Danko Antolovic

Radiolocation in Ubiquitous Wireless Communication

Springer

Radiolocation in Ubiquitous Wireless Communication

Danko Antolovic

Radiolocation in Ubiquitous Wireless Communication

Springer

Danko Antolovic
Indiana University
University Information Technology Services
2711 East 10th Street
Bloomington, IN 47408
USA
dantolov@indiana.edu

ISBN 978-1-4419-1631-0 e-ISBN 978-1-4419-1632-7
DOI 10.1007/978-1-4419-1632-7
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009941543

In memory of
Zorka Rojc (1900–1985)
An early teacher,
and a lifelong influence

Preface

This volume has its beginnings in a laboratory project, development of a radiolocator for the Wi-Fi network that was growing by leaps and bounds on the campus of Indiana University at that time. What started as a very focused and practical attempt to improve network management, touched in its lifetime upon broader issues of the use of radio spectrum, design of system architectures for the wireless medium, and image formation outside the limits of geometrical optics.

I have intended this book mostly for the audience of engineers and system designers, in the growing field of radio communication among small, portable, ubiquitous devices that have become hybrid platforms for personal communication and personal computing. It is also a book addressed to network professionals, people to whom radio is largely a black box, a medium that they usually rely upon, but seldom fully understand.

In fact, in the course of my work in the field, I have witnessed, to my dismay, a wide disconnect between the networking world and the radio technology that networking has come to depend upon so heavily. Perhaps, because digital wireless communication is seen as digital first and wireless second, there is often a misplaced emphasis on its information-processing side, with the methodology centered around the discrete symbol, and with little intuition of the underlying physics. I had it once suggested to me, in apparent seriousness, to use radio cards for intra-system communication within a radiolocator!

Wireless communication *is* radio, plain and simple. Radio is what makes it both powerful and frustrating, and radio, an old technology after all, is the key to understanding its idiosyncrasies. In the broadest strokes, my motivation for writing this book is to help bridge this gap in knowledge, not by offering one more textbook on E&M, but by delving into the physical layer of wireless communication, through the exposition of a technology that I have had the opportunity to contribute to. The attentive reader will notice, though, that the real backdrop for this book is the physics of the electromagnetic wave, rather than digital communication.

That said, there remains a real need for a better set of diagnostic tools for wireless professionals, and for a more ambitious engineering reach into the physical layer, beyond simply making the transceiver work. Multitudes of mobile devices crowd the physical spaces, as well as the available spectrum, opening up problems in security and logistics. The fact that radio transmissions are by nature omnidirectional, and

that it is neither cheap nor easy to restrict them in space, throws the security level of a wireless network back to that of an old-fashioned Ethernet bus: everybody can hear everything, and the network relies on every player's decency not to eavesdrop or interfere. Obviously, true security in the radio medium can be found only in good encryption – this is a lesson that goes as far back as World War II and Bletchley Park – but it does help to know where your interlocutors are. It is always difficult to quantify things that did not happen, but I am inclined to believe that a lot of wired mischief is deterred by the fact that your wired connection can be traced to your desk, at least within a typical workplace!

Knowing the location, or at least the general direction of the other party, is helpful with the logistics of radio communication as well. When radiolocation is combined with the ability to steer the transmission, radio communication becomes something in the nature of a focused personal conversation. In contrast, omnidirectional broadcast appears more like yelling in an open field, and while this is the right mode for a traditional *broadcasting* station, the "personal conversation" mode is obviously more appropriate for the point-to-point communications that occur in networks. Just as in verbal communication, directed transmission lowers the volume, reduces power consumption, and improves audibility in general, by reducing the speakers' mutual interference. In wireless jargon this is called "improved quality of service," and is being applied to some extent in cellular phone systems, by using directional antennas or by configuring multiple towers as rudimentary phased arrays. Other areas of ubiquitous wireless communication still present an untouched field for the application of radiolocation.

Exposition in this book tends to follow the direction from the general and abstract toward the specific and concrete. It is centered around an engineering project, which hopefully anchors and focuses the discussion. To be more specific, the exposition centers on radiolocation by measuring signal amplitudes in multiple directions, and on an actual device that performs this radiolocation around the full circle of the horizon.

