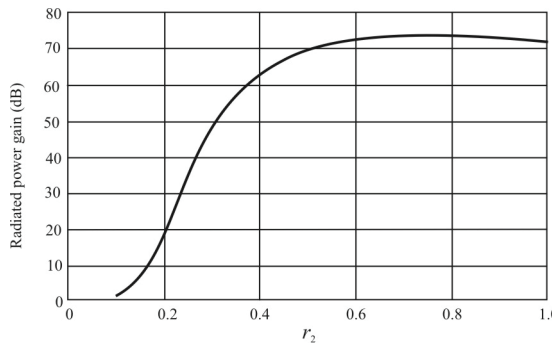


Figure 14.15 Radiated power gain of the dipole shown in Figure 14.14.

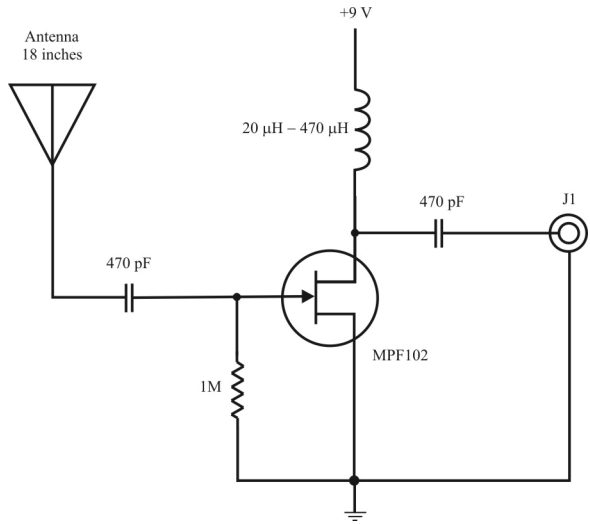


Figure 14.16 An active antenna. The output driver in this circuit converts to a low impedance, matching typical cables.

Shorter antennas can match some of the characteristics of larger (longer) antennas by using an active device (amplifier) with the antenna. Such antennas are called *active antennas*. For short wires, say, less than 10 meters or so, the impedance of the wire decreases with increasing length. As the length increases the amount of EM wave energy collected by the antenna increases. For short wire antennas, then, not much signal energy is collected. The amplifier at the output of the short antenna (consisting of BJTs, FETs, or vacuum tubes) amplifies what signal is collected, more or less to the same levels as the output of much longer wire antennas. Figure 14.16 shows an example of a short active antenna [23] for the HF and lower-frequency ranges. The value selected for the inductor depends on the frequency range of the amplifier in this figure.

The reception of RF signals in the frequency region 10 kHz–30 MHz is external noise limited with the noise originating as atmospheric, man-made, and galactic noise. Because of these significant noise levels, reducing the noise characteristics of a receiving system below some point yields diminishing returns. Since decreasing the length of a dipole or monopole increases its impedance, it is possible to maintain the external SNR as the limiting characteristic even with short, very high impedance-antennas, where $l << \lambda$. Small antennas are very desirable in many applications, as tactical communication systems are typically mobile. Therefore, the EW system typically must be mobile as well. Mobility implies the smaller the better.

An active antenna can have various additional advantages, including:

- Its transfer function can be made broadband and frequency-independent and can be accurately fixed.
- Its output impedance can be accurately matched to the characteristic impedance of the cable to the receiver.

Because an amplifier forms a part of the active antenna, noise and distortion products are inevitably added to the received signals. Design efforts must therefore be focused on minimizing these contaminating signals and maximizing signal-handling capability.

Amplifier noise minimization offers the possibility of reducing antenna length. The reduction of intermodulation distortion and the enlargement of the signal-handling capability increase the suitability of active antennas on locations where strong interfering signals are present. Other important design aspects are the accuracy of the antenna transfer function (for application in antenna arrays) and the protection against statics.

The disadvantage of active antennas is that interferers become larger in amplitude as well. Also the amplifier adds noise to what otherwise might be a low-noise signal. In addition, like all active devices and many devices that are not active, an amplifier is truly linear over only a portion of its operating range. Thus, intermodulation (IM) products can be generated in it, particularly by close, strong, unwanted signals. These IM products must be filtered to prevent significant corruption.

For use in arrays, active antennas must be amplitude and phase matched in their characteristics. If not, errors will be generated that are difficult to manage, especially over time and temperature.

The principal advantage of active amplifiers is the size of the antenna element. They can be very short relative to their passive counterparts. This is particularly important in the lower-frequency ranges (HF and less) because antennas are large in this region. The antenna shown in Figure 14.16 is only 18 inches long in the HF range. In the higher-frequency ranges, antennas are relatively small anyway, so active antennas are not used much there. Since the first amplification stage in an amplifier chain is the predominant contributor to the noise figure, it is important that the devices used in an active antenna be of the low-noise variety.

We present in Figure 14.17 another approach to active antenna design [24]. This amplifier is intended to be collocated with the antenna, and the length of the antenna is such that $l \ll \lambda$. The frequency range is 5 kHz to 30 MHz. The innovative approach for this amplifier is to use negative feedback around several amplification stages, which are enclosed within the amplifier module in the figure. This implements a virtual ground at the input to the amplifier. This virtual ground

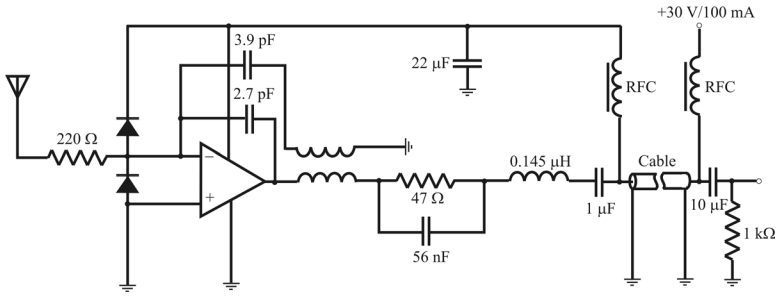


Figure 14.17 Active antenna with negative feedback around more than one stage.

facilitates some static protection as well as helps to maintain a linear amplifier transfer function.

The experimental performance achieved is illustrated in the transfer characteristics shown in Figure 14.18 [24], where P_{out} (dBm) is plotted versus the electrical field strength E_{RMS} (V m^{-1}). The antenna length in this case was 0.5 m. Extremely high dynamic ranges were achieved with second-order intercept points at 69–80 dBm; as well as high third-order intercept points: 53 dBm.

14.6.1 Matching to Small Antennas with Non-Foster Elements

We mentioned above that the impedances of electrically small antennas are quite low, which makes impedance matching to them difficult. Sussman-Fort [25–27] matured the process of using non-Foster impedance elements to address this problem [28].

The goal is to obtain broadband and efficient electrically small antennas.

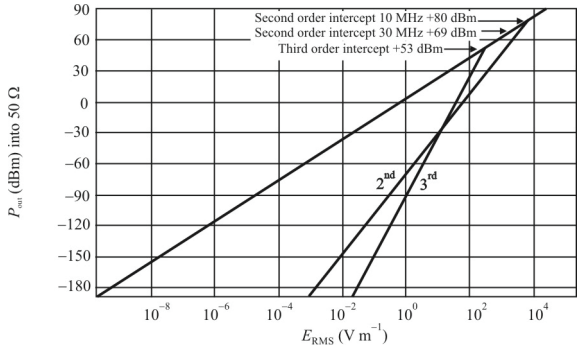


Figure 14.18 Radiated power from the active antenna shown in Figure 14.17.

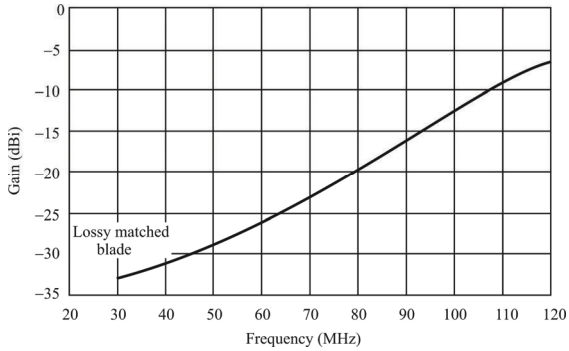


Figure 14.19 Short monopole gain characteristic. In this case the antenna is a 6 inch blade (λ/2 at about 900 MHz).

From above, an antenna is electrically small when its length l satisfies

$$l << \lambda \tag{14.20}$$

The electrically short monopole, for example, has large reactance and low radiation resistance that squares with frequency. The radiation resistance is given by (6.8). If $l = \lambda/100$ for example, $R_r \approx 0.14\ \Omega$. Also from (6.8) it is clear that $R_r = Kf^2$ for the appropriate constant K.

The antenna reactance is responsible for storing the nonpropagated energy in the near field while the radiation resistance is responsible for the radiated power. The purpose of the matching circuit is to remove the reactance by cancellation with a conjugate value, while at the same time providing a matching resistive component. Therefore, in this case, the matching resistance must be about 0.14Ω, a very small value, difficult to obtain.

The other significant problem with short antennas is their bandwidth. When maximizing gain, the shorter the antenna, the higher its Q (see Figure 14.7). The higher the Q, the narrower the bandwidth.

The noise figure of a typical receiver in the lower VHF frequency range is around 10–15 dB, while the gain of a typical small passive antenna in this range is illustrated in Figure 14.19. Even though the external noise is quite high in this frequency range, because of the low antenna gain the sensitivity is determined by the noise figure, and is not externally noise limited. Therefore, if some method could be found to increase the gain, the sensitivity would improve. That is the goal of matching with non-Foster impedance elements.

The gain-bandwidth product of passive matching circuits is limited by the Bode-Fano criterion [29]. For example, for a lossless matching network that

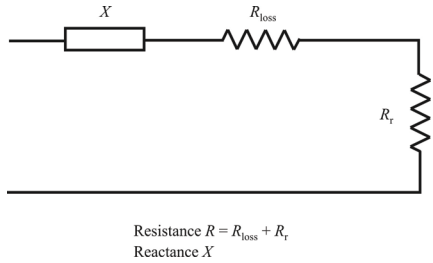


Figure 14.20 Antenna model.

matches a transmission line with reflection coefficient $\Gamma(\omega)$ to a load consisting of a capacitor of value C and resistor of value R in parallel, the Bode-Fano criterion says that

$$\int_0^\infty \ln\left|\Gamma(\omega)\right|^{-1} d\omega < \frac{\pi}{RC} \tag{14.21}$$

Matching to other load combinations of Foster elements has similar limitations. These limits mean that good passive matching to electrically small antennas cannot be accomplished.

Non-Foster impedance matching is defined as the use of *negative* inductors and *negative* capacitors to manage the transfer of power between a source and a load. It is the *negative impedance converter* (NIC), discovered over 50 years ago [30], that creates negative elements by transforming conventional passive *L*s and *C*s into their negative counterparts. For a lossless one-port with positive elements and whose driving-point impedance is given by $Z(j\omega) = R(\omega) + jX(\omega)$, Foster's reactance theorem says that the slope $dX/d\omega$ must be positive. With negative elements this slope can be negative; hence, the appellation "non-Foster" is applied to networks containing negative reactive elements.

Given a driving point impedance of any network, including an antenna that is modeled as seen in Figure 14.20, $Z(\omega) = \mathrm{Re}\{Z(\omega)\}+j\mathrm{Im}\{Z(\omega)\} = R(\omega) + jX(\omega)$ then Foster's theorem states that

$$\frac{dX(\omega)}{d\omega} > 0 \qquad \forall\omega \tag{14.22}$$

In other words, this says that the slope of the reactance versus frequency curve is always positive. The approach with non-Foster networks is to cancel the reactance by using a negative reactance of the same absolute value. Since matching networks of non-Foster elements are active, not passive, the limitations implied by (14.21)

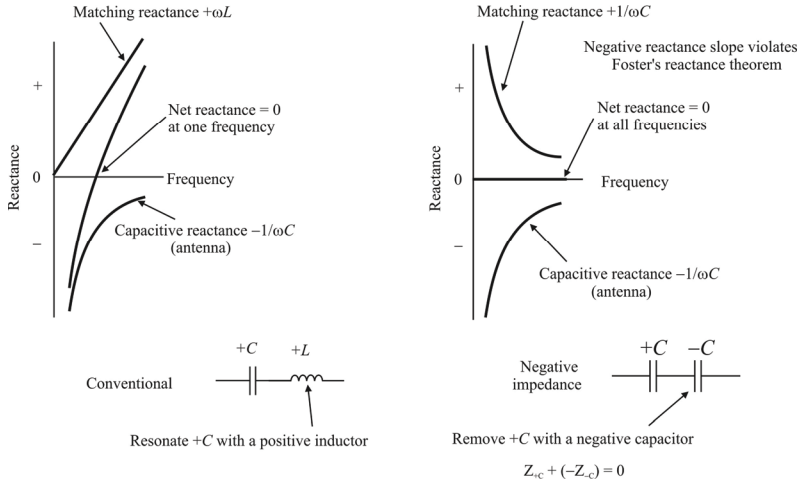


Figure 14.21 Conventional versus negative impedance matching.

do not apply. In fact, non-Foster elements are negative reactances, $-L$ and $-C$. The general notion is illustrated in Figure 14.21 [31]. In the conventional approach to impedance matching, a capacitive antenna load is matched to the transmission line by adding an inductor in series so the impedances cancel, as seen on the left. As such, the matching occurs at a single frequency as shown. If negative impedances are used instead, the capacitive antenna load is cancelled by adding a negative capacitive reactance in series as illustrated on the right. Ideally the two reactances cancel making the load look resistive at all frequencies. The characteristics of the negative slope of the negative capacitor shown on the right in Figure 14.21 violates the Foster requirement.

The floating and grounded NICs were discussed by Linville [30]. Connecting a positive impedance on one side of the NIC produces a negative impedance at the input terminals as shown in Figure 14.22. The NICs can also be used to modify the frequency dependence of the radiation resistance. The networks shown in Figure 14.22 accomplish this.

The input impedance of the dualizer networks shown in Figure 14.23 is

$$Z_{\text{in}} = \frac{L_0^2}{Z_{\text{L}}}\omega^2 \tag{14.23}$$

where $Z_{\text{L}} = K\omega^2$ is the frequency-dependent load resistance. Therefore to make the input impedance a real constant value R_0, let

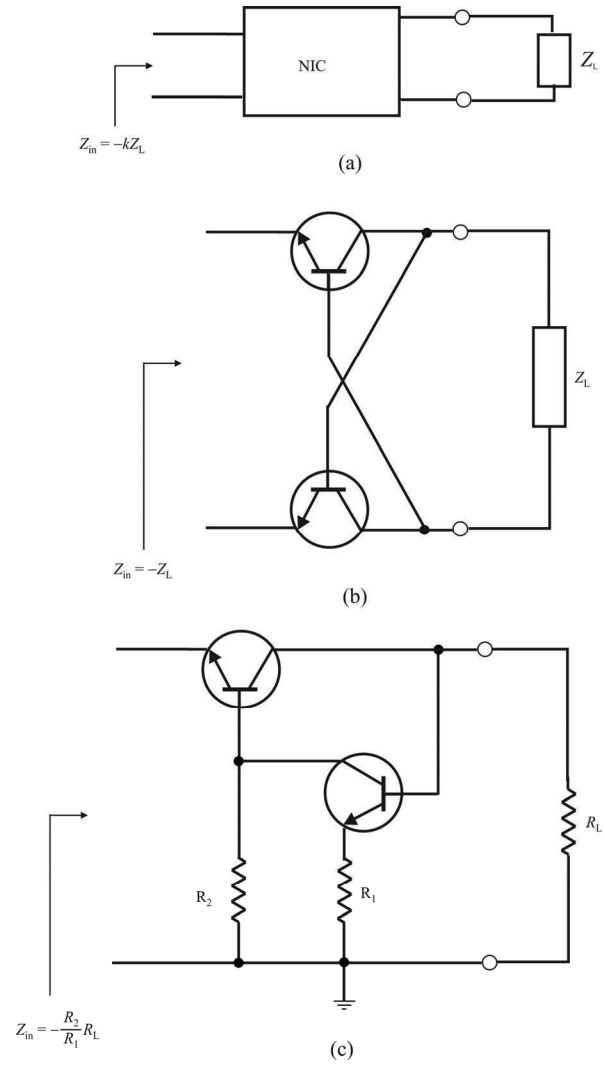


Figure 14.22(a–c) Negative impedance converters: (a) a negative element is prodded by terminating a NIC with a corresponding positive element; (b) floating negative impedance; and (c) grounded negative resistance.

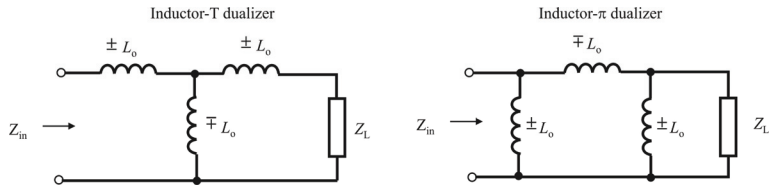


Figure 14.23 Dualizers convert the frequency-dependent radiation resistance to a real and constant value R_0.

$$L_0^2 = KR_0 \tag{14.24}$$

By negating (or minimizing) the detrimental effects of the reactive component variation with frequency combined with the effects described here on negating (minimizing) the variation with frequency of the radiation resistance, we can achieve a constant impedance antenna—at least over a limited portion of the spectrum.

The measured results that are achievable with non-Foster reactive elements are illustrated in Figure 14.24 [31]. For the tee and π networks shown

$$Z_{\text{in}} = \omega^2 L_{\text{o}}^2 / Z_{\text{L}}$$

If Z_L is the frequency-dependent radiation resistance

$$Z_{\text{L}} = k\omega^2$$

and if we want the input impedance Z_{in} to be the real and constant value R_o, we choose

$$L_{\text{o}}^2 = kR_{\text{o}}$$

The short monopole with the non-Foster elements was 6 inches while the passively matched antenna was twice as long. The system gain at the bottom of the frequency range was improved by approximately 20 dB. The improvement in SNR is illustrated in Figure 14.25 [31].

To garner a sense of what this improvement means, VHF signals propagating close to the ground suffer approximately 12 dB per octave distance loss, assuming they encounter no obstacles. Gaining about 9 dB with these active matching

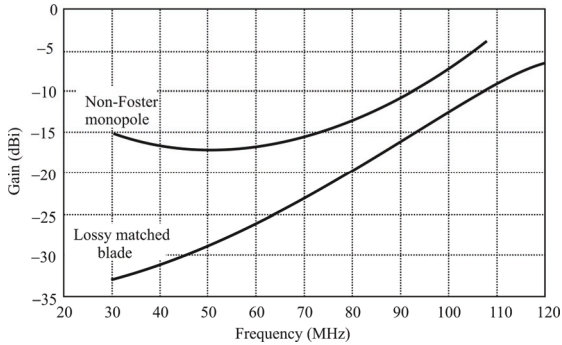


Figure 14.24 Results achievable by using non-Foster elements.

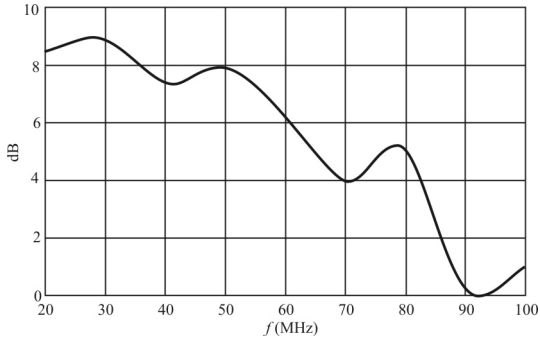


Figure 14.25 SNR advantage of non-Foster matching. Improvement in SNR: 6 inch monopole with non-Foster matching over passive 6 inch monopole.

networks almost doubles the intercept range of an EW system (again, assuming no signal blockage).

14.7 Active Transmit Antennas

Sussman-Fort and Rudish [32] reported on the application of non-Foster impedance matching techniques to eliminate or minimize the reactive impedance from transmit antennas. As above for receive antennas, non-Foster impedances are negative and, when properly configured in circuits, can cancel positive impedances. Measurements have shown that larger values of transducer power gain ($|S_{21}|^2$) can be achieved over broader bandwidths than with any comparable passive match. However, evaluation of power efficiency must be included in any comparison between non-Foster and conventional transmit matching. Ceteris paribus, whatever gain advantage that results by using non-Foster techniques can be overcome, in principle, by using larger transmitter power with conventional matching. In addition, these non-Foster matching approaches are based on the use of negative-valued components. Such components can become unstable, especially troublesome at the power levels of concern to EW systems. Instability in high power amplifiers can lead to device destruction.

14.7.1 Transmit Antennas

In EW transmitters, very large RF voltages and currents arise at the terminals of an ESA due to even moderate levels of radiated power. This causes the transistors or other active components of the NICs used in a non-Foster matching circuit to be vulnerable to damage. A channelized approach to matching is one way of dealing with this problem where [27]:

- The wide frequency band is divided into channels.
- Individual matching networks are implemented for these smaller bands.
- Negative elements are realized by terminating NICs in series-LC combinations resonant in a channel. In so doing, the voltage across any LC will be small at and near resonance, and hence the NIC devices would be protected.

Naturally the current limit of the NIC transistors must be adequate. For present purposes we consider the equivalent circuit model of an electrically small monopole shown in Figure 14.26 [31].

The variation of radiation resistance with frequency over the band of 15–30 MHz was ignored since the task of interest was to neutralize the reactive

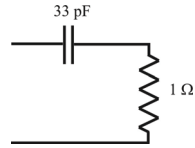


Figure 14.26 Antenna model. $Q = 250$ at 20 MHz .

impedance. The physical antenna was a two-foot vertical monopole with a Q of approximately 250 at 20 MHz. A 2-foot antenna at 20 MHz certainly qualifies as electrically small ($L \approx 0.02\lambda$).

14.7.2 Matching the Transmit Antenna

Electrically small antennas, as pointed out in Chapter 3, are normally capacitive. The antenna with the model shown in Figure 14.26, a 33 pF capacitor in series with a 1Ω resistance, was matched to a 4Ω resistance representing the output stage of a transmitter at the output transistor. The passive match for this antenna was determined to be a series 1.8 μH inductor shown in Figure 14.27. The source in this case had a characteristic impedance of 50Ω and passive matching was used to derive the 4Ω output impedance from this source as illustrated in Figure 14.26.

To cancel the series combination of the inductor and capacitor, a NIC was inserted with the load impedance consisting of the same combination of 1.8 μH inductor and 33 pF capacitor as illustrated in Figure 14.28.

14.7.2.1 Results

Figure 14.29 [31] illustrates the results achieved with this technique, where the S_{21} advantage of the NIC approach is compared with the best possible passive match. Over the frequency range considered (15–30 MHz), the NIC showed considerable variability but at worst was over 5 dB better than the passive matching. The measured advantage in S_{21} of active matching over passive exceeded 20 dB over the lower part of the band. The best possible passive match between a 33 pF-1Ω

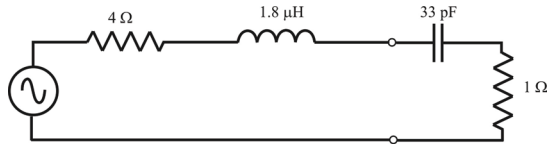


Figure 14.27 Passive match of antenna.

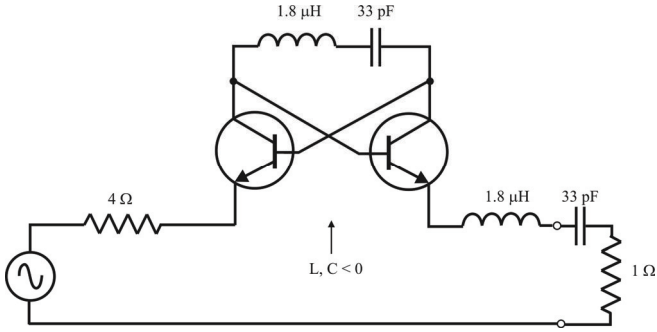


Figure 14.28 Active antenna match.

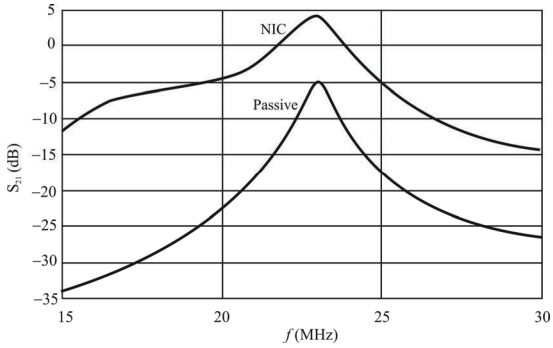


Figure 14.29 Comparison of S_{21} for passive antenna and NIC-coupled antenna.

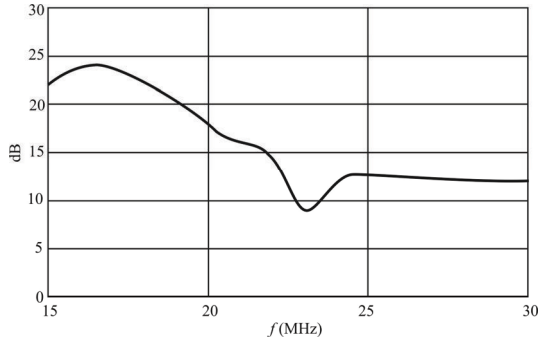


Figure 14.30 Advantage of S_{21} for NIC-coupled antenna over passive antenna.

load and 50Ω over 15–30 MHz was determined and used for charting Figure 14.29. Because of gain-bandwidth constraints and component losses, an average improvement in S_{21} of 20 dB can be expected over the band (see Figure 14.30).

While there are other considerations in the use of these non-Foster implementations in small transmit antennas, such as the high voltages and currents involved and the power handling capacity of the transistors, these results show promise for using such implementations. S_{21} is certainly an important parameter in the implementation of any PA, but it is not the only consideration.

The power efficiency of the active match still needs to be determined. The DC power required to bias the NIC was measured at 800 mW (29 dBm). The maximum signal power delivered to the load was approximately 4 dBm at 23 MHz. From Figure 14.29, for the best passive circuit to deliver as much power to the load, the input signal power would have to be about 24 dB greater or 28 dBm.

Note that the particular antenna discussed here with the NIC included also functions as a receive antenna since the NIC is symmetrical.

14.7.3 Summary

Non-Foster impedance matching bypasses the gain-bandwidth constraints of passive LC networks and allows octave bandwidth matching with a substantial increase in transducer power gain.

14.8 Noise-Suppressing Active HF Antenna[2]

In this section we discuss an efficient foreshortened antenna for the HF range. Efficient antennas ($l = \lambda / 2, \lambda / 4$) in the HF range are large—the wavelength at 1

[2] This antenna configuration has a U.S. patent pending.

MHz is 300 m and at 10 MHz, it is 30 m. These antennas are not suitable for tactical EW systems, especially for OTM operations. As discussed above, active antennas have been devised that employ small (18 inches) physical antennas in conjunction with an amplifier for operation in this frequency band, but noise gets amplified along with the desired signal as do all other signals in the passband of the amplifier; the SNR does not get enhanced. In fact, the output SNR gets worse since the noise in the amplifier is added to the input signal and noise waveforms.

To put the situation is perspective, we begin with a discussion of the signal and noise spectrum in the HF band to establish a common understanding of the problem being addressed. The noise characteristics considered consist of both the external noise sources as well as internal system noise sources that must be dealt with. These sources are the raison d'être for this antenna configuration. This is followed by a description of an approach to address the noise problems in the form of an active antenna. Such an antenna allows for the use of foreshortened antenna elements while achieving sensitivity levels of much longer antennas. The processing involved to address the noise added by the amplifier in the antenna is based on *high-order statistics* (HOS) which are known to have properties that suppress certain types of noise.

The key motivations for this antenna system are that:

- Active antennas can be used in the HF range that are small.
- External and internal noise in the HF range is accurately modeled with symmetric power spectral densities (PSDs) while thermal noise is additive Gaussian and atmospheric and man-made noises are modeled as alpha-stable processes.
- The bispectrum function in HOS signal processing [33] suppresses signals with symmetric PSDs and is identically zero for AWGN.
- Active antennas with HOS signal processing can be used to:
    - Shrink the size of HF antennas to multiple inches with the performance of long-wire antennas;
    - The signals can be de-noised allowing for enhanced SNRs.

The following sections detail the characteristics of signals in the HF frequency range as well as the noise that must be dealt in this frequency range. Also discussed is an approach to facilitate considerably improved intercept for EW systems employing this antenna structure.

14.8.1 High-Order Statistics

The moments of a stochastic signal are statistical measures that characterize the signal properties. The moments used most of the time in signal analysis are the

mean (average) and the variance (distribution of the signal about the mean). The mean and variance are the first and second moments, respectively. The moments characterize a signals probability distribution.

If the process probability distribution is Gaussian (normal), then it is completely characterized by the mean and variance. The moments higher than $n = 2$ are zero for Gaussian r.v.s. However, if the process is not Gaussian, then the (non-zero) moments of higher orders are needed to fully describe the distribution.

Traditional analysis of stochastic processes is centered on the second-order statistical properties of the process. In particular, if $x(t)$ is such a stochastic process, then

$$R_{xx}(\tau) = \mathcal{E}\{x(t)x(t+\tau)\} \tag{14.25}$$

is its autocorrelation function, where $\mathcal{E}\{\bullet\}$ denotes expectation. The PSD is given by the Fourier transform of the autocorrelation function [34]

$$S_{xx}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} R_{xx}(\tau)e^{-j\omega\tau}d\tau \tag{14.26}$$

These are the second-order functions.

While extremely useful for process analysis, the second-order functions have some limitations. For example, phase is not preserved in the second-order functions. High-order statistical functions can be used to address some of these shortcomings. We will concentrate here on the third-order functions, and in particular, the bispectrum. The key characteristic of the bispectrum is that for any i.i.d. stochastic process with a symmetric PSD, the HOS calculations reduce the magnitude of the bispectrum. Furthermore, as mentioned, the bispectrum of AWGN is identically zero. Therefore, since many noise phenomena have symmetrical PSDs, much of the corrupting noise can be removed or reduced from communication signals with such signal processing. While all spectra higher than the second for Gaussian signals are identically zero, such is not true for signals with other PSDs.

Consider a real discrete process $X(k)$ that is i.i.d., third-order stationary, and has zero mean. Let

$$R(m,n) \triangleq \mathcal{E}\{X(k)X(k+m)X(k+n)\} \tag{14.27}$$

denote its third moment sequence. The third moments have the following symmetry properties:

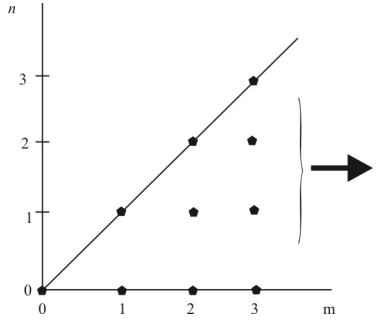


Figure 14.31 Region of support for the third order moment sequence.

$$R(m,n) = R(n,m) = R(-n,m-n) = R(n-m,-m) \tag{14.28}$$

As a consequence, knowing the third moments on the infinite triangle bounded by the lines $m = 0$ and $m = n; m,n \geq 0,$ enables us to find the entire third moment sequence (see Figure 14.31).

The bispectrum of the process $X(k)$, denoted by $B(\omega_1,\omega_2),$ is defined as the Fourier transform of its third moment sequence

$$B(\omega_1,\omega_2) = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} R(m,n)\exp[-j(m\omega_1 + n\omega_2)], \qquad |\omega_1|,|\omega_2| \leq \pi \tag{14.29}$$

In general, $B(\omega_1,\omega_2)$ is complex and a sufficient condition for its existence is that $R(m, n)$ is absolutely summable. Using the properties of $R(m, n)$ in (14.28), the following symmetry properties can be derived for the bispectrum:

$$B(\omega_1,\omega_2) = B(\omega_2,\omega_1) = B^*(-\omega_2,\omega_1) = B(-\omega_1-\omega_2,\omega_2) = B(\omega_1,-\omega_1-\omega_2) \tag{14.30}$$

Also $B(\omega_1,\omega_2)$ is periodic in ω_1 and ω_2 with period 2π. Thus knowledge of the bispectrum in the triangular region $\omega_2 \geq 0, \omega_1 \geq \omega_2, \omega_1 + \omega_2 \leq \pi$ (see Figure 14.32) is enough for its complete description.

It should be noted that some of the target signals of interest also have symmetric PSDs. In particular, AM DSB, FSK, and PSK signals all have symmetric PSDs. *Single sideband* (SSB) signals, however, by definition, have asymmetric PSDs. Therefore, the bispectrum for SSB signals is not suppressed. SSB signals are the most prolific signals in the HF range.

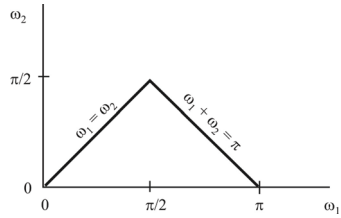


Figure 14.32 Region of support for the bispectrum.

Since it is originated from a large number of impulsive sources (lightning strikes), atmospheric noise is frequently modeled as having a Gaussian PSD. Man-made noise, while not originating from lightning strikes, is still typically impulse-like, originating, for example, from arc welders. Man-made noise thus contains broadband noise components. In both cases, large amplitude noise events frequently occur that are not adequately modeled as Gaussian. Such processes can be modeled as alpha-stable processes. A process is symmetric alpha stable (SαS) if its PSD is symmetric and it is alpha-stable. SαS processes by definition have symmetric PSDs. The spectra for a few standard SαS processes are shown in Figure 14.33 [33]. The tails of alpha-stable processes are heavier than Gaussian (a Gaussian process is an SαS process with $\alpha = 2$) as shown in Figure 14.34 [33]. Heavier tails in the PSD allow for higher probabilities of noise with large amplitudes. This is the advantage of modeling atmospheric and man-made noise as SαS processes. All of the noise sources discussed herein exhibit symmetric PSDs and are therefore suppressed with HOS processing keeping in mind that for Gaussian sources, the bispectrum is identically zero.

14.8.2 Signals in the HF Band

Both analog and digital signals can be found in the HF frequency range. For spectral efficiency reasons, the predominant analog signals are SSB, where one of

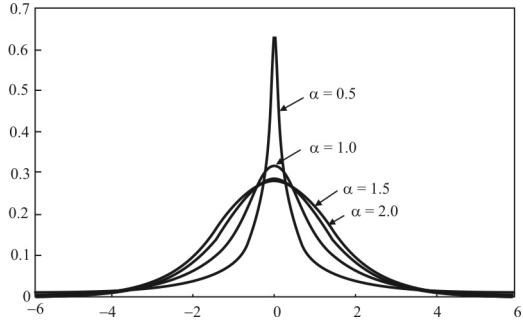


Figure 14.33 Power spectral densities of some alpha stable processes.

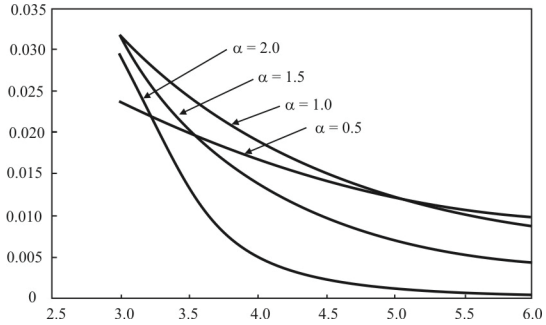


Figure 14.34 Details of the tails of a few common alpha stable processes.

the redundant sidebands in regular amplitude modulation (AM) has been removed (or suppressed). A typical PSD for USB is shown in Figure 14.35. Removal of the sideband causes the PSD of SSB signals to be asymmetric. Regular AM, to include AM with a suppressed carrier, has symmetric PSDs—these signals are rarely seen, however, outside of the radio amateur ranks. Two independent SSB signals can be combined on the same carrier, one upper sideband and the other lower sideband. This effectively doubles the throughput of the AM signal while requiring allocation of two adjacent channels. (NATO countries, by agreement, use 3 kHz channels in the HF band.) This is known as *independent sideband* (ISB) modulation and it possesses an asymmetric PSD. (See Figure 14.36).

Therefore the two predominant analog modulations found in the HF band, SSB and ISB have asymmetric psds and are not suppressed with HOS processing.

The predominant digital modulations used in the HF band by the military are phase shift key (PSK), quadrature amplitude modulation (QAM), orthogonal frequency division multiplexing (OFDM), and frequency shift key (FSK). Data rates are moderate at best, typically being 9,600 baud or less for reliable communications. The PSDs for these data signals are symmetric as illustrated in Figure 14.37. Therefore, the HOS processing would be expected to be marginally effective against these data signals.

14.8.3 Noise Sources in the HF Spectrum

Noise sources can conveniently be divided into two categories: (1) those external to the system of interest and (2) internal sources. Note that these noise sources are additive but not necessarily AWGN. (Thermal noise can be modeled as AWGN but other sources need not be.)

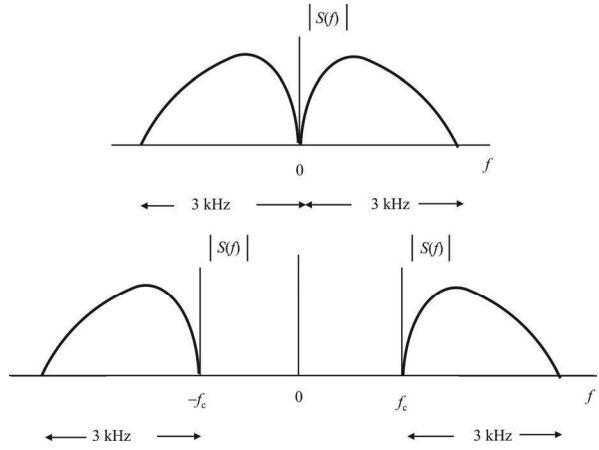


Figure 14.35 Upper sideband PSD.

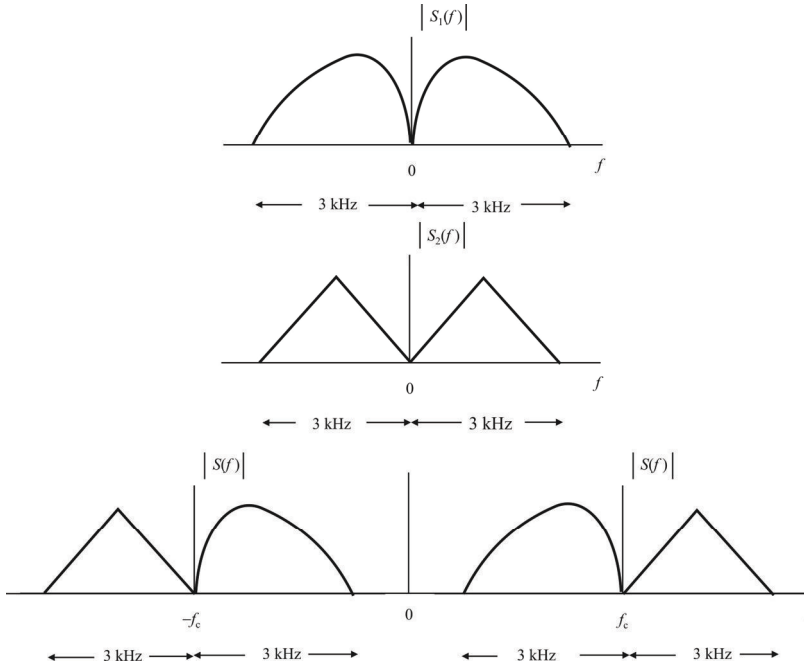


Figure 14.36 Independent sideband PSD.

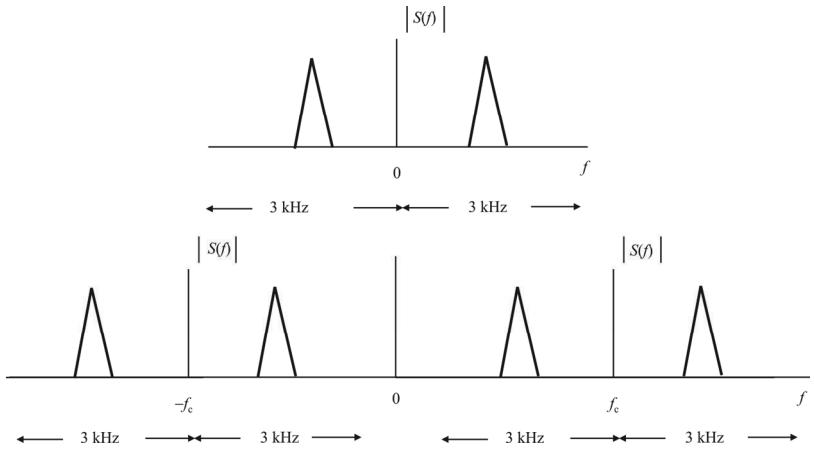


Figure 14.37 Binary data signals PSD.

Suppose the input to the antenna consists of a target signal denoted by $x(t)$ and a single noise source, denoted by $n(t)$ that are statistically independent of each other. Denote the third moment of $x(t)$ by $R_x(m, n)$ and that of the noise by $R_n(n, m)$. Since the noise is additive the input can be expressed as

$$X(t) = x(t) + n(t) \tag{14.31}$$

Then from (14.27) the third-order moment function is

$$\begin{aligned} R(m,n) &= E\{[x(t)+n(t)][x(t+m)+n(t+m)][x(t+n)+n(t+n)]\} \\ &= \mathcal{E}\{x(t)x(t+m)x(t+n)+x(t)x(t+m)n(t+n)+\ldots+n(t)n(t+m)n(t+n)\} \\ &= \mathcal{E}\{x(t)x(t+m)x(t+n)\}+\mathcal{E}\{x(t)x(t+m)n(t+n)\}+ \\ &\qquad \ldots+\mathcal{E}\{n(t)n(t+m)n(t+n)\} \end{aligned} \tag{14.32}$$

Because the noise is assumed to be independent of the signal, all terms in (14.32) except the first and last are zero, because they are third order moments of statistically independent functions. Therefore,

$$R(m,n) = R_x(m,n) + R_n(m,n) \tag{14.33}$$

Palpably, this result extends to any number of independent noise sources at the input.

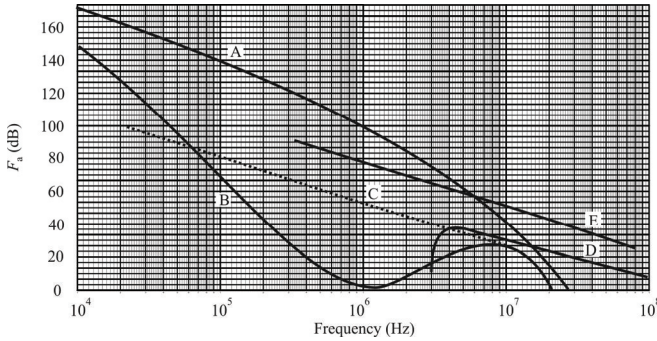


Figure 14.38 F_a for low VHF; low RF noise; A: atmospheric noise, value exceeded 0.5% of the time; B: atmospheric noise, value exceeded 99.5% of the time; C: man-made noise, quiet receiving site; D: galactic noise; and E: median business area man-made noise; minimum noise level expected. (Source: [36] © Artech House 2008. Reprinted with permission.)

14.8.3.1 External Noise Sources

External noise typically dominates the internal noise sources in the HF range. The most important external noise sources in this range are atmospheric, thermal, man-made, and galactic noises. The levels of the predominant sources of external noise in this frequency range are illustrated in Figure 14.38 [35]. F_a refers to the level of the noise source above ambient thermal noise, expressed in decibels, given by

$$F_a = 10\log\frac{P_n}{k_B TW} \tag{14.34}$$

where P_n is the noise power, k_B is Boltzmann's constant (= 1.38×10^{-23}), T is the temperature in Kelvins, and W is the noise bandwidth. F_a expresses the additional noise added by the source being considered. When the atmospheric noise starts to tail off in importance, around 8 MHz or so, man-made noise starts to dominate. It remains the dominant noise source up to about 1 GHz.

To the degree that the external noise can be reduced by the HOS processing, the internal noise sources become more important and may become the dominant noise sources to deal with once the external noise has been suppressed. Fortunately, all of the internal noise sources produce noise waveforms that can be adequately modeled as Gaussian, so they too should be reduced by the HOS processing.

Atmospheric noise is caused predominantly by lightning strikes. Because lightning strikes are very powerful and they radiate primarily in the HF range, and

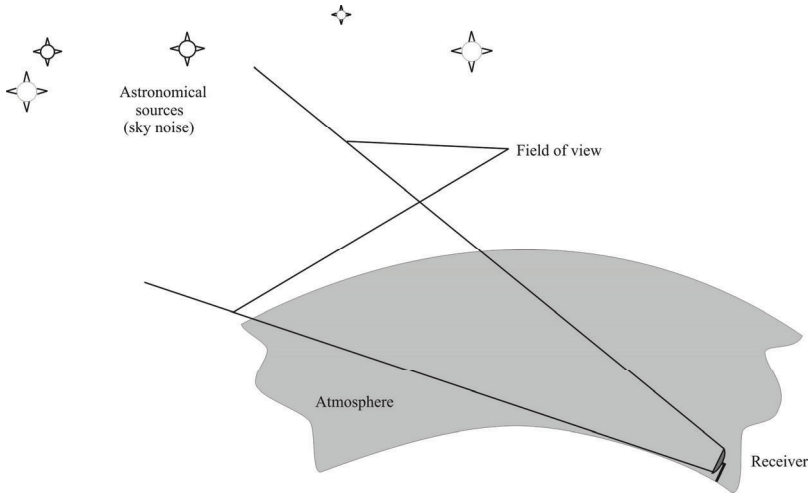


Figure 14.39 Sky noise and atmospheric noise.

signals in the HF range can propagate for considerable distances, atmospheric noise levels can be quite large. As mentioned, the statistics of atmospheric noise can be modeled as SαS processes.

The thermal radiation from a black body is simply "thermal noise" going under another name. (The sky can be modeled as a black body with intermittent bright spots due to stars and planets.) The received power in a small bandwidth, W, is

$$P_{\mathrm{n}} = 2k_{\mathrm{B}}TW \tag{14.35}$$

which increases proportionally to the temperature and the bandwidth. A receiver picking up signals in the real world will also usually pick up *background noise* from objects in its field of view. This process can be summarized as illustrated in Figure 14.39; this type of noise is sometimes referred to as sky noise.

The receiving antenna will see some thermal radiation from the atmosphere in its field of view. Unless the atmosphere is very opaque it will also collect some noise from astronomical sources in the field of view. (Even if the atmosphere were perfectly transparent and there were no bright stars and galaxies in the field of view, the receiver would still pick up the 2.7 K *cosmic background* radiation.) The total noise power reaching the receiver can be represented in terms of the noise temperature contributions from the atmosphere and the astronomical "sky."

Galactic noise has its origins in deep space. On the Earth the largest sources of such noise are the Sun and the Milky Way. In general, the amount of galactic noise received by an antenna depends on the pointing direction of the antenna. For an omnidirectional antenna of course this does not apply. Vertical monopoles and dipoles have a null at 0^o zenith, so the amount of galactic noise collected by them is strongly affected by the zenith angle.

By far the most important source of noise in the higher mobile communication bands is that radiated by electrical equipment of various kinds and is often termed *man-made noise* (MMN). MMN is usually impulsive in nature, which means that its characteristics are fundamentally different from random noise of a thermal origin. Impulsive noise is characterized by bursts of very short duration, which may have random amplitude and random time of occurrence. The impulsive nature of MMN produces a PSD that is broadband. Impulsive noise typically comes from electrical equipment where sparks occur, such as automotive ignitions. Automotive ignitions are not the only sources of MMN, however, albeit in urban areas where there is considerable road traffic it is one of the major ones. MMN, arising due to a variety of emissions from electrical discharges and other sources, may set a higher background limit than natural and internal noise.

14.8.3.2 Internal Noise Sources

The components and circuits that make up any electronic system, to include EW systems, produce noise. In particular they produce thermal noise, flicker noise, shot noise, and recombination noise.

Shot noise is due to the discrete quantum nature of electron flow through a potential barrier. Shot noise consists of additional current fluctuations that occur when a voltage is applied to a conductor and a macroscopic current starts to flow. It is most often associated with diodes and bipolar junction transistors (BJTs). Shot noise consists of random fluctuations of the electric current in an electrical conductor, which are caused by the fact that the current is carried by discrete charges (electrons). The strength of this noise increases for growing magnitude of the average current flowing through the conductor. The PSD for shot noise is illustrated in Figure 14.40. Note the wide frequency extent.

Flicker noise occurs in all active devices and depends on the DC bias current. The inverse dependence on frequency gives this term it's most common name: $1/f$ noise. In semiconductor devices, the origins of $1/f$ noise are attributed to the effects of contaminants and defects in the crystal structure. In MOS structures, $1/f$ noise is associated with oxide surface states that periodically trap and release carriers. Over the decades, advances in semiconductor processes and fabrication practices have reduced flicker noise due to these imperfections.

The PSD for flicker noise is illustrated in Figure 14.41. Flicker noise is limited to a relatively narrow frequency band around the carrier frequency and

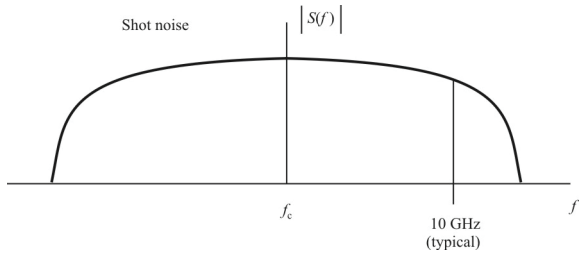


Figure 14.40 Shot noise PSD.

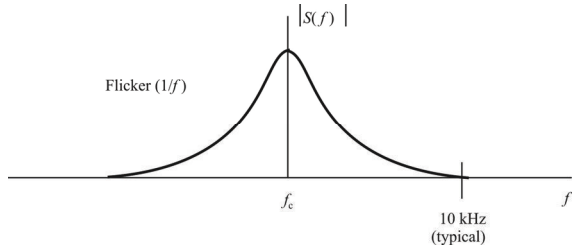


Figure 14.41 Flicker noise PSD.

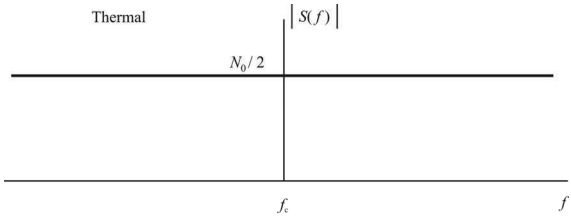


Figure 14.42 Thermal noise PSD.

falls proportionally to 1/*f*. It is, however, important to narrowband, linear EW systems such as those of interest here

Thermal or Johnson noise is due to random carrier motion within a device, producing a broadband noise signal. This term is called thermal noise because the carrier motion is thermally excited. Johnson noise has a Gaussian amplitude distribution in the time domain and is evenly distributed across the spectrum. The spectral breadth and its sources' ubiquity lead to thermal noise's domination over other noise types in many applications. The PSD of thermal noise is illustrated in Figure 14.42. It is flat, with equal levels at all frequencies of practical interest for the most part.

14.8.4 Summary

This section describes an active antenna configuration that suppresses noise in the HF frequency range. The two primary technologies used to implement the active antenna are active antennas and HOS signal processing. The former of these allows for efficient but short antennas for the HF range while the latter suppresses much of the noise present in the HF range. This noise can be externally generated as well as internal to the amplifier stage and antenna. The block diagram of the resulting configuration is shown in Figure 14.43.

It should be noted that although an antenna specifically for the HF range was discussed in this section, because that is where the problems with antenna size are particularly acute, the approach discussed here can be extended to other frequency ranges as long as the frequency of operation is within the state of the art of ADCs. At the time of this writing that would probably require frequency conversion to an appropriate IF frequency for the VHF and higher-frequency ranges.

14.9 Concluding Remarks

The fundamental characteristics of electrically small antennas for EW applications were covered in this chapter. It is a continuous struggle adapting antennas for EW

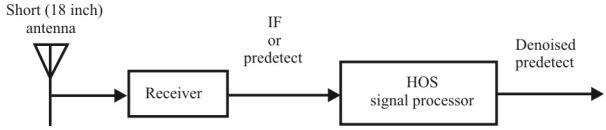


Figure 14.43 Noise suppressing antenna architecture.

systems, principally because there is usually limited space for antennas on top of masts upon which the antennas are mounted in the ground-based applications and on the body of UAS systems. Large aircraft have the luxury of not having so much of this constraint; however, it is still desirable to have antennas that are as small as possible for these systems as well. This is particularly true for tactical applications where usually there can be only one or at most two antennas on the mast. Therefore, keeping the antennas as small as possible is important.

On the other hand, the performance of an antenna (especially gain) is typically inversely proportional to its size.

We covered a few new ideas for small antennas where performance is not necessarily compromised when the antennas are shrunk. The double-negative material-loaded antenna is an example of this. So is the notion of non-Foster antennas where negative impedance converters are used to cancel the frequency dependent-behavior of small antennas. The biggest problem with the former technology is that the materials must be fabricated—they appear to not be available from nature. The problem with the latter is non-Foster circuits can easily become unstable.

References

[1] Tretykov, S. A., S. I. Maslovski, A. A. Sochhava, and C. R. Simovski, "The Influence of Complex Material Coverings on the Quality Factor of Simple Radiating Systems," *IEEE Transactions on Antennas and Propagation*, Vol. 53, No. 3, March 2005, pp. 965–970.

[2] Fujimoto, K., A. Henderson, K. Hirasawa, and J. R. James, *Small Antennas,* New York: Wiley, 1987.

[3] Volakis J. L., C.-C. Chen, M. Lee, B. Lamer and D. Psychoudakis, "Miniaturization Methods for Narrowband and Ultra Wideband Antennas," *Proceedings IEEE International Workshop on Antenna Technology*, 2005.

[4] Stutzman, W. L., and G. A. Thiele, *Antenna Theory and Design*, New York: Wiley, 1981, p. 203.

[5] Wheeler, H. A., "Small Antennas," *IEEE Transactions on Antennas and Propagation,* Vol. AP-23, No. 4, July 1975, pp. 462–469.

[6] Huang, L., W. L. Schroeder, and P. Russer, "Estimation of Maximum Attainable Antenna Bandwidth in Electrically Small Mobile Terminals," *Proceedings of the 36th European Microwave Conference,* Manchester U.K., September 2006, pp. 630–636.

[7] Yaghjian, A. D., and S. R. Best, "Impedance, Bandwidth, and Q of Antennas," *Digest of the IEEE AP-S International Symposium, Vol. 1,* Columbus OH, June 2003, pp. 501–506.

[8] Hansen, R. C., "Fundamental Limitations in Antennas," *Proceedings of the IEEE,* Vol. 69, No. 2, February 1981, pp. 170–182.

[9] Thiele, G. A., P. L. Detweiler, and R. P. Penno, "On the Lower Bound of the Radiation Q for Electrically Small Antennas*," IEEE Transactions on Antennas and Propagation,* Vol. 51, No. 6, June 2003.

[10] Fante, R. L., "Quality Factor of General Ideal Antennas," *IEEE Transactions on Antennas and Propagation,* Vol. Ap-17, No. 2, March 1969, pp.151–154.

[11] Sten, J. C.-E, A. Hujanen, and P. K. Koivisto, "Quality Factor of an Electrically Small Antenna Radiating Close to a Conducting Plane," *IEEE Transactions on Antennas and Propagation,* Vol. 49, No. 5, May 2001, pp. 829–837.

[12] Geyi, W., "Physical Limitations of Antenna*," IEEE Transactions on Antennas and Propagation,*" Vol. 51, No. 8, August 2003, pp. 2116–2121.

[13] Wheeler, H. A., "Fundamental Limits of Small Antennas," *Proceedings of the I.R.E.* (*IEEE*), December 1947, pp. 1479–1481.

[14] Wheeler, H. A., "The Radiansphere Around a Small Antenna," *Proceedings of the IRE,* August 1959, pp. 1325–1331.

[15] Stutzman, W. L., and G. A. Thiele, *Antenna Theory and Design,* New York: Wiley, 1981, p. 1014.

[16] Newman, E. H., P. Bohley, and C. H. Walter, "Two Methods for the Measurement of Antenna Efficiency," *IEEE Transactions on Antennas and Propagation,* Vol. AP-23, July 1976, pp. 457–461.

[17] Collin, R. E., and S. Rothschild, "Evaluation of Antenna Q," *IEEE Transactions on Antennas and Propagation,* January 1964, pp. 23–27.

[18] McLean, J. S., "A Re-Examination of the Fundamental Limits on the Radiation Q of Electrically Small Antennas," *IEEE Transactions on Antennas and Propagation,* Vol. 44, No. 5, May 1996, pp. 672–676.

[19] Hujanen, A., J. Holmberg, and J. C.-E., Sten, "Bandwidth Limitations of Impedance Matched Ideal Dipoles," *IEEE Transactions on Antennas and Propagation,* Vol. 53, No. 10, October 2005, pp. 3236–3239.

[20] Geyi, W., "A Method for the Evaluation of Small Antenna Q," *IEEE Transactions on Antennas and Propagation,* Vol. 51, No. 51, August 2003, pp. 2154–2159.

[21] Yaghjian, A. D., and S. R. Best, "Impedance, Bandwidth, and Q of Antenna," *IEEE Transactions on Antennas and Propagation*, Vol. 53, No. 4, April 2005, Vol. 53, No. 4, April 2005, pp. 1298–1324.

[22] Ziolkowski, R. W., and A. D. Kipple, "Application of Double Negative Materials to Increase the Power Radiated by Electrically Small Antennas," *IEEE Transactions on Antennas and Propagation,* Vol. 51, No. 10, October 2003, pp. 2626–2540.

[23] *Popular Electronics*, July 1989.

[24] Nordholt, E. H., and D. Van Willigen, "A New Approach to Active Antenna Design," *IEEE Transactions on Antennas and Propagation,*" Vol. AP-28, No. 6, November 1980, pp. 904–910.

[25] Sussman-Fort, S. E., "Matching Network Design Using Non-Foster Impedances," *International Journal of RF and Microwave Computer-Aided Engineering*, March 2006, www.ieee.li/pdf/ viewgraphs_matching_network_ design_non_foster_impedances.pdf.

[26] Rudish, R., and S. E. Sussman-Fort, "Non-Foster Impedance Matching Improves S/N of Wideband Electrically-Small VHF Antennas and Arrays," *Proceedings of the 2005 IASTED International Conference on Antennas, Radar, and Wave Propagation* (*ARP2005*), Banff, Canada, 2005.

[27] Rudish, R., and S. E. Sussman-Fort, "Progress in Use of Non-Foster Impedances to Match Electrically-Small Antennas and Arrays," *2005 Antenna Applications Symposium,* Allerton Park, IL, September 21–23, 2005.

[28] Aberle, J. T., Two-Port Representation of an Antenna with Application to Non-Foster Matching Networks," *IEEE Transactions on Antennas and Propagation*, Vol. 56, No. 5, May 2008, pp. 1218–1222..

[29] Pozar, D. M., *Microwave Engineering*, 3rd ed., New York: Wiley, 2005, pp. 261–264.

[30] Linvill, J. G., "Transistor Negative Impedance Converters," *Proceedings of the IRE,* Vol. 41, June 1953, pp. 725–729.

[31] Sussman-Fort, S. E., and R. M. Rudish, "Non-Foster Impedance Matching of Electrically Small Antennas," *IEEE Transactions on Antennas and Propagation*, Vol. 57, No. 8, August 2009, pp. 2230–2241.

[32] Sussman-Fort, S. E., and R. M. Rudish, "Non-Foster Impedance Matching for Transmit Applications," *Proceedings 2006 International Workshop on Antenna Technology Small Antennas and Novel Materials,* March 6–8, 2006, pp. 53–54.

[33] Shao, M., and C. L. Nikias, "Signal Processing with Fractional Lower Order Moments: Stable Processes and Their Applications," *Proceedings of the IEEE*, Vol. 81, No. 7, July 1993, pp. 986–1010.

[34] Kay, S. M. *Modern Spectral Estimation: Theory and Applications*, Upper Saddle River, NJ: Prentice-Hall, 1988.

[35] Poisel, R. A., *Introduction to Communications Electronic Warfare Systems*, 2nd ed., Norwood, MA: Artech House, 2008, p.109.

Chapter 15

Patch Antennas

15.1 Introduction

Microstrip patch antennas (MPAs) are a popular type of antenna for the higher-frequency ranges. Their size essentially precludes their use below the UHF range. Although they can become quite complicated, in its simplest form, MPAs consist of a radiating patch on one side of a dielectric substrate that has a ground plane on the other side as shown in Figure 15.1 [1]. The patch is generally made of conducting material such as copper or gold and can take many possible shapes. The radiating patch and the feed lines are usually photo etched on the dielectric substrate [1–6]. We introduce patch antennas and their characteristics in this chapter. An overview of the more important electrical parameters of patch antennas is given in Table 15.1.

The structure of this chapter is as follows. We begin with a discussion of the advantages and disadvantages of patch antennas; probably the worst of the latter is its bandwidth limitations. We then discuss the physical characteristics of these antennas. That is followed by an analysis of their electrical properties. Radiation patterns are discussed next. Then we include a section on feed techniques. After that we present the two dominant electrical models of patch antennas. The last topic covered is a patch antenna design process.

15.2 Advantages and Disadvantages

In this section we present some of the advantages and disadvantages of patch antennas, with particular emphasis on their use in EW and other sensor-related applications.

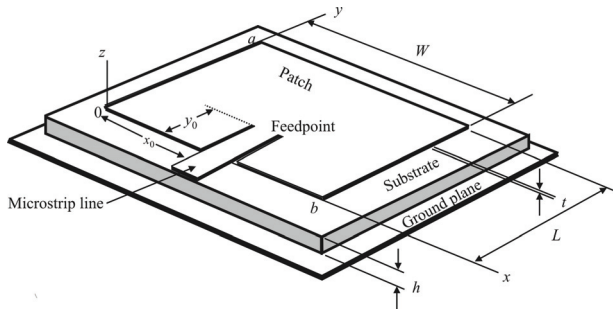


Figure 15.1 Microstrip feed.

Table 15.1 Overview of Parameters of Patch Antennas

Parameter	Classical Antenna	Aperture-Fed Antenna	Note
Impedance Bandwidth	2 to 3 %	4 to 6 %	In % of resonant frequency
Radiation Efficiency	80 %	80 %	Surface-wave losses only
Total Efficiency	40 to 60 %		Considering losses in dielectrics
Cross-Polarization	35 to 40 dB		
Patch Coupling	Indispensable	small	
Feeder Radiation	Middle (*)	small	(*) Reducible by feeding topology

Source: [1].

15.2.1 Advantages of Patch Antennas

MPAs are particularly useful in the higher-frequency ranges due to their low-profile structure. They are extremely compatible as antennas embedded in handheld wireless communication devices such as cellular phones, pagers, and PDAs. The telemetry and communication antennas on missiles, for example, need to be thin and conformal and are often MPAs. Another area where they have been used successfully is in satellite communication (GPS is an example).

Many of these applications are shared with EW applications. Perhaps the most obvious is as conformal antennas embedded in the body of EW UAS systems. This could be either in the fuselage or wings or both. They also can serve as directional EA or ES antennas for distributed sensor fields.

15.2.1.1 Volume

The footprint of microwave patch antennas is small—the patch length is on the order of half a wavelength (at 1 GHz, λ~30 cm). In addition, their overall volume is quite small, with heights on the order of single digit millimeters.

15.2.1.2 Weight

Since these antennas are small, they are also inherently lightweight. This makes them ideal as antennas for aircraft, missiles, or man-packed systems. Hand-emplaced sensor fields also can take advantage of their lightweight properties.

15.2.1.3 Planar Configurations

Patch antennas can easily be formed into curved surfaces, both uniform and irregular. This makes mechanical integration into irregular host platforms economical.

15.2.1.4 Fabrication Cost

The fabrication of patch antennas uses the same technology used to manufacture printed circuit boards, so manufacture is easy and therefore inexpensive. Thus patch antennas can be manufactured in large quantities, which further brings their cost down.

15.2.1.5 Polarization

Linear, circular, or elliptical polarization can be easily accommodated. In fact, as will be discussed in Chapter 16, cross-polarization where two orthogonal linear polarizations are transmitted simultaneously can be accommodated.

15.2.1.6 Integration

MPAs can be easily integrated with MMIC circuits, and when system requirements allow it, the feeds can be manufactured into these circuits as well. We will discuss feed lines at length later in this chapter.

15.2.1.7 Multiple Operating Frequencies

Patch antennas can be manufactured to operate at multiple (~2 or 3) frequencies.

15.2.1.8 Mechanical Characteristics

Because all the components in a patch antenna are securely attached to each other, when attached to adequately rigid surfaces, patch antennas are rugged. This makes them particularly appropriate for military applications

15.2.1.9 Cross-Polarization

Patch antennas exhibit very good cross-polarization. This parameter is not important for antennas using linearly polarized waves. For dual antennas designed to operate with cross polarization, normally significant polarization isolation is required and this parameter becomes important.

15.2.2 Disadvantages of Patch Antennas

MPAs suffer from a number of disadvantages when compared to conventional antennas. Some of these are discussed here.

15.2.2.1 Operating Bandwidth

Their operating bandwidth is severely limited, with typical values 2–5% of the operating frequency. This narrow bandwidth (high Q) is the principal disadvantage.

Because of this, the design of microstrip antenna arrays exhibiting sufficiently low sidelobe levels is difficult.

MPAs have a very high antenna quality factor (Q). Q represents the losses associated with the antenna and a large Q indicates narrow bandwidth and low efficiency.

One way of reducing the Q is by increasing the thickness of the dielectric substrate. However, increasing a fraction of the total power delivered by the source goes into a surface wave as the thickness increases. This surface wave contribution can be considered as an unwanted power loss since it is ultimately scattered at the dielectric bends and causes degradation of the antenna characteristics.

15.2.2.2 Radiation Efficiency

Ignoring ohmic losses, radiation efficiency is given by the ratio of the radiated power and real power on the input port of the antenna element. Energy, which is not radiated by the antenna, is removed in the form of surface waves along the infinite dielectric substrate (in real situations, the substrate is limited, of course).

From a practical point of view, the total efficiency is important. In addition to dielectric losses, surface-wave losses contribute to its value. Lossy dielectrics can be viewed as lossy transmission lines. Because the substrate has limited dimensions, standing waves are formed in the substrate. Higher standing-wave ratios are caused by higher losses.

The total efficiency of patch antennas is about 1 to 2 dB worse than the efficiency of reflector antennas.

15.2.2.3 Gain

The gain of patch antennas is typically quite low compared to conventional antennas. Low gain can be overcome by using an array of antennas, however. Because of their low cost due to a potentially large manufacturing base, arrays of patch antennas are extremely feasible.

15.2.2.4 Extraneous Radiation

There is unwanted radiation from the feed and junctions that distorts the antenna pattern and increases losses. This *parasitic radiation* of the feeding microstrip network can deform the directivity pattern [7] .

15.2.2.5 Power Handling Capacity

The small size dictates that patch antennas have limited power handling capacity. Low power handling capacity can be at least partially overcome by using an array configuration for the elements, discussed in Chapter 16.

15.3 Patch Antenna Physical Characteristics

Development of microstrip antenna technology has been driven primarily by system requirements for antennas with low profiles, low weights, low costs, easy integratability into arrays or with microwave integrated circuits, or polarization diversity.

Thanks to their low profile, among many other possibilities, microstrip antennas can be mounted to the walls or ceilings of buildings and to the fuselage or wings of aircraft, manned or unmanned. They can be used singly, as in GPS (or other satellite system) receive antennas, or they can be used to form arrays. (More will be said about microstrip patch arrays in Chapter 16.) Finally, depending on system requirements, microstrip antennas can be simply integrated directly into microwave circuits, which are based on microstrip technology, and results in no special transmission lines for coupling.

The first patch elements were fed with either a coaxial line through the bottom of the substrate or by a coplanar microstrip line. This latter type of excitation allows feed networks and other circuitry to be fabricated on the same substrate as the antenna element. The microstrip antenna radiates a relatively broad beam broadside to the plane of the substrate.

To facilitate analysis and design performance prediction, the patch is generally a regular structure such as a square, rectangle, circle, triangle, ellipse, or some other common shape as shown in Figure 15.2 [2]. For a rectangular patch, the length L of the patch is usually $0.3333\lambda_0 < L < 0.5\lambda_0$, where λ_0 is the wavelength in a vacuum. The patch is very thin, such that $t << \lambda_0$ while the height h of the dielectric substrate is usually $0.003\lambda_0 \leq h \leq 0.05\lambda_0$. The dielectric constant of the substrate (ε_r) is typically in the range $2.2 \leq \varepsilon_r \leq 15$ [2].

The primary materials used for the radiating element in microwave patch antennas are listed in Table 15.2 along with a few of the important performance parameters, the conductivity (σ), and relative permittivity (ε_r). Likewise, a few of the materials used as substrate material are listed in Table 15.3 along with their

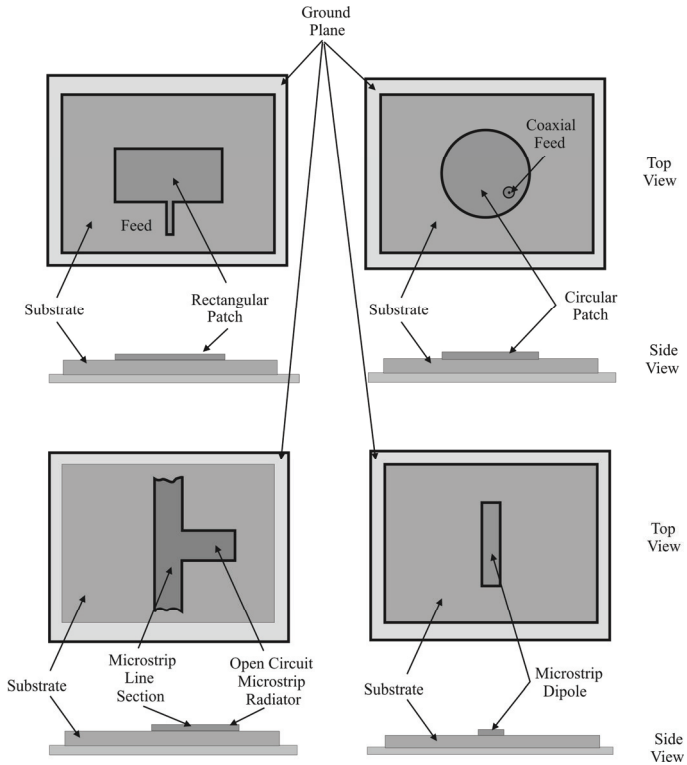


Figure 15.2 Patch antennas.

relative permittivity and loss tangent (tan δ).[1] Some of the other important parameters associated with other common dielectric materials are listed in Table 15.4.

15.4 Patch Antenna Electrical Characteristics

While there are several important parameters associated with a patch antenna, the three essential ones are:

- Dielectric constant of the substrate, ε_r;
- Height of the dielectric substrate, h;
- Frequency of operation, f_0.

15.4.1 Frequency

As mentioned, the operating frequency of patch antennas is restricted to UHF and above because lower than this yields antennas that are quite large and much more fragile than their conventional counterparts.

15.4.2 Substrate Dielectric Constant

The dielectric constant of the substrate determines, to a large extent, the operating characteristics of the antenna. We say much more about the dielectric constant later.

5.4.3 Dielectric Height

For the same reasons that make the dielectric constant of the substrate a principal parameter for determining the performance of patch antennas, the amount of material placed in the substrate, determined primarily by its height once other physical parameters such as the length and width are determined, is an important parameter.

---

[1] Power is lost in all dissipative systems, such as dielectric materials, usually, but not always, in the form of heat. The dissipation factor is a measure of the power loss and is also referred to as the loss tangent. The expression for the loss tangent is

$$\tan \delta = 2\sigma / \varepsilon f_0$$

where $\varepsilon = \varepsilon_r \varepsilon_0$ is the permittivity and σ is the conductivity of the material.

Table 15.2 Patch Antenna Material

Material	Conductivity (σ) S m^{-1}	Relative Permittivity (ϵ_r)
Aluminum	15.5×10^7	1
Brass	2.6×10^7	1
Copper	5.8×10^7	1
Gold	4.8×10^7	1

Source: [8].

Table 15.3 Patch Substrate Material

Material	Relative Permittivity (ε_r)*	Loss Tangent (tan δ)*
Alumina	~9–10	0.0003
FR-4	~4.5	0.025
Silicon	~15.9	0.004
Teflon	~2.1	0.0004

*Relative permittivity and loss tangent are frequency-dependent. The values given are typical.
Source: [8].

Table 15.4 Non-Electrical Properties of Commonly Used Substrate Materials for MPAs

Property	PTFE	Fused Quartz	Alumina	Sapphire	GaAs
Temperature Range (°C)	–55 to 260	<+1,100	<1,600	–24 to 370	–55-260
Thermal Conductivity (W/cm K)	0.0026	0.017	0.35 to 0.37	0.42	0.46
Coefficient of Thermal Expansion (ppm K^{-1})	16.0 to 108.0	0.55	6.30 to 6.40	6.00	5.70
Temperature Coefficient of Dielectric Constant (ppm K^{-1})	–350 to +480	+13.0	+136.0	+110 to +140	–
Minimum Thickness (mil)	4	2	5	4	4
Machinability	Good	Very Poor	Very Poor	Poor	Poor
Solderability	Good	Good	Good	Good	Good
Dimensional Stability	Poor for unreinforced, very good for others	Good	Excellent	Good	Good
Estimated Cost ($US cm^{-2})	1.1	560	36	–	3,000

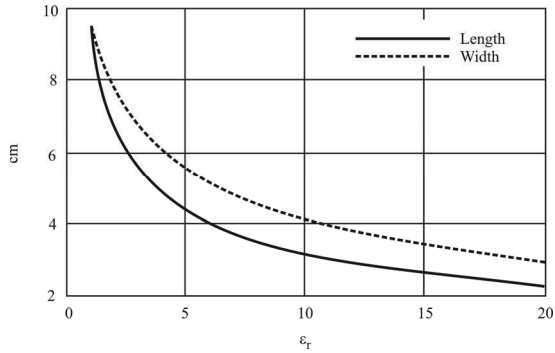


Figure 15.3 Patch dimensions versus dielectric constant for a rectangular patch.

15.4.4 Patch Dimensions

Patch dimensions (L and W) as a function of the dielectric constant are indicated in Figure 15.3 [8]. Likewise, patch dimensions as a function of frequency are indicated in Figure 15.4 [8].

15.4.5 Patch Efficiency

Patch efficiency as a function of substrate height (h/λ) is plotted in Figure 15.5 [8] for a few representative values of dielectric constant. We can see that efficiency is

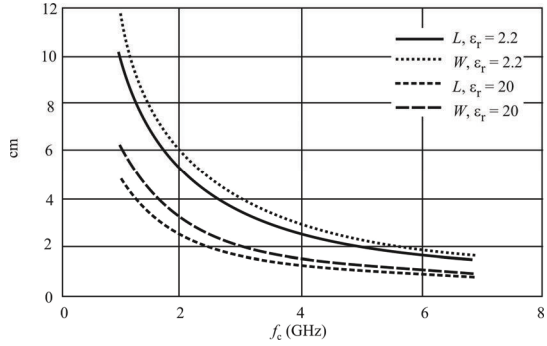


Figure 15.4 Patch dimensions versus frequency for a rectangular patch.

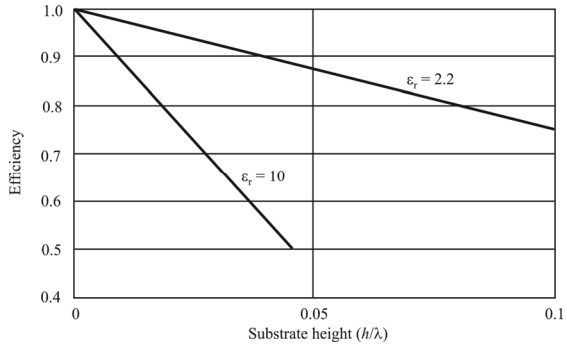


Figure 15.5 Patch efficiency versus height for a rectangular patch.

better for smaller dielectric constants for the range of substrate heights considered.

15.4.6 Patch Bandwidth

Patch bandwidth (as a percentage of design frequency) versus substrate height (h/λ) is plotted in Figure 15.6. We can see that the fractional bandwidth increases as the substrate height increases. This is counterbalanced with the decrease in efficiency as the height increases as illustrated in Figure 15.5.

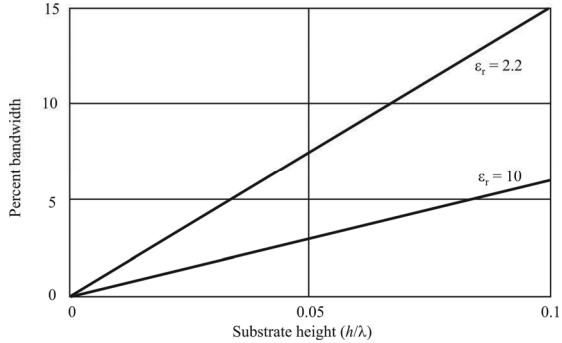


Figure 15.6 Patch bandwidth versus height for a rectangular patch.

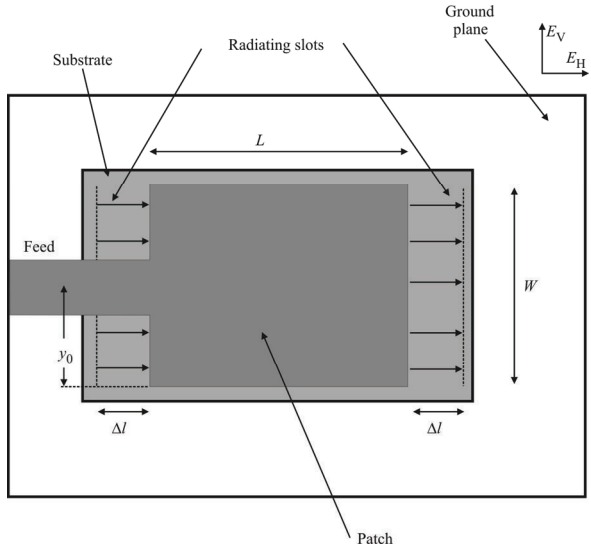


Figure 15.7 Patch top view.

15.5 Patch Antenna Analysis

Microstrip patch antennas radiate primarily because of the fringing fields between the patch edge and the ground plane, as illustrated in Figures 15.7 and 15.8. A thick dielectric substrate having a low dielectric constant is desirable since this provides better efficiency, larger bandwidth, and better radiation [4]. However, such a configuration results in larger antennas. A compromise must be reached between antenna dimensions and antenna performance because in a compact MPA, higher dielectric constants must be used, which are less efficient and result in narrower bandwidth.

The value of the effective dielectric constant, ε_{eff}, is slightly less than ε_r because the fields around the periphery of the patch are not confined in the dielectric substrate but are also spread in the air as shown in Figure 15.8. The expression for ε_{reff} is given by [9, 10]

$$\varepsilon_{eff} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left(1 + 12\frac{h}{W}\right)^{-1/2} \qquad W/h \geq 1 \qquad (15.1)$$

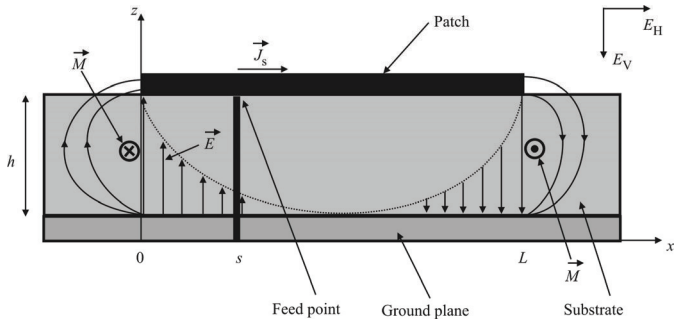


Figure 15.8 Patch side view.

where

ε_r is the dielectric constant of substrate.
h is the height of dielectric substrate.
W is the width of patch.

A rectangular microstrip patch antenna of length *L*, width *W* on a substrate of height *h* is shown in Figures 15.8. The coordinate axes are selected such that the length is along the *x* direction, width is along the *y*-direction, and the height is along the *z*-direction.

The current density varies in the *x*-direction as illustrated in Figure 15.9. We see that the current variation is along the *x*-axis, with no variation along the *y*-axis.

In order to operate in the fundamental TM_{10} mode, the length of the patch, *L*, must be slightly less than $\lambda/2$ where λ is the wavelength in the dielectric medium and is equal to $\lambda_0/\varepsilon_{eff}$ where λ_0 is the free-space wavelength. The TM_{10} mode implies that the field varies one $\lambda/2$ cycle along the length, and there is no variation along the width of the patch.

In Figure 15.7, the MPA patch antenna is represented by two slots, separated by a transmission line of length *L* and open circuited at both the ends. At both ends the voltage is maximum and current is minimum along the width of the patch due to the open circuits. The fields at the edges can be resolved into normal and tangential components with respect to the ground plane.

The normal components of the electric field at the two edges along the width are in opposite directions and therefore out of phase since the patch is $\lambda/2$ long and thus cancel each other in the broadside direction. The tangential components (seen in

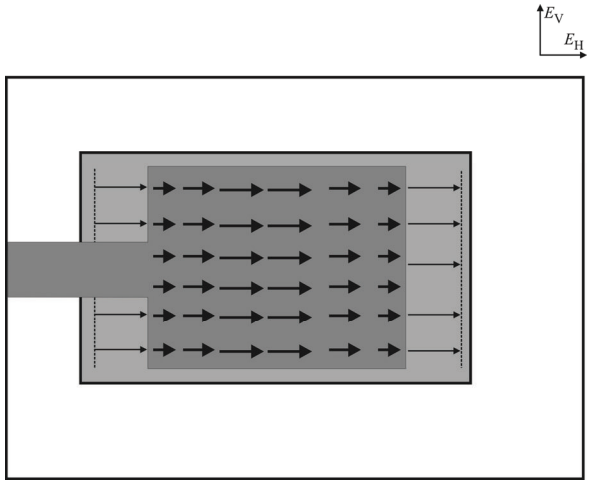


Figure 15.9 Patch current variation. The current density is shown as the dark arrows while the radiating fields are shown as the light arrows.

Figure 15.8), which are in phase, mean that the resulting fields combine to give maximum radiated field normal to the surface of the patch. Hence, the edges along the width can be represented as two radiating slots, which are λ/2 apart and excited in phase and radiating in the half space above the ground plane. The fringing fields along the width can be modeled as radiating slots and electrically the patch of the microstrip antenna looks greater than its physical dimensions. The dimensions of the patch along its length have now been extended on each end by a distance Δl, which is given by [10]

$$\Delta l = 0.412h \frac{(\varepsilon_{\text{eff}} + 0.3)\left(\frac{W}{h} + 0.264\right)}{(\varepsilon_{\text{eff}} - 0.258)\left(\frac{W}{h} + 0.8\right)} \tag{15.2}$$

The effective length of the patch L_{eff} now becomes:

$$L_{\text{eff}} = L + 2\Delta l \tag{15.3}$$

Given the resonance frequency f_0 , the effective length is given by [8]:

$$L_{\text{eff}} = \frac{\lambda_0}{2\sqrt{\varepsilon_{\text{eff}}}} \tag{15.4}$$

The resonance frequency for any TM_{mn} mode for a rectangular antenna is [11]

$$f_0 = \frac{c}{2\sqrt{\varepsilon_{\text{eff}}}}\left[\left(\frac{m}{L}\right)^2 + \left(\frac{n}{W}\right)^2\right]^{1/2} \tag{15.5}$$

where m and n are modes along L and W, respectively. For efficient radiation, the optimum width W is given by [10]

$$W = \frac{\lambda_0}{2\sqrt{\frac{\varepsilon_r + 1}{2}}} \tag{15.6}$$

15.6 Radiation Pattern

A model for a patch antenna designed for f_0 = 100 MHz (λ = 3 m) is shown in Figure 15.10. Each axis is 1 m long, and the two slots are modeled as two dipoles 1.4 m apart, each with length 1.5 m. The dielectric thickness h = 1 m and the dielectric is air ($\varepsilon_r = \varepsilon_0 = 1$).

The vertical pattern (the horizontal pattern is omnidirectional) is shown in Figure 15.11, where it can be seen that the 3 dB beamwidth is about 60^o and radiation is broadside to the patch. The 3-D pattern is shown in Figure 15.12.

15.6.1 Patch Element Far-Field Patterns

Dawood and Amjad developed analytical expressions for the field patterns of the rectangular patch antenna [13]. They used the coordinate system shown in Figure 15.13. The microstrip patch is in the *xy*-plane and the *z*-axis is normal to the patch. The center of the patch is located at the origin. The spherical coordinate system is

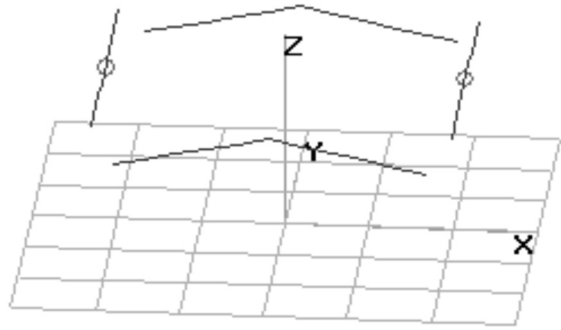


Figure 15.10 VHF patch 4NEC2 model. The two equivalent dipoles are separated from the ground plane by the thickness of the dielectric—in this case air with a thickness of 1m. The antennas are 1.5 m long (λ/2 at 100 MHz).

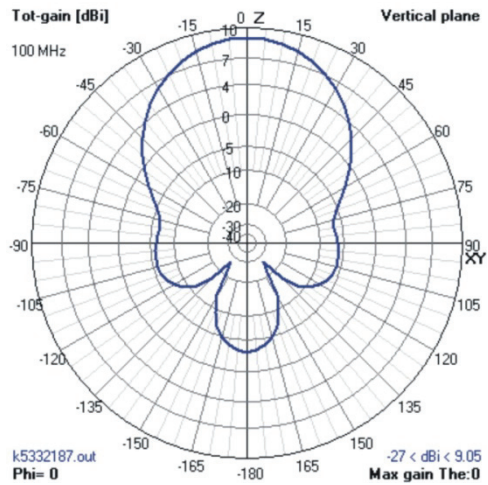


Figure 15.11 VHF patch 2-D vertical pattern. The horizontal pattern is omnidirectional.

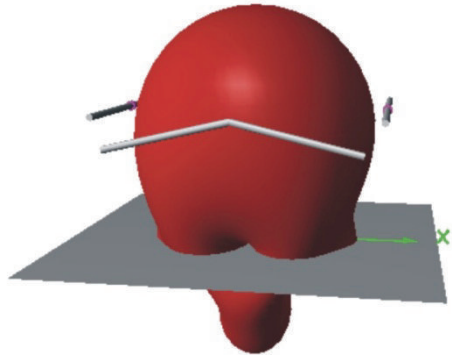


Figure 15.12 VHF patch 3-D pattern.

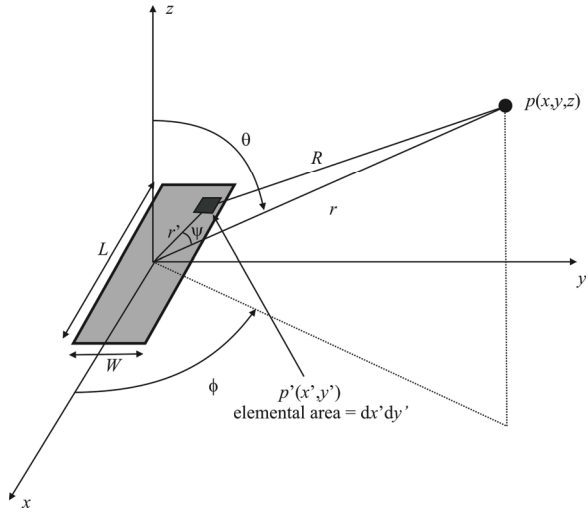


Figure 15.13 Patch coordinate system.

shown in Figure 15.13, where ϕ is the azimuth angle and θ is the elevation angle (measured from zenith). Point $p'(x',y')$ represents any point on the patch. The surface current density at that point is given by

$$J(x',y') = J(x') = \begin{cases} \bar{a}_x \dfrac{I_0}{W} \sin\left[\beta\left(\dfrac{L}{2} - x'\right)\right] & 0 \le x' \le L/2 \\ \bar{a}_x \dfrac{I_0}{W} \sin\left[\beta\left(\dfrac{L}{2} + x'\right)\right] & -L/2 \le x' \le 0 \end{cases} \tag{15.7}$$

From Maxwell's equations in Chapter 2, the fields are given by

$$E_\theta = -j\eta I_0 \frac{e^{-j\beta r}}{\pi r} \frac{\sin\left(\dfrac{\beta W}{2}\sin\theta\sin\phi\right)}{\beta W \tan\theta\tan\phi} \frac{\left[\cos\left(\dfrac{\beta L}{2}\sin\theta\cos\phi\right) - \cos\left(\dfrac{\beta L}{2}\right)\right]}{1-\sin^2\theta\cos^2\phi} \tag{15.8}$$

$$E_\phi = j\eta I_0 \frac{e^{-j\beta r}}{\pi r} \frac{\sin\left(\dfrac{\beta W}{2}\sin\theta\sin\phi\right)}{\beta W \sin\theta} \frac{\left[\cos\left(\dfrac{\beta L}{2}\sin\theta\cos\phi\right) - \cos\left(\dfrac{\beta L}{2}\right)\right]}{1-\sin^2\theta\cos^2\phi} \tag{15.9}$$

$$H_\theta = -jI_0 \frac{e^{-j\beta r}}{\pi r} \frac{\sin\left(\dfrac{\beta W}{2}\sin\theta\sin\phi\right)}{\beta W \sin\theta} \frac{\left[\cos\left(\dfrac{\beta L}{2}\sin\theta\cos\phi\right) - \cos\left(\dfrac{\beta L}{2}\right)\right]}{1-\sin^2\theta\cos^2\phi} \tag{15.10}$$

$$H_\phi = -jI_0 \frac{e^{-j\beta r}}{\pi r} \frac{\sin\left(\dfrac{\beta W}{2}\sin\theta\sin\phi\right)}{\beta W \tan\theta\tan\phi} \frac{\left[\cos\left(\dfrac{\beta L}{2}\sin\theta\cos\phi\right) - \cos\left(\dfrac{\beta L}{2}\right)\right]}{1-\sin^2\theta\cos^2\phi} \tag{15.11}$$

where

$$\eta = \frac{\omega\mu}{\beta}$$

The total electric and magnetic fields are

$$E = \sqrt{E_\theta^2 + E_\phi^2} \tag{15.12}$$

$$H = \sqrt{H_\theta^2 + H_\phi^2} \tag{15.13}$$

Equations (15.8)–(15.13) are used to find the far-field radiation patterns of the patch element. As with all other antenna arrays, this element pattern is then multiplied by the array factor to find the far-field radiation pattern of the MPA (we will use an 8×8 array for specificity). The resultant expression is the far-field radiation pattern of the microstrip antenna array without coupling. We then compare the radiation pattern without mutual coupling to that obtained after taking mutual coupling into account in Chapter 16.

15.7 Feed Techniques

The feeds for MPAs can be classified into two categories: contacting and noncontacting. In the contacting method, the RF power is fed directly to the radiating patch using a connecting element such as a microstrip line as shown in Figure 15.1. When the feed and patch antenna are noncontacting, electromagnetic field coupling transfers power between the microstrip line and the radiating patch [4]. The four most popular feed techniques used are the microstrip line, coaxial probe, aperture coupling, and proximity coupling.

Early microstrip antennas used either a microstrip feed line or a coaxial probe feed [4, 11, 14]. These two feeding methods are very similar in operation, and for a fixed patch size and substrate, offer essentially one degree of freedom in the design of the antenna element, that being the positioning of the feed point to adjust the input impedance level.

15.7.1 Microstrip Line Feed

For the microstrip line feed, a conducting strip is connected directly to the edge of the microstrip patch as shown in Figure 15.1. The conducting strip is smaller in width than the patch and this feed arrangement has the advantage that the feed can be etched on the same substrate to provide a planar structure.

The purpose of the inset cut in the patch (x_0) is to match the impedance of the feed line to the patch without the need for any additional matching element. This is achieved by properly controlling the inset position. Hence, this is an easy feeding scheme, since it provides ease of fabrication as well as impedance matching.

However, as the thickness of the dielectric substrate increases, surface waves and spurious radiation from the feed line also increase, which deleteriously affects the bandwidth of the antenna [5]. The feed radiation also leads to undesirable cross polarized radiation.

The E_z field in the patch is excited by the coupling of the J_z feed current. The coupling is given by [13]

$$\text{coupling} \sim \cos\frac{\pi y_0}{L} \tag{15.14}$$

where

L is the resonant length of the patch.
y_0 is the offset of the feed point from the patch edge.

Therefore the maximum coupling occurs when $y_0 = 0$ or L, at the edge of the patch. Such direct contacting feeding methods have the advantage of simplicity, but also have several disadvantages.

Figure 15.14 shows the equivalent circuit that applies to both the microstrip feed as well as the coaxial feed; the parallel *RLC* network represents the resonant patch, while the series inductor represents feed inductance of the coaxial probe or the microstrip feed line. These feeds excite the patch through the coupling of J, the feed current, to the E field of the dominant patch mode.

The microstrip antenna that is fed by the microstrip transmission line (Figure 15.1) can be considered as a (nonshielded) open-ended transmission line, which is significantly widened at its end. If an EM wave propagates along such a transmission line, electromagnetic energy is primarily radiated at the non-homogeneities (widening of the microstrip at the border of the feeding line, the antenna element, and the open end of this element) of the transmission line. The structure therefore behaves as a transmitting antenna. Moreover, if the length of the microstrip antenna element is equal to $\lambda/2$ on this widened transmission line, then the input impedance is purely real. Then the antenna is said to be in resonance.

15.7.2 Coaxial Feed

The coaxial feed, or probe feed, illustrated in Figure 15.14 [12], is a very common technique used for feeding microstrip patch antennas. As indicated in Figure 15.15 [21], the inner conductor of the coaxial connector extends through the dielectric and

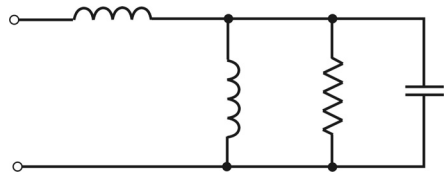


Figure 15.14 Equivalent circuit for microstrip feed and coaxial feed.

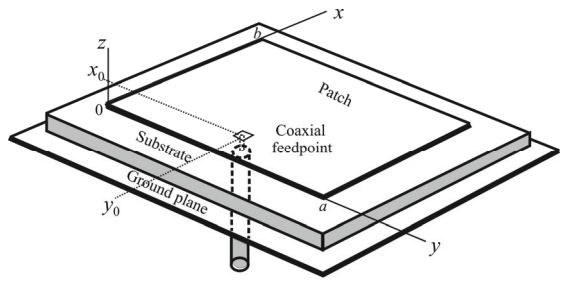


Figure 15.15 Coaxial feed.

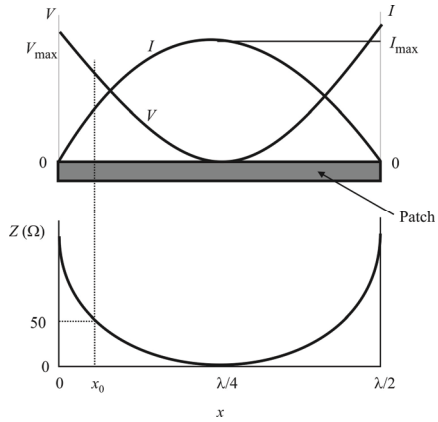


Figure 15.16 Patch characteristic impedance.

is attached to the radiating patch, while the outer conductor is connected to the ground plane. This feed method is easy to fabricate and has low spurious radiation.

The principal advantage of this type of feed is that it can be placed at any desired location inside the patch in order to match with its input impedance. This can be understood by considering Figure 15.16. This figure shows the side views of a λ/2 patch and the notional voltage and current distributions across the surface of the patch. The current is zero at the patch edges while the voltage varies across the surface. The impedance, being the ratio of the voltage to current, will vary approximately as shown at the bottom of Figure 15.16, becoming very large at the edges and some small nominal value somewhere on the patch surface. Any of the impedances can be matched simply by locating where on the surface the correct value of impedance is. Thus, the input impedance can be controlled by adjusting the position of the feed point; typical input impedances at an edge of a resonant patch range from 150Ω to 300Ω.

Both the microstrip feed and coaxial feed have some significant disadvantages. Their biggest disadvantage is that they produce narrow bandwidths. MPAs are typically narrowband in the first place, and this type of feed exacerbates that problem (if it is a problem). Such antennas are typically limited in bandwidth to

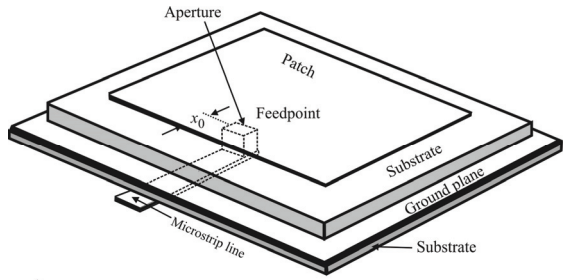


Figure 15.17 Aperture feed.

about 2–5%. These feeds suffer from the lack of effective bandwidth/feed radiation trade-off, where an increase in substrate thickness for the purpose of increasing bandwidth leads to an increase in spurious feed radiation, increased surface wave power, and possibly increased feed inductance [15].

Coaxial probes can be used to feed patch elements through the ground plane from a parallel feed substrate [7], but in an array having thousands of elements, such a large number of connections makes fabrication difficult and lowers reliability. Finally, although probe and microstrip line feeds primarily excite the dominant mode of the patch element, the inherent asymmetry of these feeds generates some higher-order modes, which produce cross-polarized radiation.

15.7.3 Aperture Coupled Feed

In this type of feed technique, the radiating patch and the microstrip feed line are separated by the ground plane as shown in Figure 15.17. This configuration uses two parallel substrates separated by the ground plane. Coupling between the patch and the feed line is made through a slot or aperture (typically a narrow rectangular slot) in the ground plane.

The amount of coupling from the feed line to the patch is determined by the shape, size, and location of the aperture. Since the ground plane separates the patch and the feed line, spurious radiation is minimized (the ground plane serves as shielding). Generally, a high dielectric material is used for the bottom substrate and a thick, low dielectric constant material is used for the top substrate to optimize radiation from the patch by allowing independent optimization of both the feed and the radiation functions [4]. In addition, the aperture is usually smaller than resonant

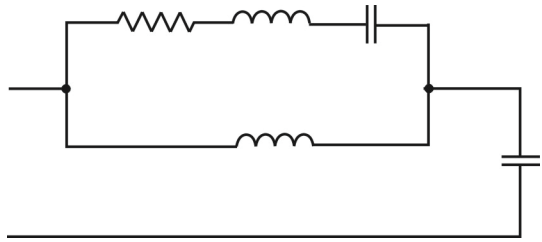


Figure 15.18 Equivalent circuit for aperture feed.

size, so the back lobe radiated by the slot is typically 15–20 dB below the forward main beam.

The major disadvantage of this feed technique is that it is more difficult to fabricate due to multiple layers, which increases the antenna thickness as well as the cost. This feeding scheme also provides narrow bandwidth.

This geometry has at least four degrees of freedom: the slot size, its position, the feed substrate parameters, and the feed-line width. Impedance matching is accomplished by adjusting the size (length) of the coupling slot, together with the width of the feed line, which is usually terminated in an open-circuited tuning stub. Coupling can occur via an equivalent electric or magnetic polarizability in the slot, but the magnetic coupling is the stronger mechanism, resulting in a coupling of the equivalent magnetic current M_y in the aperture to the dominant H_y field of the patch.

The coupling for the aperture coupling feed is [12]

$$\text{coupling} \sim \sin\frac{\pi y_0}{L} \tag{15.15}$$

Therefore, maximum coupling occurs for the aperture feed when $y_0 = L/2$, or at the center of the patch, which is where the magnetic field is maximum [15]. In that case the excitation of the patch is symmetric, which reduces the excitation of higher-order modes and leads to very good polarization purity. The aperture coupled patch with a centered feed has no cross-polarization in the principal planes.

The equivalent circuit of an aperture coupled microstrip antenna is shown in Figure 15.18 [12]. The patch resonator now appears as a series RLC network with a shunt inductance representing the coupling slot. This network allows the possibility of double tuning for increased bandwidth.

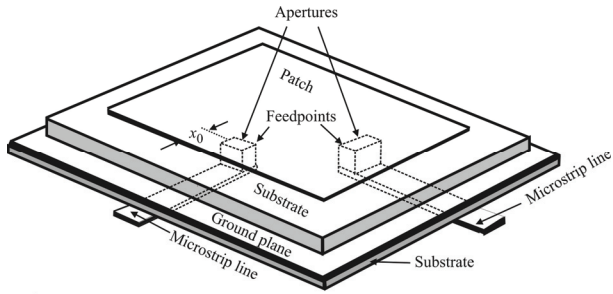


Figure 15.19 Dual aperture feed.

The above feeding techniques use a single-feed point to generate linear polarization. While it is possible to produce circular polarization with a single feed [2, 7], such techniques are usually very narrow band. Thus it is more common to generate circular polarization by using two feed points to excite two orthogonal modes on the patch with a 90^o phase difference between their excitations, as illustrated in Figure 15.19. This technique can also be applied to proximity coupled patches [16]. Yet another approach is to use sets of printed dipoles and slots proximity coupled to a microstrip feed line with the proper phasing to achieve circular polarization [12].

15.7.4 Proximity Coupled Feed

In this type of feed technique, two dielectric substrates are used such that the feed line is between the two substrates, terminating in an open stub below the patch that is printed on the upper substrate (see Figure 15.20). The main advantage of this feed technique is that it eliminates spurious feed radiation and provides higher bandwidth (as high as 13%) [4], due to overall increase in the thickness of the microstrip patch antenna. This scheme also provides choices between two different dielectric media, one for the patch and one for the feed line to optimize the individual performances. Matching can be achieved by controlling the length of the feed line and the width-to-line ratio of the patch.

The major disadvantage of this feed scheme is that it is more difficult to fabricate because of the two dielectric layers. Also, there is an increase in the overall thickness of the antenna.

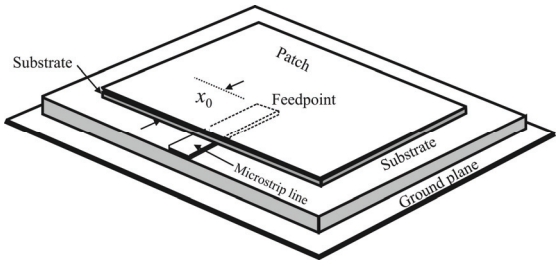


Figure 15.20 Proximity feed.

The equivalent circuit for a proximity feed is shown in Figure 15.21 [12], while Table 15.5 summarizes the salient characteristics of the different feed techniques.

15.8 Patch Antenna Models

Simple analytical models treat the microstrip antenna as a lossy cavity, or as a loaded transmission line resonator [4, 6–8]. These models involve several simplifying approximations, so they are easy to use and can provide useful information on antenna properties, such as antenna patterns, impedance, efficiency, and bandwidth.

15.8.1 Transmission Line Model

The structure is considered as a length of open transmission line along the line joining the radiating apertures loaded by impedances at its radiating edges. The radiating wall may be characterized by an equivalent admittance, the susceptance being due to the fringing field and the conductance due to the radiation loss. The transmission line model is shown in Figure 15.22. The voltage and current on this equivalent transmission line can be approximated as

$$V(x) = V_0 \cos\frac{\pi x}{L} \tag{15.16}$$

$$I(x) = \frac{V_0}{Z_0}\sin\frac{\pi x}{L} \tag{15.17}$$

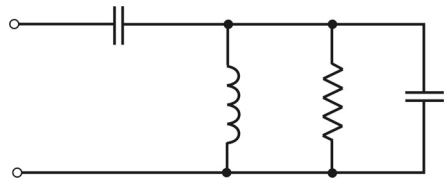


Figure 15.21 Equivalent circuit for proximity coupled feed.

Table 15.5 Characteristics of the Different Feed Techniques

Characteristics	Microstrip Line	Coaxial	Aperture Coupled	Proximity Coupled
Spurious feed radiation	More	More	Less	Minimum
Reliability	Better	Poor due to soldering	Good	Good
Ease of fabrication	Easy	Soldering and drilling needed	Alignment required	Alignment required
Impedance matching	Easy	Easy	Easy	Easy
Bandwidth*	2–5%	2–5%	2–5%	13%

*Achieved with impedance matching.
Source: [2].

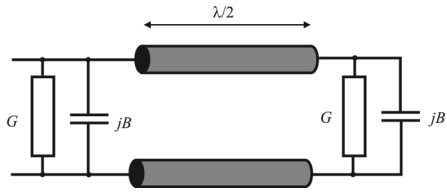


Figure 15.22 Transmission line model of a patch.

For a feed point at a radiating edge ($x = 0$ or $x = L$), the voltage is a maximum and the current is a minimum, so the input impedance is a maximum (see Figure 15.16). For a feed point at the center of the patch ($x = L/2$), the voltage is zero and the current is a maximum, so the input impedance is zero. The impedance locus is that of a half-wave open-ended transmission line resonator, which can be modeled as a parallel RLC network as illustrated in Figure 15.22.

15.8.2 Cavity Model

The region between the patch and the ground plane may be treated as a cavity, bounded by magnetic walls on the sides and electric walls on the top and bottom. The modes in this cavity are dictated by the *Helmholtz equation*

$$\frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} + \epsilon\mu\omega^2 E_z = 0 \tag{15.18}$$

Applying boundary conditions, $E_z = 0$, at the boundary walls, we get

$$H_x = \frac{j}{\omega\mu}\frac{\partial E_z}{\partial y} \tag{15.19}$$

and

$$H_y = -\frac{j}{\omega\mu}\frac{\partial E_z}{\partial x} \tag{15.20}$$

$$E_z = E_0 \cos\left(n\pi\frac{x}{a}\right)\cos\left(m\pi\frac{y}{b}\right), \qquad 0 \le x \le a, 0 \le y \le b \tag{15.21}$$

where n and m are positive integers. The frequency term in (15.18) is given by

$$\omega = \frac{\pi}{\sqrt{\epsilon\mu}}\sqrt{\left(\frac{n}{b}\right)^2 + \left(\frac{m}{a}\right)^2} \tag{15.22}$$

In the cavity model the perimeter of the patch element is approximated as a magnetic wall, for which the electric and magnetic fields of the dominant resonant mode can be expressed as

$$E_z = E_0 \cos\frac{\pi x}{L} \tag{15.23}$$

$$H_y = H_0 \sin\frac{\pi x}{L} \tag{15.24}$$

An effective dielectric loss tangent is used to account for power lost to radiation.

The magnetic wall approximation of the cavity model and the open-end approximation of the transmission line model become more realistic as the substrate becomes thinner. This implies that the Q of a patch antenna on a thin substrate is large, and that the bandwidth is small. The impedance bandwidth of a microstrip element increases with substrate thickness and decreases with an increase in substrate permittivity.

15.9 Patch Antenna Design Procedure

We present a procedure for designing a patch antenna in this section. This procedure illustrates the various parameters that can and must be addressed when examining patch antennas. The material in this section is patterned after [17].

15.9.1 Element Width

The first step is to select a dielectric substrate of appropriate thickness and properties. Several common dielectric materials that serve this function are indicated

in Table 15.3. The nonelectrical properties of some of these materials are given in Table 15.4. The three most common substrate materials used are rexolite ($\varepsilon_r = 2.6$), RT Duroid ($\varepsilon_r = 2.32$), and Alumina ($\varepsilon_r = 9.8$).

Element width is given by

$$W = \frac{c}{2f_r\sqrt{\varepsilon_0\mu_0}}\sqrt{\frac{2}{\varepsilon_r + 1}} \tag{15.25}$$

where

f_r is the resonant frequency.
ε_r is the relative dielectric constant.
$\varepsilon_0 = 8.9 \times 10^{-12}$ C^2 N^{-1} m^{-2}.
$\mu_0 = 4\pi \times 10^{-7}$ T m A^{-1}.

W is usually chosen to be in the range $1 \le W \le 2$.

Referring to Figure 15.7, the length of the resonant element is obtained from

$$L = \frac{c}{2f_r\sqrt{\varepsilon_{\text{eff}}}\sqrt{\varepsilon_0\mu_0}} - 2\Delta l \tag{15.26}$$

For a dielectric substrate with a thickness given by h, an antenna operating frequency of f_r, and for an effective radiator, a practical width W specification can be obtained from

$$\frac{\Delta l}{h} = 0.412\frac{(\varepsilon_{\text{eff}} + 0.3)\left(\frac{W}{h} + 0.264\right)}{(\varepsilon_{\text{eff}} - 0.258)\left(\frac{W}{h} + 0.8\right)} \tag{15.27}$$

15.9.2 Radiation Pattern

The E-plane and H-plane radiation patterns are given by

H-Plane

$$F(\theta) = \frac{\sin\left(\frac{\beta W}{2}\cos\theta\right)}{\frac{\beta h}{2}\cos\theta}\sin\theta \tag{15.28}$$

E-Plane

$$F(\phi) = \frac{\sin\left(\frac{\beta W}{2}\cos\phi\right)}{\frac{\beta h}{2}\cos\phi}\cos\left(\frac{\beta h}{2}\cos\phi\right) \tag{15.29}$$

For two slots spaced by a distance L apart, the E- and H-plane radiation patterns may be plotted using (15.28) and (15.29).

15.9.3 Slot Susceptance

The slot susceptance can be represented by a capacitance C

$$C = \frac{\Delta l\sqrt{\varepsilon_{\text{eff}}}}{CZ_0} \tag{15.30}$$

where Z_0 is the characteristic impedance.

15.9.4 Input Admittance

The input impedance or admittance should be accurately known so that a good match between the element and the feed can be designed. Based on the input admittance, and referring to Figure 15.16, the distance of the feed entry point to the edge of the patch (y_0) is specified by

$$Y_{\text{in}}(y_0) = 2G\left[\cos^2(\zeta y_0) + \frac{G^2 + B^2}{Y_0^2}\sin^2(\zeta y_0) - \frac{B}{Y_0}\sin(2\zeta y_0)\right]^{-1} \tag{15.31}$$

where

$Y_0 = 1/Z_0$
$\zeta = 2\pi\sqrt{\varepsilon_{\text{eff}}/\lambda_0}$
$B = \beta\Delta l\sqrt{\varepsilon_{\text{eff}}/Z_0}$
$G = 1/R_{\text{r}}$

15.9.5 Radiation Resistance/Conductance

The radiation resistance is given by

If $W \le \lambda_0$, $R_{\text{r}} = 90\lambda_0^2/W^2$
If $W > \lambda_0$, $R_{\text{r}} = 120\lambda_0/W$
and $G_{\text{r}} = 1/R_{\text{r}}$.

15.9.6 Probe Position

The equations given for calculating the feed position are very complicated, but can be simplified to give a starting point for simulation

$$R_{\text{in}} = \frac{1}{2(G_{\text{r}} \pm G_{12})} \tag{15.32}$$

where G_{r} is from Section 15.9.5.

However, G_{12} is complicated and is given by

$$G_{12} = \frac{1}{120\pi^2}\int_0^{\pi}\left[\frac{\sin\left(\frac{\beta W}{2}\cos\theta\right)}{\cos\theta}\right]J_0(\beta L\sin\theta)\sin^2\theta d\theta \tag{15.33}$$

where J_0 is the Bessel function of the first kind and order 0. Often $G_{12} << G_{\text{r}}$, so in most cases can be ignored.

We find that R_{in} is given by

$$R_{in}(y = y_0) = \frac{1}{2(G_r \pm G_{12})} \cos^2\left(\frac{\pi}{L} y_0\right) \tag{15.34}$$

and when $G_r >> G_{12}$,

$$R_{in}(y = y_0) \approx \frac{1}{2G_r} \cos^2\left(\frac{\pi}{L} y_0\right) \tag{15.35}$$

Since $\cos^2 x = 1 - \sin 2x$

$$R_{in}(y = y_0) \approx \frac{1}{2G_r}\left[1 - \sin\left(\frac{2\pi}{L} y_0\right)\right] \tag{15.36}$$

where

y_0 is the position of feed from patch edge (see Figure 15.15).
L is the length of the patch.
G_r is the radiating conductance.

15.9.7 *Q* Factor and Losses

The quality factor, Q_r, associated with the radiation resistance is

$$Q_r = \frac{C\sqrt{\varepsilon_{eff}}}{4 f_r h} \tag{15.37}$$

The equivalent resistances for the conductor loss R_c and dielectric loss R_d are then

$$R_c = 0.00027\sqrt{f_r}\,\frac{L}{W} Q_r^2 \tag{15.38}$$

$$R_d = \frac{30 \tan\theta}{\varepsilon_r} h \frac{\lambda_0}{LW} Q_r^2 \tag{15.39}$$

and the total quality factor Q_t is

$$Q_t = \frac{Q_r R_t}{R_r} \tag{15.40}$$

where R_t is the total resistance given by

$$R_t = R_r + R_d + R_c$$

15.9.8 Antenna Efficiency

The efficiency (which increases as h increases) may be expressed in terms of the equivalent resistances as

$$\eta(\%) = 100\frac{R_r}{R_t} \tag{15.41}$$

15.9.9 Bandwidth

The bandwidth may be increased by increasing the inductance of the radiators by cutting holes or slots in them or by adding reactive components to improve the match of the radiator to the feed line.

15.9.10 Directivity and Gain

As we discussed in Chapter 3, the directivity, D, of an antenna is defined as the ratio of the maximum power density in the main beam to the average radiated power density. The gain G is then given by the product of the efficiency and the directivity (3.54). The gain increases with resonant frequency and h and decreases with increase in ε_r. The directivity is independent of the resonant frequency.

15.9.11 Beamwidth

The half-power beamwidth is the angular width between directions where the gain decreases by 3 dB and is given by

$$\theta_{BH} = 2\cos^{-1}\sqrt{\frac{1}{2\left(\frac{1+\beta W}{2}\right)}} \tag{15.42}$$

$$\theta_{BE} = \cos^{-1}\sqrt{\frac{7.03}{3\beta^2 L^2 + \beta^2 h^2}} \tag{15.43}$$

Thus, the beamwidths can be increased by choosing a smaller element, thereby reducing W and L. For a given resonant frequency these dimensions may be changed, by selecting a substrate having a higher relative permittivity. As beamwidth increases, element gain and directivity decrease; however, efficiency is unaffected.

15.10 Concluding Remarks

The patch antenna was introduced in this chapter. As illustrated in Figure 15.1, a patch antenna consists of a conducting surface structured over a dielectric material with a ground plane on the reverse side.

The method of feeding a patch antenna has significant impact on its performance. There are four principal methods of feeding a patch antenna, two of which required physical contact with the feed and the antenna, and two of which do not.

A simplified model of the patch antenna was introduced to illustrate the principal operation of the antenna. This model is useful for understanding the way the patch antenna reradiates but is not very accurate for design purposes. The MoM simulation technique introduced in Chapter 2 is a better method for designing these antennas; however, the MoM does not lend itself to obtaining good insight as to how a patch antenna operates.

We introduced a procedure for designing patch antennas that is easily automated and that can be iterated to arrive at a final design.

References

[1] Carver, K. R., and J. W. Mink, "Microstrip Antenna Technology," *IEEE Transactions on Antennas and Propagation,* Vol. AP-29, No. 1, January 1981, pp. 2–24.

[2] James, J. R., "What's New in Antennas," *IEEE Antennas and Propagation Society Magazine,* February 1990, pp. 6–115.

Chapter 16

EW Applications of Patch Antennas

16.1 Introduction

The bandwidth limitations and inefficiencies of an unimproved, individual patch antenna described in Chapter 15 effectively preclude use as ES or EA antennas in many cases. There are techniques that can be done with single antennas to broaden their bandwidth, however. Also, when two or more are configured into arrays, the bandwidth can be enlarged and patch antennas can serve useful purposes in EW systems. In this chapter, we describe some bandwidth enhancement techniques for patch antennas to include the characteristics of a few of these arrays.

In the last 40 years or so, research has been focused on creating new designs or variations to the original antenna that, to some extent, produce either wider bandwidths or multiple-frequency operation in a single element. In many cases, however, it is required to add reflective or parasitic elements to the main radiating element. Also, most of these innovations bring disadvantages related to the size, height, or overall volume of the single element, and the improvement in bandwidth usually causes a degradation of one or more of these other characteristics.

This chapter is structured as follows. We begin with a discussion of bandwidth enhancement methods on a single MPA. That is followed by a section on the use of parasitic elements placed in the near vicinity of the MPA and their effects on bandwidth and other parameters. Lastly, we present a discussion on MPAs arranged into arrays: linear, log-periodic, and 2-D, including the effects of mutual coupling in MPA arrays.

16.2 Bandwidth Enhancement Techniques for a Single MPA

In MPAs the pattern bandwidth is usually many times larger than the impedance bandwidth and therefore, the discussion of bandwidth here will concentrate on

impedance rather than patterns. For a single element operating at the fundamental lowest mode, the typical bandwidth is from less than one to several percent for thin substrates. We will discuss some techniques for expanding the inherent bandwidth of a single patch element in this section [1].

As mentioned, in most cases the narrow bandwidth (2–5%) of the traditional MPA antenna element is its most serious disadvantage. Thus techniques have been devised for improving the bandwidth of the microstrip antenna. Most of the techniques applicable for bandwidth expansion of a single element can be classified as either using an impedance matching network or as parasitic elements.

16.2.1 The *Q* Factor

The antenna is a resonant circuit; therefore, energy will be stored in the region around it. This energy stored is inversely proportional to the dielectric height h. We can also describe the energy stored by the quality factor, Q [2]. Losses in the antenna will allow energy to leak away and such an antenna has a lower Q factor. The Q factor can be calculated with

$$Q = \frac{\pi f_r W_t(h)}{P_r + P_d + P_c + P_{sw}} \tag{16.1}$$

where:

f_r is the resonant frequency.
$W_t(h)$ is the stored energy in the cavity.
P_r is the energy lost through radiation.
P_d is the dielectric loss.
P_c is the conductor loss.
P_{sw} is the surface wave loss (very small).

The antenna efficiency is given by the ratio of the loss due to radiation to the total losses

$$\eta(\%) = \frac{P_r}{P_d + P_c + P_r} \times 100 \tag{16.2}$$

Typically η can be between 80% and 90% for a patch antenna with a dielectric constant of 2.3 at 10 GHz.

Typical Q-factors are in the range 20–200 and the bandwidth is given by

$$B = \frac{100(S-1)}{Q\sqrt{S}} \quad \% \tag{16.3}$$

where S is the VSWR. As we see in (16.1)–(16.3), as normally, losses reduce the Q-factor, which results in an increased bandwidth and reduced efficiency.