The medium of radio is, of course, fundamentally analog, and radiolocation is really an exercise in quantitative measurement. For that reason, I have emphasized the quantitative and real-time aspects of the architecture and design. As the book progresses toward the full description of the implemented architecture, the discussion explores design alternatives, and seeks to justify the engineering choices that were made along the way.

Chapter 1 offers an overview of physical phenomena underlying radio communication: it describes the electromagnetic wave and its interactions with matter, leading to the highly important topic of antenna physics. This chapter will be of most use to those engineering denizens of the digital world who do not routinely venture below the data link layer.

Chapters 2 and 3 deal with the mathematical issues that are at the core of reconstructing the direction of the radio wave. Chapter 2 develops the direction-finding algorithm and investigates its numerical aspects; this chapter is fundamental to understanding the implementation of radiolocation, as we describe it in subsequent chapters. Chapter 3 develops a broader method of radio imaging, including multi-

ple sources and radiolocation over the full sphere of directions. This chapter lays theoretical groundwork for future imaging architectures, and the reader may choose to omit it on first reading.

Chapters 4 to 7 build the architecture of the Wi-Fi radiolocator, from the antenna to the complete radiolocation transceiver; these chapters form the engineering core of the book. Chapter 8 aggregates various implementation details that are peripheral to the main topic, but are nevertheless essential for the device design. Chapter 9 looks forward, toward applying the lessons of this work to other, potentially more intricate wireless protocols.

In following this book, the reader will benefit from some previous familiarity with the concepts in wireless communication, and from college-level physics and mathematics. The discussion of radiolocation is supplemented with topics in radio fundamentals, and the exposition is grounded in basic physics, but the reader will not be hampered by the lack of specialized background knowledge. Where it seemed necessary, I have provided introductions to some well-established side topics, in the form of appendices. These appendices contain additional mathematical details, or give concept summaries, but they are not meant to be complete expositions. Throughout the book, I have also provided references to standard, and hopefully approachable, textbooks in the field.

My greatest expectation of the reader may be that of familiarity with board-level electronics and embedded systems design. Practical experience in these areas will be helpful in envisioning some of the implementation details, details that I left out in order not to drift too far away from the main topic. Attempting to cover that background would have resulted in an entirely different book!

Bloomington, IN *Danko Antolovic*

Acknowledgments

Every book draws upon the contributions and help of many people and institutions. I wish to acknowledge here Steven Wallace who, as director of the Advanced Network Management Lab at Indiana University, recognized the importance of radiolocation in wireless communication early on. Building the prototypes described in this book was greatly facilitated by the efforts of John Poehlman and the excellent staff of Electronic Instrument Services of the Chemistry Department, at Indiana University.

I would also like to thank Douglas Palmer of Cal(IT)^2, and to the CalRadio team at University of California at San Diego, for making an early prototype of their wireless platform available to me.

Scott Wayne and Larry Hawkins of Analog Devices Inc. have graciously agreed to review Section 5.2, and I wish to thank Analog Devices for the permission to reproduce their copyrighted materials in the book. Likewise, thanks are extended to ACM and IEEE associations for the permission to use copyrighted materials.

My thanks go to Springer science editor Ephraim Suhir, and to Springer editors Jennifer Mirski, Ciara Vincent and Brett Kurzman for their help during the preparation of the manuscript.

Finally, a warm word of recognition goes to my colleagues Bryce Himebaugh and Caleb Hess, and to my former mentor Steve Johnson of the IU Computer Science Department. I appreciate the years of our collaboration and friendship, for knowledge is indeed advanced best through a free and generous exchange of ideas.

Contents

Chapter 1
Physical Principles of Radio Communication

1.1 Introduction

Electrodynamics is an established science. Ever since James Clerk Maxwell spelled out its fundamental equations in 1865, its tenets have been verified and reverified by measurements, its formalism developed and made more elegant. It is also a science with wide application, since, except for the force of gravity, the vast majority of phenomena with which we come into contact every day is electrical in nature. Light and radiated heat, radio waves of all kinds, X-rays and UV rays are all manifestations of the same basic entity: the electromagnetic wave/photon. The differences between these phenomena stem entirely from different ways in which photons of different wavelength interact with matter.

In conjunction with quantum-mechanical principles, electrical force underlies the structure of atoms and molecules, and therefore all of chemistry, crystallography, and molecular biology as well. Solid and liquid state, all that we perceive as bulk or extension in space, is maintained by a quantum-mechanical balance of electromagnetic forces.

The same electrical force provides a remarkably flexible and efficient method for the transport and distribution of energy, the electric grid. It is the basis of our communication and digital information technology, not to even speak of consumer electronics.

We cannot do justice to this vast field in this book, nor is that our purpose; excellent textbooks on the subject abound (Jackson 1998, Pozar 1998, Russer 2006, Born and Wolf 2006). In this introductory chapter we review topics in electrodynamics that form the foundations for understanding radio technology. We hope to convey to the reader some of the subtlety of the formalism of electrodynamics, although we relegate one or two of the more intricate mathematical topics to the Chap. 10. We begin with that simple and remarkable phenomenon, the electromagnetic wave in empty space.

D. Antolovic, *Radiolocation in Ubiquitous Wireless Communication*,
DOI 10.1007/978-1-4419-1632-7_1,

1.2 Electromagnetic Wave in Empty Space

In the absence of charges, which are always physically associated with material particles, electromagnetism is a disturbance in the vacuum, disturbance which can carry energy over distance, and which manifests itself by exerting mechanical force on charges that it encounters. That is the classical pre-quantum view, which is perfectly adequate for our purposes.

This disturbance is physically described as the presence of two vector fields, electric field **E** and magnetic field **H**, at any point in space. These two fields are mutually dependent, and their development in time and space is governed by four vector differential equations, which are known as the Maxwell's equations.

Maxwell's equations can be expressed in several notational conventions and unit systems, leading to occasional confusion.[1] We do not need to consider the equations in their full generality yet, and we will start the discussion with the following simple form, which applies in the vacuum, that is in the absence of either charges or a material medium:

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t} \tag{1.1}$$

$$\nabla \times \mathbf{H} = \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} \tag{1.2}$$

$$\nabla \cdot \mathbf{E} = 0 \tag{1.3}$$

$$\nabla \cdot \mathbf{H} = 0 \tag{1.4}$$

Let us first review the physical meaning of the two spatial differential operators that appear in the above equations.

Divergence operator, $\nabla\cdot$ acts on vector fields, and yields scalar fields. Imagine an infinitesimal volume τ, surrounding the point in space: divergence is the total flux of the vector field **V** through the boundary surface σ of that volume, divided by the volume:

$$\nabla \cdot \mathbf{V} = \lim_{\Delta\tau \to 0} \frac{1}{\Delta\tau} \int_{\sigma} \mathbf{V} \cdot \mathrm{d}\boldsymbol{\sigma} \tag{1.5}$$

This is obviously a scalar, and it measures the "strengthening" or "weakening" of the field at the point. If the flux into the volume equals the flux out of the volume, the field is "conserved" and its divergence is zero. Otherwise, the small volume contains a source or a sink of the field, such as an electric charge.

Curl operator, $\nabla\times$ acts on vector fields and results in new vector fields. Now, imagine an infinitesimal area σ around, and including, the point in space; unit vector **n**

[1] In this book, we follow the SI (also known as the MKSA) unit system.

is perpendicular to that surface. Curl yields the circulation integral of the vector field $\mathbf{V}$ around the boundary curve λ of the surface, divided by the surface area:

$$(\nabla \times \mathbf{V}) \cdot \mathbf{n} = \lim_{\Delta\sigma \to 0} \frac{1}{\Delta\sigma} \oint_{\lambda} \mathbf{V} \cdot \mathrm{d}\boldsymbol{\lambda} \tag{1.6}$$

When the circulation integral is different from zero, the field has a rotational component: If $\mathbf{V}$ were the field of velocities in a moving fluid, and σ a small paddlewheel, the paddlewheel would rotate, since the overall torque exerted on its boundary would be nonzero. Notice that the circulation integral, which is a scalar, is the component of the curl in the direction $\mathbf{n}$, and changes with the orientation of the surface σ in space. The curl $\nabla \times \mathbf{V}$ itself is a vector which gives the magnitude and direction of the vorticity, i.e., the "curliness," of the field $\mathbf{V}$. Inherent in the definition of the curl is the intrinsic orientation of the surface σ; therefore, curl is an axial or pseudovector, which changes sign if the coordinate system is switched between left- and right-hand orientations.