We know that the volume occupied by an antenna is inversely proportional to the Q of that antenna in terms of wavelengths [3]. Lowering the Q, and thereby increasing the bandwidth, can be achieved by an introduction of nonradiating losses. This is often undesirable; as we see from (16.2) it will reduce antenna efficiency, which is undesirable when transmitting and will elevate the noise figure when receiving. This leaves the aspect of broadening the bandwidth by increasing patch antenna size [4] as an alternative.

16.2.1.1 Increased Volume

In the case of the single element, width and thickness can be increased while the length is a function of the resonant frequency. Patch length can be adjusted by using a substrate with a lower dielectric constant. An additional benefit of this method is a reduction in surface wave excitation. Surface waves exist and are guided within the dielectric but can evolve to a greater unwanted condition at the microstrip level due to signals on microstrip circuits and cause a loss of power as it tends to propagate away from the microstrip [5].

16.2.2 Dielectric Substrate

One way to increase the bandwidth of the MPA is to use a thick, low dielectric constant substrate. But, as mentioned, this leads to unacceptable spurious feed radiation, surface wave generation (increased losses), or feed inductance.

A simplified method of increasing bandwidth is to adjust to a substrate of greater thickness. The limitation is realized with the elevated percentage of surface wave excitation compared to desired space wave content. The result again is a reduction in power efficiency. Table 16.1 [6] presents a comparison of bandwidth, radiation efficiency, and substrate thickness.

The bandwidth can be increased by decreasing the relative permittivity. We see from (15.40) that lowering ε_{eff} lowers the Q, which means the bandwidth increases. This, however, has an obvious limitation based on size.

16.2.2.1 Surface Waves

We mentioned previously that signals on microstrip circuits could excite surface waves. This is usually in the form of the TM_0 mode and can be minimized by

Table 16.1 Bandwidth and Efficiency of Patch Antennas as a Function of Height

h / λ_0	Bandwidth (%)	Radiation Efficiency (%)
0.010	1.54	96.3
0.020	3.01	92.9
0.030	4.60	89.7
0.040	6.28	86.8
0.050	8.08	84.1
0.060	10.11	81.6

Source: [6].

selection of thin substrates. Surface waves can be excited at the dielectric-to-air interface and they give rise to end-fire radiation (off the ends/sides of the antenna), resulting in increased sidelobes. In addition, they can lead to unwanted coupling between array elements.

The excitation of surface waves in a dielectric slab backed by a ground plane has been well documented [7]. The lowest-order TM mode, TM_0, has no cutoff frequency, while the cutoff frequency for the higher-order modes (TM_n and TE_n) are given by

$$f_c^{(n)} = \frac{nc}{4h\sqrt{\varepsilon_r - 1}}, \; n = 1, 2, \ldots \tag{16.4}$$

(see Chapter 4 for a discussion of these modes).

The cutoff frequencies for the TE_n modes are given by the odd n = 1, 3, 5,…, and the cutoff frequencies for the TM_n modes are given by even n. For the TE_1 mode the calculated values of $h / \lambda_c^{(1)}$ are $[\lambda_c^{(1)} = c / f_c^{(1)}, \; h / \lambda_c^{(1)} = n / (4\sqrt{\varepsilon_r - 1})]$:

- 0.217 for duroid ($\varepsilon_r = 2.32$);
- 0.0833 for alumina ($\varepsilon_r = 10$).

Thus, the lowest-order TE_1 mode is excited at 41 GHz for 0.16 cm thick duroid substrate, and at about 39 GHz for 0.0635 cm thick alumina substrate. The substrate thickness must be chosen so that [6]

$$h < \frac{c}{4 f_U \sqrt{\varepsilon_r - 1}} \tag{16.5}$$

is satisfied, where f_U is the highest frequency in the band of operation. Note that h should be chosen as high as possible, under the constraint of (16.5), so that maximum efficiency is achieved (increases in h cause increases in efficiency). Of course, the chosen height h must to conform to the commercially available substrates, as well.

The TM_0 mode has no cutoff frequency and is always present to some extent. The surface TM_0 wave excitation becomes appreciable when $h/\lambda > 0.09$ when $\varepsilon_r \cong 2.3$ (see Appendix D for a discussion of dielectric materials), and when $h/\lambda > 0.03$ when $\varepsilon_r \cong 10$. Generally, to suppress the TM_0 mode, the dielectric constant should be low, as should the substrate height. Unfortunately, decreasing ε_r increases the antenna size, while decreasing h leads to lower antenna efficiency and operating frequency band.

The practice of limiting width to less than a wavelength in the dielectric is explained by referring to the behavior of a microstrip transmission line (Chapter 4). The electric field below the microstrip, within the dielectric, is essentially uniform across the line (in the y-direction). With an increase in frequency with a particular line-width or similarly, an increase in line width at a particular frequency, it is possible to excite a transverse resonant mode on the line [4]. The MPA will behave as a resonator developing a standing wave along its width. The electric field is at a maximum at both ends of the line (or patch) and is at a minimum at the center. As a transmission line this is undesirable, as there will be a loss of power. Signal distortion can also result as energy splits into two modes that travel at different velocities [5].

16.2.2.2 Increasing the Thickness of the Substrate

Another way of increasing the bandwidth is to simply increase the thickness of the substrate [8, 9]. However, this technique introduces various problems. A thicker substrate tends to support larger surface waves, which deteriorates the radiation patterns and reduces the radiation efficiency. Also, feeding the antenna becomes more difficult. Additionally, higher-order modes may arise that exacerbate the problems with pattern distortions and impedance characteristics.

16.2.3 Impedance Matching

Since the bandwidth of the element is usually dominated by the impedance variation, it is generally possible to implement a planar impedance matching network to increase bandwidth. Using this technique, a bandwidth of 13% was achieved for a proximity coupled patch element with a stub tuning network [10]. Spurious radiation from the matching network must be addressed if the matching network is coplanar with the antenna element. Figure 16.1 shows one method of

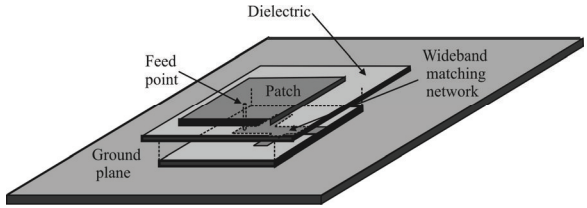


Figure 16.1 Wideband matching network.

inserting such a matching network, where the network is plated onto a dielectric layer between the ground plane and patch.

16.2.4 Slotted Patches

Slots etched into a patch antenna near the edges have some of the characteristics of patch antennas constructed with added elements. The idea is to introduce two slots in the radiating edges of the patch, as shown in Figure 16.2**.** These slots, which are etched close to the radiating edges, do not change the first resonance frequency and radiation pattern of the patch significantly while introducing another resonance with similar radiating properties that is determined by the slot length. Dual-frequency operation is possible with two slots or only one slot printed on the patch. Dual-frequency operation can also be achieved by employing dual slots beneath a single antenna. Typical performance is illustrated in Figure 16.3.

16.2.5 An Adjustable Air Gap Between the Substrate and the Ground Plane

Including an air gap to tune the resonant frequency of the patch antenna allows for tuning a patch antenna. The structure is made of two layers, including the substrate of thickness h and an air region of thickness D (see Figure 16.4). The effective permittivity is reduced, tending toward the free space value ε_0 as D increases. This concept can also be applied to stacked patches, performing a tunable arrangement with two stacked discs (see Figure 16.5). In this case, the upper air gap has the effect of altering the resonant frequency of the upper resonance, while the lower air gap has more complicated impacts. The air gap does not affect the radiation fields significantly.

A variation of this concept includes a lossless matching network. This has the effect of increasing the bandwidth. The air gap width cannot be altered arbitrarily, however. There are three regions in which trade-offs between good pattern

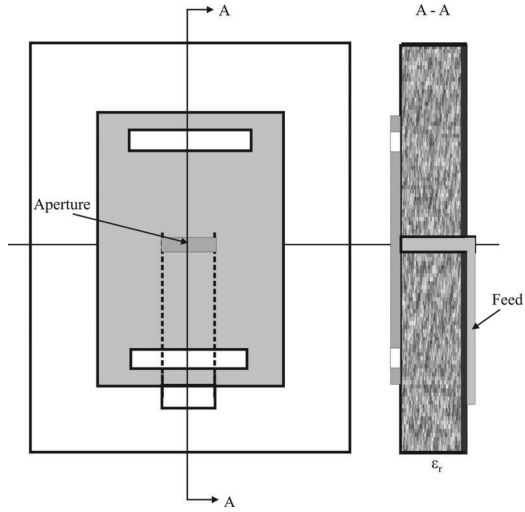


Figure 16.2 A slotted patch.

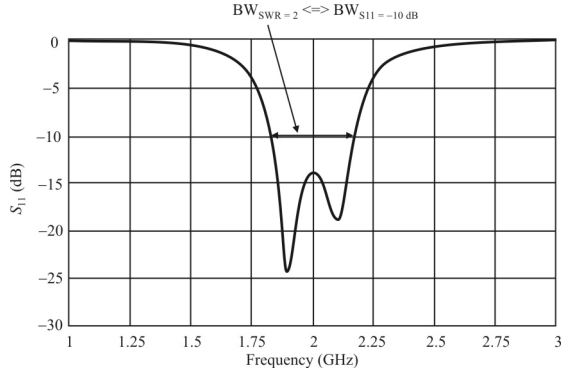


Figure 16.3 Slot patch reflection coefficient. In this example the bandwidth (VSWR = 2) is approximately 0.33 GHz or 16.5%.

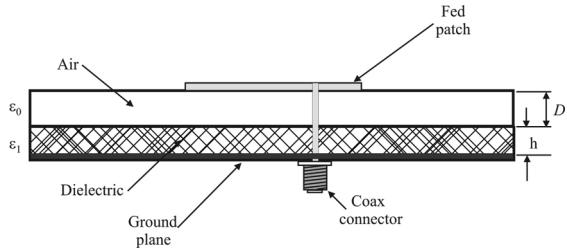


Figure 16.4 Patch bandwidth enhancement via air gap.

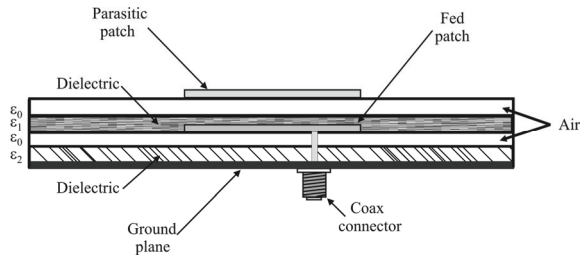


Figure 16.5 Patch bandwidth enhancement via air gap with parasitic.

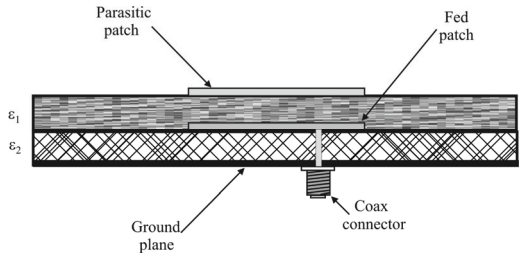


Figure 16.6 Patch bandwidth enhancement via a parasitic element.

characteristics and wider bandwidth are required. For D less than 0.14 λ_0, the patterns show good broadside features, but between 0.14 λ_0 and 0.31 λ_0 the bandwidth is reduced dramatically and the patterns showed a dip at broadside. For values of D greater than 0.31 λ_0, the pattern returns to normal, and a high gain is achieved with a narrow bandwidth.

16.3 Parasitic Patches (Multiple Patches)

Increasing the bandwidth is possible with more complicated structures. A two-patch configuration is shown in Figure 16.6 where a parasitic patch has been added in front of the fed patch. This produces a double-tuned resonance. The top patch is proximity coupled to the bottom patch, which can be fed by any of the methods to be discussed. Bandwidths of 10%–20% are possible with this structure with probe-fed stacked patches [11–13], and 18%–23% bandwidths are possible with aperture coupled stacked patches [14, 15].

Two patches (typically circles) can be vertically stacked and built like a multi-layer printed circuit board. Vertically stacking two patches creates a dual band or a wideband patch antenna, depending on the feed connections to the patches. There are three ways to connect to this antenna:

- Connecting to both patches individually (dual band, really a type of vertically arranged array);
- Connecting to the upper patch only: series fed (dual band);
- Connecting to the lower patch only: shunt fed (wideband).

Connecting to both patches is a useful technique to create a dual band patch with separate feeds for each band (see Figure 16.7). The required volume is substantially lower than when using separate patch antennas for each band. The

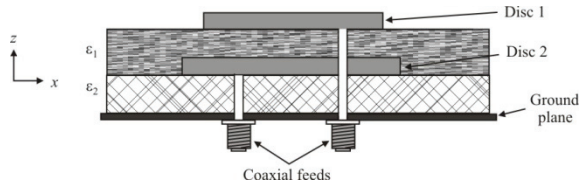


Figure 16.7 Stacked circular patches with dual feed.

antenna itself acts as a diplexer, eliminating the need for an additional discrete diplexer (see Chapter 21 for a discussion of diplexers).

Feeding the upper patch only (*series feed*) is a technique to create a dual band antenna with a single feed (see Figure 16.8). Typically, the lower frequency band uses the lower (and larger) patch, while the upper frequency band uses the upper (and smaller) patch. At the lower frequency, the cavity between the lower patch and the ground plane resonates and since the off-resonance impedance of patch antennas is effectively zero, the cavity between the upper and lower patches is electrically short circuited because it is off-resonance. At the higher frequency, the cavity between the upper patch and lower patch resonates, while the cavity between the lower patch and the ground plane is electrically short circuited because it is off-resonance. This is illustrated in the impedance versus frequency plot in Figure 16.9 and the input reflection versus frequency plot in Figure 16.10. It is important that the upper patch be sufficiently smaller than the lower patch in order to avoid fringing field coupling between the two patches. This implies that the frequency bands should be separated by at least 5% to 10%.

Two distinct resonances are observed. The lower resonant frequency is relatively steady over a range of different diameters for the upper conductor, whereas the second resonance is highly dependent on those diameters.

Series fed patches work very well (and only then) as dual frequency antennas when the two bands are well separated and the antennas present an effective short at each other's band of operation.

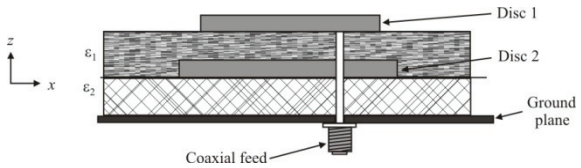


Figure 16.8 Series fed stacked circular disc upper feed.

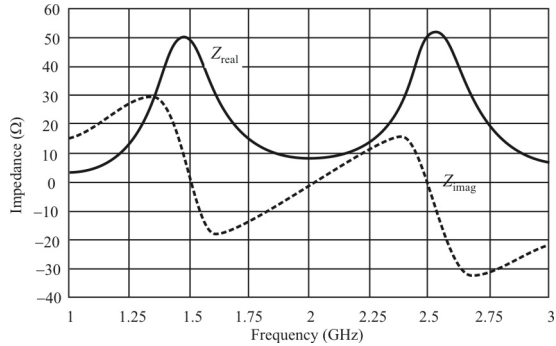


Figure 16.9 Series fed stacked patch impedance behavior versus frequency.

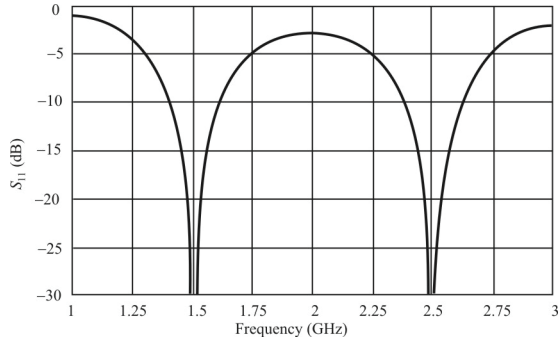


Figure 16.10 Series fed stacked patch input reflection versus frequency behavior. (SWR < 2 ⇔ S_{11} = reflection coefficient < –10 dB.)

Feeding the lower patch only (*shunt feed*) is a technique to create a wide band (3% to 10%) patch antenna where each of the two vertically stacked patches covers a subband (see Figure 16.6). Dual frequency operation does not work very well with this configuration because the upper patch is electromagnetically coupled with the lower patch. This puts a constraint on the dimensions of the patches making this a wideband rather than dual band structure. It is, however, a successful technique for widening the frequency response for EW applications where wide bandwidth is frequently desirable.

Shunt fed patches work well as wideband antennas when (and only when) both resonances are within a couple percent of each other. Typical shunt-fed self-impedance performance is illustrated in Figure 16.11. The reflection coefficient characteristics for this arrangement are shown in Figure 16.12.

16.3.1 Coplanar Strips

Another bandwidth enhancement technique for MPAs is to incorporate parasitic coplanar metallic strips coupled to the main patch. The parasitic resonators are gap coupled to either of the edges of the main patch. When the resonant lengths of the two *coupled* patches are different, two separate loops appear in the input impedance Smith chart locus, and the gap width s can be changed to yield a broader bandwidth, up to 5.1 times that of a single rectangular patch antenna.

Any of the feed-point techniques discussed in Chapter 15 can be used for the driven patch placed in an appropriate inset from the edge to provide proper impedance matching. The configuration described is illustrated in Figure 16.13.

Fringing electromagnetic fields couple the parasitic patch elements to the driven element. The parasitic patch lengths and widths need not be the same as the driven element, and varying them results in resonant frequencies that are marginally different from the driven patch. This effectively widens the bandwidth of the array. The amount of coupling, and therefore the resulting bandwidth, is dependent upon the spacing between the driven and parasitic elements. Separations from 0.3 to 1 times the substrate height are typical along the radiating edges, while narrower gaps can be used along the nonradiating edges. Bandwidth improvements of three to six times that of a single patch element have been produced [16, 17]. Figure 16.14 shows a coplanar microstrip patch antenna constructed on FR4 [6]. The VSWR plot is seen in Figure 16.15 [6].

16.4 Patch Antenna Arrays

MPAs that operate as a single element usually have a relatively large half power beamwidth, low gain, and low radiation efficiency. In order to improve on these

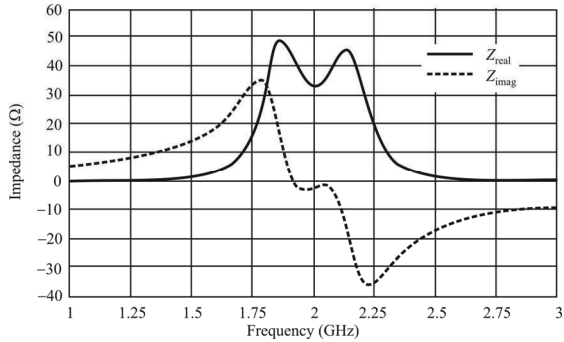


Figure 16.11 Shunt fed stacked patch impedance behavior versus frequency.

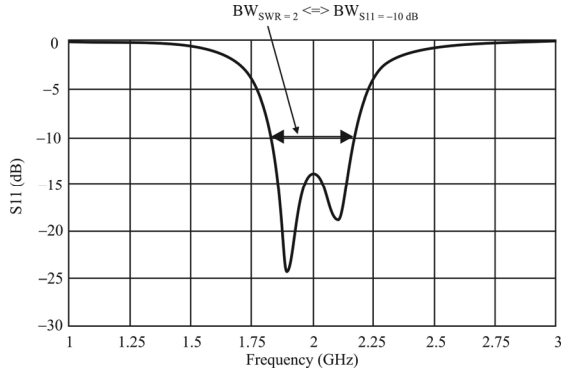


Figure 16.12 Shunt fed stacked patch input reflection versus frequency behavior.

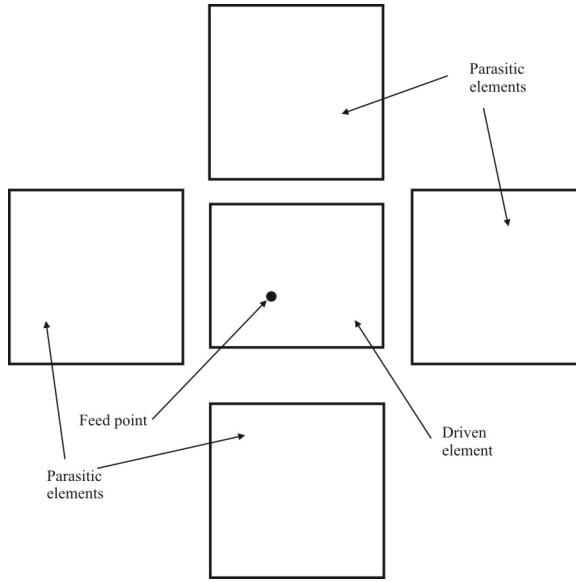


Figure 16.13 Coplanar patch antenna. One element driven with four parasitic elements.

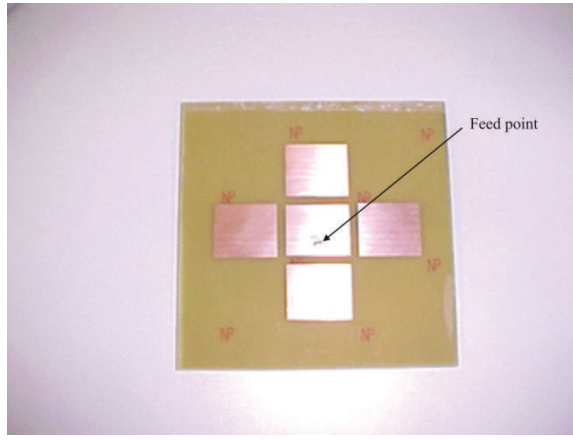


Figure 16.14 Coplanar patch antenna image. One element driven with four parasitic elements. (Source: [6].)

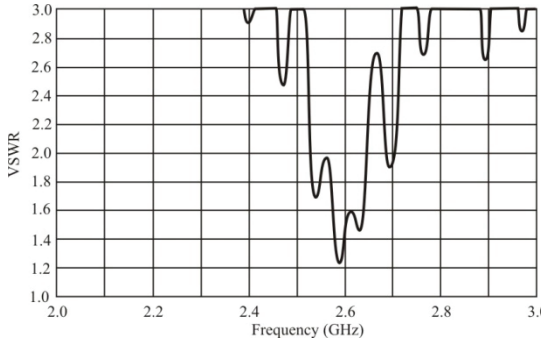


Figure 16.15 VSWR for the coplanar array. One element driven with four parasitic elements.

parameters, microstrip antennas are used in array configurations to improve the gain and range of the radiating structure.

There are many effects, such as mutual coupling between elements, that must be taken into consideration when analyzing array structures. As a result, full wave analyses are usually used to model arrays, an example of which is the MoM (see Appendix C for a discussion of the MoM).

We discuss only a few of the plethora of antenna array architectures. The possible configurations are limited only by the designer's imagination.

16.4.1 Driven Elements

Two configurations are illustrated in Figure 16.16, where both antennas are driven [18]. Figure 16.17 [19] shows the mutual coupling between the patches shown in Figure 16.16 as a function of the normalized gap size s/λ_0. Figure 16.18 [18] shows the mutual conductance as a function of normalized gap width. We notice that the coupling of the H-field dies out faster than that due to the E-field.

16.4.2 Log-Periodic Structures

16.4.2.1 Introduction

A *frequency-independent antenna* (FIA) is defined as an antenna that has no theoretical bandwidth limitations [6]. This statement implies that its geometry is self-scalable through an infinite range, such that the antenna reveals an almost constant performance. In general, the geometry of an FIA is a collection of adjoining cells. Each cell is scaled in dimensions relative to the adjacent cell, by a factor that remains constant within the structure.

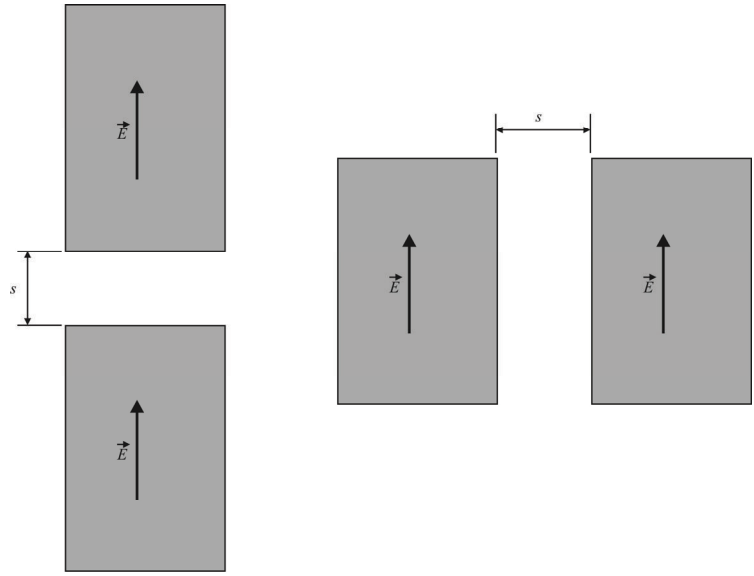


Figure 16.16 E- and H-plane configurations. (a) E-plane and (b) H-plane.

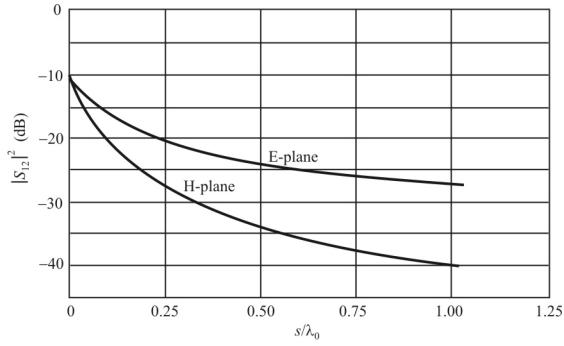


Figure 16.17 Mutual coupling between two coax-fed microstrip antennas. L = 10.57 cm, W = 6.55 cm, h = 0.1588 cm, ε_r = 2.55, f_r = 1,410 MHz.

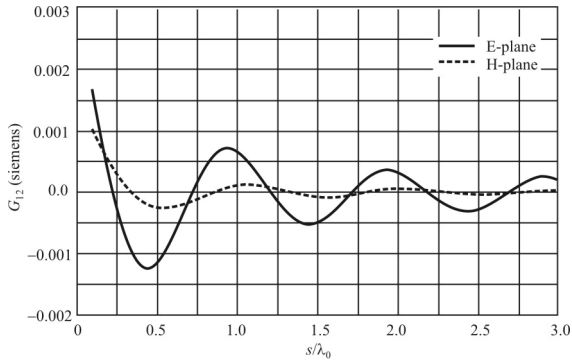


Figure 16.18 Mutual conductance.

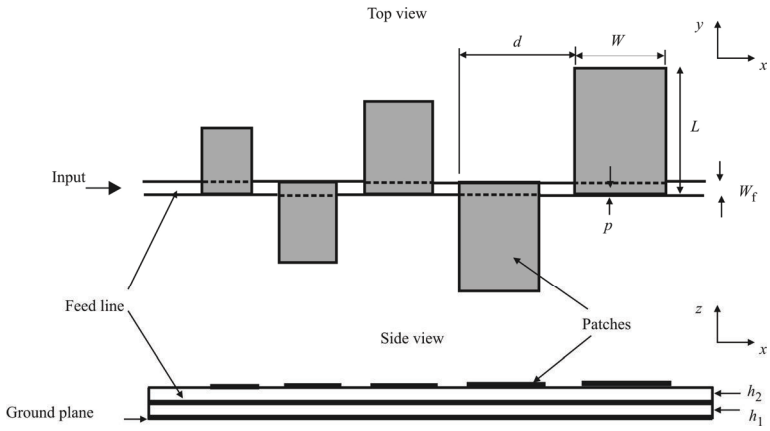


Figure 16.19 Microstrip log-periodic antenna.

One approximation to an FIA is the log-periodic architecture. In a log-periodic MPA array the resonators are directly fed by microstrip lines in a one-layer structure (see Figure 16.19 [20]). The low- and high-frequency limits of the bandwidth are set by the largest and smallest dimensions of the structure; based on the log-periodic idea, the radiators are arrayed at different narrow bandwidths, each having its own frequency band of operation, spaced and sized with a logarithmic relationship. The main advantages are the absence of an array effect in the E-plane and the fact that the antenna can be designed for a specified degree of matching by the proper choice of spacing between the resonant frequencies, namely, the log-periodic expansion factor.

A considerable part of the input power can reach the end of the feed line when is not terminated. This is an attempt to maintain high efficiency. However, other options could improve the efficiency better than an open circuit.

Application of the log-periodic principles to an array involves several inherent problems, with the constant substrate thickness the most apparent. If more than a few elements are used, the feed line will become too long for some elements and the performance will deteriorate. Yet the increase in size may not justify what this antenna can offer; the radiation patterns also vary significantly with frequency.

16.4.2.2 Log-Periodic Microstrip Patch Antenna Array

The LPA is a broadband antenna that has impedance and radiation characteristics that are regularly repetitive as a logarithmic function of the excitation frequency. The length and spacing of the elements of log-periodic antennas increase logarithmically from one end to another [16]. The principle of LPAs requires all the dimensions of the array elements to be scaled in a log-periodic manner, to achieve a corresponding scaling in frequencies. The quasi log-periodic antenna is achieved by arranging different narrow bandwidth radiators, each having its own frequency band of operation [17]. The largest and smallest dimensions of the structures can be selected. Bandwidths up to multi-octave frequency have been achieved using this technique [17].

The LPA structure is similar to a proximity coupled antenna; however, the elements are designed such that they are a log size and spacing apart. These structures have relatively broad bandwidth, some on the order of 40%. The length, width, and spacing between patches (d) are given by the expression

$$\tau = \frac{L_{m+1}}{L_m} + \frac{W_{m+1}}{W_m} + \frac{d_{m+1}}{d_m} \tag{16.6}$$

where τ is the scale factor.

We can note that all frequencies beyond the first are related by the scale factor as

$$f_1, f_{n+1} = \tau^n f_n, \qquad n = 1, 2, 3, \ldots \tag{16.7}$$

so that

$$\ln \frac{f_2}{f_1} = \ln \tau, \qquad \ln \frac{f_3}{f_1} = 2 \ln \tau, \qquad \ln \frac{f_3}{f_1} = 3 \ln \tau, \qquad \text{etc.} \tag{16.8}$$

That is why τ is called log-periodic. The height of both the substrate layers and feed-line width should be kept constant.

The basic guidelines for the design of such antennas are as follows:

- Select the substrate layers.
- Determine the upper or lower patch length and then use (16.6) to find all the corresponding patch widths and element distances.
- For the initial patch the width is chosen such that $W = 0.8 \times L$, to prevent higher-order modes.
- Find the number of patches (M) required which, palpably, is the ratio of the desired bandwidth to the bandwidth of an individual patch:

$$M = \frac{W_{\text{Desired}}}{W_{\text{SinglePatch}}} \tag{16.9}$$

- The input return loss or bandwidth can be improved by changing the patch spacing, d. The value for d is usually very close to the length L so for the initial design a d of around $d = 1.05 \times L$ can be used as an initial estimate then varied in the simulation to determine the optimum spacing.

Series-Fed Linear Patch LPA Array

A series-fed linear array is illustrated in Figure 16.20. Each patch element is connected in series by a microstrip patch transmission line of the appropriate impedance. Note the range in widths of the patches. This will control the magnitude with which each element is excited and tends to minimize losses in the system. Each of these patch elements will exhibit its own unique frequency characteristic.

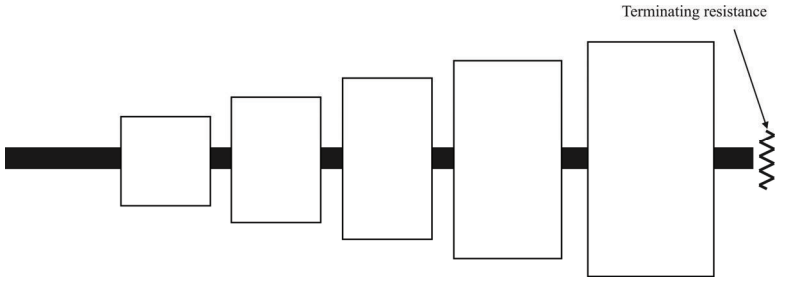


Figure 16.20 Series-fed linear array. This is an example of a microstrip patch traveling wave antenna due to the terminating resistance at the end of the feed line.

There are two types of series-fed arrays. One arrangement develops a 360° phase shift between elements. The interconnecting lines are therefore a half-wavelength long in the dielectric. A (relatively) broadband field pattern that is broadside to the antenna will be generated from this configuration. The second type of series feed is for beamforming discussed subsequently.

The series-fed broadside array is a resonant array due to the phase shift between patches. For the beam steering type of array, the elements are not a wavelength apart so they are nonresonant, or traveling wave, arrays. In both cases, however, they must be terminated with a load resistance, typically a chip resistor shunted to the ground plane by way of a via hole, in order to reduce reflections caused by the presence of nonradiated power.

The power drawn by an element is a function of its impedance and that of the line. For most of the input power to be radiated, a patch must present itself as a very low impedance. This tends to set up a condition where each element introduces a mismatch to the line and with the low impedances there will be reflections. This is a reason for using a large number of elements and therefore a distribution that offers less power per element keeping element mismatches low. Figure 16.21 shows the results of VSWR versus frequency for a series-fed patch array of the style illustrated in Figure 16.20. We can see from Figure 16.21 that an improved bandwidth over that of a single element patch antenna ensues. Excluding two locations where the VSWR exceeds 2 slightly, a very wide bandwidth is usable.

One such antenna was constructed with a 22% measured bandwidth for an SWR of ≤ 2.6, which is an improvement of approximately 10 times over the basic patch antenna. A nine-element completely log-periodic structure was also constructed with a 30% measured bandwidth for an SWR of ≤ 2.2, and 70% efficiency. In this case, the feed-line was terminated with a matched load in order to prevent reflections degrading the radiation patterns at the edges of the frequency band.

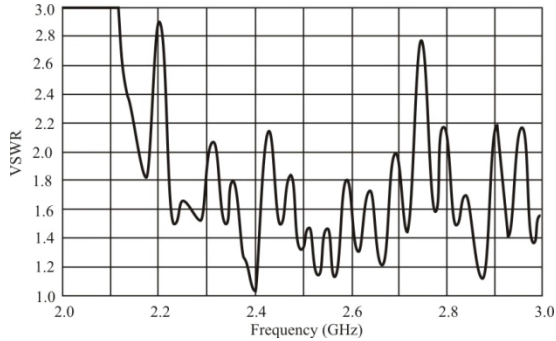


Figure 16.21 VSWR for series-fed linear patch LPA array.

Eleven-Element LPA MPA Array

A second LPA configuration was described by Rahim and Gardner [22]. This architecture, illustrated in Figure 16.22, uses inset feeds on the MPA and the MPA elements are square. Eleven elements were considered in the design. The inset feed distance of the microstrip antenna is chosen for 50Ω input impedance and it is scaled log-periodically. This model can be used for any element of log-periodic patch antenna. Calculation of design parameters for the square patch microstrip antenna is shown in Table 16.2. The substrate used is FR4 with dielectric constant of 4.7 and a height of 1.6 mm. The scaling factor τ is 1.05. The loss tangent is 0.019.

The bandwidth performance of this antenna is illustrated in Figures 16.23 and 16.24 [22]. Figure 16.23 shows S_{11} from 1 GHz to 5 GHz; both simulated and measured data are included. The bandwidth from the measurement result is 63% and from the simulation result is 57%. The resonances of the antenna can be seen

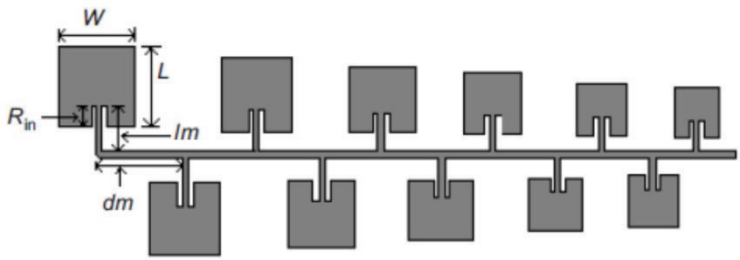


Figure 16.22 Eleven-element LPA MPA.

Table 16.2 Design Parameters for LPA with $\tau = 1.05$

	Frequency (GHz)	$W = L$ (mm)	Quarter Wave Length (mm)	R_{in} (50 Ω) (mm)
f_1	2.03	35.67	19.38	10.80
f_2	2.14	32.35	18.45	10.40
f_3	2.24	30.81	17.55	9.80
f_4	2.35	29.34	16.71	9.55
f_5	2.47	27.94	15.90	9.00
f_6	2.60	36.61	15.14	8.60
f_7	2.73	25.34	14.40	8.20
f_8	2.87	24.13	13.70	7.80
f_9	3.01	22.98	13.04	7.50
f_{10}	3.16	21.89	12.45	7.20
f_{11}	3.32	20.84	11.83	6.90

Source: [22].

by observing the dip in the return loss, at approximately 1.5, 2.0, 2.3, 2.6, 3.3, and 45.3 GHz in the simulated data. There is a close agreement between the simulation and measurement result for the bandwidth. The simulation result gives a good approximation for the measurement even though the frequency has been shifted slightly from the simulation result. The shifting of the frequency is due to the substrate of FR4, whose dielectric constant varies between 4.0 and 4.7. In this design, the dielectric constant of 4.7 has been selected. Therefore, any variations of the dielectric constant will change the frequency of the antenna. The shifting of the frequency is also from the fabrication process of the hardware. Any changes in the dimension will affect the resonance frequencies of the antenna.

The gain (S_{21}) relative to dipole across this same frequency range is shown in Figure 16.24. The gain of this antenna is from –2 to 7 dB, as shown in Figure

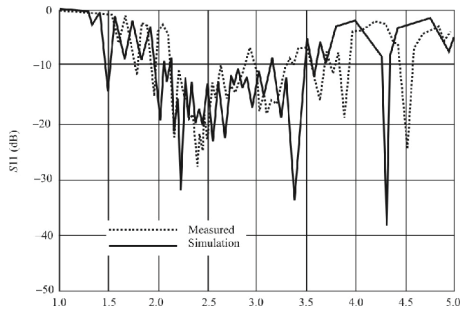


Figure 16.23 Reflection coefficient for an eleven-element LPA MPA.

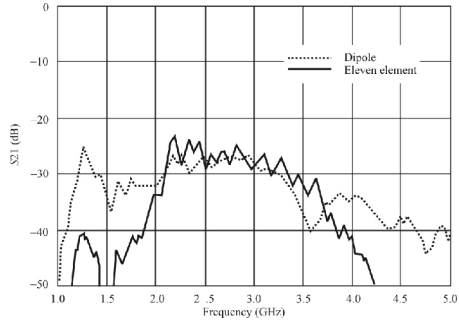


Figure 16.24 Forward voltage gain of eleven-element LPA MPA relative to dipole.

16.24. The highest gain (over a dipole) is at $f =$ 2.3 GHz. The lowest gain is –2 dB is at 3.2 GHz. As in other LPA architectures, at any frequency, the number of array elements contributing to the radiation is dependent on the bandwidth of the individual radiators, and the scaling factor used for the array. Therefore, adding more elements does not increase the peak gain. Instead, a reduction in peak gain is seen, primarily because the efficiency is degraded when more elements are added.

An eleven-element LPA MPA array was analyzed in this section. A bandwidth up to 1.8 GHz was achieved. The bandwidth has been increased from 2.8% for a single-element antenna into 63% for the eleven-element array. The typical beamwidth of the LPA is 40° for the E-plane and 60° for the H-plane. The gain of the antenna relative to the dipole antenna ranges from –2 to 7 dB.

16.5 Microstrip Array Feed Techniques

There are several ways to feed MPAs. We discussed one above—the serial feed. We introduce an additional one in this section after presenting the second use for the serial feed: beamforming.

16.5.1 Serial Feed: Beamforming

The second series feed method is fueled by a need to achieve beam steering. This requires the lengths of interconnecting lines to be adjusted in nonuniform lengths, such that the phase differences of the elements relative to each another would steer the main beam in the direction desired. Element spacing must be restricted to less than a wavelength, as grating lobes will develop. As described in Chapter 11, a grating lobe is an unwanted second main beam. It is undesirable because it is

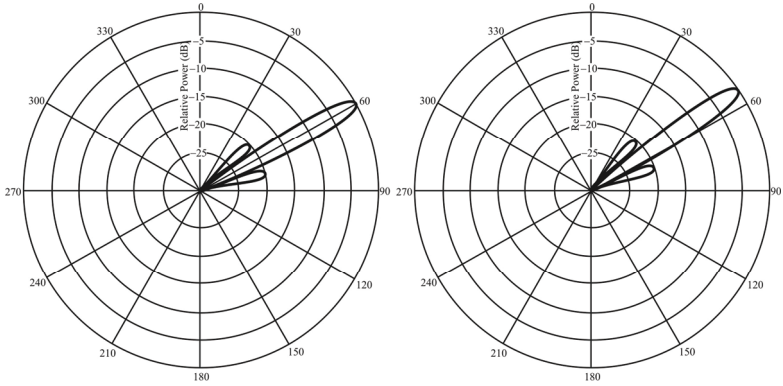


Figure 16.25 MPA beam steering. In (a) a beam is centered at 60°. In (b) the same beam has been shifted 5° by beam steering to point at 65°.

impossible to distinguish whether a signal originated from the main beam or a grating lobe. A grating lobe will not develop if the element spacing adheres to

$$d \leq \frac{\lambda_0}{1+\sin\theta_0} \tag{16.10}$$

In order to control the angle that the main beam is directed off-axis, the phase of the signal applied to the patch elements is controlled in order to achieve the desired direction of the main beam (see Chapter 11). This can be achieved by making the interconnecting horizontal line lengths unequal. Note that the main beam is 5° off axis in Figure 16.25(b) compared to Figure 16.25(a).

16.5.2 Corporate Feed

A corporate feed, also known as a parallel feed, is another of the constrained class of feeds. All elements are connected in parallel as seen in Figure 16.26, where a typical distribution of feed lines to a four-element array can be viewed. This configuration can also represent a subarray a 16-element 2-D corporate-fed antenna where four such identical arrangements are connected in a square. This case has a uniform distribution of the signal.

Although the diagram indicates that all the feed lines are of the same widths and lengths, this is not the case for all applications. There may be an application where we want to direct as much energy to the main lobe as possible and reduce the amount of power developed in the sidelobes. This is achieved through

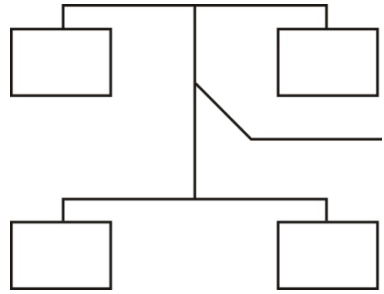


Figure 16.26 Corporate feed.

manipulation of the signal amplitudes distributed to the individual patches. The amount of power delivered to each patch element is controlled by altering the widths of transmission lines over the correct lengths of the line, thus fashioning them as power dividers (see Chapter 21). The darker feed-lines in Figure 16.27 represent wider lines than the lighter lines.

Consider the four-element corporate fed patch array shown in Figure 16.26, with uniform distribution of power and the main beam on axis. The input to the main feed-line is 50Ω. We first consider that all patches are in parallel and that the impedance at the edges are quite high, in fact at a maximum and typically between 200Ω and 300Ω. If we calculate the width of the line to have a characteristic impedance required to match the patch edge impedance, we find that the line width is very thin and impractical for both power transfer as well as manufacturability. Feed points on the interior have lower impedance values and, in pariticular, an inset point can be found that has a lower impedance, thus allowing a wider line width (see Figure 15.15). In addition, when connecting two shunt impedances together symmetrically, the total admittance must be taken into

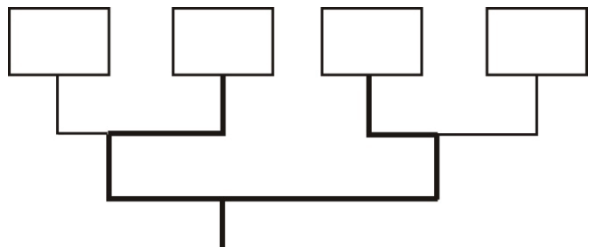


Figure 16.27 Corporate feed with different power levels delivered to the elements. The darker microstrip lines are wider lines than the lighter lines.

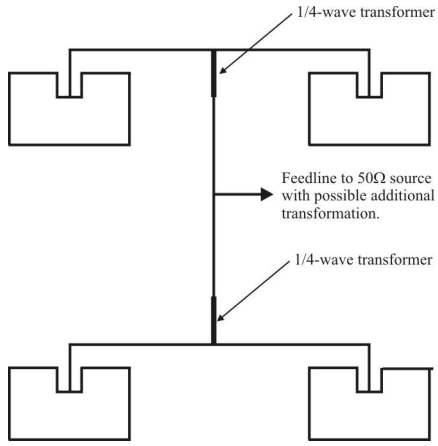


Figure 16.28 Matching with transmission lines.

consideration. We must also take into account the additonal parallel combination of each set of element pairs. The quarter-wave transformer can be used to establish appropriate admittance values to match to the 50Ω source impedance.

Assume a suitable inset impedance value has been selected at one of the patch elements that presents a feed line of practical width (> 0.3 mm). Where the patch pair and their feed-lines meet, we can use a quarter-wave transformer at the resonant frequency (in the dielectric) of the antenna to step the impedance to a value that, when both antennas are considered, matches the characteristic impedance of the source. Figure 16.28 gives an illustration of the process.

When calculating the impedance of the matching transformer, we assume a nonreactive impedance at resonance so that

$$Z_0 = \sqrt{Z_S Z_L} = \sqrt{R_S R_L} \tag{16.11}$$

The permissible element spacings can be determined by (16.10); however, in practice a separation between elements of 0.5 to 0.75 of the guide wavelength is typically selected, as otherwise feed-line lengths could become too long. This will also ensure that mutual coupling between elements is low. Figure 16.29 [6] shows an image of an antenna similar to the example of the four-element corporate-fed array. Figure 16.30 [6] is the measured VSWR versus frequency of this array. We see from Figure 16.30 that a bandwidth of approximately 8% has been achieved.

16.6 Mutual Coupling in MPA Arrays

Just as for any antenna array, when two patch elements are brought in the vicinity of each other, the amplitude and phase of the currents in both elements change.

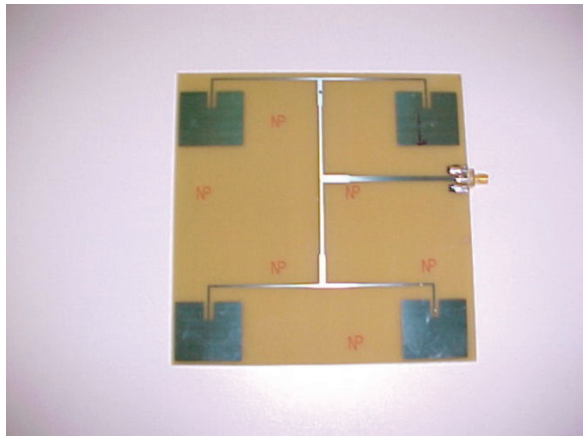


Figure 16.29 Four-element corporate-feed image. (Source: [6].)

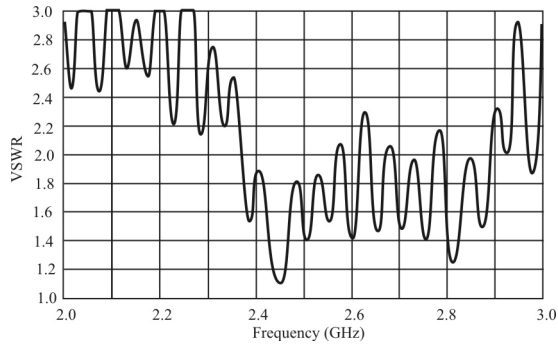


Figure 16.30 VSWR for the corporate feed.

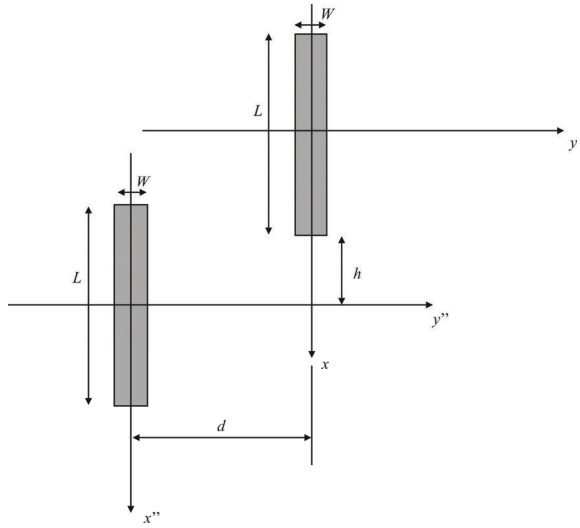


Figure 16.31 Mutual coupling between MPA elements.

The amount of change depends on the mutual coupling between the elements. The far fields of these elements change due to these changes in current. To find the resultant far-field radiation pattern, the variation in the phase and amplitude of the currents in these two individual patch elements are calculated first, and then the resultant far-field pattern can be found. Self- and mutual impedances can be derived from the near-field calculations. These impedance expressions can be used to find the effective current distributions on the individual patch elements. The current distributions are then used for calculating the far field, which is the required field in the presence of mutual coupling.

We focus on 2-D MPA arrays in this section, as that is the more difficult problem to address. Mutual coupling, however, is also present in the linear array structures discussed previously. The results presented here can be extended to those cases in a straightforward manner.

Dawoud and Amjad analyzed patch antenna arrays to determine the effects of mutual coupling between the patch elements. To find both the self- and mutual impedances, consider the arrangement of the two patches shown in Figure 16.31 [23]. Using the current density equations in Section 15.6, the near fields can be calculated using the vector potential equation (2.92), repeated here for convenience as

$$\vec{A}(\vec{r}) = \frac{\mu}{4\pi}\int\int_{V}\int \vec{J}(\vec{r}\,')\frac{e^{-j\beta R}}{R}d\vec{v}\,' \tag{16.12}$$

where

$$R = |\vec{r} - \vec{r}\,'| = \sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2} \tag{16.13}$$

If we consider only the plane of the microstrip element (that is the only place where there is any current) then

$$\vec{A} = \frac{\mu}{4\pi}\iint_{S} \vec{J}(x', y')\frac{e^{-j\beta R}}{R}d\vec{s}\,' \tag{16.14}$$

Only the A_x component exists as the current variation is only along the *x*-axis (see Figure 15.9). Using Maxwell's equation, we can find the H field from the vector potential $\vec{A}$ via (2.80) as

$$\vec{H} = \frac{1}{\mu}\nabla \times \vec{A} \tag{16.15}$$

The near fields are calculated in cylindrical coordinates. We can deduce that only the H_θ component exists and is given by

$$H_\theta = \frac{1}{4\pi}\int_{-W/2}^{W/2}\int_{-L/2}^{L/2}\frac{\partial}{\partial\rho}\left(J(x', y')\frac{e^{-j\beta R}}{R}\right)dx'dy' \tag{16.16}$$

where ρ is the distance parameter in the cylindrical coordinate system.

Manipulating (16.16) produces an expression for the mutual impedance between the patches. The resulting expression is quite complicated and does not provide much intuitive insight into the characteristics of the mutual impedance. For those interested in the derivation and the expressions for the mutual impedances, [23] is recommended. The results, however, are plotted in Figures 16.32 and 16.33 [23]. In these figures note that, as pointed out in Chapter 11, the real part of the mutual impedance can be negative. Of course, when the reactance is positive, it is inductive, while when it is negative, it is capacitive. We can see that either can be implemented with patch arrays.

Similarly, the self-impedance expression can be derived using the formulas from [23]. The final expression for the self-impedance, however, like that for the

mutual impedance, is complicated and yields little intuition. An example is plotted in Figure 16.34 [23].

The oscillatory nature of these curves is due to the near-field interactions between the two patches. The mutual impedance values fall rather rapidly as the separation distance increases. Indeed, when the MPAs are coplanar, the mutual impedance is quite small even at close distances.

The amplitude of the oscillations of the self-impedance, particularly the reactive part, remains quite large as the length of the patch (L) increases. The resistive part, while slightly oscillatory, increases its amplitude as the length increases.

16.6.1 Far-Field Array Pattern of MPA Array

As discussed in Chapter 9, the array pattern of an antenna array can be obtained by multiplying the array factor of the array and the element pattern of the microstrip patch. When there is mutual coupling between the closely placed microstrip patch elements, the array factor must take this mutual coupling into account.

The current distributions in the microstrip array elements are known after the mutual coupling effects are taken into account. For this, we first obtain the **Z**-matrix, which has mutual and self-impedance entries for each array element.

After obtaining the **Z**-matrix, for given input voltage excitations at the array elements, the current distribution on the elements in the array can be obtained that are then used to find the array pattern. The variation of mutual impedance between two elements placed side by side and close to each other is examined first as a function of separation between the centers of the patch elements (d/λ). Similarly, the variation of mutual impedance as a function of distance (s/λ) for collinearly placed patches is determined (where $s = h - L/2$).

These results are simplified in that they only consider the two patches in a pair-wise fashion, without considering if there are effects on the resulting fields of one or more intervening patch(es) between the two under consideration. With different sized patches, such as in a log-periodic array, the frequencies of concern are sufficiently far apart that the intervening antennas would have little influence. This can be seen by examining Figure 16.35, which is the reflection coefficient for a typical, unwidened patch antenna. We can see that the return loss is so high (very little signal is radiated) just a short distance from the design frequency that a patch placed in the close proximity to this patch with a different design frequency would act as a short circuit. However, if the antennas are of the same size, or approximately so, the intervening antennas will have an influence on the energy coupled between the two antennas for which we are determining the mutual impedance

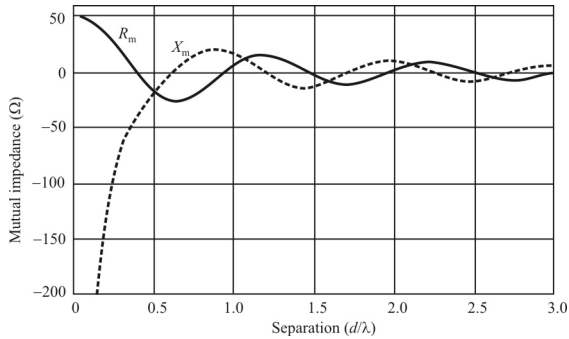


Figure 16.32 Mutual impedance when the patches are placed side by side ($h = L/2$, $d = 0$).

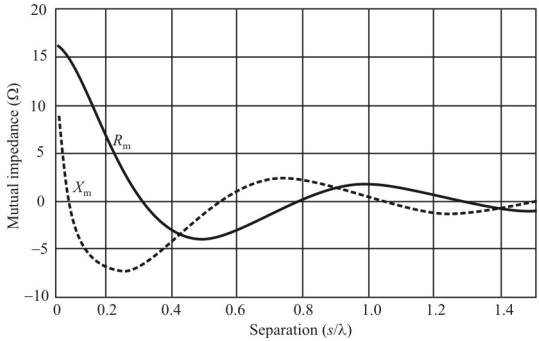


Figure 16.33 Mutual impedance between patches when the patches are colinear ($s = h - L/2$).

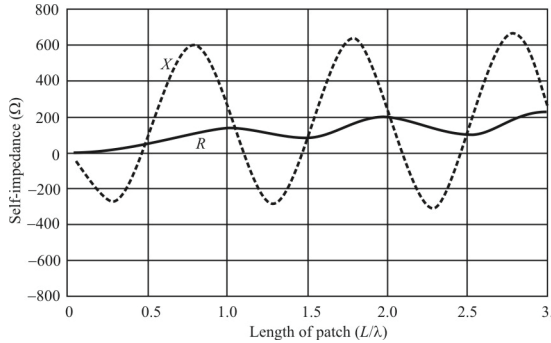


Figure 16.34 Variation in self-impedance.

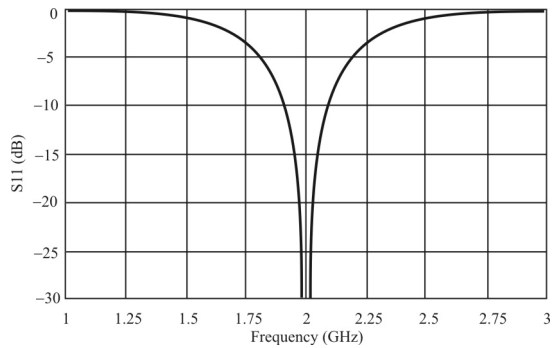


Figure 16.35 Reflection coefficient single frequency.

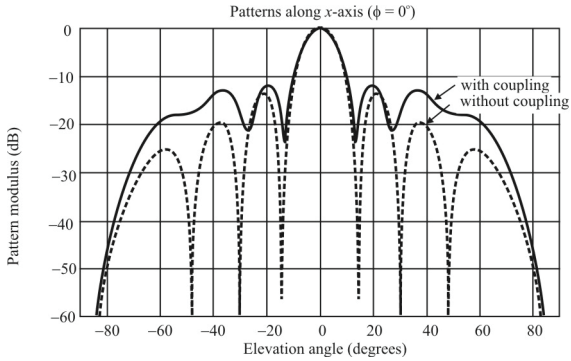


Figure 16.36 Far-field pattern variation along *x*-axis.

Ignoring this complexity, the mutual impedances for the case of side by side and co-linear placements are shown in Figures 16.32 and 16.33 respectively. The variation of self-impedance as a function of its length (L/λ) of the patch is presented in Figure 16.34. There is more variation along the *x*-axis than along the *y*-axis. This variation can be further understood from the 2-D plots of the far-field patterns taken along the *x*-axis (where $\varphi = 0°$) and comparing them to see the variation in the radiation pattern due to mutual coupling. This comparison is shown in Figure 16.36 [23]. A similar comparison of the plots is made in Figure 16.37 [23], taking radiation results along the *y*-axis (where $\varphi = 90°$). Comparisons of Figures 16.36 and 16.37 show the variation in the pattern due to coupling along the *x*-axis compared to coupling along the *y*-axis, respectively. This is caused by the difference in the antenna functioning along the two axes. The currents on the patches vary in the *x*-direction while they do not on the *y*-axes (see Figure 15.15). Furthermore, the radiating edges are oriented along the *x*-axis while the non-radiating edges are oriented along the *y*-axis. It would be expected that the

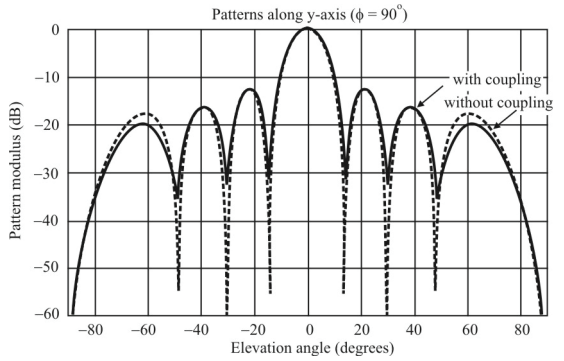


Figure 16.37 Far-field pattern along *y*-axis.

radiating slot interactions would be greater than the nonradiating edges.

16.6.2 Summary

Mutual impedances between microstrip patch antenna elements in rectangular planar arrays have been explored in this section. Explicit analytical expressions for the mutual impedance and self-impedance were developed from Maxwell's equations using the magnetic vector as the starting point. The results show that the mutual coupling between the elements of the microstrip patch array substantially changes the far-field pattern. The variation in the pattern is more pronounced along the *x*-axis than along the *y*-axis.

16.7 Concluding Remarks

This chapter presented a survey of several of the different structures used to accomplish dual-frequency or wideband operation in either a single-element patch antenna or a multi-element architecture. EW applications of antennas typically require broadband performance from the antennas. However, some of these examples create problems in the design or manufacturing stage, along with an increase in size or degradation in some of the other characteristics. By introducing slots in the patch, dual-frequency operation can be achieved in a single-element patch antenna. This dual-frequency operation is not accomplished at the expense of any other feature such as cross-polarization or distorted radiation patterns.

An extra improvement of 4% in bandwidth can be achieved with slotted patches, which means an overall bandwidth of approximately 7.5%. This improvement is much less than the 20% possible with multilayer structures or the 50 percent possible with log-periodic arrangements, but no deterioration is observed on any of the other characteristics, such as size, height, or radiation patterns.

References

[1] Pozar, D. M., "Microstrip Antennas, *Proceedings of the IEEE*, Vol. 80, No. 1, January 1992, pp. 79–91.
[2] Silver, J. P., "Micro-strip Patch Antenna Primer," www.Rfic.co.uk.
[3] Chu, L. J., "Physical Limitations of Omni-Directional Antennas," *Journal Applied Physics*, Vol. 19, December 1948, pp. 1160–1170.
[4] James, J. R., and P. S. Hall, *Handbook of Microstrip Antennas*, London, U.K.: Peter Peregrinus Ltd. 1989, Ch. 6.
[5] Sainati, R. A., *CAD of Microstrip Antennas for Wireless Applications*, Norwood, MA: Artech House, 1996, p. 139.
[6] Swick, G., technology.niagarac.on.ca.

[7] Collin, R. E., *Field Theory of Guided Waves*, 2nd ed., New York: Wiley, 1990.
[8] Sainati, R. A., *CAD of Microstrip Antennas for Wireless Applications*, Norwood, MA: Artech House, 1996, p. 25.
[9] Sanchez-Hernandez, D., "A Survey of Broadband Microstrip Patch Antennas," *Microwave Journal*, September 1996.
[10] http://www.orbanmicrowave.com/assets/Pdfs/Technical_Articles/Orban-Patch-Antennas009-rev.pdf.
[11] Pozar, D. M., and B. Kaufman, "Increasing the Bandwidth of a Microstrip Antenna by Proximity Coupling," *Electronics Letters*, Vol. 23, April 1987, pp. 368–369.
[12] Sabban, A., "A New Broadband Stacked Two-Layer Microstrip Antenna," in *IEEE Antennas and Propagation Symposium Digest*, 1981, pp. 63–66.
[13] Chen, D. H., A. Tulintseff, and M. Sorbello, "Broadband Two Layer Microstrip Antenna," *IEEE Antennas and Propagation Symposium Digest*, 1984, pp. 251–254.
[14] Lee, R. Q., K. F. Lee, and J. Bobinchak, "Characteristics of a Two-Layer Electromagnetically Coupled Rectangular Patch Antenna," *Electronics Letters*, Vol. 23, September 1987, pp. 1070–1072.
[15] Tsao, C. H., Y. M. Hwang, F. Kilburg, and F. Dietrich, "Aperture-Coupled Patch Antennas with Wide-Bandwidth and Dual Polarization Capabilities," *IEEE Antennas and Propagation Symposium Digest,* 1981, pp. 936–939.
[16] Ittipiboon, A., B. Clarke, and M. Cuhaci, "Slot-Coupled Stacked Microstrip Antennas," *IEEE Antennas and Propagation Symposium Digest,* 1990, pp. 1108–1111.
[17] Kumar, G., and K. C. Gupta, "Broadband Microstrip Antennas Using Additional Resonators Gap–Coupled to the Radiating Edges," *IEEE Transactions on Antennas and Propagation*, Vol. 32, No. 12, December 1984 pp. 1375–1379.
[18] Kumar, G., and K. C. Gupta, "Non Radiating Edges and Four Edges Gap-Coupled Multiple Resonator Broad-Band Microstrip Antennas," *IEEE Transactions on Antennas and Propagation,* Vol. 33, No. 2, February 1985, pp. 173–178.
[19] Balanis, C. A., *Antenna Theory Analysis and Design,* 3rd ed., New York: Wiley, 2005.
[20] Pozar, D. M., "Input Impedance and Mutual Coupling of Rectangular Microstrip Antennas," *IEEE Transactions on Antennas and Propagation*, Vol. AP-30, No. 6, November 1982, pp. 1192–1196.
[21] Leung from P. S. Hall, "Multioctave Bandwidth Log-Periodic Microstrip Antenna Array," *Proceedings IEE,* Vol. 133, Part H, April 1986, pp. 127–136.
[22] Rahim, M. K. A., and P. Gardner, "Microstrip Bandwidth Enhancement Using Log Periodic Technique with Inset Feed," *Jurnal Teknologi*, Vol. 41 (D), 2004, pp. 53–66. http://www.penerbit.utm.my/onlinejournal/41/D/JTDIS41D5.pdf.
[23] Dawoud, M. M., and M. K. Amjad, "Analytical Solution for Mutual Coupling in Microstrip Patch Antenna Arrays," *The Arabian Journal for Science and Engineering,* Vol. 31, No. 1 B, April 2006, pp. 47–60.

[3] Richards, W. F., Y. T. Lo, and D. D. Harrison, "An Improved Theory for Microstrip Antennas and Applications," *IEEE Transactions on Antennas and Propagation,* Vol. AP-29, No. 1, January 1981, pp. 38–46.

[4] Mailoux, R. J., J. F. McIlvenna, and N. P. Kernweis, "Microstrip Array Technology," *IEEE Transactions on Antennas and Propagation,* Vol. AP-29, No. 1, January 1981, pp. 25–37.

[5] James, J. R., P. S. Hall, C. Wood, and A. Henderson, "Some Recent Developments in Microstrip Antenna Design," *IEEE Transactions on Antennas and Propagation,* Vol. AP-29, No. 1, January 1981, pp. 124–125.

[6] Derneryd, A. G., and I. Karlson, "Broadband Microstrip Antenna Element and Array," *IEEE Transactions on Antennas and Propagation,* Vol. AP-29, No. 1, January 1981, pp. 140–141.

[7] James, J. R., and P. S. Hall, *Handbook of Microstrip Antennas,* London, U.K.: Peter Peregrinus, 1981.

[8] Sharawi M. S., "Use of Low-Cost Patch Antenna in Modern Wireless Technology," *IEEE Potential*, Vol. 25, No. 4, 2006, pp. 35–47.

[9] Schneider, M. V., "Microstrip Dispersion," *Proceedings IEEE*, Vol. 60, No. 1, January 1972, pp. 144–146.

[10] Bahl, I. J., P. Bhartia, and S. Stuchley, "Design of Microstrip Antennas Covered with a Dielectric Layer," *IEEE Transactions on Antennas and Propagation*, Vol. AP-30, No. 2, March 1982, pp. 314–318.

[11] James, J. R., P. S. Hall, and C. Wood, *Microstrip Antenna Theory and Design,* Stevenage, Herts., U.K.: Peter Peregrinus, Ltd. (IEE Electromagnetic Waves Series, Volume 12), 1981.

[12] Pozar, D. M., "Microstrip Antennas," *Proceedings of the IEEE*, Vol. 80, No. 1, January 1992, pp. 92–102.

[13] Dawoud M. M., and M. K. Amjad, "Analytical Solution for Mutual coupling in Microstrip Patch Antenna Arrays," *The Arabian Journal for Science and Engineering,* Vol. 31, No. 1, B, April 2006, pp. 47–60.

[14] Bahl, I. J., and P. Bhartia, *Microstrip Antennas.* Dedham, MA: Artech, 1980.

[15] Schaubert, D. H., D. M. Pozar and A. Adrian, "Effect of Microstrip Antenna Substrate Thickness and Permittivity: Comparison of Theories and Experiment," *IEEE Transactions on Antennas and Propagation*, Vol. 37, June 1981*,* pp. 677–682.

[16] Adrian, A., and D. H. Schaubert. "Dual Aperture-Coupled Microstrip Antenna for Dual or Circular Polarization," *Electronics Letters,* Vol. 23, November 1987, pp. 1226–1228.

[17] Silver, J. P., *Microstrip Patch Antenna Primer*, www.rfic.co.uk.

Chapter 17

Adaptive EW Antenna Arrays

17.1 Introduction

Adaptive arrays serve several useful purposes in EW systems. Principally, for performing spatial filtering on receive, a useful method of separating signals when frequency filtering or time filtering (gating) cannot be applied. They can provide DF information on signals of interest, and they can detect signals; they can reject multipaths as well as suppress cochannel interference. For transmitting purposes, adaptive arrays direct transmitted/jamming energy in specific directions as determined by associated ES capabilities. We discuss both receiving and transmitting adaptive arrays in this chapter. For our purposes, beamformer is another term for adaptive arrays, and we will use both interchangeably [1–3].

We introduce the fundamentals of adaptive arrays in this chapter, describing how signals are typically modified in amplitude and phase. We then introduce probably the most basic type of adaptive array, the Bartlett beamformer, also known as the delay and multiply beamformer. Then we will delve into more robust beamformers that are not as sensitive to calibration errors or deviations from the assumptions used to derive the analytic expressions.

The approaches to adaptive arrays included in this chapter are not exhaustive. The literature on adaptive arrays is extensive, and there are several good technical books/texts available. We present just a few of the more popular approaches so that the reader gleans a passing familiarity with the subject. For those wishing to delve further, [4, 5] are recommended.

The two primary functions of receiving beamforming algorithms are (1) power estimation and (2) signal estimation. In both cases, information on the signal angle of arrival is generated. The MOE for the former is generally the SINR.

The receiving algorithms discussed in this chapter include:

- Delay and sum beamformer, also known as the Bartlett beamformer;

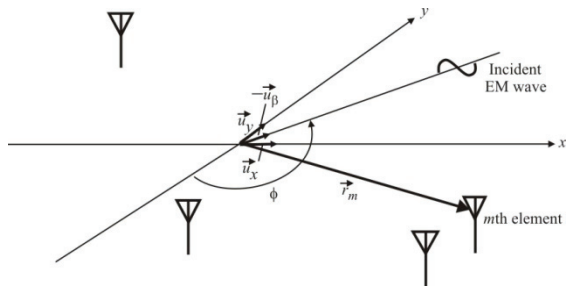


Figure 17.1 Arbitrary 2-D array.

- Standard Capon beamformer (SCB);
- Robust Capon beamformer (RCB);
- Norm-constrained Capon beamformer (NCCB);
- Eigendecomposition algorithm.

There is a popular application of beamformers where fixed beams are switched back and forth. The primary application of this is smart antennas in cellular telephone and PCS systems.

17.2 Adaptive Antenna Arrays

17.2.1 General Characteristics

An antenna sensor is represented as a point receiver at given spatial coordinates. In the 2-D case as shown in Figure 17.1, we have $\vec{r}_m = \begin{bmatrix} x_m & y_m \end{bmatrix}^{\mathrm{T}}$. The EM field at sensor m due to a source at azimuthal DOA ϕ is given by

$$E(\vec{r}_m, t) = s(t)e^{[j\omega t - \beta(x_m \cos\phi + jy_m \sin\phi)]} \tag{17.1}$$

If a flat frequency response, say, $g(\phi)$, is assumed for antenna m over the signal bandwidth, its measured output will be proportional to the field at $\vec{r}_m$. We drop the carrier dependence, recognizing that the dependence on the carrier frequency ωt is always present in all terms. The antenna output is given by

$$x_m(t) = g_m(\phi)e^{-j\beta(x_m \cos\phi + y_m \sin\phi)}s(t) = a_m(\phi)s(t) \tag{17.2}$$

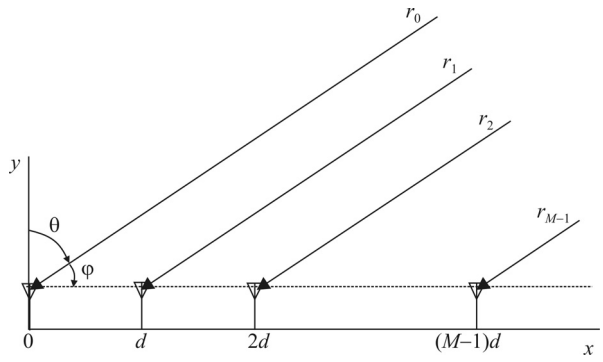


Figure 17.2 Uniform linear array.

Equation (17.2) requires that the array aperture (i.e., the physical size measured in wavelengths) be much less than the inverse relative bandwidth (f /W)—this is referred to as the *narrowband assumption.* For an M-element antenna array of arbitrary geometry, the array output vector is obtained as

$$\vec{x}(t) = \vec{a}(\phi)s(t) \tag{17.3}$$

A single signal at DOA ϕ thus results in a scalar multiple of the *steering vector* $\vec{a}(\phi) = [a_0(\phi),\ldots,a_{M-1}(\phi)]^{\mathrm{T}}$ as the array output.

Common array geometries are the ULA and UCA. In the ULA, shown in Figure 17.2, $\vec{r}_m = \begin{bmatrix}(m-1)d & 0\end{bmatrix}^{\mathrm{T}}$, and assuming that all elements have the same directivity $g_0(\phi) = \ldots = g_{M-1}(\phi) = g(\phi)$, the ULA steering vector takes the form

$$\vec{a}_{\mathrm{ULA}}(\phi) = g(\phi)\begin{bmatrix}1 & e^{-j\beta d\cos\phi} & \cdots & e^{-j(M-1)\beta d\cos\phi}\end{bmatrix}^{\mathrm{T}} \tag{17.4}$$

where d denotes the interelement distance.

The radius vectors of the UCA have the form

$$\vec{r}_m = R\begin{bmatrix}\cos\left(2\pi(m-1)/M\right) & \sin\left(2\pi(m-1)/M\right)\end{bmatrix}^{\mathrm{T}} \tag{17.5}$$

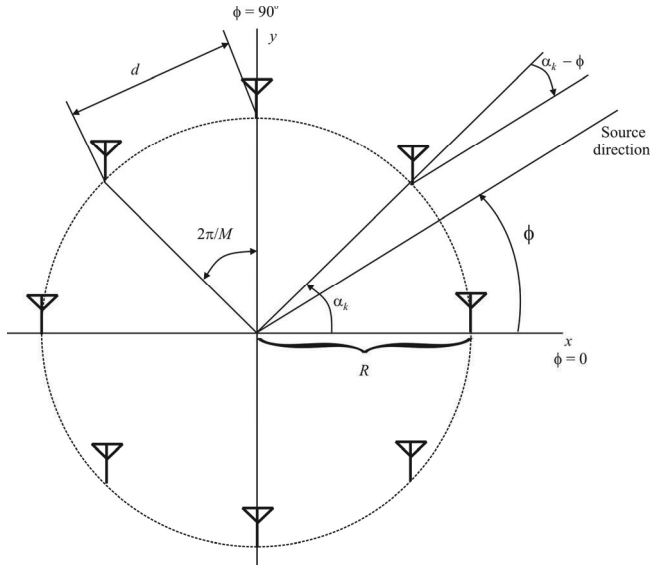


Figure 17.3 Uniform circular array for $M = 8$.

from which the UCA steering vector can be derived [5]. For a UCA illustrated in Figure 17.3, the steering vector for an M-element array for a signal arriving at angle ϕ is given by [6]

$$\begin{aligned}\vec{a}_{\text{UCA}}(\phi) &= \left[a_0(\phi) \quad a_1(\phi) \quad \cdots \quad a_{M-1}(\phi)\right]^{\text{T}} \\ &= \left[e^{-j2\pi\frac{R}{\lambda}\sin(\theta)\cos(\phi-\alpha_0)} \quad e^{-j2\pi\frac{R}{\lambda}\sin(\theta)\cos(\phi-\alpha_1)} \quad \cdots \quad e^{-j2\pi\frac{R}{\lambda}\sin(\theta)\cos(\phi-\alpha_{M-1})}\right]^{\text{T}}\end{aligned} \tag{17.6}$$

where R is the radius of the array, θ is the elevation (zenith) angle, and λ is the wavelength.

Assuming a linear receiving system (a safe assumption for our purposes), the superposition principle is applicable. If L signals impinge on an M-dimensional array from distinct DOAs $\phi_0, \ldots, \phi_{L-1}$, the output vector takes the form

$$\vec{x}(t) = \sum_{l=0}^{L-1} \vec{a}(\phi_l) s_l(t) \tag{17.7}$$

where $s_l(t)$, $l = 0, ..., L–1$ denote the baseband signal waveforms. This equation can be put in a more compact form by defining a *steering matrix* and a vector of signal waveforms as

$$\mathbf{A}(\vec{\phi}) \triangleq \left[\vec{a}(\phi_0) \quad \cdots \quad \vec{a}(\phi_{L-1})\right] \qquad M \times L \qquad (17.8)$$

$$\vec{s}(t) \triangleq \left[s_0(t) \quad \cdots \quad s_{L-1}(t)\right]^{\mathrm{T}} \qquad L \times 1 \qquad (17.9)$$

In the presence of additive noise $\vec{n}(t)$ we get

$$\vec{x}(t) = \mathbf{A}(\vec{\phi})\vec{s}(t) + \vec{n}(t) \qquad (17.10)$$

$\mathbf{A}(\vec{\phi})$ is also referred to as the *array manifold*.

The methods we discuss all require that the number of signals is less than the number of antennas, that is, $L < M$. In the noiseless case, the array output is then confined to an L-dimensional subspace of the complex M-space, which is spanned by the steering vectors. This is the signal subspace, and this observation forms the basis of subspace-based methods.

The antenna outputs are downconverted to an appropriate IF (but not demodulated) and sampled at (usually) regular time instances, labeled $t = 1, 2, ..., N$. With noise present, the process $\vec{x}(t)$ can be viewed as a multichannel random process, whose characteristics can be understood from its first- and second-order statistics determined by the constituent signals and noise. Preprocessing of the signal is often done on the signal in such a way that $\vec{x}(t)$ can be regarded as temporally white.

The *spatial covariance matrix* is given by

$$\begin{aligned}
\mathbf{R} &= \mathcal{E}\{\vec{x}(t)\vec{x}^{\mathrm{H}}(t)\} \\
&= \mathcal{E}\{[\mathbf{A}(\vec{\phi})\vec{s}(t) + \vec{n}(t)][\mathbf{A}(\vec{\phi})\vec{s}(t) + \vec{n}(t)]^{\mathrm{H}}\} \\
&= \mathcal{E}\{[\mathbf{A}(\vec{\phi})\vec{s}(t) + \vec{n}(t)][\vec{n}^{\mathrm{H}}(t) + \vec{s}^{\mathrm{H}}(t)\mathbf{A}^{\mathrm{H}}(\vec{\phi})]\} \\
&= \mathcal{E}\{\mathbf{A}(\vec{\phi})\vec{s}(t)\vec{n}^{\mathrm{H}}(t) + \mathbf{A}(\vec{\phi})\vec{s}(t)\vec{s}^{\mathrm{H}}(t)\mathbf{A}^{\mathrm{H}}(\vec{\phi}) + \vec{n}(t)\vec{n}^{\mathrm{H}}(t) + \vec{n}(t)\vec{s}^{\mathrm{H}}(t)\mathbf{A}^{\mathrm{H}}(\vec{\phi})\} \\
&= \mathcal{E}\{\mathbf{A}(\vec{\phi})\vec{s}(t)\vec{n}^{\mathrm{H}}(t)\} + \mathcal{E}\{\mathbf{A}(\vec{\phi})\vec{s}(t)\vec{s}^{\mathrm{H}}(t)\mathbf{A}^{\mathrm{H}}(\vec{\phi})\} \\
&\qquad + \mathcal{E}\{\vec{n}(t)\vec{n}^{\mathrm{H}}(t)\} + \mathcal{E}\{\vec{n}(t)\vec{s}^{\mathrm{H}}(t)\mathbf{A}^{\mathrm{H}}(\vec{\phi})\}
\end{aligned} \qquad (17.11)$$

but because the signals and noise are assumed to be uncorrelated, the first and last terms in (17.11) vanish, leaving

$$\mathbf{R} = \mathcal{E}\{\mathbf{A}(\vec{\phi})\vec{s}(t)\vec{s}^{\mathrm{H}}(t)\mathbf{A}^{\mathrm{H}}(\vec{\phi})\} + \mathcal{E}\{\vec{n}(t)\vec{n}^{\mathrm{H}}(t)\} \qquad (17.12)$$

Since the array manifold is independent of time and time delay, we have

$$\mathbf{R} = \mathbf{A}(\vec{\phi})\mathcal{E}\{\vec{s}(t)\vec{s}^{\mathrm{H}}(t)\}\mathbf{A}^{\mathrm{H}}(\vec{\phi}) + \mathcal{E}\{\vec{n}(t)\vec{n}^{\mathrm{H}}(t)\} \tag{17.13}$$

The source covariance matrix is defined as

$$\mathbf{P} \triangleq \mathcal{E}\{\vec{s}(t)\vec{s}^{\mathrm{H}}(t)\} \tag{17.14}$$

and the noise covariance matrix is defined as

$$\mathbf{N} \triangleq \mathcal{E}\{\vec{n}(t)\vec{n}^{\mathrm{H}}(t)\} = \sigma^2\mathbf{I} \tag{17.15}$$

where it has been assumed that the antenna noises are uncorrelated with each other and have identical statistics (zero mean and variance σ^2). Thus, we have

$$\mathbf{R} = \mathbf{A}(\vec{\phi})\mathbf{P}\mathbf{A}^{\mathrm{H}}(\vec{\phi}) + \mathbf{N} \tag{17.16}$$

These assumptions about statistical processes are commonly referred to as being *independent and identically distributed*, i.i.d. Such noise is termed *spatially white*, and is a reasonable model for receiver noise in our case. However, manmade noise sources generally do not result in spatial whiteness, in which case the noise must be pre-whitened[1]. This is accomplished by multiplying the sensor outputs by $\mathbf{N}^{-1/2}$.

The source covariance matrix, **P,** is often assumed to be nonsingular (coherent signals cause a rank-deficient **P**, which, although sometimes important, we will not consider here). A factorization of **R** is useful for later discussions. Since **P** is nonsingular, **R** can be factored as

$$\mathbf{R} = \mathbf{A}(\vec{\phi})\mathbf{P}\mathbf{A}^{\mathrm{H}}(\vec{\phi}) + \sigma^2\mathbf{I} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{H}} \tag{17.17}$$

with **U** unitary[2] and $\mathbf{\Lambda} = \mathrm{diag}[\lambda_0 \quad \lambda_1 \quad \cdots \quad \lambda_{M-1}]$, a diagonal matrix of real eigenvalues ordered such that $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{M-1} > 0$. Observe that any vector orthogonal to **A** is an eigenvector of **R** with the eigenvalue σ^2. There are $M - L$ such linearly independent vectors. Since the remaining eigenvalues are all larger than σ^2, we can partition the eigenvalue/vector pairs into noise eigenvectors

[1] $\mathbf{N}^{-1/2}$ is the Hermitian square root of $\mathbf{N}^{-1}$.

[2] **U** is a unitary matrix if and only if $\mathbf{U}\mathbf{U}^{\mathrm{H}} = \mathbf{I}$.

(corresponding to eigenvalues $\lambda_L = \ldots = \lambda_{M-1} = \sigma^2$) and signal eigenvectors (corresponding to eigenvalues $\lambda_0 \geq \ldots \geq \lambda_{L-1} > \sigma^2$). Hence, we can write

$$\mathbf{R} = \mathbf{U}_s \Lambda_s \mathbf{U}_s^H + \mathbf{U}_n \Lambda_n \mathbf{U}_n^H \tag{17.18}$$

where $\mathbf{\Lambda}_n = \sigma^2 \mathbf{I}$. Since all noise eigenvectors are orthogonal to **A,** the columns of $\mathbf{U}_s$ must span the range space of **A,** whereas those of $\mathbf{U}_n$ span its orthogonal complement (the null space of $\boldsymbol{A}^H$). The projection operators onto these signal and noise subspaces are defined as

$$\mathbf{\Pi} = \mathbf{U}_s \mathbf{U}_s^H = \mathbf{A}(\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H \tag{17.19}$$

$$\begin{aligned} \mathbf{\Pi}^\perp &= \mathbf{U}_n \mathbf{U}_n^H = \mathbf{I} - \mathbf{U}_s \mathbf{U}_s^H \\ &= \mathbf{I} - \mathbf{A}(\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H \end{aligned} \tag{17.20}$$

provided that the inverse in the expressions exists. Clearly

$$\mathbf{\Pi} + \mathbf{\Pi}^\perp = \mathbf{I} \tag{17.21}$$

The EW problem of central interest to us is that of estimating the DOAs of emitter signals impinging on a receiving array, when given a finite data set $\{\mathbf{x}(t)\}$ observed over $t = 1, 2, \ldots, N$. We will primarily focus on reviewing one of these algorithms based on second-order statistics of data.

These formulations assume the existence of exact quantities, which in statistical terms implies infinite observation time. It is clear that in practice only sample estimates, which we denote by a hat, that is, $\hat{\bullet}$, are available. A natural estimate of **R** is the sample covariance matrix

$$\hat{\mathbf{R}} = \frac{1}{N} \sum_{t=1}^{N} \vec{x}(t) \vec{x}^H(t) \tag{17.22}$$

for which a spectral representation similar to that of **R** in (17.18) is defined as

$$\hat{\mathbf{R}} \triangleq \hat{\mathbf{U}}_s \hat{\Lambda}_s \hat{\mathbf{U}}_s^H + \hat{\mathbf{U}}_n \hat{\Lambda}_n \hat{\mathbf{U}}_n^H \tag{17.23}$$

Table 17.1 Estimators Applicable to Arbitrary Antenna Arrays

Method	Consistency	Coherent Signals	Statistical Performance	Computations
Bartlett	$L = 1$	–	–	1-D Search
Capon*	No	No	Poor	1-D Search
MUSIC	Yes	No	Good	EVD, 1-D Search
Min-Norm	Yes	No	Good	EVD, 1-D Search
DML	Yes	Yes	Good	L-D Search
SML	Yes	Yes	Efficient	L-D Search
WSF	Yes	Yes	Efficient	EVD, L-D Search

Source: [7].

We will use this representation extensively in the explanation of the subspace-based estimation algorithms. In fact, if $\vec{x}(t)$ is a stationary white Gaussian process with unknown structure [i.e., the data model (17.10) is not used], then $\hat{\mathbf{R}}$ and its corresponding eigen-elements are maximum likelihood estimates of the exact quantities.

We assume here that the number of underlying signals, L, in the observed process is known. This is a reasonable assumption since there are good and consistent techniques for estimating L in the event that such information is not available.

There are two general categories of parameter estimation techniques, respectively classified as *spectral-based* and *parametric* methods. These algorithms can be distinguished by whether they are applicable to arbitrary array geometries or whether they require a ULA. Tables 17.1[3] and 17.2 summarize the algorithm performances [7]. The major computational requirements for each method are also included, assuming that the sample covariance matrix $\hat{\mathbf{R}}$ has

[3]This is the standard Capon beamformer. We discuss several extensions in the sequel that are not necessarily described by these categorizations.
Coherent Signals: Two signals are coherent if one is a scaled and delayed version of the other.
Consistency: An estimate is consistent if it converges to the true value when the number of data tends to infinity.
Statistical efficiency: An estimator is statistically efficient if it asymptotically attains the Cramer-Rao bound (CRB) which is a lower bound on the covariance matrix of any unbiased estimator.
1-*D* Search: the parameter estimates are computed from L one-dimensional searches over the parameter space.
L-D Search: the parameter estimates are computed from L one-dimensional searches over the parameter space.
A "good" statistical performance is when the theoretical mean square error of the estimates is close to the CRB, typically within a few decibels in practical scenarios.
EVD: Eigenvalue Decomposition
DML: Deterministic Maximum Likelihood
SML: Statistical Maximum Likelihood
WSF: Weighted Subspace Fitting
MUSIC: Multiple Signal Classification

Table 17.2 Estimators Applicable to ULAs

Method	Consistency	Coherent Signals	Statistical Performance	Computations
Root-MUSIC	Yes	Yes	Good	EVD, polynomial
ESPIRIT	Yes	Yes	Good	EVD
IQML	Yes	Yes	Good	Iterative
Root-WSF	Yes	Yes	Efficient	EVD, LS

Source: [7].

already been computed, which would be the same computations irrespective of the algorithm.

17.2.2 Narrowband Signals

A beamformer for narrowband signals samples the signal in space. The general architecture for such a beamformer is illustrated in Figure 17.4. Depending on the implementation and application, a sample of the signal to be matched [$d(n)$] may be required. The output of the array in Figure 17.4 at time interval n is given by

$$y(n) = \sum_{l=1}^{L} w_l^* x_l(n) = \vec{w}^{\mathrm{H}} \vec{x}(n) \tag{17.24}$$

Premultiplication by the conjugate transpose of the weight vector simplifies some

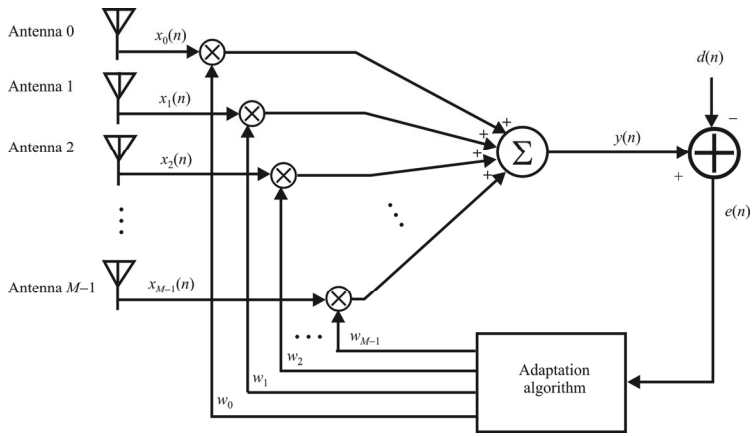


Figure 17.4 Architecture of a narrowband adaptive array. w_i are complex weights, so both the amplitude and phase are adjusted.

of the derivations. This signal is compared to a representation of the SOI and an error signal is generated, so for time interval n,

$$e(n) = y(n) - d(n) \tag{17.25}$$

$e(n)$ is then used in the adaptation algorithm to compute the weights for the next time interval $n + 1$. In fact, several iterations may occur during time interval n in order to optimize $\vec{w}$.

A signal can be considered narrowband if the reciprocal of the bandwidth is much larger than the propagation time across the array.

17.2.3 Wideband Signals

Wideband signals cannot be handled by the adaptive array described in Section 17.2.1. This is because there are frequency components in such signals that violate the assumption that the reciprocal of the bandwidth is much larger than the propagation time across the array. In other words, the propagation time across the array cannot be assumed to be the same for all frequency components.

A beamformer for wideband signals samples the signal in both time and space; the structure of such an array is illustrated in Figure 17.5. The output in this case is

$$y(n) = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} w_{m,k}^{*} x_m(n-k) \tag{17.26}$$

where K is the number of delays in each of the M antenna channels. Time sampling is required due to the broad frequency extent of the signal (and therefore typically short time constants). A signal can be considered broadband if the reciprocal of the bandwidth is on the order (or less) of the propagation time across the array.

To achieve high spatial and temporal resolution, large arrays and time windows must to be employed, thus resulting in considerable complexity in a broadband beamformer [8]. A number of techniques have been developed to find numerically efficient schemes such as DFT-based beamforming in the frequency domain [1, 6, 9–11] or subband processing [12]. Both frequency domain and subband methods apply a filter bank to the sensor data. The traditional choice of filter bank is a discrete Fourier transform (DFT). This approximation as an *independent frequency bin* (IFB) processor offers computational efficiency but also suffers from drawbacks due to the DFTs, relatively poor frequency resolution. We discuss digital beamforming more extensively later.

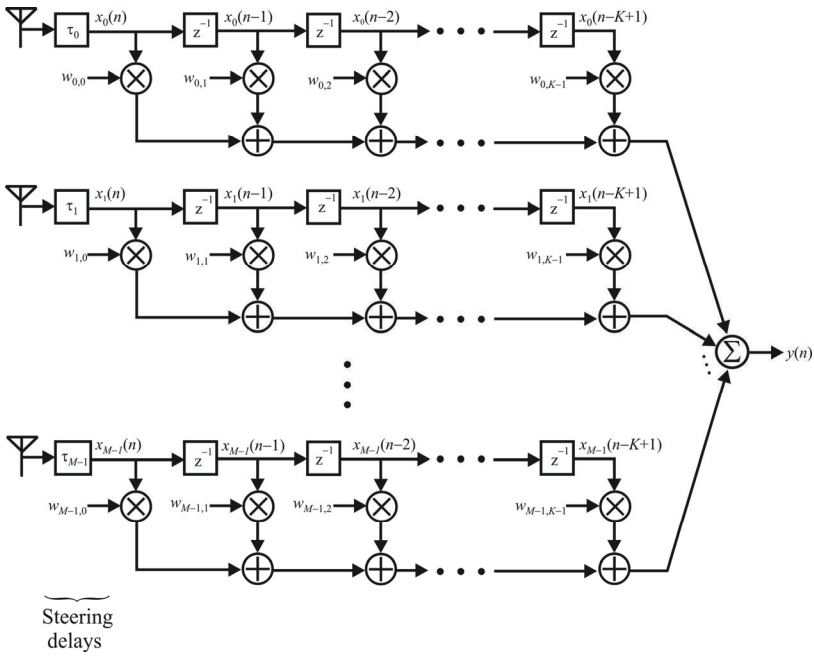


Figure 17.5 Wideband adaptive array. Sampling over time is also necessary in such an array.

17.2.4 Fixed and Adaptive Beamformers

A fixed beamformer is used when the SOI and interfering signal environment do not change or do not change rapidly, at any rate. Adaptive beamformers address the opposite case. They are applied when the signal environment is not statistically stationary and can change rapidly. The latter is normally the case of interest to our problems in EW systems.

A type of fixed beamformer was used in early applications of beamforming to cellular phone systems. Several beams were formed at the base station and as a mobile user moved throughout the cell, it would be switched from one beam to another. The idea was to increase the number of phones that a single base station could handle. The earliest versions of these fixed beams were sectored in that a beam was quite wide—120^{o} was typical, covering one-third of the total omnidirectional pattern of a base station. In this way, frequency reuse was possible by a factor of (theoretically) 3 at each base station, thus tripling the number of mobile users that could be handled.

A better way to accomplish this, in the sense of increasing the number of users per base station, is to adaptively track a mobile user, assigning a beam to each one. These are so-called "smart antennas." A single frequency is all that would be necessary at a base station. In fact, that is how the third generation systems work when CDMA is implemented. For CDMA, a single frequency is used for up to 64 users. This separation is not due primarily to beam steering, however; each user has its own code that provides the separation. Adding beams nevertheless increases the number of possible mobile users per base station even when CDMA is employed.

17.2.5 Data-Independent Beamformer

The taxonomy of beamformers can be divided into two broad categories: (1) data-independent and (2) data-dependent. The distinction is obvious from the names of the categories. Data-independent beamformers do not use any properties of the SOIs to form their beams. On the other hand, data-dependent beamformers use one or more properties inherent in the SOIs to adjust their beams [13].

The standard data-independent beamformers include the delay-and-sum (Bartlett) approach, which is known to suffer from poor resolution and high sidelobe problems, as well as methods based on various data-independent weight vectors for sidelobe control, which can have even poorer resolution.

17.2.6 Data-Dependent Beamformer

The data-dependent beamformer theory was developed to apply to the case when the SOI is known and fixed, or can be very closely approximated. In fact, if its DOA is known as well as the DOAs of the interfering signals, then a fixed, data-

dependent beamformer can be used. The array steering vector is determined by calibrating the array and stays fixed. Its adaptive nature is therefore only exploited once. If the geometry does not change, then the steering vector can be estimated fairly well.

The popular data-dependent *Capon beamformer* adaptively selects the weight vector to minimize the array output power subject to the linear constraint that the SOI does not suffer from any distortion. The Capon beamformer has better resolution and much better interference rejection capability than the data-independent beamformer, provided that the array steering vector corresponding to the SOI is accurately known.

However, knowledge of the SOI steering vector can be imprecise, which is often the case in practice due to differences between the assumed signal arrival angle and the true arrival angle or between the assumed array response and the true array response (array calibration errors). Whenever this happens, the Capon beamformer may interpret the SOI as interference and attempt to suppress it, which results in significantly underestimated SOI output power and drastically reduced array output SINR. Then the performance of the Capon beamformer may become worse than that of the standard beamformers.

Similar results ensue when the number of snapshots is relatively small (i.e., on the order of the number of antennas). Many approaches have been developed to improve the robustness of the Capon beamformer, and the literature on robust adaptive beamforming is extensive (see, for example, [13–18] and the many references therein). Among these robust approaches, diagonal loading (including its extended versions) has become a popular and widely used approach to improve the robustness of the Capon beamformer. However, for most of the diagonal loading methods, it is not clear how to choose the diagonal loading level based on information about the uncertainty of the array steering vector.

Most of the adaptive beamforming theory is based on the requirement to allow an SOI into a receiving system while nulling out one or more interfering signals. In that case the direction in which to look for the SOI is known and fixed while the interferers are allowed to be unknown. The SOI waveform is assumed to be completely known and the receiving array is perfectly calibrated with that waveform [17].

The EW applications under consideration here include using a beamformer to look in a particular direction to see if there is any signal there, while rejecting possible interfering signals from other directions. Thus, for that application we assume that there is an SOI, so in the look direction and the calibration signal is sufficiently closely matched to the SOI that significant distortion is avoided.

A beamformer can also be used to focus a jamming signal in a particular direction to maximize the power applied to the target.

The traditional approach to the design of adaptive beamformers assumes that the desired (known) signal components are not present in training data [19]. In this case, several rapidly converging techniques have been developed [20–23] that are

applicable to problems with small training sample size. Although the assumption of signal-free training snapshots may be true in some cases (such as radar), in our case the SOI is either not present at all or always present. The question being asked is whether the signal is present—it is not known, so it cannot be assumed that the signal is not there.

The presence of the signal of interest in training data may dramatically reduce the convergence rates of adaptive beamforming algorithms as compared to when the training is SOI free [24]. This is a particular problem when the training base is small.

In practical applications of adaptive beamforming, the performance degradation may become pronounced because some of the underlying assumptions on the environment, sources, or antenna array can be violated, and this may cause a mismatch between the nominal (presumed) and actual steering vectors. Adaptive array techniques are very sensitive even to slight mismatches of such type that can easily occur in practical situations as a consequence of flawed array calibration and imperfect antenna shape [25]. Other common causes of model mismatch include array response distortion due to source wavefront curvatures resulting from environmental inhomogeneities [26, 27], the near-far problem [28], source spreading, and local scattering [29–32], as well as other effects [33]. In such cases, robust approaches to adaptive beamforming are required [34, 35].

Some approaches are known to be able to at least partly overcome the problem of steering vector mismatches. The quadratically constrained beamformer (whose implementation is based on the diagonal loading of the sample covariance matrix mentioned earlier [36]) and the eigenspace-based beamformer [37] are two of the most popular. The principal shortcoming of the former approach is that it is not always clear how to obtain the optimal value of the diagonal loading factor, whereas the latter approach is essentially ineffective at low SNRs.

One of the first approaches to the space-time processing of data sampled at an array of sensors was the application of Fourier-based spectral analysis to spatio-temporally sampled data. The conventional beamformer (due to Bartlett), which has limited resolution, dates back to World War II [38–40]. Time-delay estimation techniques [41] were applied to improve the ability to resolve closely spaced signal sources. This approach, however, suffers from performance limitations due to its size. Its performance is directly dependent upon the aperture of the array (its physical size), irrespective of the amount of data available and the SNR [7].

Extension of the time-delay estimation methods to more than one signal is possible, which requires the addition of more sensor elements. In addition, the resolution limits of conventional beamforming can be improved. Resolution refers to the ability to distinguish closely spaced signal sources. Whenever there are two peaks in the spatial spectrum near two actual emitters, the targets are said to be resolved.

The parameter estimation approach to space-time processing includes the *maximum entropy* (ME) spectral estimation method in geophysics [42] and

elsewhere, and applications of the maximum likelihood principle [43, 44] can be applied to improve the resolution.

The introduction of subspace-based estimation techniques [45, 46] brought a new approach to array processing. The subspace-based approach relies on certain geometrical properties of the assumed data model, in that there are two subspaces associated with the data: the signal subspace and orthogonal noise subspace.[4] These subspace techniques resulted in a resolution capability that is not limited by the array aperture. This is provided that the data collection time and/or SNR are sufficiently large.

17.3 Narrowband Beamforming

The output of the narrowband[5] beamformer in Figure 17.4 is given by (17.24) and can be expressed as

$$y(t) = \vec{w}^{\mathrm{H}}\vec{x}(t) \tag{17.27}$$

where t is the time index, $\vec{x}(t) = [x_0(t) \quad \ldots \quad x_{M-1}(t)]^{\mathrm{T}} \in \mathbb{C}^{M\times 1}$ is the complex vector of array observations, $\vec{w}(t) = [w_0(t) \quad \ldots \quad w_{M-1}(t)]^{\mathrm{T}} \in \mathbb{C}^{M\times 1}$ is the complex vector of beamformer weights, M is the number of antennas in the array, $[\cdot]^{\mathrm{T}}$ denotes transpose, and $[\cdot]^{\mathrm{H}}$ denotes conjugate transpose. The observation vector is given by

$$\vec{x}(t) = \vec{s}(t) + \vec{i}(t) + \vec{n}(t) \tag{17.28}$$

which can also be rewritten as

$$\vec{x}(t) = s(t)\vec{a}(\theta_0) + \vec{i}(t) + \vec{n}(t) \tag{17.29}$$

where $\vec{s}(t)$, $\vec{i}(t)$, and $\vec{n}(t)$ are the desired signal, interference, and noise components, respectively. Here, $s(t)$ is the signal waveform impinging on the array from an AOA ϕ_0, and $\vec{a}(\phi_0)$ is the array steering vector[6] in that direction so that

[4] The space referred to here, from which the two subspaces are obtained, is the mathematical construct of eigenvectors associated with a matrix. The eigenvectors are said to *span* the space.

[5] When considering digital implementations, which almost all modern systems are, this could also be a narrowband, single channel of a wideband signal.

$$\vec{s}(t) = s(t)\vec{a}(\phi_0) \tag{17.30}$$

As previously mentioned, the array steering vector $\vec{a}(\phi_0)$ is the response of the antenna array to a signal arriving at AOA ϕ_0.

The attempt to automatically localize signal sources using antenna arrays is through beamforming techniques. The idea is to "steer" the array in one direction at a time and measure the output power. The steering locations that result in maximum power, or some other statistical parameter, yield the AOA estimates. The array response is steered by forming a linear combination of the antenna outputs

$$y(t) = \sum_{m=0}^{M-1} w_m^* x_m(t) = \vec{w}^{\mathrm{H}}\vec{x}(t) \tag{17.31}$$

Given samples $y(1), y(2), \ldots, y(N)$, the output power is given by

$$P(\vec{w}) = \frac{1}{N}\sum_{n=1}^{N}\left|y(n)\right|^2 \tag{17.32}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\vec{w}^{\mathrm{H}}\vec{x}(n)\vec{x}^{\mathrm{H}}(n)\vec{w}$$

$$= \vec{w}^{\mathrm{H}}\left[\frac{1}{N}\sum_{n=1}^{N}\vec{x}(n)\vec{x}^{\mathrm{H}}(n)\right]\vec{w}$$

$$= \vec{w}^{\mathrm{H}}\mathbf{R}\vec{w} \tag{17.33}$$

Different beamforming approaches correspond to various statistical parameters and lead to a number of choices of the weighting vector $\vec{w}$.

We assume that one of the signals is the desired signal, say, $s_0(t)$, and treat the remaining signals as interferers [47]. Since $s_0(t)$ is uncorrelated with the noise and interferers, the data covariance matrix has the form

$$\mathbf{R} = \sigma_0^2\vec{a}(\phi_0)\vec{a}^*(\phi_0) + \sum_{l=1}^{L-1}\sigma_l^2\vec{a}(\phi_l)\vec{a}^{\mathrm{H}}(\phi_l) = \mathbf{R}_{\mathrm{s}} + \mathbf{R}_{\mathrm{i+n}} \tag{17.34}$$

---

[6] Note that this is a simplification of the concept of *array manifold* where there are multiple signals $\{s_i(t)\}$ impinging on the array. In that case $\mathbf{A} = [\vec{a}_i]$ where each $\vec{a}_i$ is a column of $\mathbf{A}$ and is a steering vector corresponding to signal $s_i(t)$.

where $\mathbf{R}_s = \sigma_0^2 \vec{a}(\phi_0)\vec{a}^*(\phi_0)$, $\sigma_l^2 = \mathcal{E}\left\{\left|s_l(t_k)\right|^2\right\}$ is the power of *l*th signal, and $\mathbf{R}_{i+n}$ is the interference plus noise covariance matrix. In practice, the covariance matrix $\mathbf{R}$ is estimated by finding the correlation matrix of the observed data

$$\hat{\mathbf{R}} = \frac{1}{N}\sum_{n=1}^{N} \vec{x}_n \vec{x}_n^{\mathrm{H}} \tag{17.35}$$

where all received signals have zero means and the *N* samples are independent. Note that, in general, $\hat{\mathbf{R}}$ will have the SOI embedded in the data samples, where $\mathbf{R}$ is comprised of noise responses only. When the SOI is present in (17.35), the performance degrades for most beamformers.

We focus on the problem of estimating the SOI power σ_0^2 from $\mathbf{R}$ (or more practically $\hat{\mathbf{R}}$) when the knowledge of $\vec{a}(\theta_0)$ is imprecise. However, the beamforming approaches we discuss can also be used for other applications, including signal waveform estimation. Once the optimum weights $\hat{\vec{w}}$ are known, the power in the SOI, σ_0^2, can be estimated with

$$\sigma_0^2 = \hat{\vec{w}}^{\mathrm{H}} \hat{\mathbf{R}} \hat{\vec{w}} \tag{17.36}$$

17.4 Conventional Beamformer

The conventional *delay-and-multiply beamformer*, also known as the *Bartlett beamformer*, is an adaptation of classical Fourier-based frequency spectral analysis to sensor array data. For an array of arbitrary geometry, this algorithm maximizes the power of the beamformer output for a given input signal. The *conventional beamformer* (CBF) attempts to maximize the expected output power of the spatial filter (which in white noise is equivalent to maximizing the output SNR)

$$\max_{\vec{w}} \mathcal{E}\{\vec{w}^{\mathrm{H}} \vec{x}(t) \vec{x}^{\mathrm{H}}(t) \vec{w}\} \qquad \text{such that } \vec{w}^{\mathrm{H}} \vec{w} = 1 \tag{17.37}$$

Suppose we wish to maximize the output power from a certain direction ϕ_0. Given a signal $s(t)$ arriving from direction ϕ_0, the array output, corrupted by additive noise, is

$$\vec{x}(t) = \vec{a}(\phi_0) s(t) + \vec{n}(t) \tag{17.38}$$

Maximizing the output power then requires

$$\max_{\vec{w}} \mathcal{E}\{\vec{w}^{\mathrm{H}}\vec{x}(t)\vec{x}^{\mathrm{H}}(t)\vec{w}\} = \max_{\vec{w}} \vec{w}^{\mathrm{H}}\, \mathcal{E}\{\vec{x}(t)\vec{x}^{\mathrm{H}}(t)\}\vec{w} \tag{17.39}$$

$$= \max_{\vec{w}} \left[\mathcal{E}\left\{ |s(t)|^2 \left| \vec{w}^{*}\vec{a}(\phi_0) \right|^2 \right\} + \sigma^2 |\vec{w}|^2 \right] \tag{17.40}$$

where spatially white noise is assumed. The norm of $\vec{w}$ is constrained to $|\vec{w}| = 1$ so that

$$\vec{w}_{\mathrm{CBF}} = \frac{\vec{a}(\phi_0)}{\sqrt{\vec{a}^{\mathrm{H}}(\phi_0)\vec{a}(\phi_0)}} \tag{17.41}$$

Equation (17.41) can be interpreted as a spatial filter, which has been matched to the impinging signal. Inserting the weighting vector (17.41) into (17.33), the classical *spatial spectrum* results in

$$P_{\mathrm{CBF}}(\phi_0) = \frac{\vec{a}^{\mathrm{H}}(\phi_0)\hat{\mathbf{R}}\vec{a}(\phi_0)}{\vec{a}^{\mathrm{H}}(\phi_0)\vec{a}(\phi_0)} \tag{17.42}$$

For a ULA of M isotropic antennas, the steering vector $\vec{a}(\theta_0)$ is

$$\vec{a}_{\mathrm{ULA}}(\phi_0) = \begin{bmatrix} 1 & e^{j\varphi_0} & \cdots & e^{j(M-1)\varphi_0} \end{bmatrix}^{\mathrm{T}} \tag{17.43}$$

where

$$\varphi_0 = -\beta d \cos\phi_0 = -\frac{\omega}{c} d \cos\phi_0 \tag{17.44}$$

is known as the *electrical angle.* By inserting (17.43) into (17.42), and noting that $|\vec{a}_{\mathrm{ULA}}(\phi_0)|^2 = M$, we obtain $P_{\mathrm{CBF}}(\phi_0)$ the spatial analog of the classical periodogram in temporal time series analysis [48, 49]. The spatial spectrum shares the same resolution limitation as the periodogram.

Two of the more important parameters for a ULA are its 3 dB beamwidth $\Delta\varphi_{3\mathrm{dB}}$, or, in angle space $\Delta\phi_{3\mathrm{dB}}$, and its first (highest) sidelobe level shown as A_{S}

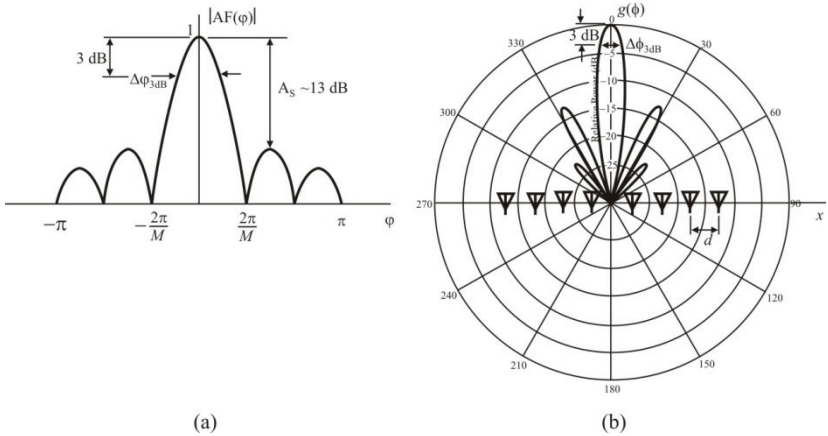


Figure 17.6(a, b) ULA main lobe width and first sidelobe level when $M = 8$: (a) rectangular plot and (b) polar plot.

in Figure 17.6. The patterns in Figure 17.6 are based on an eight-element array ($M = 8$). For M larger than ~5 or 6, A_S is relatively independent of M and $A_S \sim 13$ dB as seen in Figure 17.6(a). The beamwidth in φ-space at the half-power (–3 dB) level is

$$\Delta\varphi_{3\text{dB}} \approx 0.886\frac{2\pi}{M} \tag{17.45}$$

The first nulls shown in Figure 17.6 around the main lobe are at $\varphi_1 = \pm 2\pi / M$. If we differentiate $\varphi = \beta d \cos\phi$, that is $d\varphi = (\partial\varphi / \partial\phi)d\phi = -\beta d \sin\phi d\phi$, at broadside $(\varphi = 0)$ and assuming the main lobe is sufficiently narrow, we get

$$\Delta\varphi_{3\text{dB}} = \left|\frac{\partial\varphi}{\partial\phi}\right| \Delta\phi_{3\text{dB}} = \beta d \Delta\phi_{3\text{dB}} \tag{17.46}$$

From (17.45) and (17.46) we obtain

$$\Delta\phi_{3\text{dB}} = 0.886\frac{\lambda}{Md} \tag{17.47}$$

The quantity $A_e = Md$ is referred to as the *effective aperture* of the ULA. Equation (17.47) is the Rayleigh limit on the resolution of an optical system,

which says that the angular resolution that can be achieved by an aperture of length A_e,is essentially λ / A_e. Sources whose electrical angles are closer than $\Delta\phi_{3dB}$ will not be resolved by the conventional beamformer, regardless of the available data quality. This point is illustrated in Figure 17.7, where the beamforming spectrum is plotted versus the DOA in two different scenarios. A ULA of $M = 10$ antennas of $d = \lambda/2$ inter-element spacing[7] is used to separate two uncorrelated emitters, based on a batch of $N = 100$ data samples. The SNR for both sources is 0 dB. The beamwidth for such an array is $\Delta\phi_{3dB} = 0.886\lambda / (10\lambda / 2) \sim 0.1772$ radians (10^o), implying that sources need to be at least 10^o apart in order to be separated by the beamformer. This is also verified in the figure, because for 10^o separation, the sources are nearly (but not quite) resolved.

17.5 Standard Capon Beamformer

We will discuss a data-dependent beamformer that was first derived by Capon [39] in this section. The array response is steered to the directions which result in minimum power output of the array, subject to conditions. This steering is accomplished by forming a linear combination of the antenna outputs

$$y(t) = \sum_{m=0}^{M-1} w_m^* x_m(t) = \vec{w}^{\mathrm{H}} \vec{x}(t) \tag{17.48}$$

Given samples $y(1), y(2), ..., y(N)$, the output power is given by (17.33) [17].

The weight vector can be found from the maximum of the SINR [6]

$$\mathrm{SINR} = \sigma_s^2 \frac{\left|\vec{w}^{\mathrm{H}} \vec{a}\right|^2}{\vec{w}^{\mathrm{H}} \mathbf{R}_{\mathrm{i+n}} \vec{w}} \tag{17.49}$$

where

$$\mathbf{R}_{\mathrm{i+n}} = \mathcal{E}\left\{[\vec{i}(k) + \vec{n}(k)][\vec{i}(k) + \vec{n}(k)]^{\mathrm{H}}\right\} \tag{17.50}$$

is the $M \times M$ interference-plus-noise covariance matrix, and σ_s^2 is the signal power. From (17.49) it is clear that to maximize the output SINR, the response to the interference and noise, given by $\vec{w}^{\mathrm{H}} \mathbf{R}_{\mathrm{i+n}} \vec{w}$, should be minimized. The solution for

[7] Such a ULA is termed a *standard ULA*, because $d = n/k$ is the maximum allowable element separation to avoid ambiguities.

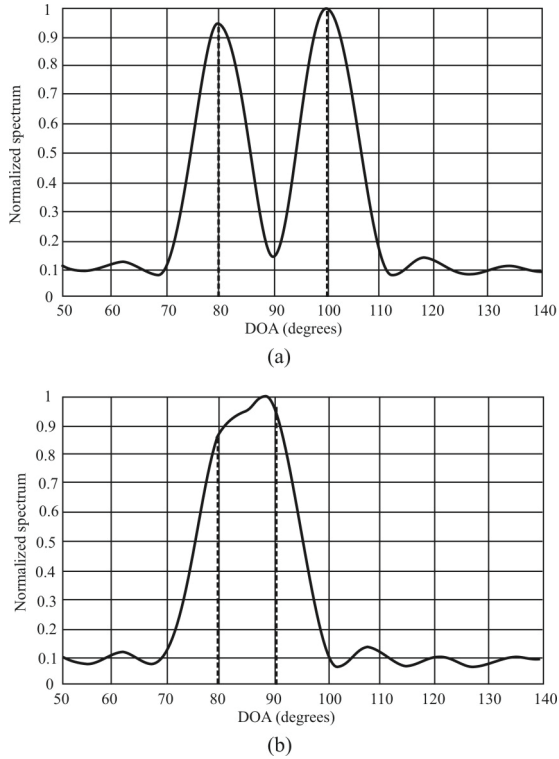


Figure 17.7(a, b) Beamforming spectra: (a) 20° separation, greater than resolution limit, and (b) 10° separation, less than the resolution limit.

the weight vector is found by maintaining a distortionless response toward the desired signal while minimizing the output interference-plus-noise power.

Hence, the maximization of (17.49) is equivalent to

$$\min_{\vec{w}} P(\vec{w}) \qquad \text{subject to} \qquad \vec{w}^{\mathrm{H}}\vec{a} = 1 \tag{17.51}$$

From (17.51), the following well-known solution can be found for the optimal weight vector:

$$\vec{w}_{\mathrm{opt}} = \alpha \mathbf{R}_{\mathrm{i+n}}^{-1}\vec{a} \tag{17.52}$$

where $\alpha = (\vec{a}^{\mathrm{H}}\hat{\mathbf{R}}^{-1}\vec{a})^{-1}$ is a normalization constant that does not affect the output SINR (17.49) and therefore will be omitted in the following. The solution to (17.52) is commonly referred to as the *minimum variance distortionless response* (MVDR) beamformer [50] or the *standard Capon beamformer* (SCB).

However, in general, the exact interference-plus-noise covariance matrix is unavailable. Therefore, the sample covariance matrix (17.35) is used instead of (17.50). Here N is the number of training snapshots (also termed the *training sample size*). In this case, (17.51) should be rewritten as

$$\min_{\vec{w}} \vec{w}^{\mathrm{H}}\hat{\mathbf{R}}_{\mathrm{i+n}}\vec{w} \qquad \text{subject to} \qquad \vec{w}^{\mathrm{H}}\vec{a} = 1 \tag{17.53}$$

The solution to this problem is commonly referred to as the *sample matrix inversion* (SMI) algorithm, whose optimum weight vector, optimum in the sense of maximizing the output SINR [after omitting the irrelevant constant $\alpha = 1/(\vec{a}^{\mathrm{H}}\mathbf{R}_{\mathrm{i+n}}^{-1}\vec{a})$], in a stationary environment is given by the Weiner equation (see Appendix 17A)

$$\vec{w}_{\mathrm{SMI}} = \hat{\mathbf{R}}^{-1}\vec{a} \tag{17.54}$$

As mentioned, when the SOI component is present in the collected data, the use of the sample covariance matrix (17.35) in place of the true interference-plus-noise covariance matrix (17.50) affects the performance of the SMI algorithm dramatically. It is well known that, in the case of when the training data is SOI-free, the use of weight vector (17.54) provides rapid convergence of the output SINR to its optimal value given by

$$\mathrm{SINR}_{\mathrm{opt}} = \sigma_s^2 \vec{a}^{\mathrm{H}}\hat{\mathbf{R}}_{\mathrm{i+n}}^{-1}\vec{a} \tag{17.55}$$

so that the average performance losses relative to (17.55) are less than 3 dB if $N \geq 2M$. However, this may no longer be true if the training data frames are contaminated by the SOI component. It was shown in [24] that in the latter case the convergence to (17.55) becomes much slower and generally requires $N >> M$.

Another shortcoming of the SMI algorithm is that it is not robust against a mismatch between the presumed and actual signal steering vectors, $\bar{\bar{a}}$ and $\vec{a}$. Here $\bar{\bar{a}}$ denotes the calibrated (estimated) steering vector, while $\vec{a}$ is the true steering vector that characterizes the spatial signature of the signal. In the mismatched case

$$\bar{\bar{a}} = \vec{a} + \vec{\Delta} \neq \vec{a} \tag{17.56}$$

where $\vec{\Delta}$ is an unknown complex vector that describes the effect of steering vector mismatches. As a result, the SMI beamformer tends to interpret the signal components in array observations as an interfering signal and tries to suppress these components the same way that true interferers are suppressed—by means of adaptive nulling instead of maintaining distortionless response toward $\vec{a}$ (see [23, 33]).

In the mismatched case, (17.49) and (17.55) should be rewritten as

$$\text{SINR} = \sigma_s^2 \frac{\left|\vec{w}^{\text{H}} \bar{\bar{a}}\right|^2}{\vec{w}^{\text{H}} \hat{\mathbf{R}}_{\text{i+n}} \vec{w}} \tag{17.57}$$

and

$$\text{SINR}_{\text{opt}} = \sigma_s^2 \bar{\bar{a}}^{\text{H}} \hat{\mathbf{R}}_{\text{i+n}}^{-1} \bar{\bar{a}} \tag{17.58}$$

respectively.

From (17.54) it is clear that inversion of $\hat{\mathbf{R}}$ is required in order to compute the optimum weights. This can best be done recursively, since determining a matrix inverse is computationally expensive. The recursive solution is given by the *matrix inversion lemma* as [51]

$$\hat{\mathbf{R}}^{-1}(k) = \hat{\mathbf{R}}^{-1}(k-1) - \frac{\hat{\mathbf{R}}^{-1}(k-1)\vec{x}(k)\vec{x}^{\text{H}}(k)\hat{\mathbf{R}}^{-1}(k-1)}{1 + \vec{x}^{\text{H}}(k)\hat{\mathbf{R}}^{-1}(k-1)\vec{x}(k)} \tag{17.59}$$

with

$$\hat{\mathbf{R}}^{-1}(0) = \frac{1}{\varepsilon_0}\mathbf{I} \tag{17.60}$$

Updating the autocorrelation matrix can also be made to be more computationally efficient than recomputing the average with each new sample. This can be done with

$$\hat{\mathbf{R}}(k+1) = \frac{k\hat{\mathbf{R}}(k) + \vec{x}(k+1)\vec{x}^{\mathrm{H}}(k+1)}{k+1} \tag{17.61}$$

Numerous modifications of the conventional beamformer have been developed to alleviate the limitations, such as its resolving power of two sources spaced closer than a beamwidth. One of those was developed by Capon [39] which was posed as

$$\begin{aligned} &\underset{\vec{w}}{\text{minimize}} && P(\vec{w}) \\ &\text{subject to} && \vec{w}^{\mathrm{H}}\vec{a}(\phi_0) = 1 \end{aligned} \tag{17.62}$$

where $P(\vec{w})$ is defined in (17.33). The *standard Capon beamformer* (SCB), also with the appellation MVDR filter, attempts to minimize the power contributed by noise and any signals coming from directions other than ϕ_0, while maintaining a fixed gain in the "look direction," ϕ**.** The optimal $\vec{w}$ can be found using Lagrange multipliers [52, 53] (see Appendix 17A), resulting in

$$\vec{w}_{\mathrm{CAP}} = \frac{\hat{\mathbf{R}}^{-1}\vec{a}(\phi_0)}{\vec{a}^{\mathrm{H}}(\phi_0)\hat{\mathbf{R}}^{-1}\vec{a}(\phi_0)} \tag{17.63}$$

Inserting (17.63) into (17.33) leads to the "spatial spectrum"

$$P_{\mathrm{CAP}} = \frac{1}{\vec{a}^{\mathrm{H}}(\phi_0)\hat{\mathbf{R}}^{-1}\vec{a}(\phi_0)} \tag{17.64}$$

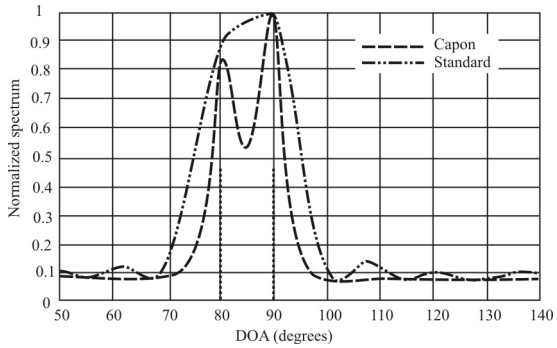


Figure 17.8 Comparison of standard and Capon. ULA, 10 antennas, lambda/2 apart. Targets 10º apart.

Capon's beamformer outperforms the classical one given in (17.42) because the former uses every available degree of freedom to concentrate the received energy along one direction, namely the AOA of interest. This is reflected by the constraint given in (17.62). The resolution capability of the Capon beamformer is still dependent upon the array aperture and clearly on the SNR, however. A comparison of the resolution of the Capon beamformer and the standard beamformer for a ULA is illustrated in Figure 17.8.

The SCB has better resolution and much better interference rejection capability than the standard beamformer, provided that the presumed array steering vector of the SOI match the actual array steering vector precisely [54]. In practice, the exact steering vector of the SOI is unavailable or its measure/estimation is imprecise; therefore, we use $\bar{\vec{a}}$ instead of the actual $\vec{a}(\phi_0)$ in the Capon beamformer, and the mismatch between the exact steering vector and the presumed one may drastically degrade the performance of the Capon beamformer. An example of this degradation is illustrated in Figure 17.9 [55]. We can see that a major shift in the SINR (19.70 to –5.10 dB) occurs for only a minor mismatch difference in the calibrated steering vector and the measured response.

We can quantify the impact of the steering vector mismatch as follows. The array output SINR can be written as

$$\text{SINR} = \frac{\mathcal{E}\left\{\left|\vec{w}_{\text{CAP}}^{\text{H}} s_0(t)\right|^2\right\}}{\vec{w}_{\text{CAP}}^{\text{H}} \mathbf{R}_{\text{i+n}} \vec{w}_{\text{CAP}}} = \sigma_0^2 \frac{\left|\vec{w}_{\text{CAP}}^{\text{H}} \vec{a}(\phi_0)\right|^2}{\vec{w}_{\text{CAP}}^{\text{H}} \left[\sum_{l=1}^{L-1} \sigma_l^2 \vec{a}(\phi_l)\vec{a}^{\text{H}}(\phi_l) + \mathbf{R}_{\text{n}}\right] \vec{w}_{\text{CAP}}} \tag{17.65}$$

where $\sigma_0^2 = \mathcal{E}\{|s_0(t)|\}$. Inserting (17.52) into (17.65) yields

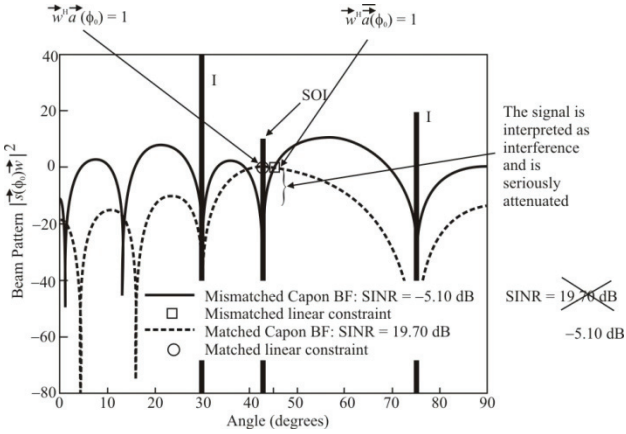


Figure 17.9 Mismatched steering vectors. I: interferers.

$$\text{SINR} = \sigma_0^2 \frac{\left|\bar{\vec{a}}^{\text{H}}(\phi_0)\mathbf{R}_{\text{i+n}}^{-1}\vec{a}(\phi_0)\right|^2}{\bar{\vec{a}}^{\text{H}}(\phi_0)\mathbf{R}_{\text{i+n}}^{-1}\bar{\vec{a}}(\phi_0)} \tag{17.66}$$

where $\vec{a}(\phi_0)$ is the actual steering vector, then (17.66) can be rewritten as:

$$\text{SINR} = \sigma_0^2\vec{a}^{\text{H}}(\phi_0)\mathbf{R}_{\text{i+n}}^{-1}\vec{a}(\phi_0)\times\frac{\left|\vec{a}^{\text{H}}(\phi_0)\mathbf{R}_{\text{i+n}}^{-1}\bar{\vec{a}}(\phi_0)\right|^2}{[\vec{a}^{\text{H}}(\phi_0)\mathbf{R}_{\text{i+n}}^{-1}\vec{a}(\phi_0)][\bar{\vec{a}}^{\text{H}}(\phi_0)\mathbf{R}_{\text{i+n}}^{-1}\bar{\vec{a}}(\phi_0)]} \tag{17.67}$$

$$= \text{SNIR}_{\text{m}}\cos^2[\vec{a}(\phi_0),\bar{\vec{a}}(\phi_0);\mathbf{R}_{\text{i+n}}^{-1}] \tag{17.68}$$

where $\text{SINR}_{\text{m}} = \sigma_0^2\vec{a}^{\text{H}}(\phi_0)\mathbf{R}_{\text{i+n}}^{-1}\vec{a}(\phi_0)$ and $\cos^2(\cdot)$ is

$$\cos^2(\vec{a},\vec{b};\mathbf{Z}) = \frac{\left|\vec{a}^*\mathbf{Z}\vec{b}\right|^2}{(\vec{a}^{\text{H}}\mathbf{Z}\vec{a})(\vec{b}^{\text{H}}\mathbf{Z}\vec{b})} \tag{17.69}$$

Since $0 \le \cos^2(\vec{a},\vec{b};\mathbf{Z}) \le 1$, the array output SINR is reduced due to mismatch between the presumed steering vector of the SOI and its true value.

17.5.1 Discussion

Since $\vec{a}(\phi_0)$ is usually close to $\bar{\vec{a}}$ (ideally it is identical to it), the Euclidean norm of the resulting weight vector (which equals the white noise gain at the array output)

can become very large in order to satisfy the distortionless constraint $\vec{w}^{\mathrm{H}}\vec{\bar{a}} = 1$ and at the same time cancels the SOI (i.e., $\vec{w}^{\mathrm{H}}\vec{a}(\phi_0) = 1$; note that the previous two conditions on $\vec{w}$ imply $\vec{w}^{\mathrm{H}}[\vec{a} - \vec{a}(\phi_0)] \approx 1$, which can hold only if $\|\vec{w}\|^2 >> 1$ whenever $\|\vec{\bar{a}} - \vec{a}(\phi_0)\|^2$ is relatively small).

17.6 Robust Capon Beamformers

Several techniques have been investigated and reported to address the most significant problem identified for the Capon beamformer, which is sensitivity to errors in the array manifold. These errors arise because of calibration errors, multipath (coherency), inaccurate estimates of the location for the antennas, local scattering, near-far signature mismatch, differences between the measured signal and the calibration signal, source spreading, and distorted antenna shape.

We discuss first two techniques for making the SCB more robust against the mismatch errors. The first is to load the diagonal elements of the correlation matrix. The second is to use eigenvalue decomposition of the correlation matrix.

17.6.1 Diagonal Loading

Diagonal loading (DL) is one way to enhance the robustness of the SBF to array manifold mismatches. The DL approach yields the *loaded SMI* (LSMI) algorithm, which is based on the diagonal loading of the sample covariance matrix [22, 34]. The essence of this approach is to replace the conventional sample covariance matrix by the so-called diagonally loaded covariance matrix

$$\hat{\mathbf{R}}_{\mathrm{dl}} = \hat{\mathbf{R}} + \xi\mathbf{I} \tag{17.70}$$

in the SMI algorithm [8]. Here, ξ is a diagonal loading factor, and $\mathbf{I}$ is the identity matrix. Using (17.70), we can write the LSMI weight vector in the following form [22]:

$$\vec{w}_{\mathrm{LSMI}} = \hat{\mathbf{R}}_{\mathrm{dl}}^{-1}\vec{a} = (\hat{\mathbf{R}} + \xi\mathbf{I})^{-1}\vec{a} \tag{17.71}$$

The main issue with the LSMI method is how to choose the diagonal loading factor ξ. There are two fundamental different ways to make this selection. It can be a fixed value [yielding the *fixed diagonal loading* (FDL) algorithm], or it can be allowed to vary and determined by some algorithm.

Cox et al. [34] proposed to use the so-called *white noise gain constraint* to obtain reasonable values of this parameter. Unfortunately, it is not clear how to

relate the parameters of the white noise gain constraint and the level of uncertainty of the signal steering vector. Furthermore, the relationship between the diagonal loading factor and the parameters of the white noise gain constraint is not simple, and to satisfy this constraint, a multistep, iterative procedure is required to adjust the diagonal loading factor [34]. Each step of this iterative procedure involves an update of the inverse of the diagonally loaded covariance matrix, and as a result, the total computational complexity of adaptive beamforming with the white noise gain constraint may be higher than that of the SMI algorithm. Because of this, the diagonal loading factor is usually chosen in a more ad hoc way, typically about $10\sigma^2$, where σ^2 is the noise power from a single antenna.

We present an approach in the next section that belongs to the general set of DL algorithms that addresses the problem identified here. The technique incorporates the selection as part of the algorithm. Here we present an approach to the selection problem posited by Du [55] that includes calculation of the loading factors automatically as part of the algorithm.

17.6.2 General Linear Combination–Based Robust Capon Beamforming

This section is extracted from [56] with permission. It is shown in [56] how this algorithm can be used to detect and geolocate CDMA signals. It is included here for completeness.

As mentioned, the sample covariance matrix $\hat{\mathbf{R}}$ can be a poor estimate of the true covariance matrix $\mathbf{R}$ when N is small, relative to the array dimension M. An improved estimate of $\mathbf{R}$ is provided by the GLC-based covariance matrix estimator [24]. This estimator uses the sample covariance matrix $\hat{\mathbf{R}}$ and the identity matrix $\mathbf{I}$ to obtain a more accurate estimate of $\mathbf{R}$ than $\hat{\mathbf{R}}$ and is given by

$$\tilde{\mathbf{R}} = \gamma\hat{\mathbf{R}} + \alpha\mathbf{I} \tag{17.72}$$

where $\tilde{\mathbf{R}}$, which should be positive semidefinite $(\tilde{\mathbf{R}} \geq 0)$, is the enhanced estimate of $\mathbf{R}$. α and γ in (17.72) are chosen by minimizing (an estimate of) the MSE of $\tilde{\mathbf{R}}$, where

$$\text{MSE}\{\tilde{\mathbf{R}}\} = \mathcal{E}\{\left\|\tilde{\mathbf{R}} - \mathbf{R}\right\|\} \tag{17.73}$$

α and γ can be constrained such that $\alpha \geq 0$ and $\gamma \geq 0$ so that $\tilde{\mathbf{R}} \geq 0$ can be guaranteed.

The values of α and γ that minimize an estimate of the MSE of $\tilde{\mathbf{R}}$ are computed in [24] with the results being that

$$\hat{\alpha}_0 = \min\left[\hat{\nu}\frac{\hat{\rho}}{\left\|\hat{\mathbf{R}} - \hat{\nu}\mathbf{I}\right\|^2}, \hat{\nu}\right] \tag{17.74}$$

and

$$\hat{\gamma}_0 = 1 - \frac{\hat{\alpha}_0}{\hat{\nu}} \tag{17.75}$$

where

$$\hat{\rho} = \frac{1}{N^2}\sum_{k=1}^{N}\left\|\vec{y}(k)\right\|^4 - \frac{1}{N}\left\|\hat{\mathbf{R}}\right\|^2 \tag{17.76}$$

and

$$\hat{\nu} = \frac{\mathrm{tr}\{\hat{\mathbf{R}}\}}{M} \tag{17.77}$$

with the observed data snapshots given by $\{\vec{y}(k)\}_{k=1}^{N}$.

17.6.2.1 The Beamformer

Based on (17.72)–(17.77), we have a diagonally-loaded enhanced estimate of the covariance matrix

$$\tilde{\mathbf{R}}_{\mathrm{GLC}} = \hat{\gamma}_0\hat{\mathbf{R}} + \hat{\alpha}_0\mathbf{I} \tag{17.78}$$

Using the enhanced estimate $\tilde{\mathbf{R}}_{\mathrm{GLC}}$ in lieu of $\hat{\mathbf{R}}$ in the SCB formulation yields the following *general linear combination* (GLC)-based robust adaptive beamformer

$$\vec{w}_{\mathrm{GLC}} = \frac{\tilde{\mathbf{R}}_{\mathrm{GLC}}^{-1}\vec{a}}{\vec{a}^{\mathrm{H}}\tilde{\mathbf{R}}_{\mathrm{GLC}}^{-1}\vec{a}} \tag{17.79}$$

Assuming that $\hat{\gamma}_0 \neq 0$, the estimate of (17.79) can be written as

$$\tilde{\vec{w}}_{\text{GLC}} = \frac{\left[\frac{\hat{\alpha}_0}{\hat{\gamma}_0}\mathbf{I} + \hat{\mathbf{R}}\right]^{-1}\vec{a}}{\vec{a}^{\text{H}}\left[\frac{\hat{\alpha}_0}{\hat{\gamma}_0}\mathbf{I} + \hat{\mathbf{R}}\right]^{-1}\vec{a}} \tag{17.80}$$

The output SINR of the resulting beamformer is given by

$$\text{SINR} = \frac{\sigma_0^2 \left|\vec{w}_{\text{GLC}}^{\text{H}}\vec{a}\right|^2}{\vec{w}_{\text{GLC}}^{\text{H}}\mathbf{R}_{\text{n+i}}\vec{w}_{\text{GLC}}} \tag{17.81}$$

and the SOI power estimate is

$$\hat{\sigma}_0^2 = \vec{w}_{\text{GLC}}^{\text{H}}\tilde{\mathbf{R}}_{\text{GLC}}\vec{w}_{\text{GLC}} = \frac{1}{\vec{a}^{\text{H}}\tilde{\mathbf{R}}_{\text{GLC}}\vec{a}} \tag{17.82}$$

An example of the performance of the GLC method when there are mismatch errors is shown in Figure 17.10 [47]. The optimal performance as well as the performance of the SCB are also shown. We see that the SINR degrades much more slowly than the SCB as the mismatch error increases.

Example [48]: We examine a performance comparison, as determined by the output SINR, as N increases for the SCB, FDL, and GLC. We also include the case when the true **R** is used (True R). For the FDL we set the fixed DL level to equal the noise power, which is assumed known. We assume a ULA with M =10 sensors and with $\lambda/2$ inter-element spacing. The far-field narrowband source waveforms and the additive noise are assumed to be temporally white, circularly symmetric, complex Gaussian random processes with zero-mean and variances according to the SNRs. The noise is further assumed to be spatially white, and its covariance matrix is the identity matrix **I**. Three sources with SNRs 10 dB, 20 dB, and 20 dB are assumed to be present at 0^o, 20^o, and 60^o, respectively. We assume that there are no errors in the array steering vector. The SOI is the first (10 dB) signal and the other two sources are interferers. For each scenario 1,000 Monte Carlo trials were performed. Figure 17.11 [55] shows the mean of the output SINRs versus the number of snapshots N. As shown, the SCB converges to the

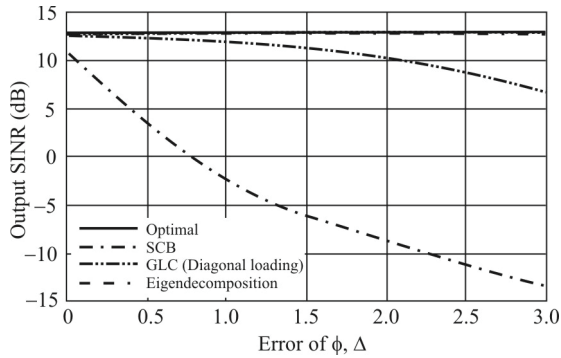


Figure 17.10 Output SINR comparison of beamforming methods. (SNR = 0 dB, INR = 20 dB, SOI at 30^{o}, $M = 10$, $N = 50$.)

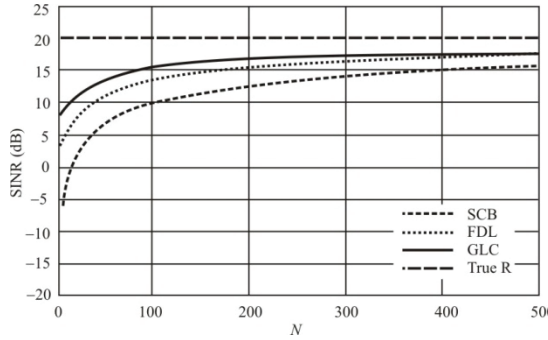


Figure 17.11 Beamformer output SINR versus snapshot number N in absence of array steering vector errors.

optimal SINR value very slowly, since $\hat{\mathbf{R}}$ contains the SOI and the SOI power is not small. GLC converges to the optimal value faster than all the other algorithms. FDL gives the second best performance.

17.6.2.2 Summary

In this section we have described a GLC-based robust adaptive beamformer, where the sample covariance matrix $\hat{\mathbf{R}}$, used in the SCB formulation, is replaced by an enhanced covariance matrix estimate $\tilde{\mathbf{R}}$ given by (17.72). GLC is a robust DL approach where the loading elements need not be fixed or the same. We can efficiently obtain the enhanced covariance matrix estimate completely automatically (i.e., without specifying any user parameter). Numerical examples demonstrated the performance of GLC compared with that of the SCB and FDL.

17.6.3 Eigendecomposition Method

The most robust RCB method illustrated in Figure 17.10 is based on eigendecomposition of the covariance matrix.[8] As before, there are two (mathematical) subspaces associated with **R**: the signal subspace and the noise subspace. The basis vectors of the signal subspace are orthogonal to the noise subspace and the basis vectors of the noise subspace are orthogonal to the signal subspace. The covariance matrix can be expressed as

$$\mathbf{R} = \mathbf{U}_s \mathbf{\Sigma}_s \mathbf{U}_s^H + \mathbf{U}_n \mathbf{\Sigma}_n \mathbf{U}_n^H \tag{17.83}$$

where $\mathbf{U}_s$ is a matrix representing the signal subspace and contains the steering vectors, $\vec{a}(\theta)$, corresponding to the AOAs of the signals; $\mathbf{\Sigma}_s$ is a diagonal matrix containing the eigenvalues of the signal subspace; $\mathbf{U}_n$ is a matrix that represents the noise subspace and contains the remaining steering vectors, at angles where no signals are arriving; and $\mathbf{\Sigma}_n$ is a diagonal matrix containing the eigenvalues of the noise subspace.

---

[8] The *eigenvalue decomposition*, also known as the *spectral decomposition* and *eigendecomposition*, of a square matrix **A** is given by

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$$

where $\mathbf{\Lambda}$ is a diagonal matrix where each diagonal entry is an eigenvalue of **A**, and the columns of **U** are the corresponding eigenvectors of **A**. If the eigenvalues are all normalized, **A** is square and positive semi-definite, as all correlation matrices are, and then $\mathbf{U}^{-1} = \mathbf{U}^H$ and the eigenvalue decomposition is also given by

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$$

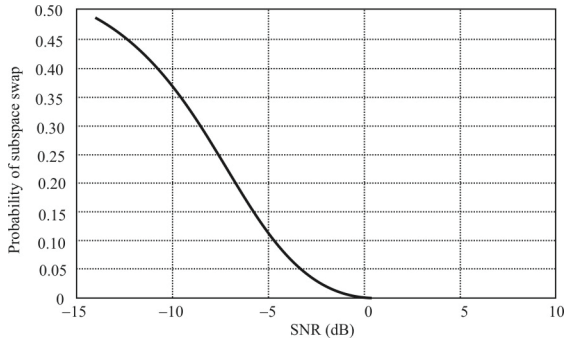


Figure 17.12 Probability of subspace swap.

First, we estimate the value of the covariance function. Then the eigendecomposition of $\hat{\mathbf{R}}$, given by (17.83), is found (there are standard software packages available, for example, in MATLAB and MathCad, for this function). The angular spectrum, given by

$$P(\theta) = \frac{1}{\left|\vec{a}(\theta)^{\mathrm{H}} \mathbf{U}_{\mathrm{n}}\right|^2} \tag{17.84}$$

is then calculated. Since the $\vec{a}(\theta)$ are orthogonal to $\mathbf{U}_{\mathrm{n}}$ (the noise subspace eigenvectors) at the values of θ corresponding to the AOAs of the signals, $\{\theta_k\}$, then $\left|\vec{a}(\theta_k)^{\mathrm{H}} \mathbf{U}_{\mathrm{n}}\right| \to 0$ and $P(\theta_k) \to \infty$, or at least peaks at these values of θ. The example results of this method are also shown in Figure 17.10. There is essentially no degradation of SINR as the error is increased, at least for the parameters corresponding to this example.

A drawback of the eigendecomposition method is the requirement to frequently compute the eigendecomposition of $\hat{\mathbf{R}}$. Such an operation requires $O(L^3)$ floating point computations. Methods have been devised, however, to avoid these computations.

It has been observed that RCB eigendecomposition techniques for spatial filtering tend to perform poorly at low SNRs. This is due primarily to the probability of a *subspace swap*, that is, the switching of the noise subspace with the signal subspace, as the value of the signal eigenvalues approach the value of the noise eigenvalues. An analysis was performed by Thomas, Scharf, and Tufts on the susceptibility of the approach to such swapping. Some of those results are shown in Figure 17.12 [57]. As can be seen, good performance is maintained down to around 0 dB SNR before significant problems occur.

G_p for WCDMA is about 22 dB, as is the processing gain for the other 3G+ standards. This is the maximum processing gain. In the worst case, if the post-detect SNR required is 10 dB (a reasonable goal) then the predecorrelation SNR that is required is –12 dB, which puts the probability of subspace swap at about 0.4. If the spreading code is unknown at the intercept receiver, then about 40% of the time there will be a subspace swap. With the *orthogonal variable spreading factor* (OVSF) coding in the standards, G_p can be as low as 6 dB (4 chips per symbol). With a coding gain of 6 dB, the predecorrelation SNR required is 4 dB, and the probability of subspace swap is low (essentially zero). If the code can be determined, as discussed in Chapter 9 of [58], then the postdecorrelation SNR is about 10 dB, and essentially no swapping will occur in either case.

17.6.4 Subspace-Based Beamformer

A different RCB based on eigendecomposition is presented in this section. As above, robust adaptive beamforming in the case of a steering vector mismatch is accomplished using the eigenspace-based beamformer [24, 37] where the idea is to use, instead of the presumed steering vector $\bar{\bar{a}}$, the projection of $\bar{\bar{a}}$ onto the sample signal-plus-interference subspace.

The eigendecomposition of (17.35) yields

$$\hat{\mathbf{R}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{H}} + \mathbf{V}\boldsymbol{\Gamma}\mathbf{V}^{\mathrm{H}} \tag{17.85}$$

where the matrix $\mathbf{U} \in \mathbb{C}^{M\times(L+1)}$ contains the $L + 1$ signal-plus-interference subspace eigenvectors of $\hat{\mathbf{R}}$, and the diagonal matrix $\boldsymbol{\Lambda} \in \mathbb{C}^{(L+1)\times(L+1)}$ contains the corresponding eigenvalues of $\hat{\mathbf{R}}$. Similarly, the matrix $\mathbf{V} \in \mathbb{C}^{M\times(M-L-1)}$ contains the $M–L–1$ noise-subspace eigenvectors of $\hat{\mathbf{R}}$, whereas the diagonal matrix $\boldsymbol{\Gamma} \in \mathbb{C}^{(M-L-1)\times(M-L-1)}$ is built from the corresponding eigenvalues. Here, L is the number of interfering sources (or, mathematically, the rank of the interference subspace), which is assumed to be known.[9] The weight vector of the eigenspace-based beamformer is given by

$$\vec{w}_{\mathrm{eig}} = \hat{\mathbf{R}}\vec{v} \tag{17.86}$$

where

[9] If the number of interfering signals is not known there are methods of estimating the quantity, that is, setting an eigenvalue threshold and assuming that all eigenvalues above that threshold belong to signals. The number of these eigenvalues is the estimated quantity.

$$\vec{v} = \mathbf{P}_{\mathrm{E}}\bar{\vec{a}}, \tag{17.87}$$

and

$$\mathbf{P}_{\mathrm{E}} = \mathbf{U}\mathbf{U}^{\mathrm{H}} \tag{17.88}$$

are the projected steering vector and the orthogonal projection matrix onto the signal-plus-interference subspace, respectively. Inserting (17.87) and (17.88) into (17.86), then the optimum weights can be rewritten as

$$\vec{w}_{\mathrm{eig}} = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^{\mathrm{H}}\bar{\vec{a}} \tag{17.89}$$

The eigenspace-based beamformer is known to be one of the most powerful robust techniques applicable to arbitrary steering vector mismatches [37]. However, an essential shortcoming of this approach is that it is limited to high SNR[10] cases because at low SNR the estimation of the projection matrix onto the signal-plus-interference subspace breaks down due to a high probability of subspace swaps [57]. Moreover, the eigenspace-based beamformer is efficient only if the dimension of the signal-plus-interference subspace is low and known exactly.

17.6.4.1 Subspace Decomposition Using Uncertainty Ellipse Capon Beamforming

Another method of dealing with steering vector mismatches is to recognize that the exact steering vector is unknown, but is known to belong to a (possibly infinite) set of steering vectors. We demonstrate the method of subspace decomposition in this section where the only known characteristic of the steering vector is that it lies within the range of an uncertainty ellipse. The derivation here follows [47] closely and demonstrates the use of the Lagrange multiplier technique.

We attempt to improve the array output SINR using a (potentially) imprecise steering vector $\vec{a}(\phi_0)$. We assume that $\vec{a}(\phi_0)$ belongs in the range defined by the uncertainty ellipse [23]

$$[\vec{a}(\phi_0) - \bar{\vec{a}}(\phi_0)]^{\mathrm{H}}\mathbf{C}^{-1}[\vec{a}(\phi_0) - \bar{\vec{a}}(\phi_0)] \le 1 \tag{17.90}$$

where **C** is a specified positive definite matrix.

The actual steering vector of the SOI is orthogonal to the noise subspace and we will project the signal steering vector onto the noise subspace. The steering

[10] High SNR can be as low as 0 dB [51]. See Figure 17.12.

vector is normed as $\|\vec{a}(\phi_0)\|^2 = \vec{a}^*(\phi_0)\vec{a}(\phi_0) = M$. We use the following constrained optimization

$$\begin{aligned} &\hat{\vec{w}} = \underset{\vec{a}}{\text{argmin}}\, \vec{a}^{\mathrm{H}}(\phi_0)\mathbf{V}\mathbf{V}^{\mathrm{H}}\vec{a}(\phi_0) \\ &\text{such that } [\vec{a}(\phi_0) - \bar{\vec{a}}(\phi_0)]^{\mathrm{H}}\mathbf{C}^{-1}[\vec{a}(\phi_0) - \bar{\vec{a}}(\phi_0)] \le 1 \text{ and } \|\vec{a}(\phi_0)\|^2 = M \end{aligned} \tag{17.91}$$

where $\bar{\vec{a}}(\phi_0)$ is the calibrated array steering vector corresponding to ϕ_0, but with probable errors (mismatch to the actual steering vector of the SOI). **V** is the noise subspace, which is obtained by the eigendecomposition of $\hat{\mathbf{R}}$ given by (17.35). As with other subspace techniques, we assume that the number of plane waves impinging on the array, *L*, is known a priori. This assumption is removed later. Note that we can improve the estimation accuracy of the actual steering vector of the SOI from (17.91), and then obtain optimal weight $\vec{w}_0$ by the Capon method.

We solve (17.91) for the case in which $\mathbf{C} = \varepsilon\mathbf{I}$ (ε is a user parameter), and then (17.91) becomes

$$\begin{aligned} &\hat{\vec{w}} = \underset{\vec{a}}{\text{argmin}}\ \vec{a}^{\mathrm{H}}(\phi)\mathbf{V}\mathbf{V}^{\mathrm{H}}\vec{a}(\phi) \\ &\text{such that } \left\|[\vec{a}(\phi) - \bar{\vec{a}}(\phi)]\right\|^2 \le \varepsilon \text{ and } \|\vec{a}(\phi)\|^2 = M \end{aligned} \tag{17.92}$$

We use the Lagrange multiplier methodology, which is based on the function

$$\begin{aligned} g[\vec{a}(\phi), \lambda, \mu] = {} & \vec{a}^{\mathrm{H}}(\phi)\mathbf{V}\mathbf{V}^{\mathrm{H}}\vec{a}(\phi) \\ & + \mu[2N - \varepsilon - \bar{\vec{a}}^{\mathrm{H}}(\phi)\vec{a}(\phi) - \vec{a}^{\mathrm{H}}(\phi)\bar{\vec{a}}(\phi)] + \lambda[\vec{a}^{\mathrm{H}}(\phi)\vec{a}(\phi) - M] \end{aligned} \tag{17.93}$$

where $\mu \ge 0, \lambda \ge 0$ are the Lagrange multipliers. This technique adds the two constraint functions in (17.92) that evaluate to zero as the last two terms in (17.93)—this is the heart of the Lagrange multiplier technique.

Now, the unconstrained minimization of (17.93) for fixed μ, λ, is given by

$$\frac{\partial g[\vec{a}(\phi), \mu, \lambda]}{\partial \vec{a}} = 2\mathbf{V}\mathbf{V}^{\mathrm{H}}\vec{a}(\phi) - 2\mu\bar{\vec{a}}(\phi) + 2\vec{a}(\phi) = 0 \tag{17.94}$$

Palpably, the optimal solution for $\vec{a}$ is

$$\hat{\vec{a}}(\phi) = \mu(\mathbf{V}\mathbf{V}^{\mathrm{H}} + \lambda\mathbf{I})^{-1}\bar{\vec{a}}(\phi) \tag{17.95}$$

Inserting $\hat{\bar{a}}(\phi)$ into (17.93), minimizing $g(\bar{a},\mu,\lambda)$ with respect to μ gives

$$\frac{\partial g[\hat{\bar{a}}(\phi),\mu,\lambda]}{\partial \mu} = 2M - \varepsilon \bar{\bar{a}}^{\mathrm{H}}(\phi)\hat{\bar{a}}(\phi) - \hat{\bar{a}}^{\mathrm{H}}(\phi)\bar{\bar{a}}(\phi) = 0 \tag{17.96}$$

Then we obtain

$$\hat{\mu} = \frac{2N - \varepsilon}{2\bar{\bar{a}}^{\mathrm{H}}(\phi)(\mathbf{V}\mathbf{V}^{\mathrm{H}} + \lambda \mathbf{I})^{-1}\bar{\bar{a}}(\phi)} \tag{17.97}$$

Inserting $\hat{\mu}$ into (17.93) and minimizing the Lagrange function with respect to λ yield

$$\frac{\partial g[\hat{\bar{a}}(\phi),\mu,\lambda]}{\partial \lambda} = \hat{\bar{a}}^{\mathrm{H}}(\phi)\hat{\bar{a}}(\phi) - M = 0 \tag{17.98}$$

and the following ensues,

$$\frac{\bar{\bar{a}}^{\mathrm{H}}(\phi)(\mathbf{V}\mathbf{V}^{\mathrm{H}} + \hat{\lambda}\mathbf{I})^{-2}\hat{\bar{a}}(\phi)}{\left[\bar{\bar{a}}^{\mathrm{H}}(\phi)(\mathbf{V}\mathbf{V}^{\mathrm{H}} + \hat{\lambda}\mathbf{I})^{-1}\bar{\bar{a}}(\phi)\right]^2} = \frac{M}{\left(M - \varepsilon/2\right)^2} \tag{17.99}$$

Then the solution for $\hat{\lambda}$ can be obtained from (17.99) by some simple manipulations.

Substituting (17.97) into (17.95) yields

$$\hat{\bar{a}}(\phi) = \left(M - \frac{\varepsilon}{2}\right)\frac{(\mathbf{V}\mathbf{V}^{\mathrm{H}} + \hat{\lambda}\mathbf{I})^{-1}\bar{\bar{a}}(\phi)}{\bar{\bar{a}}^{\mathrm{H}}(\phi)(\mathbf{V}\mathbf{V}^{\mathrm{H}} + \hat{\lambda}\mathbf{I})^{-1}\bar{\bar{a}}(\phi)} \tag{17.100}$$

To summarize, this RCB consists of following steps.

A Subspace RCB Algorithm

1. Calculate data covariance matrix (17.35).
2. Compute the eigendecomposition of $\hat{\mathbf{R}}$ and obtain the noise subspace $\mathbf{V}$.
3. Solve for $\hat{\lambda}$ in (17.99).
4. Use $\hat{\lambda}$ from step 3 to calculate

$$\hat{\vec{a}}(\phi) = \left(M - \frac{\varepsilon}{2} \right) \frac{(\mathbf{V}\mathbf{V}^{\mathrm{H}} + \hat{\lambda}\mathbf{I})^{-1} \bar{\vec{a}}(\phi)}{\bar{\vec{a}}^{\mathrm{H}}(\phi)(\mathbf{V}\mathbf{V}^{\mathrm{H}} + \hat{\lambda}\mathbf{I})^{-1} \bar{\vec{a}}(\phi)} \tag{17.101}$$

5. Compute the optimal weight by the Capon method, that is,

$$\vec{w}_0 = \alpha \hat{\mathbf{R}}^{-1} \hat{\vec{a}}(\phi) \tag{17.102}$$

where

$$\alpha = \frac{1}{\hat{\vec{a}}^{\mathrm{H}}(\phi) \hat{\mathbf{R}}^{-1} \hat{\vec{a}}(\phi)} \tag{17.103}$$

This RCB belongs to the class of diagonal loading algorithms, but the problem of determining the optimal amount of the diagonal loading level has been solved and is based on the uncertainty set of the steering vector.

Avoiding the Eigendecomposition Step

Since the eigendecomposition in Step 2 is computationally expensive [it requires $\mathcal{O}(L^3)$ floating point operations], in order to avoid that computation as well as avoiding the requirement of knowing the number of signals a priori, we use a different technique to obtain the noise subspace [9].

$\mathbf{R}$ is decomposed with the EVD

$$\mathbf{R} = \begin{bmatrix} \mathbf{U} & \mathbf{V} \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda}_{\mathrm{s}} + \sigma_v^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_v^2 \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{U}^{\mathrm{H}} \\ \mathbf{V}^{\mathrm{H}} \end{bmatrix} \tag{17.104}$$

where $\mathbf{\Lambda}_{\mathrm{s}} = \mathrm{diag}\{\delta_1^2, \delta_2^2, \ldots, \delta_L^2\}$, and $\mathbf{U}$ denotes the signal subspace. We approximate the noise subspace of $\mathbf{R}$ based on $\mathbf{R}^{-m}$ (m is a positive integer). Accordingly,

$$\sigma_v^{2m} \mathbf{R}^{-m} = \mathbf{V}\mathbf{V}^{\mathrm{H}} + \mathbf{U}\mathrm{diag}\left[\left(\frac{\sigma_v^2}{\delta_i^2 + \sigma_v^2} \right)^m \right] \mathbf{U}^* \tag{17.105}$$

Palpably, $[\sigma_v^2 / (\delta_i^2 + \sigma_v^2)]^m$ is less than 1 and converges to 0 for sufficiently large m. Thus,

$$\lim_{m\to\infty} \sigma_v^{2m}\mathbf{R}^{-m} = \mathbf{V}\mathbf{V}^{\mathrm{H}} \tag{17.106}$$

Therefore, we can modify the criterion (17.91) and consider the following cost function

$$\begin{aligned} &\underset{\vec{a}}{\text{minimize}} && \vec{a}^{\,*}(\phi)\mathbf{R}^{-m}\vec{a}(\phi) \\ &\text{such that} && \left\|\vec{a}(\phi)-\bar{\vec{a}}(\phi)\right\|^2 \le \varepsilon \text{ and } \left\|\vec{a}(\phi)\right\|^2 = M \end{aligned} \tag{17.107}$$

By contrast, (17.107) avoids estimating that dimension directly. Moreover, as $m \to \infty$, the beamforming method in (17.107) converges to the subspace one in (17.91), and we will show in the next section through computer simulation that the performance of this method for finite *m* will converge to the subspace one.

17.6.4.2 Simulation Results

The simulation results in [47] assume a ULA with L = 20 antennas with half-wavelength spacing. The signals impinging on the array are mutually independent, narrowband waveforms. In all cases 50 Monte Carlo trials were executed.

First, consider the effect of the pointing error of the SOI on array output SINR. The exact direction of arrival of the SOI is θ_0, and we assume the value to be $\phi_0 + \Delta$, that is, $\vec{a}(\phi_0) = \vec{a}(\phi_0 + \Delta)$. We assume that $\vec{a}(\phi_0)$ belongs to the uncertainty set

$$\left\|\vec{a}(\phi_0)-\bar{\vec{a}}(\phi_0)\right\|^2 \le \varepsilon \tag{17.108}$$

In this example, the directions of the SOI and an interference source are $\phi_0 = 30^{\circ}, \phi_1 = -30^{\circ}$, respectively. The assumed direction of the SOI is $\phi_0 + \Delta = 33^{\circ}$, which results in exact ε_0 = 5.7750. The *interference-to-noise ratio* (INR) is 40 dB.

Figure 17.13 [47] shows the array output SINR versus the SNR of the SOI when the number of snapshots is set to be N = 100 and ε = 0.7. It is observed that the RCB algorithm (17.91) performs better than the other two algorithms at all input SNR. Also, since the error in the steering vector of SOI is relatively large and cannot be neglected, the SCB and the DL version suffer from severe performance degradation as the SNR increases. However, the RCB has a 5 dB SINR loss when SNR = 20 dB. The modified RCB method in (17.107) for different m = 1 and m = 3 over various input SNRs is also illustrated in Figure 17.9. Obviously, the output SINR for m = 3 converges to the subspace approach (17.91), the counterpart for m = 1 [24] has the largest output SINR loss. These

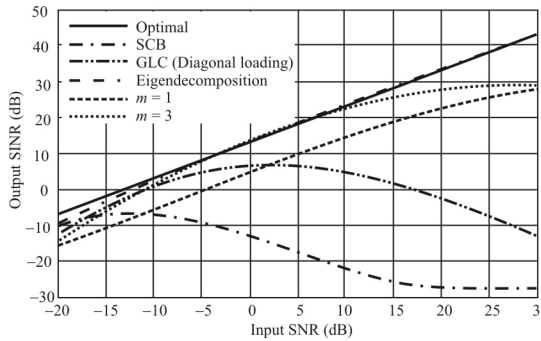


Figure 17.13 Output SINR versus input SNR. $\Delta = 3^{o}$, $\varepsilon = 0.7$, INR = 40 dB.

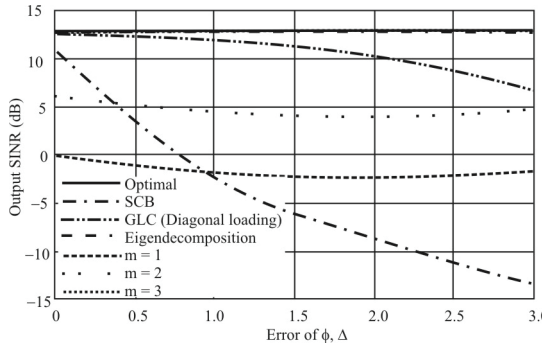


Figure 17.14 Output SINR versus pointing errors. $\varepsilon = 0.7$, SNR = 0 dB, INR = 20 dB.

results for $\varepsilon = 7$ are virtually identical, illustrating that the sensitivity to the selection for ε is small.

Figure 17.14 [47] shows the array output SINR curve versus the pointing errors, in which case SNR = 0 dB, INR = 20 dB, $N = 100$, and $\varepsilon = 0.7$. In Figure 17.14, the excellent performance achieved by the RCB (17.91) algorithm is observed, which shows the robustness to the pointing errors. The RCB results are virtually identical to the optimal, as are the results for (17.107) with $m = 3$. It is noted that, similar to other robust approaches, our method will worsen if there is strong interference spatially closed to the SOI. The reason is that for a given uncertainty region (17.90), the solution for $\vec{a}$ in (17.91) will converge to the strong interference source. Also it can be seen that the output SINR of the proposed modified method increases as m increases.

As evidenced by these simulation results, the RCB is effective at detecting signals at very low SNRs—conceivably below the noise floor. This makes the approach very attractive for CDMA signal detection.

The hypersensitivity of the SMI to steering vector mismatches is due to the projection of the signal onto the noise subspace. When the errors are considered part of the noise, the optimum weights that result from the Weiner equation [59] can be written as

$$\vec{w} = \sum_{l=1}^{L} \frac{1}{\lambda_l} \vec{u}_l \vec{u}_l^{*} \vec{a} + \sum_{j=1}^{N-L} \frac{1}{\sigma^2} \vec{v}_j \vec{v}_j^{*} \vec{a} \tag{17.109}$$

where λ_l are the eigenvalues of the signal subspace, $\vec{u}_l$ are the corresponding eigenvectors, σ^2 is the level of the noise subspace eigenvalues, $\vec{v}_j$ are the corresponding eigenvectors, $\vec{a}$ is the steering vector that maximizes the SINR. Typically $\lambda_l >> \sigma^2$ so the division by σ^2 in the second term in (17.109) causes the noise eigenvectors to have much greater influence over $\vec{w}$.

Therefore, a way to reduce the sensitivity to array manifold mismatches is to ignore the projection of the signal onto the noise subspace [54]. Shown in Figure 17.15 [60] are comparisons of the performance with such projection and without in Figure 17.16 [60]. An M-element, half-wavelength ULA is the antenna used for this simulation with M varying from 2 to 17. The SOI arrives at 0^o with an SNR of 10 dB. The single interferer arrives at 15^o with an INR of 30 dB. The hypersensitivity of the SINR to the steering vector is clearly evident in Figure 17.15, whereas by ignoring the noise subspace projection the SINR remains constant over a much larger range. The trade-off being made here is that for each iteration an eigendecomposition is required to obtain the signal subspace, whereas for the SMI algorithm, a single matrix inversion is required.

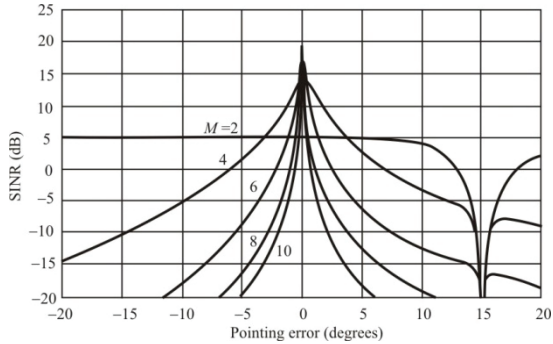


Figure 17.15 Hypersensitivity effect. SINR versus pointing error for the Weiner solution. *M* is the number of antennas.

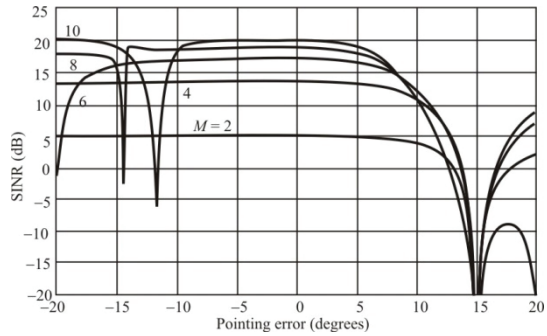


Figure 17.16 Hypersensitivity effect 2. SINR versus pointing error for a hypersensitivity free solution. *M* is the number of antennas.

17.6.5 Norm-Constrained Capon Beamforming

We present another algorithm in this section where the steering vector errors are recognized and are assumed unknown. However, the steering vector is known to lie within a sphere [61].

Consider an array comprised of M antennas, and let $\mathbf{R}$ denote the theoretical covariance matrix of the array output vector. We assume that $\mathbf{R}$ (positive definite) has the following form

$$\mathbf{R} = \sigma_0^2 \vec{a}(\phi_0)\vec{a}(\phi_0)^{\mathrm{H}} + \sum_{l=1}^{L} \sigma_l^2 \vec{a}(\phi_l)\vec{a}(\phi_l)^{\mathrm{H}} + \mathbf{N} \tag{17.110}$$

where $\left\{\sigma_0^2, \{\sigma_l^2\}_{l=1}^{L}\right\}$ are the powers of the $(L + 1)$ uncorrelated signals impinging on the array, $\left\{\vec{a}(\phi_0), \{\vec{a}(\phi_l)\}_{l=1}^{L}\right\}$ are the steering vectors that depend on the array geometry and are functions of DOAs, $(.)^{\mathrm{H}}$ denotes conjugate transpose, and $\mathbf{N}$ is the noise covariance matrix [the "noise" comprises nondirectional signals, and hence, $\mathbf{N}$ usually has full rank as opposed to the other terms in (17.110) whose rank is equal to 1]. The first term in (17.110) corresponds to the SOI and the remaining rank-one terms to L interferences. To avoid ambiguities, we assume that

$$\left\|\vec{a}(\phi_0)\right\|^2 = M \tag{17.111}$$

where $\|\bullet\|$ denotes the Euclidean norm.

We note that the above expression for $\mathbf{R}$ holds for both narrowband and wideband signals; in the former case, $\mathbf{R}$ is the covariance matrix at the carrier frequency, and in the latter case, $\mathbf{R}$ is the covariance matrix at the center of a given frequency bin. In practical applications, $\mathbf{R}$ is replaced by the sample covariance matrix $\hat{\mathbf{R}}$, given by (17.35).

Similarly to the eigendecomposition RCB we considered in the last section, we assume that the only knowledge we have about the steering vector is that it belongs to the following uncertainty sphere specified by

$$\left\|\vec{a}(\phi_0) - \bar{\vec{a}}\right\|^2 \le \epsilon \tag{17.112}$$

where $\bar{\vec{a}}$ and ϵ are given (as above, $\bar{\vec{a}}$) is the assumed steering vector of the SOI, determined by calibration of the array, and ϵ is a user parameter (whose choice

will be discussed later). We also assume that the steering vector satisfies the same norm constraint (17.111)

$$\left\|\bar{\vec{a}}\right\|^2 = M \tag{17.113}$$

The assumption that $\left\|\vec{a}(\phi_0)\right\|^2 = M$ (we choose $\bar{\vec{a}}$) is reasonable for many scenarios including the cases of the look direction error and phase perturbations. While being violated when the array response vector also has gain perturbations, if these gain perturbations are small, the norm constraint still approximately holds.

The RCB algorithm described in this section, called the *norm-constrained Capon beamformer* (NCCB) [13], in addition to (17.62), imposes an additional constraint on the Euclidean norm of $\vec{w}$ for the purpose of improving the robustness of the Capon beamformer against SOI steering vector errors and noise. Thus, the beamforming problem is formulated as

$$\min_{\mathbf{w}} \vec{w}^{\mathrm{H}}\mathbf{R}\vec{w} \qquad \text{subject to} \qquad \vec{w}^{\mathrm{H}}\bar{\vec{a}} = 1 \text{ and } \left\|\vec{w}\right\|^2 \le \zeta \tag{17.114}$$

The choice of ζ is not easy to make, however. In particular, this choice is not directly linked to the ϵ in (17.112) or the uncertainty of the SOI steering vector.

Let $\mathcal{S}$ be the set defined by the constraints in (17.114). In addition, let the Lagrange function be defined as

$$g(\vec{w},\lambda,\mu) = \vec{w}^{\mathrm{H}}\hat{\mathbf{R}}\vec{w} + \lambda(\left\|\vec{w}\right\|^2 - \zeta) + \mu(-\vec{w}^{\mathrm{H}}\bar{\vec{a}} - \bar{\vec{a}}^{\mathrm{H}}\vec{w} + 2) \tag{17.115}$$

where λ and μ are the real-valued Lagrange multipliers with μ being arbitrary and $\lambda \ge 0$ satisfying $\mathbf{R} + \lambda\mathbf{I} > 0$ so that $g(\vec{w},\lambda,\mu)$ can be minimized with respect to $\vec{w}$.

Consider

$$\frac{\bar{\vec{a}}^{\mathrm{H}}\hat{\mathbf{R}}^{-2}\bar{\vec{a}}}{[\bar{\vec{a}}^{\mathrm{H}}\mathbf{R}^{-1}\bar{\vec{a}})]^2} \le \zeta \tag{17.116}$$

When (17.116) is satisfied, then the optimum weights are given by

$$\hat{\vec{w}} = \frac{\hat{\mathbf{R}}^{-1}\bar{\vec{a}}}{\bar{\vec{a}}^{\mathrm{H}}\hat{\mathbf{R}}^{-1}\bar{\vec{a}}} \tag{17.117}$$

and $\lambda = 0$ and the norm constraint in (17.114) has no role. Otherwise, we have

$$\frac{\bar{\bar{a}}^{\mathrm{H}}\hat{\mathbf{R}}^{-2}\bar{\bar{a}}}{[\bar{\bar{a}}^{\mathrm{H}}\mathbf{R}^{-1}\bar{\bar{a}}]^2} > \zeta \tag{17.118}$$

yielding

$$\hat{\mu} = \frac{1}{\bar{\bar{a}}^{\mathrm{H}}(\mathbf{R}+\lambda\mathbf{I})^{-1}\bar{\bar{a}}} \tag{17.119}$$

The function for determining $\hat{\lambda}$ is given by

$$\frac{\bar{\bar{a}}^{\mathrm{H}}(\mathbf{R}+\lambda\mathbf{I})^{-2}\bar{\bar{a}}}{\left[\bar{\bar{a}}^{\mathrm{H}}(\mathbf{R}+\lambda\mathbf{I})^{-1}\bar{\bar{a}}\right]^2} = \zeta \tag{17.120}$$

Then the optimum weights are

$$\hat{\bar{w}} = \frac{(\mathbf{R}+\hat{\lambda}\mathbf{I})^{-1}\bar{\bar{a}}}{\bar{\bar{a}}^{\mathrm{H}}(\mathbf{R}+\hat{\lambda}\mathbf{I})^{-1}\bar{\bar{a}}} = \zeta \tag{17.121}$$

It can be shown [13] that a lower bound on ζ is given by

$$\zeta \geq 1/M \tag{17.122}$$

The eigenvalue decomposition of $\hat{\mathbf{R}}$ is

$$\hat{\mathbf{R}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{H}} \tag{17.123}$$

where the columns of $\mathbf{U}$ contain the eigenvectors of $\hat{\mathbf{R}}$, and the diagonal elements of the diagonal matrix $\boldsymbol{\Lambda}$, $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_M$ are the corresponding ordered eigenvalues. Let

$$\bar{\bar{z}} = \mathbf{U}^{\mathrm{H}}\bar{\bar{a}} \tag{17.124}$$

and let z_{m} denote the mth element of $\bar{\bar{z}}$. Then (17.120) can be rewritten as

$$\frac{\sum_{m=0}^{M-1} \frac{|z_m|^2}{(\gamma_m + \hat{\lambda})^2}}{\left[\sum_{m=0}^{M-1} \frac{|z_m|^2}{(\gamma_m + \hat{\lambda})^2}\right]^2} = \zeta \tag{17.125}$$

Hence, we have

$$\zeta \le \frac{\frac{\|\vec{a}\|^2}{(\gamma_M + \hat{\lambda})^2}}{\frac{\|\vec{a}\|^4}{(\gamma_1 + \hat{\lambda})^2}} = \frac{(\gamma_1 + \hat{\lambda})^2}{L(\gamma_M + \hat{\lambda})^2} \tag{17.126}$$

yielding an upper bound on $\hat{\lambda}$ given by

$$\hat{\lambda} \le \frac{\gamma_1 - \sqrt{M\zeta}\gamma_M}{\sqrt{M\zeta} - 1} \tag{17.127}$$

Note that the computational complexity of NCCB as determined by the decomposition in (17.123) is comparable with that of the SCB, which also requires $\mathcal{O}(M^3)$ floating point operations.

NCCB RCB Algorithm

1. Compute $\hat{\mathbf{R}}$ with (17.35).
2. Find the eigenvalue decomposition of $\hat{\mathbf{R}}$ using (17.123).
3. If (17.118) is satisfied, solve (17.125) for $\hat{\lambda}$ using Newton's method, using the knowledge that the solution is unique and it is lower bounded by 0 and upper bounded by (17.127); otherwise, set $\hat{\lambda} = 0$.
4. Use the $\hat{\lambda}$ from step 3 to get

$$\hat{\vec{w}} = \frac{\mathbf{U}(\mathbf{\Lambda} + \hat{\lambda}\mathbf{I})^{-1}\mathbf{U}^{\mathrm{H}}\vec{a}}{\vec{a}^{\mathrm{H}}\mathbf{U}(\mathbf{\Lambda} + \hat{\lambda}\mathbf{I})^{-1}\mathbf{U}^{\mathrm{H}}\vec{a}} \tag{17.128}$$

where the vector $\vec{z} = \mathbf{U}^{*}\vec{a}$ is available from step 3.

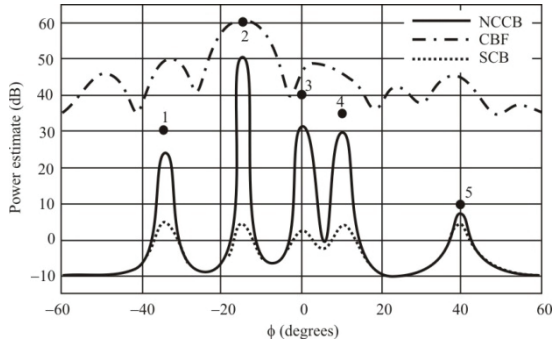


Figure 17.17 Power estimates versus ϕ (AOA) in decibles from (17.129). The true powers are indicated with a circle and the true AOAs are -35^{o}, 15^{o}, 0^{o}, 10^{o}, and 40^{o}, and $\varepsilon = 1.0$.

5. Compute the SOI power estimate of NCCB from (17.36) using (17.128) yielding

$$\hat{\sigma}_0^2 = \frac{\bar{\bar{a}}^{\mathrm{H}}\mathbf{U}(\boldsymbol{\Lambda}+\hat{\lambda}\mathbf{I})^{-2}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{H}}\bar{\bar{a}}}{\left[\bar{\bar{a}}^{\mathrm{H}}\mathbf{U}(\boldsymbol{\Lambda}+\hat{\lambda}\mathbf{I})^{-1}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{H}}\bar{\bar{a}}\right]^2} \tag{17.129}$$

Note that this algorithm works even when $\hat{\mathbf{R}}$ is singular. Therefore, $N < M$ is allowed.

A comparison of the CBF, the SCB, and the NCCB algorithms is shown in Figure 17.17 [13] for five signals. The power of these signals is 30, 60, 40, 35, and 10 dB at the AOAs shown. As expected, the CBF did not perform well at all, and

the SCB peaks are discernible, but the power accuracy is limited. The NCCB algorithm peaks are much more discernible and considerably more accurate than either of the other two algorithms.

17.7 Digital Beamforming

Traditionally beamforming was accomplished in the analog signal processing domain. Almost all modern implementations utilize digital signal processing, however, because digital beamforming provides for much greater flexibility [62]. Any of the algorithms discussed in Section 17.6 can be implemented digitally. The operations of phase shifting and amplitude scaling are still required. In digital beamforming, however, these operations, as well as summation for receiving for each antenna element, are done digitally. Either general-purpose DSPs or dedicated beamforming chips are used (perhaps implemented with ASICs).

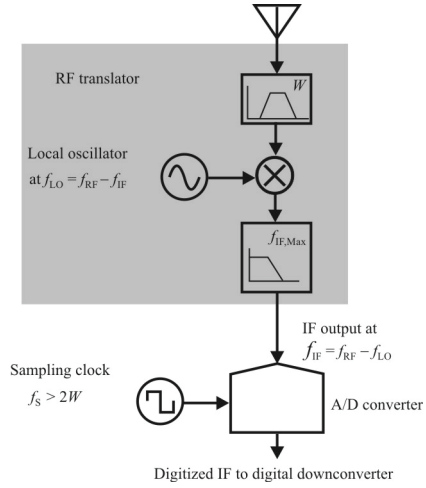


Figure 17.18 RF translator.

Although both receiving and transmitting beamformers can be implemented digitally, we will focus on beamforming receivers for now. Although beamforming can be performed digitally at IF (RF digital beamforming is generally too limited in frequency range for EW applications), we will focus on baseband beamforming here. The steps are essentially the same.

Digital processing requires that the signal from each antenna element is digitized using an A/D converter. Since the frequencies of radio signals above HF (>30 MHz) are generally too high to be directly digitized at a reasonable cost and dynamic range, digital beamforming receivers use analog RF translators to shift the signal frequency down before the ADC. Figure 17.18 [62] shows the architecture of a translator that shifts the bandwidth W down to a lower frequency range.

Once the antenna signals have been digitized, they are passed to digital downconverters that shift the signal's center frequency down to 0 Hz and pass only the bandwidth required for one channel. The downconverters produce a quadrature baseband output at a low sample rate (see Figure 17.19 [62]).

The quadrature baseband I and Q components can be used to represent a signal as a complex vector (phasor) with real and imaginary parts. Two components are required so that both positive and negative frequencies (relative to the channel center frequency) can be represented. So

$$s(t) = x(t) + jy(t) \tag{17.130}$$

where

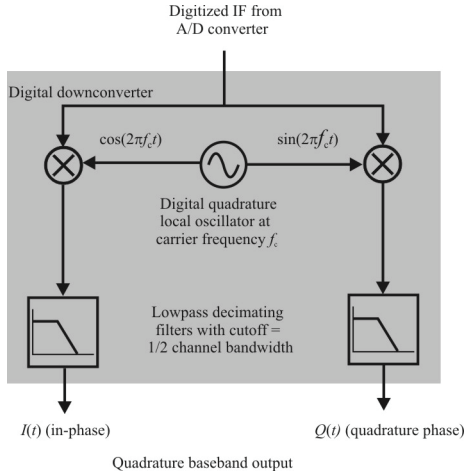


Figure 17.19 Digital downconverter.

$s(t)$ is the complex baseband signal.
$x(t) = I(t)$ is the real part.
$y(t) = -Q(t)$ is the imaginary part.

The complex baseband signals are multiplied by complex weights to apply the phase shift and amplitude scaling required for each antenna element.

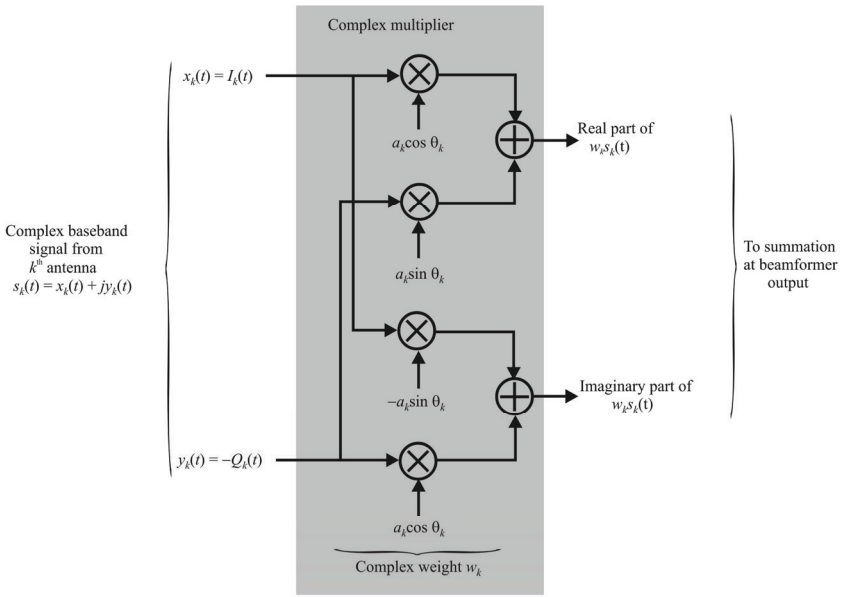
$$w_k = a_k e^{j\theta_k} = a_k \cos\theta_k + j a_k \sin\theta_k \tag{17.131}$$

w_k is complex weight for the kth antenna element, a_k is the relative amplitude of the weight, and θ_k is the phase shift of the weight. A general-purpose DSP can implement the complex multiplication for each antenna element, which is given by

$$s_k(t)w_k = a_k\{x_k(t)\cos\theta_k - y_k(t)\sin\theta_k + j[x_k(t)\sin\theta_k + y_k(t)\cos\theta_k]\} \tag{17.132}$$

The flow diagram for this calculation is illustrated in Figure 17.20 [62].

Figure 17.21 [62] shows a digital beamforming receiver. One set of antenna elements, RF translators, and ADC can be shared by a number of beamformers as implied in the figure. The RF translators and A/D converters share common oscillators and clocks so that they all (ideally) produce identical phase shifts. The digital downconverters share a common clock, are set for the same center frequency and bandwidth, and their digital local oscillators are in-phase so that all phase shifts are identical. Digital local oscillators are easier to keep in phase



Figuer 17.20 Flow diagram for complex coefficient multiplier.

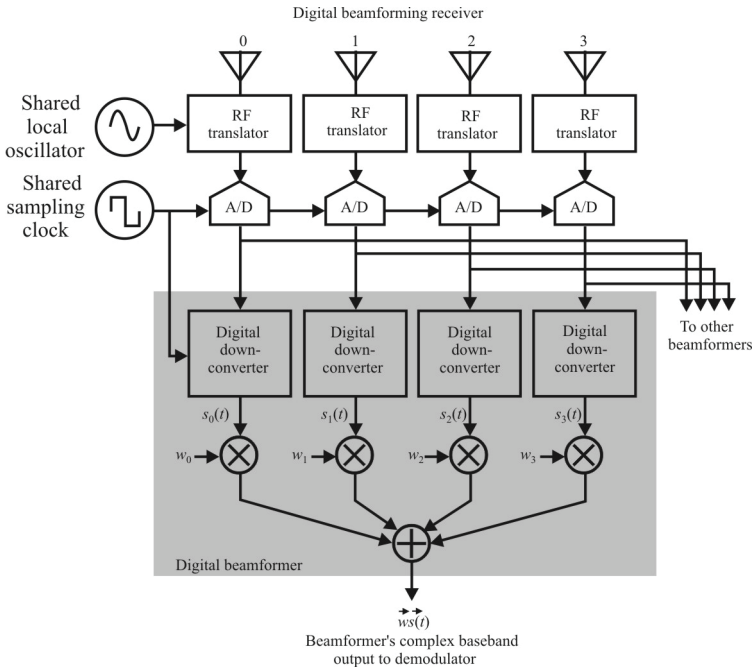


Figure 17.21 Digital beamforming receiver.

synchronization than their analog counterparts. Each DDC's baseband output is multiplied by the complex weight for its antenna element, and the results are summed to produce one baseband signal. A demodulator may then follow to recover information from the signal, although that function is not part of the beamformer.

17.7.1 FFTs in Beamforming

As mentioned, many beamformers can share one set of antenna elements, RF translators, and A/D converters. The beamformers may have their central beams pointed in different directions. In situations where a fixed set of nonoverlapping beams must be formed simultaneously (e.g., DF), an FFT can implement many beamformers efficiently.

Figure 17.22 [62] shows an FFT beamformer implementation with M antenna elements. Each element feeds a digital downconverter. All DDCs produce a baseband sample simultaneously, and all of these are passed at once to an M-point complex FFT.

The FFT then produces a set of N complex outputs, each of which is the next baseband sample for a different beam. The FFT is processing a set of samples that are separated in space (not in time). Therefore, its outputs are a set of samples that are separated in direction (not in frequency).

FFT beamforming, as shown above, implements fixed beams and therefore is not flexible. For a linear array, the N beams are fixed and equally spaced in direction. They range from -90^{o} to $+90^{o}$ degrees from broadside of the array. The beams are orthogonal: the central peak of any beam lies in a null on all other beams. Such a set of beams is useful for cellular phone towers where fixed beams are used but not very useful for tactical mobile communications or EW.

It is possible to use FFTs for flexible beamforming. A set of FFT outputs can be combined, using complex weights and sums as before, to form arbitrary radiation patterns. This is illustrated in Figure 17.23 [62].

17.8 Adaptive Transmit Arrays

Up until now we have been focused on adaptive receiving arrays and optimizing in some sense the quality of the signal being received. It is also possible to form adaptive arrays for transmitting—indeed, that is precisely what an electronic steerable (phased) array (ESA) for radar applications does. For EW applications a transmitting adaptive array may be used for EA.

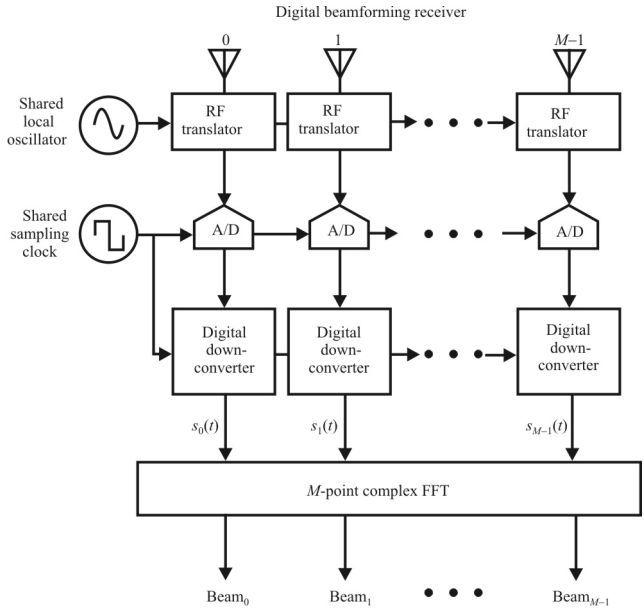


Figure 17.22 FFT beamforming.

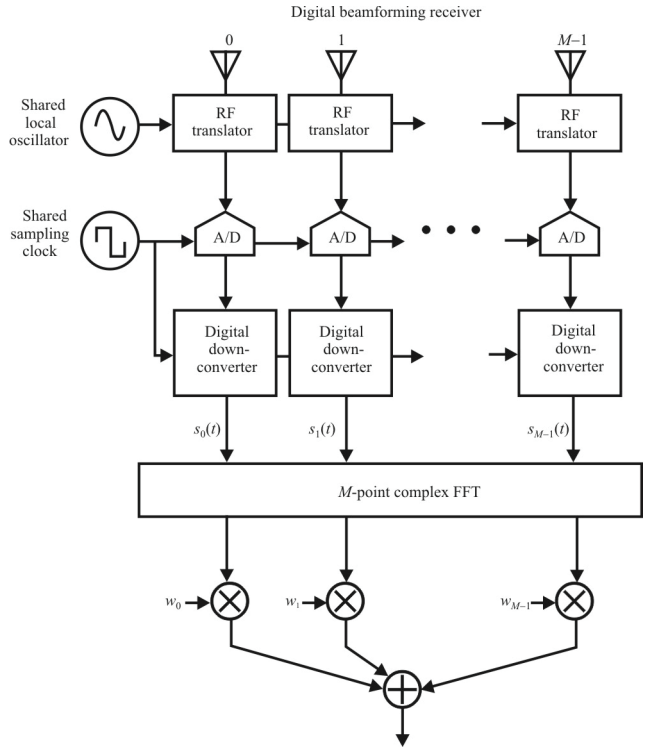


Figure 17.23 FFT beamspace forming.

The general principles are similar to receiving arrays, except that the pointing direction for the beam(s) are known a priori (perhaps as determined by an associated ES system) and the complex weights need to be applied to steer the beam(s) in the desired direction. A block diagram of such an arrangement when the array consists of four antennas is illustrated in Figure 17.24 [62].

The complex weights are adjusted in exactly the same way they are adjusted for receiving applications. The phase shifters/amplitude adjusters are set when the jamming signal is at low power, before the power amplifiers. The PAs elevate the signal power to each antenna. It is required that these PAs be as linear as possible in order to avoid spilling jamming power over into adjacent channels (causing fratricide and wasted power) [63].

The pattern for the eight-element linear array shown in Figure 17.25 [62] is shown in Figure 17.26 [62].

17.8.1 Beam Steering

The beam of the array can be steered electronically by varying the amplitudes and phases of the array weights [64]. Consider a linear array of equally spaced elements as illustrated in Figure 17.27. Adjacent elements are separated by a distance specified by d, and the amplitudes of the signals at each element are assumed equal. The position of the main beam will be broadside to the array at an angle $\theta = 0$ when the same phase delay is applied to all elements because the relative phase difference between adjacent elements is zero. If the relative phase difference between elements is something other than zero, the main beam will point in a direction other than broadside. The direction of the main beam is steered to an angle θ_0 when that phase difference is

$$\phi = 2\pi \frac{d}{\lambda} \sin\theta_0 \tag{17.133}$$

Using the first antenna element as the reference, the phase at each element is therefore $(m-1)\phi$, where $m = 1, 2, \ldots, M$.

The normalized radiation pattern of the array when the phase difference between adjacent elements is given by (17.133) is given by [64]

$$G(\theta) = \frac{\sin^2[M\pi(d/\lambda)(\sin\theta - \sin\theta_0)]}{M^2 \sin^2[\pi(d/\lambda)(\sin\theta - \sin\theta_0)]} \tag{17.134}$$

When $\sin\theta = \sin\theta_0$, both the numerator and denominator in (17.134) are zero producing an indeterminate form. However, it can be shown (using L'Hopital's rule) that the maximum of this pattern occurs when $\sin\theta = \sin\theta_0$. Obviously by

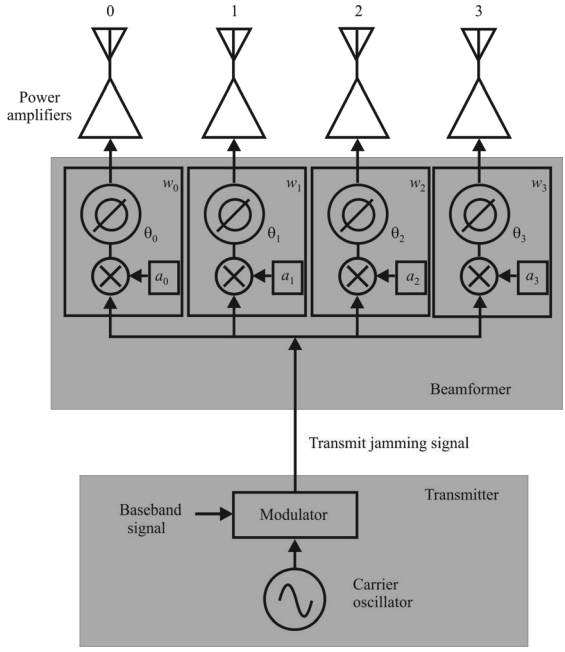


Figure 17.24 Transmit beamformer. The baseband signal is an appropriate jamming waveform (noise, tone, multiple tones, and so forth).

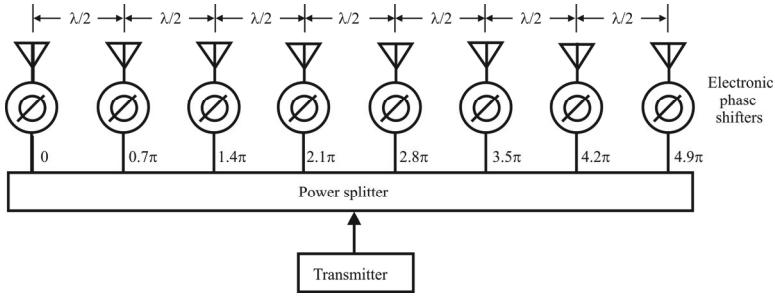


Figure 17.25 Eight-element linear array.

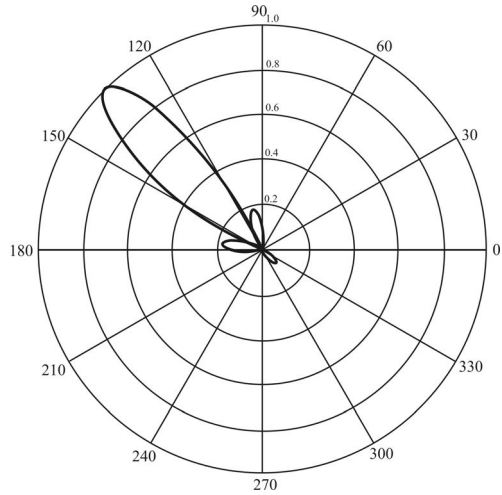


Figure 17.26 Eight-element transmit pattern. Horizontal plane. 3.5 wavelength, eight-element linear array, progressive 0.7π phase shift.

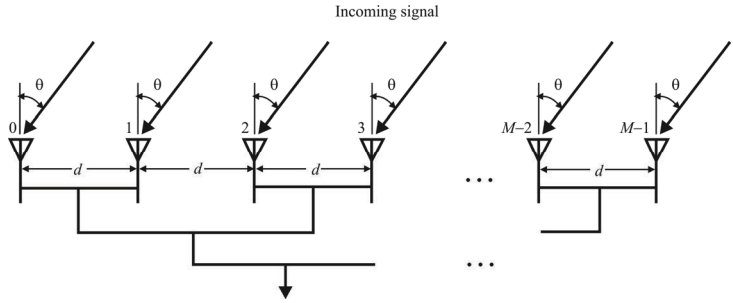


Figure 17.27 Linear array. Shown is a corporate feed where all elements are driven in parallel.

inserting variable ϕ in (17.133) and using (17.134), the beam may be steered as the relative phase between elements is changed.

As usual, grating lobes appear at an angle θ_g whenever the denominator of (17.134) is zero, or when

$$\frac{d}{\lambda}(\sin\theta_g - \sin\theta_0) = \pm n\pi, \quad n = 0,1,\ldots \tag{17.135}$$

or

$$\left|\sin\theta_g - \sin\theta_0\right| = \pm n\frac{\lambda}{d}, \qquad n = 0,1,\ldots \tag{17.136}$$

The element spacing must not be larger than a half-wavelength if the beam is to be steered over a wide angle without having undesirable grating lobes appear. A grating lobe will occur at -90^o when the beam is steered to $+90^o$ if $d = \lambda/2$, thus establishing this limit. For practical reasons, array antennas do not scan $\pm 90^o$. A typical limit is $\pm 60^o$, where (17.136) indicates that the element spacing should be less than 0.54λ.

17.8.2 Beamwidth as a Function of Steering Angle

In a linear array, the width of the beam changes as it is steered off boresight. An approximate expression for this variation is given by [64]

$$\theta_B \approx \frac{0.886\lambda}{Md\cos\theta_0} \tag{17.137}$$

Equation (17.137) obviously varies as $\cos^{-1}\theta_0$ and is most accurate for small θ_0 (boresight), while the accuracy gets worse the larger θ_0 becomes, and is not accurate at all at endfire.

The above analysis applies to linear arrays, but similar results can be derived for equispaced planar arrays.

17.9 Concluding Remarks

The principles of operation of several Capon beamformers were presented in this chapter and compared with the standard beamformer. Capon beamformers are not the only form of beam steering algorithm available, but they have been extensively researched and their performance is well understood. We focused on these in particular because they are representative of the types of processing available to accomplish beam steering.

We also covered how the beamforming steps can be accomplished digitally. Digital signal processing for this purpose has the advantages it has in other types of signal processing—it is considerably more flexible and the system performance characteristics are more stable over time, as well as over temperature, device tolerances, and so forth.

References

[1] Godara, L. C., "Applications of Antenna Arrays to Mobile Communications, Part I: Performance Improvement, Feasibility, and System Considerations," *Proceedings of the IEEE*, Vol. 85, No. 7, July 1997, pp. 1031–1060.

[2] Godara, L. C., "Applications of Antenna Arrays to Mobile Communications, Part II: Beam-Forming and Direction-of-Arrival Considerations," *Proceedings of the IEEE,* Vol. 85, No. 8, August 1997, pp. 1195–1245.

[3] Allen, B., and M. Ghavami, *Adaptive Array Systems Fundamentals and Applications*, New York: Wiley, 2005.

[4] Johnson, D. H., and D. E. Dudgeon, *Array Signal Processing, Concepts and Techniques,* Upper Saddle River, NJ: Prentice Hall, 1993, Chapter 4.

[5] Van Trees, H. L., *Detection, Estimation, and Modulation Theory, Part IV, Optimum Array Processing,* New York: Wiley, 2002, Section 6.2.

[6] Tsai, J-A., and B. D. Woerner, "Adaptive Beamforming of Uniform Circular Arrays (UCA) for Wireless CDMA System," *Conference Record of the Thirty-Fifth IEEE Asilomar Conference on Signals, Systems and Computers*, 2001, Vol.1, 04–07 Nov 2001 Pacific Grove, CA, pp. 399–403.

[7] Krim, H., and M. Viberg, "Two Decades of Array Signal Processing Research," *IEEE Signal Processing Magazine*, July 1996, pp. 67–94.

[8] Weiss, S., and I. K. Proudler, "Comparing Efficient Broadband Beamforming Architectures and their Performance Trade-Offs," Internal Report Qinetiq, Ltd, Great Malvern, U.K.

[9] Nunn, D., "Suboptimal Frequency-Domain Adaptive Antenna Processing Algorithm for Broad-Band Environments," *IEE Proceedings of Radar and Signal Processing*, Vol. 134, No. 4, July 1987, pp. 341–351.

[10] Compton, R. T., "The Bandwidth Performance of a Two-Element Adaptive Array with Tapped Delay-Line Processing," *IEEE Transactions on Antennas and Propagation*, Vol. 36, No.1, January 1988, pp. 4–17.

[11] Godara, L. C., Application of the Fast Fourier Transform to Broadband Beamforming. *Journal of the Acoustic Society of America*, Vol. 98, No. 1, July 1995, pp. 230–240.

[12] Weiss, S., R.W. Stewart, M. Schabert, I. K. Proudler, and M. W. Hoffman, "An Efficient Scheme for Broadband Adaptive Beamforming," *Asilomar Conference on Signals, Systems, and Computers*, Volume I, Monterey, CA, November 1999, pp. 496–500.

[13] Li, K., P. Stoica, and Z. Wang, "On Robust Capon Beamforming and Diagonal Loading," *IEEE Transactions on Signal Processing*, June 2004, pp. 1702–1717.

[14] Van Trees, H. L., *Detection, Estimation, and Modulation Theory, Part IV, Optimum Array Processing,* New York: Wiley, 2002.

[15] Steele, A. K., "Comparison of Directional and Derivative Constraints for Beamformers Subject to Multiple Liner Constraints," *Proceedings of the IEEE*, Vol. 130, No. 1, February 1983, pp. 41–45.

[16] Gershman, A. B., "Robust Adaptive Beamforming in Sensor Arrays," *Institute Journal Electronics Communications,* Vol. 53, No. 6, 1999, pp. 305–314.

[17] Vorobyov, S. A., A. B. Gershman, and Z.-Q. Luo, "Robust Adaptive Beamforming Using Worst Case Performance Optimization: A Solution to the Signal Mismatch Problem," *IEEE Transactions on Signal Processing*, Vol. 51, February 2003, pp. 313–324.

[18] Lorenz, R. G., and S. P. Boyd, "Robust Minimum Variance Beamforming," *IEEE Transactions on Signal Processing,* 2001, pp. 1684–1696.

[19] Reed, I. S., J. D. Mallett, and L. E. Brennan, "Rapid Convergence Rate in Adaptive Arrays," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES–10, November 1974, pp. 853–863.

[20] Griffiths, L. J., and C. W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beamforming," *IEEE Transactions on Antennas and Propagation*, Vol. AP–30, January 1982, pp. 27–34.

[21] Hung, E. K., and R. M. Turner, "A Fast Beamforming Algorithm for Large Arrays," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-19, July 1983, pp. 598–607.

[22] Carlson, B. D., "Covariance Matrix Estimation Errors and Diagonal Loading in Adaptive Arrays," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 24, July 1988, pp. 397–401.

[23] Monzingo, R. A., and T. W. Miller, *Introduction to Adaptive Arrays*. New York: Wiley, 1980.

[24] Feldman, D. D., and L. J. Griffiths, "A Projection Approach to Robust Adaptive Beamforming," *IEEE Transactions on Signal Processing*, Vol. 42, April 1994, pp. 867–876.

[25] Jablon, N. K., "Adaptive Beamforming with the Generalized Sidelobe Canceller in the Presence of Array Imperfections," *IEEE Transactions on Antennas and Propagation*, Vol. AP-34, August 1986, pp. 996–1012.

[26] Gershman, A. B., V. I. Turchin, and V. A. Zverev, "Experimental Results of Localization of Moving Underwater Signal by Adaptive Beamforming," *IEEE Transactions on Signal Processing*, Vol. 43, October 1995, pp. 2249–2257.

[27] Ringelstein, J., A. B. Gershman, and J. F. Böhme, "Direction Finding in Random Inhomogeneous Media in the Presence of Multiplicative Noise," *IEEE Signal Processing Letters*, Vol. 7, October 2000, pp. 269–272.

[28] Hong, Y. J., C. C. Yeh, and D. R. Ucci, "The Effect of a Finite-Distance Signal Source on a Far-Field Steering Applebaum Array—Two Dimensional Array Case," *IEEE Transactions on Antennas Propagation*, Vol. 36, Apr. 1988, pp. 468–475.

[29] Pedersen, K. I., P. E. Mogensen, and B. H. Fleury, "A Stochastic Model of the Temporal and Azimuthal Dispersion Seen at the Base Station in Outdoor Propagation Environments," *IEEE Transactions on Vehicular Technology*, Vol. 49, March 2000, pp. 437–447.

[30] Goldberg, J., and H. Messer, "Inherent Limitations in the Localization of a Coherently Scattered Source," *IEEE Transactions on Signal Processing*, Vol. 46, Dec. 1998, pp. 3441–3444.

[31] Besson, O., and P. Stoica, "Decoupled Estimation of DOA and Angular Spread for a Spatially Distributed Source," *IEEE Transactions on Signal Processing*, Vol. 48, July 2000, pp. 1872–1882.

[32] Astely, D., and B. Ottersten, "The Effects of Local Scattering on Direction of Arrival Estimation with MUSIC," *IEEE Transactions on Signal Processing*, Vol. 47, December 1999, pp. 3220–3234.

[33] Gershman, A. B., "Robust Adaptive Beamforming in Sensor Arrays," *International Journal of Electronic Communications*, Vol. 53, December 1999, pp. 305–317.

[34] Cox, H., R. M. Zeskind, and M. H. Owen, "Robust Adaptive Beamforming," *IEEE Transactions on Acoustics, Speech, Signal Processing*, Vol. ASSP-35, October 1987, pp. 1365–1376.

[35] Bell, K. L., Y. Ephraim, and H. L. Van Trees, "A Bayesian Approach to Robust Adaptive Beamforming," *IEEE Transactions on Signal Processing*, Vol. 48, February 2000, pp. 386–398.

[36] Er, M. H., and T. Cantoni, "An Alternative Formulation for an Optimum Beamformer with Robustness Capability," *Proceedings of the Institute of Electrical Engineering and Radar, Sonar, and Navigation*, October 1985, pp. 447–460.

[37] Chang, L., and C. C. Yeh, "Performance of DMI and Eigenspace-Based Beamformers," *IEEE Transactions on Antennas and Propagation*, Vol. 40, November 1992, pp. 1336–1347.

[38] Applebaum, S. P., "Adaptive Arrays," *IEEE Transactions on Antennas and Propagation,* Vol. Ap-24, No. 9, September 1976, pp. 585–598.

[39] Capon, J., "High-Resolution Frequency-Wavenumber Spectrum Analysis," *Proceedings of the IEEE,* Vol. 57, No. 8, August 1969, pp. 1408–1418

[40] Gabriel, W. F., "Spectral Analysis and Adaptive Array Superresolution Techniques," *Proceedings of the IEEE,* Vol. 68, No. 6, June 1980, pp. 654666.

[41] Eight Articles, "Time Delay Estimation," *IEEE Transactions on Acoustics, Speech and Signal Processing,* Vol. ASSP-29, No. 3, June 1981, Special Issue.

[42] Burg, J., "Maximum Entropy Spectral Analysis," *Proceedings 37th Meeting Society Exploration Geophysicists,* 1961.

[43] MacDonald, V. H., and P. M. Schultheiss, "Optimum Passive Bearing Estimation in a Spatially Incoherent Noise Environment," *Journal of the Acoustic Society of America,* Vol. 46, No. 1, 1969, pp. 37–43.

[44] Schweppe, F. C., "Sensor Array Data Processing for Multiple Signal Sources," *IEEE Transactions on Information Technology,* Vol. IT-l4, 1968, pp. 294–305.

[45] Bienvenu, G., and L. Kopp, "Adaptivity to Background Noise Spatial Coherence for High Resolution Passive Methods," *Proceedings International Conference on Acoustics, Speech and Signal Processing,* 1980, pp. 307–317.

[46] Schmidt, R. O., A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation. Ph.D. Thesis, Stanford Univ., Stanford, CA, November 1981.

[47] Liao, G. S., Q. Liu, and J. Li, "A Subspace-Based Robust Adaptive Capon Beamforming," *Progress In Electromagnetics Research Symposium*, 2006, Cambridge, USA, March 26–29, 2006, pp. 374–379.

[48] Poisel, R. A., *Target Acquisition in Communication Electronic Warfare Systems*, Norwood, MA: Artech, 2004, Ch 6.

[49] Poisel, R. A., *Target Acquisition in Communication Electronic Warfare Systems*, Norwood, MA: Artech, 2004, pp. 101–108.

[50] Zoltowski, M. D., "On the Performance of the MVDR Beamformer in the Presence of Correlated Interference," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 36, June 1988, pp. 945–947.

[51] Allen, B., and M. Ghavami, *Adaptive Array Systems,* New York: Wiley, 2005, p. 117.

[52] Poisel, R. A., *Foundations of Communication Electronic Warfare Systems,* Norwood, MA: Artech, 2008, pp. 144–145.
[53] Bryson, A. E., and Y-C. Ho, *Applied Optimal Control,* London: Blaisdell, 1969.
[54] Vaidyanathan, P. P., and C.-Y Chen, www.systems.caltech.edu/cyc/presentations/pizza_meeting06.ppt.
[55] Du, L., J. Li, and P. Stoica, "Fully Automatic Computation of Diagonal Loading Levels for Robust Adaptive Beamforming," *IEEE Transactions on Aerospace and Electronic Systems*, January 2010, pp. 449–458.
[56] Poisel, R. A., *Modern Communications Jamming Principles and Techniques,* 2nd Ed., Norwood, MA: Artech, 2011, pp. 828–833.
[57] Thomas, J. K., L. L. Scharf, and D. W. Tufts, "The Probability of a Subspace Swap in the SVD," *IEEE Transactions on Signal Processing*, Vol. 43, No. 3, March 1995, pp. 730–736.
[58] Poisel R. A., *Modern Communications Jamming Principles and Techniques*, 2nd Ed., 2011.
[59] Weiner, N., *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*, New York: Wiley, 1949.
[60] Bull, J. F., and L. R. Burgess, "Hypersensitivity Effects in Adaptive Antenna Arrays," *Proceedings IEEE MILCOM*, 1990, pp. 396–399.
[61] Li, J., P. Stoica, and Z. Wang, "Doubly Constrained Robust Capon Beamformer," *IEEE Transactions Signal Processing*, Vol. 52. No. 9, September 2004, pp. 2407–2423.
[62] Haynes, T., "A Primer on Digital Beamforming," March 26, 1998, http://www.spectrumsignal.com.
[63] Kennington, P. B., *High-Linearity RF Amplifier Design,* Norwood MA: Artech House, 2000.
[64] Skolnick, M. I., *Introduction to Radar Systems*, 2nd Ed., New York: McGraw-Hill, 1980, pp. 282–284.

Appendix 17A Weiner Filter

17A.1 Introduction

The Weiner filter, as applied to antennas as the source of the input time sequence, is shown in Figure 17A.1. The estimation error at time slot n, denoted by $e(n)$, is defined as the difference between the desired response $d(n)$ and the filter output $y(n)$. The goal is to make the estimation error $e(n)$ as small as possible in some statistical sense by adjusting the weight vector $\vec{w}$. Specifically, for the Weiner filter, the mean-square value of the estimation error is minimized.

17A.2 Basic Principles

17A.2.1 Orthogonality

The filter output $y(n)$ at discrete time n is defined by the linear convolution sum.

$$y(n) = \sum_{k=0}^{\infty} w_k^* x(n-k), \qquad n = 0,1,2,\ldots \tag{17A.1}$$

The estimation error is defined by the difference

$$e(n) = y(n) - d(n) \tag{17A.2}$$

We define the cost function as the mean square of the error $e(n)$ as

$$\begin{aligned} J &= \mathcal{E}\{e(n)e^*(n)\} = \mathcal{E}\{|e(n)|^2\} \\ &= \mathcal{E}\left\{\left[\sum_{k=0}^{\infty} w_k^* x(n-k) - d(n)\right] e^*(n)\right\} \qquad (17A.3) \\ &= \sum_{k=0}^{\infty} w_k^* \mathcal{E}\left\{x(n-k)e^*(n)\right\} - \mathcal{E}\left\{d(n)e^*(n)\right\} \qquad (17A.4) \end{aligned}$$

Note that the filter coefficients are in general complex, so both the amplitude and phase are adjusted.

Applying the gradient operator of the kth filter coefficient to the cost function J in (17A.3), we get

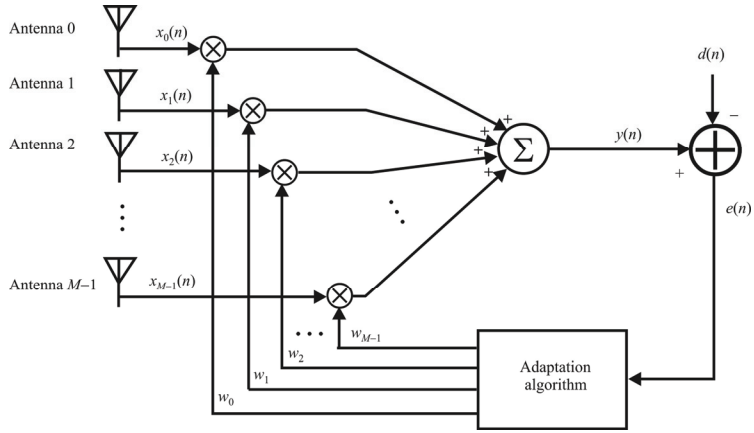


Figure 17A.1 Generic adaptive array. Weiner filter applied to antenna inputs.

$$\nabla_k J = \frac{\partial}{\partial w_k^*}\left[\sum_{k=0}^{\infty} w_k^* \mathcal{E}\left\{x(n-k)e^*(n)\right\} - \mathcal{E}\left\{d(n)e^*(n)\right\}\right] \tag{17A.5}$$

$$= \sum_{k=0}^{\infty} \frac{\partial}{\partial w_k^*}\left[w_k^* \mathcal{E}\left\{x(n-k)e^*(n)\right\}\right] - \frac{\partial}{\partial w_k^*}\mathcal{E}\left\{d(n)e^*(n)\right\}$$

$$= \mathcal{E}\{x(n-k)e^*(n)\} \tag{17A.6}$$

Therefore, when the filter is optimized, we get

$$\nabla_k J = \mathcal{E}\{x(n-k)e_0^*(n)\} = 0$$

which leads to the orthogonality principle.

Orthogonality Principle

$$\mathcal{E}\{x(n-k)e_0^*(n)\} = 0 \tag{17A.7}$$

Therefore, the necessary and sufficient condition for the cost function J to attain its minimum value is that the corresponding value of the estimation error $e_0(n)$ is orthogonal to each input sample that enters into the estimation of the desired response at time n.

Note that

$$\mathcal{E}\{y(n)e^*(n)\} = \mathcal{E}\left\{\sum_{k=0}^{\infty} w_k^* x(n-k)e^*(n)\right\}$$
$$= \sum_{k=0}^{\infty} w_k^* \mathcal{E}\left\{x(n-k)e^*(n)\right\} \quad (17A.8)$$

which proves the corollary to the orthogonality principle.

Corollary to the Orthogonality Principle

$$\mathcal{E}\left\{y_0(n)e_0^*(n)\right\} = 0 \quad (17A.9)$$

When the filter is optimized, the estimate of the desired response defined by the filter output, $y_0(n)$, and the corresponding estimation error, $e_0(n)$, are orthogonal to each other.

For the optimum filter, the minimum mean-square error equals the difference between the variance of the desired response and the variance of the estimate that the filter produces at its output.

$$e_0(n) = y_0(n) - d(n)$$
$$= \hat{d}(n|X_n) - d(n) \quad (17A.10)$$

where X_n is the input data space. Hence, evaluating the mean-square values of both sides of (17A.10), we get

$$J_{\min} = \mathcal{E}\left\{|e_0(n)|^2\right\} = \sigma_{\hat{d}}^2 - \sigma_d^2 \quad (17A.11)$$

17A.2.2 Weiner-Hopf Equations

The principle of orthogonality specifies the necessary and sufficient condition for the optimum operation of the filter as follows:

$$\mathcal{E}\left\{x(n-k)e_0^*(n)\right\} = 0$$
$$= \mathcal{E}\left\{x(n-k)\left[\sum_{i=0}^{\infty} w_{i0}x^*(n-i) - d^*(n)\right]\right\} = 0, \quad k = 0,1,2,\ldots \quad (17A.12)$$

Expanding (17A.12) and rearranging terms, we get

$$\sum_{i=0}^{\infty} w_{i0}\mathcal{E}\left\{x(n-k)x^*(n-i)\right\} = \mathcal{E}\left\{x(n-k)d^*(n)\right\}, \qquad k = 0,1,2,\ldots \tag{17A.13}$$

The two expectations in (17A.13) are

$$r(i-k) = \mathcal{E}\left\{x(n-k)x^*(n-i)\right\} \tag{17A.14}$$

is the autocorrelation of the filter input for a lag of $i - k$, and

$$p(k) = \mathcal{E}\left\{x(n-k)d^*(n)\right\} \tag{17A.15}$$

is the cross-correlation between the filter input and the desired response for a lag of k. Equations (17A.14) and (17A.15) are called the *Weiner-Hopf equations.* The normal equations are given as

$$\sum_{i=0}^{\infty} w_{i0}r(i-k) = p(k), \qquad k = 0,1,2,\ldots \tag{17A.16}$$

17A.2.3 Matrix Formulation of the Weiner-Hopf Equations

Let **R** denote the $L \times M$ correlation matrix of the tap inputs $x(n)$, $x(n-1),\ldots, x(n-L+1)$

$$\mathbf{R} = \mathcal{E}\left\{\vec{x}(n)\vec{x}^{\mathrm{H}}(n)\right\} \tag{17A.17}$$

where $\vec{x}(n)$ is the L-by-1 tap-input vector:

$$\vec{x}(n) = \left[x(n), x(n-1),\ldots, x(n-L+1)\right]^{\mathrm{T}} \tag{17A.18}$$

In expanded form, we have

$$\mathbf{R} = \begin{bmatrix} r(0) & r(1) & \cdots & r(L-1) \\ r^*(1) & r(0) & \cdots & r(L-2) \\ \vdots & \vdots & \ddots & \vdots \\ r^*(L-1) & r^*(L-2) & \cdots & r(0) \end{bmatrix} \tag{17A.19}$$

Let $\vec{p}$ denote the $L \times 1$ cross-correlation vector between the tap inputs of the filter and the desired response $d(n)$

$$\vec{p} = \mathcal{E}\left\{\vec{x}(n)\vec{d}^{\,\mathrm{H}}(n)\right\} \tag{17A.20}$$

$$= \left[p(0), p(1), \ldots, p(1-L)\right]^{\mathrm{T}} \tag{17A.21}$$

Rewrite the Weiner-Hopf equations in the matrix form:

$$\mathbf{R}\vec{w}_0 = \vec{p} \tag{17A.22}$$

where $\vec{w}_0 = \left[w_{0,0}, w_{0,1}, \ldots, w_{0,L-1}\right]^{\mathrm{T}}$. The optimum tap weight vector is therefore

$$\vec{w}_0 = \mathbf{R}^{-1}\vec{p} \tag{17A.23}$$

Properties of the Correlation Matrix

Property 1: The correlation matrix **R** is symmetric

$$\mathbf{R}^{\mathrm{T}} = \mathbf{R} \tag{17A.24}$$

Therefore,

$$\mathbf{R} = \begin{bmatrix} r(0) & r(1) & \cdots & r(L-1) \\ r(1) & r(0) & \cdots & r(L-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(L-1) & r(L-2) & \cdots & r(0) \end{bmatrix} \tag{17A.25}$$

Property 2: The correlation matrix **R** is Toeplitz, that is, the elements on its main diagonal are equal and so are the elements on any other diagonal parallel to the main diagonal.

Property 3: The correlation matrix **R** is positive semi-definite and is almost always positive definite. Let

$$a = \vec{u}^{\mathrm{T}}\vec{x}(n) = \vec{x}^{\mathrm{T}}(n)\vec{u}$$

$$\mathcal{E}\{a^2\} = \mathcal{E}\{\vec{u}^{\mathrm{T}}\vec{x}(n)\vec{x}^{\mathrm{T}}(n)\vec{u}\} = \vec{u}^{\mathrm{T}}\mathcal{E}\{\vec{x}(n)\vec{x}^{\mathrm{T}}(n)\}\vec{u} = \vec{u}^{\mathrm{T}}\mathbf{R}\vec{u}$$

Since $E\{a^2\} \geq 0$ it follows that $\vec{u}^{\mathrm{T}}\mathbf{R}\vec{u} \geq 0$. In practice, with probability close to 1, we can say that

$$\vec{u}^{\mathrm{T}}\mathbf{R}\vec{u} > 0$$

17A.2.4 The Correlation Matrix in Terms of Eigenvalues and Eigenvectors

Since the autocorrelation matrix **R** is positive definite, the eigenvalues of **R** are both real and positive. The traditional eigenvalue-eigenvector relationship is given by

$$\mathbf{R}\vec{q}_k = \lambda_k \vec{q}_k \tag{17A.26}$$

By defining

$$\mathbf{Q} = [\vec{q}_0, \vec{q}_1, \ldots, \vec{q}_{L-1}] \tag{17A.27}$$

and

$$\mathbf{A} = \mathrm{diag}[\lambda_0, \lambda_1, \ldots, \lambda_{L-1}] \tag{17A.28}$$

we can write

$$\mathbf{RQ} = \mathbf{Q\Lambda} \tag{17A.29}$$

When the eigenvalues $\{\lambda_0, \lambda_1, \ldots, \lambda_{L-1}\}$ are distinct, the associated eigenvectors are orthogonal and thus the matrix of eigenvectors **Q** is nonsingular. So we can multiply both sides by $\mathbf{Q}^{-1}$ yielding

$$\mathbf{Q}^{-1}\mathbf{RQ} = \mathbf{\Lambda} \tag{17A.30}$$

If the eigenvectors $\{\vec{q}_0, \vec{q}_1, \ldots, \vec{q}_{L-1}\}$ are normalized to have unit length, then

$$\mathbf{Q}^{-1} = \mathbf{Q}^{\mathrm{T}} \tag{17A.31}$$

which represents a unitary similarity transformation through which

$$\mathbf{Q}^{-1}\mathbf{RQ} = \mathbf{Q}^{T}\mathbf{RQ} = \mathbf{\Lambda} \tag{17A.32}$$

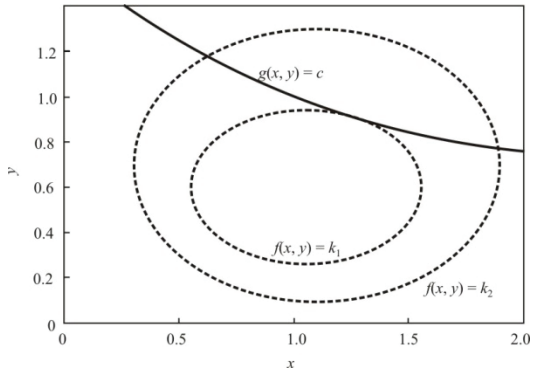


Figure 17A.2 Lagrange multipliers.

17A.3 Applications to Antenna Arrays

17A.3.1 Linearly Constrained Minimum Variance Beamformer

Our goal is to find the optimum set of filter coefficients, **w**, that minimizes the MSE value of the filter output $e(n)$, subject to the linear constraint as

$$\sum_{k=0}^{M-1} w_k^* e^{-jk\phi_0} = g$$

$$\vec{w}^{\mathrm{H}} \vec{a}(\phi_0) = g \tag{17A.33}$$

where g is the complex-value gain. To solve this constrained optimization formulation we use the method of Lagrange multipliers.

17A.3.1.1 Lagrange Multipliers

Lagrange multipliers find the extremals of a function $f(x, y)$ subject to the constraint $g(x, y) = c$. At an extremal, the gradient $\nabla_{\perp}$ of f and g align, and the tangential component $\nabla_{\|}$ of the derivative is zero (see Figure 17.A.2):

$$\nabla f(x,y) = \lambda \nabla \left[g(x,y) - c \right]$$

where λ is a scaling constant, called the Lagrange multiplier. Define the *Lagrangian* as

$$\mathcal{L} = f(x,y) - \lambda\left[g(x,y) - c\right]$$

Then

$$\frac{\partial \mathcal{L}}{\partial x} = 0, \qquad \frac{\partial \mathcal{L}}{\partial y} = 0, \qquad \frac{\partial \mathcal{L}}{\partial \lambda} = 0$$

The last equation is true as it reduces to the original constraint.

Generalizing to *N*-dimensional space, in *N* dimensions, with

$$\vec{x} = [x_1, x_2, \ldots, x_N]^{\mathrm{T}}$$

and with *K* constraints, we define the Lagrangian $\mathcal{L}$ as

$$\mathcal{L}(\vec{x}, \lambda_1, \lambda_2, \ldots, \lambda_K) = f(\vec{x}) + \sum_{k=1}^{K} \lambda_k g(\vec{x})$$

where the constant *c* of $g(\vec{x})$ is now included in the definition of $g(\vec{x})$. This yields $N + K$ equations and $N + K$ unknowns:

$$\nabla_{\vec{x}} \mathcal{L} = 0, \qquad \nabla_{\vec{\lambda}} \mathcal{L} = 0$$

with $\vec{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_K]^{\mathrm{T}}$.

17A.3.2 Linearly Constrained Minimium Variance Beamformer Revisited

Given the *L* zero-mean antenna element random signals for the LMVDR beamformer $\vec{x}(t) = [x_0(t), x_1(t), \ldots, x_{M-1}(t)]^{\mathrm{T}}$, with $\mathbf{R} = \mathcal{E}\{\vec{x}(t)\vec{x}(t)^{\mathrm{H}}\}$. Let

$$y(t) = \vec{w}^{\mathrm{H}} \vec{x}(t) \tag{17A.34}$$

be the beamformer output signal with beamformer weights $\vec{w}$. In this case we minimize the beamformer signal variance

$$P = \min_{\vec{w}} \mathcal{E}\{|y(t)|^2\} = \min_{\vec{w}} (\vec{w}^{\mathrm{H}} \mathbf{R} \vec{w}) \tag{17A.35}$$

with constraint of gain *g* in the look direction *s*,

$$\vec{w}^{\mathrm{H}}\vec{a} = g \tag{17A.36}$$

where

$$\vec{a} = e^{-j2\pi\vec{\mathcal{R}}\vec{s}fc^{-1}} \tag{17A.37}$$

where

$$\vec{\mathcal{R}} = [b_0, b_1, \ldots, b_{M-1}]^{\mathrm{T}} \tag{17A.38}$$

Now define the Lagrangian

$$\mathcal{L} = \vec{w}^{\mathrm{H}}\mathbf{R}\vec{w} + \lambda(\vec{w}^{\mathrm{H}}\vec{a} - 1) \tag{17A.39}$$

Using

$$\nabla_{\vec{w}}(\vec{w}^{\mathrm{H}}\mathbf{R}\vec{w}) = \mathbf{R}^{\mathrm{T}}\overline{\vec{w}} \tag{17A.40}$$

and

$$\nabla_{\vec{w}}(\vec{a}^{\mathrm{H}}\vec{w}) = \overline{\vec{a}} \tag{17A.41}$$

yields

$$\nabla_{\vec{w}}\mathcal{L} = \mathbf{R}^{\mathrm{T}}\overline{\vec{w}} + \lambda\overline{\vec{a}} = 0 \tag{17A.42}$$

This results in $\vec{w} = -\overline{\lambda}\mathbf{R}^{-1}\vec{a}$; inserting this into the constraint equation (17A.36) yields

$$\lambda = -\frac{1}{\vec{a}^{\mathbf{H}}\mathbf{R}^{-1}\vec{a}} \tag{17A.43}$$

and to the MVDR beamformer weights

$$\vec{w} = \frac{\mathbf{R}^{-1}\vec{a}}{\vec{a}^{\mathrm{H}}\mathbf{R}^{-1}\vec{a}} \tag{17A.44}$$

17A.3.3 Conventional Beamformer

Given M zero-mean antenna element random signals

$$\vec{x}(t) = [x_0(t), x_1(t), \ldots, x_{M-1}(t)]^{\mathrm{T}} \tag{17A.45}$$

with

$$\mathbf{R} = \mathcal{E}\{\vec{x}(t)\vec{x}(t)^{\mathrm{H}}\} \tag{17A.46}$$

Let

$$y(t) = \vec{w}^{\mathrm{H}}\vec{x}(t) \tag{17A.47}$$

be the beamformer output signal with beamformer weights $\vec{w}$. For the conventional beamformer we want to maximize the beamformer output power

$$P = \max_{\vec{w}} \mathcal{E}\left\{\left|y(t)\right|^2\right\} = \max_{\vec{w}}(\vec{w}^{\mathrm{H}}\mathbf{R}\vec{w}) \tag{17A.48}$$

with the constraint of unit norm of $\vec{w}$:

$$\left\|\vec{w}\right\|^2 = 1 \tag{17A.49}$$

Solving the gradient equation yields

$$\vec{w} = \frac{\vec{a}}{\sqrt{\vec{a}^{\mathrm{H}}\vec{a}}} \tag{17A.50}$$

The vector $\vec{w}$ can be interpreted as a spatial filter matched to the impinging signal.

Assuming observational data is used, the spatial spectrum is given by inserting the weight into the beamformer output power equation, yielding

$$P_{\mathrm{conv}}(s) = \frac{\vec{a}^{\mathrm{H}}\hat{\mathbf{R}}\vec{a}}{\vec{a}^{\mathrm{H}}\vec{a}} \tag{17A.51}$$

and

$$P_{\text{mvdr}}(\mathbf{s}) = \frac{1}{\vec{a}^{\text{H}}\hat{\mathbf{R}}\vec{a}} \tag{17A.52}$$

17A.3.4 MUSIC Beamformer

Recall the data model

$$\mathbf{R} = \mathbf{A}\mathbf{B}\mathbf{A}^{\text{H}} + \sigma_{\text{n}}^2\mathbf{I} = \mathbf{U}\Lambda\mathbf{U}^{\text{H}} + \sigma_{\text{n}}^2\mathbf{V}\mathbf{V}^{\text{H}} \tag{17A.53}$$

For all $\vec{a}$ in $\mathbf{A}$, define the ad hoc MUSIC beamformer spatial spectrum by

$$P_{\text{music}}(s) = \frac{\vec{a}^{\text{H}}\vec{a}}{\vec{a}^{\text{H}}\mathbf{\Pi}^{\perp}\vec{a}} \tag{17A.54}$$

with

$$\mathbf{\Pi}^{\perp} = \hat{\mathbf{V}}\hat{\mathbf{V}}^{\text{H}} \tag{17A.55}$$

with $\hat{\mathbf{V}}$ the noise subspace. This leads to improved sidelobe suppression and smaller beamwidth as compared to the conventional beamformer. The drawback is that these are data dependent and computation scales with L^3, as compared to L^1 for the conventional beamformer.

17A.3.5 Minimum Variance Distortionless Response Beamformer

For the special case of MVDR beamformer, the complex constraint constant g is unity. The optimum solution is

$$\vec{w}_0 = \frac{\mathbf{R}^{-1}\vec{a}(\phi_0)}{\vec{a}^{\text{H}}(\phi_0)\mathbf{R}^{-1}\vec{a}(\phi_0)} \tag{17A.56}$$

This beamformer is constrained to produce a distortionless response along the look direction corresponding to ϕ_0. In other words, the optimum beamformer is constrained to pass the target signal with unit response, while at the same time minimizing the total output variance. This variance minimization process attenuates interference and noise not originating at the AOA θ_0.

17A.4 Concluding Remarks

We have presented the fundamental principles underlying the Weiner filter in this appendix. That was followed by a discussion of applications of that filter to antenna arrays.

Chapter 18

Fractal Antennas

18.1 Introduction

A relatively recent entry into the field of small antenna design is based on fractal geometry. The term fractal comes from the Latin *fractus*, which means broken or irregular fragments, was originally coined by Mandelbrot [1] to describe a family of complex shapes that possess an inherent self-similarity in their geometrical structure. Since the pioneering work of Mandelbrot and others, a wide variety of applications for fractals continues to be found in many branches of science and engineering. One such area is fractal electrodynamics [2–7], in which fractal geometry is combined with electromagnetic theory for the purpose of investigating a new class of fractal electrodynamics-based antennas. We will outline the application of fractal structures to these antennas in this chapter.

This chapter is structured as follows. We begin with an introduction to the notion of what fractals are and how they relate to antennas. We focus mostly on monopoles, dipoles, and loops, and how these fundamental antenna shapes are modified when fractals are included. We briefly mention one of the more exotic antenna results—the Sirpinski gasket fractal antenna. Again, we focus on small antennas, as that is the focus of most of the fractal antenna research and that is why discussions on fractal antennas are included in a book on EW antenna systems.

18.2 Fractals

Fractal geometries are generated in an iterative fashion, leading to what are called self-similar structures [8]. This iterative generating technique can best be conveyed pictorially, as illustrated in Figure 18.1. The example shown is a

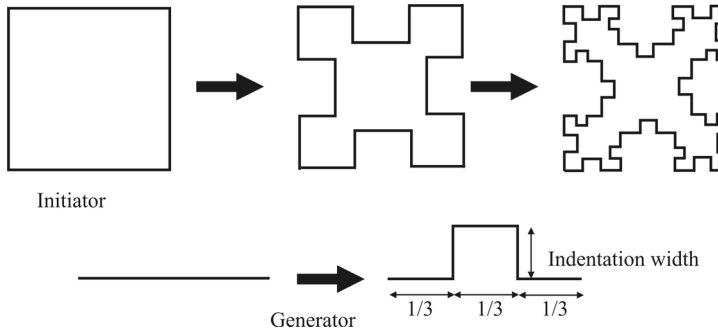


Figure 18.1 The iterative-generation procedure for a Minkowski island fractal. Each straight segment of the geometry is replaced with the generator. The initiator, the square, is shown, along with the first two generating iterations or prefractals. The indentation length need not be the same as the indentation width in the generator.

Minkowski island fractal. The starting geometry of the fractal, called the initiator, is a Euclidean square: each of the four straight segments of the starting structure is replaced with the generator, which is shown at the bottom of Figure 18.1. This iterative generating procedure continues for an infinite number of times. The final result is a curve with an infinitely intricate underlying structure that is not differentiable at any point. The iterative generation process creates a geometry that has intricate details on an ever-shrinking scale.

A second example of fractal generation in this manner is illustrated in Figure 18.2 using the Koch curve as the initiator. K3 is a monopole, vertically oriented when placed above a ground plane as illustrated in Figure 18.3.

The term "polarization" of higher order fractal antennas tends to lose its meaning. The bends and turns in a fractal antenna are the source of the accelerating and decelerating electrons that cause the radiation. Since these bends and turns are at odd angles, the resulting radiated EM wave will have a polarization pattern that is the vector sum of these sources, and vertical, horizontal, or elliptical, no longer carries the same meaning.

Even though a fractal is mathematically defined to be infinite in intricacy, this is obviously not a usable concept for the design of antennas or any other practical device. There normally reaches a point where adding more detail is unnecessary for practical application. Truncating the fractal generating process at some point less than infinity generates *prefractals* [9]. A prefractal drops the details that are not distinguishable for the particular applications.

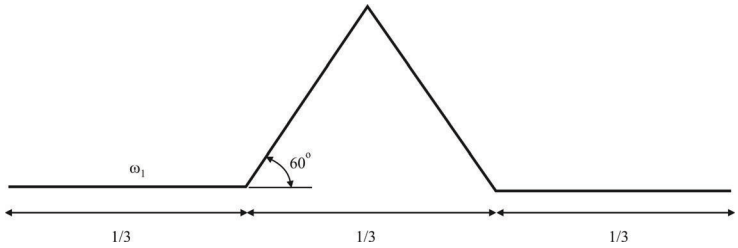


Figure 18.2 The Koch generator consists of two of three sides of an equilateral triangle. The two tails on the triangle are the same length as the sides of the triangle.

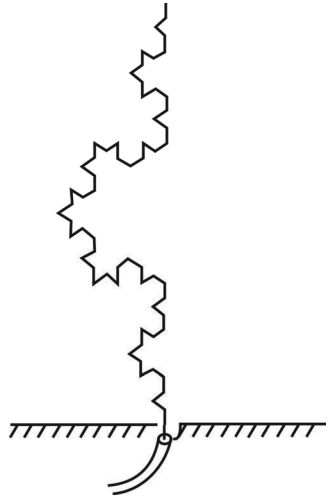


Figure 18.3 Vertically oriented Koch monopole.

18.3 Fractal Antennas

While Euclidean geometries are limited to points, lines, sheets, and volumes (even in n dimensions), fractals include the geometries that fall between these distinctions. Therefore, a fractal can be a line that approaches a sheet. The line can meander in such a way as to effectively almost fill the entire sheet. These space-filling properties lead to curves that are electrically very long, but fit into a compact physical space. This property can lead to the miniaturization of antenna elements.

For antennas, the intricacies that are much, much smaller than a wavelength in the band of useable frequencies can be dropped. Typically, the number of generating iterations required to reap the benefits of miniaturization is only a few before the additional complexities become indistinguishable.

18.3.1 Fractals as Wire Antenna Elements

A fractal can fill the volume occupied by the antenna in a more effective manner than the traditional Euclidean antenna. This leads to more effective coupling of energy from feeding transmission lines to free space in less volume.

18.4 Fractal Loop Antennas

The space-filling abilities of fractals fed as loop antennas can exhibit two benefits over Euclidean antennas. The first benefit is that the increased space-filling ability of the fractal loop means that more electrical length can fit in a smaller physical area. The increased electrical length leads to a lower resonant frequency, which effectively miniaturizes the antenna. The second benefit is that the increased electrical length can raise the input resistance of a loop antenna when it is used in a frequency range as a small antenna.

18.4.1 Miniaturization of a Loop Antenna at Resonance

The resonant frequency of a loop decreases as the generating iterations are increased. The fractal used to demonstrate this miniaturization of antennas is a Minkowski square loop. The generation procedure for the fractal is depicted in Figure 18.1. The initiator is a square, and the generator is depicted in the figure. The fractal is formed by displacing the middle third of each straight segment by some fraction of 1/3. This is called the indentation width, w, which can vary from 0 to 1. The resulting structure has five segments for every one of the previous

Table 18.1 The Fractal Dimension, Perimeter, and Height for the Various Cases of Indentation Widths and Generating Iterations.

Indentation Width Scaling Factor	Generating Iteration	Fractal Dimensions	Perimeter at Resonance (λ)	Height at Resonance (λ)
0	0	1	1.1100	0.2775
0.200	1	1.0982	1.2149	0.2680
0.200	2	1.0982	1.5372	0.2640
0.333	1	1.1562	1.2432	0.2543
0.333	2	1.1562	1.7980	0.2462
0.500	1	1.2280	1.2688	0.2379
0.500	2	1.2280	1.1239	0.2240
0.666	1	1.3019	1.2838	0.2222
0.666	2	1.3019	2.4411	0.2025
0.800	1	1.3640	1.2862	0.2097
0.800	2	1.3640	2.6850	0.1862
0.900	1	1.4132	1.2864	0.2010
0.900	2	1.4132	2.8361	0.1731

Source: [9].