We have presented divergence and curl in their integral forms, which are highly intuitive, but these are true differential operators in spatial coordinates. As (1.5) and (1.6) show, they are limit ratios of infinitesimal quantities defined by volumes, surfaces, and curves. These two operators have more commonly used differential forms, which are suggested by their vector notation: They can be expressed as formal scalar and vector products of the vector $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$, called del or nabla, with the operand vector field $\mathbf{V}$. For further exposition of this topic, the reader should consult any standard textbook on electrodynamics or mathematical physics (Jackson 1998; Byron and Fuller 1992; Morse and Feshbach 1953).

Returning to Maxwell's equations, (1.1) and (1.2) represent Faraday's and Ampere's laws of induction in the absence of currents, stating that the time change in one field induces spatial change in the other. Equations (1.3) and (1.4) say that both fields are divergenceless, as they must be in the absence of charges.

Let us mention here that there is an asymmetry in the electrodynamics of the world, in the sense that there are no known magnetic charges. In the presence of electric charges, (1.3) acquires the form $\nabla \cdot \mathbf{E} = \rho$, where ρ is the charge density, but (1.4) always holds. The formalism of electrodynamics would allow magnetic charges, and there are no known reasons for them not to exist – they have simply not been observed so far.

Quantities μ_0 and ε_0 are known as the permeability and the permittivity of the vacuum. They are related to the speed of light in the following way:

$$c = (\mu_0 \varepsilon_0)^{-1/2} \tag{1.7}$$

and their values are, in SI units: $\mu_0 = 4\pi \times 10^{-7}\,\mathrm{H/m}$, $\varepsilon_0 = 8.85 \times 10^{-12}\,\mathrm{F/m}$.

Equations (1.1) and (1.2) are first-order differential equations, coupling the fields $\mathbf{E}$ and $\mathbf{H}$. We can separate the fields into two equations, at the price of having equations of the second order: We first act upon (1.1) with

$\nabla\times$ and upon (1.2) with $\partial/\partial t$ and subtract the equations; subsequently we reverse the operators and repeat the steps. The results are:

$$\nabla \times (\nabla \times \mathbf{E}) = -\mu_0 \varepsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} \tag{1.8}$$

$$\nabla \times (\nabla \times \mathbf{H}) = -\mu_0 \varepsilon_0 \frac{\partial^2 \mathbf{H}}{\partial t^2} \tag{1.9}$$

The curl-of-curl terms can be brought into a more familiar form by invoking this vector identity[2]:

$$\nabla \times (\nabla \times \mathbf{V}) = \nabla(\nabla \cdot \mathbf{V}) - \nabla^2 \mathbf{V} \tag{1.10}$$

and the divergence conditions (1.3) and (1.4). The resulting equations for $\mathbf{E}$ and $\mathbf{H}$ are:

$$\nabla^2 \mathbf{E} = \frac{1}{c^2} \frac{\partial^2 \mathbf{E}}{\partial t^2} \tag{1.11}$$

$$\nabla^2 \mathbf{H} = \frac{1}{c^2} \frac{\partial^2 \mathbf{H}}{\partial t^2} \tag{1.12}$$

The operator $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ is known as the Laplacian. It acts on scalar fields, as well as component-wise on vector fields, and yields scalars or vectors correspondingly. Physically, Laplacian of a field at a point in space is proportional to the difference between the average value of the field over an infinitesimal volume around the point, and the actual value of the field at that point (see Sect. 10.1).

Equations (1.11) and (1.12) have the familiar form of the wave equation. The fields change in time by accelerating toward their average values, but since their first time derivative is not zero at equilibrium, they overshoot and continue in an oscillatory motion. From these equations we see that electromagnetic phenomena in empty space, away from the material anchors of electric charges and currents, must be wavelike disturbances propagating at the speed of light.

1.3 The Plane Wave

It follows from (1.11) and (1.12) that every component of the fields $\mathbf{E}$ and $\mathbf{H}$ satisfies the wave equation

$$\nabla^2 u = \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} \tag{1.13}$$

[2] This identity makes no intuitive sense in terms of the properties of vector differential operators, until we realize that it is an application of a well-known *algebraic* formula for the vector product, $\mathbf{x} \times (\mathbf{y} \times \mathbf{z}) = \mathbf{y}(\mathbf{x} \cdot \mathbf{z}) - \mathbf{z}(\mathbf{x} \cdot \mathbf{y})$, to the operator ∇.

As we mentioned in Sect. 1.2, its form indicates oscillatory behavior of its solutions. This equation is also homogeneous and linear, which means that any linear combination of any solutions (if solutions exist) is also a solution. Naturally, this applies to (1.11) and (1.12) as well.

Equation (1.13) has, in fact, great many solutions, but in this section we shall explore one simple solution in the form:

$$u(\mathbf{x}, t) = u_0 \exp(i\mathbf{k} \cdot \mathbf{x} - i\omega t) \tag{1.14}$$

In this equation, $\mathbf{x}$ is the position vector, and $\mathbf{k}$ is known as the wave vector or the propagation vector. The magnitude of the wave vector

$$|\mathbf{k}| = k = \left(k_x^2 + k_y^2 + k_z^2\right)^{1/2} \tag{1.15}$$

is called the wave number, and for (1.14) to be a solution of the wave equation (1.13), the relation

$$k = \frac{\omega}{c} \tag{1.16}$$

must be true. Equation (1.14) represents, of course, a sine wave with frequency $\nu = \omega/(2\pi)$, and wavelength $\lambda = 2\pi/k$. Surfaces of constant phase, $\mathbf{k}\cdot\mathbf{x} - \omega t = const$, are planes perpendicular to $\mathbf{k}$, and we conclude that this is a plane wave, traveling in the direction $\mathbf{k}$ in space.

So far, nothing in this discussion is specific to electrodynamics: for example, when applied to two dimensions, this formalism describes oscillations of an elastic membrane as well. Let us now introduce the electromagnetic fields in the form.

$$\mathbf{E} = \mathbf{E}_0 \exp(i\mathbf{k} \cdot \mathbf{x} - i\omega t) \tag{1.17}$$

$$\mathbf{H} = \mathbf{H}_0 \exp(i\mathbf{k} \cdot \mathbf{x} - i\omega t) \tag{1.18}$$

In these equations, $\mathbf{E}_0$ and $\mathbf{H}_0$ are constant vectors, unchanging in time and same everywhere in space. In analogy with the analysis for u in (1.13), fields $\mathbf{E}$ and $\mathbf{H}$ are plane-wave solutions of (1.11) and (1.12), respectively. We may ask, naively, whether the two fields must really have the same wave vector and the same frequency, and whether there are any restrictions on the vectors $\mathbf{E}_0$ and $\mathbf{H}_0$. Wave equations (1.11) and (1.12) permit arbitrary waves, as long as (1.16) holds; to fully understand the physics of the electromagnetic wave, we must go back to (1.1)–(1.4).

We notice, first, that because of the differentiation properties of the exponential function, vector differential operations become algebraic operators for vectors of the type $\mathbf{V} = \mathbf{V}_0 \exp(i\mathbf{k} \cdot \mathbf{x} - i\omega t)$:

$$\frac{\partial}{\partial t}\mathbf{V} = -i\omega\mathbf{V} \tag{1.19}$$

$$\nabla \times \mathbf{V} = i\mathbf{k} \times \mathbf{V} \tag{1.20}$$

$$\nabla \cdot \mathbf{V} = i\mathbf{k} \cdot \mathbf{V} \tag{1.21}$$

For this more restricted type of fields, Maxwell's equations become

$$\mathbf{k} \times \mathbf{E} = \mu_0 \omega \mathbf{H} \tag{1.22}$$

$$\mathbf{k} \times \mathbf{H} = -\varepsilon_0 \omega \mathbf{E} \tag{1.23}$$

$$\mathbf{k} \cdot \mathbf{E} = 0 \tag{1.24}$$

$$\mathbf{k} \cdot \mathbf{H} = 0 \tag{1.25}$$

Immediately we see that the two fields must behave as a single wave. Suppose that there were separate electric and magnetic frequencies and wave vectors; it follows from (1.22), or similarly from (1.23) that

$$(\mathbf{k}_\mathrm{E} \times \mathbf{E}_0) \exp(i\,\mathbf{k}_\mathrm{E} \cdot \mathbf{x} - i\,\omega_\mathrm{E} t) = \mu_0 \omega_\mathrm{H} \mathbf{H}_0 \exp(i\,\mathbf{k}_\mathrm{H} \cdot \mathbf{x} - i\,\omega_\mathrm{H} t) \tag{1.26}$$

Oscillatory terms on the two sides are always proportional, which is possible only if $\mathbf{k}_\mathrm{E} = \mathbf{k}_\mathrm{H}$ and $\omega_\mathrm{E} = \omega_\mathrm{H}$.