We can see how the miniaturization increases as the fractal dimension increases.

iterations, but not all of the same scale. The fractal dimension can be determined by using the formula to calculate the self-similar dimension

$$k_1\left(\frac{1}{h_1}\right)^D + k_2\left(\frac{1}{h_2}\right)^D = 1 \tag{18.1}$$

The subscripts are required to differentiate between the two scales present in the generating iterations. For this example, $k_1 = 3$ and $h_1 = 3$, representing the horizontal segments, and $k_2 = 2$ and $h_2 = 3/w$, representing the vertical segments in the generator. For closed fractals that are formed by replicating the structure around the sides of a shape, the fractal dimension needs to be calculated based on the geometry of one side only, since the self-similarity is valid for straight segments and not the entire square. The fractal dimensions of the Minkowski fractal for various values of the indentation width and number of generating iterations are tabulated in Table 18.1.

The perimeter length of the Minkowski fractal loop is given by

$$perimeter_n = \left(1 + \frac{2}{3}w\right)^n perimeter_{n-1} \tag{18.2}$$

where n is the number of generating iterations. Because the term inside the parentheses is greater than 1, the perimeter of the fractal grows exponentially with increasing iterations. These values are tabulated in Table 18.1.

A corner-fed square loop antenna is shown in Figure 18.4. The first-order Minkowski fractal loop antenna is illustrated in Figure 18.5. The patterns for the x-y plane for these antennas are illustrated in Figure 18.6 [8]. The patterns in these two cases are similar, illustrating that the fractal antenna has similar performance as the larger square loop antenna. The resonant frequencies of the fractal loops were calculated using the MoM (Appendix C) at each generating iteration, and using various indentation widths. It can be seen that the resonant frequency of the loop antennas decreases as the number of generating iterations are increased, or as the indentation width is increased. The resulting heights of the loop antennas are tabulated in Table 18.1.

Two conclusions can be drawn from the analysis of the Minkowski loop antennas. The first is that the perimeter is not the only factor in determining the first resonance of the antenna. The perimeter increases at a rate faster than the decreasing resonant frequency. This is the expected result, since it is known that the perimeter of the fractal goes to infinity for an ideal fractal, and that lowering the resonant frequency saturates after a few generating iterations.

The second conclusion that can be drawn is that the resonant frequency of the antenna decreases as the fractal dimension increases. This is clearly the case when the indentation width is increased. Also, the dimension of the prefractal can generally be assumed to be the same as the ideal fractal, if enough iterations are used in the generation and if the increased intricacy is not discernable. It can be assumed that as the number of generating iterations is increased, the dimension increases from a one-dimensional line to the dimension calculated of the ideal fractal. Thus, the increase in the iterations increases the fractal dimension, which lowers the resonant frequency.

18.4.2 Increasing the Input Resistance of a Small Loop Antenna

An additional facet of the reduction in size of the resonant antenna is that the input resistance of small loops can be increased using fractals. As we discussed in Chapter 7, one problem with a small loop is that the input resistance is very low, making it difficult to couple power to the antenna. By using a fractal loop, the antenna can be brought closer to resonance, which raises the input impedance. **A** Koch island was used as a loop antenna to increase the input resistance. A Koch island is generated by forming a Koch curve on the three segments of an equilateral triangle as illustrated in Figure 18.7. The process is similar to the formation of the Minkowski loop, except that the generator is comprised of only four segments of equal length, instead of five segments of two different scales.

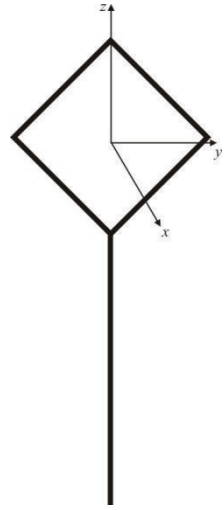


Figure 18.4 Corner-fed square loop antenna.

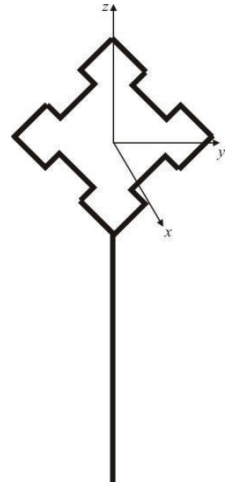


Figure 18.5 First-order Minkowski fractal loop antenna. This configuration is about 15% smaller than the antenna shown in Figure 18.4.

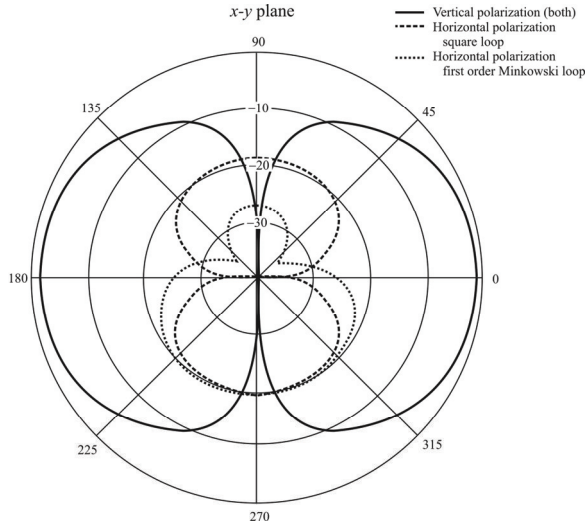


Figure 18.6 Corner-fed loop and first-order Minkowski fractal loop antennas *x-y* plane patterns. We can see that the vertical polarization patterns are virtually equivalent even though the Minkowski loop is smaller.

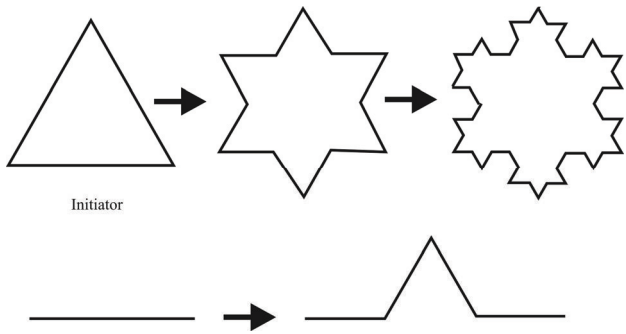


Figure 18.7 First three iterations of a Koch island fractal antenna.

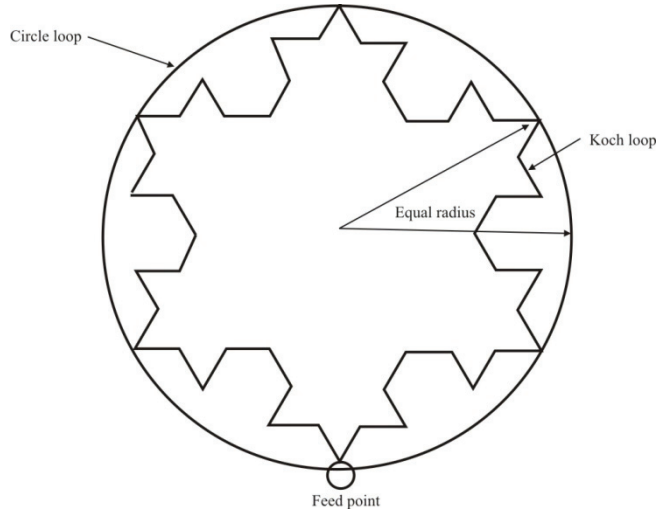


Figure 18.8 Fractal loop antenna design using a Koch island as the initiator. The input resistance of the circular loop with a perimeter of 0.26λ is 1.17Ω. Although only the first three iterations are shown here, after four iterations on the Koch island, the perimeter is 0.68λ with an input resistance of 26.7Ω.

For comparison, a circular loop with equal radius is shown in Figure 18.8, along with the fractal loop, which circumscribes the fractal loop. Numerical simulations of the two loop antennas shown in Figure 18.8 predicted an input resistance of 1.17Ω for the circular loop, and 26.7Ω for the fourth iteration of the Koch island. The perimeter lengths of the circular and fractal loop were 0.26λ and 0.68λ, respectively. Even though the loops require the same space, it is much easier to couple power to the fractal loop.

18.5 Fractal Dipole Antennas

In a way similar to how the loop antennas were studied in the previous section, this section will include a systematic look at the various geometries of fractal dipole antennas, with varying fractal dimensions at various stages of growth [10].

As mentioned in the previous section for Figure 18.8, the Koch fractal antenna is generated by replacing the middle third of each segment with two sides of an equilateral triangle. The resulting curve is comprised of four segments of equal length. As calculated above, the fractal dimension of the Koch curve is 1.2618. The first four iterations for the Koch dipole are illustrated in Figure 18.9. In this case each element of the dipole is treated as a Koch curve separately.

The resonant frequencies of the Koch dipole, as predicted by the numerical simulations, are plotted in Figure 18.10 [8]. Lowering the resonant frequency has the same effect as miniaturizing the antenna at a fixed resonant frequency. We can see how the planar structures have a similar miniaturization response.

18.6 Koch Fractal Monopole

The monopole is a well-studied antenna configuration and is included frequently in RF transmission designs. However it is problematic when its size is reduced to be compatible with EW systems. When the monopole is very short compared to the wavelength, the radiation resistance decreases, the stored reactive energy increases, and as a result the radiation efficiency decreases. Matching circuitry can become quite complicated. Fractal designs for monopoles is one solution to these problems. The fractal antenna not only has a large effective length, but the contours of its shape can generate a capacitance or inductance that can help to match the antenna to the circuit.

18.6.1 Koch Monopole

Constructing a Koch fractal monopole antenna from the iterative application of the Koch curve is illustrated in Figure 18.11. The length of the curve grows by a factor of 4/3 at each stage, yet always maintains exactly the same height. The limit object obtained after an infinite number of iterations is an ideal fractal curve with no derivative at any point.

Wheeler and Chu stated a lower bound for the quality factor (Q) of an antenna which is considered small when it can be fitted into a radiansphere, that is, a sphere with radius $r = \lambda/2\pi$ (see Chapter 14). However, application of the same arguments to fractal shaped antennas is not straightforward; in particular, some fractal antennas can be resonant even when considered small, and the spherical mode series used by Chu may not apply for ideal fractal shapes. It could be argued that an ideal fractal antenna is not feasible, but such an ideal object can be approached with a succession of feasible shapes which increase their length (but not their size) at each iteration. Whether the behavior of such a succession converges to the fundamental limit or not is an as yet unsolved question.

Five antennas (K1–K5) corresponding to the first five iterations of the Koch monopole were implemented [10] and their input impedances measured. Also, a standard monopole (K0) of the same height ($h = 6$ cm), but, of course, not the

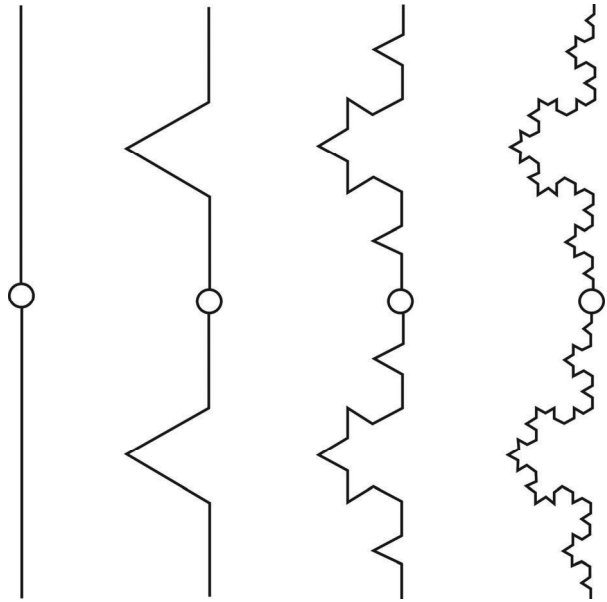


Figure 18.9 Construction of the Koch dipole. The first four iterations are shown.

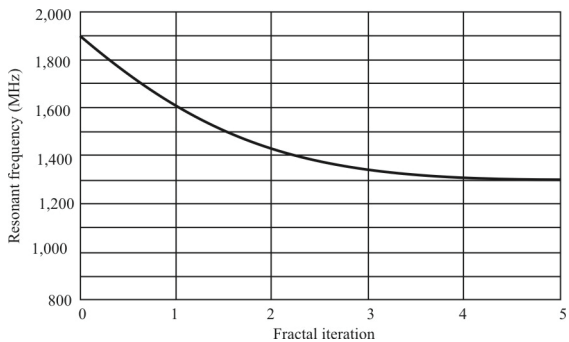


Figure 18.10 The computed resonant frequency for the three types of fractal dipoles as a function of the number of iterations used in the generating procedure. The reduction of the resonant frequency can be correlated with miniaturizing the antenna for a fixed resonant frequency.

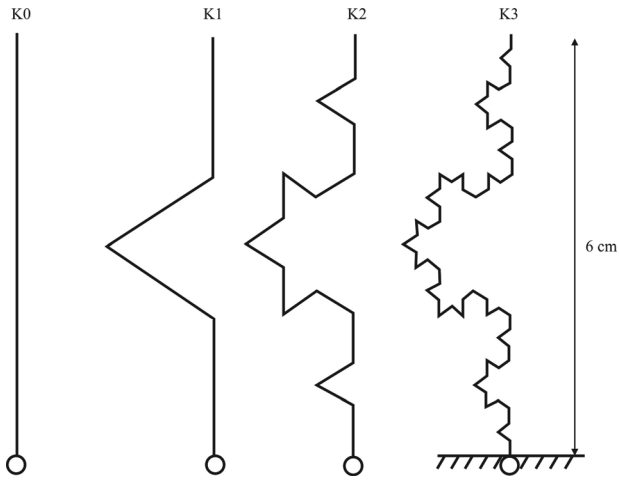


Figure 18.11 Iterative construction of fractal Koch monopole antenna. The first four iterations (K0–K3) of the curve were constructed and mounted in a monopole configuration over a ground plane (right). A linear monopole (K0) was built under the same conditions for comparison.

same length, was constructed for comparison. In addition, the first three iteration antennas together with the linear monopole (K0–K3) were simulated using the MoM method. Figure 18.12 [10] shows the input impedances of the six antennas, while their *Q*-factors with respect to frequency are shown in Figure 18.13. Notice in Figure 18.13 the resonant behavior of the fractal antennas below the small antenna limit: input resistance is increased and reactance is reduced at each iteration.

We see from these figures that the input resistance is increased each time the length (not the size) of the antenna is increased. Although both the radiation and ohmic resistance grow, it can be shown [11] that the latter is kept at least an order of magnitude below the radiation resistance over a wide range of frequencies below the small antenna limit frequency ($\beta h < 1$, $\beta = 2\pi/\lambda$). Also, the resonance frequency is shifted lower becoming resonant antennas even in the small antenna region. Also the *Q*-factor (Figure 18.13 [10]) is reduced at each iteration, broadening the bandwidth, tending to the fundamental limit. Although the ohmic losses of the antennas contribute to a reduction in the overall *Q,* the numerical results, which do not include ohmic losses, demonstrate the same tendency, and a comparison of experimental and numerical data leads to the conclusion that the ohmic losses do not play a major role in reducing the antenna *Q*.

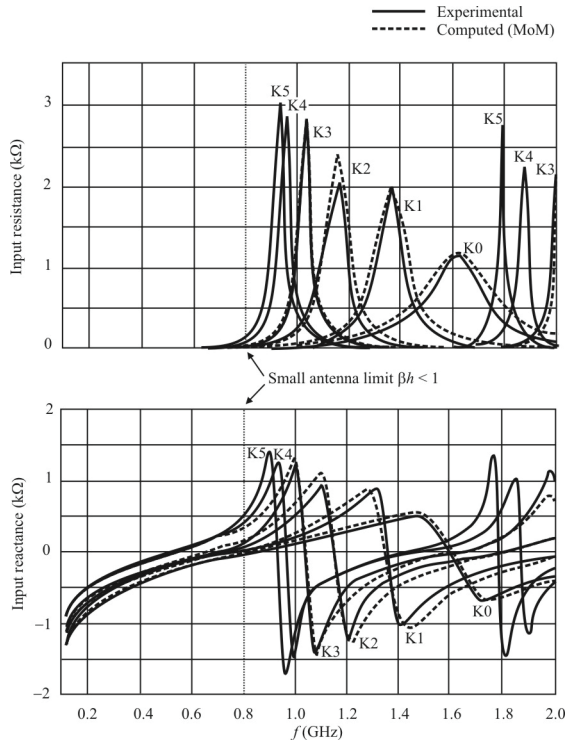


Figure 18.12 Input impedance K0–K5. Koch monopole.

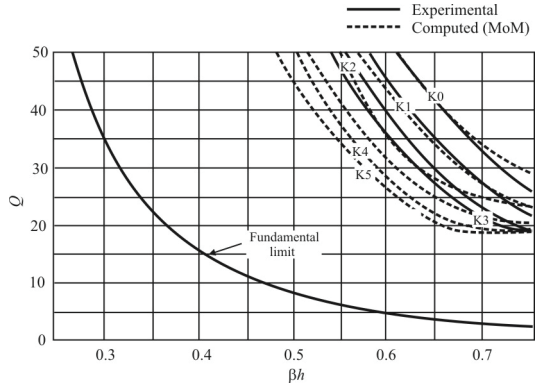


Figure 18.13 *Q*-factor K0–K5.

A likely explanation of such a behavior is found in the increasing number (unbounded as $K \to \infty$) of sharp corners and bends of the fractal monopole, which would enhance radiation, since antenna radiation is due to the acceleration (or deceleration) of electrons. Actually, the improvements are not surprising if the space filling properties of fractal objects are considered; while the 1-D linear monopole uses the available volume inside the radiansphere inefficiently, the Koch monopole with a dimension $D > 1$ ($D = 1.26$) better uses such a small space.

18.6.1.1 Summary

Both the experimental and numerical data in Figures 18.12 and 18.13 show that the Koch fractal monopole can improve the behavior of a small monopole. The radiation resistance is increased and the reactance is reduced even in the small antenna frequency region. The Q is reduced at each iteration as well, tending to the Chu limit.

18.6.2 The Iterative Function System Algorithm

The shape of the fractal monopole antenna can be determined by an iterative mathematical process described by an *iterative function system* (IFS) algorithm, which is based upon a series of affine transformations[1] [12, 13]. An affine transformation in the plane ω can be described by

$$\omega\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix} = \mathbf{A}\vec{x} + \vec{t} \tag{18.3}$$

This affine transformation can also be written in a short and convenient way as

$$\omega = [a, b, c, d, e, f] \tag{18.4}$$

The matrix **A** can also be written as:

$$\mathbf{A} = \begin{bmatrix} r\cos\theta & -r\cos\theta \\ r\sin\theta & r\cos\theta \end{bmatrix} \tag{18.5}$$

---

[1] An affine transformation in a plane is a rotation in the plane described by $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ followed by a possible translation described by $\vec{t} = \begin{bmatrix} e & f \end{bmatrix}^{\mathrm{T}}$.

where r is a scaling factor and θ is the rotation angle of the rotation. The vector $\vec{t}$ is a translation on the plane. Therefore, the affine transformation can be written as

$$\omega = [r\cos\theta, -r\sin\theta, r\sin\theta, r\cos\theta, e, f] \tag{18.6}$$

The Koch monopole fractal is based on the transformation illustrated in Figure 18.2, which has a scaling factor of $r = 1/3$ and rotation angles of $\theta = 0^o$, 60^o, -60^o, and 0^o. There are four basic segments that form the basis of the Koch fractal curve used to design the antenna, which are shown in Figure 18.2.

The IFS for the Koch curve IFS can be written as

$$\omega_1 = [1/3, 0, 0, 1/3, 0, 0] \tag{18.7}$$

$$\omega_2 = [1/3\cos 60, -1/3\sin 60, 1/3\sin 60, 1/3\cos 60, 1/3, 0] \tag{18.8}$$

$$\omega_3 = [1/3\cos 60, 1/3\sin 60, -1/3\sin 60, 1/3\cos 60, 1/2, 1/(2\sqrt{3})] \tag{18.9}$$

$$\omega_4 = [1/3, 0, 0, 1/3, 2/3, 0] \tag{18.10}$$

The fractal shape in Figure 18.2 represents the first iteration of the Koch fractal curve. Additional iterations are performed on each segment by applying the IFS. The effective length of a Koch fractal curve increases by (4/3) for each iteration. Therefore, for an initial height h, the effective length for each antenna is

$$l = h\left(\frac{4}{3}\right)^n \tag{18.11}$$

18.6.3 Resonant Frequency

In this section, we analyze the Koch fractal monopole by using the MoM provided in the NEC program. Four antennas over a ground plane were investigated: A straight monopole, Koch fractal antennas with one, two, and three iterations. The effective height for each antenna is summarized in Table 18.2. Diagrams of the Koch curves for $n = 0$ through 3 are illustrated in Figure 18.11. For each case, the total physical height of the antenna was $l = 6$ cm. The feed point for all of the antennas was located at the antenna/ground plane interface. The wire was considered to be a perfect conductor with a radius of 100 μm.

Table 18.2 Effective Length of Low-Order Koch Monopoles with a 6 cm Physical Length

Antenna	Effective Length
K0	6.00 cm
K1	8.00 cm
K2	10.67 cm
K3	14.22 cm

When a quarter-wavelength monopole is fed at the base, the largest current should occur at the base and no current should exist at the end. From the MoM, the program was able to compute the real and imaginary parts of the input impedance over a frequency range from 250 MHz to 3 GHz. It is shown that as the effective length of the antenna increases, the resonant frequency decreases. The resulting input impedances are shown in Figure 18.14.

From Figure 18.14, we can see that as the number of iterations of the fractal antenna increases, the resonant frequency decreases. The resonant frequency of the simulated plots also agrees quite well with the frequency of the effective length of the antenna. Since the antennas were a quarter wavelength placed over a ground plane, the corresponding frequency is given by

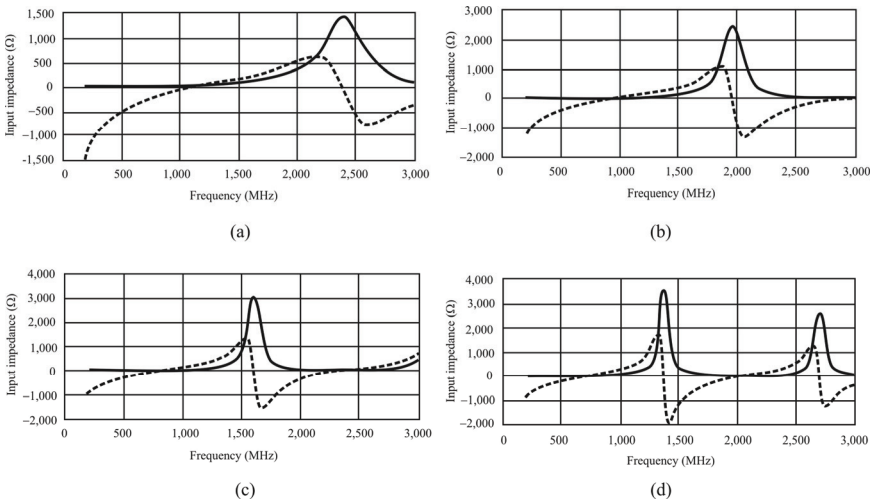


Figure 18.14(a–d) Input impedance Koch monopole. Input impedances of various iterations of the Koch fractal antenna (solid = real, dashed = imaginary): (a) K0, (b) K1, (c) K2, and (d) K3.

Table 18.3 Table to Demonstrate How Resonant Frequency Decreases as the Iterations Increase

Antenna	Predicted Resonant Frequency	Resonant Frequency from Quarter-Wavelength Effective Length
K0	1.203 GHz	1.250 GHz
K1	983 MHz	938 MHz
K2	836 MHz	703 MHz
K3	740 MHz	527 MHz

$$f = \frac{c}{4l} \tag{18.12}$$

where c is the speed of light and l is the effective length. At that frequency, the fractal antenna should be similar to the quarter-wave monopole since the effective length is a quarter wavelength. The results are summarized in Table 18.3, where we see that the correspondence is reasonably good.

Just as for the Koch dipole, whose resonant frequency characteristic is seen in Figure 18.10, the frequency does not decrease as might be expected as the number of iterations increased. It is also interesting to note that, as the number of iterations of the fractal increases, there are more resonant frequencies, thus resulting in a multiband antenna.

18.6.4 Input Impedance

According to Balanis [14], the input impedance of a small linear dipole of length l and wire radius a can be approximated with

$$Z_{\text{in}}^{\text{dipole}} \approx 20\pi^2 \left(\frac{l}{\lambda}\right)^2 - j120 \frac{\ln\left(\frac{l}{2a}\right) - 1}{\tan\left(\pi \frac{l}{\lambda}\right)} \tag{18.13}$$

For comparison purposes, the length of the dipole was set to 12 cm, which is twice that of the Koch monopoles described in the previous section. The wavelength varied according to the resonant frequencies of the Koch monopoles. The radius of the wire remained the same at 100 μm. Finally, we divide (18.13) by 2, since (18.13) is for a small linear dipole:

$$Z_{\text{in}}^{\text{monopole}} = \frac{1}{2} Z_{\text{in}}^{\text{dipole}} \tag{18.14}$$

Table 18.4 The Input Impedance of a 6 cm Long Koch Fractal Monopole and a 6 cm Long Electrically Short Monopole

Frequency	Input Impedance for Koch Monopole	Input Impedance for Electrically Short Monopole
983 MHz	23.1Ω, K1	13.89 – *j*134Ω
836 MHz	17.0Ω, K2	7.80 – *j*266Ω
740 MHz	13.4Ω, K3	4.39 – *j*415Ω

The resulting computed values, in comparison with various iterations of the Koch fractal monopole, are shown in Table 18.4. For all three cases, the real part of the input impedance of the electrically short monopole is less than that for the Koch monopole. Also, for all three cases, the electrically short monopole has a significant negative reactive component, which becomes quite large at lower frequencies. The small real impedance and large negative reactance of the electrically short monopole can make the matching difficult.

18.6.5 Current Distribution

As expected, when the fractal antenna K3 reached its first resonance at 740 MHz, straight monopole K0 still acted like a electrically short monopole with a significant negative reactance. Since antenna K3 is at resonance at 740 MHz, the current distribution resembled a sinusoid pattern. However, K0 resembled a triangular shaped distribution since its length was less than the effective length of K3. The normalized current distributions for these two antennas are shown in Figure 18.15.

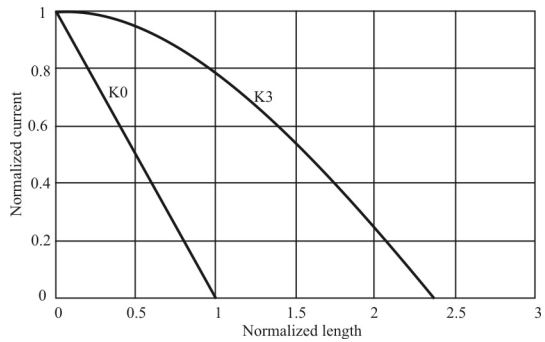


Figure 18.15 Normalized current distributions. Approximate current distributions for the K3 fractal antenna and the K0 straight monopole at 740 MHz.

18.6.6 Radiation Patterns

Using the straight monopole K0 and the three iteration fractal K3, the corresponding radiation patterns were plotted. Each radiation pattern was generated at the resonant frequency of the antenna, which was given in Table 18.3. The computed patterns are shown in Figures 18.16 and 18.17.

It is interesting to note that for the E-plane radiation pattern, the fractal antenna also has a null at 90^o. Due to the symmetry of the fractal antenna, the electric and magnetic fields cancel in this direction. The contours of the fractal antenna in the H-plane are pointing into the 0 degree direction. As shown in the radiation pattern, this fractal antenna K3 radiates omnidirectionally. Due to the symmetry of the fractal antenna at resonance, the electric and magnetic fields are added and cancelled in the far field to give a symmetrical pattern. Also note that the fractal antenna has slightly less gain than the straight monopole. This is due to the fact that the fractal antenna is slightly less efficient than the straight monopole.

18.6.7 Quality Factor and Bandwidth

As discussed previously, the quality factor, Q, is used to measure the ratio of an antenna's radiated energy to its stored energy. Small antennas, such as an electrically short monopole, tend to have a high Q (see Chapter 14). In the case of a small monopole, as the physical length decreases, the radiation resistance tends to decrease and the reactive component tends to increase, and therefore the Q increases. In general, the Q of an antenna is inversely proportional to its bandwidth. Narrow bandwidth antennas are usually not preferred because of the difficulty of matching. In addition, an inordinate number of them would be required for most EW applications.

The limits of an electrically small antenna can be analyzed when it is assumed that the antenna is enclosed within a radiansphere of radius r. Due to the variations of the current inside the radiansphere, the field outside of the radiansphere can be described as a set of orthogonal spherical vector waves. The real power radiated is due to the propagating modes, while the reactive power is due to all modes.

However, when the radiansphere becomes very small, there are no propagating modes, and thus there is very little real radiated power.

As discussed in Chapter 14, the limit for the smallest possible Q for any antenna within the radian sphere, regardless of its shape, is given by

$$Q = \frac{1 + 2(\beta r)^2}{(\beta r)^3 [1 + (\beta r)^2]} \tag{18.15}$$

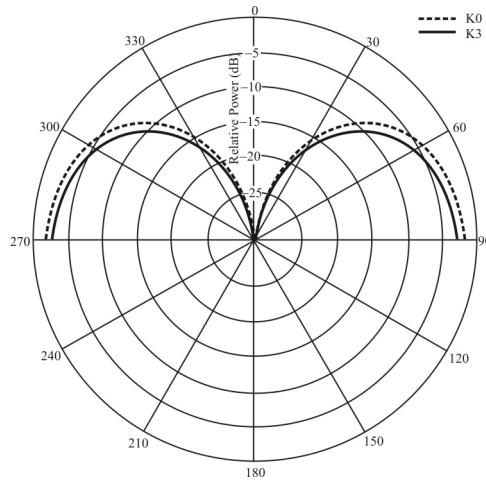


Figure 18.16 Koch radiation E-pattern for K0 and K3.

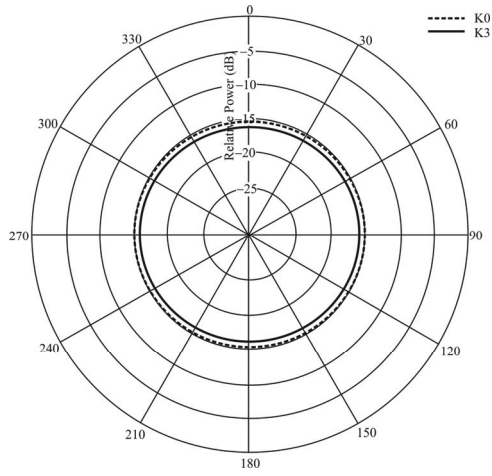


Figure 18.17 The H-plane radiation pattern for the straight monopole K0 and the fractal antenna K3 (dashed = K0, solid = K3).

According to Chu [15], the Q of a small antenna may be described according to the stored electric and magnetic energies W_e and W_m, respectively, the frequency ω, and the average radiated power P_r as

$$Q = \omega \frac{2W_e}{P_r}, \qquad W_e >> W_m \tag{18.16}$$

$$Q = \omega \frac{2W_m}{P_r}, \qquad W_m >> W_e \tag{18.17}$$

The stored electric and magnetic energies can be related to the input impedance of a lossless one-port network with input reactance X_{in} and susceptance B_{in}. The current and voltage at the input terminals of the antenna are I and V, respectively. Then

$$W_e = \frac{|I|^2}{8}\left(\frac{\partial X_{in}}{\partial \omega} - \frac{X_{in}}{\omega}\right) = \frac{|V|^2}{8}\left(\frac{\partial B_{in}}{\partial \omega} + \frac{B_{in}}{\omega}\right) \tag{18.18}$$

$$W_m = \frac{|I|^2}{8}\left(\frac{\partial X_{in}}{\partial \omega} + \frac{X_{in}}{\omega}\right) = \frac{|V|^2}{8}\left(\frac{\partial B_{in}}{\partial \omega} - \frac{B_{in}}{\omega}\right) \tag{18.19}$$

The power dissipated by the antenna is given by

$$P_L = \frac{1}{2}|I|^2 R_{in} = \frac{1}{2}|V|^2 G_{in} \tag{18.20}$$

Substituting (18.18), (18.19), and (18.20) into (18.16) and (18.17), the following ensue

$$Q = \frac{\omega}{2R_{in}}\left(\frac{\partial X_{in}}{\partial \omega} - \frac{X_{in}}{\omega}\right), \qquad W_e >> W_m \tag{18.21}$$

$$Q = \frac{\omega}{2R_{in}}\left(\frac{\partial X_{in}}{\partial \omega} + \frac{X_{in}}{\omega}\right), \qquad W_m >> W_e \tag{18.22}$$

Finally, the equivalent quality factor can by be found by

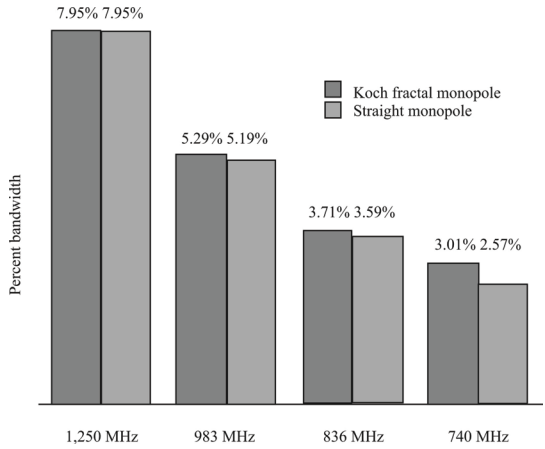


Figure 18.18 The bandwidth of a 6 cm long Koch fractal antenna in comparison to a 6 cm long straight monopole.

$$Q = \frac{\omega}{2R_{\text{in}}}\left(\frac{\partial X_{\text{in}}}{\partial \omega} + \left|\frac{X_{\text{in}}}{\omega}\right|\right) \tag{18.23}$$

It has been shown that the quality factor for the fractal Koch monopole is quite low for a small antenna. As the number of iterations of the fractal increases, the trend is that the quality factor approaches the fundamental lower limit described by (18.15). A plot demonstrating the effect of the number of iterations and the quality factor is shown in Figure 18.13 [10].

The four antennas that were investigated in the previous sections were analyzed to find their respective bandwidths. For each case, the bandwidth was determined according to the frequency range where the return loss S11 was less than –10 dB, which corresponds to a VSWR of less than ~2. In addition, it was assumed that a matching network was made corresponding to the input impedance at the respective frequencies. For comparison purposes, the bandwidth of four straight monopoles was found. All antennas had a physical length of 6 cm. Therefore, for frequencies below 1,250 MHz, the antennas can be considered electrically short. The resulting bandwidths are summarized in Figure 18.18. The results show that the Koch fractal antenna does indeed have a larger bandwidth compared to its equivalent electrically short monopole, although the increase is quite small. As the frequency decreases, this becomes more apparent.

In order to improve the bandwidth of an antenna and thus minimize the Q, the antenna must have a geometrical configuration that is spatially efficient within the volume of the radiansphere. For example, a monopole antenna is not very spatially

efficient since it consists of just one thin wire within the radiansphere. On the other hand, the fractal antenna is fairly spatially efficient within the radiansphere, and thus has a lower Q. In theory, as the antenna becomes more spatially efficient, the Q of the antenna approaches the fundamental limit as shown in Figure 18.13. However, in reality, the Q of the antenna rarely approaches the fundamental limit.

18.6.8 Multiband Characteristics

Not only are fractal antennas broadband, but they also possess multiband characteristics. This is seen in Figure 18.14(d). Over a large frequency span, the fractal antenna can have many different resonant frequencies. This is due to the coupling between the wires. As more contours and iterations of the fractal are added, the coupling becomes more complicated and different segments of the wire resonate at different frequencies. It can be shown that as the number of iterations of the fractal increases, the antenna has more resonant frequencies. This is a clear advantage over the electrically short monopole, which tends to have fewer resonant frequencies.

18.6.9 Radiation Efficiency

One way to determine the radiation efficiency of an antenna is to use the efficiency, the Wheeler cap method [16]. In this method, the total resistance is measured under two cases: the antenna is measured in free space, or the antenna is measured within a cylinder of metal, which is called a Wheeler cap. In free space, when electromagnetic waves encounter metal, it effectively looks as a short circuit and the waves are reflected. Therefore, when measuring the total resistance of an antenna with metal surrounding it, the radiation resistance is 0 while the total resistance is comprised of only loss resistance. Formulas for the Wheeler cap method are described by

$$\text{Resistance with Wheeler cap: } R_{\text{WC}} = R_{\text{loss}} \tag{18.24}$$

$$\text{Resistance in free space: } R_{\text{fs}} = R_{\text{loss}} + R_{\text{radiated}} \tag{18.25}$$

$$R_{\text{radiated}} = R_{\text{fs}} - R_{\text{loss}} \tag{18.26}$$

$$\text{Efficiency: } \eta = \frac{R_{\text{radiated}}}{R_{\text{radiated}} + R_{\text{loss}}} \tag{18.27}$$

The measured efficiency results are shown in Figure 18.19 [13]. All four measured antennas had the same height of 6 cm. Through these measurements, it was found that with each additional iteration of the fractal, the efficiency of the antenna

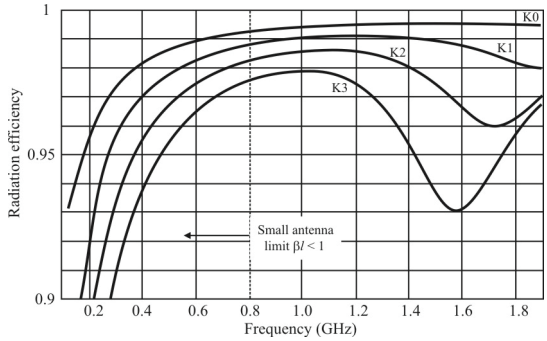


Figure 18.19 Radiation efficiency measurements for various iterations of the Koch fractal monopole.

slightly decreased. This is due to the fact that as the number of iterations increased, the radiation resistance decreased.

18.6.10 Applications

The Koch fractal monopole can be used in many applications that require a small antenna, which is generally true for most EW applications. Some key advantages of the fractal antenna include its ability to operate on multiple bands, fairly wide bandwidth, and its size. Also, fractal antennas can sometimes fit within the casing of the device, thus eliminating the protruding antenna.

18.7 The Sirpinski Gasket Fractal Antenna

Although the Koch fractal antenna described in the last section is a great approach to minimizing the size of a monopole, its radiation pattern is omnidirectional, which may not be preferable in some applications. If the application needs an antenna with a directional radiation pattern, a fractal array may be implemented. A fractal array can consist of elements that are arranged in a fractal spatial arrangement [17]. Also, the fractal array can consist of many uniformly spaced elements, where the elements are fed according to a fractal distribution [18]. An example of a commonly used fractal array is the *Sierpinski gasket*, which is shown in Figure 18.20.

A major advantage of the Sierpinski fractal array is its low sidelobes and easy matching to 50Ω. At lower frequencies, the Sierpinski gasket has almost an omnidirectional pattern, which could serve as an alternative to the Koch fractal antenna. The measured radiation patterns of the Sierpinski gasket at a high

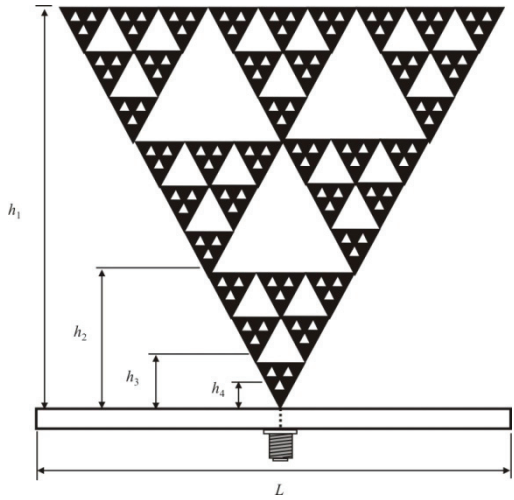


Figure 18.20 The Sierpinski gasket fractal array.

frequency of 7.46 GHz and a low frequency of 0.92 GHz are shown in Figure 18.21 [19].

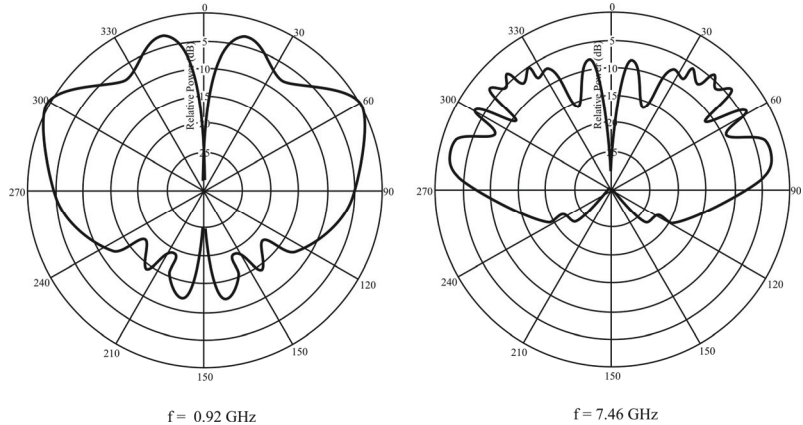


Figure 18.21 Radiation patterns for the Sierpinski gasket fractal array.

18.8 Concluding Remarks

We briefly described the characteristics of fractal antennas in the chapter. Only the fundamental properties were presented, as this is a fairly new field and many of these properties are just emerging. It is hoped that the reader can grasp some of the nuances of this developing diverse and interesting field.

References

[1] Mandelbrot, B. B., *Fractal Geometry*, New York: Freeman, 1977.

[2] Yamaguti, M., M. Hata, and J. Kigami, *Translations of Mathematical Monograpohs Volumne 167: Mathematics of Fractals*, Providence, RI.: American Mathematical Society, 1993.

[3] Kritikos, H. N., and D. L. Jaggard (eds.), *Recent Advances in Electromagnetic Theory*, New York: Springer-Verlag, 1990.

[4] Wemer, D. H., and R. Mittra, *Frontiers in Electromagnetics*, New York: IEEE Press, 2000.

[5] Wemer, D., R. I. Haupt, and P. L. Wemer, "Fractal Antenna Engineering: The Theory and Design of Fractal Antenna Arrays," *IEEE Antennas and Propagation Magazine*, Vol. 41, No. 5, 1999, pp. 37–59.

[6] Gianvittorio, J., "Fractal Antenna Design Characterization and Applications," Masters Thesis, University of California, Los Angeles, 2000.

[7] Hodges, R. E., and Y. Rahmat-Samil, "An Iterative Current-Based Method for Complex Structures," *IEEE Transactions on Antennas and Propagation*, Vol. AP-15, No. 2, 1997, pp. 265–276.

[8] Gianvittorio, J. P., and Y. R. Samil, "Fractal Antennas: A Novel Antenna Miniaturization Technique and Applications," *IEEE Antenna and Propagation Magazine*, Vol. 44, No. 1, February 2002, pp. 20–36.

[9] Gouyet, J.-F., *Physics and Fractal Structures,* New York: Springer, 1996.

[10] Puente, C., J. Romeu, R. Pous, J. Ramis, and A. Hijazo, "Small but Long Koch Fractal Monopole," *Electronics Letters*, Vol. 34, No. 1, January 8, 1998, pp. 9–10.

[11] Puente, C., "Fractal Antennas," Ph.D. Dissertation, Department of Signal Theory and Communications, Universitat Politecnica de Catalunya, June 1997.

[12] Werner, D. H., and S. Ganguly, "An Overview of Fractal Antenna Engineering Research*," IEEE Antennas and Propagation Magazine*, Vol. 45, No. 1, February 2003, pp. 38–56.

[13] Baliarda, C. P., J. Romeu, and A. Cardama, "The Koch Monopole: A Small Fractal Antenna," *IEEE Transactions on Antennas and Propagation*, Vol. 48, November 2000, pp. 1773–1781.

[14] Balanis, C. A., *Antenna Theory: Analysis and Design*, 2nd ed., New York, John Wiley & Sons., 1997

[15] Chu, L. J., "Physical Limitations on Omni-Directional Antennas," *Journal of Applied Physics*, Vol. 19, December 1948, pp. 1163–1175.

[16] Agahi, D., and W. Domino, "Efficiency Measurements of Portable-Handset Antennas Using the Wheeler Cap," *Applied Microwave and Wireless*, Vol. 12, No. 6, June 2000, pp. 34–42.

[17] Baliarda, C. P., "Fractal Design of Multiband and Low Side-Lobe Arrays," *IEEE Transactions on. Antennas and Propagation*, Vol. 44, No. 5, May 1996, pp. 730–739.

[18] Baliarda, C. P., C. B. Borau, M. N. Rodero, and J. R. Robert, "An Iterative Model for Fractal Antennas: Application to the Sierpinski Gasket Antenna," *IEEE Transactions on Antennas and Propagation*, Vol. 48, No. 5, May 2000, pp. 713–718.

[19] Romeu, J., and J. Soler, "Generalized Sierpinski Fractal Multiband Antenna," *IEEE Transactions on Antennas and Propagation*, Vol. 49, No. 8, August 2001, pp. 1237–1239.

Chapter 19

Genetically Designed Antennas

19.1 Introduction

Charles Darwin discovered the processes by which nature selects and optimizes organisms fit for life in the mid-1800s [1]. With the advent of computers and powerful computational techniques, we can now apply Nature's optimization processes in the form of *genetic algorithms* (GAs) to devices built using Maxwell's equations covered in Chapter 2.

A new antenna design technique that has been developed relies on genetic algorithms for determining the structure. In that technique, structure seeds are given to the algorithm and antenna elements are added, or moved in space, one at a time and at random, to the successor structure, the first of which is the seed. A measure of performance is required so that the structures can be evaluated at each stage. If the performance of the new structure is better than the last stage, then the design is kept for the next pass. If not, it is discarded. The normal antenna parameters, such as frequency coverage and bandwidth, are specified. Some of the performance parameters used to evaluate each antenna configuration are gain and directionality [2–9].

These algorithms are called genetic because they mimic the way biological entities evolve. Bad traits typically die off after time, while good traits tend to last. The resultant structures can have very interesting shapes, some of which are counterintuitive.

The structure of this chapter is as follows. We begin with a discussion of genetic algorithms and their fundamental characteristics. That is followed by a section on an antenna design example using the GA approach. Then the design of a crooked wire antenna is presented, using a GA algorithm that leads to a surprising result. Next the design of broadband loaded monopole antennas using conventional broadband RLC designs and GA designs. The two resulting designs are compared. Finally, we include a section on simulated annealing, which is similar to genetic algorithms, but fundamentally different.

19.2 Genetic Algorithms

19.2.1 Introduction to Genetic Algorithms

These GAs all have one feature in common: they all employ objective functions that are used as the measure of "goodness" of a particular structure. In our case here, that "structure" is a particular physical configuration of an antenna. Such a function might calculate the gain of an antenna if that is the important parameter. The function may incorporate more than one parameter, perhaps weighted by some method to apply more weight to some parameters over others.

The objective functions that classically have arisen in electromagnetic optimization problems are often highly nonlinear, inflexible, multiextremal, and nondifferentiable. In addition, they almost always have employed *deterministic optimization methods* (DOMs) [10]. DOMs are known to have important drawbacks. GAs (along with Monte Carlo techniques and simulated annealing) belong to a small but growing class of so-called global optimizers, which are stochastic in nature and, therefore, less prone to converge to a weak local optimum than DOMs [11–16].

19.2.2 Genetic Algorithms

GAs are function optimizers, that is, methods for extracting extrema (minima or maxima) of a given objective function $f(\vec{x})$ where $\vec{x} = \{x_l \,|\, l = 1, 2, \ldots, N_x\}$. In design problems, the x_l are parameters describing salient features of a design, and $f(\vec{x})$ is a measure of system performance to be either maximized or minimized. The x_l can be continuous or discrete, and complex, real- or integer-valued.

GAs differ from most traditional optimization methods in two ways. First, they do not necessarily operate directly on the design parameters $\vec{x}$, and second, they simultaneously optimize entire *populations* of designs at once, not a single design at a time. In general, GAs do not operate directly on the parameter vector $\vec{x}$, but on a symbolic representation of $\vec{x}$, known as a *chromosome.* A chromosome is a concatenation of *genes* that decode to the $\vec{x}$, and is symbolically denoted as

$$\vec{p} = \{g_j \,|\, j = 1, 2, \ldots, N_{gl}\} \tag{19.1}$$

where N_{gl} is the genetic length and there is a correspondence between the x_l and the g_j given by

$$\vec{p} \leftrightarrow \overbrace{g_1 g_2 \cdots g_{N_1}}^{\vec{x}_1} \overbrace{g_{N_1+1} g_{N_1+2} \cdots g_{N_2}}^{\vec{x}_2} \cdots \overbrace{g_{N_{N_x-1}} g_{N_{N_x-1}+1} \cdots g_{N_{N_x}}}^{\vec{x}_{N_x}} \tag{19.2}$$

where $N_{gl} \triangleq N_{N_g}$. In allelic[1] GAs, the genes g_j are selected from a finite alphabet (most often the binary alphabet {0, 1}), whose members are known as *alleles*, which together decode to a discrete, or perhaps discretized continuous, parameter. For example, in binary-coded GAs, real-valued parameters can be decoded through a simple linear transformation

$$x_l = x_l^{\min} + 2^{-N_{l-1}} \frac{x_l^{\max} - x_l^{\min}}{2^{N_l - N_{l-1}} - 1} \sum_{j=N_{l-1}+1}^{N_l} g_j 2^{j-1} \tag{19.3}$$

where x_l takes values ranging from $x_l^{\min}$ to $x_l^{\max}$, and $N_0 = 0$.

The second difference between GAs and classical optimizers is that GAs do not work on a single chromosome at a time, but on a whole population of chromosomes

$$\mathbf{P} = \{\vec{p}_i \,|\, i = 1, 2, \ldots, N_{\text{pop}}\} \tag{19.4}$$

The GA is composed of operators that produce a succession of populations whose members will have generally improving objective function values. Given a population $\mathbf{P}^k = \{\vec{p}_i^{\,k} \,|\, i = 1, 2, \ldots, N_{\text{pop}}\}$, a single GA iteration starts by evaluating the vector $\vec{F}^k = \{f_i^k \,|\, i = 1, 2, \ldots, N_{\text{pop}}\}$ of objective function values f_i^k associated with chromosomes $\vec{p}_i^{\,k}$. The GA then applies the genetic operators of *selection, crossover*, and *mutation* to $\mathbf{P}^k$ to produce $\mathbf{P}^{k+1}$ (see Figure 19.1).

To commence this iteration, the GA generates either a random initial population $\mathbf{P}^0$, or an initial population base on some a priori knowledge to seed $\mathbf{P}^0$ with good chromosomes (see Figure 19.1). After the creation of $\mathbf{P}^0$, the main GA loop is entered which is iterated on each successive population $\mathbf{P}^k$.

[1] Allelic: one member of a pair or series of genes that occupy a specific position on a specific chromosome.

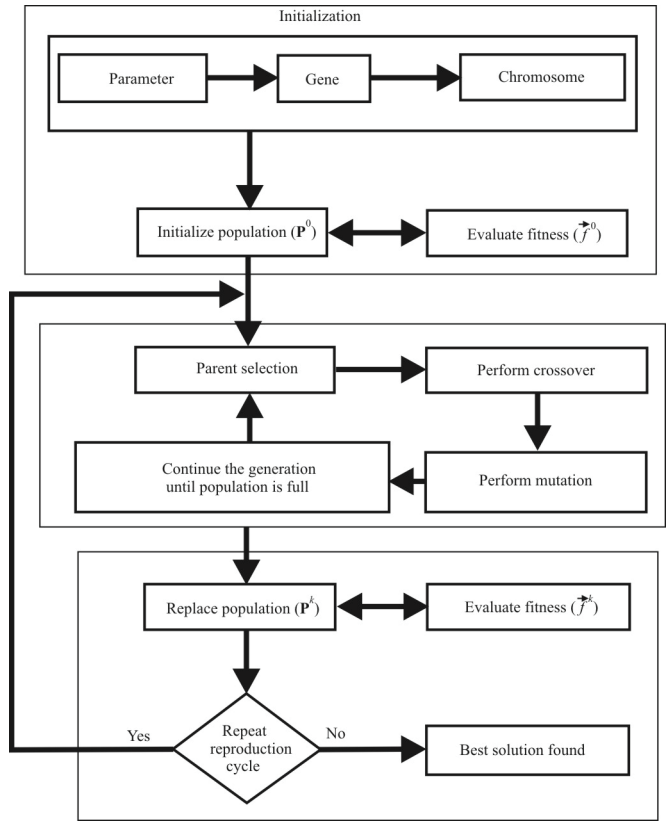


Figure 19.1 Simple GA flowchart.

19.2.2.1 Selection

Manipulation of $\mathbf{P}^k$ begins with the selection of good chromosomes based on their objective function values f^k. The selection operator produces a new population $\mathbf{P}_S^k = S(\mathbf{P}^k)$, also of N_{pop} designs, which, on average, will be better than those in $\mathbf{P}^k$. The most common methods of producing better designs are *roulette-wheel*, *ranking*, and *stochastic binary tournament selection*. This discussion describes maximization of objective functions, but minimization is just as viable—it depends on how the objective function is defined.

Roulette-wheel selection randomly places each design in the next population with a probability proportional to its fitness value given by

$$S(\mathbf{P}^k) = \bigcup_{i=1}^{N_{\text{pop}}} s(\mathbf{P}^k) \tag{19.5}$$

where s with probability

$$\Pr\{s(\mathbf{P}_k) = \mathbf{P}_j^k\} = \frac{f_j^k}{\sum_{n=1}^{N_{\text{pop}}} f_n^k} \tag{19.6}$$

A popular version of ranking selection places N copies of the best N_{pop}/n designs into $\mathbf{P}_{\text{S}}^k$. The parameter N controls the greediness of the ranking algorithm; a large N forces a fast convergence at the likely expense of a thorough global search.

Faster than both ranking and roulette-wheel selection is stochastic binary tournament selection which iteratively chooses pairs of chromosomes from $\mathbf{P}^k$ and places the better one in $\mathbf{P}_{\text{S}}^k$ until it is satiated.

The selection operator is responsible for the algorithm's convergence, and is the only operator that involves the objective function values f^k.

19.2.2.2 Crossover

Crossover follows selection, which creates a new population $\mathbf{P}_{\text{C}}^k = C(\mathbf{P}_{\text{S}}^k)$, again of size N_{pop}. Crossover is denoted as the union over a composition of operators:

$$C(\mathbf{P}_{\mathrm{S}}^{k}) = \bigcup_{i=1}^{N_{\mathrm{pop}}/2} c[ch(\mathbf{P}_{\mathrm{S}}^{k}), ch(P_{\mathrm{S}}^{k})] \tag{19.7}$$

The operator, *ch*, chooses a random chromosome from **P** and the operator, *c*, maps a pair of chromosomes $\vec{p}_1 = \{g_{1j} \,|\, j = 1,2,\ldots,N_{gl}\}$ and $\vec{p}_2 = \{g_{2j} \,|\, j = 1,2,\ldots,N_{gl}\}$ to another pair according to the rule

$$c(\vec{p}_1, \vec{p}_2) = \begin{cases} \hat{\vec{p}}_1, \hat{\vec{p}}_2 \text{ with probability } p_{\mathrm{cross}} \\ \vec{p}_1, \vec{p}_2 \text{ with probability } 1 - p_{\mathrm{cross}} \end{cases} \tag{19.8}$$

where the hybrids are given by

$$\hat{\vec{p}}_1 = \{g_{11}, g_{12}, \cdots, g_{1k}, g_{2(k+1)}, \cdots, g_{2N_{gl}}\} \tag{19.9}$$

$$\hat{\vec{p}}_2 = \{g_{21}, g_{22}, \cdots, g_{2k}, g_{1(k+1)}, \cdots, g_{1N_{gl}}\} \tag{19.10}$$

The crossover point k in (19.9) and (19.10) is chosen randomly from integers between 1 and $N_{gl}-1$. This specific single locus implementation is known as a one-point crossover.

19.2.2.3 Mutation

Finally, the mutation operator creates a new population $\mathbf{P}_{\mathrm{M}}^{k} = M(\mathbf{P}_{\mathrm{C}}^{k})$ of size N_{pop} by randomly perturbing genes. If we denote the members of the population after crossover as $\tilde{\vec{p}}_i^{\,k}$, mutation for allelic algorithms can be defined as

$$M(\mathbf{P}_{\mathrm{C}}^{k}) = \bigcup_{i=1}^{N_{\mathrm{pop}}} m(\tilde{\vec{p}}_i^{\,k}) \tag{19.11}$$

where, given a chromosome $\vec{p} = \{g_j \,|\, j = 1,2,\ldots,N_{gl}\}$, $m(\vec{p}) = \{\mu(g_j) \,|\, j = 1,2,\ldots,N_{gl}\}$, and

$$\mu(g) = \begin{cases} \hat{g} \text{ with probability } p_{\mathrm{mut}} \\ g \text{ with probability } 1 - p_{\mathrm{mut}} \end{cases} \tag{19.12}$$

where $\hat{g}$ is any other legal allele. In real-coded algorithms, mutation is very similar to that described for allelic GAs, except that $\hat{g}$ is, in general, a small random perturbation of its initial value g [15].

The population resulting from the mutation operator serves as the starting point for the next generation or $\mathbf{P}^{k+1} = \mathbf{P}_{\mathrm{M}}^{k}$; hence, a single GA generation can be functionally represented as $\mathbf{P}^{k+1} = M\{C[S(\mathbf{P}^{k})]\}$. This process is iterated until no improvement is observed, a set goal is met, or a set number of generations has passed.

Although we have described a particular sequence of functional operations on the populations, there are a plethora of other steps that can be included to improve performance for a particular problem. In general, though, the steps described or steps that are similar to these, are part of most, if not all, genetic algorithms.

19.2.3 Summary

GAs have been applied successfully to the synthesis of novel antennas, inexpensive arrays, filters, absorbers, and FSSs, and have been used to tackle inverse scattering problems. We describe one such application in the next section. They require no initial guesses to start the optimization process and often give rise to counterintuitive designs. As parallel computing resources become more widespread, more advantage can be taken of the innate parallelism of the GA structure as objective function values of whole populations of designs can be calculated in parallel [17].

19.2.4 Genetic Algorithm Antenna Design

This example is similar to one in [7]. The goal is to design a wire monopole antenna mounted on a complex structure (in our case, a Bradley fighting vehicle; see Figure 19.2 [7]) and to match it across a broad bandwidth. The antenna is to include a number of parallel RLC loads (shown in the inset in Figure 19.2) whose locations and element values are to be determined by the GA to produce maximum gain over the frequency band of interest. In addition, the GA simultaneously designs a matching network to achieve an acceptable VSWR over the band (shown in the upper left in Figure 19.2).

We assume a general matching network template. By allowing the most extreme values of the groups of genes describing the capacitances and inductances in the network to take on values of zero and infinity, elements may evolve into shorts or opens. The number of elements in the network may be kept to a minimum by penalizing designs with an excessive number of elements. The matching network part of the gene would now require

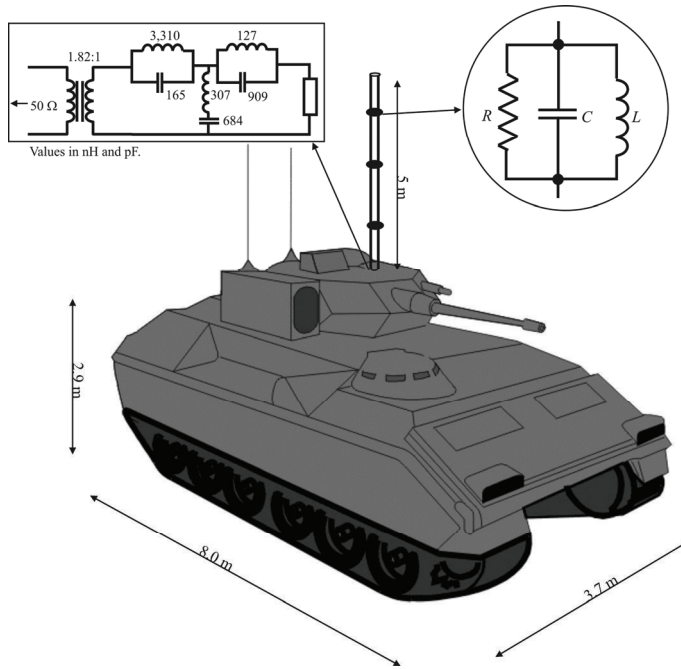


Figure 19.2 An illustration of the broadband antenna design problem.

$$N_{mn} = 4N_s \times (N_c + N_d) \qquad \text{bits} \tag{19.13}$$

where N_s is the number of matching-network sections and N_d is the number of bits added to allow for elements with zero and infinity values.

With this coding, the problem becomes a straightforward application of a simple GA. The objective function to be maximized involves a sum of the system gain and penalties described more fully in [18]. A monopole antenna mounted on top of a tank (APC) was designed to radiate between 9 and 60 MHz with maximum gain at the horizon on the side of the Bradley (Figure 19.2). This problem covers a 7:1 band, and the antenna is only 3/20 of a wavelength long at 9 MHz. The specific GA implemented to accomplish this used stochastic binary tournament selection with $N_{RLC} = 3$, $N_c = 7$, $N_s = 2$, and $N_d = 1$, resulting in a 148 bit chromosome. The GA parameters used were $N_{pop} = 500$, $p_{cross} = 0.9$. Capacitance, inductance, and resistance values were allowed to vary between 1 and 110 pF, 3 and 5,000 nH, and 1 and 1,500Ω, respectively. The optimized antenna gain and VSWR after 100 generations are shown in Figure 19.3 [7]. The

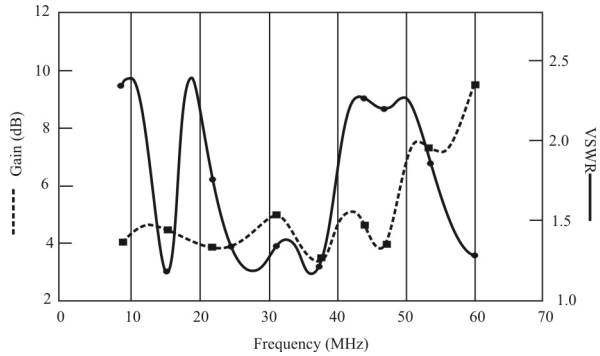


Figure 19.3 Genetic antenna gain and VSWR.

matching network component values are shown in Figure 19.2, and the load locations and component values for the loads selected by the GA are shown in Table 19.1.

We see that the gain varied between 1 and 10 dB while the VSWR < 2.5. This is acceptable gain for this application since the example was for a VHF communication transceiver, which typically is made with enough output power so that a high gain antenna is not required. The VSWR performance was excellent.

19.3 Design of Crooked Wire Antenna with Genetic Algorithm

Linden reported on an antenna designed with a GA that produced an unconventional antenna for operation at 1,600 MHz [19]. The antenna is illustrated schematically in Figure 19.4 and the fabricated antenna is shown in Figure 19.5. The antenna was to be used for satellite communications, with omnidirectional coverage and right-hand circular polarization. The goal of the cost function was to

Table 19.1 Load Locations and Component Values Selected by the GA

Load Number	Distance from Source (cm)	Resistance (Ω)	Capacitance (pF)	Inductance (nF)
1	369	66.9	103	1,830
2	331	168	49.1	695
3	361	356	107	150

Source: [7].

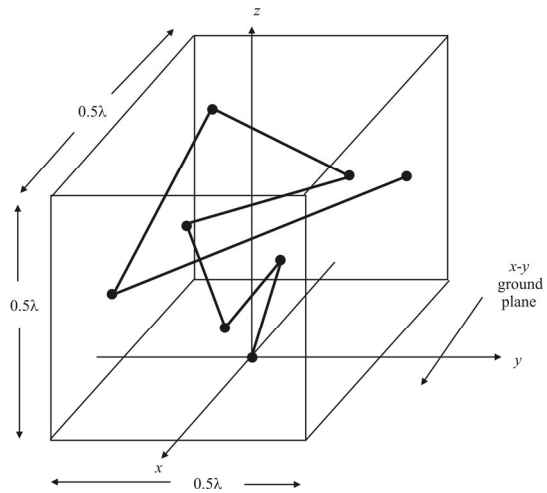


Figure 19.4 Crooked wire genetic antenna.

Figure 19.5 Crooked wire antenna photograph. (Source: [20] © Microwave Journal 1999. Reprinted with permission.)

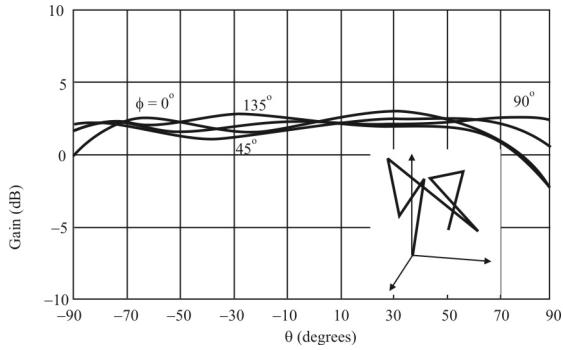


Figure 19.6 Crooked wire radiation pattern. f = 1,600 MHz.

obtain right-hand circular polarization 10° above the horizon at a frequency of 1,600 MHz. A good measure of that desired performance is the sum of the squares of the deviation of all calculated gains from the mean. That is:

$$Fitness = \sum_{\text{all } \theta,\phi} [Gain(\theta,\phi) - Gain_{\text{avg}})^2 \tag{19.14}$$

The computed radiation patterns of the antenna assuming an infinite ground plane are shown [in Figure 19.6] for azimuth angles of 0º, 45º, 90º, and 135º at a frequency of 1,600 MHz. This pattern varies by less than 4 dB for angles over 10° above the horizon—excellent performance, especially since the antenna is so simple.

Although this antenna was only designed to operate at a single frequency, its performance was also investigated for the range of 1,300 to 1,900 MHz, and it was found to have bandwidth of over 30%, which is excellent for a circularly polarized antenna having near hemispherical coverage.

We can see from Figure 19.6 that the resulting optimized antenna has a very strange shape compared with conventional designs. It is also quite surprising that the useful bandwidth is as large as it is.

19.4 Design of Broadband Loaded Monopoles

As stated many times up to this point, antennas for EW systems must have as broad a bandwidth as possible. The simple monopole antenna has several attractive features, such as an omnidirectional radiation pattern and ease of construction. Unfortunately, the conventional monopole has a narrow impedance

bandwidth and losses in the power gain at higher frequencies. However, by placing suitable loading, the monopole can be an effective ultra-broadband omni-directional antenna. By loading the monopole with appropriate resonant circuits, the drawbacks of the simple monopole can be minimized so that the antenna can effectively operate over a broadband of frequencies. The question arises as to the locations, resonant frequencies, and component values of the resonant circuits for a prescribed frequency band [20].

Several methods of designing loaded monopole antennas are available. For example, by using resistive-capacitive loadings, it is possible to obtain a band of operation with a frequency ratio of 3:1. This method approximates the antenna impedance characteristics with an $M + 1$ port network and designs a matching network so that the VSWR is at an acceptable level [21]. While efficient, this method is not adequate for very wideband (15:1) designs. We discuss two techniques in this section to produce a wideband antenna based on the monopole.

The first of these, with the appellation of conventional design, utilizes several parallel L-C circuits that are initially placed along the monopole such that the effective electrical length of the antenna remains approximately one-quarter wavelength over the frequency range of interest. This method is based upon a two-step procedure as follows: (1) design a loaded antenna that meets the gain requirements, and (2) devise its matching network to obtain the desired VSWR.

The second method uses a genetic algorithm to design both the loaded monopole and the matching network simultaneously in an optimal manner.

19.4.1 Conventional Design of Loaded Monopole

The antenna is designed to operate over the 30–450 MHz band and is 1.75 m long. The specifications for the antenna-matching network require that the system gain be greater than –5 dB and have a VSWR less than 3.5 across the frequency band. In the first design, L-C resonator circuits are placed at various locations along the antenna, and their resonant frequencies are chosen such that their distances to the base correspond to a quarter wavelength at their respective frequencies. The strategy is to force the current to go to zero at these locations and hope that it will be small above these locations as well. However, when this configuration is tested using NEC [22], it is found that the current does not go to zero above the resonant circuits, and the gain does not meet the design requirements. Therefore, the initial locations and resonant frequencies must be adjusted to obtain the desired gain. After a low-level optimization, four resonant circuits are chosen, and their resonant frequencies, quarter wavelengths, and actual distances are shown in Table 19.2.

Table 19.2 Load Characteristics for the Conventional Design

Resonant Frequency (MHz)	Quarter Wavelength (m)	Actual Distance from Base (m)
230	0.326	0.475
245	0.306	0.425
260	0.288	0.375
285	0.263	0.325

Source: [31].

We see that the locations chosen are not exactly one-quarter wavelength. Also, the resonant frequencies are relatively close together instead of being distributed over the band of frequencies. These results show that the load locations determined by the intuitive approach must be viewed merely as a starting point, as they do not resemble the final result. Once acceptable gain is achieved, a matching network is designed. A gradient optimization method with random initial guess is used to minimize the maximum VSWR of the antenna matching network system. Since the VSWR specification is not met with the resonant circuit loading alone, a resistor is added to the top portion of the antenna to rectify the situation. It is necessary to ascertain, however, that the gain requirements are still satisfied even after the introduction of this resistor.

19.4.2 GA Design of Loaded Monopoles

As discussed above, GAs are probabilistic optimization processes that start from a population of randomly-generated potential solutions and gradually evolve toward better solutions through a repetitive application of genetic operators as illustrated in Figure 19.1. In contrast to the conventional design approach outlined above, the GA-based synthesis technique simultaneously optimizes parallel RLC circuit parameters, their locations along the antenna, as well as the matching network. To this end, the GA considers a large population of antenna and matching network configurations, each of which is evaluated for the gain and VSWR performance over the frequency band of interest using an MoM-based analysis technique.

19.4.3 Comparison of the Two Design Methods

The calculated impedance plots for each of the designs are shown in Figures 19.7 and 19.8 [20]. The gains of the antenna-matching networks are shown in Figure 19.9 [20]. The drop in the gain of the antenna designed by the conventional method at the lower end of the band is attributable to the loading resistor. We can see that the genetic algorithm produces a more desirable gain curve because, in contrast to the conventional loading approach, it designs both the antenna and the matching network in one step. The VSWR characteristics for both antenna-

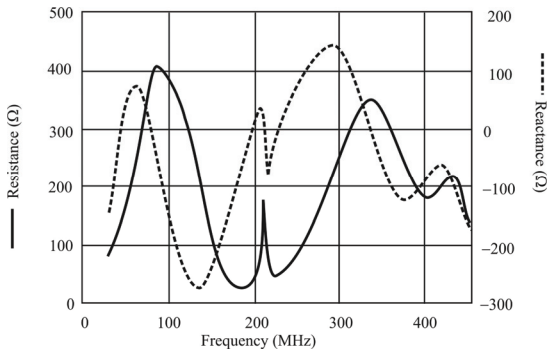


Figure 19.7 Impedance of antenna designed by conventional method.

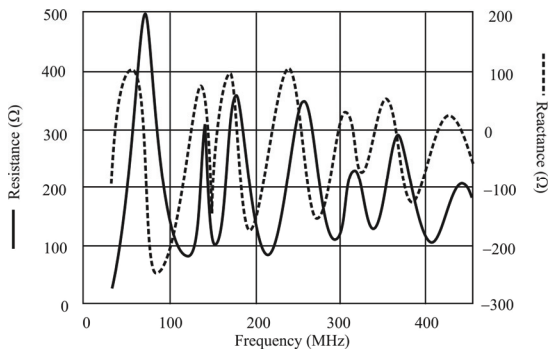


Figure 19.8 Impedance of antenna designed by genetic algorithm.

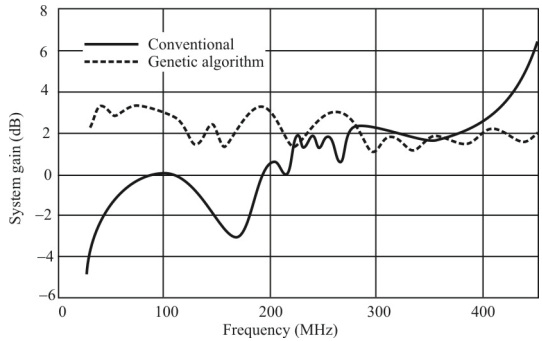


Figure 19.9 Gains of antenna-matching network systems.

matching network systems are shown in Figure 19.10 [20]**.** These numerical results are supported, with only minor variations, by measured data.

19.5 Simulated Annealing

As its name implies, *simulated annealing* (SA) exploits an analogy between the way in which a metal cools and freezes into a minimum energy crystalline structure (the annealing process) and the search for a minimum in a more general system [23]. This process involves heating a solid material up to a high temperature and then allowing it to cool at a very slow rate. The result is that the particles in the solid arrange themselves in the lowest energy state configuration, usually into an ordered lattice of some sort [24]. The optimization algorithm effectively does the same thing.

The algorithm is based upon that of Metropolis et al. [25], which was originally proposed as a means of finding the equilibrium configuration of a collection of atoms at a given temperature. The connection between this algorithm and mathematical minimization was first noted by Pincus [26], but it was Kirkpatrick et al. [27] who proposed that it form the basis of an optimization technique for combinatorial (and other) problems.

SAs major advantage over other methods is an ability to avoid becoming trapped at local minima. The algorithm employs a random search that not only accepts changes that decrease objective function f, but also some changes that increase it. The latter are accepted with a probability

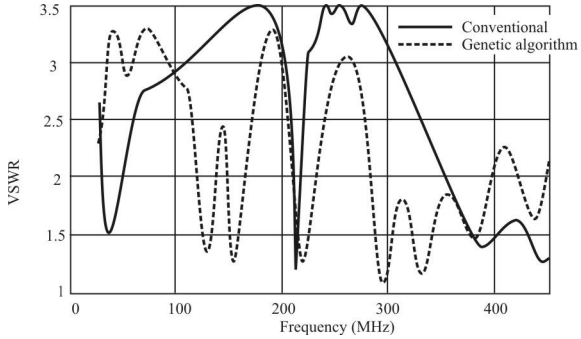


Figure 19.10 VSWR characteristics of antenna-matching network systems.

$$p = \exp\left\{-\frac{\delta f}{T}\right\} \tag{19.15}$$

where δf is the increase in f and T is a control parameter, which by analogy with the original application is known as the system "temperature," irrespective of the objective function involved.

The implementation of the SA algorithm is remarkably easy. Figure 19.11 [23] shows its basic structure. The following elements must be provided:

- A representation of possible solutions;
- A generator of random changes in solutions;
- A means of evaluating the problem functions;
- An *annealing schedule*—an initial temperature and rules for lowering it as the search progresses.

19.5.1 Solution Representation and Generation

When attempting to solve an optimization problem using the SA algorithm, the most obvious representation of the control variables is usually appropriate. However, the way in which new solutions are generated may need some thought. The solution generator should introduce small random changes, and allow all possible solutions to be reached.

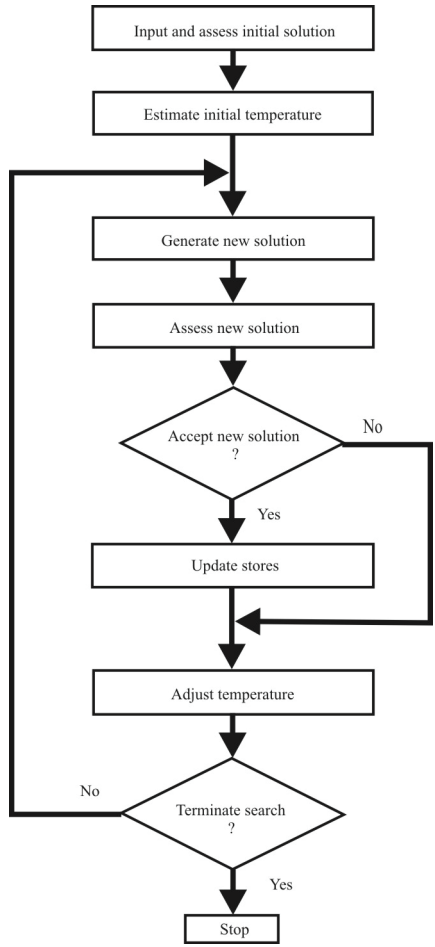


Figure 19.11 Simulated annealing algorithm.

For problems with continuous control values, Vanderbilt and Louie [28] recommend that new trial solutions be generated according to

$$\vec{x}_{i+1} = \vec{x}_i + \mathbf{Q}\vec{u} \tag{19.16}$$

where $\vec{u}$ is a vector of random numbers in the range $(-\sqrt{3}, \sqrt{3})$, so that each has zero mean and unit variance, and **Q** is a matrix that controls the step size distribution. In order to generate random steps with a covariance matrix **S**, **Q** is found by solving

$$\mathbf{S} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}} \tag{19.17}$$

by the Cholesky decomposition,[2] for example, **S** should be updated as the search progresses to include information about the local topography:

$$\mathbf{S}_{i+1} = (1-\alpha)\mathbf{S}_i + \alpha w\mathbf{X} \tag{19.18}$$

where matrix **X** measures the covariance of the path actually followed and the damping constant α controls the rate at which information from **X** is folded into **S** with weighting *w*. One drawback of this scheme is that the solution of (19.17), which must be done every time **S** is updated, can represent a substantial computational overhead for problems with high dimensionality. In addition, because the probability of an objective function increases being accepted, given by (19.15), does not reflect the size of the step taken, **S** must be estimated afresh every time the system temperature is changed.

An alternative strategy suggested by Parks [29] is to generate solutions according to

$$\vec{x}_{i+1} = \vec{x}_i + \mathbf{D}\vec{u} \tag{19.19}$$

where $\vec{u}$ is now a vector of random numbers in the range (–1, 1) and **D** is a diagonal matrix which defines the maximum change allowed in each variable. After a successful trial, that is, after an accepted change in the solution, **D** is updated:

$$\mathbf{D}_{i+1} = (1-\alpha)\mathbf{D}_i + \alpha w\mathbf{R} \tag{19.20}$$

[2] If **A** is a nonsingular symmetric matrix, then it can be expressed as the product $\mathbf{L}\mathbf{L}^{\mathrm{H}}$, where **L** is a lower triangular matrix and $\mathbf{L}^{\mathrm{H}}$ is the conjugate transpose of **L**. This is the *Cholesky decomposition* of **A**.

where α and w perform similar roles and **R** is a diagonal matrix, the elements of which consist of the magnitudes of the successful changes made to each control variable. This tunes the maximum step size associated with each control variable towards a value giving acceptable changes.

When using this strategy, it is recommended that the probability of accepting an increase in f be changed from that given by (19.15) to

$$p = \exp\left(-\frac{\delta f / \bar{d}}{T}\right) \tag{19.21}$$

where $\bar{d}$ is the average step size, so that $\delta f / \bar{d}$ is a measure of the effectiveness of the change made. As the size of the step taken is considered in calculating p, **D** does not need to be adjusted when T is changed.

For problems with integer control variables, the simple strategy whereby new trial solutions are generated according to

$$\vec{x}_{i+1} = \vec{x}_i + \vec{u} \tag{19.22}$$

where $\vec{u}$ is a vector of random integers in the range (–1, 1) often suffices.

19.5.2 Simulated Annealing Example

Bevelacqua and Balanis discussed using simulated annealing to optimize an antenna array geometry in the presence of interfering signals [30]. We discuss their approach to this problem in this section as an example of the application of simulated annealing to antenna design. The approach is to formulate optimum beamformers to place nulls in the direction of interfering signals.

As we discussed in Chapter 17, spatial filtering is done with an adaptive antenna array, which is an array of spatially separated antennas in an arbitrary pattern. The received signals on each antenna are weighted with complex weights to account for amplitude and phase shifts and then summed together to produce the output of the spatial filter. The weights are dynamically adjusted to develop an appropriate beam pattern using an adaptive signal processing algorithm. This adaptation can be either blind or nonblind, the latter if the structure of the SOI is known or can be accurately approximated, the former if not.

We assume that there is an interference environment present that can be statistically characterized, parameterized by the expected DOAs and relative power of the interference.

The performance of an antenna array with interference depends primarily on three strictures: the interference environment, the weight vector used to minimize this interference, and the geometry of the array. The literature is replete with techniques for determining the optimum weights—we covered several in Chapter 17. The interference environment that an EW system is expected to operate in normally cannot be a priori selected, for tactical scenarios anyway. The remaining parameter, the array geometry, is the question to be addressed in this section. Optimization of array geometry has been attempted in [31], but this was done with respect to parameter estimation and results presented only for three-element arrays.

19.5.2.1 The Objective Function

Assume that a plane wave of the SOI is incident upon an array of M elements, with the direction given by the wave vector $\vec{\beta}$. This wave produces a steering vector, $\vec{v}$, whose ith element given by

$$v_i = e^{-j\vec{\beta}^{\mathrm{T}}\vec{r}_i} \tag{19.23}$$

where $\vec{r}_i$ represents the position of the ith antenna element (see Figure 19.12), and $\vec{\beta}^{\mathrm{T}}$ is the transpose of the wave vector. The sources consisting of the SOI and interference are incident upon the array, so the input to the array is given by

$$\vec{X} = \sum_{n=0}^{Q} g(\vec{\beta}_n) s_n v_n \tag{19.24}$$

(a single SOI and Q interferers) where $g(\vec{\beta}_n)$ is the individual antenna's element factor for direction $\vec{\beta}_n$, and s_n is the signal of the nth source. The autocorrelation matrix of the input signal is

$$\mathbf{R}_{XX} = \mathcal{E}_t\{\vec{X}\vec{X}^{\mathrm{H}}\} \tag{19.25}$$

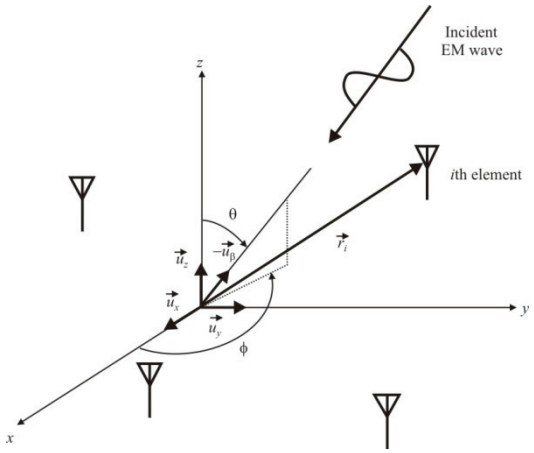


Figure 19.12 Arbitrary 3-D array.

where the expectation is over time as indicated by t.

The expected autocorrelation matrix, $\mathbf{S}_X$, is

$$\mathbf{S}_X = \mathcal{E}_I\{\mathbf{R}_{XX}\} \tag{19.26}$$

Here the expectation is over the possible interference situations (which defines the interference environment).

One of the weights is restricted to be unity so that the power minimization algorithm does not set all the weights to zero. We assume that, without loss of generality, s_0 is the SOI. In addition, for practical reasons we require that the separation between elements to be at least $\lambda/4$, where λ is the wavelength of interest. (The separation must be less than λ, however, in order to avoid grating lobes—see Chapter 10.) Let r_{ij} be the separation between elements i and j. The problem of finding an optimal array for interference suppression can be expressed as an optimization problem as

$$\begin{aligned} &\text{minimize} && \vec{w}^{\mathrm{H}}\mathbf{S}_X\vec{w} && \\ &\text{subject to} && \vec{w}^{\mathrm{H}}\vec{e}_0 = 1 && \\ & && r_{ij} \geq \lambda/4 && \text{for } i \neq j \end{aligned} \tag{19.27}$$

The minimization variables are the complex weights, w_i, and the values of r_{ij}. Note that $\vec{e}_0^{\,\mathrm{T}} = [1, 0, \ldots, 0]$.

Since the locations of the antenna elements are assumed to be fixed and known, the optimal weight vector for this problem can be found by using Lagrange multipliers (see [32], Appendix 17A). The resulting optimum weights are given by (see Chapter 17)

$$\vec{w}_{\text{opt}} = \frac{\mathbf{S}_X^{-1}\vec{e}_0}{\vec{e}_0^{\,H}\mathbf{S}_X^{-1}\vec{e}_0} \tag{19.28}$$

which, of course, assumes that $\mathbf{S}_X$ is nonsingular. The minimum value of the objective function of (19.27) is

$$\min_w\{w^{\text{H}}\mathbf{S}_X\vec{w}\} = \frac{1}{\vec{e}_0^{\,\text{H}}\mathbf{S}_X^{-1}\vec{e}_0} \tag{19.29}$$

Equation (19.29) is a function of the antenna locations and the interference environment. Since we have no control over the interference environment, the performance of the array is determined by the elements' positions. Therefore, the optimization problem given by (19.27) can be restated as

$$\begin{array}{ll} \text{maximize} & \vec{e}_0^{\,H}\mathbf{S}_X^{-1}\vec{e}_0 = [\mathbf{S}_X^{-1}]_{00} \\ \text{subject to} & r_{ij} \geq \lambda/4, \qquad i \neq j \end{array} \tag{19.30}$$

where $[\mathbf{X}]_{mn}$ denotes the element of matrix $\mathbf{X}$ in the *m, n* position. Note that $\mathbf{S}_X$ is the expected autocorrelation matrix and is therefore positive semidefinite. Palpably, (19.30) will not have a unique solution since $\mathbf{S}_X$ is invariant to translation (shifting the elements uniformly).

19.5.2.2 Simulated Annealing

As noted above, simulated annealing optimizes over a space of points that we denote by χ. We require an initial feasible point, which can be any element $x \in \chi$ that satisfies the constraints, a perturbation mechanism that generates the next point to be tested, and a cost function, $C : \chi \to \Re$. The set of all arrays that satisfy the constraints in (19.30) is the space to be optimized over; each specific geometry is a point in the space. The cost function is the objective function in (19.30).

From the current point x_i, the next point x_{i+1} is chosen at random using the perturbation mechanism. If

$$\Delta C_i = C(x_{i+1}) - C(x_i) > 0 \tag{19.31}$$

then the current solution point is updated from x_i to x_{i+1}. As discussed above, if $\Delta C_i < 0,$ then the probability of choosing the current solution point to be x_{i+1} over x_i is given by

$$\Pr\{\text{transition from } x_i \text{ to } x_{i+1}\} = \exp\left(\frac{\Delta C_i}{T}\right) \tag{19.32}$$

where T represents the current "temperature" of the system.

The simulated annealing algorithm starts the optimization procedure at a high "temperature" (sufficiently high such that most transitions occur) and lowers the temperature slowly enough so that a global maximum is found.

19.5.2.3 Planar Array with Uniform Interference

For specificity, we consider a planar array of elements with the same physical orientation. We assume that the interfering signals are mutually independent and arrive from a uniform distribution in the azimuth direction but from a fixed elevation angle. This assumption also permits the dropping of the element factor $[g(\vec{\beta}_n)]$ in (19.24), since in an EW application we would not normally a priori know the AOAs of the interfering signals, the element patterns typically selected for this problem will not vary significantly with the azimuth angle (omnidirectional coverage in the azimuth plane—dipoles, monopoles, and loops). Assuming the elements are located at positions in the *x*-*y* plane given by

$$\vec{r}_i = (x_i, y_i, 0), \qquad i = 1, 2, \ldots, M \tag{19.33}$$

and assuming Q interferers

$$\mathbf{R}_{XX} = \mathcal{E}\{\vec{X}\vec{X}^{\mathrm{H}}\} = \mathcal{E}\left\{\left(\sum_{n=0}^{Q} s_n v_n\right)\left(\sum_{n=0}^{Q} s_n v_n\right)^{\mathrm{H}}\right\} \tag{19.34}$$

it follows that

$$\mathcal{E}\{(s_n v_n)(s_n v_n)^{\mathrm{H}}\}_{ab} = \sigma_n^2 e^{-j\beta \sin\theta[\cos\phi_n (x_a - x_b) + \sin\phi_n (y_a - y_b)]} \tag{19.35}$$

where σ_n^2 is the signal power of source n, θ_n is the elevation angle of the interference (fixed), and ϕ_n is the azimuth angle for signal n. Since we assume that the interfering signals are independent of each other and of the SOI, $\mathcal{E}\{v_n v_p^{\mathrm{H}}\} = 0$ for $n \neq p$, we have

$$[\mathbf{R}_{XX}]_{ab} = \sum_{i=0}^{Q} \sigma_i^2 e^{-j\beta \sin\theta_n [\cos\phi_i (x_a - x_b) + \sin\phi_i (y_a - y_b)]} \tag{19.36}$$

Finding the expected value of (19.36) with respect to ϕ (recall that the interference is uniform in ϕ), we get

$$[\mathbf{S}_x]_{ab} = \left(\sum_{i=0}^{Q} \sigma_i^2 \right) J_0(\beta r_{ab} \sin\theta_i) \tag{19.37}$$

where J_0 is the Bessel function of the first kind of order 0. We seek to find a global optimum for (19.30), where the variables are r_{ij}. From (19.37) we see that the results depend only on the total power incident on the array, and not on the number of interferers. We will take the elevation angle of the signals as on the horizon so that $\theta_i = \pi / 2$.

The initial array configuration is a UCA. The initial array chosen should have no effect on the optimization result. The perturbation is achieved by choosing a random vector in $\Re^{2Q}$ that has Euclidean norm given by a small variance. The variance is chosen such that the average perturbation for each element is on the order of 0.01λ. The components of this random vector are added to the x and y coordinates of the elements. If the perturbation moves the elements too close together, then the perturbation is discarded. Another constraint is imposed such that all elements stay within 0.75λ of the origin (center of initial circular array) to keep the search space finite. Note that this distance can put two antennas up to 1.5λ apart and grating lobes will result.

The initial temperature T_0 is chosen such that virtually all (>99%) of perturbations are accepted. The temperature is held constant for a fixed number (P) of perturbations. The temperature is then multiplied by a factor $u < 1$. The solution array is then again perturbed P times. This process is performed until T is small enough that no perturbations that decrease the objective function are accepted; once this happens, the solution has converged upon a local maximum. If P is sufficiently large and the temperature decreased sufficiently slowly, this method will converge to the global optimum [33]. In the solution for $M = 6$, the

parameters used were $u = 0.99$, $P = 50{,}000$, and $T_0 = 12$. See [30] for the method of determining these numbers.

The optimum array configurations for four, five, and six elements are shown in Figure 19.13. The results suggest that the interference suppression capabilities are best for arrays spaced as closely as possible. They all have a center element and are surrounded by a circular array of radius 0.25λ (the minimum distance allowed). This suggests a trade-off between interference suppression and largely spaced arrays used for diversity or to minimize mutual coupling. With the constraints used, minimizing the size of the array forces element separation distances to be less than a wavelength, thereby avoiding grating lobes.

19.5.2.4 Summary

We showed how simulated annealing can be applied to find optimum adaptive antenna array configurations. Equation (19.30) represents the optimization problem for determining an optimum array for a particular interference environment. A specific interference environment was analyzed and the optimization problem was solved via the simulated annealing algorithm. The particular interference environment analyzed consisted of Q interfering signals uniformly distributed over $(0, 2\pi]$, and a single SOI all at an elevation angle of $\theta = \pi/2$. It is found that array geometry has a significant effect on the interference suppression capabilities of an adaptive antenna array. The results also suggest that an array's interference suppression capabilities increase with small interelement spacing.

19.6 Concluding Remarks

An introduction to genetic algorithms and their application to EW antenna design was provided in this chapter. Applying GAs to this problem can lead to some unusual antenna architectures as we saw.

GAs are not confined to prescribed structures, but operate on different principles. Designs are optimized based on performance parameters, not on architectures that are well understood. Because of this, effective designs can ensue.

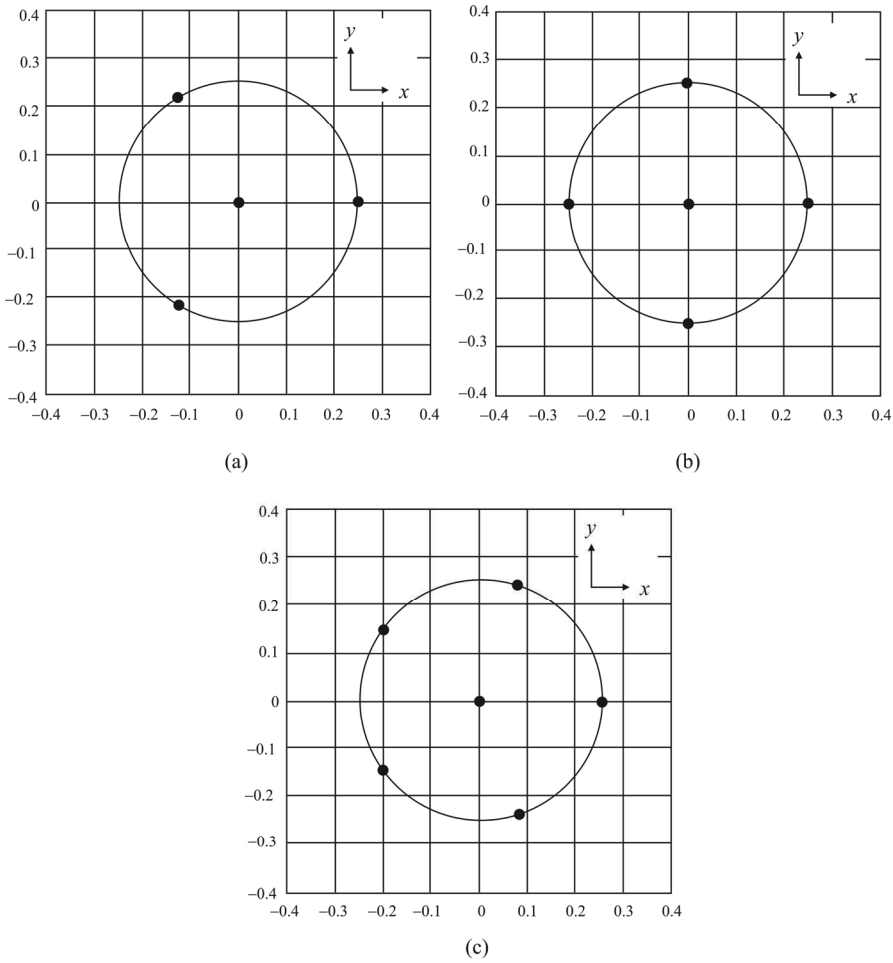


Figure 19.13(a–c) Optimum arrays (measured in units of λ). Circles are of radius 0:25λ: (a) $M = 4$; (b) $M = 5$; and (c) $M = 6$.

References

[1] Darwin, C., *The Origin of Species*, New York: Prometheus Books, 1991. Originally published in 1859 by John Murray in the U.K.

[2] Altshuler, E. E., and D. S. Linden, "Wire-Antenna Designs Using Genetic Algorithms," *IEEE Antennas and Propagation Magazine*, Vol. 39, No. 2, April 1997, pp. 33–43.

[3] Jones, E. A., and W. T. Jones, "Genetic Design of Linear Antenna Arrays," *IEEE Antenna and Propagation Magazine*, Vol. 42, No. 3, June 2000, pp. 92–100.

[4] Haupt, R. L., "An Introduction to Genetic Algorithms for Electromagnetics," *IEEE Antennas and Propagation Magazine*, Vol. 37, No. 2, April 1995, pp. 7–15.

[5] Varlamos, P. K., P. J. Papakanellos, S. C. Panagiootou, and C. C. Capsalis, "Multi-Objective Genetic Optimization of Yagi-Uda Arrays with Additional Parasitic Elements," *IEEE Antennas and Propagation Magazine*, Vol. 47, No. 4, August 2005, pp. 92–97.

[6] Marcano, D., and F. Duran, "Synthesis of Antenna Arrays Using Genetic Algorithms," *IEEE Antennas and Propagation Magazine*, Vol. 42, No. 3, June 2000, pp. 12–19.

[7] Johnson, J. M., and Y. Rahmat-Samii, "Genetic Algorithms in Engineering Electromagnetics," *IEEE Antennas and Propagation Magazine*, Vol. 39, No. 4, August 1997, pp. 7–19.

[8] Weile, D. S., and E. Michielssen, "Genetic Algorithm Optimization Applied to Electromagnetics: A Review," *IEEE Transactions on Antennas and Propagation*, Vol. 45, No. 3, March 1997, pp. 343–353.

[9] Boag, A., A. Boag, E. Michielssen, and R. Mittra, "Design of Electrically Loaded Wire Antennas Using Genetic Algorithms," *IEEE Transactions on Antennas and Propagation*, Vol. 44, No. 5, May 1996, pp. 687–695.

[10] Fletcher, R., *Practical Methods of Optimization.* New York: Wiley , 1980.

[11] Davis, L., *Genetic Algorithms and Simulated Annealing*, London, U.K.: Pittman, 1987.

[12] DeJong, K. A., "An Analysis of the Behavior of a Class of Genetic Adaptive Systems," Ph. D. Dissertation, University of Michigan, 1975.

[13] Goldberg, D. E., *Genetic Algorithms in Search, Optimization and Machine Learning.* Reading, MA: Addison-Wesley, 1989.

[14] Holland, J. H., *Adaptation in Natural and Artificial Systems*, Ann Arbor, MI: Univ. Michigan, 1975.

[15] "Genetic Algorithms," *Scientific American.,* Vol. 267, July 1992, pp. 66–72.

[16] Forrest, S., "Genetic Algorithms: Principles of Natural Selection Applied to Computation," *Science,* Vol. 261, August 1993, pp. 872–878.

[17] Michielssen, E., A. Boag, J. M. Sajer, and R. Mittra, "Design of Frequency Selective Surfaces Using Massively Parallel Genetic Algorithms," *Proceedings URSI Radio Science Meeting,* Seattle, WA, June 1994, p. 441.

[18] Weile, D. S., "Genetic Algorithm Applications in Electromagnetics," Master's thesis, University of Illinois, Urbana-Champaign, August, 1995.

[19] Linden D. S., and E. E. Altshuler, "Automating Wire Antenna Design Using Genetic Algorithms," *Microwave Journal*, Vol. 39, No. 3, March 1996, pp. 74–86.

[20] Bahr, M., A. Boag, E. Michielssen, and R. Mittra, "Design of Ultra-Broadband Loaded Monopoles," *Proceedings IEEE Antenna and Propagation Society International Symposium*, Seattle, WA, June 1994, pp. 1290–1293.

[21] Popovich, B. D., *CAD* of *Wire Antenna and Related Radiating Structures*, Somerset, England Research Studies Press, 1991.

[22] Burke, G. J., and A. J. Pogio, *Numerical Electromagnetics Code—Method* of *Moments,* San Diego, Califomia: Naval Ocean Systems Center, 1981.

[23] http://www.phy.ornl.gov/csep/CSEP/MO/NODE28.html.

[24] Altshuler, E. E., "Small Wire Antenna Design Using a Genetic Algorithm," *Proceedings of the URSI General Assembly*, Toronto, Ontario, Canada, August 13–21, 1999.

[25] Metropolis, N., A. W. Rosenbluth, M. N., Rosenbluth, A. H. Teller, and E Teller, "Equations of State Calculations by Fast Computing Machines," *Journal of Chemistry and Physics*, Vol. 21, 1958, pp. 1087–1092.

[26] Pincus, M., "A Monte Carlo Method for the Approximate Solution of Certain Types of Constrained Optimization Problems," *Operations Research*, Vol. 18, 1970, pp. 1225–1228.

[27] Kirkpatrick, S., C. D. Gerlatt, Jr., and M. P. Vecchi, "Optimization by Simulated Annealing," *Science*, Vol. 220, 1983, pp. 671–680.

[28] Vanderbilt, D., and S. G. Louie, "A Monte Carlo Simulated Annealing Approach to Optimization over Continuous Variables," *Journal of Computational Physics*, Vol. 56, 1984, pp. 259–271.

[29] Parks, G. T., "An Intelligent Stochastic Optimization Routine for Nuclear Fuel Cycle Design," *Nuclear Technology*, Vol. 89, 1990, pp. 233–246.

[30] Bevelacqua, P. J., and C. A. Balanis, "Optimizing Antenna Array Geometry for Interference Suppression," *IEEE Transactions on Antennas and Propagation*, Vol. 55, No. 3, March 2007, pp. 637–641.

[31] Ang, C. W., C. M. See, and A. C. Kot, "Optimization of Array Geometry for Identifiable High Resolution Parameter Estimation in Sensor Array Signal Processing," in *Proceedings International Conference on Information and Communication Signal Processing*, Singapore, September 1997, pp. 1613–1617.

[32] Poisel, R. A., *Target Acquisition in Communication Electronic Warfare Systems*, Norwood, MA: Artech House, 2002.

[33] van Laarhoven, P. J. M., and E. H. L. Aarts*, Simulated Annealing: Theory and Applications,* Dordrecht, Holland: D. Reidel, 1987.

Part III
Related Antenna Topics

Chapter 20

Antenna Matching

20.1 Introduction

Matching the output of an amplifier stage to the load antenna ensures that power transfer is maximized. At the power levels involved with high power transmissions from EW systems, even 1 dB of mismatch can amount to considerable inefficiencies. It is normally worthwhile to ensure that such mismatches are minimized.

The same can be said for matching a receiving antenna to the input to the transmission line connected between the antenna and receiver. Maximum power is transferred when the impedances match, although typically the power levels are not as large. Maximizing this power transfer ensures that maximum sensitivity is achieved.

As we will show, matching can be accomplished with lumped elements or with transmission line sections; the advantages of each method will be demonstrated.

Most modern electrical circuit analysis is performed with computers, including circuit operation simulation. In this chapter, the discussion begins with a short section on the fundamental manual tool used in transmission line analysis—the Smith chart. The Smith chart not only produces useful results, but it also provides an intuition of how transmission lines, and other RF circuits, operate.

As we discussed in Chapter 4, transmission line stubs can be, and are, used to match the impedances of transmission lines to antennas, a prerequisite for maximum power transfer to the antenna. This matching is facilitated by adjusting the length of the stub as well as terminating it with either a short or open circuit. The Smith chart also comes into use for this analysis.

This chapter is organized as follows. After this introduction we present some general concepts of impedance matching. That is followed by a brief discussion on the Smith chart. This is primarily to refresh the reader's memory. We then cover antenna matching with two and three lumped elements. Next is matching with

distributed lines, as discussed in Chapter 4. Then we cover some common techniques for actually coupling the transmission line and matching network to the antenna. Lastly, we briefly mention a few of the plethora of software tools (free) that are available for designing simple matching networks.

20.2 Impedance Matching

There are several reasons why impedance matching of a transmission line to an antenna is important. A few of the more important are:

- Impedance matching maximizes the power transfer from sources to loads, or, in this case, from the transmission line into the antenna (or vice versa).
- Losses in the transmission line are minimized when they are matched.
- The SNR is maximized in input stages of receivers when the antenna is matched to the line.
- Distortion in transmission lines is minimized.

20.2.1 Passive Matching Networks

Matching networks can be either passive or active, although the former is more common. There are several methods to implement passive matching networks:

- Resistive networks are used for extremely wide bandwidths; however, they result in lossy networks and are only used when such losses can be tolerated.
- For wide bandwidth applications, transformer matching can be used (in particular transmission line transformers). Such matchings are not as noisy as resistive networks but they typically are not as broadband.
- If the requirements are more narrowband, reactance (inductive and capacitive) matching is used with high efficiency.
- Coaxial lines, stubs, and transmission line are used for matching for narrow bandwidth, very high frequencies, and high efficiency.
- Combined matching with these techniques is also possible.
- Lastly there is no (extra) matching required if we use already matched components (usually 50Ω amplifiers).

We will delve mostly into implementing matching networks with L-sections, π sections, and Tee sections using passive reactive components in this chapter. We also will cover how matches between transmission lines and antennas can be implemented with feed lines.

It is important to reiterate that the matching networks discussed here only match perfectly (within component tolerances) at a single frequency. At all other frequencies there are mismatches to some degree.

There are two major issues when constructing the feed circuit: impedance matching and balanced-unbalanced matching.

20.2.2 Impedance Matching

The VSWR, in terms of the reflection coefficient is given by (4.60)

$$\text{VSWR} = S = \frac{1+|\Gamma|}{1-|\Gamma|} \tag{20.1}$$

From (20.1), the fractional reflected power in terms of VSWR is given by

$$|\Gamma|^2 = \left(\frac{S-1}{S+1}\right)^2 \tag{20.2}$$

Finally, the fractional transmitted power is given by

$$|T|^2 = 1-|\Gamma|^2 \tag{20.3}$$

Impedance mismatch is undesirable for several reasons. Not only is the power transfer inefficient, causing larger than necessary losses but in high-power transmitting systems such as the EW systems we are considering, high VSWR leads to maximation of the standing wave, which can cause arcing. Sometimes the frequency of the transmitter can be affected by severe impedance mismatch (this is called "frequency pulling").

The effects of impedance mismatch between the drive circuit and the antenna can be envisioned by considering Table 20.1, which shows the amount of power actually transmitted as a function of the VSWR.

20.3 Smith Chart

The Smith chart appeared in 1939 [1] as a graph-based method of manipulating the complex-number mathematics involved with describing the characteristics of high frequency electrical circuits. The Smith chart, like other graphical engineering aids [2], remains a valuable tool even though calculators and computers can now easily solve the problems the Smith chart was designed to solve. It is the insights that can

Table 20.1 VSWR and Transmitted Power

VSWR	$\|\Gamma\|^2$ (%)	$\|T\|^2$ (%)
1.0	0	100
1.1	0.2	99.8
1.2	0.8	99.2
1.5	4.0	96.0
2.0	11.1	88.9
3.0	25.0	75.0
4.0	36.0	64.0
5.0	44.4	55.6
5.83	50.0	50.0
10.0	66.9	33.1

be derived from the Smith chart's graphical representations that keep the chart relevant for today's engineering purposes.

The Smith chart is a special plot of the complex S-parameter s_{11}, which is equivalent to the complex reflection coefficient Γ for single-port microwave components (see Figure 20.1) [3]. Note that in general the reflection coefficient is complex:

$$\Gamma = \Gamma_r + j\Gamma_i = |\Gamma| e^{j\theta} \tag{20.4}$$

which is often expressed as $\Gamma \angle u$. It is customary that in complex-notation formats that include the angle sign (∠), the preceding variable or constant is assumed to represent magnitude. Figure 20.2 shows the specific case of a complex Γ value 0.6 + j0.3 plotted in Cartesian as well as polar coordinates (0.67∠26.7°).

The reflection coefficient is defined as the ratio between the reflected voltage wave and the incident voltage wave:

$$\Gamma = \frac{V_{\text{Refl}}}{V_{\text{Inc}}} \tag{20.5}$$

(see Figure 20.3). The amount of reflected signal from the load is dependent on the degree of mismatch between the source (line) impedance and the load impedance. Its expression is (4.14)

$$\Gamma = \frac{V_{\text{Refl}}}{V_{\text{Inc}}} = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{20.6}$$

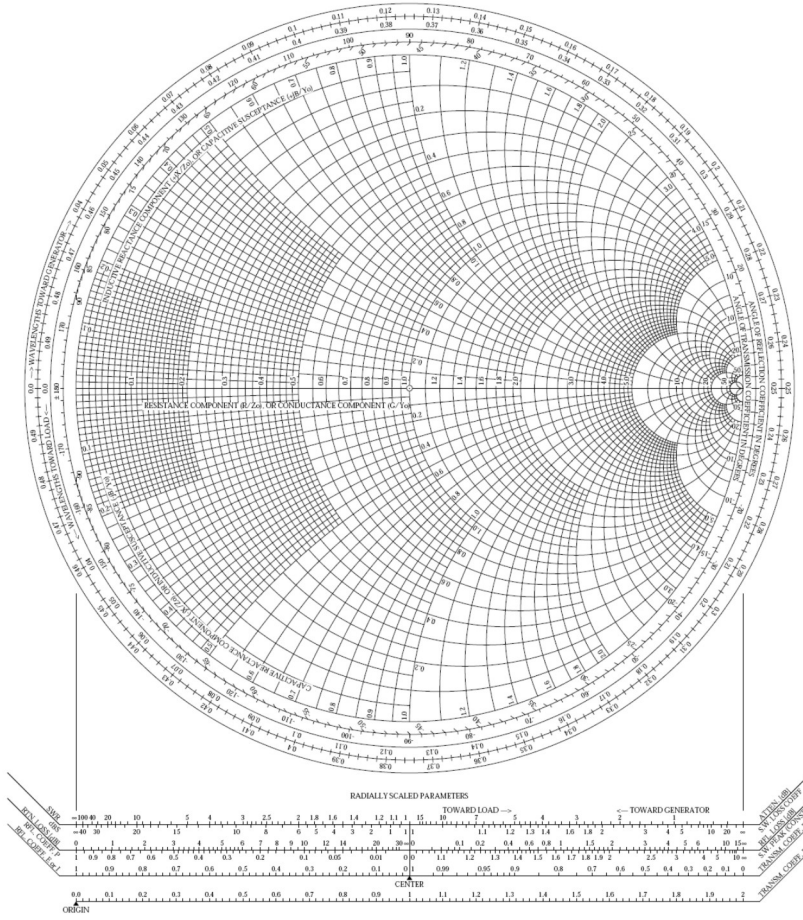


Figure 20.1 Smith chart. (Source: [3] © 2000 Atlanta: Noble Publishing. Reprinted with permission.)

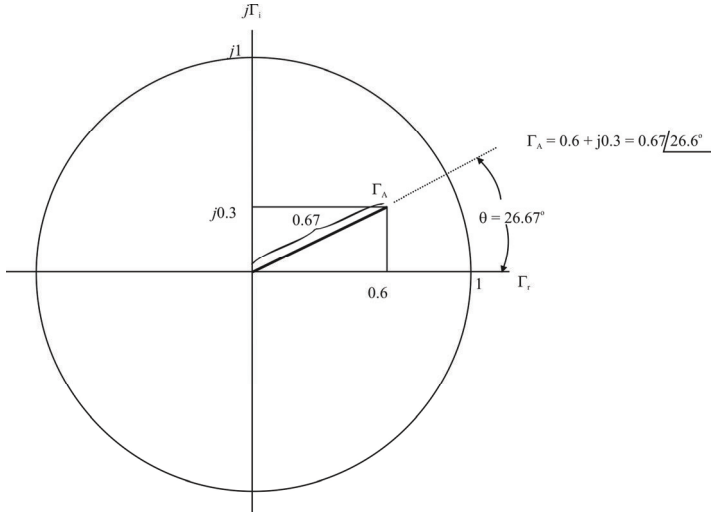


Figure 20.2 Example Smith chart.

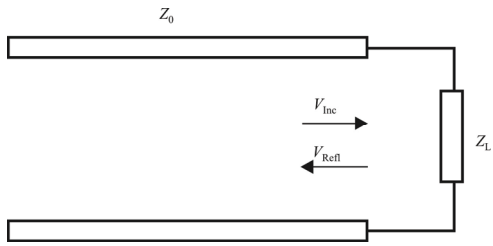


Figure 20.3 Impedance at the load.

Since the impedances are complex numbers, the reflection coefficient will be a complex number as well. Here Z_0 (the characteristic impedance) is often a constant and a standardized value that is: 50Ω, 75Ω, 100Ω, 600Ω, and so forth.

Nonzero reflection coefficients arise when a propagating wave encounters an impedance mismatch—for example, when a transmission line having a characteristic impedance $Z_0 = R_0 + jX_0$ is terminated with a load impedance $Z_L = R_L + jX_L \neq Z_0$. In that case, the reflection coefficient is

$$\Gamma = \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{\frac{Z_L}{Z_0} - 1}{\frac{Z_L}{Z_0} + 1} = \frac{z_L - 1}{z_L + 1} \tag{20.7}$$

where the impedances have been normalized to Z_0:

$$z_L = \frac{Z_L}{Z_0} \tag{20.8}$$

Analysis and design of matching a transmission line to a load are efficiently accomplished with the Smith chart. The (blank) complete Smith chart is shown in Figure 20.1. The circles on the Smith chart can be seen better in Figure 20.4. The outer circle shown solid represents where $|\Gamma| = 1$. We can see from (20.7) that the reflection-coefficient magnitudes must be 1 or less because we cannot get more reflected energy than the incident energy we apply. Therefore regions outside this circle have no significance for the physical systems the Smith chart is designed to represent (although these regions do come into play when circuit stability is analyzed).

Placing a rectangular grid overlay on the Smith chart, as shown in Figure 20.5, allows for easy determination of the reflection coefficient. The real part of Γ is shown on the *x*-axis and the imaginary part is on the *y*-axis. A graph such as Figure 20.5 provides for convenient plotting of complex reflection coefficients, but plots of reflection coefficients are not particularly useful by themselves. Typically, we want to relate reflection coefficients to complex impedances (source, line, and load impedances). The Smith chart transforms the rectangular grid of the complex *impedance* plane $Z = R + jX$ into a pattern of circles that can directly overlay the complex *reflection coefficient* plane of Figure 20.5. As such, the circles shown on Figure 20.3 represent these impedance values. Circles of

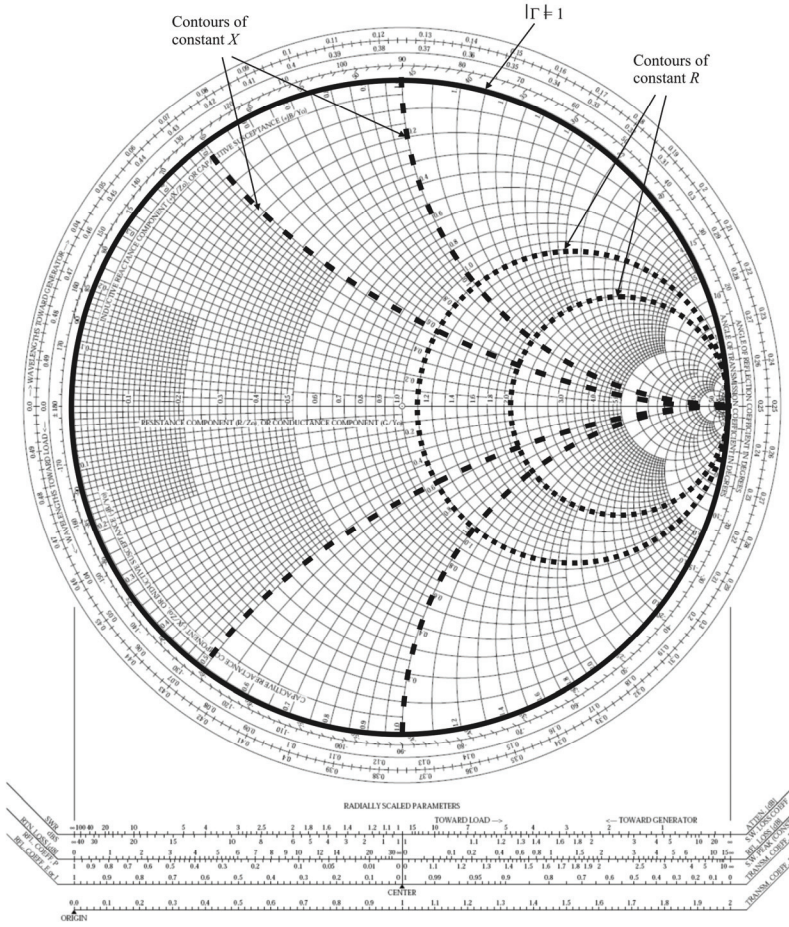


Figure 20.4 Smith chart circles. The dotted lines are contours of constant *R* while the dashed lines are contours of constant *X*.

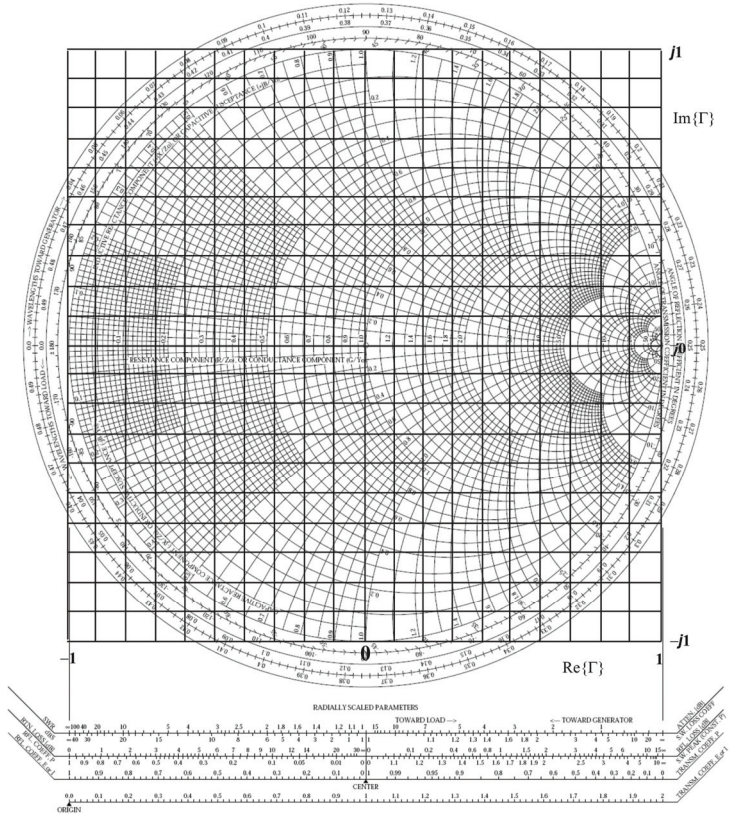


Figure 20.5 Smith chart with overlay.

constant resistance are shown as dotted lines, while circles of constant reactance (only a portion of which are shown in the Smith chart) are shown as dashed lines.

The constant resistance circles corresponding to some values of resistance and reactance are illustrated in Figure 20.6. Note that the segments lying in the top half of the complex-impedance plane represent inductive reactances ($X > 0$); those lying in the bottom half represent capacitive reactances ($X < 0$).

Overlaying the complex-impedance and complex-reflection-coefficient information on a single graph allows for easy manipulation of the values and to see how changes in one parameter affects the others. A typical situation would be to determine what complex reflection coefficient would result from connecting a particular load impedance to a system having a given characteristic impedance Consider, for example, the normalized load impedance $1 + j2$. You can locate the point representing that value on the Smith chart at the intersection of the $r_L = 1$ constant-resistance circle and the $x_L = 2$ constant-reactance contour; the intersection is point A in Figure 20.7. At this point, the reflection coefficient is $\Gamma = 0.5 + j0.5$, or $\Gamma = 0.707\angle 45°$.

As another example, the complex-impedance value $1 - j1$ is located at point B in Figure 20.7. At this point B, you can read off the corresponding reflection coefficient $\Gamma = 0.2 - j0.4$, or $\Gamma = 0.45\angle -63°$.