Equations (1.22) and (1.23) also show that the vectors **k**, **E** and **H** are always mutually orthogonal, and that, in that order, they form a right-handed coordinate system. The electromagnetic wave in vacuum is transversal in both of its components, and the components are mutually orthogonal as well (see Fig. 1.1).

Here, we invoke without proof the Poynting theorem, a general theorem about the energy flow in electromagnetic fields (see Jackson 1998). This theorem states that the power flux in the field equals $\mathbf{S} = \mathbf{E} \times \mathbf{H}$, where the vector **S**, called the Poynting vector, represents the flux. We see that, for the plane wave, **S** points in the same direction as **k**: The wave carries energy in the direction of its propagation, as one would expect.

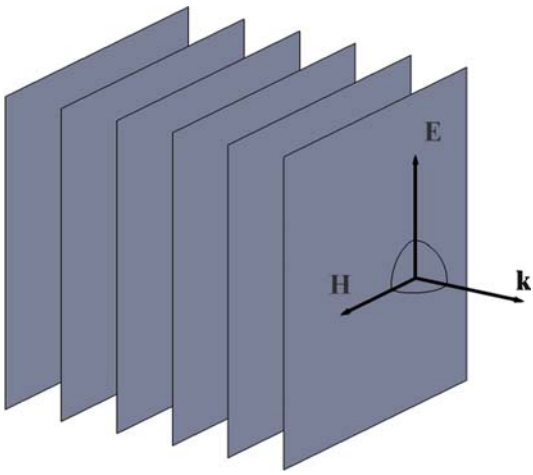

Fig. 1.1 Electromagnetic fields in the plane wave. Vectors **k**, **E** and **H** form a right-handed coordinate system

Finally, if we consider only the magnitudes of the vector fields, it follows from (1.22) (or 1.23), and from (1.16) and (1.7), that the ratio of electric and magnetic amplitudes is the following constant:

$$\frac{E}{H} = \sqrt{\frac{\mu_0}{\varepsilon_0}} = Z_0 \tag{1.27}$$

This constant has the dimensions of resistance; it is called the impedance of free space, and its numerical value is ca. 377 Ω.

1.4 Electromagnetic Wave Within Matter

In the two previous sections, we discussed electromagnetic waves in the simplest environment – empty space – but actual radio communication must contend with material objects. As we already mentioned in passing, the structure of our everyday matter is determined by the electric force, and it is the richness of this structure that gives rise to the variety of interactions between the electromagnetic wave and material objects. These remarks are true for all of the electromagnetic radiation, but we concentrate here on the part of the spectrum that is commonly known as radio waves.

On the largest terrestrial scales, radio waves encounter the surface of the sea and land, interact with the ionized upper atmosphere, and traverse large distances around the globe. Gases interact very little with radio waves, but all condensed matter does, and even mist in the air has an effect over long distances.

Coming closer to the human scale, radio waves are blocked by mountains, reflected by tall buildings, attenuated by walls, absorbed by foliage. On the size scale of an office, short-range microwave communication is affected by the presence of metallic structures inside walls, by furniture, and even by people passing by. Since microwave radio in this cluttered setting is our focus of interest, we devote some attention in this chapter to understanding the two main types of materials encountered: dielectrics and conductors.

When describing electrodynamics within material media, it is common to introduce two new fields: electric displacement **D** and magnetic induction **B**. They are related to the electric and magnetic field:

$$\mathbf{D} = \varepsilon \mathbf{E} \tag{1.28}$$

$$\mathbf{B} = \mu \mathbf{H} \tag{1.29}$$

Quantities ε and μ describe the aggregate response of the material to the fields **E** and **H**. In the vacuum they equal ε_0 and μ_0 by definition, and they are frequency-dependent scalars for most materials, but they become tensors in anisotropic crystalline solids.

In conductive media, Ohm's law also applies,

$$\mathbf{J} = \sigma \mathbf{E} \tag{1.30}$$

where $\mathbf{J}$ is the current vector, and σ is the conductivity, another aggregate response quantity.