It is important to remember that normalized values are used on the Smith chart. For a system characteristic impedance of 50Ω, the respective values of load impedances at points A and B would be $50 + j100\Omega$ and $10 - j20\Omega$, respectively.

20.3.1 Impedance Matching and the Smith Chart: The Fundamentals

The other common use for transmission lines, especially at higher frequencies, is to create desired impedance levels in circuits [4]. A typical characteristic of the impedance of a shorted transmission line versus length (βl in electrical degrees) is illustrated in Figure 20.8 [5]. Clearly, the impedance looking into the line depends on how long the line is. Similar characteristics are observed for transmission lines terminated in an open circuit. Such a characteristic is illustrated in Figure 20.9 [5]. Examination of either Figure 20.8 or 20.9 indicates that the line input impedance can be positive or negative, depending on the length of the line. A negative impedance corresponds to a capacitive reactance, while a positive impedance corresponds to an inductive reactance. Therefore another way to view the impedance characteristics of shorted and open transmission lines is given in Table 20.2.

In addition to these impedance variation characteristics of terminated transmission lines, the phases of voltages and currents are affected depending on

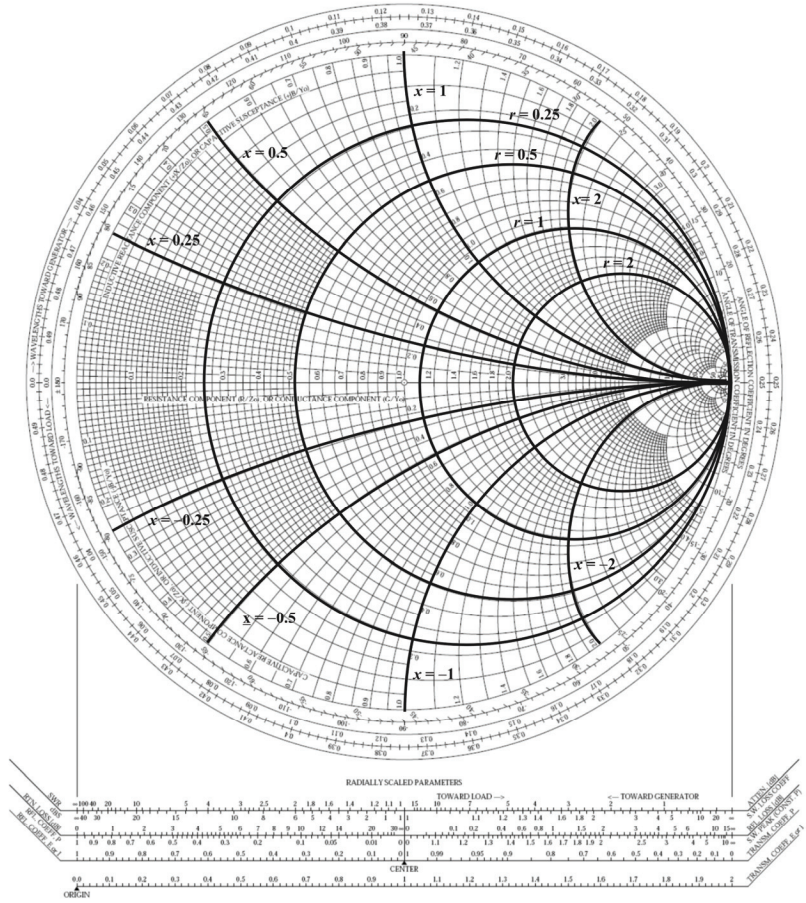


Figure 20.6 Contours of constant normalized resistance and reactance superimposed on the same chart.

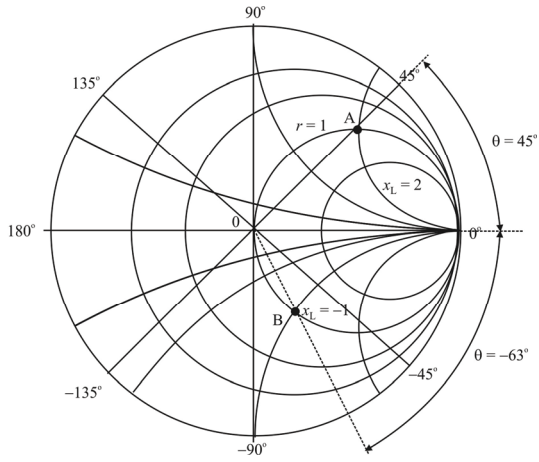


Figure 20.7 Manipulating on the Smith chart.

whether the line is terminated in a short circuit or is open (infinite impedance). These phase characteristics are indicated in Table 20.3.

It is well known that to get the maximum power transfer from a source to a load, the source impedance must equal the complex conjugate of load impedance, or

$$R_S + jX_S = R_L - jX_L \qquad \text{Maximum power transfer} \qquad (20.9)$$

For this condition, the energy transferred from the source to the load is maximized. In addition, for efficient power transfer, this condition is required to avoid the reflection of energy from the load back to the source. This is particularly true for high frequency environments of concern to us here, like RF and microwave networks.

Notable points about the Smith chart are:

- All the circles have one same, unique intersecting point at the coordinate (1, 0).
- The zero Ω circle where there is no resistance ($r = 0$) is the largest one.
- The infinite resistor circle is reduced to one point at (1, 0).
- There should be no negative resistance. If one (or more) should occur, we will be faced with the possibility of oscillatory conditions.

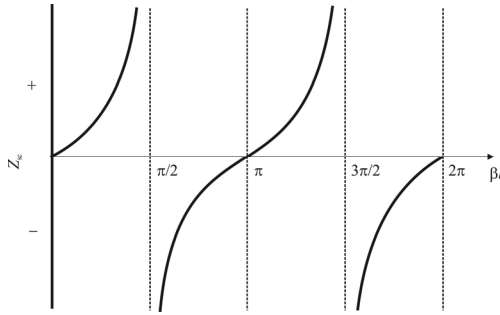


Figure 20.8 Short circuit loading. Impedance characteristics of shorted transmission line.

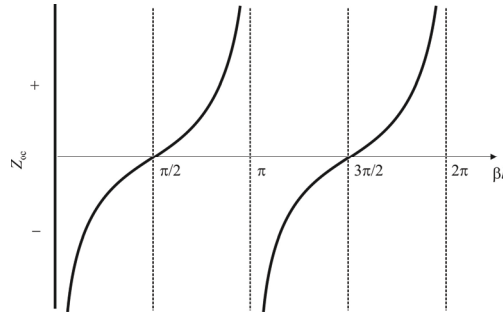


Figure 20.9 Open circuit loading.

Table 20.2 Input Impedance Characteristics of Short and Open Transmission Lines

Length	Short Circuit	Open Circuit
$0 \leq l < \lambda / 4$	Inductive	Capacitive
$\lambda / 4 \leq l < \lambda / 2$	Capacitive	Inductive
$\lambda / 2 \leq l < 3\lambda / 4$	Inductive	Capacitive
$3\lambda / 4 \leq l < \lambda$	Capacitive	Inductive

Table 20.3 Reflection Properties of Open and Shorted Transmission Lines

Transmission Line Termination	Effect on Voltage	Effect on Current
Open circuit	100% reflection no phase shift	100% reflection 180^{o} phase shift
Short circuit	100% reflection 180^{o} phase shift	100% reflection no phase shift

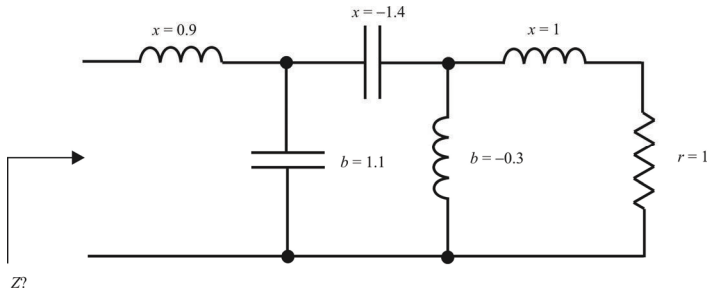


Figure 20.10 A multi-element circuit.

- Another resistance value can be chosen by simply selecting another circle. Corresponding to the new value an impedance point is plotted (the intersection point of a constant resistance circle and of a constant reactance circle).

Once the Smith chart is built, it can be used to analyze both the network series and parallel parameters. Adding elements in series is straightforward. New elements can be added and their effects determined by simply moving along the circle to their respective values. However, summing elements in parallel is another matter. It requires considering additional parameters. Often, it is easier to work with parallel elements as admittances. We know that, by definition, $Y = 1/Z$ and $Z = 1/Y$. The admittance is expressed in mhos, or Ω^{-1}. In addition, since Z is complex, Y must also be complex.

20.3.1.1 Circuit Analysis Using the Smith Chart

When solving problems where elements in series and in parallel are mixed together, we can use the same Smith chart and rotate it around any point where conversions from z to y or y to z exist. Consider the network of Figure 20.10 [6], where the elements are normalized with $Z_0 = 50\Omega$. The series reactance (x) is positive for inductance and negative for capacitors. The susceptance (b) is positive for capacitance and negative for inductance. Starting at the right, the series inductor and resistor form impedance $1 + j1$. That point is plotted on the Smith chart in Figure 20.11 as point A. The next element is in shunt so the chart is flipped into an admittance chart. This is accomplished by rotating the plane by 180^o, or, alternately reflecting point A through the $1 + j0$ point (center of the chart) to point A', which we do here. The chart is now an admittance mode. The shunt element can be added by moving along the constant conductance circle by a distance corresponding to 0.2. The direction must be counterclockwise because the

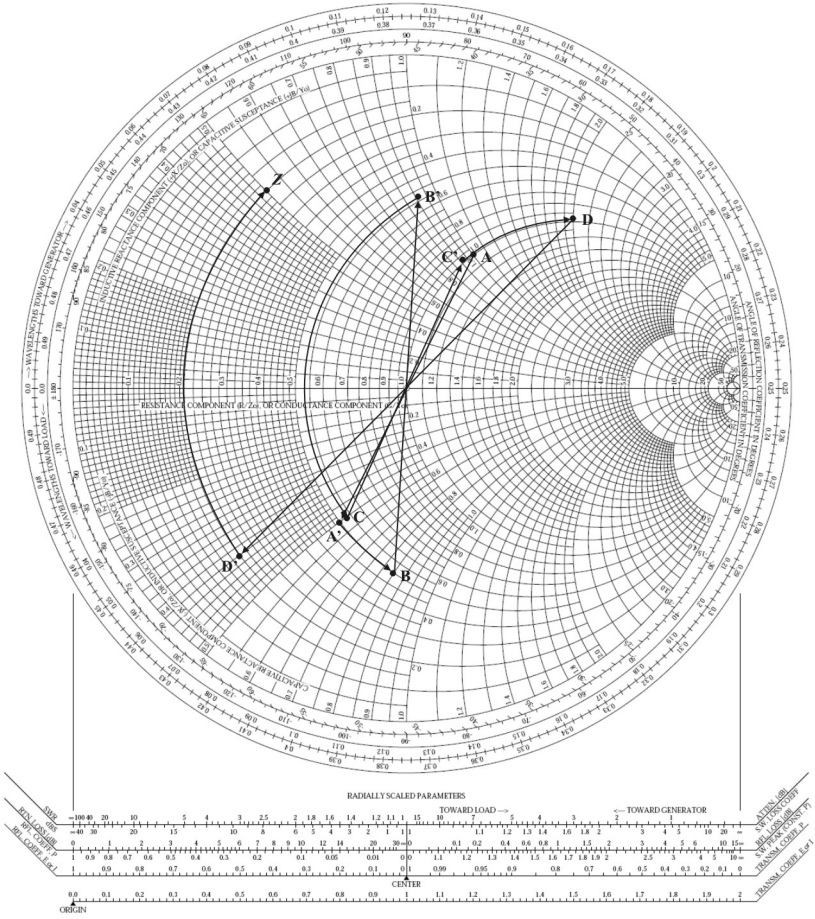


Figure 20.11 Example network analysis.

shunt value is negative, which produces point B. Next, we have another series element. We again switch back to the impedance Smith chart by reflecting through the center of the chart producing point B'. The series element is added by following along the resistance circle by a distance corresponding to 1.4 and marking point C. This has to be counterclockwise, again because of the negative value. Next, conversion to admittance by reflecting through the center of the chart produces point C'. Then move the prescribed distance (1.1), in a clockwise direction (since the value is positive), along the constant conductance circle. This point is marked D. Finally, we reconvert back to impedance mode by reflecting through the center to produce D' and add the last element, the series inductor by moving the correct distance (0.9) in a clockwise direction. The value of Z is located at the intersection of resistor circle 0.2 and reactance circle 0.5. Thus Z is determined to be $0.2 + j0.5$, which, upon denormalization yields $Z = 10 + j25\ \Omega$.

The Smith chart that contains both impedances and admittances (immittances) is also available. See Figure 20.12. In this case it is not necessary to reflect through the origin nor flip the chart over to switch from impedances to admittances and vice versa. It is simply necessary to use the alternate curves on the chart.

20.3.1.2 Transmission Lines and Impedance Matching

Another use of the Smith chart is to determine the values of elements in a matching network. This is the reverse operation of finding the equivalent impedance of a given network. This is illustrated in this section by way of an example. Here, the impedances are fixed at the two access ends (the source and the load) as shown in Figure 20.13. The objective is to design a matching network to insert between them so that proper impedance matching, and thus maximum power transfer, occurs. Series and shunt elements are added on the Smith chart until the desired impedance is achieved.

The objective is to conjugate match a source impedance, Z_S, to a load, Z_L, at the working frequency of 60 MHz, as illustrated in Figure 20.13. It is assumed that the network structure has been fixed as a lowpass, L type, as shown (for other reasons not important here).

First, the impedances are normalized. Assume Z_0 to be 50Ω. Thus $z_S = 0.5 - j0.3$, $z_S^* = 0.5 + j0.3$ and $z_L = 2 - j0.5$. Next, position the two points for z_S and z_L on the chart. Mark A for z_L and D for z_S^*. Then identify the first element connected to the load which is a shunt capacitor in Figure 20.13, which must be converted to an admittance. This gives us point A' as shown in Figure 20.14. Determine the arc portion where the next point will appear after the connection of the capacitor C. Since we do not know the value of C, we don't know where to

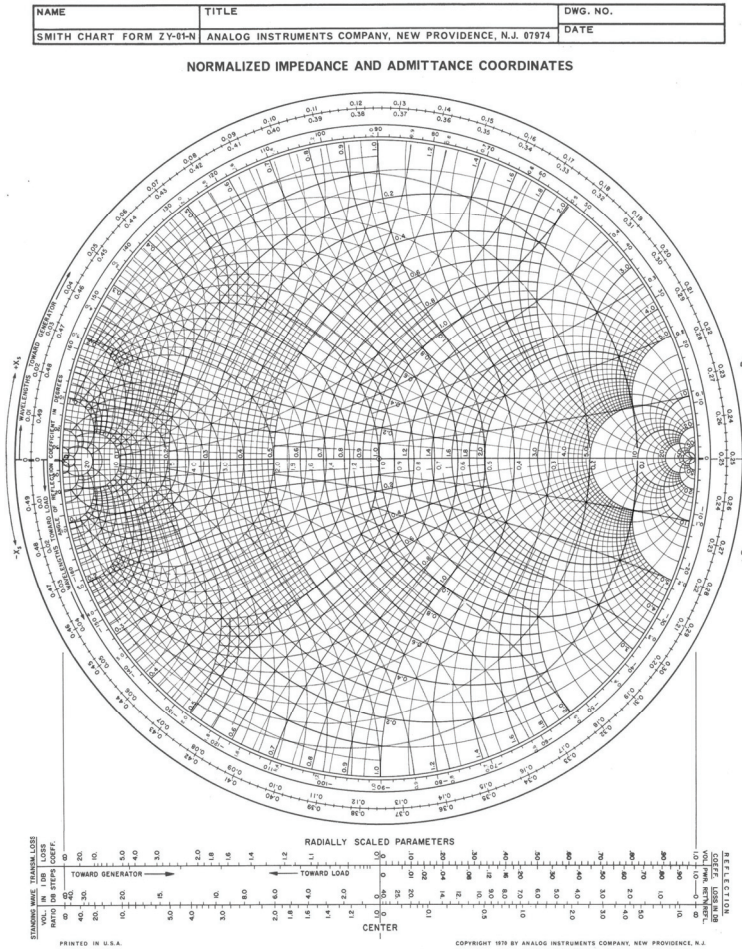


Figure 20.12 Immittance chart.

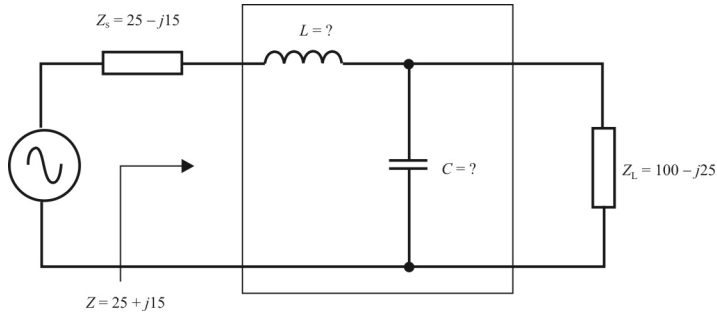


Figure 20.13 Determining the values of elements in a matching network.

stop. We do, however, know the direction. A C in shunt means move in the clockwise direction on the admittance Smith chart until the value is found. This will be point B (an admittance). Any stopping point will supposedly suffice, so pick one based on practical reasons, such as realizability. Since the next element is a series element, point B has to be converted to the impedance plane. Point B' is thus obtained. Point B' must be located on the same resistor circle as D. The intermediate points B and B' will need to be verified. After having found points B and B', we can measure the lengths of arc A' – B and arc B' – D. The first gives the normalized susceptance value of C; the second gives the normalized reactance value of L. The arc A' – B measures $b = 0.78$ and thus $B = 0.78 \times Y_0 = 0.0156$ mhos. Since $\omega C = B$, then $C = B/\omega = B/(2\omega f) = 41.4$ pF. The arc B –- D measures $x = 1.2$, thus $X = 1.2 \times Z_0 = 60\ \Omega$. Since $\omega L = X$, then $L = 159$ nH.

20.4 Antenna Matching with Lumped Elements

20.4.1 Lumped Element Matching

Although there is an infinite variety of circuits that can be used to implement matching sections, using the adage that simpler is better, simple methods of achieving the matched condition using two element L-sections are presented in this section [7]. The Smith chart introduced in Section 20.2 is used as a tool for

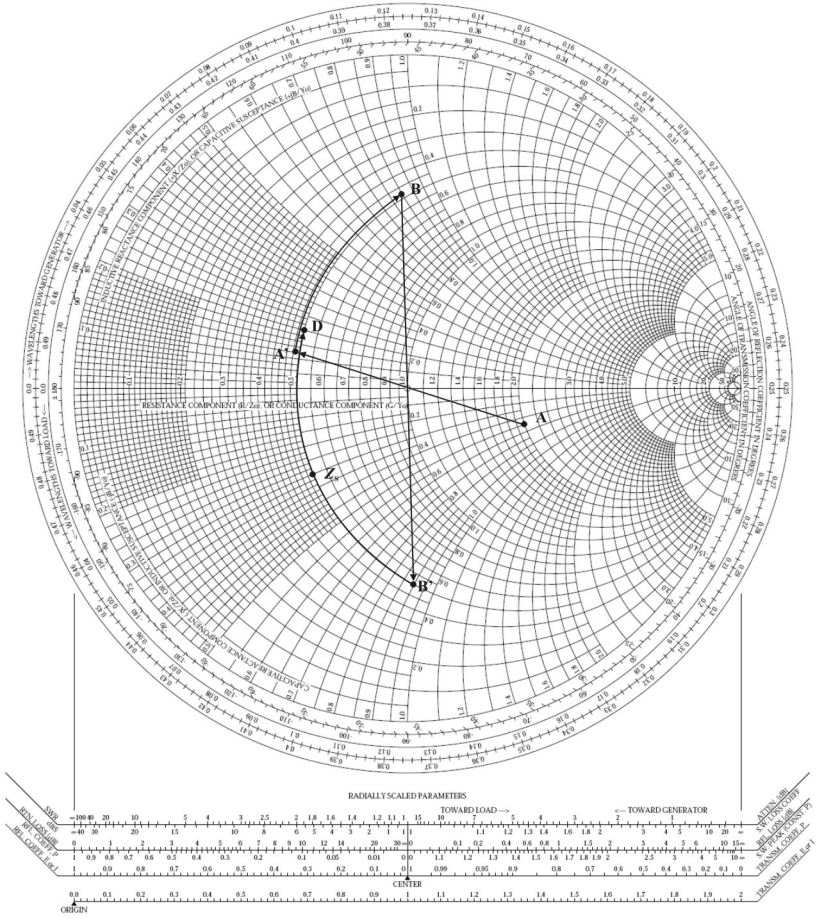


Figure 20.14 Analysis of the matching network in Figure 20.13.

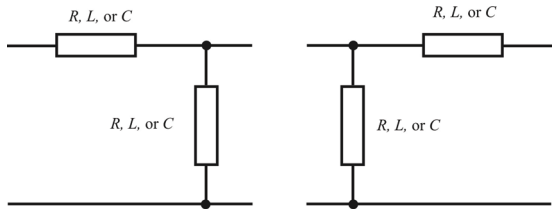


Figure 20.15 L-networks.

graphically displaying impedances and analytical expressions are given as the basis for determining the matching element values.

It is important to remember that the discussion in this section refers to matching for the explicit purpose of maximum power transfer. While important, that is not the only reason to address matching networks together. It is also important to remember that simple two-element matching circuits match a source to load at one frequency. At all other frequencies, there is impedance mismatch to more or less degree. Higher-order (more than two-element) networks can be used to match a source to a load over a broader bandwidth.

There are essentially three simple types of passive, lumped-element matching networks. The first is an L-configuration that takes the forms shown in Figure 20.15. The second is referred to as a π network and the third is referred to as a Tee network for obvious reasons. We first discuss the L-configurations and then the π and Tee networks.

20.4.1.1 Two-Element L-Network Matching

This section introduces L-networks for matching both real and complex impedances. At a single frequency, any complex impedance with a positive real part can be matched to any other impedance with a positive real part using no more than two reactive elements (capacitors and inductors).

Illustrated in Figure 20.16 are the eight unique topologies comprised of two L-C elements that can be used to match a complex impedance to a real impedance [3, 7]. Each topology is capable of matching certain complex load terminations on the network's right to a real source resistance on the network's left. The charts are normalized to the source resistance. The matchable space is enclosed by the dark curves in Figure 20.16.

Certain impedances are matched by only two topologies. For example, impedances with real components greater than the reference impedance are matched by only type 1 and 2 networks. Other impedances are matched by four topologies. For example, $Z = 0.5 + j1$ is matched by types 1, 4, 5, or 7. That is not

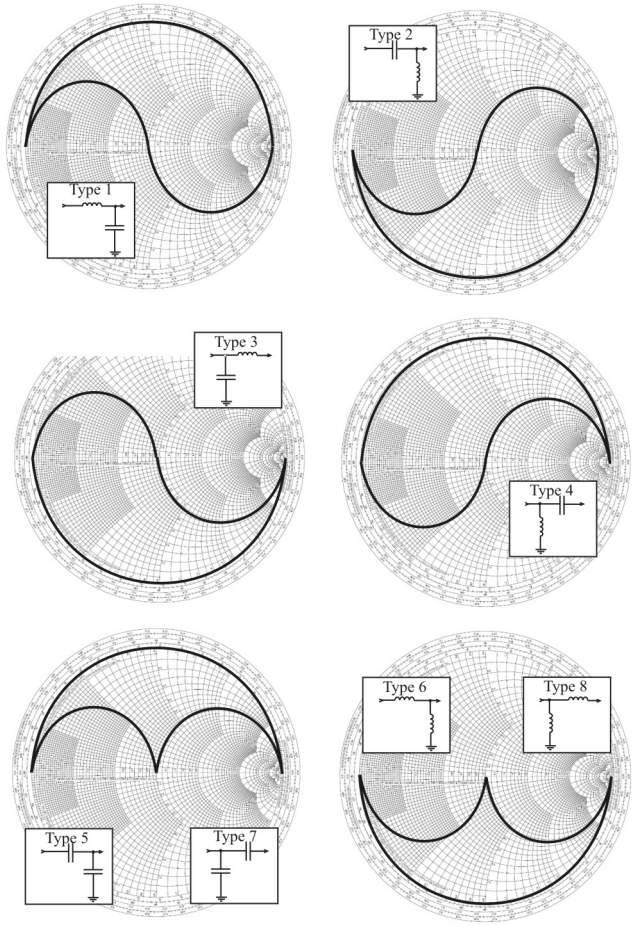


Figure 20.16 Matching Smith chart.

to say that these matchings are all equivalent. Some topologies are better at matching than others and some may produce component values that are hard to manufacture.

The reference impedance is given by $Z_0 = R_0 - j0$, the load impedance by $Z_L = R_L - jX_L$, and by straightforward algebra, the load admittance is

$$Y_L = G_L + jB_L = \frac{R_L}{R_L^2 + X_L^2} - j\frac{X_L}{R_L^2 + X_L^2} \tag{20.10}$$

Let

$$A = \sqrt{\frac{G_L}{R_0} - G_L^2} \tag{20.11}$$

$$B = \sqrt{R_L(R_0 - R_L)} - X_L \tag{20.12}$$

$$C = -\sqrt{R_L(R_0 - R_L)} - X_L \tag{20.13}$$

then the component values are (the number subscripts correspond to the corresponding circuit in Figure 20.16 [7].[1]

$$L_1 = \frac{A}{2\pi f(G_L^2 + A^2)} \tag{20.14}$$

$$C_1 = \frac{A - B_L}{2\pi f} \tag{20.15}$$

$$L_2 = \frac{1}{2\pi f(A + B_L)} \tag{20.16}$$

$$C_2 = \frac{G_L^2 + A^2}{2\pi f A} \tag{20.17}$$

$$L_3 = \frac{B}{2\pi f} \tag{20.18}$$

$$C_3 = \frac{B - X_L}{2\pi f\left[R_L^2 + (B + X_L)^2\right]} \tag{20.19}$$

$$L_4 = -\frac{\left[R_L^2 + (C + X_L)^2\right]}{2\pi f(C + X_L)} \tag{20.20}$$

[1] These are easily derived or they can be obtained from [7].

$$C_4 = -\frac{1}{2\pi fC} \tag{20.21}$$

$$C_{5a} = \frac{G_L^2 + A^2}{2\pi fA} \tag{20.22}$$

$$C_{5b} = \frac{A - B_L}{2\pi f} \tag{20.23}$$

$$L_{6a} = \frac{A}{2\pi f\left(G_L^2 + A^2\right)} \tag{20.24}$$

$$L_{6b} = -\frac{1}{2\pi f\left(A - B_L\right)} \tag{20.25}$$

$$C_{7a} = -\frac{1}{2\pi fB} \tag{20.26}$$

$$C_{7b} = -\frac{B - X_L}{2\pi f\left[R_L^2 + \left(B + X_L\right)^2\right]} \tag{20.27}$$

$$L_{8a} = \frac{C}{2\pi f} \tag{20.28}$$

$$L_{8b} = -\frac{\left[R_L^2 + \left(X_L + C\right)^2\right]}{2\pi f\left(X_L + C\right)} \tag{20.29}$$

Example: Matching with Two-Element Matching Networks

Suppose the load impedance is $Z_L = 25 + j43.33\Omega$ and it is to be matched to a 50Ω source at 100 MHz. Then the normalized load impedance is $z_L = 0.5 + j0.87$, which falls on the Smith chart as illustrated in Figure 20.17. Referring to Figure 20.17 we see that types 1, 4, 5, or 7 can be used to match this load. Equations (20.14)–(20.27) generated the values for the elements contained in Table 20.4.

Which of these topologies is best? The amplitude transmission responses are given in Figure 20.18 [7]. Notice that all four provide full transmission at the design

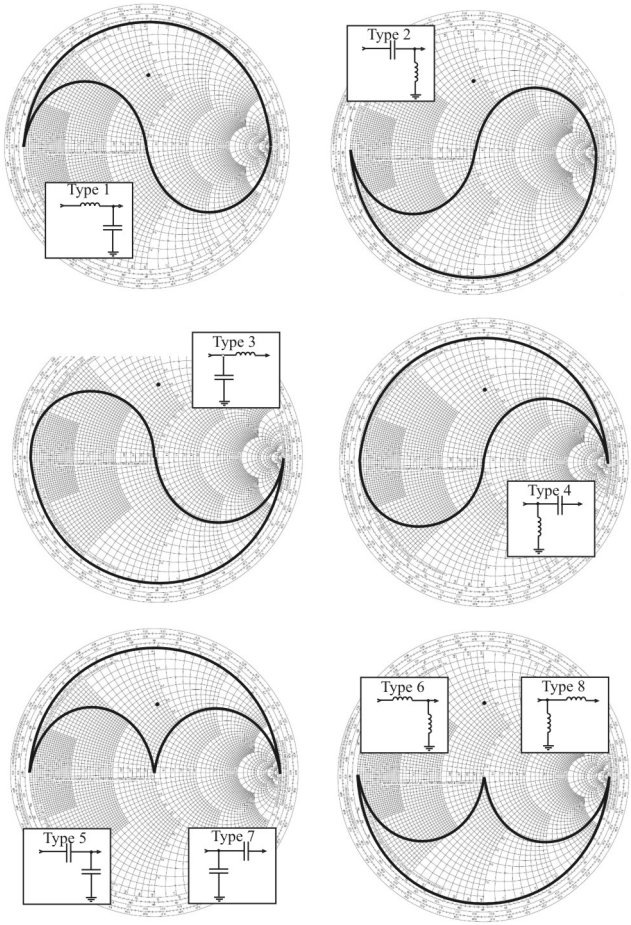


Figure 20.17 Matching example on the Smith chart. Impedance $0.5 + j0.87$ is shown as the small dot.

Table 20.4 Element Values for the Matching Example

Type 1	Type 4	Type 5	Type 7
$L_1 = 79.7$ nH	$L_1 = 79.6$ nH	$C_{5a} = 31.8$ pF	$C_{7a} = 86.3$ pF
$C_1 = 43.5$ pF	$C_1 = 20.3$ pF	$C_{5b} = 43.5$ pF	$C_{7b} = 78.5$ pF

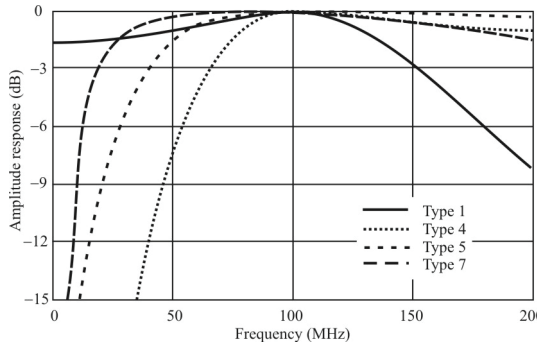


Figure 20.18 Network responses of types 1, 4, 5, and 7 that match $Z_L = 25 + j43.33\Omega$ to a 50Ω source.

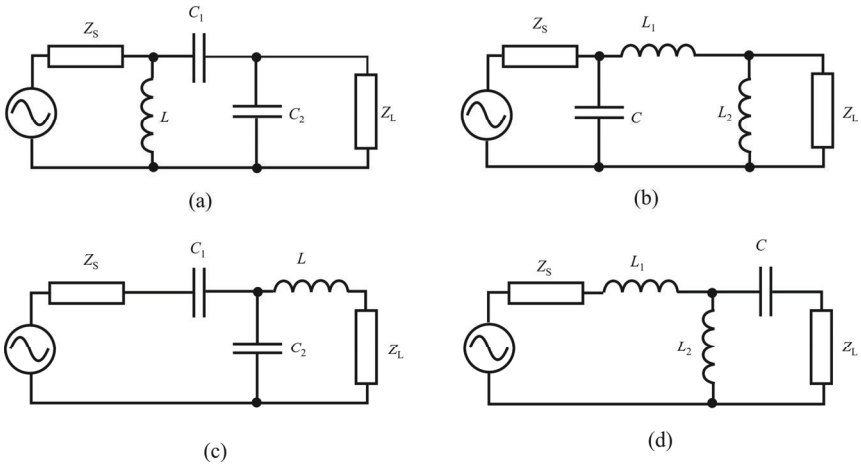


Figure 20.19(a–d) Bandpass matching networks: (a) and (b) are π networks while (c) and (d) are Tee networks.

frequency, 100 MHz. Also notice that the type 1 network is somewhat lowpass in character and that the others are somewhat highpass. The response characteristics are a function of the load impedance. In this case, type 7 provides the widest bandwidth. In general, responses become more bandpass in character and narrower for larger load reflection coefficients. In some applications wide bandwidth is desired, while for others the rejection of interfering signals by a narrow response is desired.

20.4.1.2 Three-Element π and Tee Networks

These networks [8] consist of three R, L, or C elements in shunt-series-shunt (π) shown in Figure 20.19(a, b) [9] or series-shunt-series (Tee) shown in Figure 20.19(c, d). The impedances shown in these figures can be single components or they can be networks themselves. The three-element form usually has at least one L and one C.

Bandpass arrangements are generally preferred relative to lowpass or highpass ones because of their higher gain-bandwidth product, less noise bandwidth, and their harmonics rejection capability.

For frequencies up to about 100 MHz, wound coils and lumped capacitors are usually used for implementing the L and C circuit elements, respectively. In the microwave range, for MIC and MMIC applications, short-electric-length transmission lines, spiral printed inductors and chip or printed interdigital

capacitors may be preferred. We discuss such distributed matching networks in the next section.

A matching circuit should be as simple as possible. A true bandpass design must have a minimum of three elements, for presenting a given impedance level transformation. We present in this section a complete family of three-element bandpass network topologies. Complete means that they can match any source with complex impedance to any complex load.

The set of networks is depicted in Figure 20.19, consisting of four circuits. In Figure 20.19 (a, b), there are two π networks. The first one is referred to as a capacitor divider, while the second, an inductor divider. In Figure 20.19(c, d), two Tee networks are presented—the former is a capacitor divider and the latter is an inductor divider.

Resistive Source and Load Impedances

Initially we assume that the source and load impedances are resistive with no reactive components. We will remove this restriction later. When designing a bandpass matching circuit, the designer usually knows: (1) the desired center frequency, ω_0, at which the impedance match is to be performed, and (2) the existing impedance terminations, at both sides. The matching is to be performed with the lowest insertion loss as possible; ideally it should be 0 dB at ω_0.

Additionally, the matching bandwidth would normally be one of the specifications and therefore known. This bandwidth is the –3 dB bandwidth specified by ω_1 and ω_2. In some cases, the –3 dB insertion loss is too much to be tolerated within band. In that case the –3 dB points are then moved farther apart. The relative frequency bandwidth is given by $W = 1/Q_0$, where Q_0 is the loaded quality factor of any one of the two Figure 20.19(a) or (b) arrangements.

The design equations for the two circuits depicted in Figures 20.19(a) and (b), are nearly equivalent to a parallel L-C, resonating at ω_0. This is then followed by an ideal transformer, possessing an impedance transformation ratio $n^2 = (R_S/R_L) > 1$. For the Figure 20.19(a) network, the impedance transformation ratio is obtained from the existing voltage division mechanism between C_1 and C_2, namely: $n = (C_1 + C_2) / C_1$. Additionally, the series topology of the capacitors C_1 and C_2 yields equivalent capacitance C, which, in turn, resonates with L at ω_0. Similarly, in Figure 20.19(b), $L = L_1 + L_2$, with $n = (L_1 + L_2) / L_2$.

When the load R_L is reflected through the aforementioned transformer, a value of n^2R_2 ensues, which is parallel with R_S. The perfect match condition is, at ω_0, $n^2R_2 = R_1$. In both cases, the network loaded Q_0 becomes[2]

[2] In (20.30) the notation ‖ refers to a parallel configuration, for which we recall that $R_{eq} = R_1R_2/(R_1 + R_2)$.

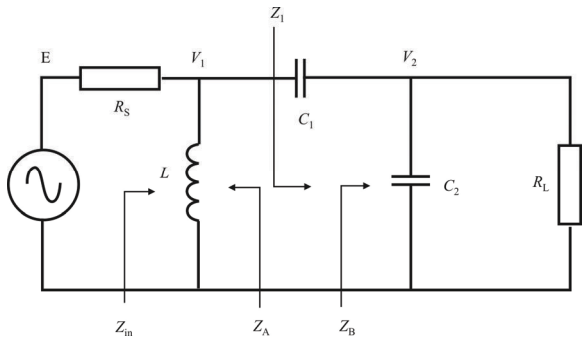


Figure 20.20 π capacitor divider three-element matching circuit with intermediary impedances and node voltages being denoted.

$$Q_0 = \frac{R_S \| (n^2 R_L)}{\omega_0 L} \tag{20.30}$$

Determining Component Values

To determine the element values for the four circuits of Figure 20.19, we will use the *Q*-based conjugate matched method [10]. Consider the circuit of Figure 20.19(a). This circuit is shown in Figure 20.20 [9], where some additional parameters are denoted as

$$Q_1 = \frac{R_S}{\omega_0 L} \tag{20.31}$$

$$Q_2 = R_L \omega_0 C_2 \tag{20.32}$$

$$n^2 = \frac{R_S}{R_L} \tag{20.33}$$

Next, we recognize that Z_A represents the parallel equivalent impedance of R_S and L, while Z_B the parallel equivalent impedance of R_2 and C_2.

$$Z_A = R_A + jX_A = R_S \frac{1}{Q_1^2 + 1} + jR_S \frac{Q_1}{Q_1^2 + 1} \tag{20.34}$$

$$Z_B = R_B + jX_B = R_L \frac{1}{Q_2^2 + 1} + jR_L \frac{Q_2}{Q_2^2 + 1} \tag{20.35}$$

Assuming a matched condition where Z_A^* is the conjugate of Z_A, and, with $Z_{C_1} = -jX_{C_1}$ being the impedance of the capacitor C_1 we get

$$Z_1 = Z_A^*$$

but

$$Z_1 = Z_{C_1} + Z_B$$

so that

$$Z_A^* = Z_{C_1} + Z_B \tag{20.36}$$

The real and imaginary parts of (20.36) are

$$n^2 = \frac{Q_1^2 + 1}{Q_2^2 + 1} \tag{20.37}$$

$$X_A + X_B + X_{C_1} = 0 \tag{20.38}$$

From (20.37), n^2 is related to Q_1 and Q_2. This allows us to obtain the Q_2 value, as we will show. In (20.38), we must remember that X_A is a positive number (an inductive reactance) while X_B and X_C are both negative (capacitive reactance). We can also see from (20.38) that there is a series resonance of X_A with X_B and with X_{C_1} at ω_0.

A power relation will be now derived for precisely dealing with the value of the network loaded quality factor, Q_0. For doing so, recall that Q_0 is the ratio of the reactive power to the active power being dissipated in the circuit [11].

$$Q_0 = \frac{\text{reactive power}}{\text{active power}} = \frac{P_X}{P_R} \tag{20.39}$$

where

$$P_X = \frac{|V_1|^2}{X_L} = |V_2 - V_1|^2 = B_{C_1} + |V_2|^2 B_{C_2} \tag{20.40}$$

$$P_R = \frac{|E - V_1|^2}{R_S} = \frac{|V_2|^2}{R_S} = 2\frac{|V_1|^2}{R_S} \tag{20.41}$$

In (20.41), we note that under the condition of conjugate matching, the input impedance of the Figure 20.20 Pi is R_S; justifying that $|V_1| = |V_E| / 2$, or $|V_E - V_1| = |V_1|$. Additionally, we must remember that the network is lossless. The power entering the π should be totally dissipated in the load. It then can be written as

$$Q_0 = \frac{|V_1|^2}{X_L}\frac{R_S}{2|V_1|^2} = \frac{R_S}{2\omega_0 L}\frac{Q_1}{2} \tag{20.42}$$

Combining the above equation with (20.37), and performing some manipulations, the Q_2 value is immediately determined

$$Q_2 = \left[\frac{4Q_0^2 + 1 - n^2}{n^2}\right]^{1/2} \tag{20.43}$$

The square root argument, in (20.43), imposes an exact solution designability condition, involving the bandwidth (Q_0) and the impedance ratio (n^2):

$$Q_0 > \left[\frac{n^2 - 1}{4}\right]^{1/2} \tag{20.44}$$

Now the component values may be easily obtained:

$$L = \frac{R_S}{2\omega_0 Q_0} \tag{20.45}$$

$$C_2 = \frac{Q_2}{\omega_0 R_L} \tag{20.46}$$

$$C_1 = \left\{\omega_0\left[\frac{R_S Q_1}{Q_1^2 + 1} - \frac{R_L Q_2}{Q_2^2 + 1}\right]\right\}^{-1} \tag{20.47}$$

We can say that there exists a hypothetical capacitor that resonates with L, at ω_0, given by:

Table 20.5 Design Equations for Any Three-Element Bandpass Matching Circuit (R_1, R_2, ω_0, and Q_0 are given)

Figure 2(a) π Capacitor Divider	**Figure 2(b) π Inductor Divider**
Obtain Q parameters	Obtain Q parameters
$Q_1 = Q_{sup} = 2Q_0$	$Q_1 = Q_{sup} = 2Q_0$
$Q_2 = Q_{inf} = [(4Q_0^2 + 1 - n^2)/n^2]^{1/2}$	$Q_2 = Q_{inf} = [(4Q_0^2 + 1 - n^2)/n^2]^{1/2}$
Obtain design elements	Obtain design elements
$L = R_1 / 2\omega_0 Q_0$	$C = 2Q_0 / \omega_0 R_1$
$C_1 = \{\omega_0[R_1Q_1/(Q_1^2+1)] - \omega_0[R_2Q_2/(Q_2^2+1)]\}^{-1}$	$L_1 = \{[R_1Q_1/(Q_1^2+1)][R_2Q_2/(Q_2^2+1)]\}/\omega_0$
$C_2 = Q_2 / \omega_0 R_2$	$L_2 = R_2 / \omega_0 Q_2$
Figure 2(c) Tee Capacitor Divider	**Figure 2(d) Tee Inductor Divider**
Obtain Q parameters	Obtain Q parameters
$Q_1 = Q_{inf} = [(4Q_0^2 + 1 - n^2)n^2]^{1/2}$	$Q_1 = Q_{inf} = [(4Q_0^2 + 1 - n^2)n^2]^{1/2}$
$Q_2 = Q_{sup} = 2Q_0$	$Q_2 = Q_{sup} = 2Q_0$
Obtain design elements	Obtain design elements
$L = 2R_2Q_0 / \omega_0$	$C = 2Q_0 / \omega_0 R_1$
$C_1 = 1/\omega_0 R_1 Q_1$	$L_1 = R_1Q_1 / \omega_0$
$C_2 = \{[Q_2 / R_2(Q_2^2+1)] - [Q_1 / R_1(Q_1^2+1)]\}/\omega_0$	$L_2 = \{\omega_0[Q_2 / R_2(Q_2^2+1)] - \omega_0[Q_1 / R_1(Q_1^2+1)]\}^{-1}$

Source: [9].

$$C = \frac{2Q_0}{\omega_0 R_S}$$

However, as already said, the above capacitor is different from the series association of C_1 and C_2.

Solutions for the Three Other Circuits

The solution of the π inductor divider is easily obtained as in the last section. It is best to use a current source as the generator for solutions for the Tees. To obtain the equivalent expressions for (20.31) to (20.47) for the Tees, it is advisable to work with the admittances, instead of the impedances. At the power relations (20.40) and (20.41) the currents, instead of voltages should be used. The design equations of the four Figure 20.19 circuits are given in Table 20.5 [10]. We can see that for the π circuits, we have $Q_1 > Q_2$, while for the Tee circuits, $Q_2 > Q_1$. We

use the notation for the respective Q factor that presents the greater value as Q_{sup}, while the other, as Q_{inf}.

Example: Three Component Matching

Consider a circuit for matching 1,250Ω to 5Ω, that is: $n^2 = 250$. The frequency is 2 MHz, with $Q_0 = 8$. The solution provides the following values: $L = 6.217$ μH, $C_1 = 1.033$ nF, and $C_2 = 2.663$ nF. In the solution, we have $Q_1 = Q_{sup} = 16$ and $Q_2 = Q_{inf} = 0.16$. The minimum insertion loss is zero, occurring at $f_0 = 2.0$ MHz. The –3 dB bandwidth goes from $f_1 = 1.885$ MHz to $f_2 = 2.135$ MHz, being 250 kHz wide, as desired, and still keeping a fair geometrical symmetry, that is, $f_0^2 = f_1 f_2$.

Matching Complex Loads

When matching complex loads, it is important to quantify the amount of reactance the circuit may absorb. The circuit's relevant parameters, for first-order complex loads, are Q_{sup} and Q_{inf}. A limit situation occurs when the circuit reactive element that touches the load is completely absorbed. The Q_{sup} and Q_{inf} values are merit factors, the higher being the better.

Example: Matching Complex Loads

We will provide for matching a 50Ω source to an 800Ω load, at the center frequency of 10.7 MHz, and using a loaded $Q_0 = 10$. We will use the Figure 20.19(a) circuit: inverted. The C_2 capacitor will be placed near the 50Ω resistance, while L will be placed near the 800Ω one, for suitably performing the step-up impedance transformation. The design values, obtained from Table 20.5, are: $L = 0.5950$ μH, $C_1 = 493.9$ pF and $C_2 = 1{,}459$ pF. The Q values are: $Q_2 = Q_{inf} = 4.905$ (near the 50Ω resistor) and $Q_1 = Q_{sup} = 20.0$ (near the 800Ω resistor).

Matching Dual Complex Loads

Only in certain cases will a given circuit in Figure 20.19 topologically conform with both terminations of an arbitrarily double terminated complex load. However, it will exist in some cases. This occurs when two sections of Figure 20.19 circuits are cascaded, forming a Butterworth shaped bandpass transformer. This is only possible to be done when a π follows a Tee, or vice versa. Further, this procedure is only worthwhile when the ports touching the loads are those presenting Q_{sup}, in both cases.

This kind of cascade must be such that we sequentially experience two step-up (or step-down) transformations, accordingly. The total resistance transformation ratio, n^2, is then split in two intermediary ones, n_A^2 and n_B^2, such as: $n_A^2 = n_B^2$ and $n_A^2 n_B^2 = n^2$.

We will center the frequency at the desired ω_0, using each one: $Q_{0A} = Q_{0B} = Q_0 / 1.4142$, to get a total quality factor of Q_0. By doing so, each block will present a less steep transformation ratio, compared with that of the total cascade. This is contributing to increase the designability range.

There are only two possibilities of arranging the circuits to fulfill the above requirements. The first one is obtained by cascading, respectively, the π and the Tee capacitive dividers, as shown in Figure 20.21(a) [10]. The second option is a cascade of the π and the Tee inductor dividers, in this order, like shown in Figure 20.21(b). In these cases, it is only possible to match a shunt inductive (or capacitive) high impedance to a series inductive (or capacitive) low impedance, by using Figure 20.21(a, b) arrangements, respectively.

Although the above arrangement is limited to only a few types of loads, when applicable, it will possess a significant capability of matching a very steep impedance ratio, together with a capability to absorb very high-Q loads, as demonstrated in the next example.

Example: Matching Dual Complex Loads

Suppose two capacitive loads are to be simultaneously matched at 10 MHz. The first one is 5 kΩ shunted by 45.02 pF. The other is 5Ω having a capacitor of 225.1 pF in series. Both loads present quality factors Q = 14.142. The step impedance is n^2 = 1,000. The –3 dB matching bandwidth is 1 MHz. This matching problem is indeed a troublesome one.

The design parameters for the first block are: R_{SA} = 5 kΩ, R_{LA} = 158.11Ω, Q_{0A} = 7.071, and $\omega_{0A} = 2\pi \times 10 \times 10^6$ rad s^{-1}. From Table 20.5, C_A = 45.02 pF, L_{1A} = 4.683 μH and L_{2A} = 1.087 μH. For the second network R_{SB} = 158.11Ω; R_{LB} = 5Ω, Q_{0B} = 7.071, and $\omega_{0B} = 2\pi \times 10 \times 10^6$ rad s^{-1}, yielding L_{1B} = 5.824 μH, L_{2B} = 1.352 μH, and C_B = 225.1 pF.

We see that the reactances were totally absorbed. The remainder circuit consists of four inductors, which may be simplified yielding the circuit depicted in Figure 20.22 [9].

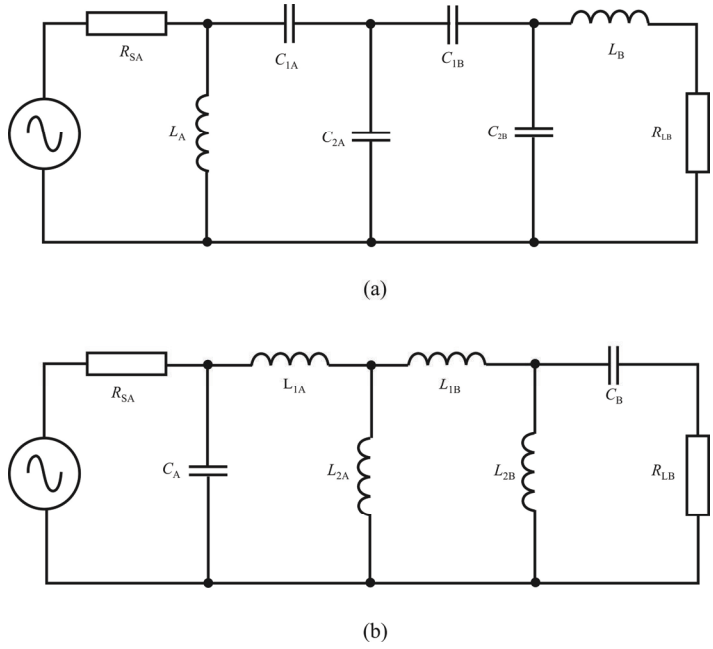


Figure 20.21(a, b) Cascaded networks: (a) for inductive loads and (b) for capacitive loads.

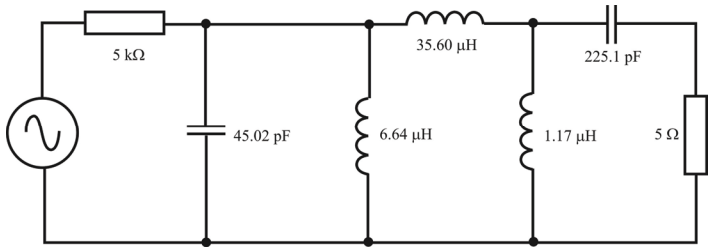


Figure 20.22 When π and Tee inductor dividers are cascaded for matching two capacitive loads; the capacitors are totally absorbed by the loads.

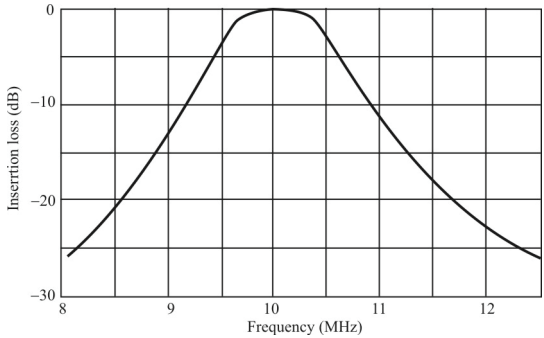


Figure 20.23 Insertion frequency response. A cascade of two three-element bandpass filters yields a Butterworth shape while matching a step impedance of $n^2 = 1{,}000$.

The frequency response for this last network insertion loss is given in Figure 20.23 [9]. The maximum flat shape is there recognized, perfectly matched at the center frequency.

Summary

The use of bandpass circuits for matching resistive and reactively constrained terminated loads has been discussed. A family of four three-element bandpass circuits has been focused. These are the simplest such circuits capable of providing a true bandpass shape together with an impedance transformation ratio. Together, they are able to match any kind of first-order reactively constrained load, for rendering either a step-up or step-down impedance scaling factor.

For complex loads, it has been shown that these circuits are able to absorb load reactance in a greater amount than others–lowpass–previously presented. Application fields for these circuits include high power EW systems. In the last stage of an EW transmitter, the match usually deals with a significant impedance step; typically from several kilo-ohms to the 25–100Ω range. The matching circuit, however, must possess a moderate Q_0, in order to keep a reasonable flat frequency response, within the useful bandwidth.

In addition, the high power level calls for a low-loss matching circuit, with a minimum of inductors as inductors typically present greater losses than capacitors. The matching condition, which usually implies reactive absorption, should be perfectly fulfilled for avoiding mismatch, which leads to the presence of very high voltages.

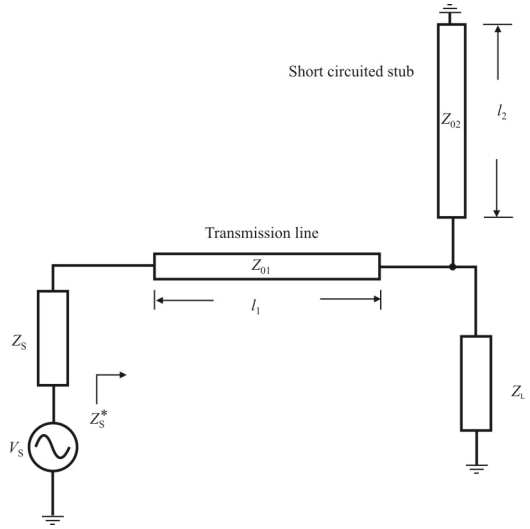


Figure 20.24 Transmission line matching.

20.5 Distributed Matching Networks

20.5.1 Stubs for Impedance Matching

A popular method of matching at microwave frequencies uses two transmission lines. One of the lines is either a short-circuited or open-circuited stub as illustrated in Figure 20.24. The matching procedure determines the lengths and the characteristic impedances of the lines so that the load is matched to the source.

The stub can either be a short-circuited or an open-circuited stub placed at either the source or the load side. This depends on the values of the two impedances. We will illustrate this method to match the transmitter and antenna with an example.

Assume that we use transmission lines with a characteristic impedance of 75Ω. With reference to the Smith chart shown in Figure 20.25, we start by locating the two impedances that we wish to match, namely, $z_T = 2 + j1$ and $z_A^* = 1 - j0.2$. Since in this case we will be adding a stub in parallel, it is more convenient to work with admittances rather than impedances. We then find the corresponding admittances $y_T = 0.4–j0.2$ and $y_A^* = 0.96 - j0.2$. From y_T we trace a constant reflection coefficient circle, representing the cascade line, until the admittance has the same real part as y_A^*. This is represented by $y_1 = 0.9+j1$. From the angles for y_T and y_1 we calculate the length l_1 of the cascade line. Next, we want to reduce the

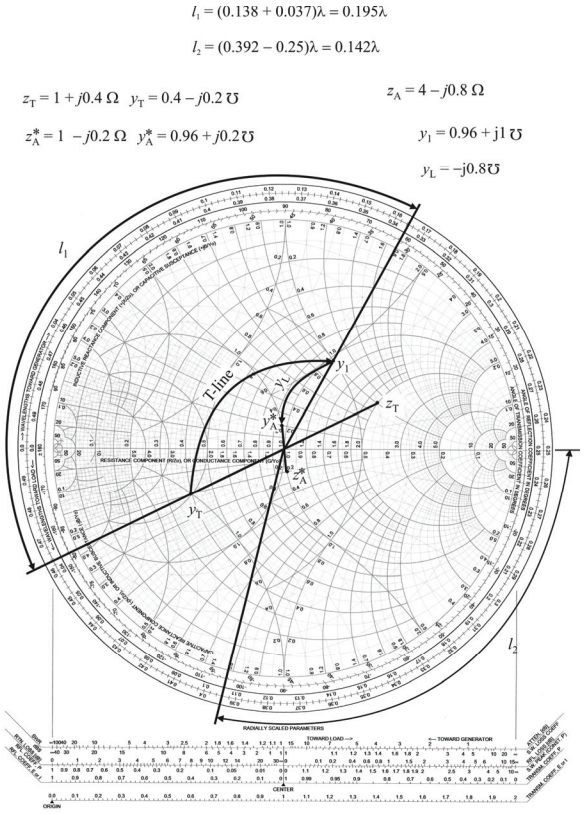


Figure 20.25 Example of stub matching.

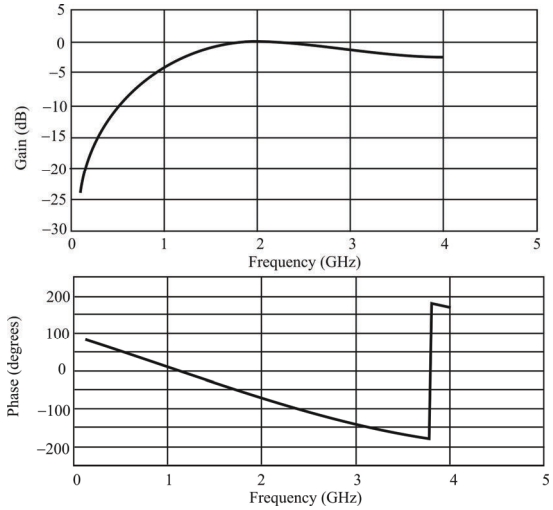


Figure 20.26 Frequency response of the stub matching.

imaginary part of y_1 from $j1$ to $j0.2$. This requires an inductance whose susceptance is $-j0.8$ at 2 GHz. This inductance is realized by a short-circuited stub of length l_2. From the Smith chart l_1=(0.158+0.195)λ = 0.195 λ and l_2 = (0.392+0.25)λ = 0.142λ. At 2 GHz, these lengths correspond to delays of T_1 = 0.195/(2x10^9) = 97.5 ps and T_2= 0.142/(2×10^9) = 71 ps. The frequency response of the stub matching is shown in Figure 20.26.

Alternately, we could replace the short-circuited stub with an inductor with the appropriate value to give the same reactance at 2 GHz. Impedance of stub = jZ_0 tan(ωT) = j92.12Ω. Value of equivalent inductor = 92.12/ω = 7.4 nH. The resulting frequency response is shown in Figure 20.27.

Distributed networks (transmission lines) are useful for matching as well [8]. The impedance at the input of a transmission line of characteristic impedance Z_0 and electrical length θ terminated in a load, Z_L, is given by (6.32) (in Chapter 6) repeated here for convenience

$$Z_S = Z_0 \frac{Z_L + jZ_0 \tan\theta}{Z_0 + jZ_L \tan\theta} \tag{20.48}$$

Suppose Z_L is a short, $Z_L = 0 + j0$. Then

$$Z_S = jZ_0 \tan\theta \tag{20.49}$$

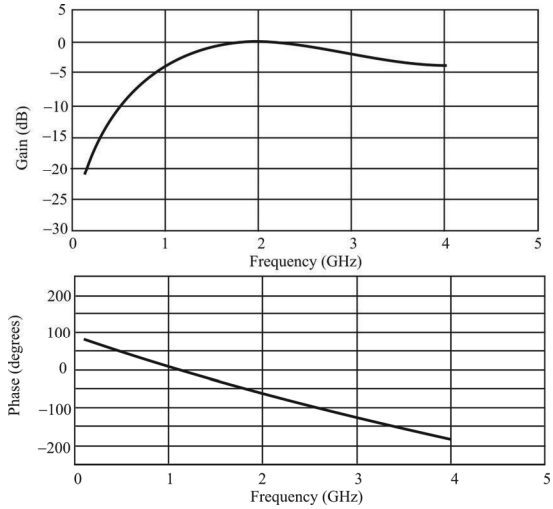


Figure 20.27 Frequency response with a 7 nH inductor in lieu of stub.

This is the impedance of an inductor of reactance

$$X_L = Z_0 \tan\theta \tag{20.50}$$

In other words, the reactance of an inductor in a network may be replaced with a transmission line of characteristic impedance Z_0 and (electrical) length θ.

An alternate approach is to design the matching network with lumped elements and replace each component with an equivalent distributed network. Figure 20.28 [7] presents the equivalent relationships between lumped and distributed circuits. Of course, these equivalents are exact only at the design frequency. The reactance of an inductor increases linearly with increasing frequency while the reactance of a shorted line increases as tan θ. If the line is short, that is, $\theta << 90°$, then $\tan\theta \approx \theta$ and the input reactance of a shorted line increases linearly with frequency. Therefore, a shorted line behaves like an inductor over a range of frequencies where the line is much less than 90° long, preferably less than 30°.

Note that at microwave frequencies, reliable shorts and opens are difficult to realize—thus the emergence of the S-parameters.

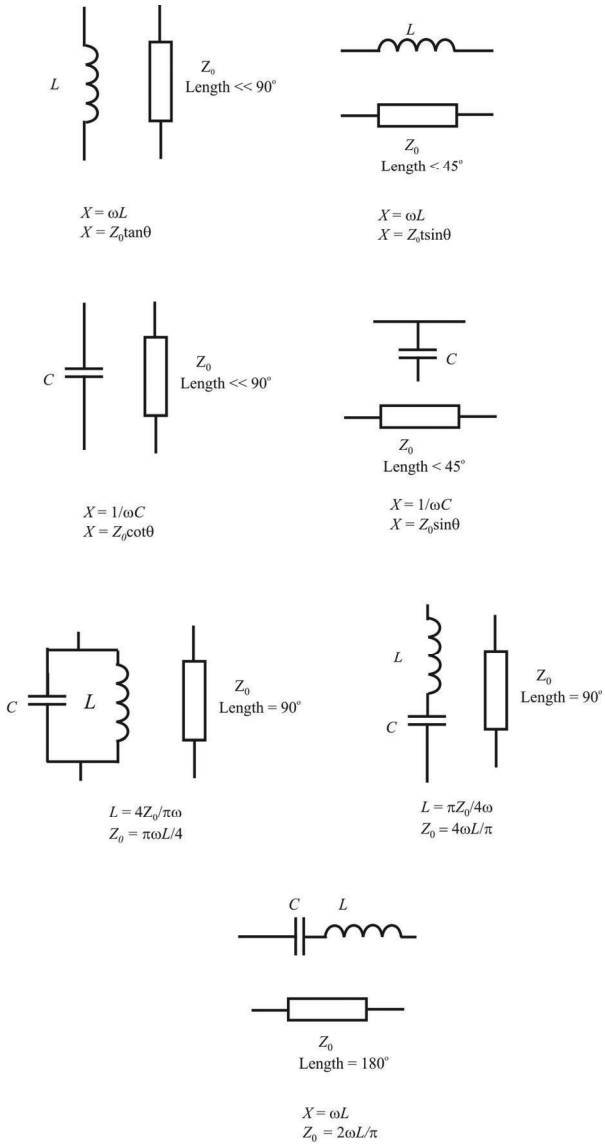


Figure 20.28 Equivalent relationships between lumped and distributed elements and resonators.

On the other hand, suppose the load is an open circuit. Then $Z_L = \infty + j\infty$ and

$$Z_S = -j\frac{Z_0}{\tan\theta} \tag{20.51}$$

which, for small θ, will act like a capacitance, which gets larger as θ gets smaller.

The distributed networks in Figure 20.28 may be used to replace lumped elements with distributed elements. Generally, the equivalence is better with higher impedance and shorter length lines for inductors, and with lower impedance and shorter length lines for capacitors. For example, a 47 nH inductor has a reactance of 29.53Ω at 100 MHz. A shorted 30.57°, 50Ω line provides the same reactance. However, a shorter (20.45°) 100Ω line would behave like this inductor over a wider frequency range.

20.5.1.1 Distributed L-Networks

Absent in Figure 20.28 is a transmission line equivalent for a series capacitor. Series capacitors are difficult to realize in distributed form. A microstrip gap requires intolerably close spacing to achieve significant capacitance and the line ends have parasitic capacitance to ground as well as the series capacitance. Edge coupled lines have even larger capacitance to ground. Therefore, the most popular form of distributed highpass filters use transmission lines for the shunt inductors and lumped-element series capacitors. However, we can see that L-network types 1, 3, 6, and 8 in Figure 20.16 do not require series capacitors. Using the equivalents in Figure 20.28, these types may be realized in distributed form.

In summary, once the lumped values are computed,

$$Z_{0,\text{SeriesLines}} = \frac{\omega L_{\text{Lumped}}}{\sin\theta} \tag{20.52}$$

$$Z_{0,\text{ShuntShortedStub}} = \frac{\omega L_{\text{Lumped}}}{\tan\theta} \tag{20.53}$$

$$Z_{0,\text{ShuntOpenStub}} = \frac{\tan\theta}{\omega C_{\text{Lumped}}} \tag{20.54}$$

20.5.1.2 Single-Stub Tuner

Single-stub tuners with shorted and open stubs and their matchable space plotted on the Smith chart are shown in Figure 20.29. The shorted-stub tuner is similar to L-network type 8, while the open-stub tuner is similar to type 3. However, by using transmission line theory embodied in (20.48), a more general solution is

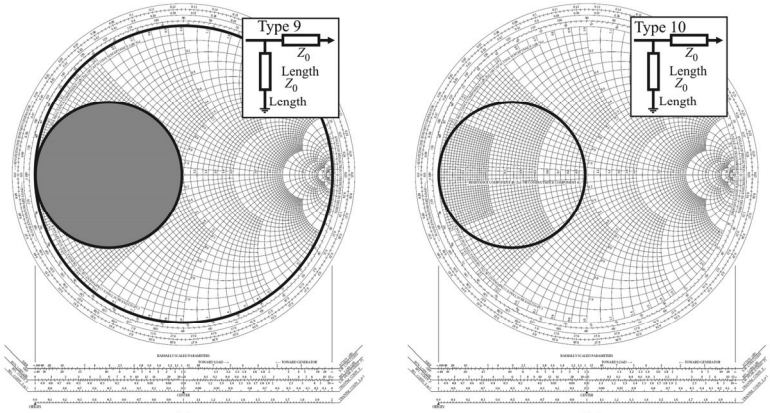


Figure 20.29 Single stub tuners: (a) a single-stub transmission line tuner with a shorted stub and (b) a tuner with an open stub. The matchable space of the shorted stub includes much of the chart, but excludes the smaller unity conductance circle in (a).

achieved that does not require short line lengths. Also, the more general solution matches a larger space than a lumped type l network. The formulas given here are for the case where the characteristic impedance of the lines in the network equals the reference impedance of the system.

When the load impedance is normalized to Z_0,

$$z_{\mathrm{L}} = \frac{Z_{\mathrm{L}}}{Z_0} = r_{\mathrm{L}} + jx_{\mathrm{L}} \tag{20.55}$$

Then the length of the series transmission line for type 9 networks, θ_{9a}, is

$$\theta_{9a} = \tan^{-1}\frac{-b - \sqrt{b^2 - 4ac}}{2a} \tag{20.56}$$

and if $\theta_{9a} < 0^{\mathrm{o}}$ then $\theta_{9a} = \theta_{9a} + 180^{\mathrm{o}}$. In (20.56)

$$a = 1 - r_{\mathrm{L}} \tag{20.57}$$

$$b = 2x_{\mathrm{L}} \tag{20.58}$$

$$c = r_{\mathrm{L}}^2 + x_{\mathrm{L}}^2 - r_{\mathrm{L}} \tag{20.59}$$

The length of the series line for type 10 networks is

$$\theta_{10a} = \tan^{-1}\frac{-b+\sqrt{b^2-4ac}}{2a} \tag{20.60}$$

The required susceptance of the stub is found by

$$b_S = \frac{(r_L^2 + x_L \tan\theta_{Series} + x_L^2 - 1)\tan\theta_{Series} - x_L}{r_L^2 + (x_L + \tan\theta_{Series})^2} \tag{20.61}$$

Then

$$\theta_{9b} = \tan^{-1}\frac{1}{b_S} \tag{20.62}$$

$$\theta_{10b} = -\tan^{-1} b_S \tag{20.63}$$

Suppose we want to match an antenna with an input impedance of 35.5 – *j*107Ω to 50Ω using the single-stub tuner. Then the series 50Ω line is 47.8° long and the shorted 50Ω stub is 20.3° long.

20.6 Antenna Coupling

We discuss the characteristics of antenna coupling between the transmission line and the antenna in this section.

20.6.1 Offset Driving Point

One way to manipulate the driving point impedance of an antenna is to offset the feed point from the center position. A common way to find the proper feed location along a half-wavelength dipole or quarter-wavelength monopole is to feed off-center by a distance *h*, which provides an increase of the input impedance with respect to the center-feed impedance, Z_s, according to

$$Z_{as}(h) = Z_s \frac{\sin^2\left(\beta\frac{l}{2}\right)}{\sin^2\left[\beta\left(\frac{l}{2}-h\right)\right]} \tag{20.64}$$

The input resistance of a half-wavelength dipole is approximately 73Ω, which is well suited for standard 75Ω coaxial lines. The quarter-wavelength monopole has

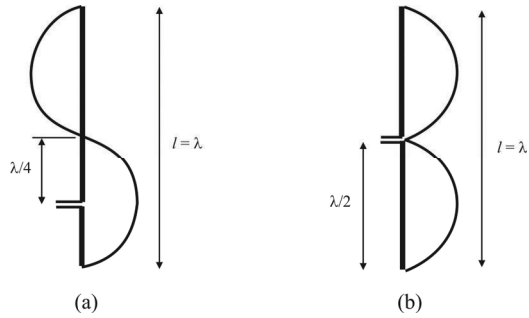


Figure 20.30(a, b) Offset feed point: (a) $h = \lambda/4$ and (b) $h = 0$. Effects of an offset feed point on the radiation pattern of a full-wavelength dipole.

an input resistance of approximately 37Ω, which is not all that bad a match to 50Ω cables but usually the sleeve-type of feed is used to achieve greater values of the antenna input impedance. As discussed in Chapter 5, the folded dipole couples well to a feed with 300Ω twin lead line.

However, the off-center feed is not symmetrical and can lead to undesirable phase reversal in the antenna if $L < \lambda/2$. This may profoundly change the radiation pattern. This effect is illustrated in Figure 20.30.

20.6.2 Symmetrical Coupling

To avoid the effects of offset coupling, symmetrical feeds for increased impedance are used. A few forms of shunt matching (or shunt feed) are shown in Figure 20.31. These dipoles are formed without the break in the middle. The balanced match is formed by spacing the points at which the feed cable satisfies a particular impedance. The attachment points are equidistant from the center of the dipole, however.

We base this discussion on the Tee match only, which gives the basic idea for all shunt feeds. The Tee-match interconnection can be viewed as two shorted

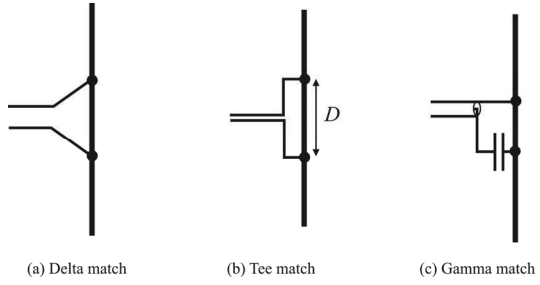


Figure 20.31(a–c) Symmetrical feeds.

transmission lines and a very-wide-gap dipole in parallel with respect to the twin-lead cable. The shorted transmission lines are less than a quarter-wavelength long, and, therefore, they have an inductive reactance. This reactance is usually greater than the capacitive reactance of the wide-gap dipole, and an additional tuning lumped capacitor might be used to achieve better match. As the distance D increases [Figure 20.31(b)], the input impedance increases, too. It has a maximum at about $D = l/2$ (half the dipole's length). Then it starts decreasing again, and when $D = l$, as expected it equals the folded-dipole input impedance. Sometimes in practice when the installation allows for them (typically not the case for wideband EW systems), sliding contacts are made between the shunt arms and the dipole for impedance adjustments and tuning. Note that shunt matches may radiate.

The Gamma-match is essentially the same as the Tee-match, but we can see in Figure 20.30(c) that it is designed for unbalanced-balanced connection. Additional matching devices are sometimes used such as quarter wavelength impedance transformers, reactive stubs for compensating antenna reactance, and so forth.

Sometimes, when high-frequency devices are connected, their impedances might be well matched, and still we may observe significant reflections. A typical example is the interconnect between a coaxial line of $Z_c = 75\,\Omega$ and a half-wavelength dipole of $Z_{in} = 73\,\Omega$. The reflections are much more severe than we would predict using (11.26). This is because the field and the current distributions in the coaxial line and at the input of the wire dipole are very different. A coaxial cable is an unbalanced source while the dipole is a balanced load. The unequal currents on the dipole's arms unbalance the antenna and the coaxial feed. To balance the currents, *balanced-to-unbalanced transformers* (baluns) are used.

20.6.3 Baluns

Most electronic sources are unbalanced—that is, the signal is carried on a wire and the other side of the signal path is ground. Meanwhile, many transmission lines are balanced, as indicated in Chapter 4. Thus, conversion between the two architectures is sometimes necessary. The conversion from an unbalanced signal from the source to a balanced signal on the line and back to an unbalanced signal again at the load is done with baluns [12]. There are three principal types of broadband baluns, shown in Figure 20.32.

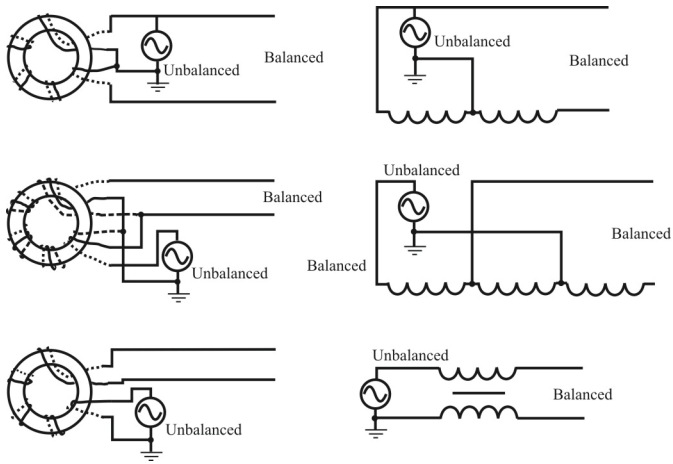


Figure 20.32 Three types of baluns. The wires in these baluns are wrapped around ferrite cores as shown on the left.

These three can be divided into two distinct categories. The first two baluns in Figure 20.32 are voltage baluns causing equal and opposite voltages on the two output terminals. The third balun is a current balun, also called a *choke balun*, forcing equal and opposite currents to flow on the line. If the line and the terminating load are perfectly symmetrical, the voltage baluns will force the voltages to be equal and opposite and thus the currents flowing from the balun output terminals will be equal and opposite and there will be no *common mode* (CM) current on the line. If, however, the line is not perfectly symmetrical, unequal currents, resulting in CM currents, will appear leading to line radiation.

A solution for reducing line radiation is to use a current balun, which causes opposite currents to flow and CM currents are rejected. Another possibility is to use a voltage balun in combination with a current balun. Notice that even if the voltage is perfectly balanced at the ends of the line, some CM currents may still appear due to nonsymmetrical line and load. When using a current balun on a balanced line, it will function as a CM suppressor. The wire is wrapped around a ferrite core as in Figure 20.32. Then the differential currents of opposite phase will produce magnetic fluxes of opposite phase in the core that cancel each other and there will be no remaining reactance in the core for differential-mode signals. The CM current will, however, produce a magnetic flux that appears as a reactance in series with the line. Ferrite cores are well suited as core materials in this application, but the characteristics vary considerably with frequency and with the ferrite material as illustrated in Figure 20.33.

In Figure 20.34, a coaxial cable is connected to a dipole antenna. For a dipole antenna to operate properly, the currents on both arms of the dipole should be equal in magnitude. When a coaxial cable is connected directly to a dipole antenna

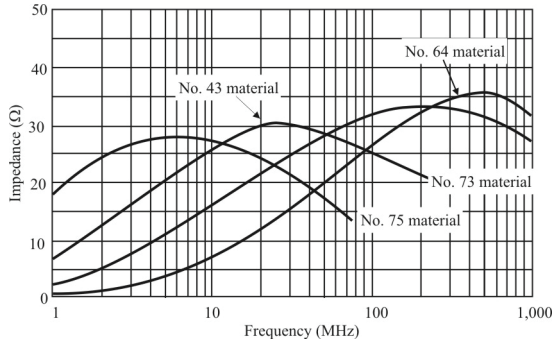


Figure 20.33 Impedance of some ferrite cores. The composition and construction of the core material dictate its numbered specification. No. 43 is a very common material for ferrite cores due to its broad applicability.

however, the currents will not necessarily be equal. To see this, note that the current along a transmission line should be of equal magnitude on the inner and outer conductors, as is typically the case. We can observe what happens when the coax is connected to the dipole: the current on the center conductor (labeled I_1) has nowhere else to go, so must flow along the dipole arm that is connected to it (down the inner conductor in this case). However, the current that travels along the inner side of the outer conductor (I_2) has two options: it can travel down the dipole antenna, or down the reverse (outer) side of the outer conductor of the coaxial cable (labeled I_3 in Figure 20.34).

Ideally, we would like the current I_3 to be zero. In that case, the current along the dipole arm connected to the outer conductor of the coax will be equal to the current on the other dipole arm. Because the dipole wants equal or balanced currents along its arms, it is the balanced section. The coaxial cable does not necessarily give this condition, however—some of the current may travel down the outside of the outer coax, leading to unbalanced operation. This is the unbalanced section.

The solution to this problem is a balun. A balun forces an unbalanced transmission line to properly feed a balanced component. In Figure 20.34, this would be done by forcing I_3 to be zero somehow. This is often called choking the current or a current choke.

There are many baluns that have been developed to choke off the outer current and restore balanced operation. Some of the most popular methods are described in the following. These are the:

- Bazooka balun (sleeve balun);
- Folded balun;

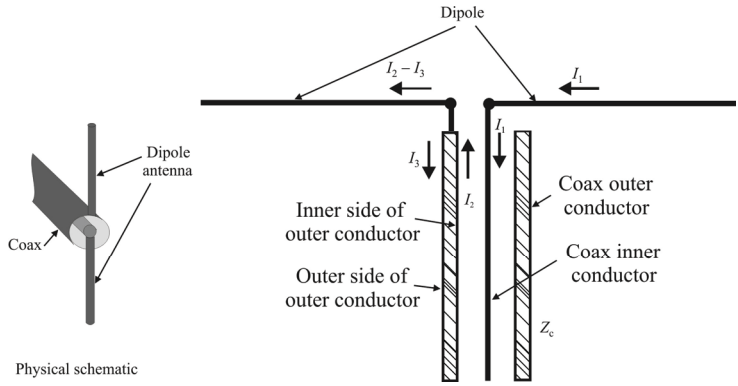


Figure 20.34 Antenna matching. An unbalanced transmission line (coaxial cable) connected to a dipole antenna.

- Tapered balun.

Suffice it to remember that a balun forces unbalanced lines to produce balanced currents, despite their inherent asymmetry. Also recall that it is the (accelerating/decelerating) currents in an antenna structure that cause radiation.

20.6.3.1 Sleeve (Bazooka) Balun 1:1

In the sleeve antenna shown in Figure 20.35, the sleeve and the outer conductor of the coaxial feed form another coaxial line, which has a characteristic impedance of Z_c'. This line is shorted a quarter wavelength away from the antenna input terminals (see Figure 20.35). This is a narrowband balun, which does not have impedance-transforming capability (1:1 balun).

20.6.3.2 Folded Balun 1:1 (split-coax balun, λ/4-coax balun)

A folded balun is depicted in Figure 20.36. This 1:1 balun is easy to make and is narrowband. The outer shield of the feeding coaxial line and the additional coax-line section form a twin-lead transmission line, shorted a distance ≈ λ/4 away from the antenna input. This line is in parallel with the antenna but does not affect the overall impedance because it has infinite impedance at the antenna terminals. The additional piece of coaxial line re-directs a portion of the I_1 current, which induces the twin-lead current I_4. The currents I_3 and I_4 are well balanced ($I_3 = I_4$) because the current of wire #1 ($I_2 - I_3$) would induce as much current at the outer coaxial shield I_3, as the current of wire #2 ($I_1 - I_4$) would induce in the outer shield of

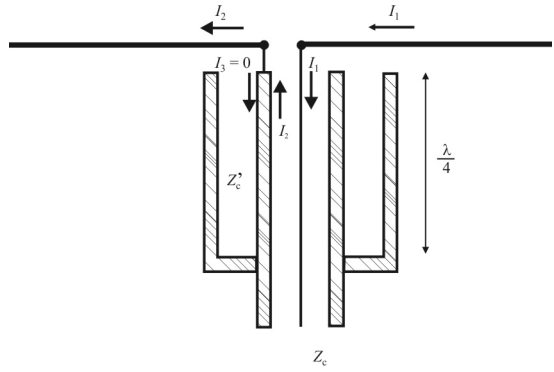


Figure 20.35 Bazooka balun.

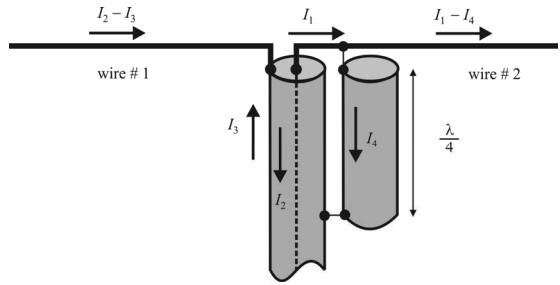


Figure 20.36 Folded balun.

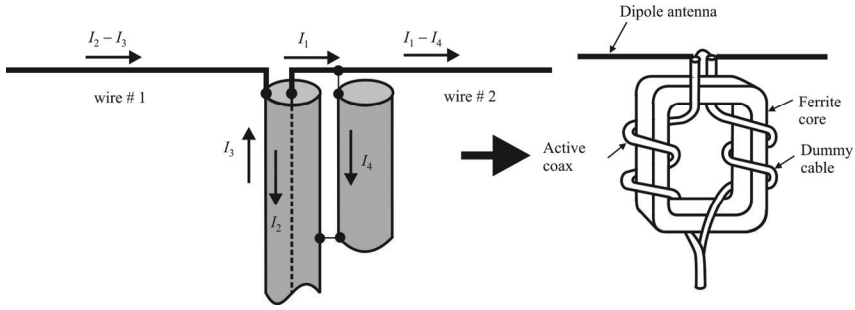


Figure 20.37 Broadband balun.

the auxiliary coaxial piece I_4 (note the geometry similarity of the interconnects), that is,

$$\frac{I_3}{I_2 - I_3} = \frac{I_4}{I_1 - I_4}$$

Since $I_1 = I_2$ in the feeding coaxial line, it is also true that $I_3 = I_4$. Therefore, the current at the outer coaxial shield is effectively canceled from a certain point on ($\approx \lambda/4$).

20.6.3.3 Broadband Baluns

The baluns described in the previous two sections are narrowband because of the critical dependence on the wavelength of the auxiliary transmission-line sections. Broadband baluns for high-frequency applications can be constructed by tapering a balanced transmission line to an unbalanced one very gradually, over a distance of at least several wavelengths (microstrip to twin lead, coax to twin lead).

At lower frequencies (below UHF), tapered baluns are impractical, and transformers are used for impedance adjustment and balancing the feed, as depicted in Figure 20.37. Often, ferrite core bifilar wound-wire baluns are preferred for their small dimensions and broadband characteristics (bandwidths of 10:1 are achievable). A ferrite core transformer 1:1, which is equivalent to the folded balun 1:1, but is much more broadband, is shown in Figure 20.37.

The transmission line formed by the outer shields of the two coaxial lines is now a very high-impedance line because of the high relative permeability of the ferrite core. Thus, its length does not depend critically on λ, in order not to disturb the antenna input impedance.

20.6.4 Broadband Antenna Matching with Transformers

Transmission line transformers (TLT) are extremely flexible devices for impedance matching and other purposes [13, 14]. The following excerpt is from [15]:

> …it [a TLT] remains the most powerful impedance matching network available in the 10 kHz to 2 GHz frequency range. When properly designed and constructed, the versatility of this network is unparalleled:
>
> - Bandwidth ratios in excess of 22,000:1;
> - Capable of low loss operation with multi-octave bandwidths at frequencies from audio to microwave (10 kHz to over 2.0 GHz);
> - High power handling ability in the multi-kilowatt level;
> - Configurable as a balun, converting balanced networks to unbalanced and vice versa;
> - Highly efficient (>95%).

A legitimate question to ask is: What good is a device that retains its characteristics from 10 kHz to 2 GHz when the antenna does not [antenna impedance can vary from very low (10Ω) to high (1 kΩ) as the frequency varies over 2:1]? At least the varying transformer characteristics become irrelevant and we can pay more attention to what the antenna is doing.

We discuss the TLT in much more detail in Chapter 21 where antenna multicouplers are presented.

20.6.5 Coupling to a Monopole Antenna

A relatively simple technique to couple to a monopole antenna is shown in Figure 20.38. In this case, a transformer is used to provide the coupling. A monopole antenna is inherently unbalanced and the transformer can be configured as shown so that it is unbalanced as well.

20.7 Matching Component Computation Software

4NEC2 computes matching filter component values for L-matching networks as well as π matching networks and Tee matching networks. Suppose we want to match the dipole antenna shown in Figure 20.39 to a 5Ω transmission line and a

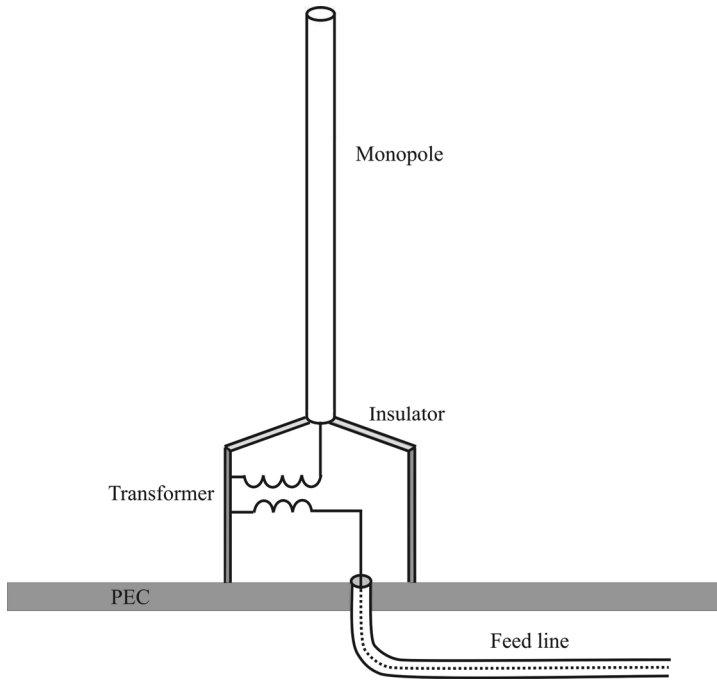


Figure 20.38 Monopole feed.

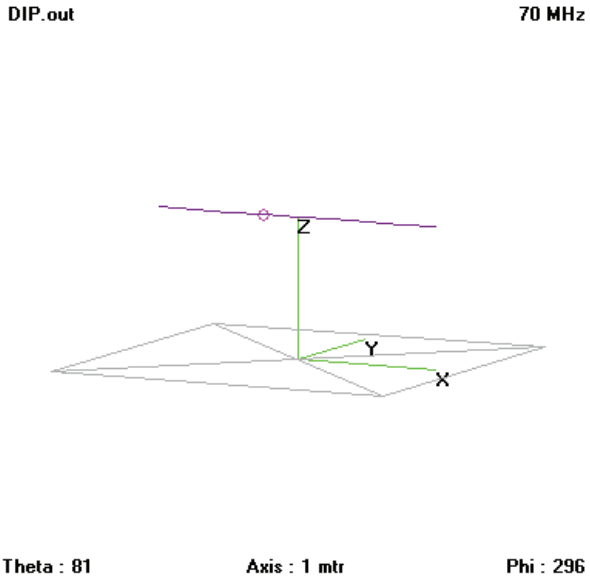


Figure 20.39 Dipole antenna model.

75Ω stub at 70 MHz. The antenna impedance at this frequency (l = 2 m) is 84.7 – j40Ω. The antenna power pattern for this antenna is illustrated in Figure 20.40.

The calculated RLC matching networks are shown in Figure 20.41 for the RLC matching networks, while the stub matching network is shown in Figure 20.42.

20.8 Concluding Remarks

We examined the basic principles of matching transmission lines to antennas in this chapter. Such matching is required for maximizing the power transfer from a source to a load. We showed two and three passive component networks that, as a set, can match any source to any load. The Smith chart was the primary tool used to investigate these matching techniques.

Methods of coupling the transmission line to antennas were also presented.

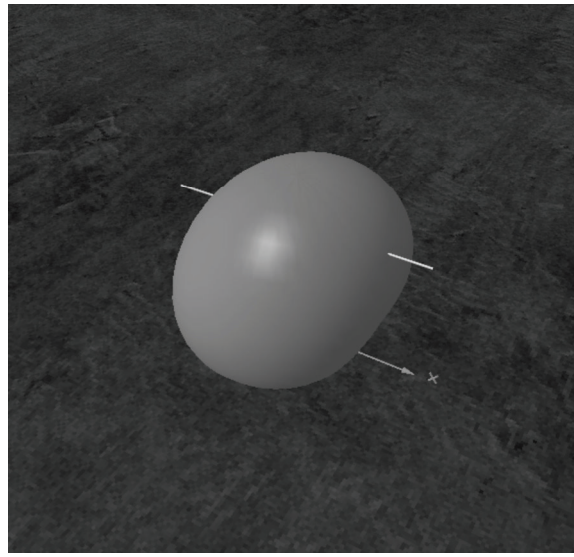


Figure 20.40 Dipole 3-D pattern.

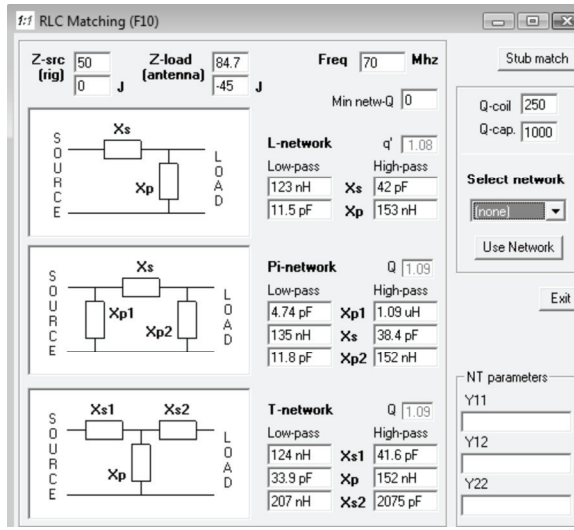


Figure 20.41 RLC matching screen in 4NEC2.

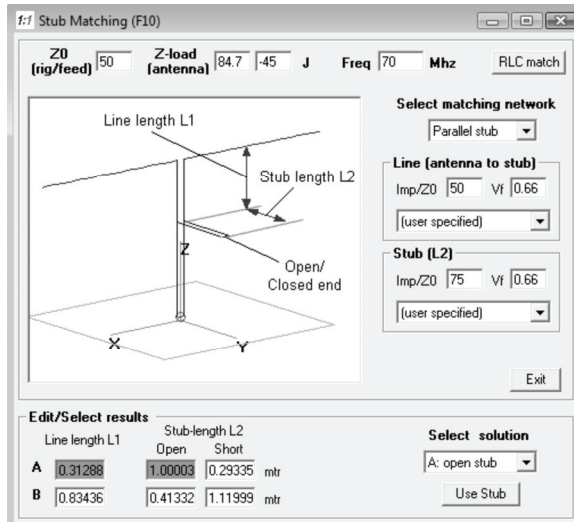


Figure 20.42 Stub matching screen in 4NEC2.

References

[1] "Philip Smith, Electrical Engineer," an oral history conducted in 1973 by Frank A. Polkinghorn, IEEE History Center, Rutgers University, New Brunswick, NJ. www.ieee.org/organizations/history_center/oral_histories/transcripts/smith2.html.

[2] "Parallel Resistance Nomograph,"www.tmworld.com/reference/nom_oddsends.htm.

[3] Smith, P. H., *Electronic Applications of the Smith Chart,* 2nd ed., Atlanta, GA: Noble Publishing, 2000.

[4] Chan, K.-C., and A. Harter, "Impedance Matching and the Smith Chart—The Fundamentals," www.rfdesign.com, July 2000, pp. 52–66.

[5] Potter, J. L., and S. Fich, *Theory of Networks and Lines*, Englewood Cliffs, NJ: Prentice-Hall, 1963, p. 336.

[6] www.rfdesign.com, July 2000.

[7] Rhea, R., "The Yin-Yang of Matching: Part 1—Basic Matching Concepts," *High Frequency Electronics,* March 2006, pp. 16–25.

[8] http://www.unimep.br/phpg/editora/revistaspdf/rct13art11.pdf.

[9] Podcameni, A., "A Complete Family of Simple Bandpass LC Matching Circuits," *Revista De Ciencia and Technologia*, June 1999, pp. 119–130.

[10] Clark, K., and D. Hess, *Communications Circuits: Analysis and Design,* 2nd ed., Reading, MA: Addison-Wesley, 1978.

[11] Van Valkenburg, M. E., *Introduction to Modern Network Synthesis*, New York: Wiley, 1967.

[12] http://www.antenna-theory.com/definitions/balun.php.

[13] Sevic, J., *Transmission Line Transformers*, 4th ed., Rayleigh, NC: Scitech Publishing, 2001.

[14] Abrie, P. L. D., *The Design of Impedance-Matching Networks for Radio-Frequency and Microwave Amplifiers*, Dedham, MA: Artech House, 1985, Ch. 5.

[15] Walker, J. L. B., D. P. Myer, F. H. Raab, and C. Trask, Eds., *Classic Works in RF Engineering*, Norwood, MA: Artech House, 2006.

Chapter 21

Multicouplers, Combiners, and Diplexers

21.1 Introduction

A receiving multicoupler and its transmit companion, the combiner, are ways of sharing antenna assets among several receivers or transmit amplifiers, respectively. We discuss multicouplers, combiners, and diplexers together in this chapter because these devices are often the first (multicoupler), last (combiner), or both (diplexer) between the antenna and the rest of the EW system. Other components that interface antennas with systems are circulators and isolators. We discuss those at the end of the chapter.

A multicoupler that drives eight receiver channels is depicted in Figure 21.1. The simplest way to share a receiving antenna is simply to split the signal from the antenna and send it to as many receivers as required. The problem with this approach is that for each fan-out of two there is at least a 3 dB loss in signal strength for each output. That is, a fan-out of two is 3 dB, of four is 6 dB, of eight is 9 dB, and so on (see Figure 21.2).

A better approach is to amplify the antenna signal prior to splitting it to make up for these losses as illustrated in Figure 21.1. If there are eight receivers to provide signals for, then a 9 dB amplifier placed between the antenna and the splitter provides the same signal level to all receivers as if there were only one receiver present.

Space at the top of an antenna mast is always an extremely limited commodity for EW systems. This creates the dual problems of placing antennas and increased risk of interference from overcrowding. Thus, the established options of transmitter antenna combining and receiver antenna multicoupling are as essential to land mobile and naval EW today as ever. One solution to the problem of fewer available sites is to make do with what you have by optimizing the number of antennas used. A combiner and multicoupler system can also reduce the level of

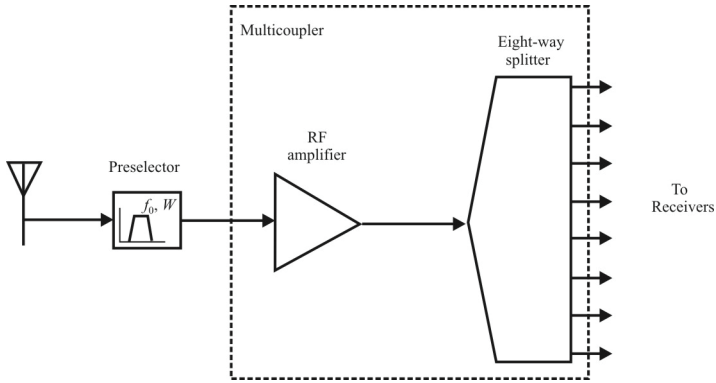


Figure 21.1 Basic multicoupler.

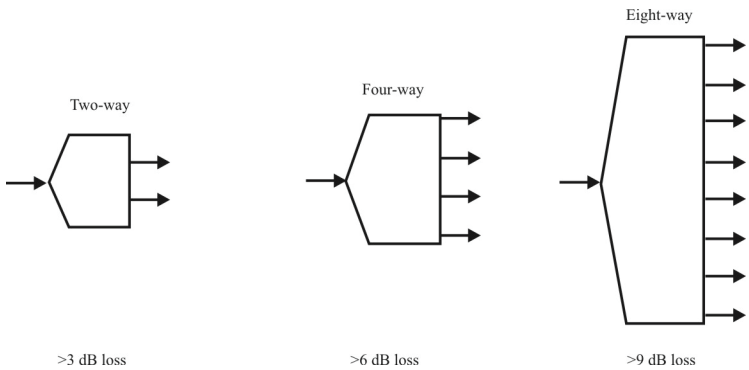


Figure 21.2 Splitter losses.

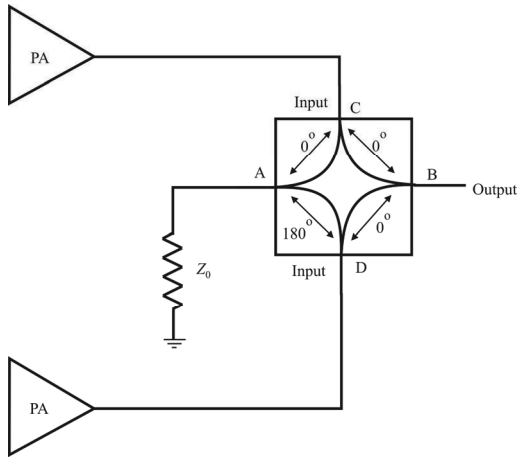


Figure 21.3 Two-way combiner.

intermodulation interference between some transmitters or receivers at a receiving site. Isolation also needs to be considered. Multicoupler configurations can be transmitter multicouplers, receiver multicouplers, or transmitter/receiver multicouplers.

A basic two-way combiner is illustrated in Figure 21.3 where the outputs of two PAs are added together. In ideal conditions when the input signals are identical, the two inputs are perfectly added at port B and completely cancelled at port A, yielding no current flow out of port A into Z_0. Practically, some power is dissipated in Z_0 due to nonideal components. Note that Z_0 is typically contained within the combiner. As we discuss later, operation of a combiner like this depends on the two input signals and their characteristics, such as same frequency or different frequencies.

This chapter is structured as follows. We begin with a discussion of multicouplers, which are used in receive configurations to feed multichannel receiving architectures. That is followed with a presentation of the characteristics of transmitter multicouplers, also known as combiners. They are often used to combine the outputs of two or more PAs to feed a common antenna. We conclude the chapter with discussions of RF diplexers and RF circulators and isolators, components associated with the interface to antennas, both for receiving and transmitting.

21.2 Receiving Multicouplers

There are several benefits from a well-designed multicoupler system. The proper preselector can provide optimum filtering of signals. A lower noise figure can be

provided for all receivers in the system (discussed later). Amplification provides wider dynamic range than otherwise possible. If properly done, coupling of two or more receivers into one antenna should yield equivalent or superior performance compared to having a dedicated antenna for each receiver.

Multicouplers for virtually every frequency are readily available from HF on up. Depending on the application, multicoupling can be as simple as two channels from one antenna to as many as 192. The most common systems have geometrically increasing numbers of available receiver outputs: two, four, eight, 16, 32, and 64. Custom sizes are also available. Some systems are expandable, allowing further branching into more receivers.

Receiver multicoupler systems can be rack-mounted in an equipment shelter, or, in some cases, there is an advantage to using a mast-top system directly adjacent to the antenna. Figure 21.1 shows a block diagram of the layout of a basic receiver multicoupler system. The transmission line from the system receiving antenna connects to the input of the preselector. The preselector output is routed to an RF amplifier, typically a *low-noise amplifier* (LNA) the output of which feeds the signal power divider. Characteristics of each of the components of the multicoupler are summarized in the following.

The most fundamental use of a receiver multicoupler is to enable several radio receivers to utilize a single antenna system without interaction that may degrade each receiver's performance. Many variations in the design concept exist as well as circuit options; however, signal splitting and the distribution of an RF signal are the basic needs to be accomplished in all cases.

Receiver multicouplers are available in two basic categories, the passive and active versions. Both will perform the fundamental task of splitting an RF signal and feeding several receivers without degrading performance of each receiver in the system. The primary difference between a passive and an active unit is in how signal losses due to signal splitting are corrected for. As indicated in Figure 20.2, passive couplers have loss and effectively add noise, which increases with the number of splits required. Due to design topology, splits are normally binary 2, 4, 8, etc., and, as indicated in Figure 21.2, with a loss that is equivalent to the number of splits. The actual losses are 0.5–1 dB above these theoretical figures. Thus, when using a passive splitter, losses to an incoming signal can be quite high.

Dependent on the range of frequencies to be split, different approaches are taken. The splitting elements themselves tend to have characteristics that limit the band to a range of frequencies that are efficiently split.

21.2.1 Antennas

Although the antenna is supported by, and is technically not part of, the multicoupler system, it requires consideration first. Antennas that feed multicouplers require sufficient bandwidth for the range of frequencies involved. If the antenna is not appropriate to the application, everything in line behind it is

wasted technology. Receive antennas should be removed far enough from all transmit antennas to prevent amplifier overload or the coupling of wideband transmitter noise into the multicoupler passband range. A minimum of 30 dB, and preferably 60 dB, of isolation is recommended between the receive antenna and any transmit antenna. In EW systems it is common to blank the receiver antennas when the transmitter is on and vice versa—to blank the transmitter output (remove power and short to ground) when the receiver chain is enabled.

For example, it is frequently necessary to enable the receiver during a jamming mission to check and see if the target being jammed is still on the identified frequency. This is a short duration look-through, and switching between the receive mode and transmit mode and back can be accomplished very rapidly [1].

21.2.2 Cabling

Receiving antenna feed lines should be routed separate from transmit feed lines. The lines should be as far away as possible from each other, on opposite mast legs, exiting the equipment building/shelter from different places, and so forth. No coaxial cable provides infinite outer conductor shielding and any mismatch of antenna and line will result in outer conductor standing waves and signal radiation.

21.2.3 Unused Ports

Any unused power divider output port should be terminated with a resistive load of value equivalent to the system characteristic impedance so that the best balance is maintained in gain levels between all receivers. Most multicoupler specifications are based on the assumption that all ports are terminated this way.

21.2.4 Expansion

Most multicouplers can be expanded to feed additional receivers. Up to a total of 64 receivers, and more, with frequencies falling within the preselector passband may be included with the addition of appropriate power dividers, the removal or replacement of existing pads, and the possible changeover of amplifier and power supply.

21.2.5 Receivers

A receiver's sensitivity is its ability to produce a specified demodulated signal output level compared to a reference modulated signal of *x*-strength. For FM narrowband systems, the 12 dB *signal to noise and distortion* (SINAD) ratio measurement method is a standard, relating recovered modulation referenced to noise and distortion. Beyond the ambient RF noise at a site, all of the elements of a

receiving system antenna, transmission line, connectors, and receiver circuits, anything that has a resistance, generate some thermal noise as well. *Noise figure* (NF) is the capability of a receiving system to detect a signal against a reference level of noise. The actual system is compared to a theoretical, noiseless receiver. NF is a combination of two ratios, the signal-to-noise power ratio of the receiving system and the noise power ratio of the theoretical system, given in decibels. The lower the noise figure, the better the receiver.

The *third-order intercept point* (TOIP) indicates the receiver's ability to linearly process a range of signal power levels without generating *intermodulation distortion* (IMD) products within the system itself. The TOIP defines a level where two signals, A and B, applied simultaneously to the receiver's input, will push the first-stage amplifier into nonlinear operation and creates an identifiable third-order (2A2B) IMD product. TOIP is expressed in decibels, referenced to 1 mW of signal power, or dBm. The higher the rating, the better the dynamic range of the amplifier. A typical performance specification for (2A2B) is 80 dB below the input levels of signals A and B.

The noise environment should be assessed before committing to a multicoupler design if possible. For tactical EW systems this is unlikely, but for fixed installations it is possible. Site noise is the difference in 12 dB SINAD sensitivity of the receiver as measured with a 50Ω dummy load and then with the antenna feed line in place. Noise figure is the ratio in decibels of the noise output divided by the noise input. A typical value for a receiver multicoupler for VHF is 6.0 dB. Multicouplers that use tower-mounted preselectors and amplifiers to overcome the loss in effective noise figure caused by long transmission lines are readily available.

Voltage standing-wave ratio (VSWR) represents the impedance match of the input and output of the multicoupler to 50Ω. A typical VSWR for a receiver multicoupler is 1.5 for both input and output.

21.3 Preselectors

Transmitter combiners and receiver multicouplers are basically filter configurations, and the preselector is a filter that prevents overloading of the receiver system by strong signals. The preselector shapes the passband and reject band responses to signals reaching the input of the amplifier stages of the receiving system. The type of preselector used, suboctave, bandpass, or band-reject, depends on the band of operation desired, the site requirements and the amount of antenna isolation that needs to be provided. Preselector passbands from as narrow as 0.5 MHz to more than 20 MHz may be used, depending on specific site and equipment requirements.

Where sufficient antenna isolation is available, a bandpass preselector can provide a 5 MHz passband and 3 dB to 4 dB per Megahertz of added rejection

above and below the desired receiving frequency range. The preselector should pass the target range of frequencies of the system's receivers with a flat response and low insertion loss. Simultaneous rejection of all other frequencies is desirable. Preselectors can be single or dual-cavity resonator combinations or multistage inductive, aperture-coupled, or combination filters. Preselectors should be frequency-stable over a range of environmental conditions, including temperature, humidity, and vibration. For EW systems, usually the preselectors are tunable since EW systems typically have extended frequency coverage requirements. Preselectors are used primarily as image reject filters ahead of the first mixer (or other nonlinear module) in a receiving chain.

The preselector prevents out-of-band high power signals from driving the multicoupler amplifier into its nonlinear region. If the amplifier has a nonlinear transfer characteristic, IMD results and the signals subsequently processed have unwanted spurs. They can be tuned bandpass filters or switch selected suboctave filters. Switches for routing RF signals are discussed in Appendix B.

Preselector bandwidth is measured at the frequencies where the attenuation is 3 dB. A typical preselector bandwidth for VHF highband (100–300 MHz) is 1 MHz. A typical value for 450 MHz is 6 MHz.

21.4 Amplifiers

As previously mentioned, an RF amplifier is often used ahead of the splitter to reduce losses incurred in the splitting process. This amplifier needs to cover the whole frequency band with a flat response.

We provide here a brief discussion of the amplifier parameters that are important for examination of their impacts on multicouplers. A more complete presentation on RF amplifiers is provided in Appendix A.

21.4.1 Amplifier Types

The taxonomy of amplifiers can be broken into three categories:

- *Small signal amplifiers*: These amplifiers typically have low output power and linearity but good low noise performance.
- *Medium power amplifiers*: These are amplifiers that can exhibit high linearity, while the noise figure may be somewhat higher than the small signal.
- *High power amplifiers*: In EW systems, these amplifiers are used at the system output to generate the power necessary to emit into target receivers. They are characterized by good linearity, while the NF is not much of a concern since typically noise is the modulation signal on the

jamming carrier. Noise at the input to the amplifier simply adds to the noise intentionally modulated onto the carrier.

Generally, the market has considered that an amplifier with a TOIP of +35–39 dBm and NF of 5 dB to be acceptable at HF since the external noise sources dominate those generated within the system itself. At higher frequencies the NF must be lower than this because the internal noise sources dominate the external noise sources.

21.4.2 Amplifier Noise Figure

Essentially any loss ahead of a receiver directly adds to the system NF. For example, a receiver with a 5 dB noise figure which is preceded by a four-way splitter with say 7 dB loss will have a system noise figure of 12 dB (neglecting cable losses, which add to this number). This is highlighted in the cascaded NF Friis formula

$$f = f_1 + \frac{f_2 - 1}{G_1} + \frac{f_3 - 1}{G_1 G_2} + \frac{f_4 - 1}{G_1 G_2 G_3} + \cdots + \frac{f_n - 1}{G_1 G_2 G_3 \cdots G_{n-1}} \tag{21.1}$$

where f_n is the noise factor (NF = $10\log_{10}f_n$) for the nth stage, and G_n is the gain of the nth stage.

At low frequencies less than 1MHz, this may not be a problem due to external atmospheric noise prevalent at these frequencies. As we go higher in frequency the background atmospheric noise drops off, and by the time we reach 30 MHz the system would be impacte; we really need the system NF to be less than 4–5 dB. As we move higher in frequency, system NF is all important and figures of 1–2 dB (or less) are required for any high-performance receiving system.

21.4.3 Dynamic Range.

When using a broadband amplifier, its overall linearity, the ability to amplify and not distort, is paramount. In a high-linearity amplifier this means that multiple signals are amplified at the same time with little distortion. Any non-linearity present would result in “mixing” of the multiple signals creating IMD. The presence of IMD in a wideband system means that when trying to receive low level signals, intermodulation products from large signals may fall on wanted low level signals, making them undetectable. Dynamic range refers to the lowest signal that may be detected at the output of the amplifier compared to the highest. The dynamic range is related to the NF and the amplifier’s ability to amplify without causing any IMD products.

A high dynamic range amplifier therefore requires good low noise performance and high linearity. In amplifier design, due to the geometries

involved with the raw devices (BJTs, FETS, and so forth) low noise figure and high TOIP are in opposition. The figure of merit for the dynamic range is referred to as the *spur-free dynamic range* (SFDR) and is the total span of the input powers that produce no IMD.

21.4.4 VHF/UHF Amplifiers

The RF amplifier feeds a series of power dividers, and it compensates for losses in that division process. Standard EW system amplifiers for the 30–3,000 MHz and higher have gains of from 15 to 26 dB, noise figures from 1.8 to 3.5 dB, TOIP ratings of up to +38 dBm and 1dB compression points of up to +26 dBm. SFDR ratings for amplifiers can be as much as 135 dBm. Most amplifiers are optimized for the best NF and linearity at a fixed gain.

21.4.5 HF Amplifiers

HF, which generally is considered to cover 1–30 MHz, is often crowded with multiple users, from military to commercial and radio amateurs. Some users have very high power, while others are relatively low. The bands themselves tend to vary in propagation throughout the day, the season and rotate on an 11 year sunspot cycle. The signal levels can range from fractions of a microvolt per meter to tens of millivolts per meter. Because of this, receiving systems at HF require amplifiers that are wideband, have reasonably low NF (<5 dB), and have good linearity resulting in high SFDR.

HF multicouplers tend to be eight-way or more. As we have seen an 8-way has approximately 10 dB loss. To overcome this loss, the preceding amplifier has to have a gain of 10 dB or slightly more. It is also important to note that the TOIP of the amplifier is reduced decibel for decibel by any loss after it. Therefore, to have a 35 dBm TOIP multicoupler (eight-way) we need the amplifier to have a TOIP of more than 45 dBm, while maintaining a good NF.

21.5 Power Dividers

A broadband power divider divides the signal from the amplifier into discrete, matched-impedance feeds while isolating the output ports from each other. Each output port feeds a receiver (or a 50Ω load, if the output port is not being used). To feed *N* receivers from a common antenna, the input is divided *N* ways. Impedances must be matched at all ports. Isolation needs to be maintained between the receivers to avoid any IMD interference generated in any one receiver from reaching the other receivers. Signal power dividers, or "splitters," based on hybrid coupler principles, provide both the coupling and the isolation simultaneously. A

signal at receiver frequency A will be attenuated by a certain number of decibels at the port for receiver frequency B. A typical specification is 25 dB or more of isolation between all receivers fed by the common multicoupler.

Any unused ports can be terminated in a resistive load, such as a low-power (1/4 W) 50Ω termination. This maintains impedance match and balance throughout the divider system and prevents signal leakage from the open port. Terminations are usually BNC or Type N. As mentioned, each two-way split creates 3.1 dB to 3.3 dB of loss, plus small conducted losses of 0.1 dB to 0.2 dB.

Various combinations of splitters can be used depending on the number of receivers to be fed and the system requirements in terms of cabling runs and equipment positions. Multiple cable runs and cabling costs should be minimized, as this also reduces undesirable cable signal and noise coupling. Again, the number of outputs from the power divider, as well as the frequency range, affects the amount of system gain realized.

21.5.1 Binary Power Dividers—RF Frequencies

A binary power divider at RF frequencies (up to 2 GHz) is an internally terminated 180^{o} hybrid. Figure 21.4 shows the standard diagram for a 180^{o} hybrid with a termination at Port A.

Physically, the two-way power divider is a three-terminal device, since the Z_0 termination at Port A is normally mounted inside the package. Also, although a conventional 180^{o} hybrid can be used as a power divider, the usual form of a two-way power divider does not have a Z_0 impedance level at all four ports. Higher-order binary power dividers, such as four-way or eight-way power dividers, are realized by cascading 2-way power dividers of various configurations. The functional diagram for a four-way device is shown in Figure 21.5, while the eight-way diagram would simply have the B port of additional two-way dividers connected at ports 1, 2, 3, and 4.

21.5.2 Binary Power Dividers—Microwave Frequencies

At microwave frequencies (500 MHz and higher), the binary divider is typically realized as a strip-line or microstrip Wilkinson power divider. A Wilkinson power divider consists of a series of cascaded quarter-wavelength transformers that transform the impedance of a single input, which is typically 50Ω, to the impedance associated with the parallel combination of the multiple outputs. The input VSWR of this divider is primarily influenced by the quality of these impedance transformers and the VSWRs of the loads that terminate the outputs of the device. To facilitate low output VSWRs and provide isolation between these outputs, resistors are placed at the ends of each of the quarter-wavelength transformers. The simplest Wilkinson divider is the single-section two-way. This

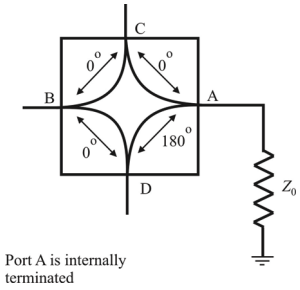


Figure 21.4 Two-way power divider. Signal input is at B while C and D are the two outputs.

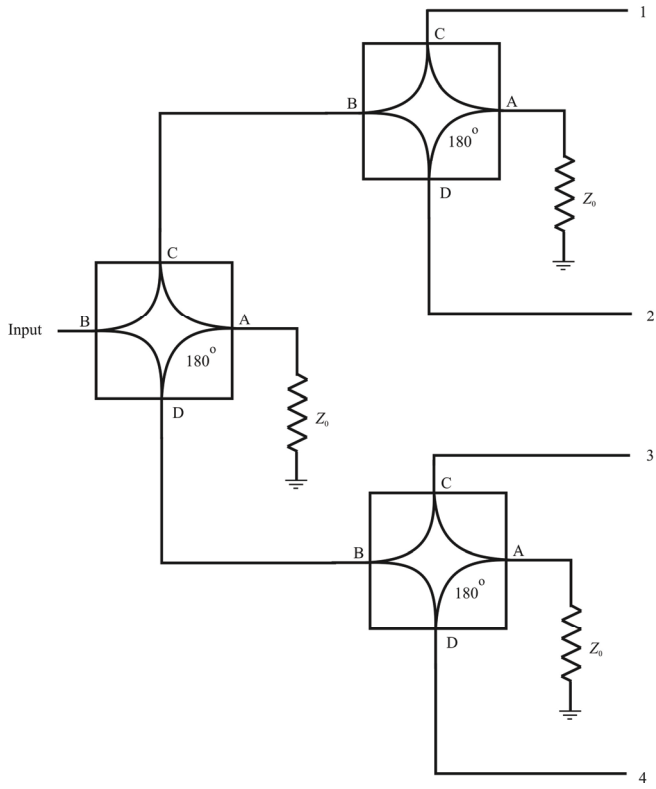


Figure 21.5 Four-way power divider.

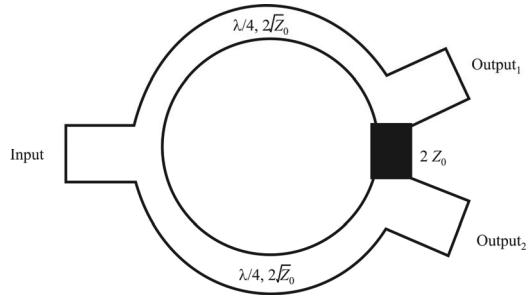


Figure 21.6 Classical Wilkinson two-way power divider for the microwave region.

device consists of two 70.7 $(=\sqrt{2}\times 50)\,\Omega$ quarter-wavelength transformers with a 100Ω resistor between them (see Figure 21.6).

The simplest Wilkinson divider is useful for limited bandwidth applications. To increase the useful bandwidth, additional quarter-wave transformers and isolation resistors are added. The general rule is that the greater the bandwidth, the greater the number of required transformers. Consequently, the unit becomes larger with higher associated loss.

Practical isolated power dividers are available with octave bandwidth and also extremely wide bandwidths of decades and greater. Octave band units are of the Wilkinson type, using discrete lumped resistors. VSWR increases outside of the band specified. The broadband units are of a tapered design, which provides semi-infinite bandwidth in principle. That is, there is no upper frequency limit to operation; however, in practice there is an actual upper limit due to increasing VSWR and loss.

The Wilkinson and tapered line dividers are cascaded to create higher order devices when such dividers are required by the application. These devices are typically realized as printed structures on low-loss Teflon-based laminates with chip resistors.

21.5.3 Performance Parameters

The critical parameters of a power divider are normally its frequency range, insertion loss, isolation (port-to-port), and VSWR.

21.5.3.1 Isolation

Probably the most important parameter for power dividers is the port isolation. The output ports of a divider (or the input ports if it is being used as a combiner) are isolated from each other. When a typical divider is splitting a signal or combining coherent signals, the voltage present on each side of the isolation

resistor (see Figure 21.6) is of equal potential and therefore no current flows through the resistor and no power is dissipated. However, when combining non-identical signals, the residual signal that does not appear at the input travels toward the other outputs and reaches the opposite side of the isolation resistor 180^o out of phase with the incident signal. This voltage differential causes a current flow through the resistor and power is dissipated. The signal is attenuated significantly, achieving isolation.

A two-way device used to combine two non-coherent signals will deliver half of each signal of the common port and half of each signal will be dissipated in the isolation resistor. Consequently, a four-way device combining four noncoherent signals will deliver one-fourth of each signal at the inputs and dissipate the rest.

21.5.3.2 Insertion Loss

The insertion loss is the loss experienced through the device, excluding the power division factor (3, 6, 9 dB, and so forth). It is the ratio of power output to power input, expressed in decibels and assumes that matched source and load are used for terminations.

Microwave frequency dividers have line losses that increase with frequency. The values specified have a minimum value at the lowest frequency and linearly increase at a rate determined by the power divider length. The RF dividers tend to exhibit a similar characteristic except the loss tends to increase at the lowest frequencies approaching the low-frequency cutoff of the device.