In order to discuss the presence of material medium, we quote Maxwell's equations in their general form:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{1.31}$$

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J} \tag{1.32}$$

$$\nabla \cdot \mathbf{D} = \rho \tag{1.33}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{1.34}$$

Ampere's law (1.32) now includes currents, i.e., moving charges, and the electric divergence equation (1.33) allows electric charges. The equations contain quantities which describe aggregate properties of the matter and are not part of the electromagnetic theory proper. These quantities must be determined by empirical observation, or derived from a theoretical description of the medium.

Most materials fall into one of two broad classes: dielectrics, within which the macroscopic current $\mathbf{J}$ is identical to zero, and conductors, in which it is not. Dielectrics, crystalline or amorphous, consist of rigid covalent structures, in which every electron is confined to a small region, either an atom or a chemical bond; in the presence of an external field, an electron can depart from its average position only very little.

Conductors are characterized by freely moving charges. In the crystalline lattice of metals, a portion of the available electrons resides in quantum states that are distributed over the bulk of the solid. These electrons can move relatively freely from one end of the piece of metal to the other, giving rise to macroscopic currents. Likewise, conductive ionic solutions (such as seawater) contain moving charges in the form of charged molecular species (ions), which again move relatively freely through the nonrigid structure of the liquid.

This classification is somewhat crude, because the distinction is a matter of degree in, for example, semimetals and semiconductors, but it nevertheless accounts well for many observed phenomena. Now, we examine the behavior of the electromagnetic wave within these two types of material (Fowles 1989).

For our purposes we can assume that the material is non-ferromagnetic, i.e., $\mu \approx \mu_0$, and electrically neutral, ρ=0. It is also convenient to express the electric effect of the medium as the polarization vector.

$$\mathbf{P} = \mathbf{D} - \varepsilon_0 \mathbf{E} \tag{1.35}$$

With that, we derive the wave equation for the electric field, following steps analogous to the derivation of (1.8) and (1.11)

$$\nabla^2\mathbf{E} - \frac{1}{c^2}\frac{\partial^2\mathbf{E}}{\partial t^2} = \mu_0\left(\frac{\partial^2\mathbf{P}}{\partial t^2} + \frac{\partial\mathbf{J}}{\partial t}\right) \tag{1.36}$$

Relative magnitude of the two source terms on the right side is what distinguishes dielectrics from conductors.

1.4.1 Dielectrics

In line with earlier discussion, we describe an electron in a dielectric as a charged, massive particle bound to its location by an elastic force. Its complex interactions with the neighboring matter are simplified into a frictional force, proportional to the speed of its movement. This is a very simplistic description of a dielectric, but it suffices to account qualitatively for the properties important in radio wave propagation. The strategy is to determine how the polarization $\mathbf{P}$ depends on $\mathbf{E}$, and to solve the wave equation (1.36) to obtain both the electric field $\mathbf{E}$ and a dispersion relation analogous to (1.16).

Classical equation of motion of an electron, under the influence of the electric force due to an external wave $\mathbf{E}$, is as follows:

$$m\left(\frac{d^2\mathbf{r}}{dt^2} + \gamma\frac{d\mathbf{r}}{dt} + \omega_0^2\mathbf{r}\right) = -e\mathbf{E} \tag{1.37}$$

where m is the electron's mass, e is its charge, $\mathbf{r}$ is the displacement from its unperturbed position, γ is the frictional drag, and $\omega_0^2 = K/m$ is the resonant frequency of an oscillating mass bound by a force with an elastic constant K. In practice, γ and K are not easy to evaluate, since they are derived from quantum-mechanical descriptions of the medium.

Assuming the harmonic time behavior $\exp(-i\omega t)$ for both $\mathbf{E}$ and $\mathbf{r}$ (a reasonable guess for the latter), the equation of motion becomes

$$m\left(-\omega^2 - i\omega\gamma + \omega_0^2\right)\,\mathbf{r} = -e\mathbf{E} \tag{1.38}$$

Polarization is the measure of charge displacement in the medium. If N is the charge density,

$$\mathbf{P} = -Ne\,\mathbf{r} \tag{1.39}$$

the wave equation (1.36) becomes, after eliminating $\mathbf{r}$ from (1.38) and (1.39), and some algebra:

$$\nabla^2\mathbf{E} = \frac{1}{c^2}\left(1 + \frac{Ne^2}{m\varepsilon_0}\cdot\frac{1}{\omega_0^2 - \omega^2 - i\gamma\omega}\right)\frac{\partial^2\mathbf{E}}{\partial t^2} \tag{1.40}$$