21.5.3.3 Output Amplitude Tracking

The difference in the signal amplitudes at the output ports is called output amplitude tracking error or output amplitude imbalance. It is the ratio of the maximum signal at any port to the minimum signal of any other port, expressed in decibels. Typically, the maximum output tracking specified is very low, for example, 0.3 dB, for two-way dividers, and increases with a higher number of output ports, to 1.0 dB for the eight-way power divider.

21.5.3.4 Output Phase Tracking

The differences in the signal phase at the output ports is called the output phase tracking error or unbalance. It is usually specified as the maximum deviation; typically the average phase is much less, especially at low frequencies.

21.5.3.5 Matched Power Rating or Input Power

This is the maximum power that can be applied to the input and still maintain other performance specifications. It is specified for RF components with Z_0

terminations on all outputs to avoid reflected signals from unbalanced loads which may exceed the limit for power dissipation in the internal termination. The microwave frequency components are specified at a maximum input power level assuming a 2:1 load VSWR on each output port. This VSWR yields a condition where ~10% of the input power is reflected and must be dissipated by the module. If load VSWRs are greater than 2:1, the module must be derated.

21.5.3.6 Internal Load Dissipation

This is simply the power rating of any one of the internal terminations. These two power parameters are related and the input power rating is normally several times larger than the internal load dissipation. The reason for this is because most of the input power is delivered to the output loads, not the internal termination.

There are three considerations affecting the amount of input power that a power divider can withstand.

- *Insertion loss*. The first is the total power dissipated in the power divider. Total power dissipation in a power divider under matched conditions can be determined to a reasonable approximation from the insertion loss and the known input power with:

$$\text{Insertion loss (dB)} = 10\log_{10} P_{\text{in}} P_{\text{out}}$$

Therefore,

$$P_{\text{out}} = \frac{P_{\text{in}}}{10^{\text{InsertionLoss}_{\text{dB}}/10}}$$

$$\begin{aligned} P_{\text{dissipated}} &= P_{\text{in}} - P_{\text{out}} \\ &= P_{\text{in}} - \frac{P_{\text{in}}}{10^{\text{Insertion Loss}_{\text{dB}}/10}} \end{aligned}$$

Example: Consider a power divider with the following conditions

Insertion loss = 0.5 dB
$P_{\text{in}} = 2$ W

then

$$\begin{aligned} P_{\text{dissipated}} &= 2 - \frac{2}{10^{0.05}} \\ &= 0.218\,\text{W} \end{aligned}$$

Most of the power will be dissipated in the wire and ferrite cores making up the transformer circuits and not in the internal load.

- *Amplitude balance*. In a power divider operating under matched conditions, a second consideration for input power dissipation is the dissipated power in the internal load. If we consider a two-way power divider similar to that shown in Figure 21.4, we can observe that ideally no power would be dissipated in the $2Z_0$ load between ports C and D, because the voltages at C and D would be equal. In practice, a small differential may occur because of imperfect amplitude balance. The approximate dissipation due to this unbalance can be shown to be

$$P_{\text{intLoad}} \approx P_{\text{out}}^{10\,\text{AmpBal}/10-1}$$

A very small amount of power normally is dissipated due to this effect.

Example: Let

AmpBal = 0.25 dB
P_{out} = 2 W

then

$$P_{\text{intLoad}} \approx 1/2^{10(0.25/20)-1}$$
$$= 0.4\,\text{mW}$$

- *Mismatched loads*. When determining input power limits the third and perhaps most important case to consider is the condition of mismatched loads at the outputs of the power divider. Reflections from these mismatches can cause a considerably larger voltage differential to appear across the internal load. If the VSWR of the two loads are S_1 and S_2, the limit on the input power P_{in}, is given as

$$P_{\text{in}} \le \frac{\text{InternalLoadRating}_{\text{W}}}{\dfrac{(S_1-1)^2}{S_1+1}+\dfrac{(S_2-1)^2}{S_2+1}}$$

Example: If the internal load rating is 0.5 W and the VSWR of S_1 and S_2 are 2:1, then

$$P_{\text{in}} \leq \frac{0.5}{\dfrac{(2-1)^2}{2+1} + \dfrac{(2-1)^2}{2+1}} = 2.25\,\text{W}$$

This is the formula for the worst case sitiuation, which assumes that the two load reflections are out of phase at the output port. If they are identical impedances, P_{in} may be several times larger than this without damage.

From the preceding discussion of power divider input power ratings, we can conclude that:

- Under the matched loading conditions (Z_0 terminations at all ports) the input power is limited by heating effects. This is especially true for the RF components that use wire wound ferrite cores as transformers. Absolute maximum temperatures for ferrite core transformers are limited to the 130°C to 500°C range. As with all good engineering practices, it is advisable to stay well below the temperature limits (20°C or more) to avoid performance degradation, particularly increased insertion loss. The actual temperature rise in the ferrite core is dependent upon the heat transfer path from the core to the heat sink or surrounding air. Heat dissipation depends on the thermal resistance, which, in turn, is highly dependent on the mounting of the power divider as well as its internal construction.

- Under conditions where mismatches are present at the power divider output, the internal load power dissipating rating may limit the input power that can be applied. A simple worst-case calculation can be performed to determine if this is the case using the provided formulas.

A final point relative to power ratings that should be considered is the application as a power combiner. In this case equal signals are normally applied to the "input" ports and little if any power is dissipated in the internal load. A condition may occur where one or more of the signal sources fails or is removed. For example, if two equal sources are applied at ports C and D and the source at D fails, 50% of the power supplied by the source at C will be dissipated in the $2Z_0$ internal load. Thus, to facilitate this possibility, the power injected at each port should not exceed twice the rating of the internal load.

21.5.3.7 Summary

Despite the functional simplicity of power dividers and generally rugged and reliable components used in the construction, their specification and application in systems can still lead to unexpected problems.

In this section we have presented some basic insights into the internal construction for power dividers, how these influence the operation of the device, and its function when imperfect matches and less than ideal physical installation occur. On this latter point, the consideration for power dividers was dealt with both analytically and in terms of expected results in the normally mounted configuration.

21.5.4 Splitter Types

Splitter taxonomy can be categorized into three types:

- *Printed hybrid*: UHF and above tends to narrow bandwidths (less than 20%). Wider bandwidths are possible, but size increases and output-to-output isolation is reduced.
- *Transformer hybrid*: The use of baluns and impedance matching transformers, often ferrite loaded cores, can cover many octaves, but are generally used below 1 GHz.
- *Resistive splitter*: The use of resistors to both split and impedance match are really not generally suited to the use in splitters for EW receivers as they offer high loss and very little isolation port to port. They also contribute to a higher noise figure in the receiver chain due to the addition of thermal noise.

For high performance LF/HF and low VHF/UHF the best choice is normally a ferrite loaded transformer**.**

21.5.5 Passive/Active

In the selection process of a receiver multicoupler the question of whether or not a passive unit is sufficient, or if an active coupler is required. In most cases it is relatively easy and straightforward to determine the best choice for a given monitoring site, as it is, generally, a question of RF signal strength (or anticipated signal strength) of the targets. For example, if strong HF broadcast stations are primarily monitored, such as BBC, VOA, and other international radio stations with substantial signal strength, a passive multicoupler will, in most cases, perform adequately. Specifically if a large antenna system is used, the passive coupler may be a good choice. If, on the other hand, weaker signals with varying

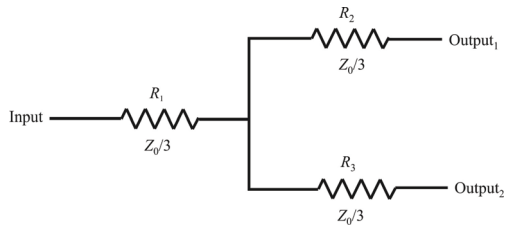


Figure 21.7 Two-way resistive divider.

strengths, such as those that may be received from tactical radio networks, is the primary goal of the EW effort, an active receiver multicoupler is the better choice.

On VHF and UHF frequencies, in most cases the active multicoupler is the better choice. Here the signal strengths of received signals are, generally, lower than what would be encountered on the HF frequencies. Also, increased losses in the antenna feed line as the frequency is increased produces lower signal levels at the receiver—specifically, from about 500 MHz and up. Depending on the coaxial cable used, these losses may be substantial. Losses ahead of a receiver's front end (such as in the coax) directly adds to the system noise figure; therefore, selection of an active multicoupler will not add to those losses as the passive unit would.

21.5.6 Resistive Power Dividers

Resistive power splitters have inherent characteristics that make them an excellent choice for certain applications but unsuitable for others. Figure 21.7 shows the layout of a simple two-way resistive power divider. Z_0 denotes the characteristic impedance of the system (50Ω, 75Ω, 300Ω, or some other value).

Some advantages and disadvantages of the resistive divider are:

Advantages:

- Generally smaller than reactive dividers because they are made utilizing lumped elements;
- Can be extremely broadband and are the only type of dividers that operate from DC.

Disadvantages:

- Poor power handling capability, usually limited by the power ratings of the resistors;
- These units tend to have high loss and poor isolation, making them a poor choice for applications where insertion loss and isolation are critical factors.

The resistor value for an *N*-way resistive divider is given by

$$R = Z_0 \frac{N-1}{N+1} \tag{21.2}$$

where Z_0 is the characteristic impedance of the system. The power transfer ratio is

$$\frac{P_o}{P_{in}} = \frac{1}{N^2} \tag{21.3}$$

where *N* is the number of outputs. On the other hand, a reactive power divider has a power ratio of 1/*N*. This means for a two-way divider with 1 W input power, a reactive unit's output port will have 0.5 W (50% or 3 dB loss) of power on each port, while a resistive divider's output will have 0.25 W (25% or 6 dB loss) of power on each port. The resistive divider's loss includes distribution loss and inherent resistive loss, unlike the "lossless" reactive dividers (which only have distribution loss). So as a general rule, insertion loss of a resistive power divider is *double* that of a similar reactive unit. In addition, in a resistive power divider the insertion loss is equal to the input/output isolation.

21.5.7 Reactive Power Dividers

Reactive power dividers come in various forms and cover a wide range of frequencies. They can have multiple output ports, but those with odd number of output ports are usually referred to as *N*-way power dividers. They can be realized using waveguide, stripline, microstrip, transformer, and various other topologies. These dividers come in in-phase, 180° out of phase (180° hybrids), 90° out of phase (quadrature hybrids), and other specialized configurations. As mentioned, one of the main advantages of using reactive power dividers is that they are "lossless."

Figure 21.8 shows the general layout of a two-way reactive divider. Higher order devices (four-, six-, eight-way) are usually realized by cascading the two-way in various configurations. Figure 21.9 shows the layout of a *N*-way power divider (a three-way is shown). Note that all ports are connected in a manner that makes them mutually isolated from one another.

21.5.8 Wilkinson Power Divider

21.5.8.1 Higher Frequencies—Microstrip Wilkinson Power Dividers

The Wilkinson power divider splits an input signal into two equal phase output signals [2]. It may also work reversed as a combiner, in which case it is used to

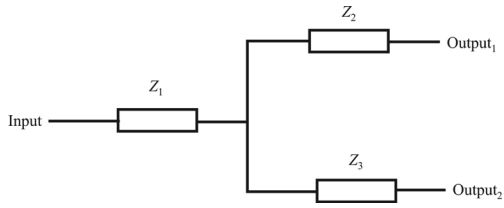


Figure 21.8 Two-way reactive divider.

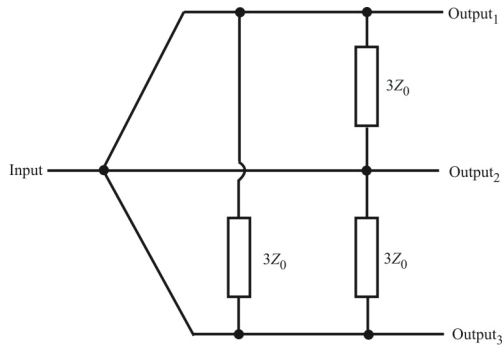


Figure 21.9 Three-way power divider.

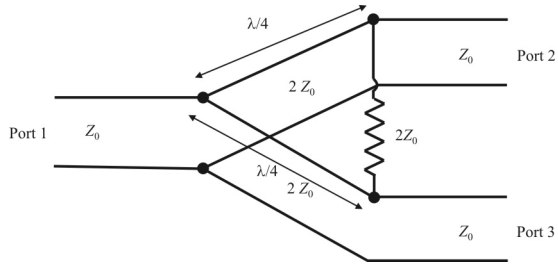


Figure 21.10 Wilkinson equivalent circuit.

combine two equal-phase signals into a single signal in the opposite direction. This divider was invented around 1960 by Ernest Wilkinson. The Wilkinson power divider is a solution of the lossless T-junction problems, where the outputs are not matched and nonisolated. It is especially useful because the output ports are simultaneously isolated and matched.

The Wilkinson power divider is a lossy three-port network and it has a property of being lossless when the output ports are matched. Wilkinson relied on quarter-wave transformers to match the split ports to the common port. Since a lossless three-port reciprocal network cannot have all three ports matched,[1] Wilkinson added one resistor. The resistor fully isolates port 2 from port 3 at the center frequency without adding resistive loss to the power split, so an ideal Wilkinson splitter is 100% efficient.

A typical Wilkinson power divider looks like the one shown in Figure 21.6 with equivalent circuit as seen in Figure 21.10. Depending upon the power split, Wilkinson power splitters can be categorized into:

- Wilkinson equal divider;
- Wilkinson unequal divider.

The latter of these can be analyzed as a simple extension of the analysis of the former, so we will focus on the first.

The equal power division concept divides the input power into two or more output paths. The most commonly used is shown in Figure 21.2 and is also called a 3 dB power divider. In this type of divider, there are four different sections:

- Input port;
- Quarter-wave transformer;
- Isolation resistors;
- Output ports.

[1] This can be mathematically proven.

The input and output ports are identical and their impedance values are Z_0. Quarter-wave transformer ports are so called because that's their length (λ/4 sections in Figure 21.10). The length of these ports is equal to the one-fourth of the wavelength of the EM, which is propagating in this three-port network. This results in matching of the output ports which provides better power transfer results. The isolation resistor isolates the output ports. If there are coupling effects between output ports, the perfect division of power cannot be possible. This isolation resistor avoids the coupling effects of the output ports.

When a signal enters port 1, it splits into equal-amplitude, equal-phase output signals at ports 2 and 3. Since each end of the isolation resistor between ports 2 and 3 is at the same potential, no current flows through it and therefore the resistor is decoupled from the input. The two output port terminations will add in parallel at the input, so they must be transformed to $2Z_0$ each at the input port to combine to Z_0. The quarter-wave transformers in each leg accomplish this; without the quarter-wave transformers, the combined impedance of the two outputs at port 1 would be $Z_0/2$. The characteristic impedance of the quarter-wave lines must be equal to 1.414 Z_0 so that the input is matched when ports 2 and 3 are terminated in Z_0.

21.5.8.2 Lumped Element Wilkinson Dividers

At higher frequencies (above 500 MHz) the Wilkinson power divider is popular and widely used [3]. These devices are usually realized as a microstrip design, as described above. However, lumped element designs are also possible, as we will show. At lower frequencies such configurations would be prohibitively large.

These power splitters generally employ quarter-wave transmission line sections at the design center frequency, which can have unrealistic dimensions at frequencies in the RF and low microwave bands, where the wavelength is large. For example, a λ/4 microstrip line with characteristic impedance $Z_0 = 70.7\Omega$ on FR-4 substrate (dielectric constant $\varepsilon_r = 4.3$, thickness $h = 1.0$ mm) is approximately 43 mm long at a frequency of 1 GHz. In some cases, it would be preferable to use lumped-element equivalent networks replacing the λ/4 transmission lines. It is possible to employ *surface mount devices* (SMD), as well as *monolithic microwave integrated circuit* (MMIC) lumped elements, which allow saving circuit area.

λ/4 transmission line segments have equivalent Tee and π lumped-element networks. The same is valid for a 3λ/4 line segment. In particular, a quarter-wave line at a frequency f_0, with characteristic impedance Z_0, can be replaced with a π LC equivalent network as shown in Figure 21.11. The element values are given by

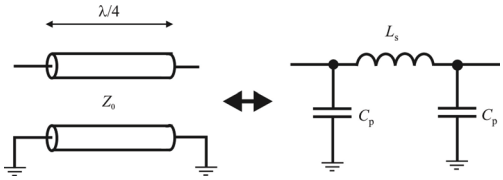


Figure 21.11 Lumped-element π equivalent of a λ/4 section of transmission line.

$$C_p = \frac{1}{2\pi f_0 Z_0} \tag{21.4}$$

$$L_s = \frac{Z_0}{2\pi f_0} \tag{21.5}$$

The π LC network is perfectly equivalent to the line section only at the center frequency f_0, but the approximation is still valid for modest bandwidths.

Figure 21.6 shows the layout of a classical microstrip Wilkinson power splitter. In the simplest form, it consists of two quarter-wave line segments at the center frequency f_0 with characteristic impedance $\sqrt{2}Z_0$, and a $2Z_0$ lumped resistor connected between the output ports. It provides low loss, equal split (ideally 3 dB), matching at all ports, and high isolation between output ports.

By replacing both λ/4 line sections by equivalent π LC networks, it is possible to obtain a lumped-element version of the Wilkinson divider, as shown in Figure 21.12. As noted above, this network is equivalent to the original only at the center frequency f_0. Consequently, the expected performance (insertion loss, return loss,

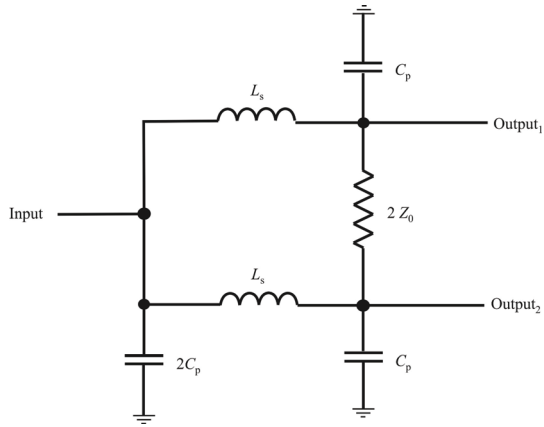


Figure 21.12 Two-way lumped-element Wilkinson power divider.

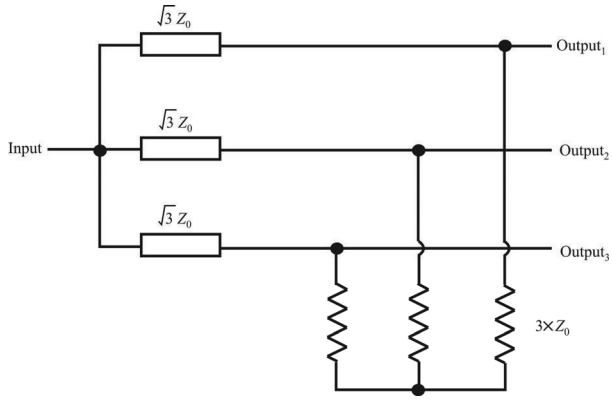


Figure 21.13 Three-way lumped Wilkinson power divider.

isolation) should be similar to that exhibited by the distributed-form power divider for a narrow bandwidth centered in f_0, wide enough for many applications.

Moreover, the π LC equivalent networks exhibit a low-pass behavior, rejecting high frequencies, while the response of the classical Wilkinson divider repeats at odd multiples of center frequency ($3f_0$ and $5f_0$, mainly). This behavior could be desirable if harmonic filtering is needed.

At 1,080 MHz, using (21.4) and (21.5) we obtain C_p = 1.8 pF and L_s = 10 nH. At port 1 we choose a 3.9 pF capacitor, and the balancing resistor is 100Ω. Standard low-cost 0805 SMD components are used, featuring 5% tolerance. The insertion loss at 1.08 GHz is about 3.6 dB, return losses result 14 dB at port 1, 16 dB at ports 2 and 3, and isolation between output ports is 20 dB. These are typical values also attainable with a microstrip power divider. However, a 1 GHz microstrip Wilkinson splitter would occupy about 6 cm^2 on FR-4, while this lumped-element version occupies less than 1 cm^2.

Three-Way Wilkinson Power Splitter

The Wilkinson divider can be generalized to an *N*-way power splitter/combiner. For example, the diagram corresponding to a three-way divider is shown in Figure 21.13. As can be seen, it requires crossovers for the balancing resistors. This makes fabrication difficult in planar form (e.g., microstrip).

For an 850 MHz splitter, we obtain C_p = 1.5 pF and L_s = 15 nH. At port 1, we choose a 4.7 pF capacitor, and the three balancing resistors are 51Ω. The shunt capacitor C_0 is used to tune out the resistor and pad parasitic to avoid performance degradation (mainly in terms of isolation between ports). Its value is determined experimentally, varying typically in the 0.5 to 2 pF range for these frequencies. Return losses were measured to be better than 12 dB at port 1 at center frequency (around 15 dB at output ports). Measured split losses from port 1 to all three

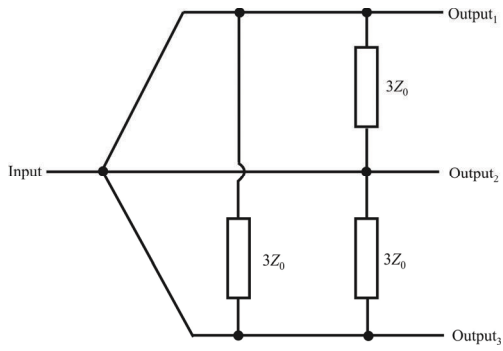


Figure 21.14 Three-way power divider.

output ports are only about 0.7 to 1.2 dB higher than in the ideal case (4.77 dB). Excellent isolation characteristics between the output ports, exceeding 25 dB, are achieved by adjusting the value of capacitor C_0.

N-Way Dividers

Power dividers having an odd number of outputs (3, 5, 7, and so forth) are sometimes called *N*-way dividers. A circuit actually used to realize a true three-way divider, as opposed to a terminated four-way power divider is shown in Figure 21.14. The transformer circuit produces three mutually isolated outputs. Three internal terminations of value $3Z_0$ must be connected between ports 1 and 2, 1 and 3, and 2 and 3 in order to maintain port match and port-to-port isolation.

21.5.9 Remote Receive Multicoupler

Mast-mounted multicoupler systems can improve receiver performance from 4 dB to as much as 9 dB in effective sensitivity. In general, the longer the feed line, the more dramatic the improvement in performance. Several factors contribute to this improvement:

- The preservation of effective noise figure that is otherwise reduced dB for decibels by feed-line loss.
- As the amplification is all in the tower-top amplifier, the level of signals conducted down the transmission line is from 16 to 26 dB higher than any signal or noise responses coupled to the line.
- More accurate overall matching is maintained between the antenna and first stage amplifier, thus establishing the best noise figure for the entire system.

In most cases, specifications of importance are port-to-port isolation and the coupler's signal loss/gain characteristics, and in the active multicoupler, front-end dynamic specs, such as P1dB and IP3, are of value to know. However, the most important specification of a receiver multicoupler (passive or active) is the port-to-port isolation. Isolation is measured in decibels and is an indication of how well an output is (signal wise) separated (isolated) from the other output(s). The issue of isolation is very important when several receivers are connected to the multicoupler. Most radio receivers will emit RF signals back out (reverse) to the antenna. These signals are, unfortunately, a natural circumstance from local oscillators, mixers, and PLL synthesizers. Depending on receiver circuit design these outgoing signals may be substantial—being strong enough in some cases to be detected by other nearby receivers. Therefore, when using several receivers in an EW system using a multicoupler, good isolation between ports is important. It ensures that each individual receiver is operating without interference from other receivers using the coupler. This isolation should be a minimum of 20 dB; more is better.

21.6 Transmit Multicouplers (Combiners)

Modern transmitters (including jammers) in the VHF and UHF range (30 MHz–3 GHz) generally employ power splitters and combiners (also called diplexers) in conjunction with a number of PA modules to increase small signals to ones powerful enough to transmit. Often a single PA is not powerful enough in and of itself to generate the power levels necessary so the signal is split into two or more paths and delivered to the corresponding number of separate PAs. The outputs of these PAs are combined together to generate the high-power signal. From a reliability point of view, employing multiple PAs can ensure continued operation, albeit at reduced performance levels, in the event that one (or more) of the component PAs fails [4].

As discussed above, a splitter is primarily concerned with dividing the power in a single signal equally into two (or more) identical signals, sometimes with different phases, for delivery to the input to PAs. This concept is illustrated in Figure 21.15.

An ideal power combiner would have the following characteristics:

- Signals are combined without loss.
- The two PAs are isolated from each other so if one fails it does not impact the operation of the other.
- A nominal (constant), resistive load to each PA for all loads.
- A nominal load to each PA regardless of output of the other PA.

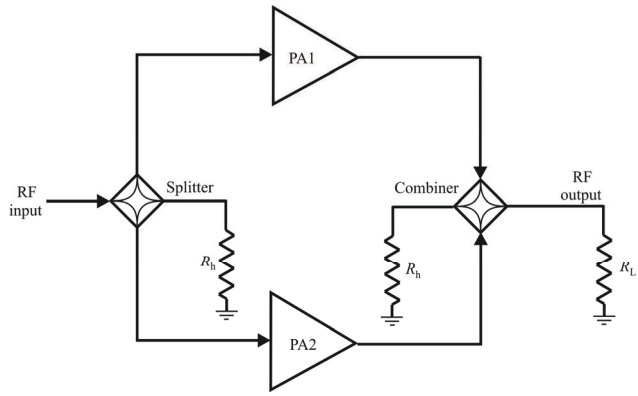


Figure 21.15 Power combining for one signal.

- Isolation of PAs from reverse-direction signals (back-IMD).

Unfortunately, an ideal power combiner does not exist. Just as in every engineering discipline, there are trade-offs that must be made among these parameters, depending on the particular situation.

The primary applications for combiners can be divided into two basic categories: combining two (or more) signals of the same frequency and amplitude, and combining two (or more) signals of different frequencies. The first application is shown in Figure 21.15, which occurs in many transmitters and allows the outputs from a number of low-power modules to be combined to produce a larger total output power. The second application, illustrated in Figure 21.16, allows simultaneous jamming of more than one target from a signal transmitter and antenna.

Combining refers to adding the outputs of two or more amplifiers together. Such ganging is fraught with design complexity, however. If the circuit components are not matched closely enough, then unbalanced amplitude and phase conditions can cause significant losses to occur. Furthermore, if one of the power amplifiers fails, in some designs the output power does not simply decrease by 50%, but can be substantially more.

Combiners, in addition to combining the signal outputs from two or more amplifiers, should provide some degree of isolation from one amplifier to another. Ideally, the characteristics of one amplifier should not influence those of another, irrespective of the circumstances. The load impedance seen by an amplifier should be constant, so the output of the amplifier can be conjugate matched to that load for maximum power transfer. This is rarely, if ever, the case, whether the output of two or more amplifiers are combined or not. Combining the outputs complicates this situation.

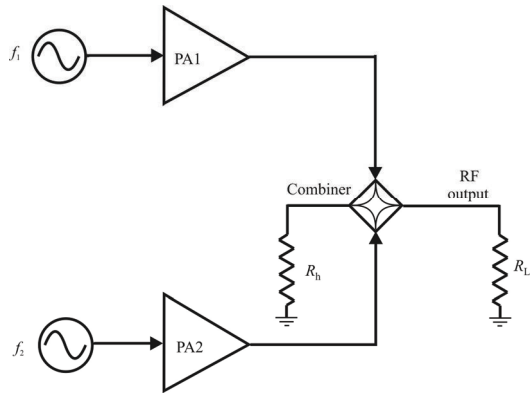


Figure 21.16 Power combining for two signals. These two signals may be at different frequencies.

The configuration shown in Figure 21.17 is perhaps the simplest way to combine the output of two amplifiers. Here, the currents from the two amplifiers are summed while the voltage output of the two amplifiers is forced to be the same. This configuration provides little isolation of the outputs, and can lead to inefficiencies in operation if the two signal paths are not identical in amplitude and phase characteristics. In addition, if one of the outputs fails as a short circuit, it will route all of the current from the other amplifier to ground and there will be no output. The current flowing through the load is $2I_{\text{out}}$, assuming that the currents from the amplifiers are equal. Since the power dissipated in the load is given by $P_{\text{load}} = I_{\text{load}}^2 R_{\text{load}}$, then the power output of this configuration is four times the power of a single amplifier configuration.

Another combining technique is shown in Figure 21.18. In this case, the voltages from the two amplifiers are added at the secondary of the transformers. Some amount of isolation is provided in this configuration.

There is a wide variety of techniques for power combining in the VHF and UHF range, including

- Direct connection combiners;
- Hybrid combiners;
- Zero phase combiners;
- Directional couplers;
- Wilkinson-hybrid (quarter-wavelength) combiners;
- Quadrature-hybrid combiners.

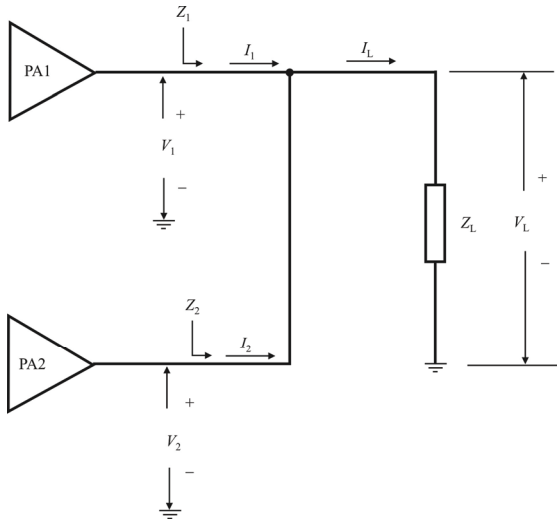


Figure 21.17 Direct-connection combiner. A simple way of combining the output of two amplifiers.

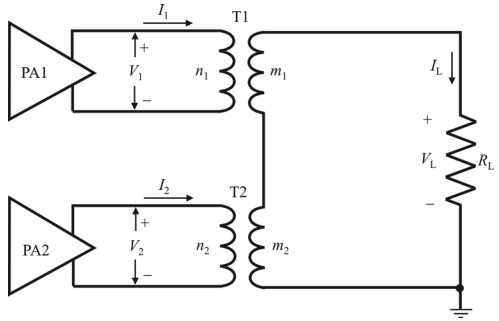


Figure 21.18 Voltage-summing combiner based upon conventional transformers. This is another way of combining the output of two amplifiers. In this case, failure of one of the amplifiers does not cause catastrophic failure of the total system.

In this section we will examine the high-level characteristics of a few of the aforementioned approaches. This discussion is not exhaustive, as with all engineering pursuits, the design of such combiners is open to the imagination of the designer.

21.6.1 PA Inefficiencies

In the typical case, every effort is made to optimize the output power in an EA transmitter. Every decibel of power gain is carefully preserved. However, power combiners can be a significant source of power loss in a transmitter. The reasons for inefficiency include:

- Resistive losses in the combiner;
- Combiner losses due to mismatch (amplitude and phase) of PA outputs;
- PA inefficiency due to non-nominal load presented to the PA by the combiner.

The resistive loses are generally small; however, the second two losses can be significant.

The characteristics of a power combiner that we will examine include its behavior with:

- Nominal load (i.e., 50Ω);
- All PAs operating;
- Class[2] A/B PA;
- Class D PA;
- One or more PA has failed.

21.6.2 Direct-Connection Combiners

A direct-connection combiner causes the output currents or output voltages of the PAs to be added. These are the most basic and simple forms of combiner possible and therefore serve as a baseline against which other types of combiners can be compared.

21.6.2.1 Current Summing Combiners

This section analyzes the behavior of a current summing combiner, which include both direct-connection, as illustrated in Figure 21.17, and transformer coupled as

[2] See Appendix A for a discussion of PA classes.

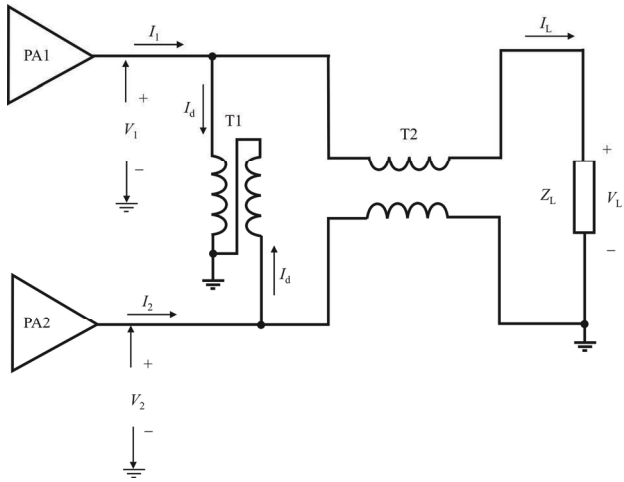


Figure 21.19 Transformer-coupled current summing combiner. T1 and T2 are transmission line transformers.

illustrated in Figure 21.19 [5] topologies.[3] We include discussions under abnormal operating conditions including the effects of one PA failing, both open and shorted. We also consider combining signals of different frequencies.

Direct Connection

Direct connection of the outputs of two PAs illustrated in Figure 21.17 is the least complex means of combining their outputs. This technique sums the current from the two PAs and therefore requires current-source PAs (e.g., operation in class A or B such as that approximated by BJTs or MOSFETs in class A) rather than voltage-source PAs (e.g., class D or saturated class C, vacuum tubes).

The load impedance presented to PA1 is

$$Z_1 = \left[\frac{I_1 + I_2}{I_1}\right] Z_L \tag{21.6}$$

The load impedance presented to PA2 has as similar form with i_1 and i_2 interchanged. Since both PA-load impedances are strongly influenced by the outputs of both PAs, it is apparent that the direct-connection coupler provides no isolation as would be expected by simple examination of Figure 21.17.

[3] In this chapter lower-case variables represent functions of time and the time variable is implied—it is not normally included. The corresponding capital letters stand for the corresponding phasor. Either of these quantities may be complex.

The power output from PA1 to the combiner is given by

$$P_1 = \frac{1}{2}\mathrm{Re}\{V_1 I_1^*\} = \frac{1}{2}\mathrm{Re}\{V_L I_1^*\} \tag{21.7}$$

The power delivered by PA2 has an analogous form. The relative contribution of each PA is determined both by the amplitude of its current output and the degree of phase mismatch between the output current and voltage. Only in theory are the two PAs exactly matched.

Transformer Coupling

The configuration shown in Figure 21.19 employs hybrid transformer T1 and balun transformer T2 (otherwise known as a transmission line transformer [6]) to combine the current outputs of the two PAs. However, grounding of the center-tap connection on T1 prevents the circuit from providing the isolation and constant impedance normally associated with a hybrid coupler.

Transformers T1 and T2 force the currents in their two respective windings to be equal in amplitude and opposite in direction. Consequently,

$$I_d = I_1 - I_L = I_2 + I_L \tag{21.8}$$

from which

$$I_1 = I_L + I_d \tag{21.9}$$

and

$$I_2 = -I_L + I_d \tag{21.10}$$

Combining (21.9) and (21.10) yields

$$I_L = \frac{1}{2}(I_1 - I_2) \tag{21.11}$$

The voltages on the opposite ends of T2 are equal, as are the voltages on opposite windings of T1. Consequently,

$$V_1 - V_2 = V_L \tag{21.12}$$

and

$$V_1 = -V_2 = \frac{1}{2}V_L \tag{21.13}$$

The impedance that is seen by PA1 is

$$Z_1 = \frac{V_1}{I_1} = \frac{V_L}{2I_1} = \frac{I_L Z_0}{2I_1} = \frac{(I_1 - I_2)Z_0}{4I_1} = \frac{1 + I_1 / I_2}{4} Z_0 \tag{21.14}$$

Thus, the transformers produce a 1:4 impedance transformation.

For both the direct connection and transformer configurations, in the nominal mode of operation, both PAs deliver identical output currents (same amplitude, same phase) to a resistive load:

$$I_L = 2I_1 \tag{21.15}$$

and

$$V_L = 2I_1 R_L \tag{21.16}$$

The power that would be delivered by one PA by itself to load Z_L is

$$P_{PA1} = \frac{I_L^2 R_L}{2} \tag{21.17}$$

However, the power delivered by both PAs is

$$P_o = \frac{(2I_1)^2 R_L}{2} = 4P_{PA1} \tag{21.18}$$

Each PA produces half of this total.

Abnormal Operation

PA Failure—Open. Suppose that PA2 fails and becomes an open circuit. Then

$$I_L = I_1 \tag{21.19}$$

$$V_1 = V_L = I_1 Z_L \tag{21.20}$$

and

$$Z_1 = Z_L \tag{21.21}$$

The power output is reduced by a factor of four, from P_o of the nominal system (21.18) to P_{PA1} (21.17) of a single PA driving load R_L. The output impedance is reduced by a factor of two from that of the nominal configuration. In addition, the power delivered by PA1 to the load is reduced by a factor of 2 from that in the nominal operating mode.

The coupler itself (theoretically) dissipates no power and therefore has no associated efficiency. However, it causes the efficiencies of conventional class-A and class-B PAs to be reduced by half when one PA fails open. The additional power is dissipated directly in the PA that remains active.

PA Failure—Short. If PA2 forms a short circuit as it fails, then

$$V_1 = -V_2 = V_L = 0 \tag{21.22}$$

and the output of power from PA1 is precluded. The direct-connection coupler therefore provides no protection against short-circuit failure on one of the PAs.

Signals of Different Frequencies

When the signals have different frequencies, that is,

$$I_1(t) = \sin \omega_1 t \tag{21.23}$$

and

$$I_2(t) = \sin \omega_2 t \tag{21.24}$$

if the load is fixed at $R_L = 1$, then each sinusoid produces a power of 0.5 W for a total output of $P_{oAVG} = 1$ W. Under the same assumptions, $I_{Lmax} = 2$A, hence $V_{Lmax} = 2$ V, hence $V_{DD} = 2$ V is required for linearity. By analogy, $V_{DD} = n$ V is required for the simultaneous production of n carriers of different frequencies and unity amplitude.

21.6.2.2 Voltage Summing Combiners

We examine *voltage-summing combiners* (VSCs) in the section, which include configurations based upon both conventional transformers shown in Figure 21.18, and transmission line transformers illustrated in Figure 21.20.

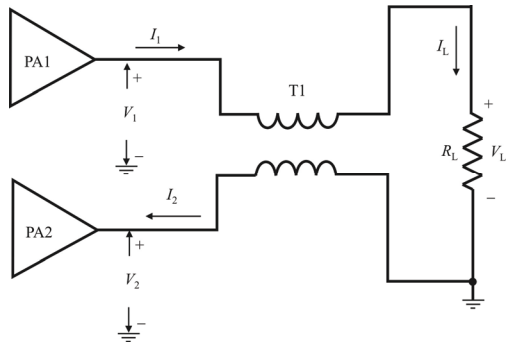


Figure 21.20 Voltage-summing combiner based on a balun.

Conventional Transformers

In the configuration shown in Figure 21.18, transformers T1 and T2 force

$$I_1 = I_2 = I_L = \frac{V_L}{R_L} \tag{21.25}$$

and

$$V_L = \frac{m_1}{n_1} V_1 + \frac{m_2}{n_2} V_2 \tag{21.26}$$

It is apparent that the PAs must act as voltage sources, allowing the output current to be determined by the sum of the PA voltages. Suitable PAs are therefore class D, saturated class B, and saturated class C (see Appendix A for a discussion of amplifier classes).

Different transformation ratios can be used to cause similar PAs operating from the same supply voltages to deliver different contributions to the load. However, it is most convenient to simplify the subsequent analysis by assuming all transformation ratios are unity.

The load impedance presented to PA1 is (assuming unity transformations)

$$Z_1 = \frac{V_1}{I_1} = \frac{V_1}{I_L} = \frac{V_1}{V_L} Z_L \tag{21.27}$$

The impedance presented to PA2 is given by an analogous relationship. If the two PAs produce equal output voltages, then each sees half of the system load impedance $Z_1 = Z_2 = Z_L/2$.

Given the voltage determined from (21.26) and the current then determined from (21.25), the power output can be computed using

$$P_o = \frac{1}{2}\mathrm{Re}\{V_L I_L^*\} \tag{21.28}$$

Baluns

In the circuit of Figure 21.20, the balun transformer T1 forces

$$I_1 = I_2 = I_L = \frac{V_L}{Z_L} \tag{21.29}$$

and

$$V_L = V_1 + V_2 \tag{21.30}$$

With the particular definitions of V_2 and I_2 used in Figure 21.20, all relationships developed for the conventional transformer configuration are also applicable to the balun configuration. Palpably, multiple baluns can be used to combine the outputs of 2, 4, 8, and so forth PAs.

When the outputs from the two PAs are identical,

$$V_L = 2V_1 = 2V_2 \tag{21.31}$$

The power output that would be delivered to the load by a single PA working by itself is

$$P_{PA1} = \frac{V_1^2}{2R_L} \tag{21.32}$$

However, the system output power

$$P_o = \frac{(2V_1)^2}{2R_L} = 4P_{PA1} \tag{21.33}$$

is four times that of one PA by itself.

Abnormal Operation

PA Failur—Open. If the output of PA2 fails and becomes an open circuit, then no current can flow and there is no signal output. Thus, this combiner configuration provides no protection against open-circuit failures of a PA.

PA Failure—Short. If PA2 fails as a short circuit, it and T2 are effectively removed from the circuit and PA1 drives the load by itself. The load seen by PA1 becomes $Z_1 = Z_L$ rather than $Z_L/2$ and delivers an output of $P_o = P_{PA1}$.

Signals of Different Frequencies

The use of a VSC for signals of different frequencies is not practical in most cases. The current coming from one PA has no path through the other PA because the different frequencies of operation implies that the output filtering is such that each PA is tuned to its own frequency.

21.6.3 Hybrid Combiners

The 180^o or hybrid power combiner is a four-port device that combines two RF signals while isolating one source from the other. This section presents the theoretical background for hybrid power combiners implemented with RF transformers

21.6.3.1 Fundamental Operation

Figure 21.21 depicts a hybrid power combiner implemented with a conventional transformer with a center-tapped primary winding ($2m$ to n turns). Hybrid power combiners can also be implemented with TLTs in a hybrid-balun configuration shown in Figure 21.22. For both cases

$$V_o = I_o Z_L \tag{21.34}$$

where V_o and I_o are complex phasors and Z_L can be matched ($Z_L = R_L$) or complex for a mismatched load.

Conventional Transformer

Both PAs are assumed to act a current sources and the directions of I_1 and I_2 are opposite, as shown in Figure 21.21. For the hybrid combiner based upon a conventional transformer,

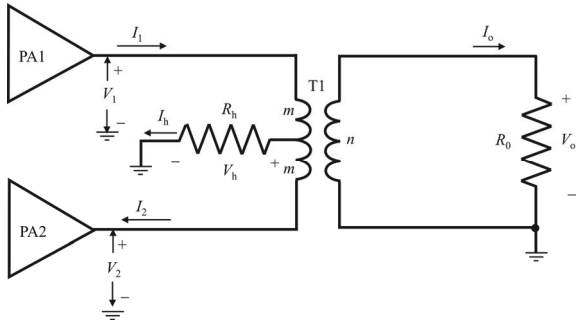


Figure 21.21 Hybrid combiner conventional transformers.

$$I_{\mathrm{L}} = \frac{n}{m}(I_1 + I_2) \tag{21.35}$$

The current in balancing resistor R_{h} is therefore

$$I_{\mathrm{h}} = I_1 - I_2 \tag{21.36}$$

When $I_1 = I_2$, $I_{\mathrm{h}} = 0$. The impedance seen by either PA is then

$$R = \left(\frac{n}{m}\right)^2 2R_{\mathrm{L}} \tag{21.37}$$

The two PAs serve as current sources and the network containing the transformer is a four-port. The S-parameter matrix is sought for this circuit, and one method to obtain it is to first compute the impedance matrix and then convert to the S-matrix. Doing so yields for the S-matrix

$$\mathbf{S} = \begin{bmatrix} \frac{R_{\mathrm{h}} - R_0}{R_{\mathrm{h}} + R_0} & 0 & 0 & 0 \\ 0 & -\frac{R_{\mathrm{h}}}{R_0} & 0 & 0 \\ 0 & 0 & -\frac{R_{\mathrm{h}}}{R_0} & 0 \\ 0 & 0 & 0 & \frac{R_{\mathrm{h}} - R_0}{R_{\mathrm{h}} + R_0} \end{bmatrix} \qquad R_L = R_0 \tag{21.38}$$

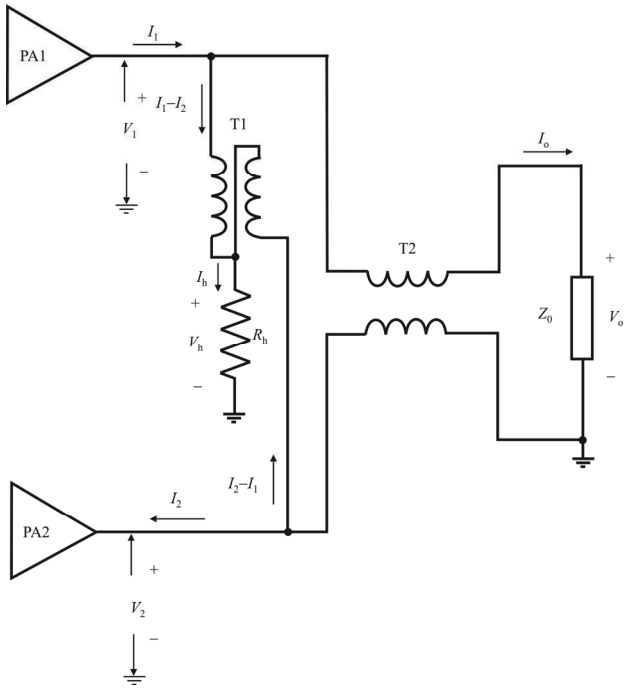


Figure 21.22 Hybrid combiner based upon transmission-line transformers T1 and T2.

When R_h is equal to the characteristic impedance R_0, then (21.38) reduces to the simple form

$$\mathbf{S} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{21.39}$$

Hybrid-Balun Transformers

Both PAs are assumed to function as current sources (as approximated by MOSFETs and BJTs) and the directions of I_1 and I_2 are opposite, as shown in Figure 21.22. A simple analysis shows that the output current is the average input current

$$I_{\mathrm{L}} = \frac{1}{2}(I_1 + I_2) \tag{21.40}$$

and

$$I_{\mathrm{h}} = I_1 - I_2 \tag{21.41}$$

as before.

When $I_1 = I_2$, $I_{\mathrm{h}} = 0$. The impedance seen by either PA is then

$$R = \frac{R_{\mathrm{L}}}{2} \tag{21.42}$$

Setting m/n = 1/2 makes all equations for the hybrid-balun configuration analogous to those for the conventional transformer configuration.

21.6.3.2 Input Impedance

The voltage at the output of PA1 is

$$V_1 = V_{\mathrm{h}} + \frac{m}{n}V_{\mathrm{o}} \tag{21.43}$$

The voltage on the balancing resistor is given by

$$V_{\mathrm{h}} = R_{\mathrm{h}} I_{\mathrm{h}} = R_{\mathrm{h}}(I_1 - I_2) \tag{21.44}$$

Substitution of (21.44) into (21.43) yields

$$\begin{aligned} V_1 &= (I_1 - I_2)R_{\mathrm{h}} + \left(\frac{m}{n}\right)^2 R_{\mathrm{o}}(I_1 + I_2) \\ &= \left(\frac{R}{2} + R_{\mathrm{h}}\right) I_1 + \left(\frac{R}{2} - R_{\mathrm{h}}\right) I_2 \end{aligned} \tag{21.45}$$

Division by I_1 then yields

$$R_1 = \left(R_{\mathrm{h}} + \frac{R}{2}\right) + \left(\frac{R}{2} - R_{\mathrm{h}}\right)\frac{I_2}{I_1} \tag{21.46}$$

We see that setting $R_{\mathrm{h}} = R/2$ makes R_1 independent of I_2 and therefore isolates PA1 from PA2. The resultant load seen by either PA is then

$$R_1 = R_2 = R \tag{21.47}$$

The analysis in this section remains valid if a reactive load Z_L replaces the resistive load R_L. It is therefore apparent that the hybrid coupler does not shield the PAs from reactive loads.

21.6.3.3 Power

The voltage on the balancing resistor is given by (21.44), hence the power dissipated in the balancing resistor is

$$P_h = \frac{|I_h|^2 R_h}{2} = \frac{|V_h|^2}{2R_h} = \frac{I_h I_h^* R_h}{2} \tag{21.48}$$

This relationship is valid for all combinations of amplitudes and phases. It is apparent that $P_h = 0$ only if $I_1 = I_2$ or $R_h = 0$.

If I_1 and I_2 have the same phase but different amplitudes, the first form in (21.48) can be used to calculate the power dissipated in the balancing resistor. It is apparent that the loss increases with the square of the magnitude of the difference current.

21.6.3.4 Different Frequencies

When the signals produced by the two PAs are of the same amplitude A but different frequencies, then they can be therefore represented by

$$i_1(t) = A\cos\omega_1 t \tag{21.49}$$

and

$$i_2(t) = A\cos\omega_2 t \tag{21.50}$$

where $\omega_1 \neq \omega_2$.

Since each PA acts as a current source, it is effectively an open circuit to the signal from the other PA. The current i_1 produced by PA1 therefore circulates from the PA through the upper m windings of T1 and then through R_h before reaching the ground. The output current is therefore

$$i_o(t) = \frac{m}{n} i_1(t) + \frac{m}{n} i_2(t) \tag{21.51}$$

and the current in the balancing resistor is

$$i_h(t) = i_1(t) - i_2(t) \tag{21.52}$$

Since the two signals are at different frequencies, the total power is simply the sum of the powers in the components; thus

$$P_o = \left(\frac{m}{n}\right)^2 A^2 R_0 \tag{21.53}$$

and

$$P_h = A^2 R_h \tag{21.54}$$

Selection of $R_h = R$ for isolation produces $R_1 = R_2 = R$ hence,

$$P_1 = P_2 = A^2 \frac{R}{2} \tag{21.55}$$

and a total (both frequencies) of

$$P_o = P_h = A^2 \frac{R}{2} \tag{21.56}$$

Half of the power at each frequency is delivered to the load. The other halves are delivered to the balancing resistor.

21.6.3.5 Summary

The advantages of hybrid combiners include:

- Isolation of one PA from the other;
- Constant, nominal load impedance;
- Continued operation at reduced power following failure of one PA.

The disadvantages include:

- Maximum of two inputs per combiner;
- No protection from reflections from the load (reactive) and therefore nonzero VSWR;

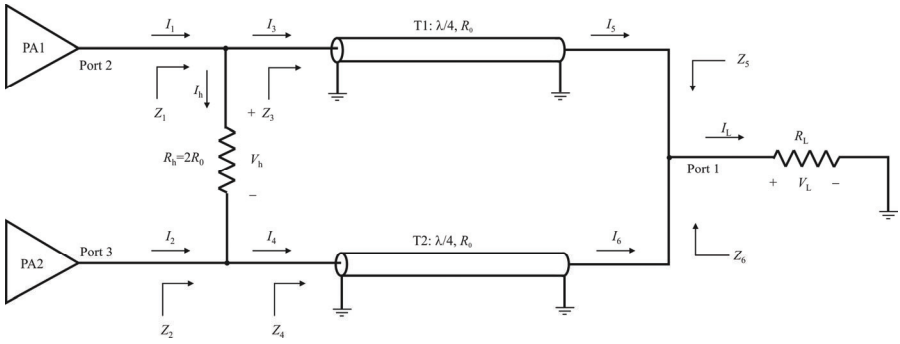


Figure 21.23 Basic Wilkinson hybrid combiner.

- Inefficient operation following failure of one PA;
- Inefficient combination of signals of different frequencies.

The conventional hybrid combiner coupler adds a resistor to the hybrid-balun transformer [4]. As a result, the average of the two PA output currents flows through the balun and the load, while the difference gets dumped into the dump load.

The hybrid combiner provides each PA with load impedance that is independent of the output of either PA. When the two PA output currents are equal in amplitude and opposite in phase, all power is delivered to the load.

Mismatch of the two PA currents can cause significant power to be dissipated in the dump resistor. For example, if PA2 fails, half of the power produced by PA1 is delivered to the load, and the other half is dissipated in the dump resistor. A similar situation occurs if the two PAs produce signals of different frequencies. In both of these cases, the efficiency of the combiner is 50%.

21.6.4 Wilkinson Combiner

The Wilkinson splitter can also be used as a combiner as illustrated in Figure 21.23. In this configuration ports 2 and 3 form the input ports and port 1 is the output port where the two input signals appear combined.

This section presents the basic theory of the Wilkinson power combiner. The basic two-PA combiner consists of the two quarter-wavelength transmission lines plus a hybrid balancing resistor as shown in Figure 21.23. This design is readily generalized to accommodate three or more inputs, as shown in Figure 21.24, where three are shown.

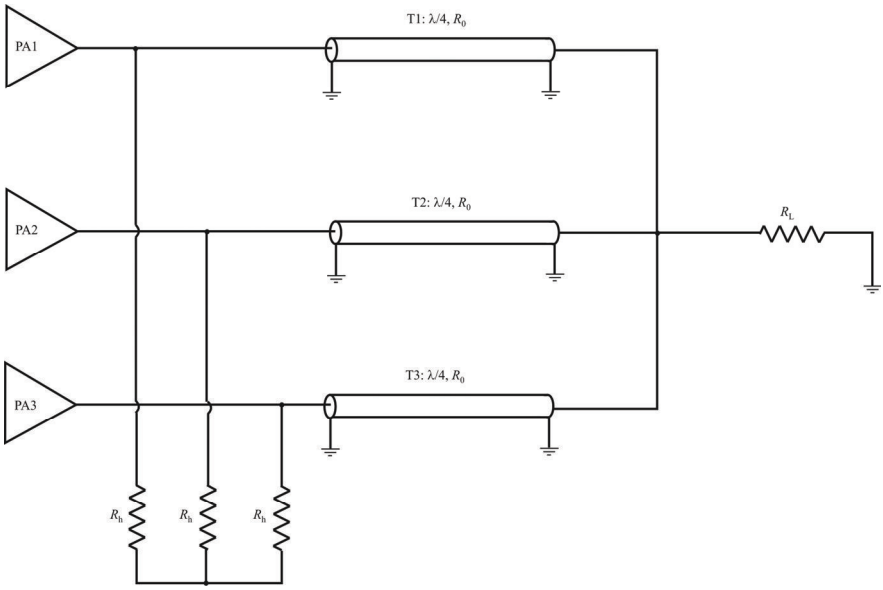


Figure 21.24 Wilkinson three-way.

21.6.4.1 Transmission-Line Coupler

The impedance at the input and output of a transmission line are related to the line impedance R_0 by Z_3 and Z_5 by (see Chapter 4)

$$R_3 R_5 = R_0^2 \tag{21.57}$$

Conservation of power at the input and output of the line requires that

$$P = \frac{1}{2} I_3^2 R_3 = \frac{1}{2} I_5^2 R_5 = \frac{1}{2} \frac{V_3^2}{R_3} = \frac{1}{2} \frac{V_5^2}{R_5} \tag{21.58}$$

Rearranging (21.58) yields

$$I_3^2 = \frac{V_5^2}{R_3 R_5} = \frac{V_5^2}{R_0^2} \tag{21.59}$$

from which

$$I_3 = \frac{V_5}{R_0} \tag{21.60}$$

Since

$$V_{\mathrm{L}} = V_5 = V_6 \tag{21.61}$$

$$I_3 = I_4 = \frac{V_{\mathrm{L}}}{R_0} \tag{21.62}$$

A key property of the quarter-wavelength transmission line combiner is therefore that the currents in all lines are forced to be equal.

21.6.4.2 Input Currents

Many PAs operating in classes A and B act as current sources (in particular MOSFET- and BJT-based PAs); thus, I_1 and I_2 are the variable inputs to the combiner. The transmission lines force $I_3 = I_4$. The average PA output current is

$$I_3 = I_4 = \frac{1}{2}(I_1 + I_2) \tag{21.63}$$

and the current in the hybrid resistor R_{h} is

$$I_{\mathrm{h}} = \frac{1}{2}(I_1 - I_2) \tag{21.64}$$

The voltage on the hybrid resistor is therefore

$$V_{\mathrm{h}} = \frac{1}{2}(I_1 - I_2)R_{\mathrm{h}} = I_{\mathrm{h}}R_{\mathrm{h}} \tag{21.65}$$

The power dissipated in the hybrid resistor is then

$$P_{\mathrm{h}} = \frac{1}{2}\frac{V_{\mathrm{h}}^2}{R_{\mathrm{h}}} = \frac{1}{2}I_{\mathrm{h}}^2 R_{\mathrm{h}} = \frac{1}{2}V_{\mathrm{h}}I_{\mathrm{h}} \tag{21.66}$$

21.6.4.3 Output

The voltage output from (21.60)

$$V_{\mathrm{L}} = I_3 R_0 \tag{21.67}$$

hence the output current is

$$I_{\mathrm{L}} = \frac{V_{\mathrm{L}}}{R_{\mathrm{L}}} = \frac{I_3 R_0}{R_{\mathrm{L}}} = \frac{1}{2}\frac{(I_1 + I_2)R_0}{R_{\mathrm{L}}} \tag{21.68}$$

The output power is therefore

$$P_{\mathrm{o}} = \frac{1}{8}\frac{(I_1 + I_2)^2 R_0^2}{R_{\mathrm{L}}} \tag{21.69}$$

21.6.4.4 Input Powers and Impedances

The impedance at the load end of a transmission line is that of the load prorated according to the fraction of the load current contributed by the line in question. Consequently,

$$R_3 = \frac{R_0^2}{R_5} = \frac{I_5 R_0^2}{(I_5 + I_6)R_{\mathrm{L}}} \tag{21.70}$$

Since the equation for R_4 is similar but has I_6 in the numerator, it is apparent that

$$R_3 + R_4 = \frac{R_0^2}{R_{\mathrm{L}}} \tag{21.71}$$

The pair of transmission lines thus acts as a sort of balance, dividing the total output load between the two PAs. The input impedance of the second line in Figure 21.23 is therefore

$$R_4 = \frac{R_0^2}{R_{\mathrm{L}}} - R_3 \tag{21.72}$$

The two PA voltages now have the forms

$$V_1 = I_3 R_3 \tag{21.73}$$

and

$$V_2 = I_4 R_4 = I_3 \left(\frac{R_0^2}{R_\mathrm{L}} - R_3 \right) \tag{21.74}$$

The voltage on the hybrid resistor is then

$$V_\mathrm{h} = V_1 - V_2 = 2 I_3 R_3 - I_3 \frac{R_0^2}{R_\mathrm{L}} \tag{21.75}$$

Combining this with (21.65) and substituting (21.63) and (21.64) for the currents yields

$$R_3 = \frac{R_0^2}{2R_\mathrm{L}} + \frac{1}{2} \frac{I_1 - I_2}{I_1 + I_2} R_\mathrm{h} \tag{21.76}$$

and

$$R_4 = \frac{1}{2} \frac{R_0^2}{R_\mathrm{L}} - \frac{1}{2} \frac{I_1 - I_2}{I_1 + I_2} R_\mathrm{h} \tag{21.77}$$

The impedance seen by PA1 is

$$\begin{aligned} Z_1 &= \frac{V_1}{I_1} = \frac{I_3 R_3}{I_1} \\ &= \frac{1}{4} \frac{R_0^2}{R_\mathrm{L}} + \frac{1}{4} R_h + \left(\frac{1}{4} \frac{R_0^2}{R_\mathrm{L}} - \frac{1}{4} R_\mathrm{h} \right) \frac{I_2}{I_1} \end{aligned} \tag{21.78}$$

An analogous expression can be derived for Z_2. We see that setting

$$R_\mathrm{h} = \frac{R_0^2}{R_\mathrm{L}} \tag{21.79}$$

causes Z_1 in (21.75) to become independent of I_2 (PA2) and equal to the half of the load impedance transformed through the transmission line; that is,

$$Z_1 = \frac{R_0^2}{R_L} \tag{21.80}$$

21.6.5 Summary

We introduced the basic concepts of power combiners and their use in EW systems in this section. Also considered were the normal failure modes of PAs used to drive these combiners, and their effects on the power output and efficiency of the EW system output stages.

21.7 RF Diplexers

An antenna diplexer or RF diplexer is a unit that in one application can be used to enable more than one transmitter to operate on a single RF antenna. Sometimes these units are called antenna *duplexers*. Typically, an antenna diplexer would enable transmitters operating of different frequencies to use the same antenna. In another application, an antenna diplexer may be used to allow a single antenna to be used for transmissions on one band of frequencies and reception on another band.

Antenna diplexers find many uses. In one common example, an antenna diplexer or RF diplexer is used in a cellular base station to allow it to transmit and receive simultaneously. The antenna diplexer enables the same antenna system to be used while preventing the transmitted signal from reaching the receiver and blocking the input. In another application, a diplexer may be used by a broadcast station transmitting on several different frequencies at the same time using the same antenna. The use of the diplexer enables a single antenna to be used, while preventing the output from one transmitter being fed back into the output of the other.

21.7.1 Basic Antenna Diplexer Concepts

There are a number of ways of implementing RF diplexers. They all involve the use of filters. In this way the paths for the different transmitters and receivers can be separated according to the frequency they use. The simplest way to implement a diplexer is to use a lowpass and a highpass filter although band-pass filters may be used (see Figure 21.25). In this way, the diplexer routes all signals at frequencies below the cutoff frequency of the lowpass filter to one port, and all signals above the cut-off frequency of the highpass filter to the other port. Also, there is no path from between the two remote connections of the filters. All signals that can pass through the lowpass filter in the diplexer will not be able to pass through the highpass filter and vice versa.

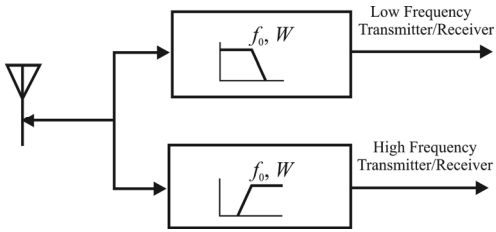


Figure 21.25 Basic concept of a highpass/lowpass filter diplexer.

A further feature of an RF diplexer is that it enables the impedance seen by the receiver or transmitter to remain constant despite the load connected to the other port. If the filters were not present and the three ports wired in parallel, neither the antenna nor the two transmitter/receiver ports would see the correct impedance.

21.7.2 RF Diplexer Filter Requirements

When designing an antenna diplexer, a number of parameters must be considered. One is the degree of isolation required between the ports labeled for the high and low frequency transmitter/receiver. If the diplexer is to be used purely for receiving, then the requirement for high levels of isolation is not so high. Even comparatively simple filters give enough isolation to ensure each receiver sees the right impedance and the signals are routed to the correct input without any noticeable loss. Even levels of isolation of 10 dB would be adequate for many installations.

The next case is when the diplexer is to be used for transmitting only. It will be necessary to ensure that the levels of power being transferred back into a second transmitter are small. Power being fed into the output of a transmitter in this way could give rise to IMD products that may be radiated and cause interference. It is also important to ensure that the transmitters see the correct impedance, and that the presence of the second transmitter does not affect the impedance seen by the first. Typically levels of isolation between the transmitter ports of 60–90 dB may be required.

The final case is where one of the ports is used for transmitting, and the other for receiving simultaneously. In this instance, very high levels of isolation are required to ensure that the minimum level of the transmitter power reaches the receiver input. If high levels of the transmitter signal reach the input to the receiver, then it will be desensitized preventing proper reception of the required signals. Levels of isolation in excess of 100 dB are normally required for these applications.

21.7.3 Bandpass Filters

Under some circumstances bandpass filters may be used. These may be used where comparatively narrow bandwidth is required for either or both of the transmitter/receiver ports. Sometimes a very high Q resonant circuit may be used. By using this approach, high degrees of rejection can be achieved. Often repeater stations that receive on one channel and transmit on another simultaneously use diplexers that utilize this approach.

21.7.4 Summary

Although antenna diplexers are mainly used in specialized applications, allowing a single RF antenna to be used by more than one transmitter or receiver, they are nevertheless a crucial element of many installations. For example, cellular technology would be significantly different if they could not be used and the cellular RF antennas for base stations would be considerably more complicated. Similarly antenna diplexers are used in many broadcast applications allowing a single large RF antenna to be used by more than one transmitter.

21.8 RF Circulators and Isolators

An RF or microwave circulator is used to pass RF signals and block returned signals. A similar statement applies to isolators—in fact, an isolator can be constructed using a circulator. While used for several purposes in systems, our interest in circulators is as interfaces between antennas and the rest of the EW system, and that is what we will focus on here.

The circulator is a three, (or more)-terminal device which passes signals input- from one port to the next port in a rotational fashion without allowing signals to pass in the opposite rotation. Circulators generally contain a microwave circuit comprising an arrangement of conductors, ferrite blocks, and a magnetic circuit providing a magnetic biasing field applied to the ferrite blocks that act as a nonreciprocal media for propagating RF signals throughout the device. RF circulators are useful as isolators, easily made by tying the third circulator port to ground through a resistor. RF circulators can be built in resonant structures such as RF resonant cavities, and in waveguides at higher frequencies. Circulators may also be realized in planar configuration using stripline or microstrip technology, which employ a planar resonating element between two ground plane conductors (stripline) or coupled to a single ground plane conductor (microstrip). RF and microwave circulators employ a DC-biasing magnetic field generated in ferrite material enveloping a conductor to provide at least one non-reciprocal transmission path between signal ports on a network. As we have discussed, it is

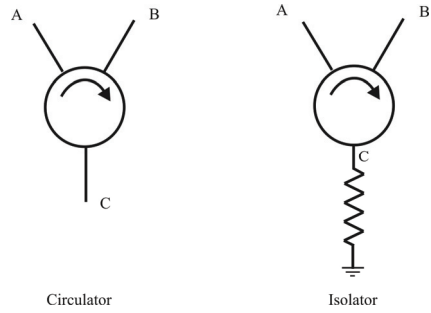


Figure 21.26 Three-way circulator and isolator symbols.

often advantageous to couple multiple devices, such as a transmitter and a receiver, to the same antenna. A common approach uses a RF circulator to isolate the transmitted signal from the received signal to avoid overloading the receiver front end. An array antenna can include a plurality of RF circulators disposed in an array in a manner in which RF signals can be received from or transmitted to the same individual radiator.

An RF isolator is a two-port ferromagnetic passive device which is used to protect other RF components from excessive signal reflection. An RF circulator is a three-port ferromagnetic passive device used to control the direction of signal flow in a circuit and is a very effective, low-cost alternative to more complex cavity diplexers.

The interaction of the magnetic field to the ferrite material inside isolators and circulators creates circulating magnetic fields. The rotary magnetic field is very strong and will cause any RF/microwave signals in the frequency band of interest at one port to follow the magnetic flow to the adjacent port and not in the opposite direction.

Figure 21.26 shows the circuit symbols for a circulator and an isolator. Notice how an isolator is a circulator with the third port terminated, usually with the system characteristic impedance R_0. The arrows indicate the direction of the magnetic fields and therefore the signal when applied to any port of these devices.

> **Example:** if a signal is placed at port A, and port B is well matched, the signal will exit at port B with very little loss (typically 0.4 dB). If there is a mismatch at port B, the reflected signal from port B will be directed to port C. As you will note, it makes no difference which port is the input of the circulator because the relationship at the outputs remains the same as these devices are electrically and mechanically symmetrical.

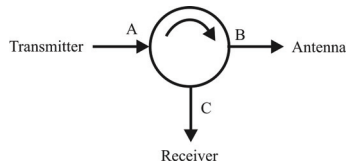


Figure 21.27 Duplexer.

21.8.1 Isolation

Isolation is an indication of the separation of signal levels on adjacent ports of a device, and is usually given in decibels. The greater the isolation, the less interference from a signal on one port is present at the other. The amount of isolation is directly affected by the VSWR presented at port C of the isolator. If the match on port C is poor, isolation will be below 10 dB, but if the match is improved to 1.10:1 by using a good termination device, then the isolation would improve to over 20 dB.

21.8.2 Insertion Loss

An important consideration is to ensure the device has minimal insertion loss when inserted in a transmission path. Generally, the insertion loss of a circulator/isolator (or any microwave device for that matter) becomes more significant at higher-frequency, namely because loss increases with frequency and as such, higher frequency power sources are considerably more expensive. Accordingly, the criteria of low insertion loss will prevent power from being wasted.

21.8.3 Diplexer

As described earlier, a circulator can function as an inexpensive diplexer (a transmitter and receiver sharing one antenna). Figure 21.27 illustrates that the signal from the transmitter goes to the antenna port and is isolated from the receiver. Good isolation is key to ensure that a high-power transmitter output signal does not get back the receiver front end as is governed by the return loss of the antenna. In this configuration, all signals from the antenna go straight to the receiver and not the transmitter because of the circular signal flow.

21.8.4 Other EW Applications of Circulators and Isolators

We briefly discuss some other (antenna-related) uses for circulators and isolators in EW systems in this section.

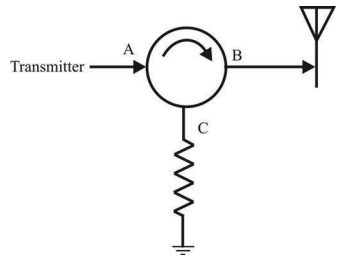


Figure 21.28 Decoupling a transmitter.

21.8.4.1 Decoupling of a PA or Receiver from its Antenna

When a PA is connected directly to an antenna it is typically influenced by impedance changes of the antenna caused by natural occurrences, such as heavy rain, or nearby reflecting obstacles. When an isolator is placed between the PA output and the antenna, these effects can be significantly reduced by about 20 dB (see Figure 21.28). In addition, other nearby antennas, perhaps sharing the same mast, may couple signals into the antenna, which travel to the amplifier.

A final stage operating in class B or C to give high-efficiency (see Appendix A) results in very non-linear behavior. Therefore, the induced signal will mix with the signal of the PA producing IMD. The intermodulation can be reduced by an isolator, but in this case, 20 dB is probably not enough. To reduce the level of the IMD products to an acceptable level, double isolators with an isolation of about 50 dB are required.

The input stage of a receiver is often an LNA whose input impedance is chosen to give a very low NF, but no match to the antenna. An isolator between the antenna and the input of the receiver will solve this problem.

21.8.4.2 Combiner for Two or More Transmitters in the VHF and UHF Bands

Combining two or more transmitters, for the operation on one antenna, is done conventionally in the following way [see Figure 21.29(a)]: pairs of transmitters are connected to a 3 dB hybrid. One output arm is terminated to a matched load, the other goes to the antenna, if we have to combine two transmitters only, or to the next 3 dB hybrid if there are more. If we have to combine four transmitters as in Figure 21.29(a), we use three 3 dB hybrids, and the signal of each transmitter passes two of them. The losses in this coupling circuit are high: we have about 0.5 dB losses in the 3 dB hybrid and convert half of the power in the matched load into heat. For our combiner for four transmitters, this results in a loss of about 7 dB. If we have to combine more than four transmitters, the losses are even higher.

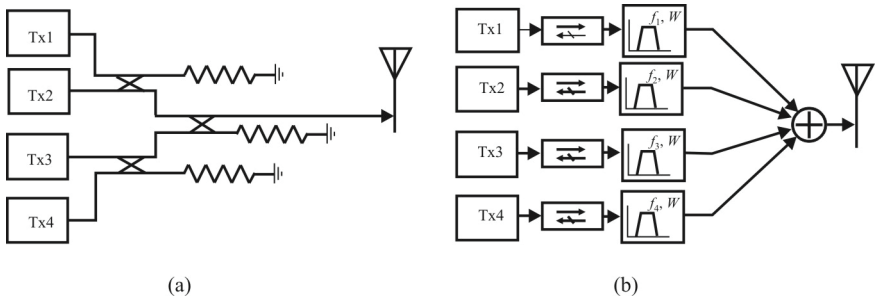


Figure 21.29(a, b) Combiner for four transmitters: (a) conventional using hybrids and (b) using isolators.

The frequency spacing of the transmitters is limited only by the bandwidth of the 3 dB hybrids.

Using isolators, we can build a completely different combiner [see Figure 21.29(b)]. The transmitter is connected to a bandpass filter, which is tuned to its signal, via a double isolator. Cables with a length of l/4 or odd multiples of it or l/2 or multiples of it, depending of the coupling of the filter, connect the different filters to the antenna.

The signal of transmitter 1 passes the isolator, which has an insertion loss of about 0.5 dB, and filter 1, and then travels to the antenna. The other transmitters are on different frequencies; therefore, the filters are tuned to different frequencies too, and the cables transform their impedances to a high impedance at the antenna. This lets the signal of the transmitter 1 travel to the antenna, attenuated about 1.5 dB by filter 1 and losses in the cables and the other filters. The losses in the other filters depend on the frequency spacing, the coupling, and the *Q*-factors of the filters. The total loss of the signal in the combining network for the four transmitters of our example is typically 2 dB and does not increase significantly if we increase the number of transmitters. Up to 18 transmitters can be combined in this way in the UHF band.

We need the double isolators to attenuate the signals of the other transmitters leaking through the filters. Otherwise we would run into intermodulation problems caused by the nonlinearity of the final stages of the transmitters.

21.8.4.3 Combining Amplifier Stages in a Solid-State Transmitter

In solid-state transmitters for high-power output, several transistor PAs are connected in parallel and their outputs have to be combined. To do this with a minimum of losses, often star point combiners are used; see Figure 21.30(a). But also 3 dB hybrids can be chosen. In normal operation the transistors see the nominal load. But as discussed above, if one of the transistor PAs fails, a big reflection is transformed to the other PAs. If they cannot tolerate this, they too will

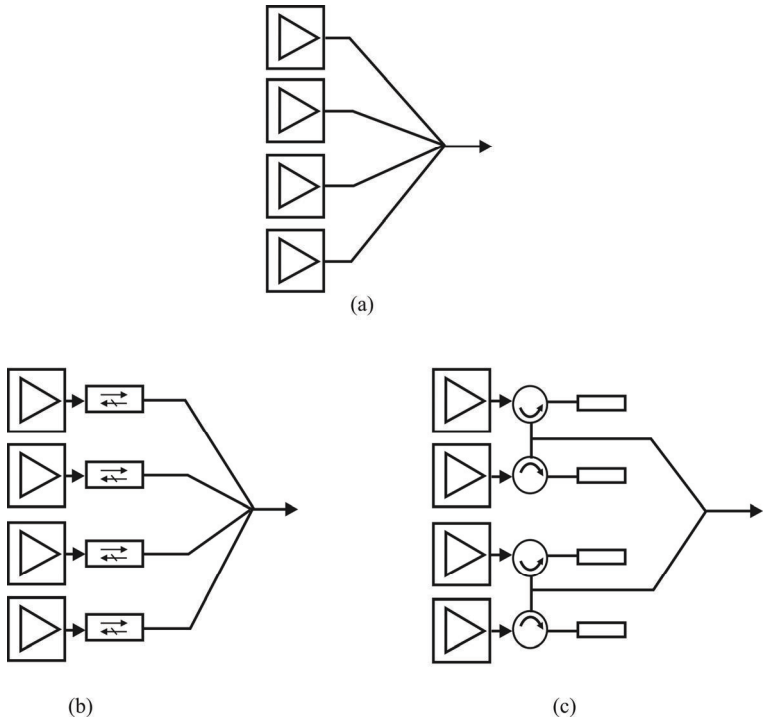


Figure 21.30(a–c) Combining amplifier stages in solid state amplifiers: (a) simple star combiner, (b) using isolators, and (c) combining isolators.

fail. To avoid this, we can use isolators at each amplifier output; see Figure 21.30(b).

Another solution is the use of special combining isolators as given in Figure 21.30(c). Such a combining isolator consists of two isolators. The PAs being combined are connected to the inputs which have an impedance of 50Ω. The output of each isolator has an impedance of 100Ω, and they are connected in parallel. Therefore the output of the combiner has 50Ω again.

For proper operation of all of these combining networks the insertion phase of the PAs and isolators or circulators must be equal within approximately $\pm 5^{o}$.

21.8.4.4 Operation of One-Port Amplifiers

If we want to use a one-port-amplifier (e.g., varactors, Gunn-amplifiers, and masers), we have to construct an input and an output port for the signal being amplified. The only possibility to do it is to use a circulator as given in Figure

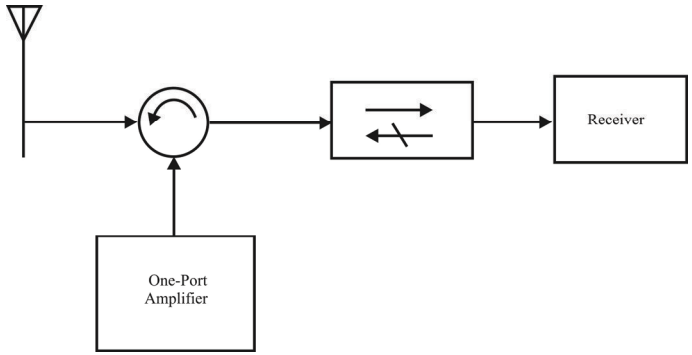


Figure 21.31 Operation of one port.

21.31. The isolation of the circulator must be higher than the amplification of the amplifier to give stable operation. An isolator between circulator and receiver avoids deterioration by the input impedance of the receiver.

21.9 Concluding Remarks

The operation of multicouplers, combiners, diplexers, and related components for dividing and combining RF signals in EW systems were discussed in this chapter. There is considerable technical literature available on each of these topics—we only scratched the surface here. Hopefully enough information was provided to adequately introduce these topics so that a basic understanding of their use in EW systems could be understood.

References

[1] Poisel, R. A., *Modern Communications Jamming Principles and Techniques*, 2nd ed., Norwood, MA: Artech House, 2011, p. 503.

[2] Wilkinson, E. J., "An N-Way Hybrid Power Divider," *IRE Transactions on Microwave Theory and Techniques*," January 1960, pp. 116–118.

[3] www.rfdesign.com, August 2002.

[4] Walker, J. L. B., D. P. Myer, F. H. Raab, and C. Trask (Eds.), *Classic Works in RF Engineering: Combiners, Couplers, Transformers, and Magnetic Materials,* Norwood, MA: Artech House, 2006.

[5] Krauss, H. L., C. W. Bostian, and F. H. Raab, *Solid State Radio Engineering*, New York: Wiley, 1980, pp. 373–382.

[6] Sevick, J., *Transmission Line Transformers*, 4th ed., Raleigh, NC: Scitech, 2006.

Chapter 22

Radomes

22.1 Introduction

The purpose of a radome is to form a protective cover between an antenna and the environment surrounding it. This protection requires that there be little impact to the electrical performance of the antenna. Ideally, there is no impact to this electrical performance so that essentially the radome is electrically invisible at RF. How well this is accomplished by a radome depends on selecting the correct material from which it is constructed and matching that to the RF range under consideration.

Radomes can come in any form factor that makes sense for the antenna they are protecting. Large, essentially round, radomes are typically used for microwave dishes placed on the ground. Aircraft avionics radomes take the shape of the antenna they cover, and are typically bent in an L-shape for aerodynamic reasons. These radomes could also be conformal to match the skin of the platform upon which the antenna is mounted. Radomes covering an antenna on a submarine would typically be a bulge on the surface of the skin.

We discuss the general characteristics of radomes in this chapter.

22.2 Radome Characteristics

Radomes are used to protect a wide range of ground-based and shipboard EW systems and radar installations as well as airborne EW system antennas [1]. The proper selection of a radome for a given antenna can actually help to improve system performance. The factors involved include:

- Wind loading minimization. This is particularly useful for highly directional antennas with narrow beamwidths such as microwave dishes. The benefit is to maintain alignment with other directional antennas.

- Minimizing the environmental factor impacts within which an antenna must operate. Radomes protect the antenna from rain, snow, hail, sand, salt spray, insects, animals, UV damage, wide temperature fluctuations, fungus, and corrosion.
- Providing cover for maintenance personnel—important for poor-environment installations. This is a seemingly pedestrian benefit unless you happen to be the technician charged with the task of maintaining the antenna.
- Preventing visual observation of system (security). This is particularly important for many military applications, to include EW. The size of an antenna is an important consideration for estimating its operating frequency range.
- Minimizing downtime and extending component and system operating life. The environment within the radome can be environmentally controlled if required which further enhances this consideration.
- Maintaining the structural integrity of an antenna. Examples of where this is important are aircraft avionics antennas and submarine communication antennas.

A variety of materials can be used for constructing radomes. Even wood has been used. Radomes are frequently constructed of composite materials, which are materials that are made of multiple constituent materials. The composite material combines strengths of multiple materials to produce a new material with better properties than the materials have individually. Composite materials such as fiberglass, quartz, and aramid fibers held together with polyester and epoxy [2] are used for manufacturing modern ground-, air-, and ship-based radomes.

Cores made of foam and low dielectric honeycombs are frequently added between inner and outer "skins" of the radome to function as a low-dielectric constant spacer material providing structural strength. These layers keep the skins at a constant separation, which provides for consistent electrical performance of the radome.

To minimize the impact to the radiation pattern and insertion loss, it is imperative to use material with a low dielectric constant ($\varepsilon_{air} = 1$). Some materials such as plastics have a dielectric constant close to 2, which is low. However, other system requirements such as strength, operating temperature, or low cost may preclude their use. Appendix D discusses the dielectric constant for other materials.

An electrically thin radome is a single layer radome where the layer is $\lambda/10$ or less thick at the frequency of interest. On the other hand, a half-wave radome is a single layer radome where the layer is $\lambda/2$ thick at the frequency of interest.

The gain of a radome is the ratio of the power density of an antenna's radiation pattern in the direction of strongest radiation to that of a reference antenna, typically isotropic or a dipole. It reflects on the ability of the antenna system to focus an RF signal at the frequency of interest.

G/T [G = gain (unit-less), T = temperature in Kelvins] is a figure of merit for satellite antennas similar to the SNR. The insertion loss of a radome is the energy loss due to reflection and absorption loss. The reflection loss is the energy lost because it is reflected by the radome, while the absorption loss is the energy lost due to absorption and converted to heat due to dielectric loss.

22.3 RF Reflections

Radomes are normally made of materials with their dominant properties given by the dielectric constant, loss tangent, and other electrical parameters. These materials are frequently dielectric. Dielectric materials are discussed in more depth in Appendix D. Dielectric materials have a characteristic impedance of

$$Z_D = \frac{\eta_0}{\sqrt{\varepsilon_r}} \tag{22.1}$$

where ε_r is the dielectric constant relative to free space. From Chapter 2 we know that the impedance of free space is

$$\eta_0 = \sqrt{\frac{\mu_0}{\varepsilon_0}} = 120\pi \approx 377\ \Omega \tag{22.2}$$

When an EM wave in free space strikes a dielectric material at normal incidence as shown in Figure 22.1, some of the energy in the signal is transmitted into the dielectric and some of it is reflected back into the direction from which it came. The reflection coefficient at this boundary is

$$\Gamma = \frac{Z_D - \eta_0}{Z_D + \eta_0} \tag{22.3}$$

Since Z_D is less than η_0, the reflection coefficient Γ is negative which means the reflected wave is 180° out of phase with the incident wave. When the wave hits the free space boundary on the other side of the dielectric, the numerator reverses and

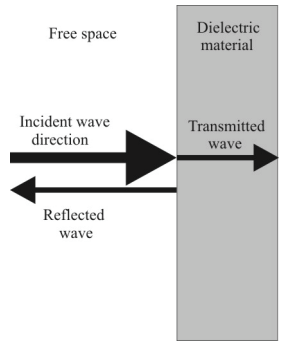


Figure 22.1 Reflected and transmitted waves.

$$\Gamma = \frac{\eta_0 - Z_D}{Z_D + \eta_0} \tag{22.4}$$

This reflection coefficient is positive, which indicates that most of the energy is transmitted out of the skin into free space again.

22.4 Radome Configurations

Several radome configurations are used to minimize RF reflections, including electrically thin, half-wave, A-sandwich, C-sandwich, and others [3]. The best configuration for a particular application depends on the mechanical requirements, operating frequency, and the surrounding environment.

A radome that is electrically thin (less than $\lambda/10$) [4], as depicted in Figure 22.2 [1], has low insertion loss because signal reflections at the free space/dielectric boundary on the left are cancelled out by out-of-phase reflections from the dielectric/free space boundary on the right side of the dielectric material. Figure 22.3 [1] shows that signal losses are low, less than 1 dB over the frequency range considered. Electrically thin radomes, however, provide little thermal insulation and are not suitable for locations with wide temperature extremes. In addition, high winds may preclude the use of thin materials.

Similar performance can be achieved from a configuration based on the half-wavelength-thick solid laminate shown in Figure 22.4 [1]. Its effectiveness is similar to the electrically thin configuration because the reflections cancel out. The

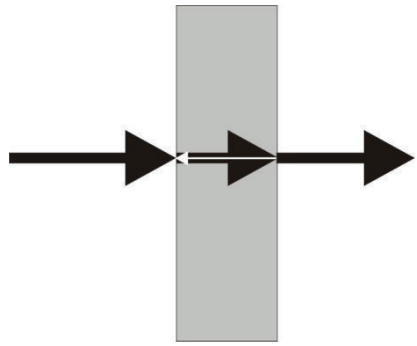


Figure 22.2 In an electrically thin dielectric layer as depicted here, reflections at the air/dielectric boundary on the left side are cancelled by the reflections from the right side of the laminate at the dielectric/air boundary, resulting in low-loss transmission.

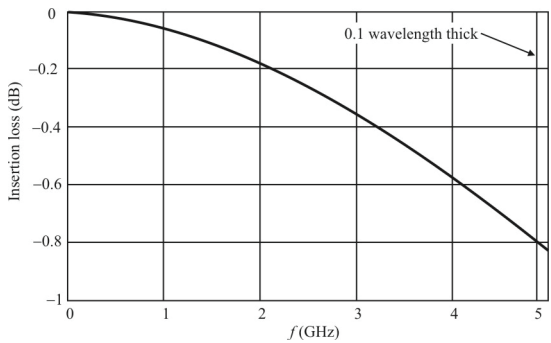


Figure 22.3 The insertion loss through a thin dielectric material at normal incidence. The laminate is 0.12 inch thick, has a dielectric constant of 4, and a loss tangent of 0.01.

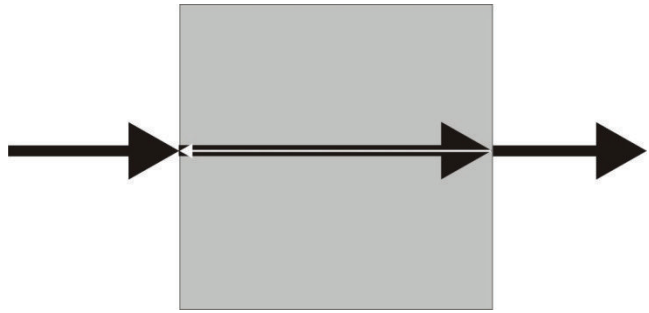


Figure 22.4 In a half-wavelength thick radome, the roundtrip of signals through and reflected by the laminate on the right side produces a 360° total phase shift. Being out-of-phase, the reflections at each interface cancel because they are out of phase resulting in low insertion loss.

wave travels 180° through the laminate, is reflected with a phase shift of –180°, and travels another 180° on the return trip to achieve the net 180° phase shift (total phase shift 540°) required for cancellation. Figure 22.5 [1] shows the performance of this laminate.

An A-sandwich radome configuration (Figure 22.6 [1]) consists of a low dielectric foam or honeycomb core sandwiched between two thin laminates. Its functioning is similar to the half-wavelength-thick solid laminate. However, it is 0.25 wavelength thick because the reflection coefficients from the skins have the same amplitude and phase. The roundtrip for the reflection from the second skin is 0.5 wavelength. Since the reflections are 180° out of phase, they cancel. Performance is summarized in Figure 22.7 [1]. Over the frequency range indicated here, insertion loss is less than 2 dB everywhere. $\lambda/2$ is achieved at 25 GHz and the insertion loss is virtually zero. Such configurations allow a great deal of flexibility in designing radomes, since the thickness and dielectric constant can be "tuned" to the operating frequency (for a communication application; a luxury not allowed for noncooperative EW intercept).

A C-sandwich radome consists of three skin layers and two foam layers, as shown in Figure 22.8 [1]. Selection of the thickness of each foam layer, as well as the thickness of the skins, allows for "tuning" this structure for the specified operating frequency for communication applications. This tuning produces virtually zero insertion loss, but even at "unturned" frequencies the insertion loss is quite low (typically less than 2 dB) so is applicable to EW intercept radomes. C-sandwich constructions provide better performance than A-sandwich radomes; however, the added complexity increases construction costs.

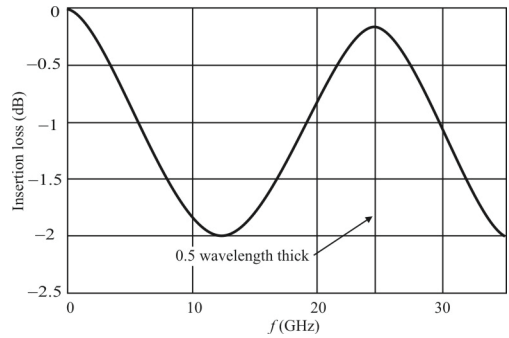


Figure 22.5 The insertion loss of the laminate in Figure 22.4 at higher frequencies. At a thickness of 0.5 wavelength, the reflections cancel and the insertion loss is low.

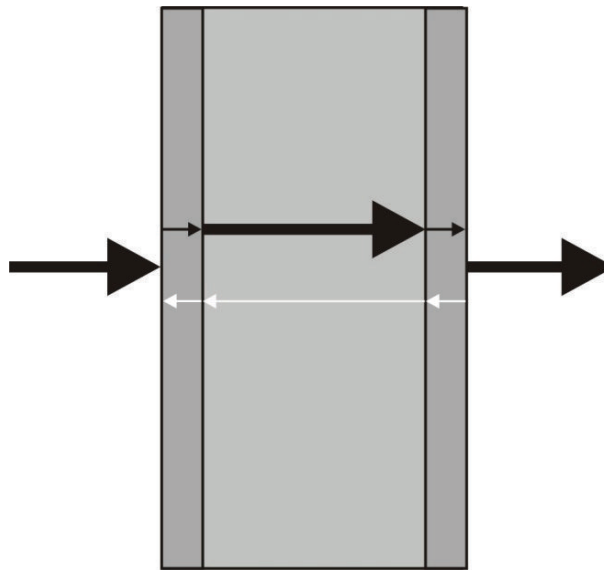


Figure 22.6 An A-sandwich radome construction consists of a foam core, surrounded on both sides with thin laminates.

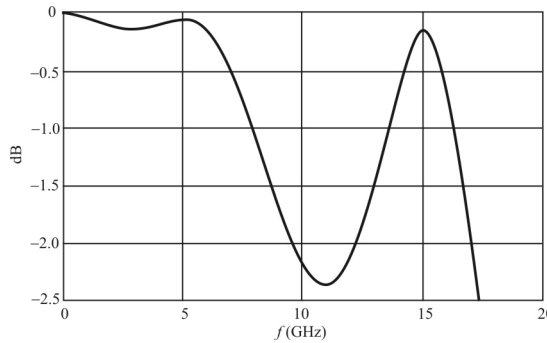


Figure 22.7 Reflections of an A-sandwich radome are plotted versus frequency. In this case, the thickness of the foam core is 0.25 wavelength at 5 GHz, which provides maximum performance at less than 7 GHz. At 15 GHz the insertion loss is also low because the phase shift is an odd multiple of 180°.

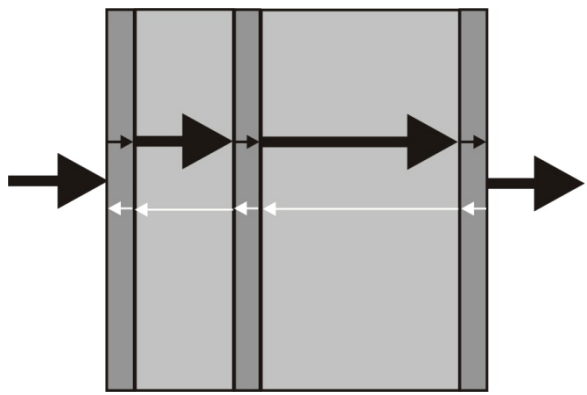


Figure 22.8 C-sandwich radome configuration.

Table 22.1 Properties of Some Radome Types

	Features				Drawbacks		
Radome Type	**Can Withstand > 150 mph Winds**	**Electrically Thin Broadband Performance**	**Tuned Multiband Performance**	**Provides Thermal Insulative Properties**	**Requires Constant Positive Pressure**	**Support Frame Adds Significant Loss**	**Notes**
Self-supporting sandwich	X		X	X			
Inflatable	X	X			X		
Metal space frame	X	X	X		X*	X	
Dielectric space frame	X	X	X		X*	X	Insertion loss ripple above 1 GHz
Solid laminate	X			X			Single band tuning

*Thin fabric membrane radomes need positive pressure to prevent wind damage in high wind conditions.
Source: [1].

22.5 Structural Support

There are several techniques for supporting radomes, when such support is required. Such support structures are almost always required for ground- and sea-based applications. Ground and shipboard radomes can range in size from very small antenna covers to massive structures tens of meters in diameter. Radomes are also used extensively on airframes and missiles to protect sensitive sensor antennas. Radomes for these antennas are usually part of the platform.

The benefits and limitations of radome support structures are summarized in Table 22.1. The A-sandwich configuration is usually used for self-supporting radomes. They are made of rigid sections that are fastened together by some means. With proper attention to how the seams are manufactured, the phase delay and insertion loss through the seams become virtually invisible to the EM wave front. A-sandwich radomes do not require that pressure be maintained inside the radomes and therefore are not dependent on a constant supply of electrical power. They also generally have lower overall operating costs.

Inflatable radomes are made of electrically thin dielectric cloth and require a constant supply of air. By being electrically thin, they are capable of achieving very low loss over wide bandwidths, however. In order to maintain adequate air

pressure, inflatable radomes must be equipped with airlocks at all doors and a stand-by power supply. Should the membrane suffer damage or if power is interrupted, it is possible for the radome to deflate and collapse. Operating costs for this type of radome usually exceed those of all other radome types.

Metal space frame radomes support the window portion of the radome consisting of the electrically thin, half-wave, or A-sandwich configuration, often in the shape of a geodesic dome. The window portion typically has very low loss. However, signal blockage and reflections from the frame reduce system gain and reflect noise back into the system. Because the frame reflects the RF wave front, it decreases the power in the main lobe and increases sidelobe levels.

Dielectric space frame radomes are supported by support members manufactured with dielectric materials, which are more electrically transparent than metal frames. However, the phase of the wavefront is delayed as it passes through the dielectric support, alternating in and out of phase depending on frequency. This can be a problem for EW applications of this type of structure (however, metal frames exacerbate this problem as well). If the delay is 180° out of phase with the phase of the incident signal, the energy that passes through the frame subtracts from the gain. This leads to a frequency-dependent ripple in the insertion loss and the lost energy goes into the sidelobes. This makes dielectric space frame radomes best suited to systems that operate at less than 1 GHz.

In order to maintain and enhance the structural integrity of their thin membrane coverings during windy conditions, both types of space frame radomes usually require the use of air blowers or compressors.

22.6 Effects of Incident Angle

The discussions above assumed normal incidence. An EM wave rarely hits the radome surface normal to that surface, but usually at an oblique angle. In the case of a spherical surface there is a continuous variation in the angle of incidence. As we indicate in Appendix D, being composed of dielectric materials, the transmission characteristics of a radome changes with the wave incidence angle and polarization. Electric fields that are parallel to the plane of incidence have much higher transmission than fields that are perpendicular to the plane of incidence.

EM wavefronts from parabolic antennas located inside spherically shaped radomes see low incident angles at the center of the wave front. The incident angle becomes higher on the edges however. If the antenna illumination pattern is symmetric, and the antenna is placed at the center of the spherical radome, the symmetric shape of the radome cancels out axial ratio degradation from the oblique incidence angles seen by the antenna.

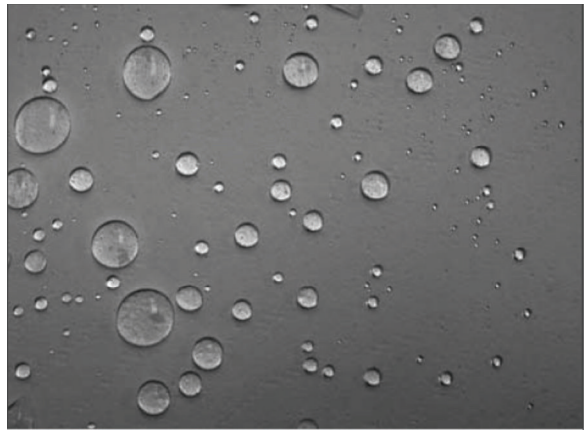


Figure 22.9 RF performance degradation is minimized with hydrophobic surfaces that cause water to bead. (Source: [1] © Microwave Product Digest 2008. Reprinted with permission.)

22.7 Radome Performance

The principal goal for a radome is to provide protection from the environment while having minimal impact on the RF performance of the antenna inside. Insertion loss is the main electrical concern for the radome. Insertion loss reduces the available signal, decreasing the ERP and G/T (the sensitivity, the ability of the antenna to receive a weak signal). Radomes can also increase antenna sidelobes, resulting in interference with other communication systems, and increasing the likelihood of signal detection and interception from unintended EW systems.

Radomes can also impact on the antenna polarization, depolarizing circularly polarized antennas, for example. Depolarization is generally very small for spherical radomes, however. Some other electrical effects of a radome on antenna performance include change in antenna beam width and shifting of the antenna boresight.

Water, rain, and snow can have an effect on radome performance as well. A thin sheet of water has a very high dielectric constant and loss tangent at microwave frequencies. This is exacerbated in sea applications of radomes because sea water is salty, producing highly conductive surfaces. Nonhydrophobic surfaces cause water to stick to the radome creating a thin film that serves as a shield to RF transmission resulting in significant signal attenuation [5]. Hydrophobic surfaces cause water to bead up and run off as shown in Figure 22.9 [1]. Even in high rain conditions, a radome with a hydrophobic surface has little additional attenuation [6].

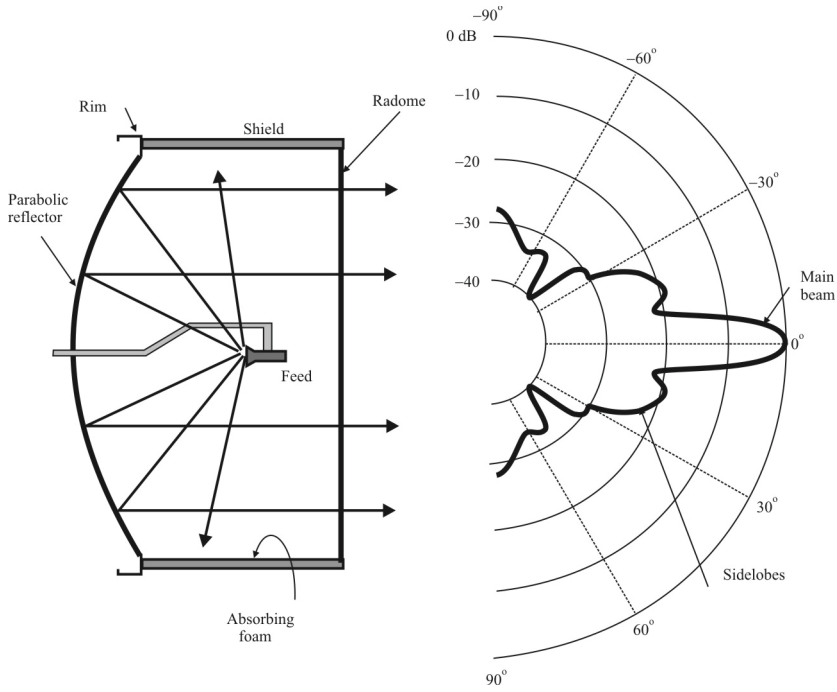


Figure 22.10 A typical microwave antenna system, with parabolic reflector, shield, and radome, will yield a strong main beam signal and lower level sidelobe signals.

Consider the microwave antenna shown in Figure 22.10 with its radiation pattern plotted in polar form on the right. It is typified by a main beam and several sidelobes. Figure 22.11 shows a typical radiation pattern of a microwave antenna in rectangular form, with the main beam at 0^o and sidelobes that are significant to about $\pm 90^o$ from the main beam. It is these sidelobes that can cause interference with adjacent point-to-point links. They also increase the EW vulnerability of the antenna, in some cases facilitating ES and EA from off-axis conditions.

Figure 22.12 shows the reflection characteristics of a plane-wall radome for different materials. Each of these materials is characterized by a relative dielectric constant of $\varepsilon_r = 2$; however, each material has a different loss parameter, $\tan \delta$, ranging from 0.0018 (low loss) to 18 (high loss). We see that for low-loss materials there exists two distinct minimum values of the reflection coefficient, for which a radome will allow maximum transmission of incident power. These correspond to design values where the ratio of radome wall thickness (d) to microwave wavelength in the sheet (λ) is close to either zero or 0.5.

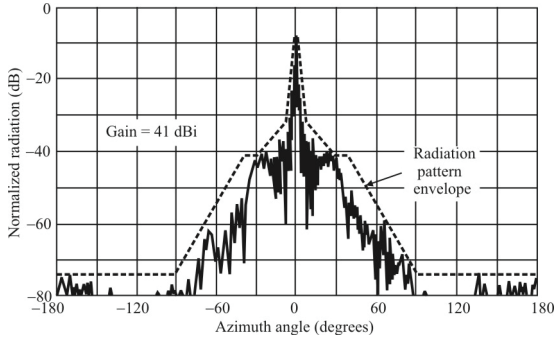


Figure 22.11 This radiation pattern is typical for a microwave antenna, with sidelobe signals being generated at significant levels at azimuth angles out to ±90°.

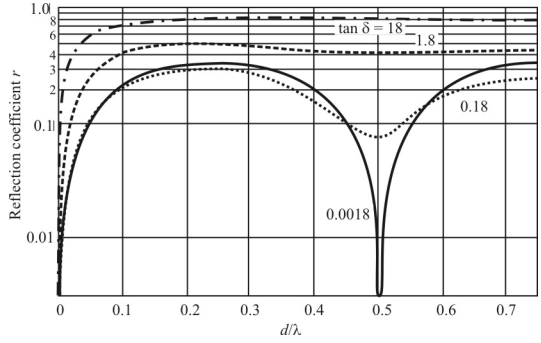


Figure 22.12 Reflection characteristics at a plane-wall radome as a function of sheet thickness d. $\varepsilon_r = 2$, $\tan\delta = 0.0018$ (low loss), $\tan\delta = 18$ (high loss).

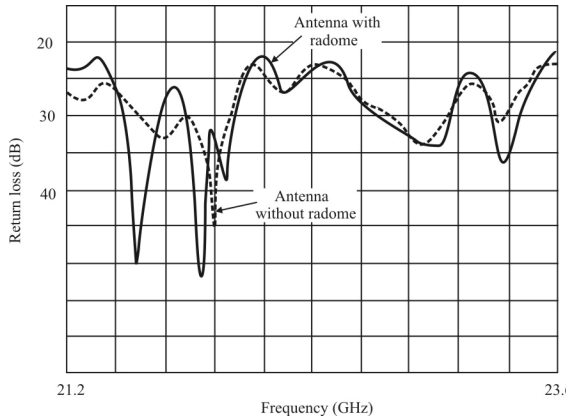


Figure 22.13 A microwave antenna with and without a radome evaluated at a center frequency of 22.4 GHz.

The first case of $d/\lambda \sim 0$ is approximated with $d < \lambda/10$, and leads to flexible radome materials with typical thicknesses of 0.4 to 0.6 mm—essentially as thin as practical. Flexible radomes are commonly used for larger antennas (greater than 4 feet), to avoid the bulk and weight of solid radomes.

The second case of $d/\lambda \sim 0.5$ is more complex and leads to solid radomes, which are more economical to produce at smaller sizes (less than 6 feet). The practical implication of $d \sim \lambda/2$ is that the thickness of solid radomes is always dependent on the wavelength (hence, frequency) of the application. Assuming a dielectric constant between 2.5 and 3, typical solid radome thicknesses would be 6 mm (14 GHz), 4 mm (22 GHz), and 2.4 mm (38 GHz).

This effect of frequency on solid radome design is illustrated in Figure 22.13, which shows return-loss measurements over a range of frequencies for an antenna designed to operate at 23 GHz. The minimum difference—corresponding to the minimum impact of the radome at $d \sim \lambda/2$—occurs at a frequency of about 22.6 GHz.

If a radome of incorrect thickness is used, the transmitted power will be reduced, and consequently the antenna gain also reduced. Greater transmit power would then be required to achieve the desired ERP, resulting in a corresponding increase of sidelobe radiation.

Figure 22.13 is valid for the "ideal" case, where wavefronts hit the wall orthogonally. When the EM wave does not have this ideal orientation, the effective wave path through the radome material is longer as it hits the wall

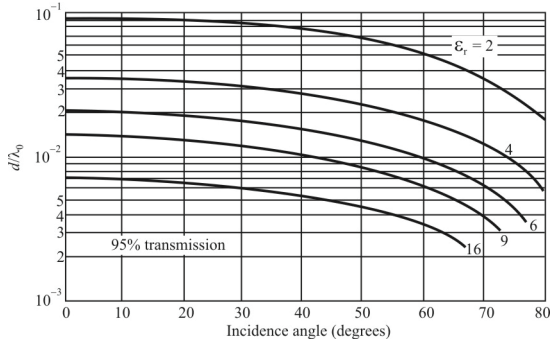


Figure 22.14 Flexible radome design depends on the incidence angle, 95% power transmission through the radome material.

obliquely, and the optimum thickness is now also dependent on the angle of incidence (θ), measured as the deviation from normal.

In practice, however, angles of incidence up to 20° have negligible effect on the optimum radome thickness. This is illustrated for flexible radome materials in Figure 22.14, which shows the relationship between angle of incidence (θ) and d/λ_o (where λ_o is the free-space wavelength), for achieving 95% power transmission through materials with different dielectric constants. For values of θ up to 20° the optimum thickness is barely impacted—particularly for low-loss materials.

A similar relationship holds for solid radomes; these relationships have been exploited in a practical fashion. A small degree of tilt of the main beam—around 5°—actually *improves* the performance of the antenna by directing spurious reflections within the antenna away from the microwave feed system.

22.8 Concluding Remarks

No matter how carefully a microwave antenna radome is designed, the potential increase in sidelobes remains. This must be taken into account during other aspects of design and installation of the antenna to minimize interference.

It is to the advantage of the EW system, however, for the target antenna to have high sidelobes. The main beam of microwave antennas is typically quite narrow, and getting inside this beam can be problematic. High sidelobes allow EW systems to be offset from the mainbeam by a considerable amount. The only practical way to place the EW system within the mainbeam is to fly the EW system on an airborne platform in many cases.

References

[1] Griffiths, L., "A Fundamental and Technical Review of Radomes," www.mpdigest.com, May 2008.
[2] Kozakoff, D. J., *Analysis of Radome-Enclosed Antennas*, Norwood, MA: Artech House, Boston, 1997.
[3] Skolnik, M., *Radar Handbook*, 2nd ed., New York: McGraw-Hill, 1990, pp. 6.44–6.45.
[4] U.S. Department of Defense, "MIL-R-7705B—General Specification for Radomes," U.S. Government Printing Office, 1975, p. 2.
[5] Anderson, I., "Measurements of 20-GHz Transmission Through a Radome in Rain," *IEEE Transactions on Antennas and Propagation*, September 1975, pp. 619–622.
[6] Dietrich, F. J., and D. B. West, "An Experimental Radome Panel Evaluation," *IEEE Transactions on Antennas and Propagation*, November 1988, pp. 1566–1570.

Appendix A

RF Amplifiers

The taxonomy of RF signal amplifiers can be divided into two broad classes: analog and digital. In this appendix we discuss the major forms of these RF amplifiers [1–14].

Although not officially one of the amplifier "classes," push-pull amplifiers are often used for high power RF applications, including those for EW. Therefore, we include a section on such amplifiers, including material on the wideband RF transformers necessary for their operation.

A.1 Analog Amplifiers

An analog amplifier class is determined by how it is biased. RF amplifier classes A, B, C, and A/B are analog amplifier classifications that are determined by the number of degrees of current flow during each 360^{o} RF cycle over which the plate, collector, or drain current flows.

A simplified configuration of analog amplifier is illustrated in Figure A.1. All of the analog amplifiers can be configured this way; the difference is in where they are biased, as we will see. R_1, R_2, and R_L are adjusted to set this bias point. Although we show a MOSFET in Figure A.1, that is but one of many active devices that can amplify signals. We also show a common source architecture in Figure A.1 (common means that the terminal is common to both the input and output). Again, that is but one of at least three configurations available. The other two are common base and common collector. Each of these architectures has its own characteristics.

A.1.1 Class A Amplifiers

Class A amplifiers operate over a relatively small portion of a tube's plate current, a transistor's collector current, or an FET's drain current range. Continuous

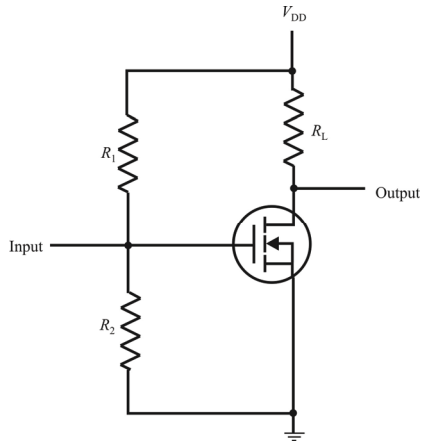


Figure A.1 One stage analog amplifier. R_1 and R_2 establish the bias condition for the active device—in this case the quiescent voltage on the gate of the MOSFET. The output is developed across R_L.

current flows throughout the entire 360° of each RF cycle. Their efficiency in converting DC source power to RF output power is poor. DC source power that is not converted to RF output power is dissipated as heat. However, in compensation, class A amplifiers have a higher linearity that is responsible for *lower output-signal distortion* than any other amplifier class. They are most commonly used in small-signal applications where linearity is more important than power efficiency, but also are sometimes used in large-signal applications where the need for high linearity outweighs cost and heat disadvantages associated with poor power efficiency.

A.1.2 Class B Amplifiers

Class B amplifiers have their tube control grid, transistor base, or FET gate biased near plate, collector, or drain current cutoff, causing the plate, or collector, or drain current to flow only during one-half (180°) of each RF cycle. That causes the DC source power to RF output power efficiency to be much higher than with class A amplifiers, but at the cost of severe output cycle waveform distortion. If the input is a sine wave, for example, only one-half of that sine wave will appear in the output current waveform.

That waveform distortion is greatly reduced in practical designs by using relatively high-Q resonant output "tank" circuits to reconstruct full RF cycles, but the signal is at higher power levels than the input signal. Low sine-wave distortion results only if the Q the tank circuit is sufficiently high. Unless the Q is infinite, which, of course, it never can be, the amplitude of one half-cycle will be larger

than the other, which is a manifestation of harmonic energy content. Of importance to EW amplifier designs, coupling an antenna system too tightly to the resonant output tank circuit of an amplifier will lower its Q, increasing the percentage of harmonic content in the output.

Another effective method commonly used to reduce class B RF amplifier output waveform distortion is to employ two amplifiers operating in "push-pull" such that one conducts on half cycles where the other is in plate, collector, or drain current cutoff. Oscillatory tank circuits are still used in the outputs of class B push-pull amplifiers to smooth switching transitions from the conduction of one amplifier to the other, and to correct other nonlinearities, but lower-Q tank circuits can be used for given percentages of harmonic content in the output.

A.1.3 Class A/B Amplifiers

As the designation suggests, class A/B amplifiers are compromises between class A and class B operation. They are biased so the plate, collector, or drain current flows less than 360^{o}, but more than 180^{o}, of each RF cycle. Any bias point between those limits can be used, which provides a continuous selection range reaching from low distortion, low efficiency on one extreme to higher distortion, higher efficiency on the other.

A.1.4 Class C Amplifiers

Class C amplifiers are biased well beyond cutoff, so that the plate, collector, or drain current flows less than 180^{o} of each RF cycle. That provides even higher power efficiency than class B operation, but with the penalty of even higher input-to-output nonlinearity, making use of relatively high-Q resonant output tank circuits to restore complete RF sine-wave cycles. High-amplifying nonlinearity makes them unsuitable to amplify signals where the amplitude is used to convey the information, such as AM, DSB, SSB, PAM, or QAM signals.

However, most class C amplifiers can be amplitude-modulated with acceptably low distortion by varying the plate, collector, or drain voltage, because they generally are operated in the region of the plate, collector, or drain saturation so that the RF output voltage is dependent upon the instantaneous DC plate, collector, or drain voltage. They also are commonly used in *continuous wave* (CW) and *frequency shift keyed* (FSK) applications and in *phase modulated* (PM) and *frequency modulated* (FM) transmitter applications where signal amplitudes remain constant, and the information is conveyed in the instantaneous phase of the carrier signal.

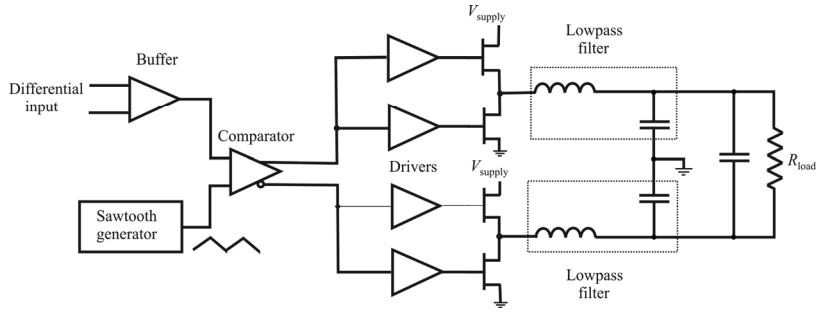


Figure A.2 Block diagram of a class D amplifier. Switching amplifiers sequentially saturate and turn off power transistors in the output stage using pulse width modulation.

A.2 Switching Amplifier Architectures

There are some additional amplifier classes that are used as RF amplifiers that rely on the ability of the active devices to rapidly switch on and off. These classes are called switching architectures and are introduced in this section.

A.2.1 Architectures

This is a class of amplifier architectures that switch rapidly to perform the amplification. Discussed here are classes D, E, F, and S architectures. The most significant advantage of these architectures is their efficiency. Generally, they are much more efficient than the analog architectures discussed above.

A.2.2 Class D Amplifiers

A class D amplifier is substantially different from those of the other classes mentioned previously. An amplifier operating in class D rapidly turns the power transistors on and off in the output stage, producing a digital output of 1s and 0s. With *pulse width modulation* (PWM), the longer the transistor is on, the greater the average output power.

A block diagram of a class D amplifier is shown in Figure A.2. The differential input of the signal to be amplified is input to one side of a comparator. A sawtooth waveform is applied to the other input as shown. When the amplitude of the input signal exceeds that of the sawtooth, the output of the comparator changes state to a logical one in this case. As seen in Figure A.3, which displays the time waveforms involved, the lower the amplitude of the input signal, the more of the sawtooth is included in that which is less than the input signal, causing the output to remain at a high state longer. In this way the circuit adjusts the width of

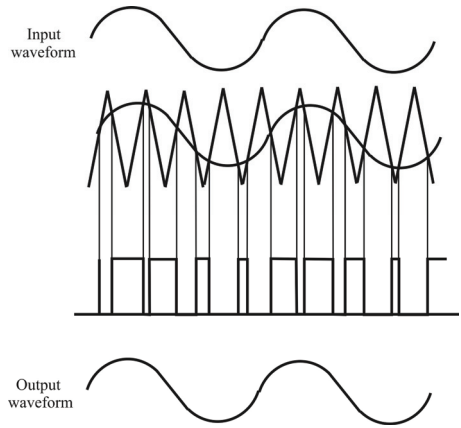


Figure A.3 Waveforms within a class D amplifier. Every time the sawtooth waveform goes above the input waveform, the output of the comparator goes to a high state and vice versa. The comparator output, through drivers, turns the output transistors on and off.

the output of the comparator as a function of the amplitude of the input signal. This PWM waveform then drives the high-power output stage that switches on and off according to this signal. The wider the pulse, the longer the power stage stays on and the more of the prime power is fed to the output lowpass filter. The output, being a square wave, would generate spurious signals if transmitted in that form. This output is lowpass-filtered prior to transmission to extract only the fundamental frequency.

The primary advantage of class D amplifiers is their efficiency compared to the other classes. Efficiencies greater than 80% are possible. However, because the transistors must be switched much faster (say, by a factor of 10) than the highest frequency in the signal, the use of class D amplifiers is limited to low frequencies with modern transistors. The most common usage of class D amplifiers is in audio amplifiers for home or automobile stereos; however, RF devices are emerging that allow operation of this architecture into the UHF range (300 MHz to 3 GHz).

A.2.3 Class E Amplifiers

Class E amplifiers are perhaps the most efficient type of amplifier available. A simplified schematic of a class E amplifier is shown in Figure A.4. The transistor here is not really amplifying at all in the normal sense, but is switching on and off. Shunt capacitor C_1 and the tuned circuit consisting of C_2, L, and R_{load} are chosen so that the phasing of the voltage across C_1 and the current through the transistor

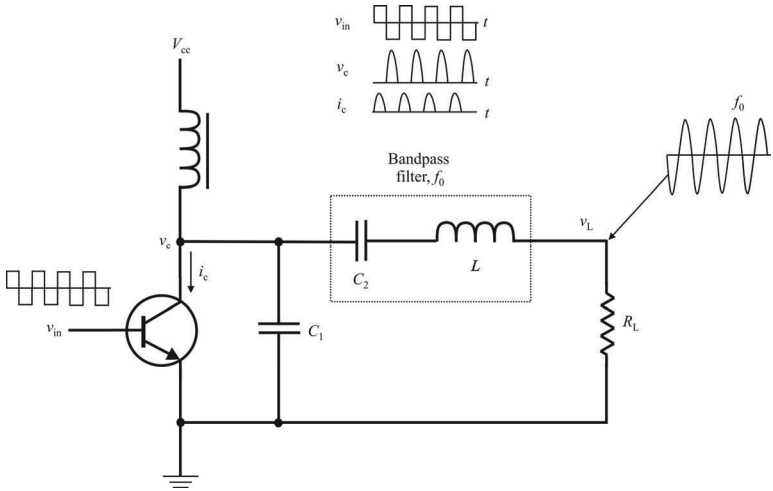


Figure A.4 Architecture of a class E amplifier. The transistor is switched with a square input waveform, while the output is tuned or lowpass, thus eliminating the spurious response.

are as close to π radians out of phase as possible. If they are exactly π radians out of phase, then the transistor dissipates no power. All of the power is delivered to the load except that lost in the other nonideal components. (The other inductor is called an RF choke and is simply there for bias purposes—it keeps the RF signal out of the power supply circuit.)

The simplified waveforms of this circuit are shown in Figure A.4. For simplicity, it is assumed that the input signal to the amplifier is a 50% duty cycle square wave, although other waveforms are possible depending on the desired effect. As the transistor is switched on and off by the input signal, the current and voltage vary as shown. The RLC circuit on the output of the transistor prevents these output waveforms from instantaneously changing the way the input signal is changing. They must change more slowly, thus forming a continuous waveform.

Class E amplifier configurations have limited utility as the power amplifier in EW systems because they are obviously tuned amplifiers. EW systems typically require amplifiers that can quickly (instantaneously) cover a broad frequency band. A class E amplifier would need its output circuit changed quickly to function as a broadband amplifier.

A.2.4 Classes F and F^{-1} Power Amplifiers

A class-F PA boosts both efficiency and output power by using harmonic resonators in the output network to shape the drain waveforms such that the load appears to be a short at even harmonics and an open at odd harmonics [15]. The

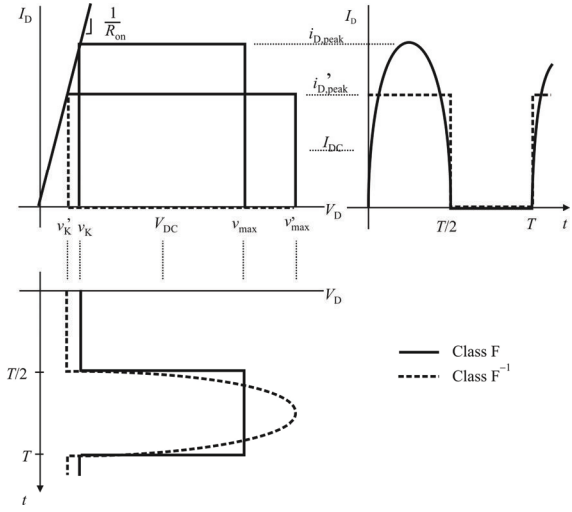


Figure A.5 Class F and F^{-1} conduction characteristics.

drain voltage waveform includes one or more odd harmonics and approximates a square wave, while the drain current waveform includes even harmonics and approximates a half sine wave. Because there is no overlap between the drain voltage and current, an ideal efficiency of 100% is achieved. In the case of its dual, that is, class F^{-1}, the shapes of drain voltage and current are swapped, that is, the drain voltage waveform is close to a half sine wave, while the drain current is close to a square wave [16]. The characteristics of class F and class F^{-1} are illustrated in Figure A.5 [17], while the time waveforms are shown in Figure A.6 [18]. In order to obtain the perfectly square waveforms, an infinite number of harmonics are required. The theoretical efficiencies achievable with class F and class F^{-1} configurations are a function of the number of harmonics included, some of which are listed in Table A.1 [15].

The output network of a classical class F PA is shown in Figure A.7, which consists of a quarter-wavelength transmission line and a parallel-tuned inductance/capacitance (*LC*) output network [19]. The output tank is tuned to resonance at the fundamental frequency f_0. At f_0, the drain sees a pure resistive load of R_L, since the output tank is an open circuit, and the quarter-wave transformer transforms the 50Ω load to R_L at the drain. The load resistance R_L is chosen as the optimum load impedance for the PA, which can lead to maximum excursions of drain current and voltages. At even harmonics $2nf_0$ ($n = 1, 2, \cdots$), the output tank becomes a short, and the transmission line appears to be a length of

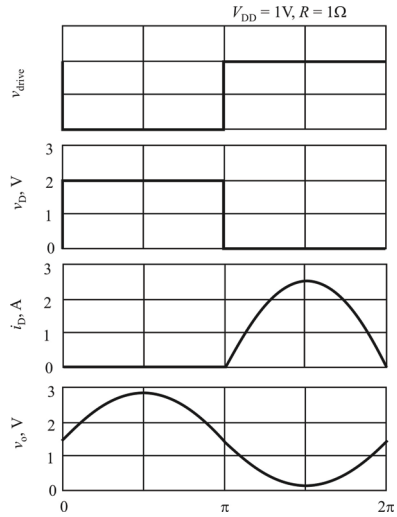


Figure A.6 Class F waveforms. For this example an infinite number of harmonics is included to produce the square-wave outputs from the FET.

Table A.1 Maximum Efficiency of Class F PAs

	$n = 1$	$n = 3$	$n = 5$	$n = \infty$
$m = 1$	50%	57.7%	60.3%	63.7%
$m = 2$	70.7%	81.7%	85.3%	90.0%
$m = 4$	75%	86.6%	90.5%	95.5%
$m = \infty$	78.5%	90.7%	94.8%	100%

m, n denote maximum order of harmonics in drain current and voltage, respectively.
Source: [15].

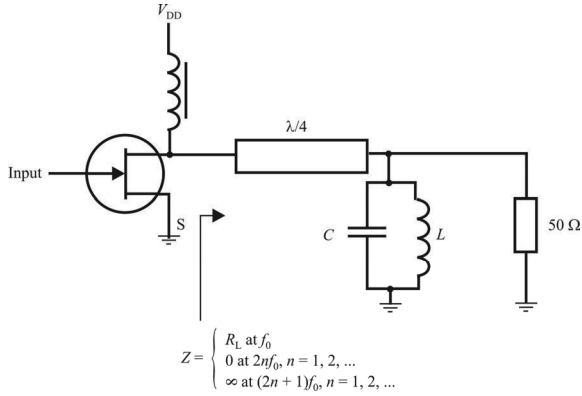


Figure A.7 Output network of a class F amplifier.

$n\lambda_g/2$ to the drain, where λ_g is the guided wavelength of the transmission line. Therefore, the drain sees a short at all even harmonics, which would result in a half-rectified sinusoid current output as desired. Conversely, at odd harmonics $(2n + 1) f_0$ ($n = 1, 2, \cdots$), the output tank also becomes a short, and the transmission line appears to be a length of $[(2n + 1)\lambda_g]/4$ to the drain. Due to the quarter-wavelength transformer, the drain sees an open circuit at odd harmonics. The drain voltage contains only the odd harmonics, while the drain current includes only the even harmonics. Thus, the harmonic impedance is zero at even harmonics and infinite at odd harmonics.

In the case of considering only the second and third harmonics (i.e., third-harmonic peaking), the maximum efficiency reaches 81.7%. An output network that includes a third-harmonic peaking circuit is shown in Figure A.8. The quarter-wavelength transformer in Figure A.7 is removed, and, instead, a parallel resonant tank at $3f_0$ is added in series with the drain output, which provides a short at $2f_0$ and an open at $3f_0$. Another parallel resonant tank at f_0 is added in parallel with the load resistance, the same as in Figure A.7, which ensures optimum load at f_0. Here R_L is the optimum load required by the drain port.

Due to the bandwidth requirements of EW PAs, several matching networks would be needed, each switched in and out with pin-diode switches (see Appendix B). Therefore, if the output tuning can provide the required filtering, the input matching networks shown in Figure A.9 [15] would probably not be considered in the tuning process, as it would double the number of networks required.

A single-stage class F amplifier is shown in Figure A.9(a) where the input is tuned at f_0 and the output is tuned for optimum load matching at f_0, and for harmonics at $2f_0$ and $3f_0$. It has been widely accepted that class F PA designs should consider both input- and output-harmonics tuning, as shown in Figure

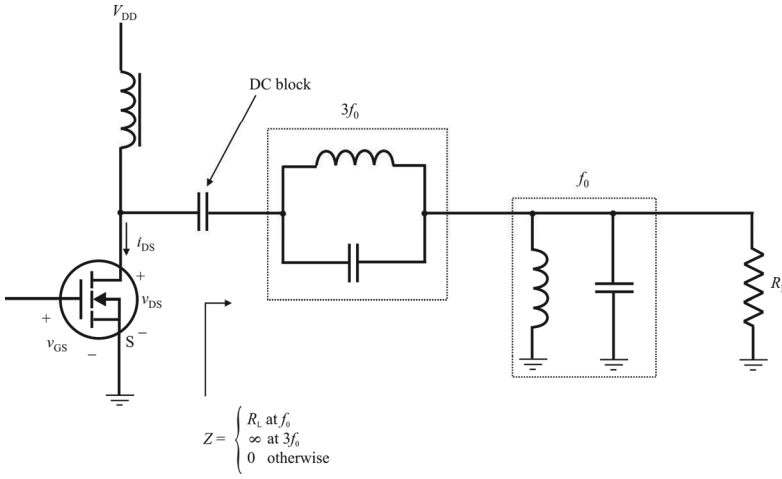


Figure A.8 Third-harmonic peaking architecture.

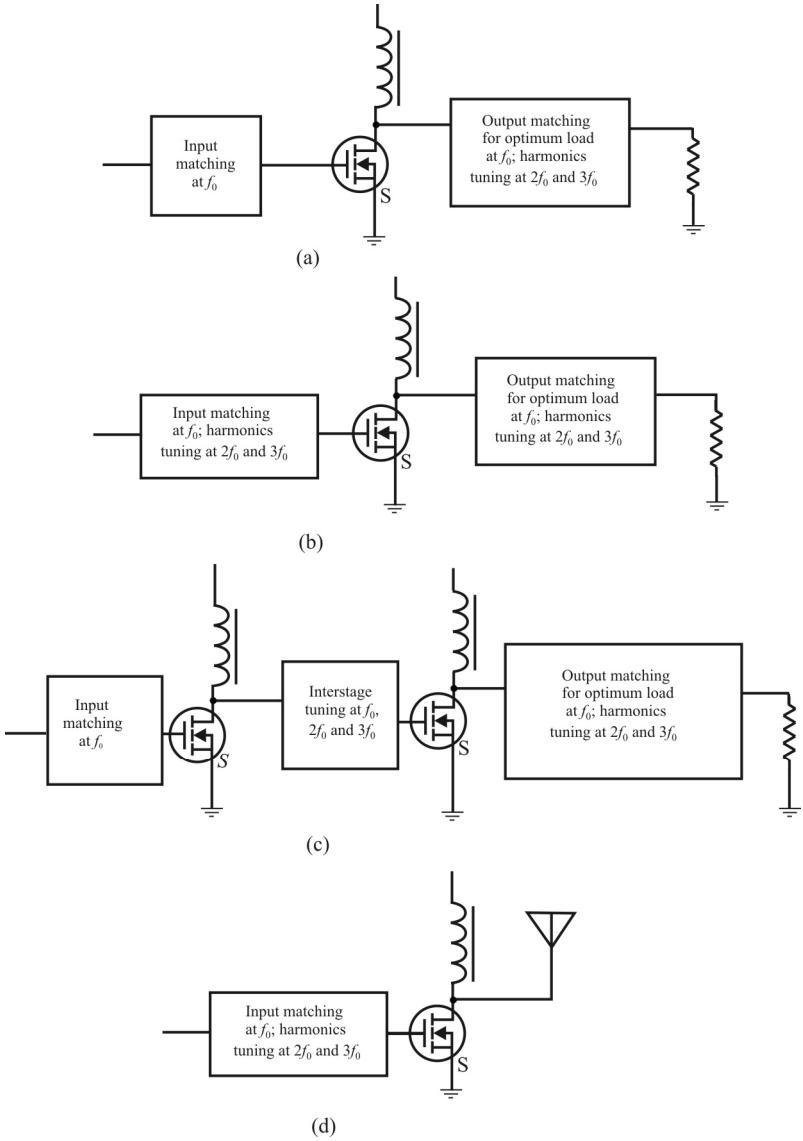


Figure A.9(a–d) Class F PA matching strategies: (a) output harmonics tuning strategy; (b) both input and output harmonics tuning; (c) multistage class F PAs; and (d) active-antenna approach.

A.9(b). Figure A.9(c) shows the two-stage amplifier, which includes a driver/waveform-shaping stage and a final stage. The first stage will serve two functions: to provide enough power at the input of the final stage and to shape the input voltage waveform for the final amplifier stage, so that the input voltage waveform is a half sinusoid at the input of the final stage. The waveform shaping could be achieved by using appropriate proportions of the fundamental component and the second harmonic, as in [20]. This is also called the harmonic-controlled amplifier, which has the advantages of the high gain of class A and the high efficiency of class F, due to the elimination of the negative V_{gs} at the input of the final stage. The problem is the complexity of its design and circuit layout, which may also introduce some losses. An alternative way of realizing the diver/waveform shaping is to use the class F^{-1} PA at the first stage and the class F PA at the final stage. The class F^{-1} PA produces the output voltage as a half sinusoid, which feeds directly into the input of the final stage.

Figure A.9(d) shows the active-antenna approach for class F PA designs [21]. The principle is to treat the antenna, the active device, and the input network as a single entity. In addition to its original function as a radiator, the antenna is used to achieve both the harmonics tuning and the filtering without using an additional output matching network between the antenna and the drain output. This leads to a functional compact design, eliminating the cable and matching network loss, thus improving the overall system efficiency and output power. The active antenna approach is very promising, enabling high efficiency, compact RF output stages.

A.2.5 Class S Amplifier

Class S amplifiers operate similarly to class D described above in that the signal to be amplified is initially in digital form and is converted to analog. In its digital state, it is fed to a delta-sigma modulator.

Delta-sigma modulators were discovered in the field of A/D and D/A converter design. One of the difficulties with D/A converters is the precision required in some of the components to convert a digital signal to an analog form.

The goal was to remove all analog components from the D/A design, and the delta-sigma modulator was discovered. A simplified block diagram of perhaps the simplest of these devices is shown in Figure A.10 [22]. The current state of the modulator, B, is subtracted from the level of the digital input signal forming $A - B$. B is a function of the output of the 1-bit quantizer. If the quantizer output is 0, then $B = 0$. If the quantizer output is 1, then $B = 2^M - 1$, where M is the number of bits in the modulator. $A - B$ is called the error signal. The second adder adds this error to the accumulated error as stored in the latch. On the next clock cycle, this updated accumulated error is clocked into the latch, forming the next state of the modulator. The 1-bit quantizer outputs a zero if the accumulated error is less than

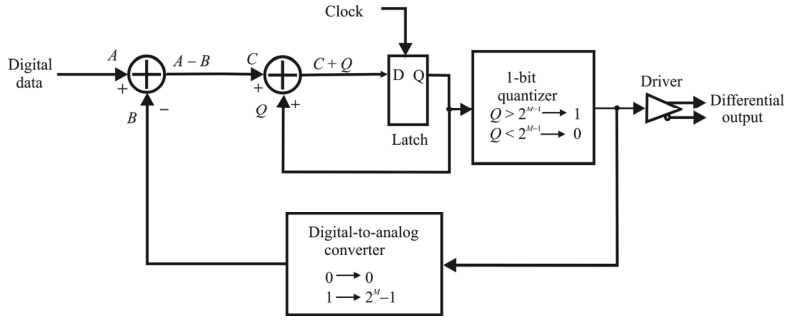


Figure A.10 Block diagram of a delta-sigma modulator.

half the maximum value allowed ($= 2^M / 2 = 2^{M-1}$), and a 1 if it is greater than half the maximum.

The net effect of the modulator is shown in Figure A.11. When the input is largely positive, the output of the quantizer is mostly 1, whereas when the input amplitude is small, this output is mostly 0. Another way of saying this is the output remains in a 1-state, on average, longer when the input signal level is greater. The output of the quantizer is periodic.

Although overall operation produces results that appear to be PWM, as with class D amplifiers, in fact, a delta-sigma modulator produces *pulse density modulation* (PDM). The output width of the pulses is not changed per se according to the amplitude of the input signal, but the number of pulses in the periodic cycle is. The differential outputs are then used to drive the output circuit as shown above for the class D amplifier. This is shown in Figure A.12. In this case, there is no need for the sawtooth generator or the comparator, and a bandpass filter has replaced the lowpass filter.

It is desirable to keep the differences between one sample and the next relatively small. In order to do this, the sample clock frequency must be much higher than the highest input frequency. This is called oversampling. Nevertheless,

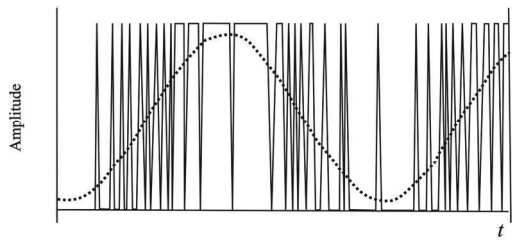


Figure A.11 Input/output of the delta-sigma modulator driving a class S output stage.

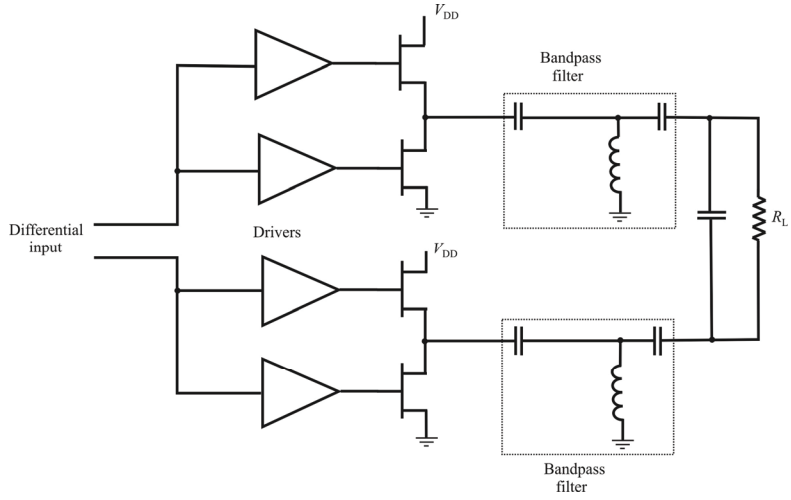


Figure A.12 Output stage of a class S amplifier. The input takes the output of the delta-sigma modulator and turns on and off the output power transistors using PWM. The output filters can be bandpass, as shown here, or lowpass for broadband operation.

much higher frequencies of operation are possible than class D, achieving cutoff frequencies in the gigahertz range. In addition, there are other forms of delta-sigma modulators. In particular, there are higher-order modulators. The order of a delta-sigma modulator is the number of delay stages it has. They tend to calculate the error better and therefore replicate the input signal better. The first-order modulator shown in Figure A.10 is one of the simplest to illustrate their basic operation.

A.3 Load Lines

Each of the amplifier classes has its own characteristic load line. These are illustrated in Figure A.13. (Although the load lines in Figure A.13 do not extend to the ordinate axis, in practice they do.)

The class A load line extends to $2V_{DD}$ on the abscissa (V_{DS}) axis. Thus, the linear output range extends to twice the supply voltage level. On the other hand, the linear portion of the class B load line terminates on the abscissa at V_{DD} and is zero after that. Current only flows for half of the input cycle. The class A/B load lines are between these two. The linear part of the class C load lines ends at a

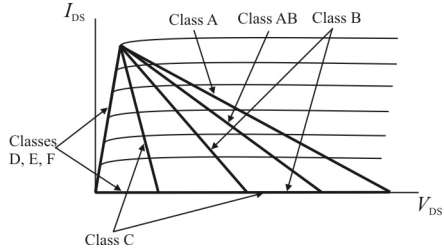


Figure A.13 Load lines of the amplifier classes.

voltage less than V_{DD}, and therefore output current conduction only occurs for less than half of the input cycle.

The load lines for the switching classes (D, E, and F) are different from those for the analog classes. Since the active device is either on or off for these classes, the load lines are either represented by an open circuit (zero current, any voltage) that corresponds to the abscissa axis or a short circuit (ideally) when the switch is on. This latter condition corresponds to the ordinate axis (ideally) or the linear region (nonsaturated) curve (practically) shown in Figure A.13. This limit is due to the fact that the on resistance of a practical device is not zero, but corresponds to the value of the slope of the linear region curve.

A.4 Push-Pull Amplifiers

Fundamentally, a push-pull amplifier architecture circuit uses a pair of effectively separate transistors operating 180^{o} out of phase with one another. If good amplitude and phase balance is maintained between the signals in each half of the device, then an RF ground will exist at the midpoint. This approach leads to several advantages over single-ended designs [23]:

- Input and output impedances doubled (easier to match);
- Reduced even-order harmonics (less out-of-band interference generated);
- Increased output power;
- Higher bandwidths possible;
- Reduced effect of common lead inductance.

Perhaps the most significant disadvantages are the need for differential RF excitation and the fact that excellent symmetry is required in both the matching circuit and the devices themselves. A generic push-pull circuit configuration is shown in Figure A.14 [23].

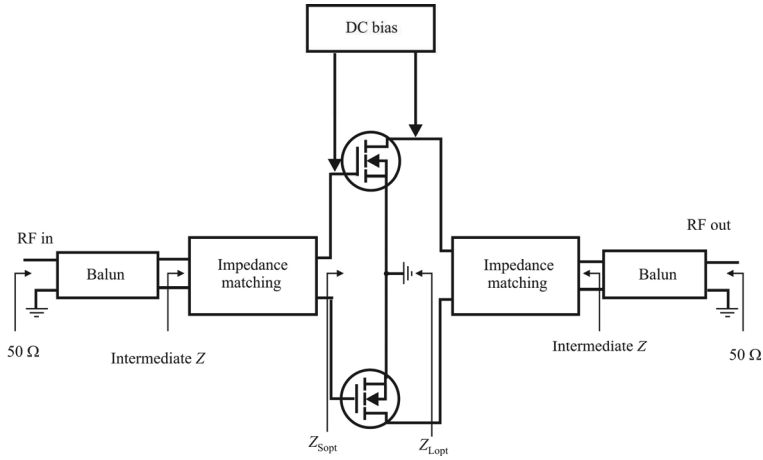


Figure A.14 Generic push-pull amplifier architecture. (Enhancement mode FETs are shown, but any active amplifying device works with suitable bias.) Note that the input feeds a balun when the actual input is unbalanced and the output is balanced. This is a balun used in reverse since it is a reciprocal device.

The RF signal at the input of push-pull amplifiers is split at the input (along with the reciprocal operation, that is, the need to recombine the output). Elements that provide this function are referred to as *baluns* (balanced to unbalanced). The ideal balun would split the signal into halves of equal amplitude without loss, along with providing a 180^o differential phase shift across all frequencies. Ideal baluns do not actually exist, but their operation can be approximated in several ways with excellent performance over wide bandwidths. Some of the methods are:

- Conventional coil transformers;
- Transmission line transformers;
- Microstrip structures (Wilkinson divider; line and ring hybrids).

The choice largely depends on frequency. For HF (up to 30 MHz) magnetically coupled coil transformers perform well. Above 30 MHz leakage inductance and parasitic capacitance degrade their performance, making them unsuitable. VHF and UHF applications (30 MHz–1 GHz) most commonly employ transmission line balun structures, typically built from coaxial cable or twisted pairs. Microstrip structures such as Wilkinson dividers also offer very good performance, but these are limited by their physical size (on the order of $\lambda/2$) and are therefore only practical at 500 MHz (λ = 60 cm) and higher frequencies. Substrates with high dielectric constants can mitigate this problem to some extent.

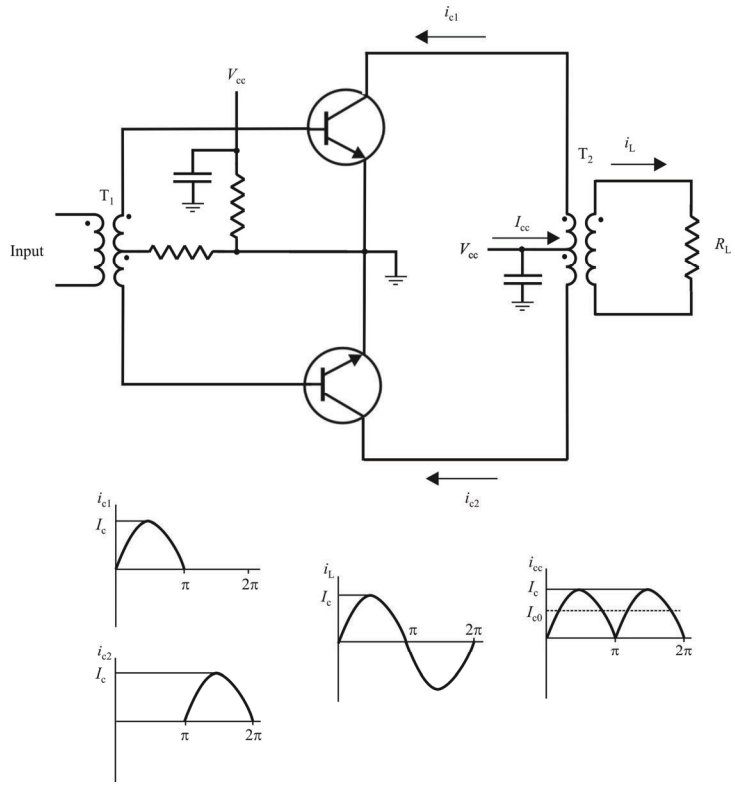


Figure A.15 Push-pull architecture.

A.4.1 With Transformers

A type of push-pull arrangement for PAs is shown in Figure A.15. As the input signal oscillates between positive and negative voltages, the transistors conduct alternately. It has the advantage that the configuration of the circuit facilitates total suppression of the even harmonics in the output (assuming ideal components). Thus, as the transistors are pushed to their linear limits and beyond, the only significant harmonic in the output is the third, which can frequently be removed by simple lowpass filters. The suppression of even-order harmonics is due to the way the output of the two channels combine in the output transformer.

The input and output impedances of push-pull PAs are higher than other types of amplifiers. For ideal class B, the duty cycle of each device is 50%. With a sinusoidal voltage as input, the first transistor collector current is

$$i_{c1} = \begin{cases} I_c \sin\theta, & 0 \le \theta < \pi \\ 0, & \pi \le \theta < 2\pi \end{cases} \tag{A.1}$$

while the second transistor collector current is

$$i_{c2} = \begin{cases} 0, & 0 \le \theta < \pi \\ -I_c \sin\theta, & \pi \le \theta < 2\pi \end{cases} \tag{A.2}$$

At the output of transformer T_2, the current through the load is

$$i_L(\theta) = i_{c1}(\theta) - i_{c2}(\theta) = I_c \sin(\theta) \tag{A.3}$$

The current flowing in the center tap of the primary winding of T_2 is

$$i_{cc}(\theta) = i_{c1}(\theta) + i_{c2}(\theta) = I_c \left| \sin(\theta) \right| \tag{A.4}$$

The total DC collector current is

$$I_{co} = \frac{1}{2\pi} \int_0^{2\pi} i_{cc}(\theta) d\theta = \frac{2}{\pi} I_c \tag{A.5}$$

Assuming zero saturation resistance when the collector voltage amplitude $V_c = V_{cc}$ and equal winding turns at the collectors so that $V_L = V_c$, DC and fundamental output powers are

$$P_0 = \frac{2}{\pi} I_c V_{cc} \tag{A.6}$$

$$P_{out} = \frac{1}{2} I_c V_{cc} \tag{A.7}$$

In ideal conditions, even-order harmonics are canceled because they are out of phase and combined in the center tap of the primary winding of T_2.

To eliminate losses, it is necessary to connect bypass capacitance to this center point. With a 50% duty cycle, third- and higher-order odd harmonics do not exist; ideally a sinusoidal signal appears in the load

$$v_L(\theta) = I_c R_L \sin(\theta) = V_L \sin(\theta) \tag{A.8}$$

Maximum collector efficiency that can be achieved in class B operation is given by

$$\eta = \frac{P_{out}}{P_0} = \frac{\pi}{4} \approx 78.5\% \tag{A.9}$$

Class E Push-Pull Architecture

The high efficiency of class E amplifier configurations can be exploited in push-pull architectures. One such configuration is illustrated in Figure A.16. The transistors are driven 180^o out of phase so when one is off the other is on, providing a current path from the DC power supply through the transformer to ground. The voltage waveforms at the collectors of the transistors are shown in Figure A.16, as well as the voltage waveform across the load. The impedance seen from the transistor looking into the transformer is given by

$$Z = \frac{1}{2}\frac{m^2}{n^2} Z_0 \tag{A.10}$$

A.4.2 Wideband Transformers

EW systems typically must operate over considerable portions of the RF spectrum. The transformers used in push-pull amplifiers therefore must have adequate bandwidth to match that required of the system. An additional requirement is low insertion loss. In this section, we discuss two primary types of wideband transformers: the conventional coupled coil transformer and the transmission line transformer [24].

The power handling capability of a transformer is largely dictated by the size of the wire used in its construction. The power handling capability and some other characteristics of wire are indicated in Table A.2 [25].

A.4.2.1 Coupled Coil (Conventional) Transformers

Conventional transformers operate by coupling signals between primary and secondary windings through a suitable magnetic material. For an ideal transformer, the voltage per turn is the same on each side, which is based on

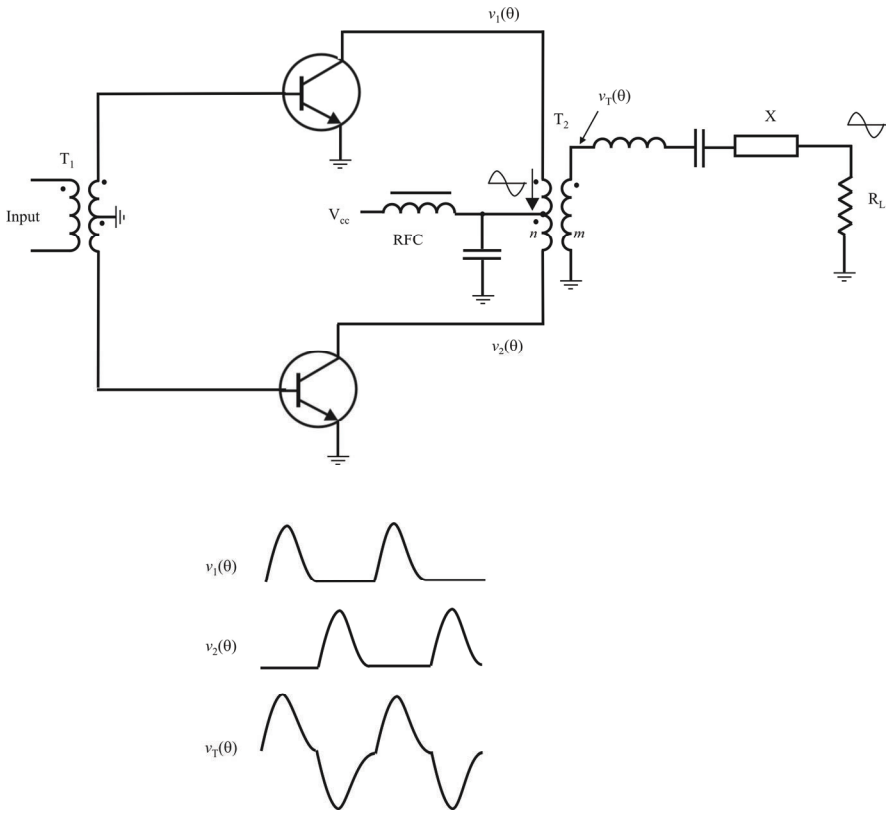


Figure A.16 Push-pull class E architecture.

Table A.2 Characteristics of Wire

AWG	Conductor Diameter (inches)	Ohms per 1,000 ft	Maximum Current for Chassis Wiring (A)	Maximum Current for Power Transmission (A)	Maximum Frequency for 100% Skin Depth for Solid Conductor Copper
10	0.1019	0.9989	55	15	2,600 Hz
11	0.0907	1.26	47	12	3,200
12	0.0808	1.588	41	9.3	4,150
13	0.072	2.003	35	7.4	5,300
14	0.0641	2.525	32	5.9	6,700
15	0.0571	3.184	28	4.7	8,250
16	0.508	4.016	22	3.7	11 kHz
17	0.0453	5.064	19	2.9	13
18	0.0403	6.385	16	2.3	17
19	0.0359	8.051	14	1.8	21
20	0.032	10.15	11	1.5	27
21	0.0285	12.8	9	1.2	33
22	0.0254	16.14	7	0.92	42
23	0.0226	20.36	4.7	0.729	53
24	0.0201	25.67	3.5	0.577	68
25	0.0179	32.37	2.7	0.457	85
26	0.0159	40.81	2.2	0.361	107
27	0.0142	51.47	1.7	0.288	130
28	0.0126	64.9	1.4	0.226	170
29	0.0113	81.83	1.2	0.182	210
30	0.01	103.2	0.86	0.142	270
31	0.0089	130.1	0.7	0.113	340
32	0.008	164.1	0.53	0.091	430
33	0.0071	206.9	0.43	0.072	540
34	0.0063	260.9	0.33	0.056	690
35	0.0056	329	0.27	0.044	870
36	0.005	414.8	0.21	.0.35	1.100 MHz
37	0.0045	523.1	0.17	0.0289	1.350
38	0.004	659.6	0.13	0.0228	1.750
Metric 1	0.00394	670.2	0.126	0.0225	1.750
39	0.0035	831.8	0.11	0.0175	2.250
40	0.0031	1049	0.09	0.0137	2.900

Source: [25].

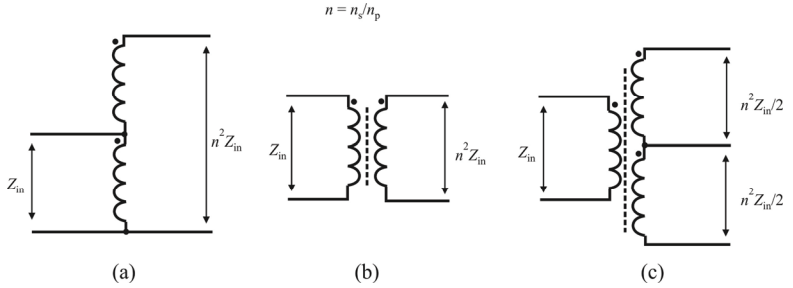


Figure A.17(a–c) Transformers: (a) autotransformer; (b) conventional transformer; and (c) center-tapped secondary.

Faraday's law (see Section 2.5.2). The relationship between impedances on the primary and secondary side is then given by

$$Z_{\text{in}} = \left(\frac{n_{\text{p}}}{n_{\text{s}}} \right)^2 Z_{\text{out}} \tag{A.11}$$

This implies that transformers are well suited to impedance transformations since they can be scaled by any factor, depending only on the turns ratio. When the parasitics are negligible (i.e. at relatively low frequencies, say, up to 30 MHz), coupled coil transformers are often used where wideband resistance transformation is required. Various winding topologies are possible and are chosen depending on whether DC isolation and/or balanced signals are required. Figure A.17 [23] shows some of the simplest configurations. The transformer in Figure A.17(a) has a tapped continuous winding and provides a DC short between input and output. Figure A.17(b) shows a more conventional transformer with DC isolation between primary and secondary windings. Such isolation is often required due to biasing reasons between two stages. Other arrangements can allow for balanced-unbalanced operation. The center tapped secondary transformer in Figure A.17(c) is a balanced signal splitter that also provides an $N^2 / 2$ impedance transformation. It also provided for DC isolation between the primary and secondary windings. Figure A.17(c) is referred to as a balanced-balanced converter.

Better coupling is provided by higher permeability ferrite materials, although it should be noted that due to the increased flux density, higher cores will saturate more easily than those with lower permeability. Permeability can, and usually does, also vary as a function of current level, frequency, and temperature.

Two-hole balun cores and toroids are the most commonly used ferrite structures. The choice of wire also affects performance and is essentially a trade-

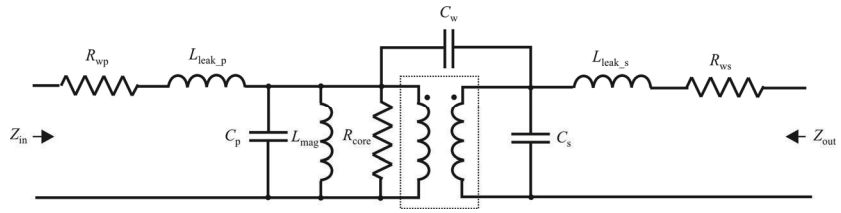


Figure A.18 Conventional transformer equivalent circuit.

off because heavier wire will increase the coupling, but it will also increase the parasitic winding capacitance. In practice, all transformers suffer from parasitic effects to some extent that cause their behavior to deviate from the ideal.

Figure A.18 [23] shows an equivalent circuit for modeling the transformer. The ideal n_p:n_s transformer is represented inside the dashed line. Series resistances R_{wp} and R_{ws} are the resistances of the primary and secondary windings, respectively. Although these are negligible at DC, the skin effect will increase the resistance in proportion to the frequency (at RF essentially all current is due to the skin effect; see Chapter 3). R_{core} models the loss of energy in the transformer's core due to eddy currents and hysteresis effects. Leakage inductances represented by L_{leak_p} and L_{leak_s} model the magnetic flux that passes outside of the core and, therefore, is not coupled between the input and output ports. L_{mag} limits the low frequency response of the transformer and has its physical origins in the finite magnetizing inductance in the coils. At RF and microwave frequencies, the capacitive effect between windings (C_w) and between the turns of each winding (C_p and C_s) also becomes important.

Single winding transformers are used extensively at RF. In one form, the primary winding consists of a single turn of hollow metal tubing threaded through suitable ferrite material shorted at one end via a copper strap or PCB. Insulated wire is then passed inside this tube to form the secondary (see Figure A.19 [23]). Since the primary is limited to a single turn, the impedance ratios are limited to 1:n^2; where n is the integer number of turns in the secondary (n = 2 in Figure A.19). The principal advantage of this construction method is the close proximity of the windings and ferrites, leading to minimal leakage inductance and improved performance at high frequency. In addition is a fairly rugged structure. A 1:4 (n = 2) transformer is shown in Figure A.19.

These types of transformers can be easily constructed from coaxial cable. By using the center conductor as the secondary, the cable can be wound to the required number of turns. The outer conductor is then removed at the open circuit end and soldered together along its remaining length.

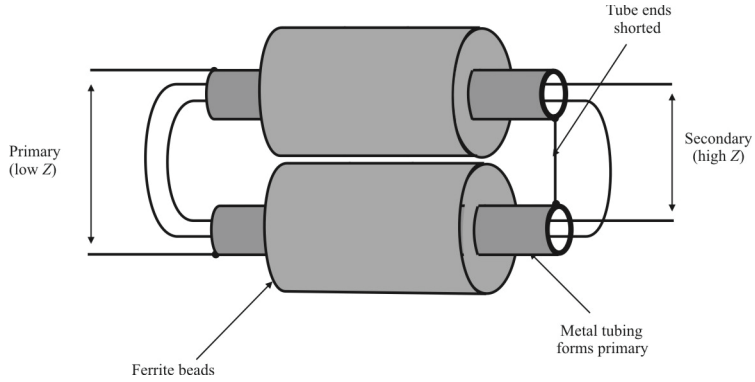


Figure A.19 Transformer implementation. Single-winding transformer implementation. In this case $n_p = 1$, $n_s = 2$, so $n = 2$ and a 1:4 transformer ensues.

Generally, conventional transformers will not perform as well as transmission line transformers (discussed next) in terms of bandwidth, losses, and power handling. They do, however, offer a wider variety of possible impedance ratios and DC isolation from primary to secondary windings.

In a conventional transformer, the energy transfer between the primary and secondary windings occurs mainly through magnetic coupling between the primary and secondary usually enhanced by a ferrite core. This coupling also accounts for the transformer's capability of providing good low-frequency response. Assuming a lossless ferrite core and pure resistive load and source impedances, and considering only the influence of its magnetization inductance, the simplified low-frequency model for the transformer can be represented by the structure in Figure A.20 [24].

The low-frequency model response under maximum power transfer condition is determined by the device's insertion loss: [24]

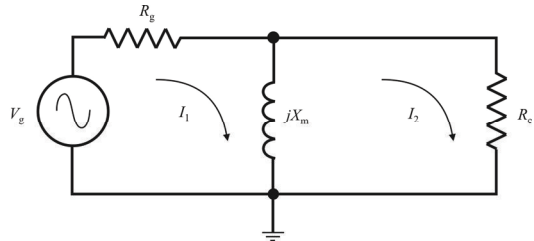


Figure A.20 Conventional transformer model, low frequency. (AC analysis model.)

$$A_{il} = \frac{P_s}{P_L} = \frac{R_s^2 + 4X_m^2}{4X_m^2} \tag{A.12}$$

where:

P_s = the maximum available power of the source.
P_L = the power delivered to the load.
R_s = the source impedance.
X_m = the magnetization reactance.

This last parameter is determined by the operating frequency, f, and the core magnetization inductance, L_m, with, of course,

$$X_m = 2\pi f L_m \tag{A.13}$$

The value of L_m depends on the number of turns for the primary winding and a core inductance factor, A_c. This factor is usually specified by ferrite core manufacturers in nH turns^{-2}. Therefore, the magnetization inductance in nanohenries is [24]

$$L_m = n_p^2 A_c \tag{A.14}$$

By applying this parameter within the corresponding reactance formula, and substituting the result of that calculation into the insertion loss equation, the lower cutoff frequency of the transformer can be determined by putting the relationship in (A.13), yielding [24]

$$f_i = \frac{R_s \times 10^9}{4\pi n_p^2 A_c} \tag{A.15}$$

This value decreases as the number of turns for the primary winding increases. This formula can also be applied for a given desired cutoff frequency in order to calculate the proper number of turns for the primary winding. The factor of 10^9 is used so that the inductance specification can be given in nanohenries.

A.4.2.2 Transmission Line Transformers

TLTs are, as the name implies, transformers that use transmission lines. They rely on a large variation in impedance being presented to the differential (balanced)

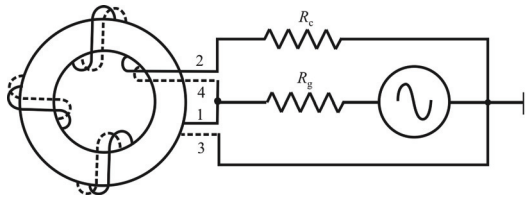


Figure A.21 Unun architecture.

mode and common (unbalanced) mode currents on the transmission line. This is readily achieved at high frequencies, and the low frequency performance can be extended using a suitable ferrite material around the transmission line.[1] It is important to note that unlike conventional transformers, the ferrite is not the medium being used to transfer RF power between ports. For this reason much larger bandwidths and greater efficiencies are achievable using the TLT. Degraded performance at high frequency is caused by the length of the line (maximum $\lambda/8$), by ground plane effects, and by nonoptimum characteristic impedance of the chosen transmission line.

In this section, we discuss *unbalanced-to-unbalanced* (unun) wideband impedance transformers with an impedance ratio of 1:4. In addition to interstage coupling of amplifier chains, other applications of these transformers are in hybrid circuits, signal combiners, and dividers (see Chapter 21).

A wideband unun impedance transformer consists of a toroidal ferrite core wrapped with a twisted bifilar transmission line, with the wires isolated by means of an enamel film. The architecture offers high efficiency for an impedance transformation ratio of 1:4 (see Figure A.21 [24]).

Figure A.22 [23] shows the basic "building block" of TLTs—the 1:1 unbalanced-to-balanced transformer first introduced by Guanella in 1944 [27]. Note that this provides no impedance transformation, but each side of the balanced load has an impedance half that is seen at the input. This is effectively an RF transmission line equivalent of the center-tapped transformer used at lower frequencies, discussed in the last section.

The choice of transmission line largely depends on the characteristic impedance as dictated by the choice of balun or transformer. Coaxial cables of fixed characteristic impedance values are readily available from numerous sources, and these can be combined to make transmission lines of varying impedances. Teflon insulated coaxial cable is preferred for most VHF/UHF

[1] Differential mode currents produce no magnetic field external to the transmission line and therefore remain unaffected by the addition of ferrites.

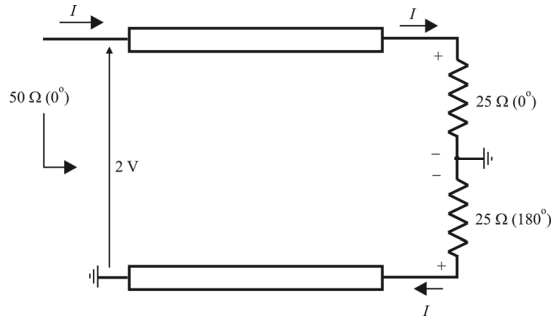


Figure A.22 Guanella transmission line balun.

applications, while microstrip lines are often used above this range. The power-handling capability is generally limited by the maximum allowable temperature, which itself is a function of dielectric material and cable diameter. Table A.3 provides a rough guide that serves as a starting point for analysis.

Twisted pairs are useful since a limitless range of characteristic impedances can be realized. We list here some guidelines for designing twisted pairs of specific Z_0:

- Increasing the number of twists per unit length decreases impedance.
- Increasing the wire diameter will reduce the impedance.
- Very low impedances are achieved by using multiple pairs in parallel.
- 50Ω wire can be made using AWG-22 enameled wire with 2.5 twists per cm.

The impedance will also be affected by the wire separation, the dielectric material between them and the proximity of any ground plane.

The simplest type of TLT is the λ/4 line. These transformers are inherently narrowband because the impedance match is dependent on the presence of a

Table A.3 Approximate Coaxial Cable Power Handling

Outer Diameter	Cable Type	1 MHz	10 MHz	100 MHz	500 MHz
1.7 mm	Flexible	1 kW	300 W	90 W	–
2.8 mm	Flexible	1 kW	800 W	250 W	–
1.1 mm	Semirigid	–	–	68 W	32 W
2.2 mm	Semirigid	–	–	330 W	140 W
6.4 mm	Semirigid	–	–	1.2 kW	515 W

Source: [23].

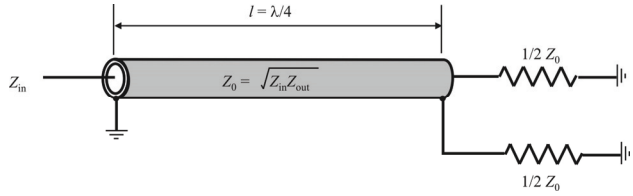


Figure A.23 Coaxial quarter-wavelength balun.

quarter-wavelength standing wave. For maximum power transfer, they also require a transmission line of characteristic impedance, which is the geometric mean of the input and output impedances to be matched, that is,

$$Z_0 = \sqrt{Z_{\text{in}} Z_{\text{out}}} \tag{A.16}$$

This can limit the practicality of this technique since many types of transmission line are only available in a limited range of impedances. However, cables can be combined to achieve Z_0 variations. For example, two paralleled 25Ω cables will produce a transmission line with a characteristic impedance of 12.5Ω.

The λ/4 line can be used as a balun in the same way as above—by grounding one conductor on the unbalanced side and connecting each conductor to half of a balanced load on the other side. Figure A.23 [23] shows an implementation of this using coaxial cable.

The classic Ruthroff 1:4 unun impedance transformer implemented with TLTs is shown in Figure A.24 [23].

A single transmission line can also be used as an *unbalanced to unbalanced* (unun) transformer. This is capable of providing wideband 1:4 impedance

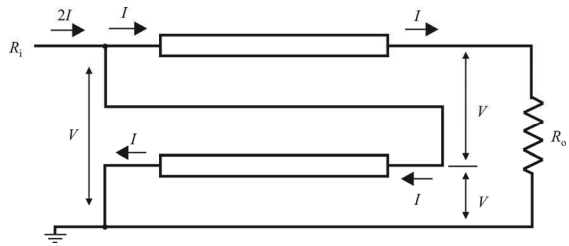


Figure A.24 Ruthroff 1:4 unun.

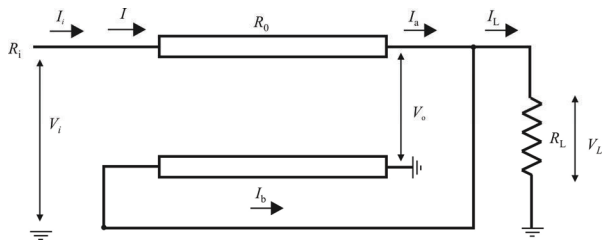


Figure A.25 Classic Ruthroff 4:1 unun impedance transformer using TLTs.

transformation ratios by connecting it in the so-called "bootstrap" configuration shown in Figure A.25 [28]. Here, the two conductors that constitute the transmission line are used as the primary and secondary windings in a similar way to the conventional autotransformer (except that there is no ferromagnetic coupling—the coupling is done with the E-field).

With voltage V across the input, the same voltage is impressed across the lower conductor of the transmission line. Charge conservation demands that the same voltage must be present across the upper conductor also. The voltage across R_0 is therefore the sum of these voltages: 2 V. If the current I is to flow through the load, it must also flow through the upper conductor. However, since both conductors have the same voltage across them, the currents through each must be identical. Therefore, since $R_0 = 2V/I$ and $R_i = V/2I$, then:

$$R_i = R_0 / 4 \tag{A.17}$$

We can show that maximum power transfer occurs for this transformer with the optimum transmission line impedance $Z_0 = 2R_i$. For best performance the transmission line should be kept as short as possible. Figure A.26 [23] shows two practical implementations of the 1:4 unun using coaxial cable and wire-wrapped toroid configurations. Palpably, these devices are bilateral and simply reversing the ports will provide 4:1 step-down transformations.

This technique can be extended to other transformation ratios. By adding extra transmission lines, other $1:n^2$ transformers can easily be implemented. Figure A.27 [23] shows a 1:9 ($1:3^2$) transmission line schematic alongside an implementation using coaxial cable; 1:16 ($1:4^2$) transformations can be realized by adding a third conductor pair and connecting them in the same way as the second.

One of the prime factors limiting high-frequency performance is the phase error caused by the arbitrary length of the transformers' interconnections. If these connections were made using a transmission line of the same length, velocity factor, and impedance of the transformer line itself, then the phase error would be

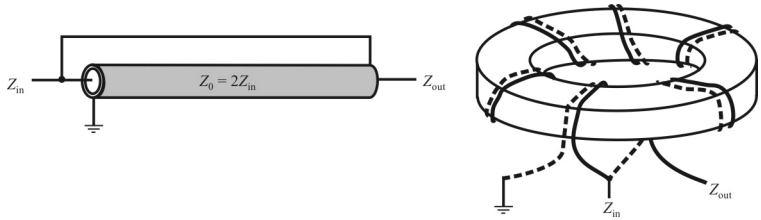


Figure A.26(a, b) Practical 1:4 unun implementations: (a) using coaxial cable and (b) toroidal implementations.

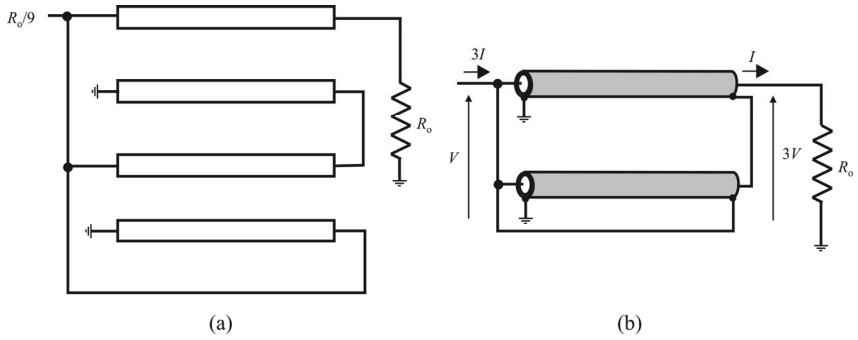


Figure A.27(a, b) 1:9 Ruthroff unun: (a) 1:9 Ruthroff unun schematic, and (b) 1:9 Ruthroff unun implementation with coaxial cables.

eliminated. An additional advantage is that the physical shape of the transformer is also no longer restricted by the need to bring connecting points close together.

Devices of this type are referred to as *equal delay transformers*. Figure A.28 [23] shows how the principle is applied to the 1:4 Ruthroff unun seen in Figure A.24. Note the absence of any magnetic material on the interconnecting line.

By combining a number of basic building blocks in a range of parallel/series combinations balanced-to-balanced impedance ratios of $1{:}n^2$ can be achieved. The diagrams in Figure A.29 [23] show 1:9 unun schematic and 1:9 transmission line transformer implementation using coaxial cables based on this principle. Like the quarter-wavelength transformer, the optimum transmission line impedance is the geometric mean of input and output.

These transformers are often constructed of coaxial cable, and it is good practice to form the cable into a suitable shape to keep the interconnects as short as possible.

The electrical coupling between the primary and secondary windings of a TLT improves the transfer of high-frequency energy. Figure A.30 [23] shows the high-frequency model for a transmission-line 1:4 unun transformer, considered without losses because of its short length. In this idealized model, the source and load impedances are assumed to be pure resistances.

As for the conventional transformer, the high-frequency response is quantified by its insertion loss. Again, the ratio between the available source power and the secondary load power is [24]

$$\begin{aligned} A_{\text{il}} &= P_{\text{s}} / P_{\text{L}} \\ &= \frac{Z_0^2\{2R_{\text{s}}[1+\cos(\beta l)]+R_{\text{L}}\cos(\beta l)\}^2+(R_{\text{s}}R_{\text{L}}+Z_0^2)^2\sin^2(\beta l)}{4R_{\text{s}}R_{\text{L}}Z_0^2[1+\cos(\beta l)]^2} \end{aligned} \tag{A.18}$$

where:

R_{s} = the source impedance.
R_{L} = the load impedance.
Z_0 = the transmission-line characteristic impedance.
βl = the phase factor.
$l = k\lambda$ = the length of the transmission line with k a fraction of wavelength λ.

Equation (A.18) shows the importance of an optimum Z_0 value to achieve good wideband high-frequency response. Note that denominator is zero for a $\lambda/2$ line indicating that the insertion loss increases without bound. Thus, there is no power transferred for a $\lambda/2$ line length. It is 1 dB less than the maximum value for a $\lambda/4$

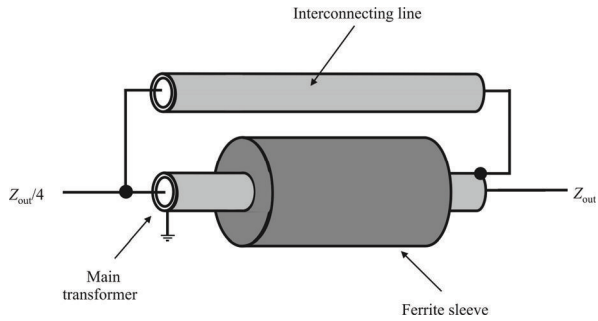


Figure A.28 1:4 equal delay unun.

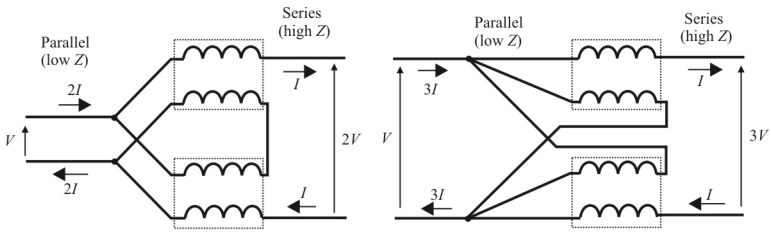


Figure A.29 Balanced-balanced transformer: (a) 1:4 balanced-balanced transformer; and (b) 1:9 balanced-balanced transformer.

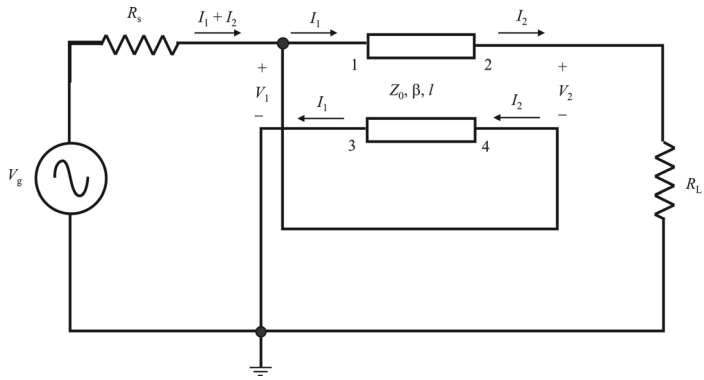


Figure A.30 High-frequency model of transmission line 1:4 unun transformer. Classic Ruthroff 4:1 unun impedance transformer.

line length. From this we can be seen that shorter lengths of line result in broader-bandwidth high-frequency responses. The optimum line characteristic impedance and the load impedance for maximum power transmission are

$$Z_{\text{opt}} = \sqrt{R_s R_L} \tag{A.19}$$

$$R_L = 4R_s \tag{A.20}$$

A 1:4 transformation is necessary between the source and load impedances for impedance matching. With this result, a relationship between line characteristic impedance and the source and load impedances can be established:

$$Z_0 = 2R_s = R_L / 2 \tag{A.21}$$

The use of twisted transmission line in a transformer makes it possible to set the characteristic impedance almost at an optimum value for a desired passband by varying the number of twists per unit length of transmission line. An increase in the number of twists per unit length results in a decrease in the characteristic impedance.

Figure A.31 [24] plots insertion loss behavior as a function of k for optimized and non-optimized values of characteristic impedance when twisted transmission lines are implemented. For a case with non-optimized characteristic impedance, the insertion loss increases and bandwidth decreases relative to a case using optimized characteristic impedance.

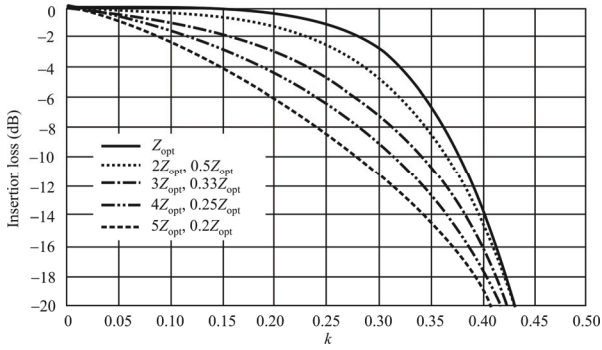


Figure A.31 Insertion loss.

- The procedure for implementation of a TLT with required specifications is as follows. In order to determine a transformer's low-frequency response, the characteristics of the ferrite core must be known since the inductance factor A_c is determined relative to a specific frequency.
- Knowing A_c as well as the source's internal impedance (R_s), we can establish the low-frequency cutoff frequency (f_i).
- Using (A.15), calculate the required number of turns for the primary winding (n_p).
- To determine the high-frequency response, information about the transmission line is required, such as its characteristic impedance (Z_0), the propagation velocity (v_p), and the phase factor (β), all at the desired operating frequency. Along with the values of the source impedance (R_g) and the load impedance (R_L), the optimum theoretical value of the characteristic impedance (Z_{opt}) can be determined from (A.19).
- Knowing the characteristics of the transmission line, the high-frequency cutoff frequency (f_s), and the true characteristic impedance of the transmission line, Z_0, it is possible to calculate the propagation velocity (v_p) and the phase factor (β).
- With the value of the true characteristic impedance, Z_0, the difference between it and Z_{opt} can be verified and the definitive insertion loss for f_s be specified. Figure A.31 shows the determination of values of k as a function of true characteristic impedance (Z_0) and insertion loss.
- With values for k, v_p, and f_s, the line length (l) required for achieving the previous specifications can be calculated with

$$l = k\lambda = kv_p / f_s \tag{A.22}$$

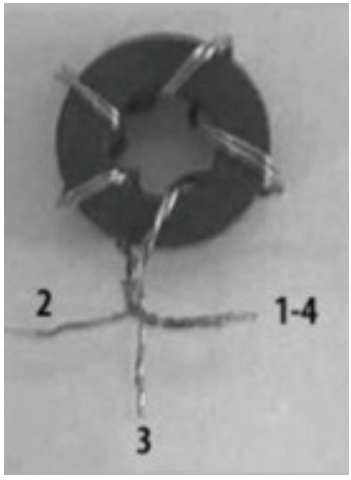


Figure A.32 Four-turn transformer implementation using a toroid.

Using this procedure, a transformer was fabricated with four turns, resulting in a 9-cm line length. This transformer corresponds to the diagram in Figure A.21. A picture of the transformer is shown in Figure A.32. Figure A.33 [24] shows the experimental frequency insertion loss behavior.

These model equations represent a simplified equivalent-circuit model for a coil transformer. More elaborate models including resistive losses and reactive effects are possible, but the model described here is a good starting point for simulations. As these simplified equations show, they can be used to design transformers over a wide frequency range (three octaves) with low insertion loss and at low cost.

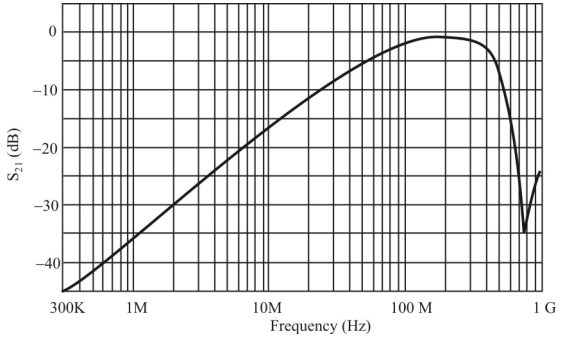


Figure A.33 S_{21} for a four-turn transformer.

A.4.3 Push-Pull Amplifiers Without Transformers

Implementation of a push-pull amplifier with BJTs is shown in Figure A.34. The biasing details are not included. We see that such an implementation requires both an npn as well as a pnp BJT. However, the limitations imposed by a transformer are avoided with this architecture.

Input and output matching networks are usually required for maximizing power transfer and minimizing distortion but also are not shown in the figure.

As the input voltage increases above 0 V, the npn (upper transistor) begins to conduct while the pnp (lower) transistor decreases its conduction. This causes the npn collector voltage to begin to decrease while the pnp collector voltage to increase. When the input reaches its maximum excursion, the npn collector voltage is at its minimum while the pnp is at its maximum. Thus, transformers are not needed for the proper operation of this push-pull architecture.

This is not a popular architecture for RF applications primarily because a pnp BJT transistor (or its FET counterpart) is required. It is widely known that the minority particle conduction in the base has a slower mobility velocity than for the npn transistor by typically a factor of 4 or more. This means that the amplifier upper frequency performance is lowered by about this factor.

A.4.4 Crossover Distortion

When class B amplifiers are used in the push-pull arrangement discussed above, there is a small amount of time when both transistors are turned off. This produces what is called “crossover distortion.” One transistor is on during half of the input cycle, and the other is on for the other half. The point where one is almost off and the other is almost on is called the crossover point. These are typically nonlinear regions for most active devices so distortion is generated which can be removed by subsequent filtering. Crossover distortion is illustrated in Figure A.35.

A.5 Concluding Remarks

Power amplifiers are a critical component in EW systems. We covered the common forms of PAs in this appendix. The architectures described, of course, can also be used in other than PA applications.

PAs come in two basic flavors: analog and digital. The former is the more common set of architectures and consists of class A, B, A/B, and C. The distinction is defined by how they are biased and primarily defines their efficiency. Class A PAs are the least efficient since they are biased to be in the on-state for

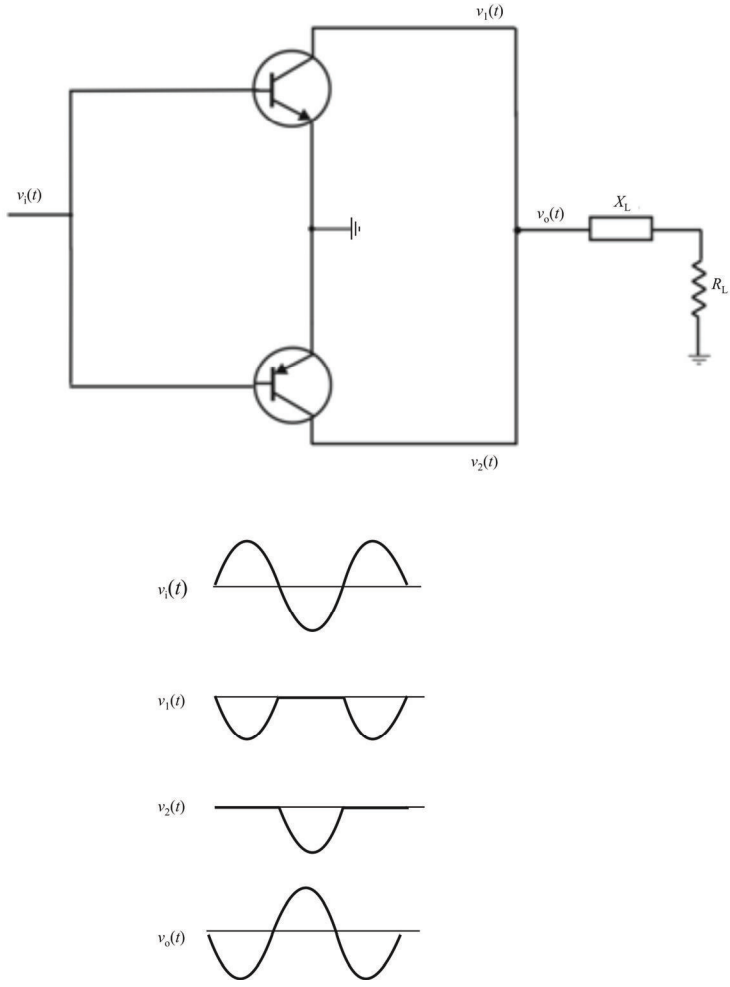


Figure A.34 Transformerless push-pull architecture using complementary BJTs. Bias circuitry is not shown.

Figure A.35 Crossover distortion.

100% of the conduction cycle. Class C PAs are the least efficient because they are biased to be on for less than 50% of the conduction cycle. The other analog classes are between these two efficiencies. On the other hand, class A generates the least amount of distortion, while class C generates the most.

There are several digital architectures. These amplifiers basically do not "amplify" the input signal at all, but the signal to be amplified is used to switch power transistors (or tubes) on and off to produce the high-power output signal. Since, typically, when the collector (BJT power transistor) or drain (MOSFET) voltage (V) is greater than zero, the corresponding current through the devices (I) is zero and vice versa: when the current is greater than zero the collector or drain voltages are zero. Since the power dissipated in the devices is given by $P = VI$, theoretically they are 100% efficient. While this is never actually achieved in practice, very efficient amplifiers are possible with digital architectures.

Linearity is a critical parameter in PAs. When operating in their linear regions, the harmonic distortion and IMD products generated are minimized. All active devices are only linear over a portion of their characteristic ranges, however. If analog amplifiers are driven into their nonlinear region, significant harmonics are generated and these signals can cause unwanted interference with other RF devices operating in the same geographical region as the EW system. These harmonics should be avoided if possible.

References

[1] Krauss, H. L., C. W. Bostian, and F. H. Raab, *Solid State Radio Engineering*, New York: Wiley, 1980, pp. 348–476.

[2] Raab, F. H., P. Asbeck, S. Cripps, P. B. Kenington, Z. B. Popovic, N. Pothecary, J. F. Sevic, and N. O. Sokal, "Power Amplifiers and Transmitters for RF and Microwave," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 50, No. 1, March 2002, pp. 814–826.

[3] Raab, F. H., P. Asbeck, S. Cripps, P. B. Kensington, A. B. Popovic, N. Pothecary, J. F. Sevic, and N. O. Sokal, "RF and Microwave Power Amplifier and Transmitter Technologies—Part 1," *High Frequency Electronics*, May 2003, pp. 22–36.

[4] Raab, F. H., P. Asbeck, S. Cripps, P. B. Kensington, A. B. Popovic, N. Pothecary, J. F. Sevic, and N. O. Sokal, "RF and Microwave Power Amplifier and Transmitter Technologies—Part 2," *High Frequency Electronics*, July 2003, pp. 22–36.

[5] Raab, F. H., P. Asbeck, S. Cripps, P. B. Kensington, A. B. Popovic, N. Pothecary, J. F. Sevic, and N. O. Sokal, "RF and Microwave Power Amplifier and Transmitter Technologies—Part 3," *High Frequency Electronics*, September 2003, pp. 34–48.

[6] Raab, F. H., P. Asbeck, S. Cripps, P. B. Kensington, A. B. Popovic, N. Pothecary, J. F. Sevic, and N. O. Sokal, "RF and Microwave Power Amplifier and Transmitter Technologies—Part 4," *High Frequency Electronics*, November 2003, pp. 38–49.

[7] Raab, F. H., P. Asbeck, S. Cripps, P. B. Kensington, A. B. Popovic, N. Pothecary, J. F. Sevic, and N. O. Sokal, "RF and Microwave Power Amplifier and Transmitter Technologies—Part 5," *High Frequency Electronics*, January 2004, pp. 46–54.

[8] Gilbert, B., "A New Wide-Band Amplifier Technique," *IEEE Journal of Solid-State Circuits*, Vol. SC-3, No. 4, December 1968, pp. 353–365.

[9] Hodges, D. A., "Darlington's Contributions to Transistor Circuit Design," *IEEE Transactions on Circuits and Systems–I: Fundamental Theory and Applications*, Vol. 46, No. 1, January 1999, pp. 102–104.

[10] Vacca, G., "Power Amplifiers for Microwaves and RF Applications with LDMOS Transistors," *Microwave Journal*, Vol. 49, No. 65, June 2006, pp. 98–101.

[11] Tayrani, R., "Broadband Monolithic S-Band Class-E Power Amplifier Design," *RF Design*, April 2004 pp. 24–28.

[12] Nezami, M., "Fundamentals of Power Amplifier Linearization Using Digital Pre-Distortion," *High Frequency Electronics*, September 2004, pp. 54–59.

[13] Pengelly, R. S., "Improving the Linearity and Efficiency of RF Power Amplifiers," *High Frequency Electronics*, September 2002, pp. 26–34.

[14] Grebennikov, A., "Load Network Design Techniques for Class E RF and Microwave Amplifiers," *High Frequency Electronics*, July 2004, pp. 18–32.

[15] Gao, S., "RF/Miocrowave Power Amplifiers," *IEEE Microwave Magazine*, February 2006, pp. 40–48.

[16] Inoue, A, A. Ohta, S. Goto, T. Ishikawa, and Y. Matsuda, "The Efficiency of Class F and Inverse Class F Amplifiers," *IEEE MTT-S International Symposium Digest*, Vol. 3, June 6–11, 2004, pp. 1947–1950.

[17] Woo, Y. Y., Y. Yang, and K. Bumman, "Analysis and Experiments for High-Efficiency Class-F and Inverse Class-F Power Amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 54, No. 5, May 2006, pp. 1969–1974.

[18] Raab, F. H., "Class-F Power Amplifiers with Maximally Flat Waveforms," *IEEE Transactions on Microwave Theory and Techniques,* Vol. 45, No. 11, November 1997, pp. 2007–2010.

[19] Raab, F. H., "Introduction to Class F Power Amplifiers," *RF Design*, Vol. 19, No. 5, May 1996, pp. 79–84.

[20] Ingruber, B. W., W. Pritzl, D. D. Smely, M. Wachutka, and G. Magerl, "Harmonic-Controlled Amplifier," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 46, No. 6, June 1998, pp. 857–863.

[21] Radisic, V., Y. Qian, and T. Itoh, "Class F Power Amplifier Integrated with Circular Sector Microstrip Antenna," *IEEE MTT-S International Symposium Digest*, June 1997, Vol. 4, pp. 687–690.

[22] Roberts, G. W., and A. Lu, *Analog Signal Generation for Built-In-Self-Test of Mixed Signal Integrated Circuits*, Appendix A, Boston, MA: Kluwer Academic Publishers, 1995.

[23] McCarthy, S. J., "SEMELAB Application Note: Push-Pull Circuits and Wideband Transformers," http://www.semelab.com/rf/documents/Push-Pull%20Circuits%20and%20Wideband%20Transformers.pdf.

[24] Ferreira, A. A., J. Antonio, J. Ribeiro, W. N. Do, and A. Pereira, "Designing Wideband RF Impedance Transformers," *Microwaves & RF Magazine*, March 2007, http://www.mwrf.com/Articles/Index.cfm?Ad=1&ArticleID=15089.

[25] http://www.powerstream.com/Wire_Size.htm.

[26] Sevick, J. "A Simplified Analysis of the Broadband Transmission Line Transformer," *High Frequency Electronics*, February 2004, pp. 48–53.

[27] Guanella, G., "New Method of Impedance Matching in Radio-Frequency Circuits," *The Brown Boveri Review*, September 1944.
[28] Carpentieri, E., "Model Characterizes Transmission-Line Transformers," *Microwaves and RF*, November 1996.

Appendix B

RF Switches

B.1 Introduction

RF switches are ubiquitous in modern RF/microwave systems. They are utilized to control signal flows, select signal sources, and to do many other applications. In this appendix we will focus on their applications to switching antenna signals.

There are two fundamental types of RF switches:

- Electromechanical;
- Solid state.

We will discuss both types here.

At RF there are many issues relating to switching performance. Some of the more important are:

- Loss versus frequency (bandwidth);
- VSWR versus frequency (impedance mismatch);
- On-off isolation (feed-through);
- Isolation to other circuits (crosstalk);
- Power handling;
- Switching speed;
- Operating lifetime (reliability, MTBF).

Other basic parameters that are important for specific applications include physical size, operating voltage, power consumption, and rated temperature range.

Many mechanical and solid state switching systems exist for controlling RF signal routing. These include diverse solid state RF switches that are preferable because they are fast acting, minimize noise and switching transients, and have no mechanical structure. Although these mechanical or solid-state devices are still in use, recent developments in *microelectromechanical systems* (MEMS) technologies and microelectronics manufacturing have made microelectrostatic

and micromagnetic relays possible. MEMS switches are termed electrostatic MEMS switches if they are actuated or controlled using electrostatic force that turns such switches on and off. Electrostatic MEMS switches are advantageous due to low power consumption because they can be actuated using electrostatic force induced by the application of a voltage with virtually no current. Electrostatic controlled RF MEMS switches are the most prolific type available.

RF switches come in a variety of configurations. The basic one is a *single pole single throw* (SPST), containing a single RF input and RF output. Other RF switches come in *single pole multiple throw* (SPMT) configurations. A typical RF switch contains two ports or outputs, which are separately controlled, that are fed from a single input port, and is referred to as a *single pole double throw* (SPDT) configuration.

RF switches consisting of solid-state devices, such as diodes and field-effect transistors (FET), are used in communication systems applications. For very high frequencies of about 1 GHz, these diode and FET devices are typically fabricated using *gallium arsenide* (GaAs) technology. Available solid-state RF switches are generally made using GaAs processes. The GaAs and possibly the *silicon*/*germanium* (Si/Ge) processes are accepted as having higher-frequency responses than silicon processes; however, Si processes are much more prolific and therefore less expensive. GaAs chips can be made reasonably small and handle sufficient power with switch characteristics that have made these processes acceptable L-band switches. PIN diodes are semiconductor devices which can be made to operate at relatively high radio frequencies and that function essentially as switched resistors, having a high or low resistance value depending upon the value of the biasing characteristics. PIN diodes are typically fabricated in Si and GaAs.

Often these devices are combined in series or shunt configurations to produce multipole-multithrow devices. Series configurations are used when minimum insertion loss is required over a broad frequency range. One of the more serious drawbacks of such switches is the necessity to provide a constant DC current. These devices typically require a relatively large reverse bias voltage to present a high impedance value, and draw a substantial forward current when forward biased to present a low impedance value. Moreover, using discrete PIN diodes increases both the size and the cost of the RF switch arrays. PIN diodes and transistors typically have an insertion loss greater than 1 dB, which is the loss across the switch when the switch is closed. Single stage transistors operating at microwave frequencies tend to have an isolation value less than 20 dB. PIN diodes and transistors have a limited frequency response and typically only respond to frequencies below 20 GHz. In addition, the insertion losses and high isolation value for these switches vary depending on the frequency of the signal passing through the switches.

This appendix is structured as follows. We begin, after this introduction, with a section on general switch characteristics, that is, attributes that are common to all the types of switches. That is followed by a section on EM switches including

MEMS. We conclude the appendix with a description of several types of solid state RF switches, including PIN diodes and FETs (MOSFETs, JFETs, and MESFETs).

B.2 RF Switch Characteristics

RF/microwave switches utilize variable impedance circuit elements to direct the flow of signals. A switch either allows a signal to propagate through a specific signal path or it blocks the signal from following this path. In an ideal switch, circuit paths are either closed via an ideal conductor with no impedance or opened by an ideal open circuit with infinite impedance. Modern semiconductor elements cannot quite meet those ideal impedance values, but they can come sufficiently close to produce excellent switch performance.

B.2.1 Reflective or Absorptive

Another taxonomy of RF switches is defined by whether they are implemented as reflective or absorptive structures. An ideal reflective switch places either an open circuit in cascade with a signal path (Figure B.1), or a short circuit across the signal path (Figure B.2).

These two extreme impedances produce maximum mismatch of impedance, which produces reflection of the entire incident signal back to its source. An ideal absorptive switch (sometimes called a "matched switch") produces isolation by placing a termination whose resistance is exactly equal to the characteristic impedance, Z_0, of the transmission line across the transmission line as shown in Figure B.3. In this case, all of the incident signal energy is absorbed by the terminating resistance and dissipated as heat, leaving no remaining energy to be reflected back to the signal source. In the other state, the termination is disconnected from the transmission line, thereby allowing all of the incident energy to propagate through the switch.

B.2.2 Common Switch Configurations

Switches may be implemented in many configurations, depending on the purpose for which the switch is to be used. These configurations are described in terms of the number of poles and the number of throws implemented in the switch. The number of poles describes the number of signal paths controlled by the switch. The number of throws indicates the number of potential directions into which a pole may be placed. For example, the simplest switch configuration is a single pole, single throw (SPST) switch (Figure B.4). This configuration has one signal

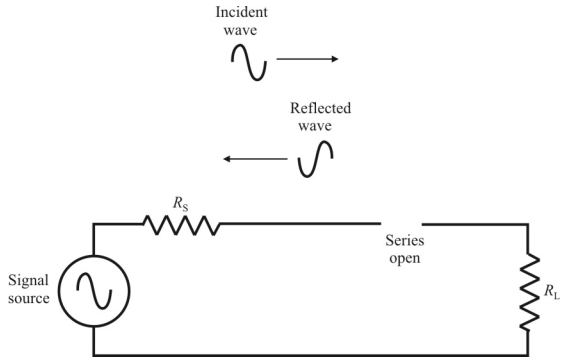


Figure B.1 Series open reflective switch.

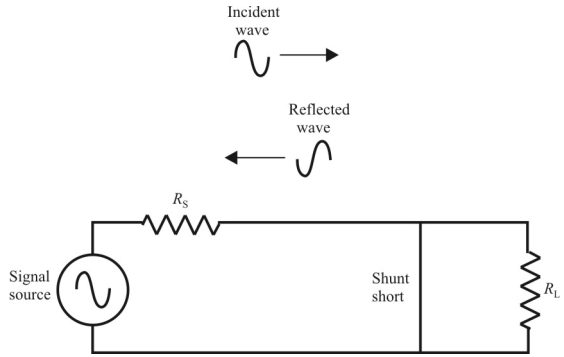


Figure B.2 Shunt short reflective switch.

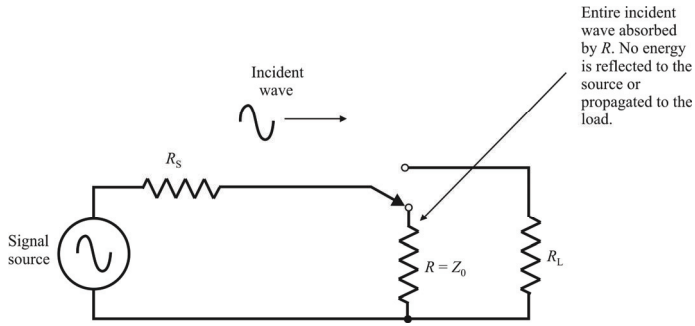


Figure B.3 Absorptive switch.

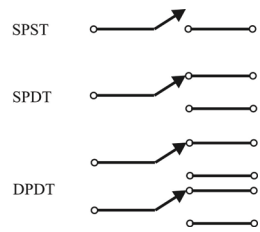


Figure B.4 Example switch configurations. Many other configurations are possible.

path which can either be completed by the switch or interrupted by the switch. A single pole double throw switch (SPDT or SP2T) can connect a single transmission line to either of two other transmission lines. The number of poles and throws and the combinations thereof are ideally unlimited, but there are practical limitations that will be described later.

B.2.3 Ideal Versus Practical Switches

In practical RF/microwave switches, it is not possible to produce a perfect open impedance or a perfect short circuit. This is particularly true with solid-state technology. Consequently, there is always some small amount of incident signal that is absorbed by the switch and a bit more reflected by the switch's nonideal impedance when the switch is in the state in which it should ideally pass all incident signal energy. This small reduction in signal amplitude is known as *insertion loss* (IL) and is typically described in terms of decibels. Insertion loss is simply the ratio of the output power to the input power.

Likewise, there is always some small amount of energy that propagates past the switch when it is in the state in which an ideal switch should produce infinite isolation. The measurement of this characteristic is known as *isolation* and is also described in terms of decibels. Isolation is also the ratio of the output power to the input power. Multiple semiconductor elements can be used in a single switch to increase the isolation that the switch produces. Often, these elements are placed in series with the signal path, and in shunt with the path.

Absorptive switches typically include multiple switching elements: some to complete or interrupt the signal path, and others to disconnect or connect the termination resistance to the signal path.

B.2.4 Harmonic Distortion

Harmonic distortion is a single-tone (single-frequency) distortion product caused by device nonlinearity. When a nonlinear device (all semiconductors, including solid-state switches inherently exhibit a degree of nonlinearity) is stimulated by a signal at a single frequency f_1, spurious output signals can be generated at the harmonic frequencies $2f_1$, $3f_1$, $4f_1$, ..., Nf_1. (Nth harmonic is the Nth-order product.) Harmonics are usually measured in decibels relative to the carrier, denoted by dBc, or decibels below the fundamental output signal in this case (see Figure B.5).

B.2.5 Intermodulation Distortion

Intermodulation distortion (IMD) arises when the nonlinearity of a device or system with multiple input frequencies causes undesired outputs at other frequencies, causing the signals in one channel to interfere with other channels. Reducing IMD has become more important as the spectrum becomes more

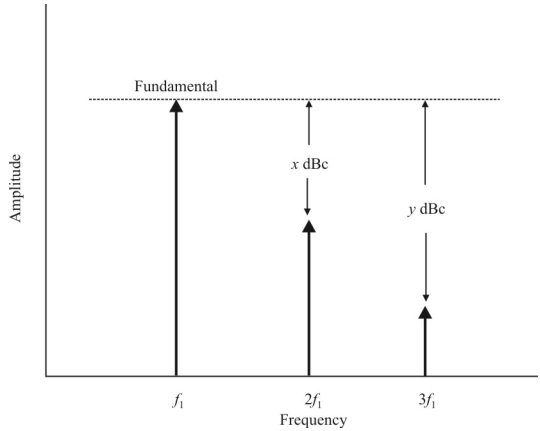


Figure B.5 Harmonics of a fundamental frequency.

crammed and channels are more tightly spaced. From an EA point of view, IMD products in the PA of a jammer can cause unintended EMI in channels other than that of the SOI. This is called *RF fratricide*, and its biggest impact is on possible interference to friendly communications.

These spurious products are mathematically related to the fundamental input signals. It is common practice to limit the analysis to two tones (two fundamental frequencies, f_1 and f_2, which are normally separated by a small offset frequency of around 1 MHz) due to the complexity of analyzing more than two input frequencies at a time. The output frequencies of the two-tone IMD are:

$$\pm pf_1 \pm qf_2$$

where p and q are integers and the combinations of p and q make sense (both cannot be simultaneously negative, for example). The order of the distortion product is the sum of $p + q$. So the third-order IMD products of the two signals, f_1 and f_2, would be $2f_1 + f_2$, $2f_1 - f_2$, $2f_2 + f_1$ and $2f_2 - f_1$, $3f_1$, and $3f_2$ (see Figure B.6). Intermodulation products greater than third order are frequently small enough to be ignored.

Broadband systems may be affected by all the nonlinear distortion products. However, unlike harmonic and second-order distortion products, third-order intermodulation distortion products (IP3) are always too close to the fundamental signals to be easily filtered (see Figure B.6). When strong signals from more than one transmitter are present at the input of the receiver, as is commonly the case in cellular telephone systems, unwanted distortion products will be generated affecting the receiver measurement.

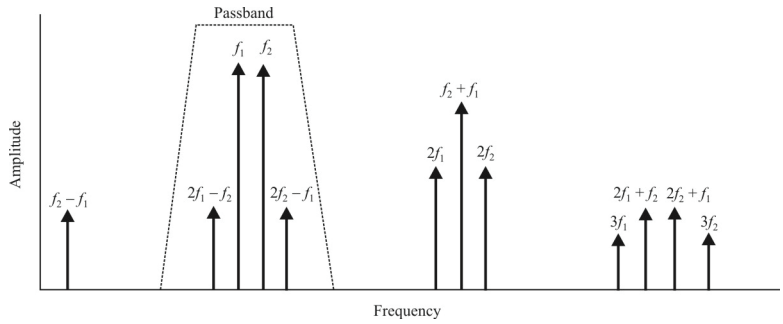


Figure B.6 Second- and third-order intermodulation distortion.

The nonlinear transfer function of a device or system can be expressed as a series expansion:

$$v_{\text{out}}(t) = a_1 A\cos(\omega t) + a_2 A^2 \cos^2(\omega t) + a_3 A^3 \cos^3(\omega t) + \ldots \tag{B.1}$$

Higher-order intermodulation products will increase in power much faster than the fundamental as shown in Figure B.7. The intermodulation signal will increase in power three times faster than the carrier signal. The increase is in dBm units, which are a logarithmic function. The same relationship holds true for other harmonics or intermodulation products. For instance, the second harmonic and the second order intermodulation products' power will increase two times (in dBm) faster than the fundamental, while the third harmonics will increase three times as fast.

B.2.6 1 dB Compression Point

Another nonlinearity of a system or device is measured by the compression point. In the linear region, a 1 dB increase in input power to a device will correspond to 1 dB increase in the output power. The nonlinearity effect becomes apparent when the output power starts to increase less than the input power. When the power difference increment is 1 dB, the device has a 1 dB compression point. See Figure B.8. If not explicitly stated, the 1 dB compression point refers to the output power (P_{out}) at that point.

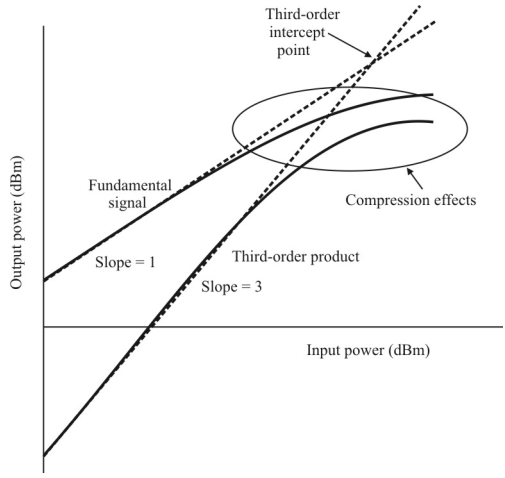


Figure B.7 Realistic representations of third-order intercept point.

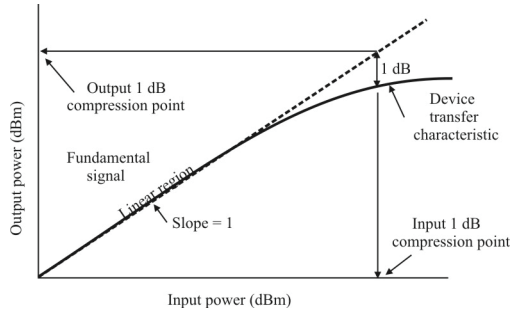


Figure B.8 1 dB compression point. Unless otherwise specified, the output 1 dB compression point is assumed.

B.3 Electromechanical Switches

Electromechanical switches are most often used where high power needs to be switched, or where the lowest possible loss is required. Both of these requirements are best met with the physical connection of metal-to-metal contacts.

There are two fundamental types of electromechanical switches. What they share in common is that some part of the device physically moves to accomplish the switching. These two are:

- Conventional switches;
- Microelectromechanical (MEM).

B.3.1 Conventional Relays

At frequencies below 500 MHz, it is often possible to use general-purpose relays that are primarily designed for control of AC/DC power. Certain manufacturing methods result in reasonably low capacitance between open contacts and low inductance in the internal connection. Although limited to the range from low frequencies through VHF, this part of the spectrum includes many communications applications.

For high power EW applications, such as those associated with EA, conventional relays may currently be the only devices that can be used for switching high voltages and currents. The solid-state switches described below cannot handle high power requirements.

B.3.1.1 RF Optimized Relays

This group is an extension to the above—manufactured like general-purpose relays, but with improvements in the internal construction that further enhance RF performance. Most relays in this group are usable up to 1 GHz and offer a price/performance choice that fits many applications.

B.3.1.2 Microwave Relays

The above relay types have limitations on some aspects of performance, usually *OFF isolation* and *crosstalk*. To handle switching at higher frequencies and with higher performance, specially designed relays are a solution. This family of relays might be better described as switches with electrical actuators, since they often bear little resemblance to general-purpose relays. To maintain low VSWR, these devices are designed with short signal paths and a physical structure that mimics either stripline or coaxial line. High isolation is obtained by grounding the signal path in the *OFF* position, perhaps augmented by actuation movement that increases the distance between disconnected contacts, lowering the capacitance.

Crosstalk is reduced with shielding between signal paths and ports, often cast or machined into the switch housing.

B.3.2 Microelectromechanical Switches

MEMS technology can be used for the fabrication of RF switches. MEMS are miniature devices that are being manufactured in a wide variety of mechanical forms. MEMS devices are inherently both mechanical and electrical devices that are subject to wear and contamination and suffer from limited lifetimes. A MEMS is fabricated using semiconductor *integrated circuit* (IC) fabrication technology.

The RF MEMS switches developed today, even if quite small, still follow the basic electromechanical laws developed a few hundred years ago by Isaac Newton and others. However, the scale and relative importance of the forces are significantly different from the macro world. Surface forces and viscous air damping dominate over inertial and gravitational forces. The switches have very low mass and therefore are not sensitive to only the most severe acceleration forces.

The advent of MEMS has allowed the creation of ultra-small switches. A variety of MEMS switches are in use in radar and communication systems, as well as other high-frequency circuits for controlling RF signals. These MEMS switches are popular, insofar as they can have a relatively high off impedance, with a low off capacitance, and a relatively low on impedance, with a high on capacitance, leading to desirable high cutoff frequencies and wide bandwidth operation. MEMS devices are small in size, and feature significant advantages in that their small size translates into a high electrical performance, since stray capacitance and inductance are virtually eliminated in such an electrically small structure as measured in wavelengths. MEMS switches have a small footprint, can operate at high RF voltages, and are compatible with conventional integrated circuit fabrication techniques.

The MEMS switches can be constructed as bistable devices and are switched by the application of an electrical voltage to an input terminal. Switches fabricated using MEMS technology normally include a substrate with one or more metal traces and control pads. A MEMS switch may be constructed that uses electrostatic force to flex a thin membrane and thereby causes the switch to be opened or closed (see Figure B.9). Since an electrostatic force is used, the switch can be controlled using only a voltage (that is, no current) and therefore consumes virtually no power. This is an important advantage for manportable systems, where power consumption is a significant operating limitation. This type of configuration is typically used in a shunt switch configuration as shown in Figure B.9 [1].

In the capacitive type MEMS switch, a dielectric layer is deposited on the first conductor in an area opposite the underside of the two-arm moveable bridge, with this area on the conductor acting as the pull-down electrode (see Figure B.9). With this arrangement, the full pull-down voltage appears across the dielectric layer resulting in a relatively high electric field across the dielectric. This type of switch is typically used in a series application.

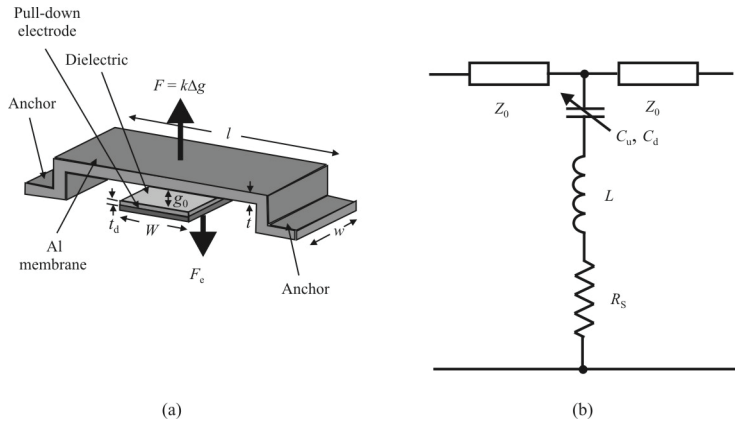


Figure B.9(a, b) MEMs capacitive shunt switch: (a) shunt switch with equivalent circuit in (b).

MEMS switches can be activated by several means, including electrostatic pull-in, magnetic pull-in, electromagnetic, piezoelectric, or thermal.

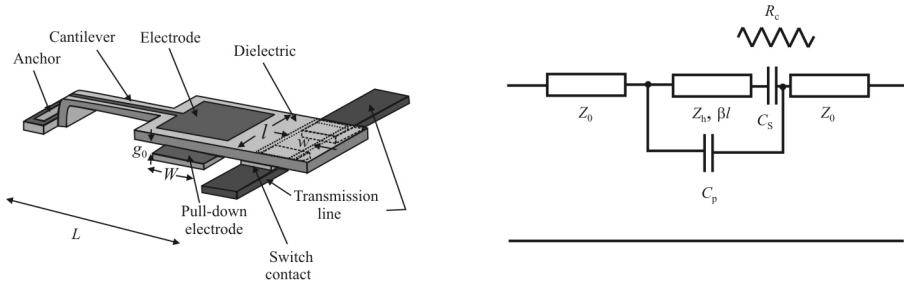
One drawback of the electrostatically controlled RF MEMS switches in use is the activation and hold voltages required for their operation. Pull-in and hold-down voltages of ~100 V and ~30 V, respectively, are required. These voltage levels are difficult to deal with for low-voltage portable EW applications. Progress has been made, however, in achieving lower operating voltages [2, 3].

A well-known expression [4] for calculating the required DC actuation voltage, that is, pull-in voltage of fixed-fixed beams or air bridges of the type shown in Figure B.9, is

$$V_p = \sqrt{\frac{8K_z g_0^3}{27\varepsilon_0 A}} \tag{B.2}$$

K_z is the equivalent spring constant of the moving structure in the direction of desired motion (typically the z-direction), g_0 is the gap between the switch and the actuation electrode, ε_0 is the free-space permittivity, and A is the switch area over which the electrostatic force is applied. For $k = 10$ N m^{-1}, $A = 100 \times 100$ μm^2, $g_0 = 2.5$ μm, the pull-down voltage calculated with (B.2) is $V_p = 23$ V [5].

Equation (B.2) suggests ways that may decrease the required actuation voltage. Reducing g_0 is one way, but it affects the high-frequency off-state switch performance by reducing the switch isolation (series switch configuration) or insertion loss (shunt switch configuration). A second approach to lower actuation voltage is by increasing the actuation area B. For IC applications, this area has to be as small as possible, however. The third approach, which offers the maximum



(a) (b)

Figure B.10(a, b) MEMS series switch: (a) broadside inline MEMs series switches with equivalent circuit (b).

design flexibility for a low-to-moderate actuation voltage, is to lower the switch spring constant. This is examined in [6].

B.3.2.1 Series MEMS Switches

There are two types of MEMS series switches [5]: the broadside series switch (Figure B.10 [1]) and the inline series switch (Figure B.11). The actuation of the broadside switch is in a plane that is perpendicular to the transmission line, while the actuation of the inline switch is in the same plane as the transmission line.

In the broadside configuration, the switch closes to connect two ends of transmission lines together. The cantilever mechanism is not part of the RF path. In the series inline switch configuration, the cantilever mechanism carries the RF signal and, when closed, the signal passes through the switch onto the lower (or upper) contact.

The most common actuation mechanism is achieved using an electrostatic force between the top and bottom electrodes and is given by [4, 5]

$$F = \frac{QE}{2} = \frac{CVE}{2} = \frac{\varepsilon A V^2}{2(g + t_d / \varepsilon_r)^2} \tag{B.3}$$

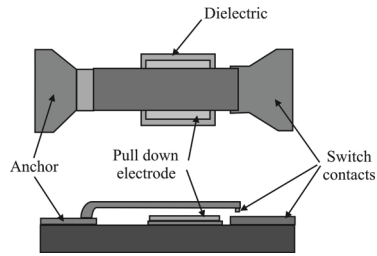


Figure B.11 MEMS inline series switch.

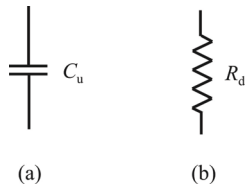


Figure B.12(a, b) MEMS inline series switch models: (a) up-state (C_u ~ 30–50 fF) and (b) down-state (R_d ~ 1–2Ω).

where V, g, and C are the voltage, gap distance, and capacitance between the lower and upper electrodes, respectively, and A is the area of the electrode.

The bottom electrode is often covered by an insulating dielectric layer (thickness t_d and a relative dielectric constant ε_r) to prevent a short circuit between the top and bottom plates.

The model of an inline MEMS series switch is a series capacitance in the up-state position and a small resistance in the down-state position, as illustrated in Figure B.12. The isolation of a series switch in the up-state position is given by [5]

$$\left|S_{21}\right|^2 = 4\omega^2 C_u^2 Z_0^2 \tag{B.4}$$

where C_u is the up-state capacitance (typically 30–50 fF [2]) and Z_0 is the transmission line impedance. The insertion loss is

$$\left|S_{21}\right|^2 = 1 - \frac{R_S}{Z_0} \tag{B.5}$$

where R_S is the contact resistance of the switch. Palpably, when $R_S << Z_0$, as would be desirable in a series switch, $\left|S_{21}\right| \approx 1$.

A MEMS series broadside inline switch is shown in Figure B.10. In some designs there might be two pull-down electrode regions. In a typical design of such a switch the gap height g_0 ~ 1.5–3 μm, total length of the cantilever L ~ 300–400 μm, and an inline length l ~ 80–160 μm. The transmission line is 20–60 μm wide. The actuation pad is placed 100–200 μm from the anchor. The net pull-down voltage with these parameters is typically 15–60 V.

The down-state performance for the switch in Figure B.10 is shown in Figure B.13 [1] when the substrate is 150 μm Si.

B.3.2.2 Shunt MEMS Capacitive Switch

There are several different types of MEMS shunt capacitive switches. One example of these switches is given in Figure B.9. The goal of shunt switching is to

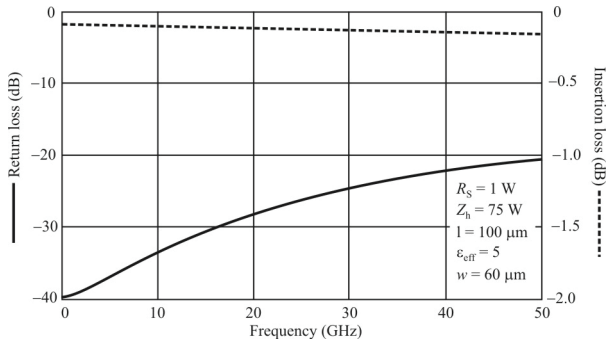


Figure B.13 MEMS broadside series switch performance.

provide an RF short circuit across the load to provide for high isolation. Usually, a shunt switch is based on a fixed-fixed beam design. The anchors are connected to the CPW ground plane, and the membrane is grounded. The center electrode provides both the electrostatic actuation and the RF capacitance between the transmission line and the ground. When the switch is in the up state, it provides low capacitance (~30–50 fF) to the ground, and it does not affect the signal on the transmission line. When the switch is actuated in the down state, the capacitance to the ground becomes substantially higher (~3 pF), which produces the RF short circuit and high isolation at microwave frequencies.

The up-state reflection coefficient of a MEMS shunt capacitive switch is [5]

$$\left|S_{11}\right|^2 = \frac{\omega^2 C_u^2 Z_0^2}{4} \tag{B.6}$$

where C_u is the up-state capacitance of the switch. The down-state isolation is

$$\left|S_{21}\right|^2 = \begin{cases} \dfrac{4}{\omega^2 C_d^2 Z_0^2}, & f << f_0 \\ \dfrac{4R_S^2}{Z_0^2}, & f = f_0 \\ \dfrac{4\omega^2 L^2}{Z_0^2}, & f >> f_0 \end{cases} \tag{B.7}$$

where f_0 is the down-state resonant frequency of the capacitive switch and C_d is the down-state capacitance (typically 3 pF).

From Figure B.9, we can see that the shunt switch impedance is given by

$$Z_S = R_S + j\omega L + \frac{1}{j\omega C} \tag{B.8}$$

where C corresponds to C_u in the OFF (up) position and C_d in the ON (down) position. The simple RLC circuit has a series resonant frequency of

$$f_0 = \frac{1}{2\pi LC} \tag{B.9}$$

Below the resonant frequency the impedance is capacitive, at the resonant frequency the impedance is purely resistive, while above the resonant frequency the impedance is inductive. Thus,

$$Z_S = \begin{cases} 1/j\omega C & f << f_0 \\ R_S & f = f_0 \\ j\omega L & f >> f_0 \end{cases} \tag{B.10}$$

For C_u = 35 fF, C_d = 2.8 pF, and L = 7 pH, the resonant frequency is f_0 = 322 GHz and 36 GHz when the switch is in the OFF state and ON state, respectively [6]. Thus, we see that in the OFF state the inductor does not come into play for most of the situations we are interested in. The case for the ON state is not the same, however, and the effects of the inductor need to be considered.

B.4 Solid-State Switches

These switches are implemented either with *positive-intrinsic-negative* (PIN) diodes or with FETs such as *pseudomorphic high electron mobility transistors* (pHEMTs), or *metal semiconductor FETs* (MESFETs) each of which offers relative advantages and disadvantages compared to the others. In this section, we present an overview of solid state RF/microwave switches and some representative switching circuits, but we will focus on their application as switches operating with antennas. We include discussions of PIN diode switch configurations, the theory of operation for FETs, some representative RF/microwave FET switch topologies, the relative advantages of PIN and FET switches, and some criteria by which one of these technologies may be selected over the other for RF switch applications.

Solid state switches are used when switching speed and reliability are key performance parameters. They also are low cost and have small physical size. The

combination of low power consumption and small size generally simplifies the design of the surrounding circuitry.

B.4.1 Forward-Biased Diodes

For RF switching, a diode biased to full conduction is quite effective. Various technologies—silicon, GaAs, and others—offer various options for loss, frequency range, and compatibility with the rest of the circuitry. A forward-biased diode has low loss and can be configured for high OFF isolation and low crosstalk levels. Since the control voltage shares the signal path, DC isolation is required for the signal (usually capacitors) and RF isolation is required for the control voltage (RF chokes).

B.4.2 PIN Diodes

The PIN diode was developed specifically for RF/microwave use, with the special property of varying resistance between the full ON and full OFF states, controlled by the current flow. They have low resistance in the ON state, making them one of the few solid state choices for switching RF power. To use the variable resistance feature, driving circuitry is required, which increases overall system complexity. Like the simple p-n junction diodes above, PINs also require DC isolation of the signal path and RF isolation of the control circuitry. When used for high power, PIN diodes require high voltage to minimize the effects of the voltage swing of the signal. For example, 100 W of power has a peak-to-peak voltage of 200 V in a 50Ω system. The control voltage must exceed this value by an amount sufficient to prevent the signal voltage from controlling the resistance of the diode and creating IMD.

B.4.2.1 PIN Diode Theory of Operation

The PIN diode is a current controlled resistor at radio and microwave frequencies. It is a semiconductor diode in which a high resistivity intrinsic I-region is sandwiched between a p-type and an n-type region as shown in Figure B.14. When the PIN diode is forward biased, holes and electrons are injected into the I-region from the *p* and the *n* layers, respectively. These charges do not immediately annihilate each other; instead they stay alive for an average time, called the carrier lifetime, τ. This results in an average stored charge, Q, in the I-region which lowers the effective resistance of the I-region to a value R_S.

When the PIN diode is at zero or reverse bias (and assuming it was not forward biased in the immediate past—more about this later), there is no stored charge in the I-region and the diode appears as a larger impedance comprising a capacitance, C_T, shunted by a parallel resistance, R_P.

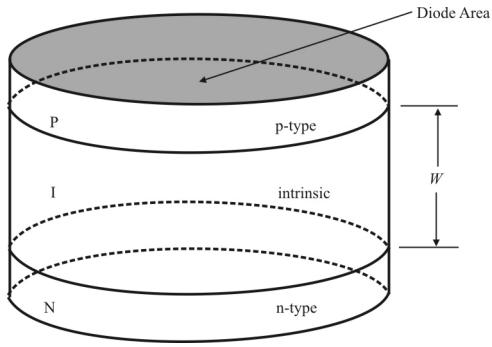


Figure B.14 PIN diode structure.

By varying the I-region width and diode area, it is possible to construct PIN diodes of different geometries to result in the same R_S and C_T characteristics. These devices may have similar small-signal characteristics. However, the thicker I-region diode would have a higher bulk, or RF breakdown voltage, and better distortion properties. On the other hand, the thinner device would have faster switching speed. There is a common misconception that τ is the only parameter that determines the lowest frequency of operation and the distortion produced by the diode. This is indeed a factor, but equally important is the thickness of the I-region which relates to the transit time frequency of the PIN diode.

Low-Frequency Model

At low frequencies (below the transit time frequency of the I-region) and DC, the PIN diode behaves like a PN junction semiconductor diode. Its *I-V* characteristic determines the DC voltage at the forward bias current level. PIN diodes are often rated for forward voltage, V_F, at a fixed DC bias. The reverse voltage rating of a PIN diode, V_R, is the reverse voltage at which the manufacturer guarantees that no more than a specified amount of reverse current, generally 10 μA, will flow. V_R is not necessarily the avalanche or bulk breakdown voltage, V_B, which is determined by the I-region width (approximately 10 V μm^{-1}). PIN diodes of the same bulk breakdown voltage may have different voltage ratings.

Large Signal Model

In order for the PIN diode to act as a current-controlled RF resistor and not as a rectifier, the PIN diode is forward-biased so that the stored charge is much greater than the incremental stored charge added or removed by the RF current, I_{RF}. To ensure this, we must have

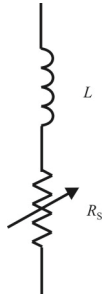


Figure B.15 PIN diode forward-biased model.

$$Q >> \frac{I_{RF}}{2\pi f} \tag{B.11}$$

where

Q is the stored charge from DC forward bias (C).
I_{RF} is the RF signal current (A).
f is the RF signal frequency (Hz).

RF Electrical Modeling of PIN Diode

Forward Bias Model

The model for a forward-biased PIN diode is illustrated in Figure B.15. The series resistance of a PIN diode, when forward biased, is given by

$$R_S = \frac{W^2}{(\mu_n + \mu_p)Q} \tag{B.12}$$

where

R_S is the RF series resistance (Ω).
Q is the DC stored charge = $I_F \times t$ (C).
W is the I-layer thickness (m).
I_F is the DC forward bias current (A).
μ_n is the electron mobility[1] $m^2/(V \times s)$.

[1] An *ambipolar* material is when $\mu_n = \mu_p$; for most semiconducting materials $\mu_n > \mu_p$.

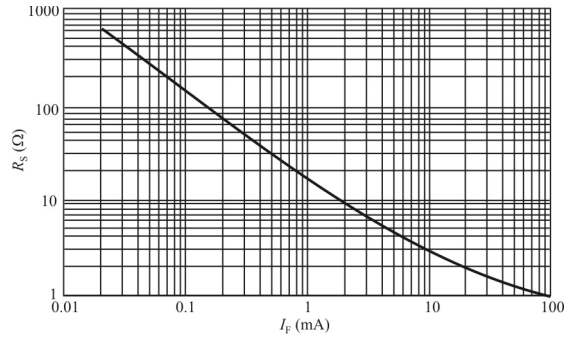


Figure B.16 PIN diode R_S versus bias current.

μ_p is the hole mobility $m^2/(V \times s)$.

This model of the forward-biased PIN provides the following characteristics:

- In a practical diode the parasitic resistance of the diode package and contact limit the lowest resistance value.
- The lowest impedance will be affected by the parasitic inductance, L, which is generally less than 1 nH.
- The equation is valid at frequencies higher than the I-region transit time frequency, that is, $f > 1{,}300/W^2$ (where f is expressed in MHz and W in μm).
- The equation is valid only if the RF signal does not affect the stored charge, that is, it is assumed that the diode is operated under small-signal conditions.

According to (B.12) the forward-biased PIN diode acts as a variable resistance, R_S, whose value depends on the forward bias current I_F through Q and t. Typical values for R_S are illustrated in Figure B.16. As seen, for low values of R_S (~1 Ω) a substantial forward bias current is required (~20–100 mA).

Zero or Reverse Bias

The model of a PIN diode when reverse biased is shown in Figure B.17. Here

$$C_T = \frac{\varepsilon A}{W} \tag{B.13}$$

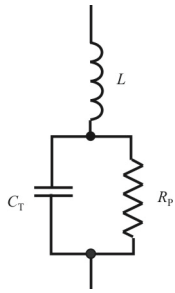


Figure B.17 PIN diode reverse bias model.

where

C_T is the total capacitance (F).
ε is the dielectric constant = $\varepsilon_r \times \varepsilon_0$ (F m^{-1}).
A is the diode area (m^2).
W is the thickness of the depletion layer ≈ thickness of the I-layer (m).

For this model:

- Equation (B.13) is valid at frequencies above the dielectric relaxation frequency of the I-region, that is,

$$f > \frac{1}{2\pi\rho\varepsilon}$$

  where ρ is the resistivity of the I-region. At lower frequencies, while under reverse bias, the PIN diode acts like an abrupt junction varactor diode.
- The value R_P is proportional to voltage and inversely proportional to frequency. In most RF applications, its value is higher than the reactance of the capacitance, C_T, and is consequently of less significance.
- The equation is valid at frequencies higher than the I-region transit time frequency, that is,

$$f > \frac{1,300}{W^2}$$

  where f is expressed in megahertz and W in micrometers.
- The equation is valid only if the RF signal does not affect the stored charge, that is, it is assumed that the diode is operated under small-signal

conditions. Also, under reverse bias, the diode should not be biased beyond its DC voltage rating, V_R. The avalanche or bulk breakdown voltage of a PIN diode is proportional to the I-region width and is always higher than V_R.

In a typical application, maximum negative voltage swing should never exceed V_B. An instantaneous excursion of the RF signal into the positive bias direction generally does not cause the diode to go into conduction because of slow reverse to forward switching speed. The DC reverse bias needed to maintain low PIN diode conductance has been analyzed [7] and is related to the magnitude of the RF signal and I-region width.

PIN Diode Switching Speed

The switching speed in any application depends on the driver circuit, as well as the PIN diode. A PIN diode can be switched from conduction to nonconduction and vice versa. There is a distinct switching speed for each of these two transitions: from forward bias to reverse bias, T_{FR}, and from reverse bias to forward bias, T_{RF}. The transition time from conducting to non-conducting is typically the slower one. The primary diode characteristic that affects T_{FR} is τ, carrier lifetime. The value of T_{FR} may be computed from the forward current and the initial reverse current, I_R with

$$T_{FR} \approx \tau \ln\left(1+\frac{I_F}{I_R}\right) \tag{B.14}$$

This relation implies that T_{FR} is dependent solely on the ratio of the forward bias current to the peak reverse current, assuming τ is constant. This is true for reasonably small magnitudes of forward current. However, as the magnitude of the forward current increases to the value which saturates the I-layer with charge carriers, the actual value of T_{FR} is somewhat less than that predicted by this relation. It is important to note that the reverse current described above is not a leakage current or the diode's reverse saturation current. The reverse current that flows during the forward-to-reverse bias transition is composed of the charge carriers that were previously injected into and stored in the I-layer while the diode was forward-biased. T_{FR} depends primarily on I-region width as indicated in Table B.1 [8] which shows typical data.

PIN Diode Thermal Characteristics

The maximum allowable power dissipation, P_D, is determined by

$$P_D = \frac{T_J - T_A}{\theta_{JA}} \tag{B.15}$$

where

P_D is the dissipated power (W).
T_J is the maximum diode junction temperature (ºC, 175°C for Si).
T_A is the ambient temperature (ºC).
θ_{JA} is the thermal resistance from junction to ambient ºC/W.

Power dissipation, which occurs as a result of Joule heating, may be computed as the product of the square of the RF current multiplied by the diode series resistance, R_S. For CW applications, the value of thermal resistance used is the average thermal resistance, θ_{AV}.

In most pulsed RF and microwave applications where the duty factor, DF, is less than 10% and the pulse width, T_P, is less than the thermal time constant of the diode, good approximation of the effective value of θ in (B.15) may be computed as

$$\theta = DF \times \theta_{AV} + \theta_{TP} \tag{B.16}$$

where θ_{TP} is the thermal impedance of the diode for the time interval corresponding to T_P. The curves in Figure B.18 show how junction temperature is affected during a pulsed RF application.

The relationship for the microwave resistance of a PIN diode contains terms that are related to the forward current, device geometry, and electronic properties of its semiconductor material [9]. These parameters are related to the PIN diode resistance by (B.12) and assuming ambipolar material so that $\mu_n = \mu_p = \mu$, can be written as [10]

$$R_S = \frac{W^2}{2\mu I_0 \tau} \tag{B.17}$$

where the factors are defined after (B.12). Of these factors, only the mobility and lifetime are functions of temperature and contribute to the dependence of the resistance on temperature. The temperature dependence of the mobility in silicon has been extensively studied [11] and it has been established that mobility decreases with increasing temperature in the temperature range of interest. This temperature dependence can be approximated by

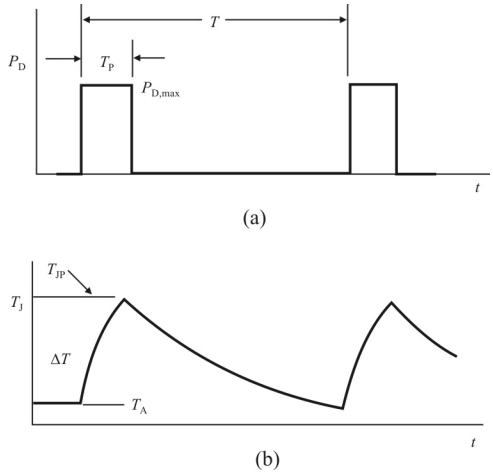


Figure B.18(a, b) PIN diode power dissipation: (a) diode power dissipation and (b) junction temperature versus time for a pulsed RF signal.

$$\mu(T) = \mu(T_0)\left(\frac{T}{T_0}\right)^{-n} \tag{B.18}$$

in the temperature range of 223 to 473 Kelvin (–50–200 ^{o}C) and with a value of n in the range of approximately 2 to 2.2.

The temperature dependence of the carrier lifetime is not as well understood and is usually measured. Based on these measurements and work performed on Si devices [12–15], carrier lifetime has been found to increase with increasing temperature. The temperature dependence of carrier lifetime may therefore be modeled as

$$\tau(T) = \tau(T_0)\left(\frac{T}{T_0}\right)^{m} \tag{B.19}$$

Combining (B.18) and (B.19) and assuming a value of n = 2, the resistance temperature dependence of a PIN diode may be approximated as

$$\frac{R_S(T)}{R_S(T_0)} = \left(\frac{T}{T_0}\right)^{2-m} \tag{B.20}$$

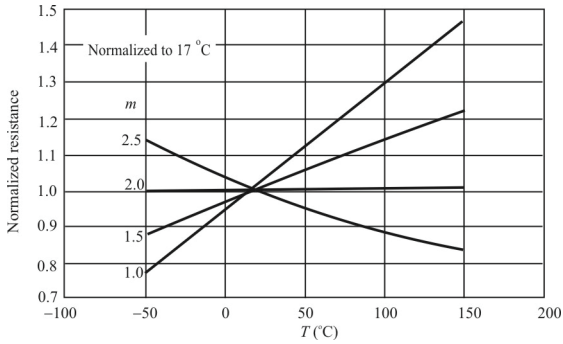


Figure B.19 PIN diode resistance versus temperature.

Figure B.19 [9] shows the resistance ratio described by (B.20) plotted versus temperature using the carrier lifetime coefficient m as a parameter. We can see that, depending on the temperature dependence of the carrier lifetime (as specified by m), the resistance of the PIN diode may increase, decrease, or remain constant over a wide variation in temperature. The value of $m = 2.0$ would indicate no change in resistance with temperature.

The value of m varies over a wide range, and a variety of factors influence the carrier lifetime temperature characteristic. Some of the most important ones are the diode geometry (specifically the I-region width and diameter), the type of material used to passivate the diode surface, and whether impurities have been intentionally introduced into the intrinsic layer in order to reduce carrier lifetime. Of these three factors, the first two are typical of PIN diodes most commonly used in microwave switch applications.

PIN Diode Switch

PIN diodes are commonly used as switching elements to control RF signals. In these applications, the PIN diode can be biased to either a high or low impedance device state, depending on the level of stored charge in the I-region. A simple untuned SPST switch may be designed using either a single series or shunt connected PIN diode, as shown in Figure B.20. The series connected diode switch is commonly used when minimum insertion loss is required over a broad frequency range. This design is also easier to physically realize using printed circuit techniques, since no through-holes are required in the circuit board to provide a connection to ground.

A single shunt mounted diode will, on the other hand, produce higher isolation values across a wider frequency range and will result in a design capable of handling more power since it is easier to heat sink the diode. This configuration does require a very low impedance connection to ground for one end of the diode.

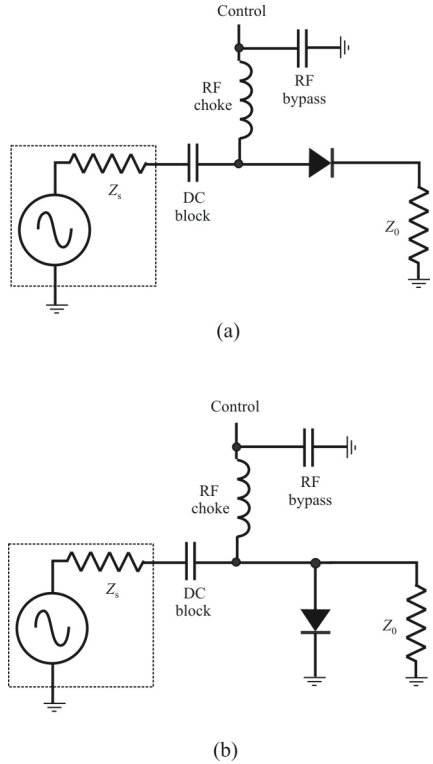


Figure B.20 Simple PIN diode SPST switches: (a) series switch and (b) shunt switch. These are both reflective configurations.

Multithrow switches are used more frequently than single-throw switches. A simple multithrow switch may be designed employing a series PIN diode in each arm adjacent to the common port, as shown in Figure B.21.

Series Connected Switch

The principal operating parameters of a series switch may be obtained using the following equations:

Insertion Loss (*Series Switch*)

$$IL = 20\log_{10}\left(1+\frac{R_S}{2Z_0}\right) \tag{B.21}$$

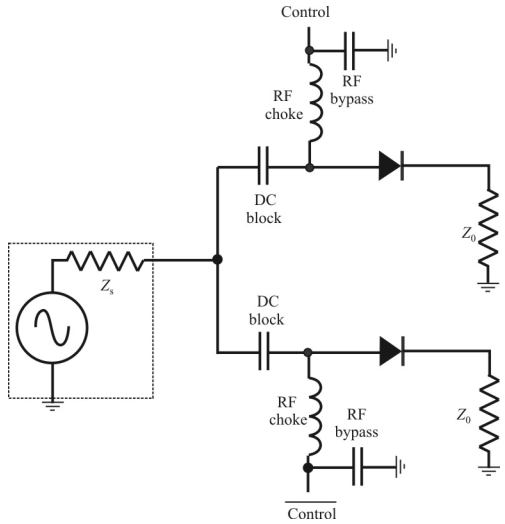


Figure B.21 Simple PIN diode SPDT switch. Again, this is a reflective switch.

This equation applies for an SPST switch and is graphically presented in Figure B.22 for a 50Ω impedance design. For multithrow switches, the insertion loss is slightly higher due to any mismatch caused by the capacitance of the PIN diodes in the OFF arms. This additional insertion loss can be determined from Figure B.21, after first computing the total shunt capacitance of all OFF arms of the multithrow switch.

Isolation (*Series Switch*)

$$ISO = 10\log_{10}\left[1+\frac{1}{(4\pi CZ_0)^2}\right], \qquad \text{dB} \tag{B.22}$$

This equation applies for an SPST diode switch. Add 6 dB for an SPMT switch to account for the 50% voltage reduction across the OFF diode, due to the termination of the generator in its characteristic impedance. Figure B.23 graphically presents isolation as a function of capacitance for simple series switches. These curves are plotted for circuits terminated in 50Ω loads.

RF Power Dissipation (*Series Switch in Forward Bias*)

$$P_{\mathrm{D}} = \frac{4R_{\mathrm{S}}Z_0}{(2Z_0+R_{\mathrm{S}})^2}P_{\mathrm{AV}} \tag{B.23}$$

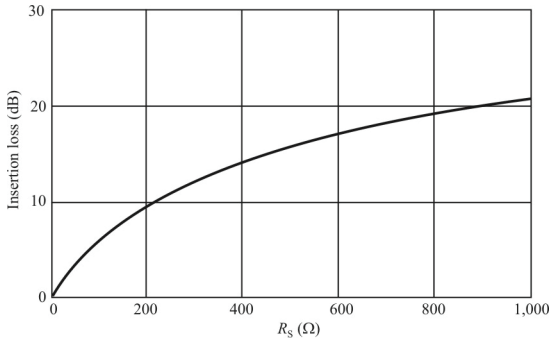


Figure B.22 PIN series insertion loss.

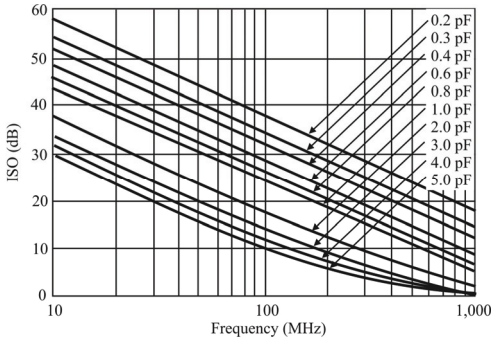


Figure B.23 Isolation for a series PIN diode. For $0.2\,\text{pF} \leq C_T \leq 5\,\text{pF}$.

For $Z_0 >> R_S$, this becomes

$$P_D = \frac{R_S}{Z_0} P_{AV} \tag{B.24}$$

where the maximum available power is given by

$$P_{AV} = \frac{V_G^2}{4Z_0} \tag{B.25}$$

It should be noted that (B.23) and (B.24) apply only for perfectly matched switches. For SWR (S) values other than unity, multiply these equations by

$$\left(\frac{2S}{S+1}\right)^2$$

to obtain the maximum required diode power dissipation rating.

Peak RF Current (*Series Switch*)

$$I_{PK} = \sqrt{\frac{2P_{AV}}{Z_0}}\left(\frac{2S}{S+1}\right) \tag{B.26}$$

In the case of a 50Ω system this becomes

$$I_{PK} = \sqrt{\frac{P_{AV}}{25}}\left(\frac{2S}{S+1}\right) \tag{B.27}$$

Peak RF Voltage (*Series Switch*)

$$V_{PK} = \sqrt{8Z_0P_{AV}} \tag{B.28}$$

For a 50Ω system this becomes

$$\begin{aligned} V_{PK} &= \sqrt{2\times 50\times P_{AV}}\left(\frac{2S}{S+1}\right) \\ &= 10\sqrt{P_{AV}}\left(\frac{2S}{S+1}\right) \end{aligned} \tag{B.29}$$

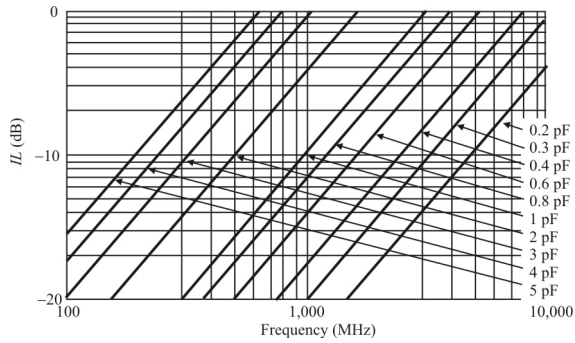


Figure B.24 PIN diode shunt insertion loss. Typical insertion loss of a shunt PIN diode in a 50Ω system for $0.2 \le C_T \le 5\,\text{pF}$.

Shunt Connected Switch

Figure B.20(b) shows a typical shunt connected PIN diode switch. This shunt diode switch offers high isolation for many applications, and since the diode may be heat sunk at one electrode it is capable of handling more RF power than a diode in a series type configuration. In shunt switch designs, the isolation and power dissipation are functions of the diode's forward resistance, whereas the insertion loss is primarily dependent on the capacitance of the PIN diode. The principal equations describing the operating parameters of shunt switches are in the following.

Insertion Loss (*Shunt Switch*)

$$IL = 10\log_{10}\left[1+(\pi fCZ_0)^2\right], \qquad \text{dB} \tag{B.30}$$

This equation applies for both SPST and SPMT shunt switches and is graphically presented in Figure B.24 for a 50Ω load impedance.

Isolation (*Shunt Switch*)

$$ISO = 20\log_{10}\left[1+\frac{Z_0}{2R_S}\right] \tag{B.31}$$

This equation, which is illustrated in Figure B.25, applies for an SPST shunt switch. Add 6 dB to these values to obtain the isolation for a multithrow switch.

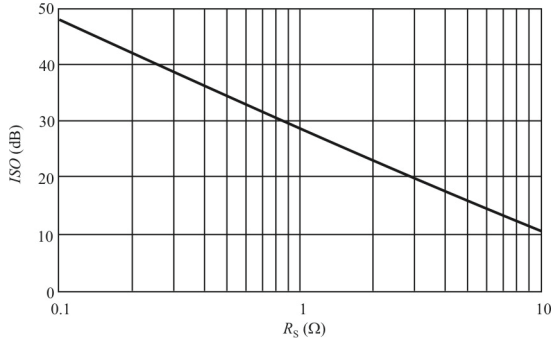


Figure B.25 Isolation of shunt PIN diode in a 50Ω system.

Peak RF Voltage (*Shunt Switch*)

$$V_{PK} = \sqrt{2Z_0 P_{AV}} \left(\frac{2S}{S+1} \right) \tag{B.32}$$

In the case of a 50 Ω system this becomes

$$V_{PK} = 10\sqrt{P_{AV}} \left(\frac{2S}{S+1} \right) \tag{B.33}$$

B.4.2.2 PIN Diode Compound SPST Switch

Improved performance is obtained by using "compound switches," which are combinations of series and shunt connected PIN diodes, in each arm of the switch. For narrowband applications, quarter-wave spaced multiple diodes may also be used in various switch designs to obtain improved operation. Figure B.26 shows such a compound SPST switch commonly used in broadband designs. The diode is in a "pass power" condition when it is forward biased and presents a low forward resistance between the RF generator and load. For the "stop power" condition, the diode is at zero or reverse bias so that it presents a high impedance between the source and load. In series connected switches, the maximum isolation obtainable depends primarily on the capacitance of the PIN diode, while the insertion loss and power dissipation are functions of the diode resistance.

B.4.2.3 PIN Diode Frequency Effects

Caverly and Hiller reported on the results of analyzing the variation of the PIN diode impedances as the frequency of operation changed [9]. In a typical EW

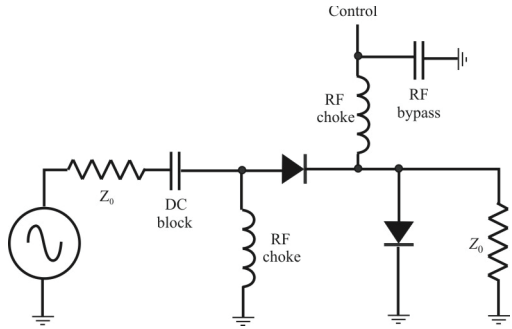


Figure B.26 PIN diode compound SPST switch.

application of RF switches, the frequency range that a switch must handle is quite wide. Variation of the switch characteristics with frequency is therefore an undesirable property. We will examine such frequency effects in this section.

When first introduced, the PIN diode was used in VHF and microwave frequency applications, but there has been significant interest in circuits containing PIN diode switches at frequencies extending significantly lower [16, 17]. In analyzing circuits containing PIN diodes, a common assumption used is that the total PIN diode impedance is controlled by the charge stored in the I-region by the AC and DC currents (I_k and I_0, respectively). At high frequencies, this assumption is valid [18]. At low frequencies, the PIN diode's distortion properties have been found to be a function of this time-varying stored charge [19–21]. In addition to these nonlinearities, resistive and reactive effects are introduced by the semiconductor junctions that may be significant enough to dominate the overall device impedance. Earlier theoretical work has indicated that the junction effects are sensitive to device parameters such as I-region carrier lifetime (τ) and I-region thickness (W) [22–24] and constitute a substantial component of the overall device impedance.

The total PIN diode impedance becomes virtually a constant resistance above a specific frequency, but at low frequencies, a significant capacitive or inductive component may be present. Analysis of the frequency dependent impedance is especially important in short-lifetime PIN diodes [25, 26], where the so-called low-frequency effects may actually extend to several hundred megahertz.

The dogma for high-frequency PIN diode series resistance is that it is a constant. While this is basically true at high frequencies, usage of PIN diodes for lower-frequency applications has increased, and at lower frequencies this dogma no longer applies. The expression for the series resistance of a PIN diode, R_S, is given by (B.12), which can be expressed as

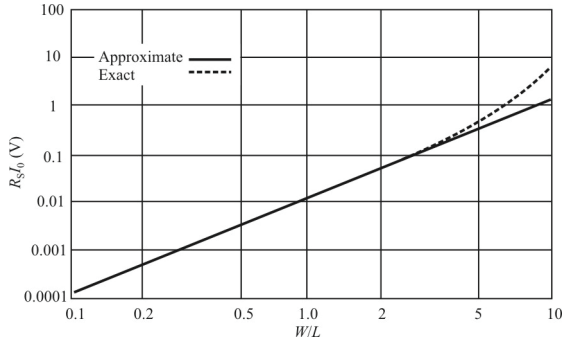


Figure B.27 PIN diode I-region resistance.

$$R_S = \left(\frac{W}{L}\right)^2 \frac{V_T}{2I_0} \tag{B.34}$$

where $V_T = kT/q$, μ is the ambipolar mobility, L (the diffusion length) $= (D\tau)^{1/2}$, D is the ambipolar diffusion constant, and Q $(= I_0\tau)$ is the I-region stored charge. Equation (B.34) applies to thin I-region PIN diodes where $W/2L$ is less than 1, and at frequencies well above $1/\tau$ (say > $10/\tau$). A more exact equation that applies to all PIN diodes is given by [27]

$$R_S = \frac{4V_T \sinh\left(\frac{W}{2L}\right)\left[\tan^{-1}\left(e^{W/2L}\right) - \pi/4\right]}{I_0} \tag{B.35}$$

Equation (B.35) reduces to (B.34) if $W/2L$ is less than unity. Both equations apply only to the I-region resistance and are high-frequency equations such that all effects of the junctions (a low-frequency phenomenon) are neglected. Figure B.27 [27] shows a plot of the I-region resistance-bias current product $R_S I_0$ versus the ratio W/L, computed using (B.34) and (B.35). Note that for values of W/L greater than approximately 2.5, (B.34) underestimates the $R_S I_0$ product.

At DC, the total PIN diode impedance is the dynamic resistance of the junctions, namely, $2V_T / I_0$, and can be determined from the DC diode characteristics. At low frequencies, both the NI and the IP junctions (which can individually be modeled as an impedance, Z_J) cause reactive effects associated with diffusion of carriers across the junction. In the intrinsic region, reactive effects caused by conductivity modulation also contribute significantly to the total PIN diode impedance, Z_T, such that Z_I may be only a small part of Z_T in the low-

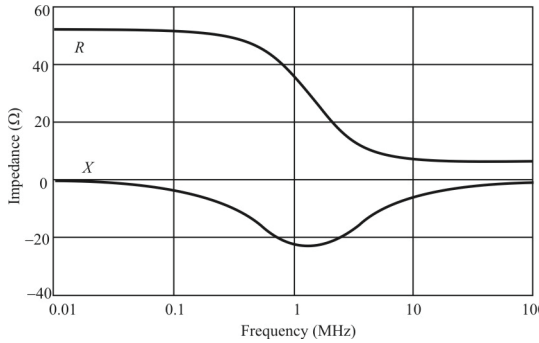


Figure B.28 Impedance PIN 1 mA. Impedance values of Si PIN diode at DC bias of 1 mA ($W = 0.42$ mils, $\tau = 116$ ns). The low-frequency reactance in this case is capacitive.

frequency limit. The total PIN diode impedance as a function of frequency may then be written as

$$Z_T(\omega) = Z_I(\omega) + 2Z_J(\omega) \tag{B.36}$$

The factor of 2 in (B.36) appears because of the two junctions in the ideal PIN diode.

Figures B.28 and B.29 show computed resistance and reactance plotted versus frequency for Si PIN diodes of different physical and electrical characteristics. Figure B.28 [27] presents results for a thin I-region, short-lifetime diode ($W = 0.42$ mils, $\tau = 116$ ns). Note the transition region in the 1–10 MHz range where the impedance changes dramatically, indicating the region where the impedance changes from being dominated by the junctions to being dominated by I-region effects. There is good agreement between the model and the experimental data.

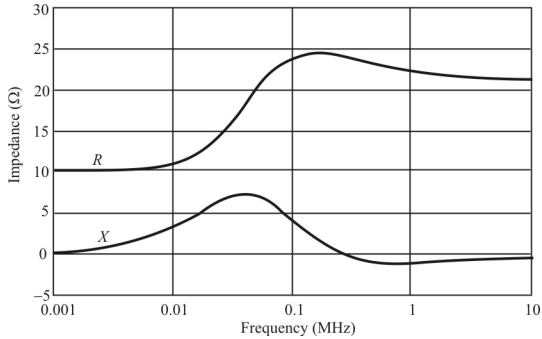


Figure B.29 Impedance PIN 5 mA. Silicon PIN diode at DC bias current of 5 mA ($W = 10.8$ mils, $\tau = 3.8$ μs). The low-frequency reactance in this case is inductive.

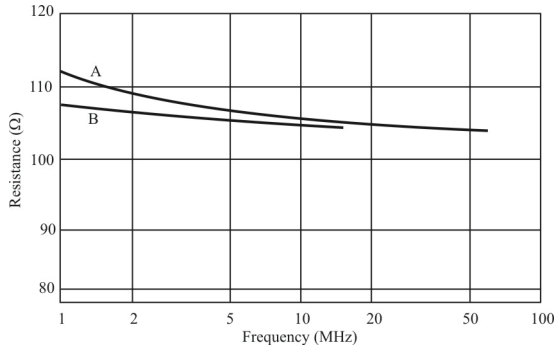


Figure B.30 Frequency characteristics two PIN diodes. Curve A: $\tau = 2.54$ μs, $W = 8.8$ mils, curve B: $\tau = 10$ μs, $W = 17.5$ mils.

Figure B.29 [27] shows the model on a thicker, longer lifetime silicon PIN diode ($W = 10.8$ mils, $\tau = 3.8$ μs). The low-frequency reactance is inductive in nature, as predicted by the model, and remains inductive up to approximately 200 kHz. At higher frequencies, the reactance becomes capacitive and completely disappears at high frequencies. The experimental data is in good agreement with these models here as well.

Resistance characteristics for two PIN diodes are shown in Figure B.30 with the parameters for the diodes indicated in the caption. Note that the thicker PIN diode (curve B) gives the flatter resistance-frequency response.

B.4.2.4 Distortion in Forward- and Reverse-Biased PIN Diodes

A PIN diode is a semiconductor device and as such there are bias and RF points that when reached will cause IMD products. Caverly and Hiller documented research results on these IMD products generated by PIN diodes when forward and reverse biased [19]. A PIN diode switch may contain both forward- and reverse-biased diodes, each contributing a measure of distortion.

Forward Bias

When forward-biased, for the simple switch illustrated in Figure B.31, the second order intercept point is given by

$$IP2 = \frac{Z_0}{2}\left(\frac{R_S + 2Z_0}{R_S}\right)^2 Q^2 \left(\frac{1}{\omega_1} + \frac{1}{\omega_2}\right)^{-2} \tag{B.37}$$

while the third order intercept point is given by

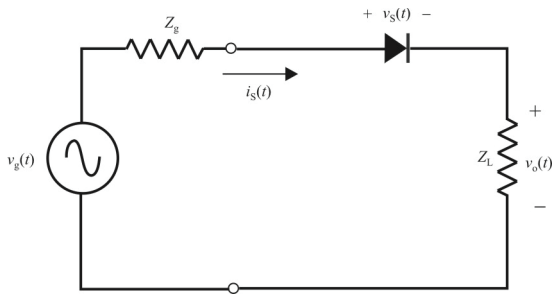


Figure B.31 High-frequency PIN diode circuit sans bias and control components.

$$IP3 = \frac{Z_0}{2}\sqrt{V_T}\,\frac{Z_0 + R_S}{R_s^{1.5}}(\omega_1 Q)^{1.5}\left[\frac{1}{4} + 1/\left(\frac{\omega_2}{\omega_1} + \sqrt{\frac{\omega_2}{\omega_1}}\right)\right]^{-1} \tag{B.38}$$

where

R_S is the series resistance of the I-region.
$Q = I_0\tau$ is the stored DC charge.
$V_T = kT/q$.

For switches (where $R_S << Z_0$) and when the frequencies are close ($\omega_1 \approx \omega_2 \triangleq \omega$), these expressions reduce to

$$IP2 = \frac{Z_0^3}{2}\left(\frac{\omega Q}{R_S}\right)^2 \tag{B.39}$$

and

$$IP3 = 0.22 Z_0^2\left(\frac{\omega Q}{R_S}\right)^{1.5} \tag{B.40}$$

In (B.39) and (B.40), the important factor controlling switch distortion at a given frequency is the ratio Q/R_S (Figure B.32). Since $Q = I_0\tau$ and $R_S = W^2/2\mu Q$ [28], then Q/R_S is proportional to $(\tau/W)^2$. In single-diode switch circuits, the distortion properties are improved if PIN diodes are selected or operated that maximize either the ratio Q/R_S or τ/W. Equations (B.39) and (B.40) show that, at a fixed

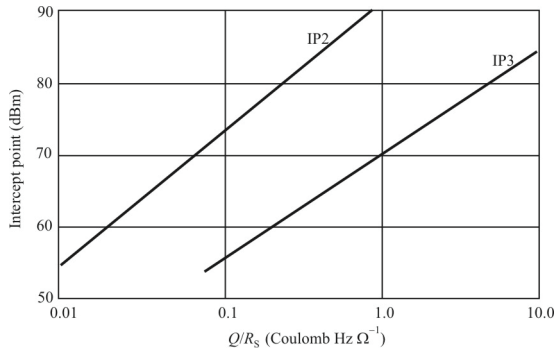


Figure B.32 PIN diode forward bias IPs.

frequency, increasing the quantity Q/R_S by a factor of 2 produces a four-fold increase (6 dB) in the second-order intercept point and a factor of 2.8 increase (4.5 dB) in the third-order intercept point. Distortion in single PIN diode switch circuits has been analyzed by Hiller and Caverly [29].

Reverse Bias

The primary cause of reverse-biased PIN diode distortion is the small but finite time varying capacitance induced by the microwave or RF signal. An ideal PIN diode operated at a reverse bias voltage beyond the punch through voltage and at frequencies higher than the dielectric relaxation frequency will exhibit no change in capacitance with increasing voltage, and therefore generate no distortion. The actual PIN diode, because of the diffusion tail phenomena, will exhibit a definite capacitance change and will generate real distortion [19].

Distortion introduced by the reverse-biased PIN diode is based on the nonlinear time varying current passing through the diode [28, 29]:

$$i_S(t) = \frac{dQ(t)}{dt} = \frac{d[C(t)v_S(t)]}{dt} = C(t)\frac{dv_S(t)}{dt} + v_S(t)\frac{dC(t)}{dt} \tag{B.41}$$

where $C(t)$ is the PIN diode reverse bias capacitance and $v_S(t)$ is the RF voltage across the diode. If $C(t)$ is a time-varying capacitance, then $i_S(t)$ will contain harmonic and intermodulation products, in addition to the fundamental components, which will produce signal distortion in a microwave or RF switching circuit consisting of a series connected PIN diode between generator and load, as illustrated in Figure B.31.

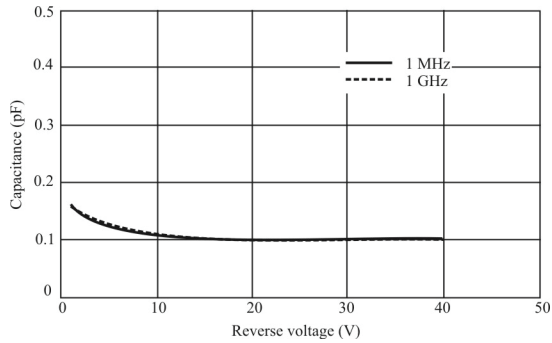


Figure B.33 Capacitance of PIN diode at 7 microns.

The capacitance of the ideal PIN diode is generally assumed constant for all reverse bias voltages beyond the punch-through voltage. In actual PIN diodes, however, the boundaries between the intrinsic layer and the end regions are not exactly abrupt, but rather exhibit a continuously increasing doping concentration that approaches the end region doping concentration. These so-called diffusion tails cause the reverse-biased PIN diode's capacitance to continue to decrease for voltages beyond punch-through, independent of frequency, since the depletion layer will advance into the end regions with increasing reverse bias voltage. The rate of capacitance change with reverse voltage is dependent on this diffusion profile shape, but may be directly determined from measured C-V data. Figures B.33 and B.34 [27] show the C-V characteristic for 7- and 100-micron I-region width PIN diodes using measurement frequencies of 1 MHz and l,000 MHz that illustrate this. Thicker PIN diodes exhibit flatter C-V profiles than thinner diodes since the depth that the depletion boundary extends into the end region is a smaller portion of the overall device thickness. This results in a lower rate of capacitance change with reverse voltage for thick PIN diodes.

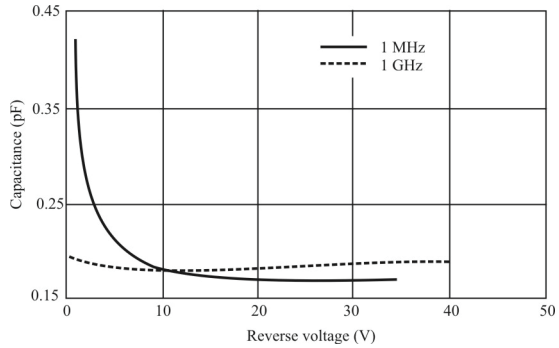


Figure B.34 Capacitance of PIN diode at 100 microns.

Table B.1 Estimated PIN Diode Intercept Points

		Reverse Bias				Forward Bias	
Width (microns)	**Frequency (MHz)**	**IP2** –10V	**IP3** –10V	**IP2** –20V	**IP3** –20V	**IP2** 10 mA	**IP3** 10 mA
7	10	94 dBm	58 dBm	102 dBm	62 dBm	39 dBm	28 dBm
	1,000	61 dBm	42 dBm	72 dBm	47 dBm	79 dBm	58 dBm
100	10	92 dBm	60 dBm	112 dBm	70 dBm	77 dBm	56 dBm
	1,000	85 dBm	57 dBm	91 dBm	60 dBm	117 dBm	86 dBm

Source: [19].

Equations for the two-tone intermodulation distortion intercept points have been derived from (B.41) for the series connected reverse-biased PIN diode. They are related to the derivatives of the *C-V* data, measured at the input frequency, and are expressed as

$$IP2 = \frac{1}{32}\frac{d^2C}{dV^2}Z_0^3(\omega_1 + \omega_2)^2 \tag{B.42}$$

and

$$IP3 = \frac{1}{12}\frac{d^2C}{dV^2}Z_0^2(2\omega_1 + \omega_2) \tag{B.43}$$

where ω_1 and ω_2 are the two-tone input frequencies. These intercept point equations are referenced to the power available from the source. We can see that when reverse-biased, the intercept point decreases (distortion increases) as frequency increases while when forward-biased, the reverse is true: as the frequency increases, the intercept point increases as well. Also, thicker I-region PIN diodes exhibiting a flatter *C-V* characteristic are expected to perform with better distortion properties than thinner diodes.

Reverse bias distortion intercept points are shown in Table B.1 based on C-V measurements and their derivatives taken at 10 MHz and 1,000 MHz. Table B.2 also illustrates forward bias distortion on the same PIN diodes based on 10 mA resistance and stored charge data [19, 29]. These results imply that for the 7 micron PIN diode the reverse bias (–10 V) third-order relative distortion (dBc) degrades by 32 dB as frequency increases from 10 to 1,000 MHz. In the same frequency range the forward bias distortion improves by 60 dB. Also, at 10 MHz the forward bias distortion is the significant contributor, whereas at 1,000 MHz it is the reverse bias distortion that predominates. In all cases, the reverse bias distortion improves with increasing reverse bias voltage.

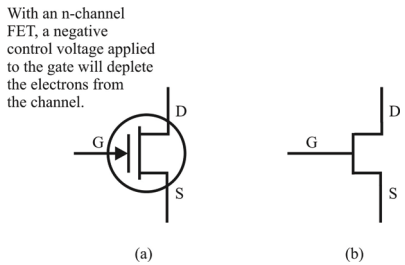


Figure B.35 (a) MOSFET depletion mode. Note that in many applications the source is connected to the source. (b) This is a simplified symbol for a MOSFET, since the source and drain are interchangeable.

Summary

The fundamental conclusion of this section is that the magnitude of the reverse bias distortion is inversely proportional to the slope of the C-V characteristic (for second-order distortion) and frequency. This is in contrast to the forward bias PIN diode switch distortion case where the distortion improves with increasing frequency. Also thick PIN diodes tend to exhibit less distortion than thin diodes.

B.4.3 FET RF Switches

B.4.3.1 Introduction

In this section we discuss the theory of operation of RF/microwave FETs, with emphasis on the pHEMT, and its implementation in RF/microwave switches, as well as MESFET switches. We will also describe criteria with which the relative advantages of PIN diode switches and pHEMT switches compare for specific types of applications.

FETs are versatile small-signal switches that are highly flexible in their application. The FET structure provides significant RF isolation between the control voltage at the gate and the signal path between the source and drain. Switching speeds in the single-digit nanosecond range can be obtained, which rivals many low-cost diode switch options. A FET switch may also be designed to have a well-controlled variable resistance between the ON and OFF states.

The primary drawbacks are relatively high ON resistance and limited signal handing capability. The latter can result in interaction of the signal and control, as noted above for high power switching in PIN diodes. In the case of the FET, signals levels should not modulate the gate-source control voltage.

FETs that are used as switches are normally of the depletion type (see Figure B.35). In this type of device the channel with no bias applied is in the conductive (ON) state and current can flow from source to drain. A negative control voltage applied to the gate depletes the carriers under the gate region, eventually cutting

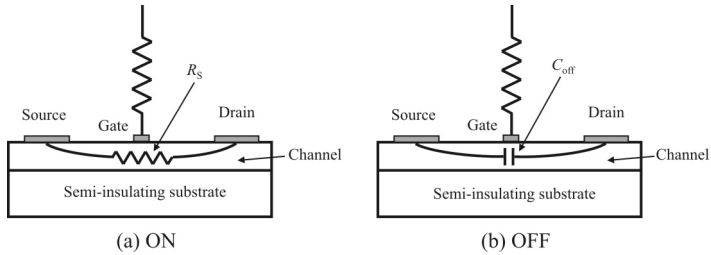


Figure B.36(a, b) FET equivalent circuits: (a) ON: channel is resistive, and (b) OFF: channel is capacitive.

the channel off, and it no longer conducts. The reason for the depletion configuration is that when a negative voltage is applied to the gate, it repels the electron carriers in the channel immediately beneath it (*n*-type semiconductor). Electrons are the preferred channel carriers because their mobility is significantly higher than holes (lack of electrons, or *p*-type semiconductor) and the operating frequency can be correspondingly higher (faster switching speed).

B.4.3.2 GaAs pHEMT Switch Theory of Operation

The FET functions as a three-port device, where the channel between source and drain ports forms a conduction path for the RF signal and the gate port controls whether an RF signal is blocked or may pass. A DC voltage applied between the gate and channel provides this control. As mentioned, most switching FETs use a depletion mode configuration, where the channel is normally in its low impedance state with no control voltage applied and in its high impedance, pinched off, state when a negative voltage with respect to the channel is applied. See Figure B.36.

FET Switch Configurations

Figure B.37 shows a single FET configured as a simplified single pole, single throw (SPST) switch. The first challenge in fabricating a switch using a GaAs FET

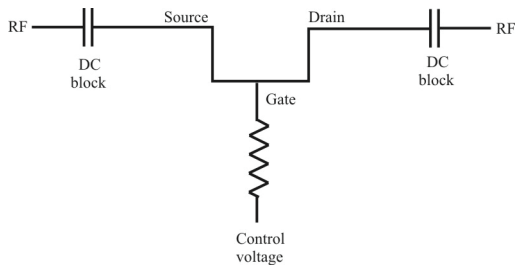


Figure B.37 FET RF switch.

is to isolate the DC gate control from the RF path. This is done by using a 5 kΩ to 10 kΩ resistor in series with the gate of the FET. This is a very simple bias network that has many advantages. The next step is to DC block the RF source and drain ports using a capacitor with adequately low capacitive reactance at the desired frequency of operation. This creates an SPST switch.

The insertion loss of the ON path of the switch will be affected by the channel resistance. Likewise the isolation of the OFF path is limited by the capacitance created by the source and drain spacing as well as the FET physical size (periphery).

Hence, a balance of channel resistance (R_S) and off capacitance (C_{off}) must be met. The following equations show the relationship of the R_S and C_{off} as expressed in insertion loss (decibels) and isolation loss (decibels) for a single series FET SPST switch. The following is an example of a simple switch. Assume the FET has the following characteristics: $R_S = 3\ \Omega$ and $C_{off} = 0.25$ pF ($X_C = 636\Omega$ at $f = 1$ GHz). Then

$$IL = 20\log_{10}\left(1+\frac{R_S}{2Z_0}\right) = 20\log_{10}\left(1+\frac{3}{2\times 50}\right) = 0.257\,\text{dB} \tag{B.44}$$

where Z_0 is the system characteristic impedance [normally 50Ω and that is what was assumed in (B.44)], and

$$ISO = 20\log_{10}\left(1+\frac{X_C}{2Z_0}\right) = 20\log_{10}\left(1+\frac{636}{2\times 50}\right) = 17.34\,\text{dB} \tag{B.45}$$

Applying the values of 3Ω and 0.25 pF would result in an insertion loss of 0.257 dB and an isolation of 17.34 dB at 1 GHz. However, if the frequency is lowered to 100 MHz, then an isolation of 36.2 dB is achieved. This is due to the fact that the isolation of the series FET is a result of its capacitive reactance which is frequency dependent, an important feature when selecting FET parameters. To improve the isolation at higher frequencies, a shunt FET can be used following the series FET. In this position, the FET must be ON in order to increase isolation and OFF to put it in the insertion loss state. This requires two separate control voltages for the switch (see Figure B.38). The resulting isolation would be an additional 19.4 dB. Since the shunt FET has no appreciable reactive component when it is biased to its *ON* state, isolation will stay relatively unchanged over frequency. Note that again, this equation provides a good approximation of the shunt FET additional isolation, assuming ideal conditions. Phase length, ground coupling, and so forth will impact actual results.

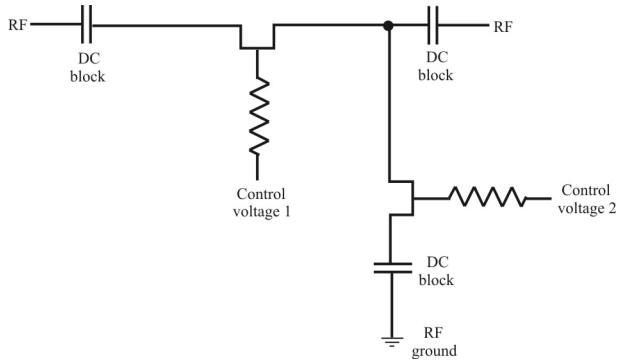


Figure B.38 Series/shunt FET SPST RF switch.

Power Handling Capabilities

FET-based devices can handle several watts while consuming very little power. This feature is very attractive for tactical applications. The *ON* condition of the FET can handle typically more power than the *OFF* state. This is due to the voltage breakdown (V_B) created by gate/source and gate/drain spacing limitations. A simple way to rationalize this is to compare the *ON* limitation being channel current (I_{dss}), which can be increased significantly by making the channel wider. This can be done by adding periphery to the FET; however, if the gate spacing is increased significantly, R_S becomes greatly affected and a significant loss penalty will result.

A way to overcome voltage limitation is by adding FETs in series (see Figure B.39). Doing so distributes the voltage across many FETs and thus reduces the voltage on the individual devices. This technique can increase power handling significantly. This SPDT can handle 10 W with only 0.1 dB of compression.

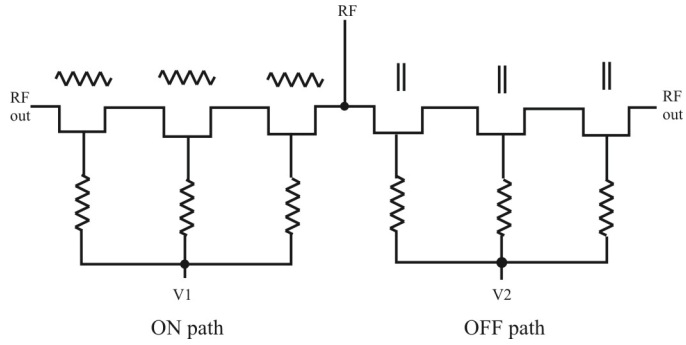


Figure B.39 High-power series FET SPDT RF switch.

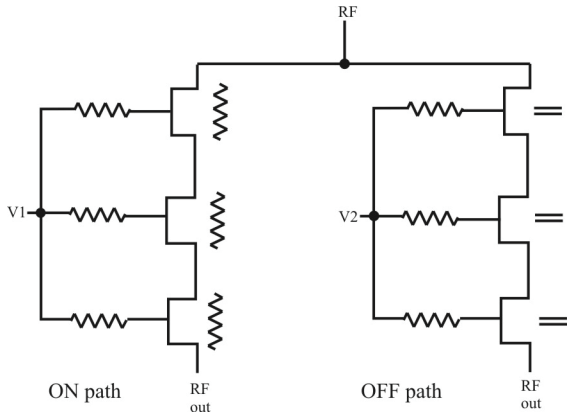


Figure B.40 High-power parallel FET SPDT RF switch.

In another way to address the handling of higher-power requirements, many FET switches can be connected in parallel and switched at the same time. Power levels of several hundreds of watts can be switched this way (see Figure B.40).

A simple series SPDT switch configured with FETs is shown in Figure B.41. Bypass, DC blocks, and choke details are not included. We can see that such a switch can be quite small. To improve the isolation, a shunt leg can be attached to the output as illustrated in Figure B.42. The isolation improves by up to 20 dB with this arrangement. Both of the switches in Figures B.41 and B.42 are reflective

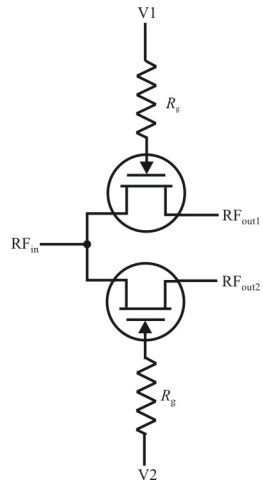


Figure B.41 FET SPDT RF switch.

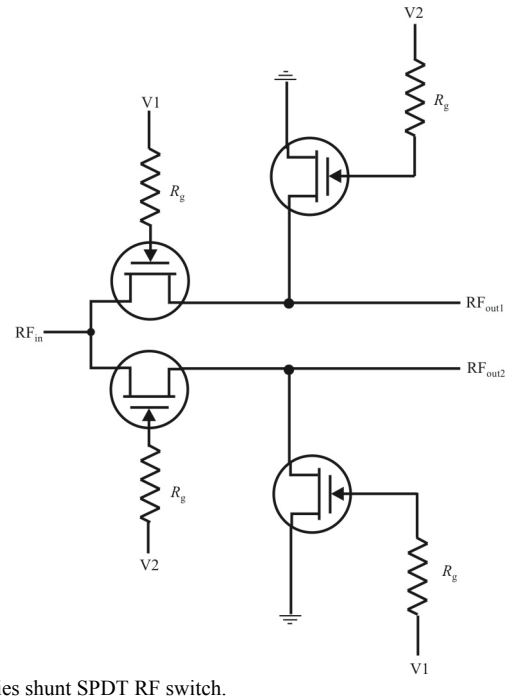


Figure B.42 FET series shunt SPDT RF switch.

arrangements since the output is shorted to ground when the switch is open. For an absorptive switch, the shunt paths across the outputs are terminated in Z_0, rather than to ground, which is illustrated in Figure B.43.

B.4.3.3 MESFETs

MESFETs are similar to JFETs in construction and terminology. The difference is that, instead of using a *p*-*n* junction for a gate, a Schottky (metal-semiconductor) interface (junction) is employed. Because of this similarity, and since the operation of a JFET is perhaps the easiest to understanding the world of FETs, we begin with a brief discussion of JFETs.

Junction FETs [30]

Junction Field-Effect Transistors (JFETs) are so named because they are formed by the junction of *p*-type and *n*-type silicon as illustrated in Figure B.44. Observe that, when no signal is applied to the gate terminal, the default depletion region

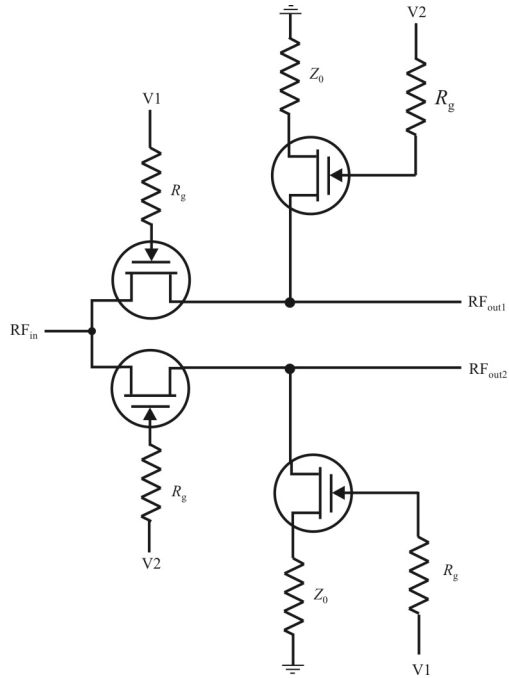


Figure B.43 FET series shunt SPDT absorptive SPDT RF switch.

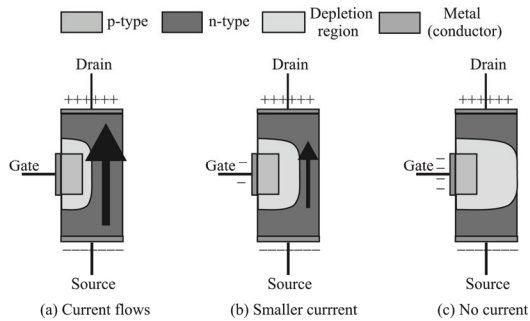


Figure B.44 GaAs pHEMT (JFET) RF switch principles of operation.

[Figure B.44(a)] leaves a channel of *n*-type silicon between the source and the drain, thereby allowing current to flow.

This means that these devices are *ON* in their quiescent state, and we have to apply a signal to the gate terminal in order to turn them *OFF*. If we apply a small negative potential to the gate terminal [Figure B.44(b)], the electrons on the gate attract the positive holes in the *p*-type silicon and repel the negative electrons in the *n*-type silicon. This reduces the size of the conducting channel between the source and the drain terminals. In turn, this increases the resistance of the channel and reduces the flow of current between the source and the drain terminals. If we keep on increasing the negative potential on the gate terminal, at some stage the depletion zone will completely block the channel, as illustrated in Figure B.44(c).

Observe that the transistor shown in Figure B.44 is an *n*-channel device. We can also create an equivalent *p*-channel component by swapping the *p*-type and *n*-type diffusion regions. JFETs are normally operated with the gate-source junction reverse biased, and are designed for the gate-source operating voltage to be negative for *n*-channel, or positive for *p*-channel. If the junction in a JFET becomes forward-biased, the junction will begin to conduct and signal distortion results.

In a MOSFET, the gate is an electrode or semiconductor layer separated from the source-drain channel by an insulating layer. This allows the FET to be designed to operate over a range of voltages that overcome the forward bias junction limitation of the JFET, as well as considerably reducing gate leakage current.

Given the possibility of positive gate voltages, the *n*-MOSFET source-drain conduction channel can be eliminated. Instead, a *p*-type substrate between the *n*-type source and drain regions is turned into an *n*-type conduction channel when a positive gate voltage is applied, which attracts electrons from the source into the region next to the gate, effectively repelling the holes away from the gate. This type of MOSFET is known as an enhancement-mode device, which does not conduct if the gate is at zero volts relative to the source, but requires a positive voltage to begin conduction (as usual, reverse *n* and *p* types and voltages for the *p*-MOSFET description).

MESFET

Figure B.45 shows a diagram of *gallium arsenide* (GaAs) MESFET, also known as a GaAsFET. A side view is provided in Figure B.46. MESFET is nothing but a JFET fabricated in GaAs, which employs a metal-semiconductor gate region (a Schottky diode). The device operates in essentially the same way as does a junction-gate FET, except that instead of a gate-channel junction, there is a gate-channel Schottky barrier. The depletion region associated with this barrier will control the effective height of the conducting channel and can thereby control the drain-to-source current of the device.

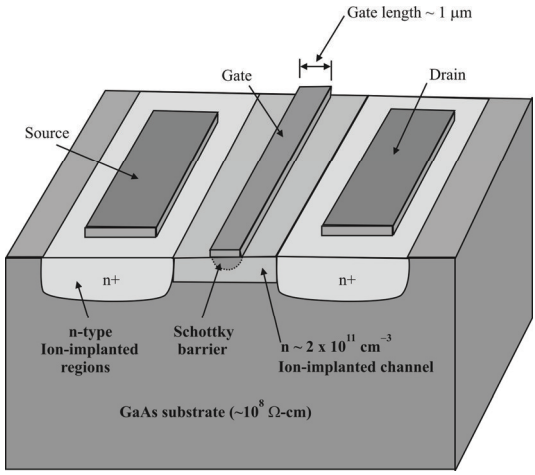


Figure B.45 GaAs MESFET structure.

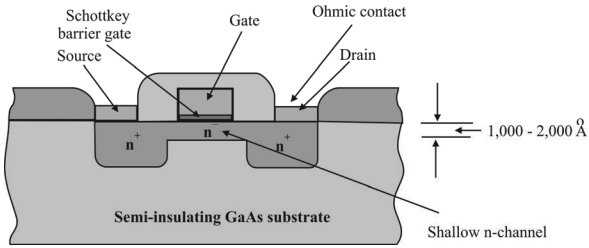


Figure B.46 Side view of MESFET.

Table B.2 Carrier Mobility in Silicon

Doping (cm^{-3})	Electron Mobility (cm^2 V-sec^{-1})	Hole Mobility (cm^2 V-sec^{-1})
10^{14}	1,500	450
10^{15}	1,500	450
10^{16}	1,400	440
10^{17}	1,200	410
10^{18}	800	200
10^{19}	200	150
10^{20}	90	50

The width of this depletion region will increase with increasing gate voltage so that we see again that the gate will be the control electrode and as long as the Schottky barrier is reverse biased, the gate current will be very small.

Electron mobility in GaAs (~8,500 cm^2 V-s^{-1}) is much higher than that of silicon and allows MESFET operation at frequencies higher than can be achieved with silicon devices (see Table B.2).

The MESFET also possesses a very short channel length (~1 μm). This results in very short channel transit times for electrons. As a result, MESFET can operate well into the 10 GHz range and higher. Thus, applications of MESFETs were initially in microwave circuits for high frequency performance.

The GaAs FET, a bulk-current-conduction majority-carrier device, is fabricated from bulk gallium arsenide by high-resolution photolithography and ion implantation into a semi-insulating GaAs substrate (see Figure B.46). Processing is relatively simple, requiring no more than six to eight masking stages.

GaAs MESFETs are similar to silicon MOSFETs. The major difference is the presence of a Schottky diode at the gate region, which separates two thin *n*-type active regions, that is, source and drain, connected by ohmic contacts (see Figure B.46). It should be noted that both D type and E type MESFETs, that is, *ON* and *OFF* devices, operate by the depletion of an existing doped channel. This can be contrasted with silicon MOS devices where the E (enhancement) mode transistor functions by inverting the region below the gate to produce a channel, while the D (depletion) mode device operates by doping the region under the gate slightly in order shift the threshold to a normally *ON* condition.

The D-MESFET is normally *ON* and its threshold voltage, V_{tdep}, is negative. The E-MESFET is normally *OFF* and its threshold V_{tenh} is positive. The threshold voltage is determined by the channel thickness, a, and concentration density of the implanted impurity, N_D. A highly doped, thick channel exhibits a larger negative threshold voltage. By reducing the channel thickness and decreasing the concentration density a normally *OFF* enhancement mode MESFET with a

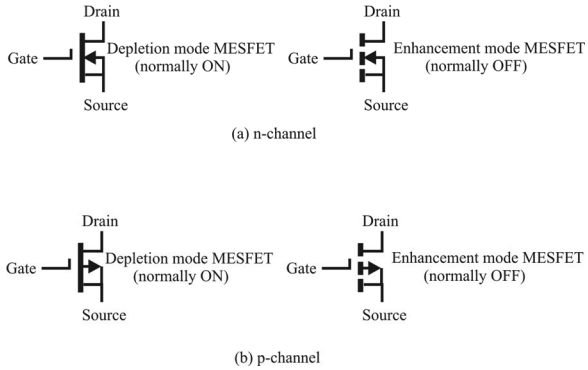


Figure B.47(a, b) MESFET circuit symbols: (a) n-channel and (b) p-channel.

positive threshold-voltage can be fabricated. Circuit symbols for the depletion and enhancement mode MESFETs are indicated in Figure B.47.

The MESFET has a maximum gate to source voltage, V, of about 0.7–0.8 V owing to the diode action of the Schottky diode gate. Since the principle underlying the operation of MESFETs is based upon the behavior of metal-semiconductor interface, we will briefly outline some of the features that characterize such an interface.

Characteristics of Schottky Barriers

An electrostatic potential barrier (known as a Schottky barrier) is created at the interface when a metal is brought into contact with a semiconductor. This barrier is the result of the difference in the work function of the two materials. To appreciate the physical nature of the barrier, we can model the interface by visualizing a situation whereby the metal is gradually brought toward the semiconductor surface until the separation becomes zero.

As this separation between the metal-semiconductor surface is reduced, the induced charge in the semiconductor increases, while at the same time the space charge layer widens. A greater part of the contact potential difference begins to appear across the space charge layer within the semiconductor. Because the carrier concentration in the metal is several orders of magnitude larger than that in the semiconductor when the separation is brought to zero, the entire potential drop then appears within the semiconductor itself. This is in the form of a depletion layer situated adjacent to the metal and extending into the semiconductor.

Depletion-Mode MOSFETs

As we see in Figure B.48, the control terminal (gate) is connected to a conducting plate, which is insulated from the silicon by a layer of nonconducting oxide. In the

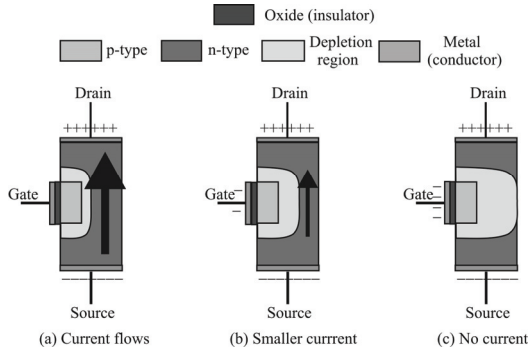


Figure B.48(a–c) MOSFET principles of operation in the depletion mode.

original devices, the conducting plate was metal—hence, the term "metal-oxide"—but this is now a misnomer because modern versions tend to use a layer of *polycrystalline silicon* (*polysilicon*). When a signal is applied to the gate terminal, the plate, insulated by the oxide, creates an electromagnetic field, which turns the transistor ON or OFF—hence the term "field effect."

When no signal is being applied to the gate, the channel of *n*-type silicon between the source and the drain allows current to flow. This means that these devices are ON by default and we have to apply a signal to the gate terminal in order to turn them OFF. If we apply a small negative potential to the gate terminal [Figure B.48(b)], the electrons on the gate repel the negative electrons in the *n*-type silicon. This leaves an area depleted of electron charge carriers.

We can think of this as pinching a garden hose to reduce the flow of water. Applying a negative potential to the gate terminal reduces the size of the conducting channel between the source and the drain. In turn, this increases the resistance of the channel and reduces the flow of current between the source and the drain terminals. If we keep on increasing the negative potential on the gate terminal, at some stage the depletion zone will completely block the channel as illustrated in Figure B.48(c).

Observe that the transistor shown in Figure B.48 is an *n*-channel device. We can also create an equivalent *p*-channel component by swapping the *p*-type and *n*-type diffusion regions.

Enhancement-Mode MOSFETs

Enhancement-mode MOSFETs (Figure B.49) are formed from two p-n-junctions, which act like two back-to-back diodes. This means that, by default, this device is OFF and we have to apply an electrical potential to its gate terminal to turn it ON.

The MOSFET illustrated in Figure B.49 is an *n*-channel device (we could, of course, create its *p*-channel counterpart by swapping the *p*-type and *n*-type

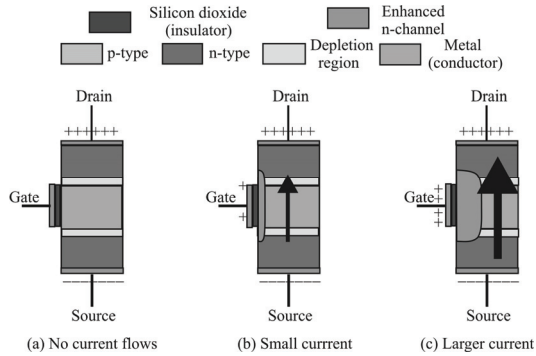


Figure B.49(a-c) Enhancement-mode N-channel MOSFET principles of operation.

diffusion regions). In order to turn the *n*-channel device *ON*, a positive voltage is applied to the gate. This positive voltage repels positive holes and attracts negative electrons in the *p*-type material. The electrons accumulate beneath the oxide layer, where they form a negative channel, hence the term "*n*-channel."

PIN Versus FET Switches

PIN diodes and FETs have relative advantages and disadvantages to each other for use in switching applications. The performance attributes that should be compared when the selection of one of these technologies is undertaken are more numerous than might be thought at first glance.

Integratability The wafer processing required for modern FET structures such as pHEMTs is largely lateral with respect to the top surface of the wafer; it lends itself quite well to the inclusion of passive component structures—such as metal-insulator-metal (MIM) capacitors, spiral inductors, and thin film resistors—all of which can be formed on or near the topmost layer of the wafer. The bottom-most layer of a FET wafer is semi-insulating material, which inherently isolates one FET structure from another. In comparison, PIN diode wafer processing is vertical with respect to the top surface of the wafer; the majority of the physical thickness of a PIN diode wafer is the cathode of the diode. It is very difficult to isolate the cathodes of PIN diodes that are on the same die. Consequently, with the exception of common-cathode pairs of PIN diodes, they are inherently discrete devices.

Power Handling The vertical structure of a PIN diode is a relative advantage for power handling. The heat that is generated by Joule heating within the I-layer of the diode can easily be conducted downwards through the diode's cathode layer to the system heat sink. RF FET structures such as pHEMTs and MESFETs are typically fabricated from III-V materials, such as GaAs, that have lower thermal conductivity than Si, which is the material utilized for most switching PIN diodes.

Switching Time The FET is a majority carrier device whose drain-source impedance is controlled by the thickness of a depletion layer that extends into the channel from the gate/source interface. The thickness of this depletion layer can be modulated very rapidly in response to a change in the gate-source control voltage. By contrast, a PIN diode stores minority carriers in its I-layer when it is under forward bias. These charge carriers must be primarily conducted out of the I-layer to change its impedance from low to high. This process is inherently slower than the change in drain-source impedance for a pHEMT or MESFET device.

The FET is a voltage-controlled device. In a practical pHEMT or MESFET, the only current that flows into the control port of the transistor is the reverse leakage current of the gate-source junction, which is very small, typically less than 10 μA. On the other hand, a PIN diode can require a significant injection of charge carriers into its I-layer to lower its impedance to the required level. The typical bias current for a PIN diode in a switch is 10 to 100 mA.

Distortion Performance The PIN diodes produce very linear series resistance when the amount of charge in its I-layer—as a result of DC forward bias current—is at least 10 times that of the charge that is alternately injected and removed by the RF signal. When the diode is nonconducting, it presents a very high resistance in parallel with its junction capacitance. This junction capacitance is independent of applied voltage for signals with sufficiently high frequency (typically greater than 100 MHz). Consequently, the PIN diode produces excellent distortion performance.

An FET structure responds very quickly to the magnitude of its applied gate-source voltage, because the gate-source junction in a MESFET and a pHEMT structure is a Schottky diode. The Schottky diode's very nonlinear impedance with respect to applied bias conditions can be a comparatively efficient distortion generation mechanism, which results in input third-order intercept for a FET switch that is roughly an order of magnitude or more lower than that of a PIN diode switch.

B.4.4 Summary

A comparison of typical performance parameters of solid-state RF switches compared with the performance of RF MEMS switches, as well as RF relay switches, is shown in Table B.3. Clearly RF MEMS have better performance for all parameters included except voltage required, power-handling capability, and switching time. The area that RF relays stand out, as expected, is in their power-handling capacity. Their biggest shortfalls are their power requirements and switching time. For example, with a switching time of 20 ms, tactical frequency hopping targets hopping at 100 hps ($T_{hop} \sim 10$ ms) could not be handled.

Table B.3 Performance Comparison of Typical RF Relays, FETs, PIN Diodes, and RF MEMS

Parameter	RF Relay	RF MEMS	PIN	FET
Voltage (V)	14	20–80	±3–5	3–5
Current (mA)	350	0	3–100	0
Power Consumption[a]	4.9 W	0.05–0.1 mW	5–500 mW	0.05–0.1 mW
Switching Time	20 ms	1–300 μs	1–100 ns	1–100 ns
C_{up} (series) (fF)	N/A	1–6	40–80	70–140
R_S (series) (Ω)	N/A	0.5–2	2-4	4–6
Capacitance Ratio[b]	N/A	40–500	10	n/a
Cutoff Frequency (THz)	N/A	20–80	1–4	0.5–2
Isolation	80 dB (100 MHz) 40 dB (2.5 GHz)	Very high (1–40 GHz) High (40–100 GHz)	High (1–10 GHz) Medium (10–100 GHz)	Medium (1–10 GHz) Low (10–40 GHz) N/A (>40 GHz)
Insertion Loss (1–100 GHz) (dB)	0.01 (100 MHz) 0.03 (2.5 GHz)	0.05–2 (1–100 GHz)	0.3–1.2 (1–100 GHz)	0.4–2.5 (1–100 GHz)
Power Handling (W)	2,000 (100 MHz) 400 (2 GHz)	<1	>100	<10
3rd Intercept Point (dBm)	N/A	66–80	+27–45	+27–45

[a] Includes voltage upconverter or drive circuitry.

[b] Capacitive switch only. A ratio of 500 is achieved with high ε_r dielectrics.

Source:].

Solid-state RF switches may be comprised of PIN diodes or FET structures such as MESFETs and pHEMTs. These two approaches to the switch makeup offer advantages and disadvantages. A PIN diode switch can typically handle higher power and produce less distortion, at the expense of longer switching time and much larger control current requirements. The very low bias current needs of MESFET and pHEMT switches make them very well suited for battery powered applications. These devices can also be integrated into complex, multithrow integrated circuit switches.

B.5 Concluding Remarks

We have provided an introduction to the various types of switches used for switching RF signals in this appendix. Such switches could be RF relays that can handle large power requirements, MEMS switches that are slow but consume little power in their operation, or one of two types of solid-state switches: PIN diodes or FETs. The former of these can switch reasonable power levels (for EW applications) while requiring significant drive current for switch control while the

latter can only handle small signals. Both of the solid-state switches change state considerably faster than their mechanical counterparts.

References

[1] Rebeiz, G. M., *RF MEMS Theory, Design, and Technology*, New York: Wiley, 2003.

[2] Peroulis, D., S. P. Pacheco, K. Sarabandi, and L. P. B. Katehi, "Electromechanical Considerations in Developing Low-Voltage RF MEMS Switches," *IEEE Transactions on Microwave Theory and Technology,* Vol. 51, No. 1, January 2003, pp. 259–270.

[3] Shen, S. C., D. T. Becher, D. C. Caruth, and M. Feng, "Development of Broadband Low-Voltage RF MEM Switches," GaAs MANTECH, Inc. Report, 2001.

[4] De Los Santos, J. J., *Introduction to Microelectromechanical* (*MEM*) *Microwave Systems*, Norwood, MA: Artech House, 1999.

[5] Rebeiz, G. M., J. B. Muldavin, "RF MEMS Switches and Switch Circuits," *IEEE Microwave Magazine,* Vol. 2, No. 4, December 2001, pp. 59–71.

[6] Rebeiz, G. M., *RF MEMS Theory, Design, and Technology*, New York: Wiley, 2003, p. 90.

[7] Caverly, R. H., and G. Hiller, "Establishing the Minimum Reverse Bias for a PIN Diode in a High Power Switch," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 38, No. 12, December 1990, pp. 1938–1943.

[8] Cory, R., *RF/Microwave Solid State Switches*, Skyworks Solutions Inc, www.MPDigest.com., http://www.skyworksinc.com/downloads/press_room/published_articles/MPD _052009.pdf.

[9] Caverly, R. H., and G. Hiller, "The Temperature Dependence of PIN Diode Attenuators," *1993 IEEE MTT-S Digest*, pp. 553–556.

[10] White, J., *Semiconductor Control*, Dedham, MA: Artech House, 1977, p. 64.

[11] Arora, N., J. Hauser, and D. Roulston, "Electron and Hole Mobilities in Silicon as a Function of Concentration and Temperature," *IEEE Transactions on Electron Devices*, Vol. ED-29, 1982, p. 292.

[12] Gamal, S., H. Morel, and J. P. Chante, "Carrier Lifetime Measurement by Ramp Recovery of PIN Diodes," *IEEE Transactions on Electron Devices*, Vol. 37, No. 8, August 1990, pp. 1921–1924.

[13] Dudeck, I., and R. Kassing, "Determination of the Temperature Dependence of the Capture Cross- Sections of the Gold Acceptor Level and of the Temperature of Current Filaments in Silicon PIN Diodes," *Physical Statistics Sol.(a),* Vol. 49, 1978, pp. 153–161.

[14] Bassett, R., "Observations on a Method of Determining the Carrier Lifetime in PIN Diodes," *Solid-State Electronics,* Vol. 12, 1969, pp. 385–391.

[15] Wilson, P., "Recombination in Silicon PIN Diodes," *Solid-State Electronics*, Vol. 10, No. 67, pp. 145–154.

[16] Bade, P., et al., "Solid State Antenna Switching," *RCA Review,* Vol. 42, December 1981, p. 752.

[17] Caulton, M., A. Rosen, P. Stabile, and A. Gombar, "PIN Diodes for Low Frequency High-Power Switching Applications," *IEEE Transactions on Microwave Theory and Techniques*, Vol. MTT-30, June 1982, p. 875.

[18] Leenov, D., "The Silicon PIN Diode as a Microwave Radar Protector at Megawatt Levels," *IEEE Tranactions on Electron Devices,* Vol. ED-11, February 1964, pp. 53–61.

[19] Caverly, R., and G. Hiller, "Distortion in PIN Diode Control Circuits," *IEEE Transactions on Microwave Theory and Techniques,* Vol. MTT-35, No. 5, May 1987, pp. 492–501.

[20] Hiller, G., and R. Caverly, "Predict PIN-Diode Switch Distortion," *Microwaves and RF,* Vol. 25, January 1986, p. 111.

[21] Caverly, R., "A Nonlinear PIN Diode Model for Use in Multi-Diode Microwave and RF Communication Circuit Simulation," *Proceedings 1988 IEEE International Symposium on Circuits and Systems,* June 1988, pp. 2295–2299.

[22] Nordman, J., and R. Greiner, "The Small-Signal Inductive Effect in a Long PIN Diode," *IEEE Transactions on Electronic Devices,* Vol. ED-10, 1963, p. 171.

[23] Prima, N., and Y. Tkhorik, "Total Impedance of a PIN Diode at High Injection Levels," *Sou. Physics-Semiconductor,* Vol. 1, No. 4, October 1967, p. 444.

[24] Varshney, R., D. Roulston, and S. Chamberlain, "Some Properties Concerning the As Impedance of PIN and PNN+ Diodes," *Solid-State Electronics,* Vol. 17, 1974, p. 699.

[25] Caverly, R., and G. Hiller, "Microwave Resistance of Gallium Arsenide and Silicon PIN Diodes," 1987 *IEEE MTT-S International Microwave Symposium Digest,* 1987, p. 591.

[26] Barratt, C., et al. "New GaAs PIN Diodes with Lower Dissipation Loss, Faster Switching Speed at Lower Drive Power," *IEEE MTT-S International Microwave Symposium Digest,* 1993, p. 507.

[27] Caverly, R. H., and G. Hiller, "The Frequency Dependence of PIN Diodes," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 37, No. 4, April 1989, pp. 787–790.

[28] White, J., *Microwave Semiconductor Engineering*, New York: Van Nostrand Reinhold, 1981.

[29] Hiller, G., and R. Caverly, "Predict PIN-Diode Switch Distortion," *Microwaves and RF*, Vol. 25, No. 1, January 1986, p. 111.

[30] http://www.techbites.com/20091006652/myblog/articles/z000c-how-it-works-jfets-mesfets-and-mosfets-part-2.html.

Appendix C

The Method of Moments

C.1 Introduction

Since the *method of moments* (MoM) technique is cited throughout the chapters, we include a brief introduction to the approach in this appendix.

Harrington introduced the MoM technique in his seminal 1967 paper, "Matrix Methods for Field Problems" [1]. It is perhaps the premier modern method for wire antenna analysis and design, as well as those antenna structures that can be approximated by wire models. This appendix reviews the MoM from a practical perspective, rather than an in-depth mathematical derivation. The discussion begins with the formulation of Pocklington's integral equation. The solution to Pocklington's equation, using the MoM, is then explained. The integral equation solution yields the current distribution on the wires which, in turn, are used to calculate the antenna's radiation characteristics and feed point impedance [2].

The task of determining the current distributions on wires in a wire antenna resulting from an arbitrary excitation may be expressed in terms of an integral equation problem. First, an integral expression is developed which defines the electric field resulting from an arbitrary current distribution on the wires. This integral expression employs a Green's function that relates the electric field at an arbitrary observation point in space to the current at an arbitrary source point. Known electric field boundary conditions are then applied to an unknown current distribution on the wire.

The MoM applies orthogonal expansions to translate the integral equation into a system of simultaneous linear equations. Basis functions are used to expand the current distribution and testing functions are used to force the known electric field boundary conditions. Matrix methods are then used to solve for the coefficients associated with the basis functions, thereby yielding the current distribution from the expansion coefficients. The antenna's radiation characteristics and feed point impedance are then derived from the calculated current distribution.

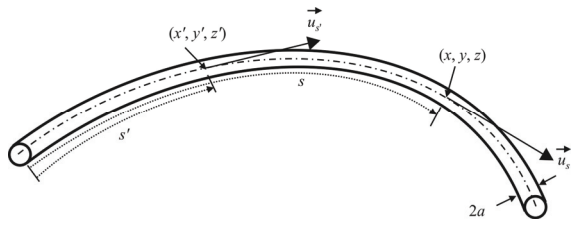


Figure C.1 Arbitrary curved wire.

C.2 Pocklington's Integral Equation

Provided that the radius of a thin wire is much less than both the wavelength and the length of the wire, in free space the equation governing the total axial current $I(s)\vec{s}$ on that thin curved wire is given by the *electric field integral equation*[1] (see Figure C.1)

$$j\frac{\eta_0}{4\pi\beta}\left\{\beta^2\int_C I(s')\vec{u}_{s'}\bullet\vec{u}_s K(s,s')ds'+\frac{d}{ds}\int_C\frac{d}{ds'}I(s')K(s,s')ds'\right\}=\vec{E}^i(s)\bullet\vec{u}_s,\ s\in C \quad \text{(C.1)}$$

where C is the wire axis contour, s denotes the arc displacement along C from a reference to a point on the wire axis, and $\vec{u}_s$ is the unit vector tangent to C at this point. The positive sense of this vector is in the direction of increasing s. $K(s,s')$ is the kernel or integral of Green's function given by [2]

$$K(s,s')=\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{e^{-j\beta r}}{r}d\phi' \quad \text{(C.2)}$$

where r is the distance between the source and observation points on the wire surface and $\vec{E}^i(s)$ is the incident electric field that illuminates the wire, evaluated in (C.1) on the wire surface at arc displacement s.

Pocklington's integral equation is a formulation for simple wire antennas based on (C.1) [3]. Figure C.2 shows a dipole from which Pocklington's equation can be derived. The dipole wire antenna is positioned along the z-axis in a cartesian coordinate system. The current is restricted to the centerline of the wire

[1] The electric field integral equation flows directly from the solution to (2.85) for $\vec{A}$ and (2.81) (in the far field so that $\nabla\phi = 0$).

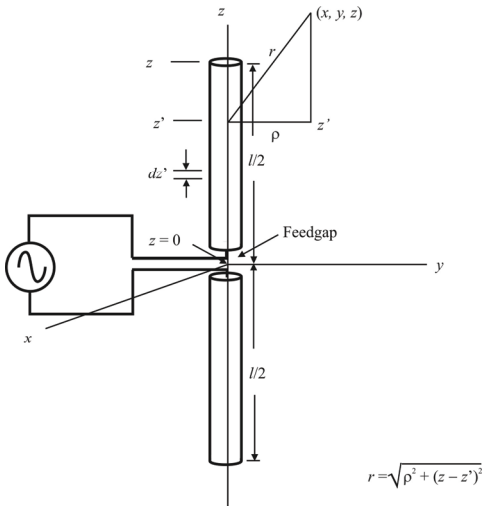


Figure C.2 Pocklington's integral equation formulation with a dipole.

and directed vertically along the z axis. Elemental current segments are located at coordinate $(0, 0, z')$. Field observation points are located at coordinates (x, y, z) and a feedgap of width Δz (small) is positioned at $z = 0$. The electric field along the surface of the wire and in the feedgap, which establishes the boundary conditions for the problem, is defined as follows: $E_z = 0$ on the surface of the wire and $E_z = V_g/\Delta z$ at the feedgap. V_g, the antenna excitation, is normally set to 2 V for input impedance calculations. Δz is commonly set equal to the diameter of the wire.

Pocklington's equation is given by

$$\int_{-l/2}^{l/2} I(z') \left[\frac{\partial^2}{\partial z^2} + \beta^2 \right] \frac{e^{-jkr}}{4\pi r} dz' = j\omega\varepsilon E_z(z) \tag{C.3}$$

where

$$r = \sqrt{\rho^2 + (z - z')^2} \tag{C.4}$$

is the distance between the current source and field observation point. The variable ρ specifies the radial distance from the wire. The current distribution $I_z(z')$ is defined along the length of the wire from $z' = l/2$ to $z' = -l/2$. The kernel $\partial^2/\partial z^2 + \beta^2$ denotes the wave equation differential operator on the free-space Green's

function $e^{-j\beta r}/4\pi r$. The constant β is the free space wave number ($\beta = 2\pi/\lambda$). $E_z(z)$ represents the electric field generated by the current on the wire.

With a specific excitation applied, as modeled through the appropriate boundary conditions, radiation characteristics and feedpoint impedances are determined from knowledge of the antenna's current distribution $I_z(z')$.

C.3 The Fundamental Concept

The fundamental concept behind the MoM employs orthogonal expansions and linear algebra to reduce the integral equation problem (C.3) to a system of simultaneous linear equations. The unknown current distribution $I_z(z')$ is expressed in terms of an orthogonal set of basis functions [4] and invoking the boundary conditions—the values of the electric field on the surface of the wire and in the feedgap—are invoked. The solution follows through the use of an inner product formulation. This inner product operation employs an orthogonal set of "testing" functions to enforce the boundary conditions, in an average sense, along the surface of the wire and in the feedgap. Moving the current's expansion coefficients to the outside of the integrodifferential operator permits the evaluation of known functions, yielding values that are loosely defined as impedances. The current's expansion coefficients, the orthogonal projections of the electric field boundary conditions and these so-called impedances are gathered into a system of simultaneous linear equations. This system of equations is solved to yield the current's expansion coefficients. The original current distribution is then determined by introducing these coefficients back into the basis function expansion.

The solution procedure begins by defining the unknown current distribution $I_z(z')$ in terms of an orthogonal set of basis functions. There are two categories of basis functions: subdomain basis functions and entire domain basis functions. Subdomain basis functions are the more popular of the two. They subdivide the wire into small segments and model the current distribution on each segment by a simple geometrical construct, such as a rectangle, triangle, or sinusoidal arc. The amplitudes of these constructs represent the expansion function coefficients. These simple constructs, illustrated in Figure C.3, often overlap to maintain continuity of the current distribution along the wire.

Entire domain basis functions employ a more formal orthogonal expansion, such as a Fourier series [5], to represent the current distribution along the entire wire. Entire domain basis functions tend to yield more complicated calculations for the so-called impedances with little gain in insight and, therefore, are not used as much. The introduction of the redefined current distribution reduces the integral equation to the form

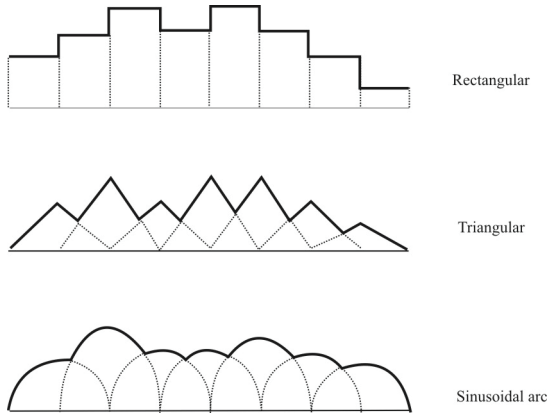


Figure C.3 Typical basis functions.

$$\sum_{n=1}^{N} C_n G_n(z) = E_z(z) \tag{C.5}$$

where

$$G_n(z) = \frac{1}{j4\pi\omega\varepsilon} \int_{-l/2}^{l/2} F_n(z') \left[\frac{\partial^2}{\partial z^2} + \beta^2 \right] \frac{e^{-j\beta r}}{r} dz' \tag{C.6}$$

C_n = current's expansion coefficient.
$F_n(z')$ = basis function.

The boundary conditions are now enforced through the use of an inner product operator with a set of orthogonal testing functions, denoted by $H_i(z)$. Each testing function is applied to both sides of the integral equation (C.5), and the inner product then enforces the boundary condition at the location described by the testing function. This operation may be thought of as simply enforcing the boundary condition at a single point on the wire. After each testing function operation, the integral equation will appear as:

$$\sum_{n=1}^{N} C_n < H_m(z), G_n(z) > = < H_m(z), E_z(z) > \tag{C.7}$$

where < > represents the inner product operator, in this case given by

$$< H_m(z), G_n(z) >= \int_{-l/2}^{l/2} H_m(z) G_n(z) dz \tag{C.8}$$

where $H_m(z)$ is a testing function which has a nonzero value for only a small segment of wire located at $z = z_m$.

There are two common approaches to formulating the orthogonal set of testing functions. The first approach, the point matching or colocation technique, defines the testing function in terms of Dirac delta functions:

$$H_m(z) = \delta(z - z_m) \tag{C.9}$$

where z_m are specific points on the wire at which the boundary conditions are enforced. The z_m are usually selected to correspond with the midpoint of each basis function.

The entire boundary condition is enforced by applying the complete set of testing functions. This operation yields a set of integral equations.

$$\mathbf{Z}\vec{I} = [Z_{mn}][I_n] = [V_m] = \vec{V} \tag{C.10}$$

where

$$Z_{mn} = \int_{-l/2}^{l/2} H_m(z) G_n(z) dz \tag{C.11}$$

$$I_n = C_n \tag{C.12}$$

$$V_m = \int_{-l/2}^{l/2} H_m(z) E_z(z) dz \tag{C.13}$$

This set of simultaneous linear equations will yield the value of C_n using

$$[I_n] = [Z_{mn}]^{-1}[V_m] \tag{C.14}$$

and (C.12).

The NEC, which is widely used to solve problems that can be defined as sets of one or more "wires" (linear elements), is perhaps the most prolific implementation of the MoM. This program is available in many versions, some of which execute on desktop PCs. 4NEC2 is one of these that is freeware and very user-friendly [6].

C.4 Concluding Remarks

The MoM is a popular solution technique for analysis and design of wire antennas. This appendix has presented an overview of this technique. The essential elements of integral equation formulation, basis and testing function definition, and reduction to a set of simultaneous linear equations, have been reviewed. For more in-depth coverage references [1, 2, 8, 9] should be consulted.

References

[1] Harrington, R. F., "Matrix Methods for Field Problems," *Proceedings of the IEEE,* Vol. 55, February 1967, pp. 136–149.

[2] Stutzman, W. L., and G. C. Thield, *Antenna Theory and Design*, New York: Wiley, 1981, Ch. 7.

[3] Pocklington, C. H. C., "Electrical Oscillations in Wires," *Cambridge Philosophical Society Proceedings*, Vol. 9, 1897, p. 324.

[4] Indritz, J., *Methods in Analysis*, New York: Macmillan, 1963, Ch. 1.

[5] Thiele, G. C., "Analysis of Yagi-Uda-Type Antennas," *IEEE Transactions on Antennas and Propagation,* Vol. AP-17, No. 1, January 1969, pp. 24–30.

[6] http://home.ict.nl/~arivoors/.

[7] Harrington, R. F., *Field Computation by Moment Methods*, New York: Wiley/IEEE Press, 1990.

[8] Balanis, C., *Antenna Theory: Analysis and Design*, 3rd ed, New York: Wiley, 2005.

[9] Kraus, J. D., *Antennas*, New York: McGraw Hill, 1988.

Appendix D

Properties of Dielectric Materials

D.1 Introduction

A dielectric medium is any nonconducting material, or insulator. In a dielectric material, the dielectric constant is different from that in a vacuum. We used the term material here rather loosely. A dielectric medium need not be a solid material. A gas, for example air, can be, and in this example is, a dielectric medium.

Some of this material is patterned after [1].

Dielectric materials are of interest when considering EW antennas through several paths. The first is a material used in the construction of cables (the material in that case may be air). The dielectric properties of the insulating material to a large extent determine the characteristics of the cables. The second is in the construction of radomes, which are essentially covers for antennas that primarily protect the antennas from the elements. In EW applications, radomes are also sometimes used to keep what is inside hidden from plain view.

Yet another use of dielectric materials in electronic systems, including EW systems, is in the makeup of capacitors, serving as the material between the plates. That is not our interest in dielectrics here. We discussed in Chapter 14 some techniques for antenna miniaturization using dielectric materials, and that is where our interest lies. As such, this appendix is tailored to this application and other applications are not discussed at length. The dielectric properties affecting radomes were discussed in Chapter 22.

D.2 Properties of the Dielectric Constant

The complex *dielectric constant*, or *permittivity*, of a region, be it a material or otherwise, is represented by the symbol ε where $\varepsilon = \varepsilon' - j\varepsilon''$ (yes, the minus sign is correct). The dimensionless *relative* complex permittivity ε_r is given by

$\varepsilon_r = \varepsilon_r' - j\varepsilon_r''$ while $\varepsilon = \varepsilon_0 \varepsilon_r$, ε_0 being the permittivity of a vacuum, a fixed real constant given approximately by $\varepsilon_0 = 8.85 \times 10^{-12}\ \mathrm{F\,m^{-1}}$. In general, ε depends on temperature and pressure. It is also frequency-dependent, although ε' and ε'' cannot vary independently with frequency, since their frequency variations are connected through the Kramers-Kronig relationship[1] [2]: a drop in ε' with increasing frequency is necessarily associated with a peak in ε''. Except for exceedingly high applied fields, ε is independent of the magnitude of the applied electric field for all dielectric materials used in practice, excluding ferroelectric materials.[2]

The velocity v_p of propagation of an EM wave through a loss-free transmitting medium of refractive index n is given by

$$v_p = \frac{c}{n} \tag{D.1}$$

where c is the velocity in free space. The velocity is also given by

$$v_p = \frac{c}{\sqrt{\mu_r \epsilon_r}} \tag{D.2}$$

where μ_r is the relative permeability. Thus for loss-free non-agnetic materials, for which $\mu_r = 1$, then $\varepsilon_r = n^2$. However, in general losses do occur, and the material is characterized by a complex refractive index $\hat{n}$ given by

$$\hat{n} = n - jk \tag{D.3}$$

where k is the *absorption coefficient*. Then $\varepsilon_r = \hat{n}^2$, or

$$\varepsilon_r' - j\varepsilon_r'' = (n - jk)^2 \tag{D.4}$$

---

[1] The Kramers-Kronig relations relate the real and imaginary parts of any complex function that is analytic in the upper half-plane and causality in any real physical system causes the analyticity condition to be satisfied. Therefore, knowing one of these functions dictates the other in all real physical systems.

[2] A ferroelectric material has a nonlinear relationship modeled as a hysteresis loop between the applied electric field and the apparent stored charge. The semi-permanent electric dipoles in the ferroelectric material cause the dielectric constant of a ferroelectric to be much higher than that of a dielectric. When an external electric field is applied across a dielectric, the dipoles naturally align themselves with the field direction. After the charge is removed, the dipoles retain their polarization state.

from which it follows that $\varepsilon_r' = n^2 - k^2$ and $\varepsilon_r'' = 2nk$. Nevertheless, when the loss is small, so that $k << n$, then $\varepsilon_r' \approx n^2$.

D.3 The Refractive Index

Consider the propagation of an EM wave through a uniform dielectric medium of real dielectric constant ε'. The dipole moment[3] per unit volume induced in the medium by the wave electric field $\vec{E}$ is given by

$$\vec{P} = \varepsilon_0'(\varepsilon' - 1)\vec{E} \tag{D.5}$$

It can be shown that Maxwell's equations for the propagation of an EM wave through a dielectric medium are the same as Maxwell's equations for the propagation of waves through a vacuum, except that the speed of propagation, *c*, is replaced by c / n, where

$$n = \sqrt{\varepsilon'} \tag{D.6}$$

n is called the *refractive index* of the medium. Because of (D.6), EM waves propagate slower through a dielectric medium than through a vacuum by the factor *n* (assuming, of course, that $n > 1$).

D.4 Dielectric Constant of a Gaseous Medium

In Section 2.8, we discussed without quantification the effects of a charged particle interacting with a magnetic wave *B*. It can be shown that the dipole moment $\vec{P}$ of an atom induced by electric field $\vec{E}$ is given by

$$\vec{P} = \frac{e^2}{m_e(\omega_0^2 - \omega^2)}\vec{E} \tag{D.7}$$

where ω_0 is the natural frequency of the atom (i.e., the frequency of one of the atom's spectral lines), ω the frequency of the incident radiation, *e* is the electron charge, and m_e is the mass of the electron. Suppose that there are *N* atoms per unit

[3] A *dipole moment* is the propensity of a system of charges to act like a dipole; in other words, the degree to which it is polarized.

volume. It follows that the induced dipole moment per unit volume of the collection of atoms takes the form

$$\vec{P} = \frac{Ne^2}{m_e(\omega_0^2 - \omega^2)}\vec{E} \tag{D.8}$$

Comparing (D.8) with (D.5) yields

$$\varepsilon' = 1 + \frac{Ne^2}{\varepsilon_0' m_e(\omega_0^2 - \omega^2)} \tag{D.9}$$

for the expression for the dielectric constant of the collection of atoms. It is, of course, necessary to sum over all species of gas in the medium, and all important spectral lines.

Equation (D.9) indicates that, in general, the dielectric "constant" of a gaseous medium is a function of the EM wave frequency, ω. Since the effective wave speed through the medium is governed by $c/\sqrt{\varepsilon'}$, it follows that waves of different frequencies traveling through a gaseous medium do so at different speeds. This phenomenon is called *dispersion*. The velocity considered here is known as the *phase velocity* since it is the speed that the phase in the wave travels.

D.5 Dielectric Constant of a Plasma

A plasma has similar characteristics to a gaseous medium, except that the electrons are *free*: that is, there is no restoring force due to nearby atomic nuclei.[4] Setting ω_0 to zero in (D.9) and setting N equal to the number density of electrons, N_e yields the permittivity of a plasma. Thus,

$$\varepsilon' = 1 - \frac{\omega_p^2}{\omega^2} \tag{D.10}$$

where the characteristic frequency

[4] An example of a plasma is the outer edges of Earth's atmosphere where, during the day, the sun strips electrons away from otherwise neutral atoms, creating the ionosphere. The particle density is so low that the nuclear force to recombine the electrons with the positively charged ions does not substantially occur until nighttime, when the sun's energy is absent.

$$\omega_{\rm p} = \sqrt{\frac{N_{\rm e} e^2}{\varepsilon_0' m_{\rm e}}} \tag{D.11}$$

is called the *plasma frequency*. We can see that (D.10) is problematic for a plasma. For frequencies above the plasma frequency, the dielectric constant of a plasma is less than 1. Hence, the refractive index $n = \sqrt{\varepsilon'}$ is also less than unity. This would seem to imply that high-frequency EM waves can propagate through a plasma with a velocity c/n, which is *greater* than the velocity of light in a vacuum. On the other hand, for frequencies below the plasma frequency, the dielectric constant is *negative*, which would seem to imply that the refractive index $n = \sqrt{\varepsilon'}$ is *imaginary*.

Consider an infinite plane wave of frequency, ω, greater than the plasma frequency, propagating through a plasma in the z-direction. The wave electric field can be written as

$$\vec{E} = E_0 e^{j(\beta z - \omega t)} \vec{u}_z \tag{D.12}$$

where the real part of (D.12) corresponds to the physical field. A point of the wave, such as a peak or trough, travels at the so-called *phase velocity*, which is given by

$$v_{\rm p} = \frac{\omega}{\beta} \tag{D.13}$$

From (D.1) we know that the phase velocity of EM waves in a dielectric medium is also given by $v_{\rm p} = c/n = c/\sqrt{\varepsilon'}$, so

$$\omega^2 = \frac{\beta^2 c^2}{\varepsilon'} \tag{D.14}$$

It follows from (D.10) that

$$\omega^2(\beta) = \beta^2 c^2 + \omega_{\rm p}^2 \tag{D.15}$$

in a plasma. The above type of expression, which determines the wave frequency, ω, as a function of the wave number, β, for the medium in question, is called a *dispersion relation* (since, inter alia, it determines how fast wave pulses disperse in the medium).

From (D.15) we get

$$\beta = \frac{1}{c}\sqrt{\omega^2 - \omega_p^2} \tag{D.16}$$

and (D.13) and (D.16) produce

$$\begin{aligned} v_p &= \frac{\omega}{\beta} \\ &= \frac{\omega c}{\sqrt{\omega^2 - \omega_p^2}} \\ &= \frac{c}{\sqrt{1 - \dfrac{\omega_p^2}{\omega^2}}} \end{aligned} \tag{D.17}$$

which, whenever $\omega_P > 0$, we see is greater than the speed of light. However, the theory of relativity does not forbid this. What the theory of relativity says is that information cannot travel at a velocity greater than c (nor can material objects, but that is not our concern here). The points on an infinite plane-wave, such as (D.12), do *not* carry any information.

D.6 Group Velocity

In this section we derive how fast wave packets actually move. Surprisingly, we find that wave packets typically move at a speed very different from the phase velocity, $v_p = \omega/\beta$, of the wave composing the wave packet. Suppose we have a wave that is the sum of two sine waves at frequencies ω_1 and ω_2, that are close but different, and with slightly different wavenumbers, β_1 and β_2, given by

$$s(t) = \sin(\beta_1 x - \omega_1 t) + \sin(\beta_2 x - \omega_2 t) \tag{D.18}$$

After some elementary algebraic and trigonometric manipulations, we find

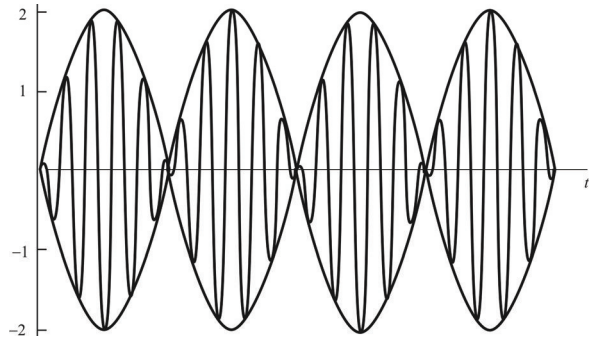


Figure D.1 Group velocity.

$$s(t) = 2\sin(\beta_0 x - \omega_0 t)\cos(\Delta\beta x - \Delta\omega t) \tag{D.19}$$

where we have

$$\beta_0 = (\beta_1 + \beta_2)/2 \tag{D.20}$$
$$\omega_0 = (\omega_1 + \omega_2)/2 \tag{D.21}$$
$$\Delta\beta = (\beta_2 - \beta_1)/2 \tag{D.22}$$
$$\Delta\omega = (\omega_2 - \omega_1)/2 \tag{D.23}$$

This can be thought of as a sine wave of frequency ω_0 and wavenumber β_0 modulated by a cosine function that is oscillating at the difference of the two original waves, presumably much slower (see Figure D.1). The modulation pattern, the second term in (D.19), moves with a speed sufficient to keep the argument of the cosine function constant:

$$\Delta\beta x - \Delta\omega t = \text{constant} \tag{D.24}$$

or

$$\Delta\beta x = \Delta\omega t + \text{constant} \tag{D.25}$$

so that

$$\frac{\partial}{\partial t}\Delta\beta x = \frac{\partial}{\partial t}\Delta\omega t \tag{D.26}$$

leading to

$$v_g = \frac{\partial x}{\partial t} = \lim_{\Delta\beta \to 0} \frac{\Delta\omega}{\Delta\beta}\bigg|_{\beta=\beta_0} = \frac{d\omega}{d\beta}\bigg|_{\beta=\beta_0} \tag{D.27}$$

The derivative of $\omega(\beta)$ with respect to β is first computed and then evaluated at $\beta = \beta_0$. Equation (D.27) is known as the *group velocity* and is different from the phase velocity in a dispersive medium. Thus, information (i.e., an individual wave pulse) travels through a dispersive medium at the group velocity, rather than the phase velocity and relativity requires that the group velocity, rather than the phase velocity, must always be less than c.

Differentiation of the dispersion relation (D.15) yields

$$2\omega\frac{d\omega}{d\beta} = 2\beta c^2 \tag{D.28}$$

since ω_p is independent of β. Hence

$$\begin{aligned} v_g &= \frac{d\omega}{d\beta} = \frac{\beta}{\omega}c^2 \\ &= \frac{c^2}{v_p} \end{aligned} \tag{D.29}$$

Incorporating (D.17), we get

$$\begin{aligned} v_g &= \frac{c^2}{\dfrac{c}{\sqrt{1-\dfrac{\omega_p^2}{\omega^2}}}} \\ &= c\sqrt{1-\frac{\omega_p^2}{\omega^2}} \end{aligned} \tag{D.30}$$

which is the group velocity for high-frequency waves propagating through a plasma, which is less than c. We thus conclude that the dispersion relation (D.15) is consistent with relativity.

Let us now consider the propagation of low-frequency EM waves through a plasma. From (D.17) and (D.30) we can see that when the wave frequency, ω, falls

below the plasma frequency, ω_p, both the phase and group velocities become imaginary. Suppose that a plane wave of frequency $\omega < \omega_p$ is propagating in the medium. According to the dispersion relation (D.15), the associated wave number is given by

$$\beta = j\frac{\sqrt{\omega_p^2 - \omega^2}}{c} = j|\beta| \tag{D.31}$$

Hence, the wave electric field takes the form

$$E_z = E_0 e^{j(j|\beta|z-\omega t)} = E_0 e^{-|\beta|z} e^{-j\omega t} \tag{D.32}$$

and the wave attenuates as it propagates. We see that for $\omega < \omega_p$, EM waves in a plasma take the form of decaying *standing waves*, rather than traveling waves. We conclude that an EM wave of frequency less than the plasma frequency, which is incident on a plasma will not propagate through the plasma, but will be *totally reflected*.

The outermost layer of the Earth's atmosphere consists of a partially ionized zone known as the *ionosphere*. The plasma frequency in the ionosphere is about 1 MHz, which lies at the upper end of the medium-wave band of radio frequencies. It follows that low-frequency radio signals (i.e., all signals in the long-wave band, and most in the medium wave band) are reflected off the ionosphere. For this reason, such signals can be detected over the horizon. Indeed, long-wave radio signals reflect multiple times off the ionosphere, with very little loss (they also reflect multiple times off the Earth, which is enough of a conductor to act as a mirror for radio waves), and can consequently be detected all over the world.

Contrast this with the HF range up to about 10 MHz. The ionosphere *refracts* such EM waves by bending them as they traverse through. Under some (many) circumstances these waves are bent over enough to propagate back to the Earth. They too can be mirrored by the Earth and *reflected* back into the ionosphere.

On the other hand, VHF radio signals (e.g., signals in the FM band) pass straight through the ionosphere. For this reason, such signals cannot be detected over the horizon, which accounts for the relatively local coverage of FM radio stations. Note, from (D.11), that the plasma frequency is proportional to the square root of the density of free electrons. Now the level of ionization in the ionosphere is maintained by ultraviolet light from the sun. Of course, there is no such light at night, and the density of free electrons in the ionosphere consequently drops as electrons and ions gradually recombine. It follows that the plasma frequency in the ionosphere also drops at night, giving rise to a marked deterioration in the reception of distant medium-wave radio stations.

D.7 Reflection at a Dielectric Boundary

An EM wave at frequency ω propagating in the direction defined by $\vec{r}$ can be written

$$\vec{E}(\vec{r},t) = \vec{E}_0 e^{j(\vec{\beta}\bullet\vec{r}-\omega t)} \tag{D.33}$$

$$\vec{B}(\vec{r},t) = \vec{B}_0 e^{j(\vec{\beta}\bullet\vec{r}-\omega t)} \tag{D.34}$$

The wave vector, $\vec{\beta}$, indicates the direction of propagation of the wave, and its phase velocity, v_p, is

$$v_p = \frac{\omega}{|\vec{\beta}|} \tag{D.35}$$

Since the wave is transverse, we must have $\vec{E}_0 \bullet \vec{\beta} = \vec{B}_0 \bullet \vec{\beta} = 0$. Finally, the familiar Maxwell equation (2.36), repeated here as

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{D.36}$$

leads us to the relation between the constant vectors $\vec{E}_0$ and $\vec{B}_0$:

$$\vec{B}_0 = \frac{1}{v_p}\vec{\beta} \times \vec{E}_0 \tag{D.37}$$

Suppose that the plane $z = 0$ forms the boundary between two different dielectric media. Let medium 1, of refractive index n_1, occupy the region $z < 0$, while medium 2, of refractive index n_2, occupies the region $z > 0$ (see Figure D.2). Let us investigate what happens when an electromagnetic wave is incident on this boundary from medium 1.

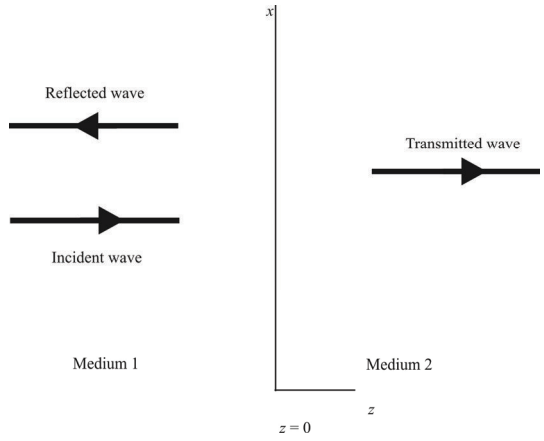


Figure D.2 Normal incidence.

D.7.1 Normal Incidence Angle

Let us first consider the case of incidence *normal* to the boundary as illustrated in Figure D.2. We have $\vec{u}_{\vec{\beta}} = +\vec{u}_z$ for the incident and transmitted waves, and $\vec{u}_{\vec{\beta}} = -\vec{u}_z$ for the reflected wave. We can assume that the incident wave is polarized in the *x*-direction. Hence, using(D.37), the incident wave can be written

$$\vec{E}_{\rm i}(z,t) = E_i e^{j(\beta_1 z - \omega t)} \vec{u}_{\rm x} \tag{D.38}$$

$$\vec{B}_{\rm i}(z,t) = \frac{E_i}{v_1} e^{j(\beta_1 z - \omega t)} \vec{u}_{\rm y} \tag{D.39}$$

where $v_1 = c/n_1$ is the phase velocity in medium 1, and $\beta_1 = \omega / v_1$. Likewise, the reflected wave takes the form

$$\vec{E}_{\rm r}(z,t) = E_r e^{j(-\beta_1 z - \omega t)} \vec{u}_{\rm x} \tag{D.40}$$

$$\vec{B}_{\rm r}(z,t) = -\frac{E_r}{v_1} e^{j(-\beta_1 z - \omega t)} \vec{u}_{\rm y} \tag{D.41}$$

Finally, the transmitted wave can be written as

$$\vec{E}_{\rm t}(z,t) = E_{\rm t} e^{j(\beta_2 z - \omega t)} \vec{u}_{\rm x} \tag{D.42}$$

$$\vec{B}_{\mathrm{t}}(z,t)=\frac{E_{\mathrm{t}}}{v_2}e^{j(\beta_2 z-\omega t)}\vec{u}_{\mathrm{y}} \tag{D.43}$$

where $v_2 = c/n_2$ is the phase velocity in medium 2, and $\beta_2 = \omega/v_2$.

For the case of normal incidence, the electric and magnetic components of all three waves are *parallel* to the boundary between the two dielectric media. Hence, the appropriate boundary conditions to apply at $z = 0$ are

$$E_{\|1}=E_{\|2} \tag{D.44}$$

$$B_{\|1}=B_{\|2} \tag{D.45}$$

The latter condition derives from the general boundary condition $H_{\|1}=H_{\|2}$, and the fact that $\vec{B}=\mu_0\vec{H}$ in both media (which are assumed to be nonmagnetic).

The boundary condition (D.44) yields

$$E_{\mathrm{i}}+E_{\mathrm{r}}=E_{\mathrm{t}} \tag{D.46}$$

while the boundary condition (D.45) gives

$$\frac{E_{\mathrm{i}}-E_{\mathrm{r}}}{v_1}=\frac{E_{\mathrm{t}}}{v_2} \tag{D.47}$$

or

$$E_{\mathrm{i}}-E_{\mathrm{r}}=\frac{v_1}{v_2}E_{\mathrm{t}}=\frac{n_2}{n_1}E_{\mathrm{t}} \tag{D.48}$$

since $v_1/v_2 = n_2/n_1$. Equations (D.46) and (D.48) yield

$$E_{\mathrm{r}}=\frac{n_1-n_2}{n_1+n_2}E_{\mathrm{i}} \tag{D.49}$$

$$E_{\mathrm{t}}=\frac{2n_1}{n_1+n_2}E_{\mathrm{i}} \tag{D.50}$$

Thus, we have the amplitudes of the reflected and transmitted waves in terms of the amplitude of the incident wave.

From (D.49) and (D.50), palpably, if $n_1 = n_2$ then $E_r = 0$ and $E_t = E_i$. That is, if the two media have the same indices of refraction, then there is no reflection at the boundary between them, and the transmitted wave is therefore equal in amplitude to the incident wave. On the other hand, if $n_1 \neq n_2$ then there is some reflection at the boundary. This has important practical consequences. We can only see a clean pane of glass in a window because some of the light incident at an air/glass boundary is reflected, due to the different refractive indices of air (1.00) and glass (1.55). It is significantly more difficult to see glass when it is submerged in water because the refractive indices of glass and water (1.33) are quite similar, and so there is very little reflection of light incident on a water/glass boundary.

From (D.49), we see that when $n_2 > n_1$, $E_r / E_i < 0$. The negative sign indicates a 180^o phase shift of the reflected wave, with respect to the incident wave.

The mean EM wave energy flux, or *intensity*, in the *z*-direction is

$$U = \frac{1}{\mu_0} < \vec{E} \times \vec{B} \bullet \vec{u}_z >= \frac{E_0 B_0}{2\mu_0} = \frac{E_0^2}{2\mu_0 v_p} \tag{D.51}$$

The *coefficient of reflection*, denoted *R*, is defined as the ratio of the intensities of the reflected and incident waves:

$$R = \frac{U_r}{U_i} = \left(\frac{E_r}{E_i} \right)^2 \tag{D.52}$$

The *coefficient of transmission*, denoted *T,* is the ratio of the intensities of the transmitted and incident waves:

$$T = \frac{U_t}{U_i} = \frac{v_1}{v_2} \left(\frac{E_t}{E_i} \right)^2 = \frac{n_2}{n_1} \left(\frac{E_t}{E_i} \right)^2 \tag{D.53}$$

Equations (D.49), (D.50), (D.52), and (D.53) yield

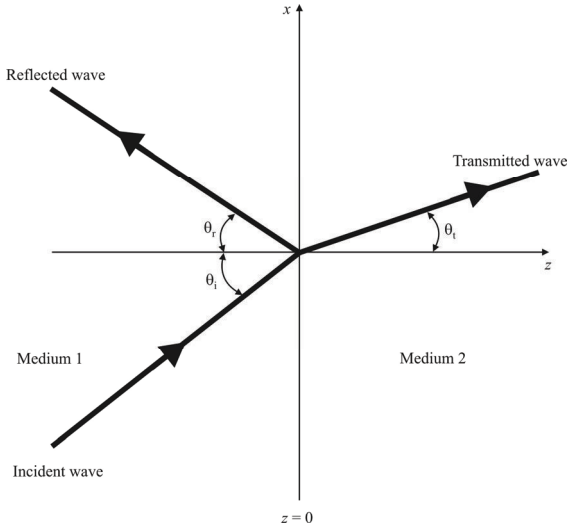


Figure D.3 Oblique incidence to the boundary.

$$R = \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2 \tag{D.54}$$

$$T = \frac{n_2}{n_1} \left(\frac{2n_1}{n_1 + n_2} \right)^2 \tag{D.55}$$

Note that $R + T = 1$. In other words, any wave energy that is not reflected at the boundary is transmitted and vice versa. Nothing is lost.

D.7.2 Oblique Incidence Angle

Let us now consider the case of incidence *oblique* to the boundary as illustrated in Figure D.3. Suppose that the incident wave subtends an angle θ_i with the normal to the boundary, whereas the reflected and transmitted waves subtend angles θ_r and θ_t, respectively.

The incident wave can be written as

$$\vec{E}_i(\vec{r}, t) = \vec{E}_i e^{j(\vec{\beta}_i \cdot \vec{r} - \omega t)} \tag{D.56}$$

$$\vec{B}_i(\vec{r}, t) = \vec{B}_i e^{j(\vec{\beta}_i \cdot \vec{r} - \omega t)} \tag{D.57}$$

with analogous expressions for the reflected and transmitted waves. Since, in the case of oblique incidence, the electric and magnetic components of the wave fields are no longer necessarily parallel to the boundary at $z = 0$, we must supplement the boundary conditions (D.44) and (D.45) to account for this. We add the boundary conditions

$$\varepsilon_1' E_{\perp 1} = \varepsilon_2' E_{\perp 2} \tag{D.58}$$

$$B_{\perp 1} = B_{\perp 2} \tag{D.59}$$

Equation (D.58) comes from the general boundary condition $D_{\perp 1}=D_{\perp 2}$.

It follows from (D.45) and (D.59) that both components of the magnetic field are continuous at the boundary. Therefore,

$$\vec{B}_i e^{j(\vec{\beta}_i \cdot \vec{r} - \omega t)} + \vec{B}_r e^{j(\vec{\beta}_r \cdot \vec{r} - \omega t)} = \vec{B}_t e^{j(\vec{\beta}_t \cdot \vec{r} - \omega t)} \tag{D.60}$$

at $z = 0$. Given that $\vec{B}_i, \vec{B}_r,$ and $\vec{B}_t$ are constant vectors, the only way in which the above equation can be satisfied for all values of x and y is if

$$\vec{\beta}_i \cdot \vec{r} = \vec{\beta}_r \cdot \vec{r} = \vec{\beta}_t \cdot \vec{r} \tag{D.61}$$

throughout the $z = 0$ plane. This, in turn, implies that

$$\beta_{ix} = \beta_{rx} = \beta_{tx} \tag{D.62}$$

and

$$\beta_{iy} = \beta_{ry} = \beta_{ty} \tag{D.63}$$

Thus, when $\beta_{iy} = 0$, then $\beta_{ry} = \beta_{ty} = 0$. This indicates that if the incident wave lies in the x-z plane, then the reflected and transmitted waves also lie in the x-z plane. Therefore, the incident, reflected, and transmitted waves all lie in the same plane, known as the *plane of incidence.*

We know that $\beta_i = \beta_r = \omega / v_1$ and $\beta_t = \omega / v_2$. Moreover,

$$\sin\theta_i = \frac{\beta_{xi}}{\beta_i} \tag{D.64}$$

with similar expressions for θ_r and θ_t. Hence, according to (D.62),

$$\sin\theta_r = \sin\theta_i \tag{D.65}$$

which implies that

$$\theta_r = \theta_i \tag{D.66}$$

(or $\theta_r = \pi - \theta_I$, which means that the reflected wave is actually transmitted; this insipid case is of no interest to us). Also,

$$\frac{\sin\theta_t}{\sin\theta_i} = \frac{v_2}{v_1} = \frac{n_1}{n_2} \tag{D.67}$$

Equations (D.66) is known as the *law of reflection* and (D.67) is known as *Snell's law of refraction* in optics.

We need to consider two independent wave polarizations separately for the case of oblique incidence. The first polarization has all the electric fields perpendicular to the plane of incidence, while the second has all the magnetic fields perpendicular to the plane of incidence.

First, consider the wave polarization. The unit vectors in the directions of propagation of the incident, reflected, and transmitted waves are given by

$$\vec{u}_i = (\sin\theta_i, 0, \cos\theta_i) \tag{D.68}$$

$$\vec{u}_r = (\sin\theta_i, 0, -\cos\theta_i) \tag{D.69}$$

$$\vec{u}_t = (\sin\theta_t, 0, \cos\theta_t) \tag{D.70}$$

The constant vectors associated with the incident wave are written

$$\vec{E}_i = E_i \vec{u}_y \tag{D.71}$$

and

$$\vec{B}_i = \frac{E_i}{v_1}(-\cos\theta_i, 0, \sin\theta_i) \tag{D.72}$$

where use has been made of (D.37). Similarly, the vectors associated with the reflected and transmitted waves are

$$\vec{E}_r = E_r \vec{u}_y \tag{D.73}$$

$$\vec{B}_r = \frac{E_r}{v_1}(\cos\theta_i, 0, \sin\theta_i) \tag{D.74}$$

and

$$\vec{E}_t = E_t \vec{u}_y \tag{D.75}$$

$$\vec{B}_t = \frac{E_t}{v_2}(-\cos\theta_t, 0, \sin\theta_t) \tag{D.76}$$

respectively.

The boundary condition (D.44) yields $E_{y1} = E_{y2,}$ or

$$E_i + E_r = E_t \tag{D.77}$$

The boundary condition (D.59) gives $B_{z1} = B_{z2}$, or

$$(E_i + E_r)\frac{\sin\theta_i}{v_1} = E_t \frac{\sin\theta_t}{v_2} \tag{D.78}$$

Using Snell's law, (D.67), the above expression reduces to (D.77). The boundary condition (D.45) yields $B_{x1} = B_{x2}$, or

$$(E_i - E_r)\frac{\cos\theta_i}{v_1} = E_t \frac{\cos\theta_t}{v_2} \tag{D.79}$$

For convenience let us define the parameters

$$\alpha = \frac{\cos\theta_t}{\cos\theta_i} \tag{D.80}$$

and

$$\nu = \frac{v_1}{v_2} \tag{D.81}$$

Equations (D.77) and (D.79) yield

$$E_r = \left(\frac{1-\alpha\nu}{1+\alpha\nu}\right)E_i \tag{D.82}$$

$$E_t = \left(\frac{2}{1+\alpha\upsilon}\right)E_i \tag{D.83}$$

These relations are known as the *Fresnel equations*.

The coefficient of reflection is

$$R = \left(\frac{E_r}{E_i}\right)^2 = \left(\frac{1-\alpha\nu}{1+\alpha\nu}\right)^2 \tag{D.84}$$

whereas the coefficient of transmission takes the form

$$T = \frac{\cos\theta_t}{\cos\theta_i}\frac{v_1}{v_2}\left(\frac{E_t}{E_i}\right)^2 = \alpha\upsilon\left(\frac{2}{1+\alpha\nu}\right)^2 \tag{D.85}$$

Note that again $R + T = 1$.

Now consider the second wave polarization. The vectors associated with the incident, reflected, and transmitted waves are written as

$$\vec{E}_i = E_i(\cos\theta_i, 0, -\sin\theta_i) \tag{D.86}$$

$$\vec{B}_i = \frac{E_i}{v_1}\vec{u}_y \tag{D.87}$$

and

$$\vec{E}_r = E_r(\cos\theta_i, 0, \sin\theta_i) \tag{D.88}$$

$$\vec{B}_r = -\frac{E_r}{v_1}\vec{u}_y \tag{D.89}$$

and

$$\vec{E}_t = E_t(\cos\theta_t, 0, -\sin\theta_t) \tag{D.90}$$

$$\vec{B}_t = \frac{E_t}{v_2}\vec{u}_y \tag{D.91}$$

respectively. The boundary condition (D.45) yields $B_{y1} = B_{y2}$, or

$$\frac{E_{\rm i} - E_{\rm r}}{v_1} = \frac{E_{\rm t}}{v_2} \tag{D.92}$$

Similarly, the boundary condition (D.44) gives $E_{\rm x1} = E_{\rm x2}$, or

$$(E_{\rm i} + E_{\rm r})\cos\theta_{\rm i} = E_{\rm t}\cos\theta_{\rm t} \tag{D.93}$$

Finally, the boundary condition (D.58) yields $\varepsilon_1 E_{\rm z1} = \varepsilon_2 E_{\rm z2}$, or

$$\varepsilon_1(E_{\rm i} - E_{\rm r})\sin\theta_{\rm i} = \varepsilon_2 E_{\rm i}\sin\theta_{\rm t} \tag{D.94}$$

Making use of Snell's law again, and the fact that $\varepsilon' = n^2$, (D.94) reduces to (D.92).

Combining (D.62) and (D.93), we obtain

$$E_{\rm r} = \left(\frac{\alpha - \nu}{\alpha + \nu}\right) E_{\rm i} \tag{D.95}$$

$$E_{\rm t} = \left(\frac{2}{\alpha + \nu}\right) E_{\rm i} \tag{D.96}$$

The associated coefficients of reflection and transmission are

$$R = \left(\frac{\alpha - \nu}{\alpha + \nu}\right)^2 \tag{D.97}$$

$$T = \alpha\nu\left(\frac{2}{\alpha + \nu}\right)^2 \tag{D.98}$$

respectively. As usual, $R + T = 1$.

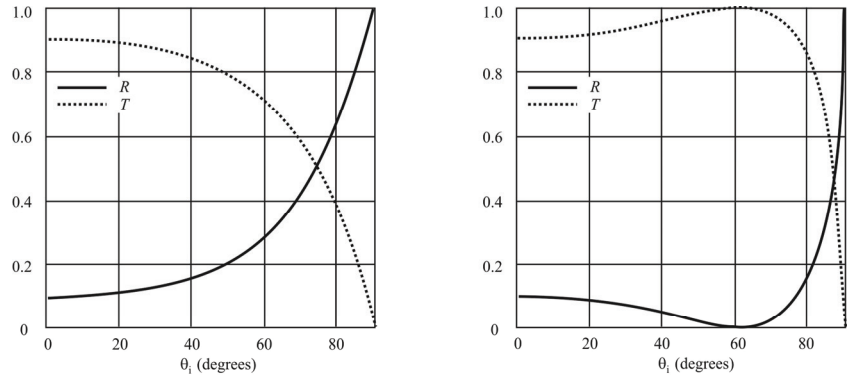


Figure D.4(a, b) Oblique incidence. Coefficients of reflection (*R*) and transmission (*T*) for (a) electric field parallel to boundary and (b) magnetic field parallel to boundary. Region 1 is air $(\varepsilon_1 = 1.0)$ and region 2 is bakelite $(\varepsilon_2 = 3.7)$.

Figure D.4 shows the coefficients of reflection and transmission for oblique incidence from air $(\varepsilon_1 = 1.0)$ to bakelite $(\varepsilon_2 = 3.7)$. Figure D.4(a) shows the wave polarization for which the electric field is parallel to the boundary, whereas Figure D.4(b) shows the wave polarization for which the magnetic field is parallel to the boundary. In general, it can be seen that the coefficient of reflection rises, and the coefficient of transmission falls, as the angle of incidence increases. Note, however, that for the second wave polarization there is a particular angle of incidence, known as the *Brewster angle*, at which the reflected intensity is zero. There is no similar behavior for the first wave polarization.

It follows from (D.95) that the Brewster angle corresponds to the condition

$$\alpha = \nu \tag{D.99}$$

or

$$\nu^2 = \frac{\cos^2\theta_t}{\cos^2\theta_i} = \frac{1-\sin^2\theta_t}{1-\sin^2\theta_i} = \frac{1-\sin^2\theta_i/\nu^2}{1-\sin^2\theta_i} \tag{D.100}$$

where use has been made of Snell's law. Equation (D.100) reduces to

Table D.1 Some Dielectric Constants

Material	Dielectric Constant	Material	Dielectric Constant
Air	1.00059	Paraffin	2–3
Aluminum silicate	5.3–5.5	Plexiglass	3.4
Bakelite	3.7	Polyethylene	2.2
Bee's wax (yellow)	2.7	Polystyrene	2.56
Butyl rubber	2.4	Porcelain	5–7
Germanium	16	Pyrex glass	5.6
Glass	4–10	Quartz	3.7–4.5
Gutta-percha	2.6	Silicone oil	2.5
Halowax oil	4.8	Steatite	5.3–6.5
Lel-f	2.6	Strontium titanate	233
Lucite	2.8	Teflon	2.1
Mica	4–8	Tenite	2.9–4.5
Micarta 254	3.4–5.4	Vacuum	1.000000
Mylar	3.1	Vaseline	2.16
Neoprene rubber	6.7	Water (distilled)	76.7–78.2
Nylon	3.00	Wood	1.2–2.1
Paper	1.5–3		

Source: [3].

$$\sin\theta_i = \frac{\nu}{\sqrt{1+\nu^2}} \tag{D.101}$$

or $\tan\theta_i = \nu = n_2 / n_1$. Hence, the Brewster angle satisfies

$$\theta_B = \tan^{-1}\left(\frac{n_2}{n_1}\right) \tag{D.102}$$

D.8 Dielectric Constants of Some Materials

The dielectric constants of some materials that are used as electrical insulators or in other applications where a dielectric material is used are given in Table D.1.

D.9 Concluding Remarks

The basic properties of dielectric materials and their effects on EM waves were discussed in this appendix. The properties of dielectric materials is such that EM waves slow when propagating through them, effectively making a small antenna have similar characteristics to larger antennas.

References

[1] http://farside.ph.utexas.edu/teaching/em/lectures/node92.html. Accessed 10/31/2008.
[2] Toll, J. S., "Causality and the Dispersion Relation: Logical Foundations," *Physical Review*, Vol. 104, 1956, pp. 1760–1770.
[3] http://www.vias.org/encyclopedia/phys_dielectric_const.htm.

List of Acronyms

2-D	2 dimensional
3-D	3 dimensional
A	ampers
AC	alternating current
ADC	analog-to-digital ocnverter
AF	array factor
AM	amplitude modulation
AOA	angle of arrival
APC	armoured personnel carrier
ASE	aircraft self-protect equipment
AWGN	additive white Gaussian noise
$\vec{B}$	magnetic field (vector)
β	wave vector
balun	balanced to unbalanced
BJT	bipolar junction transistor
BW	bandwidth
C2	command and control
CCW	counterclockwise
CDAA	circularly disposed antenna array
CDMA	code division multiple access
Ci	cosine integral
CM	common mode
CW	continuous wave

D	directivity
dB	decibel
dBi	dB relative to isotropic
DC	direct current
DF	direction finding
DFT	discreet Fourier transform
DL	diagonal loading
DN	double negative
DOA	direction of arrival
DP	double positive
$\vec{E}$	electric field
EA	electronic attack
E-field	electrical field
ELINT	electronic intelligence
EM	electromagnetic wave
EMF	electromotive force
EMI	electromagnetic interference
ERP	effective radiated power
ES	electronic support
ESA	electronic steerable array
EW	electronic warfare
F	intensity pattern function
f	frequency
FDL	fixed diagonal loading
FET	field effect transistor
FFT	fast Fourier transform
FM	frequency modulation
FNBW	first null bandwidth
FSK	frequency division multiplex
GA	genetic algorithm

GaAs	gallium arsenide
Ge	germanium
GHz	gigaHertz
GLC	general linear combination
$\vec{H}$	magnetic field (vector)
HF	high frequency
HOS	high-order statistics
HPBW	half-power beamwidth
I	current
IC	integrated circuit
IF	intermediate frequency
IFB	independent frequency bin
IFS	iterative function system
IMD	intermodulation distortion
INR	interference-to-noise ratio
IQ	inphase/quadraphase
ISB	independent sideband
$\vec{J}$	current density (vector)
JFET	junction field effect transistor
kHz	kiloHertz
λ	wavelength
LC	inductor/capacitor
LNA	low-noise amplifier
LPA	log-periodic antenna
LPDA	log-periodic dipole array
LPMA	log-periodic monopole array
LSB	lower sideband
LSMI	loaded SMI

max	maximum
ME	maximum entropy
MEMS	microelectromechanical system
MESFET	metal semiconductor field effect transistor
MF	medium frequency
M-field	magnetic field
MHz	megaHertz
MIC	microwave integrated circuit
min	minimum
MMIC	monolithic microwave integrated circuit
MMM	man-made noise
MoM	method of moments
MPA	microstrip patch antenna
MUSIC	multiple signal classification
MVDR	minimum variance distortionless response
NATO	North Atlantic Treaty Organization
NCCB	norm-constrained Capon beamformer
NEC	national electrical code
NIC	negative impedance converter
OFDM	offset frequency division multiplex
OSVF	orthogonal variable spreading factor
OTM	on the move
PA	power amplifier
PAM	pulse amplitude modulation
PC	personal computer
PCB	printed circuit board
PCS	personal communiction ystem
PDM	pulse density modulation

PE	solid polyethylene
PEC	perfect electrical conducting
PF	polyethylene foam
pHEMT	pseudomorphic high electron mobility transistor
PIN	positive-intrinsic-negative
PLF	polarization loss factor
PM	phase modulation
PSD	power spectral density
PWB	pulse width modulatoin
Q	quality factor
QAM	quadrature amplitude modulation
ρ	electric charge density
R	resistance
RCB	robust Capon beamformer
RF	radio frequency
RFID	RF identification
RLC	resistor/inductor/capacitor
RMS	root-mean-square
RWR	radar warning receiver
SA	simulated annealing
SCB	standard Capon beamformer
Si	sine integral
Si	silicon
SINAD	signal to noise and distortion ratio
SINR	signal to interference-and-noise ratio
SMI	sample matrix inversion
SNR	signal-to-noise ratio
SOI	signal of interest
SOI	surface area
SPDT	single-pole double-throw

SPMT	single-pole multiple-throw
SPST	single-pole single-throw
SSB	single sideband
T	period
TE	transverse electric
TEM	transverse electric and magnetic
TLT	transmission line transformer
TM	transverse magnetic
TOIP	third order intercept point
UAS	unattended aerial system
UCA	uniform circular array
UHF	ultra high frequency
ULA	uniform linear array
UMTS	universal mobile terminal system
unun	unbalanced to unbalanced
USD	upper sideband
V	volts
VHF	very high frequency
VP	vector potential
VSW	voltage standing wave
VSWR	voltage standing wave ratio
W	watts
WCDMA	wideband CDMA
WLAN	wide area local area network
X	reactance
Z	impedance

About the Author

Richard A. Poisel received a B.S. in electrical engineering from the Milwaukee School of Engineering in 1969 and an M.S. in the same discipline from Purdue University in 1971. He spent three years in the military service from 1971 to 1973. After his 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. During the 1993–1994 academic year, Dr. Poisel attended the MIT Sloan School of Management as a Sloan Fellow, receiving an M.B.A. Initially a research engineer, Dr. Poisel eventually rose to the role of the director of the laboratory on an acting basis from 1997 to 1999. He was appointed chief scientist in 1999 and was relocated to the Army's Intelligence Center at Ft. Huachuca, Arizona, where he served as a technical advisor to the command group. Retiring from government service in 2004, he served as a Senior Engineering Fellow at Raytheon Missile Systems in Tucson, Arizona, from 2004 to 2011. He is currently a consultant on engineering for electronic warfare applications.

Dr. Poisel holds patents in the communication Electronic Warfare subject area and is the author of the following, all published by Artech House: *Introduction to Communications Electronic Warfare Systems*, 2nd ed., 2008; *Modern Communications Jamming Principles and Techniques*, 2nd ed., 2011; *Target Acquisition in Communications Electronic Warfare Systems*, 2004; *Foundations of Communications Electronic Warfare*, 2008; and *Electronic Warfare Target Location Methods*, 2005.

Index