As in the case of (1.11), we try the plane-wave form (1.17) of the electric field. This becomes a solution for (1.40) if the following analog of the condition (1.16) is satisfied:

$$k^2 = \frac{\omega^2}{c^2}\left(1 + \frac{Ne^2}{m\varepsilon_0} \cdot \frac{1}{\omega_0^2 - \omega^2 - i\gamma\omega}\right) \tag{1.41}$$

The dispersion equation (1.41) indicates that the wave number has an imaginary component, and that the wave is attenuated as it propagates through the dielectric. This is a direct consequence of the frictional term in (1.37); this term represents the irreversible dissipation of the wave's energy, as the oscillating electron interacts randomly with many other particles in the medium, and converts the wave's energy into heat.

We can express the wave number as the complex index of refraction, $n = (c/\omega)k = \nu + i\kappa$, a "normalized" quantity that equals one in the vacuum, and we obtain from (1.41):

$$\nu^2 - \kappa^2 = 1 + \frac{Ne^2}{m\varepsilon_0} \cdot \frac{\omega_0^2 - \omega^2}{(\omega_0^2 - \omega^2)^2 - \gamma^2\omega^2} \tag{1.42}$$

$$2\nu\kappa = \frac{Ne^2}{m\varepsilon_0} \cdot \frac{\gamma\omega}{(\omega_0^2 - \omega^2)^2 - \gamma^2\omega^2} \tag{1.43}$$

Without trying to solve for ν and κ, we see from these equations that for waves at frequencies sufficiently far away from the resonance, κ is almost zero, and ν is close to one: The attenuation is low, and the dielectric is essentially transparent, bending the passing wave somewhat as glass bends the light rays.

At the resonant frequency of the oscillating electron, however, the attenuation index ν has a maximum, and the dielectric grows opaque. This is not surprising, because the amplitude of the electrons' oscillation rises rapidly at the resonance, dissipating the wave's energy to a greater degree.

Refraction index n arises naturally in the discussion of wave propagation, and since in lossless medium the phase velocity equals $u = c/n$, higher refraction index means denser material, in which the wave propagates more slowly. Of course, the name originates in Snell's law of refraction, which relates the angles of incident (θ) and transmitted (φ) waves at the boundary between air ($n = 1$) and a denser material with index n:

$$\frac{\sin\theta}{\sin\varphi} = n \tag{1.44}$$

Instead of the refraction index, electronics engineers tend to use the dielectric constant, for which a better name is relative permittivity, and which is defined as $\varepsilon_r = \varepsilon/\varepsilon_0$. It can be easily shown that the two quantities are related as

$$\varepsilon_r = n^2 \tag{1.45}$$

Dissipative media can be described by a complex dielectric constant, and the term "loss tangent" is commonly used for the ratio of the constant's imaginary and real parts. As an example, the dielectric commonly used as a substrate for printed circuit boards, FR-4 (a fiberglass and epoxy resin composite) has $\varepsilon_r \approx 4$ at 1 GHz. Signal speed in FR-4 is therefore about $c/2$. The loss tangent is small, ca. 0.01, or in the language of this section, resonant frequency of this dielectric is above the 1 GHz band.

1.4.2 Conductors

Charges in a conductor move freely, subject only to the external field, and to a dissipative frictional drag which represents their average interaction with the background structure of the conductor. As in Sect. 1.4.1, we seek to express the source term in the wave equation (1.36) in terms of the external field **E** in order to solve the equation for **E**, and to obtain the dispersion formula.

The equation of motion of the electron is

$$m\left(\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t}+\gamma\mathbf{v}\right)=-e\mathbf{E} \tag{1.46}$$

or more conveniently, in terms of the macroscopic current density $\mathbf{J}=-Ne\mathbf{v}$:

$$\frac{\mathrm{d}\mathbf{J}}{\mathrm{d}t}+\gamma\mathbf{J}=\frac{Ne^2}{m}\mathbf{E} \tag{1.47}$$

We can readily interpret the drag constant γ: first, if the field **E** is turned off instantaneously, the homogeneous equation (1.47) has the solution $\mathbf{J}=\mathbf{J}_0\exp(-\gamma t)$, a current which decays with the relaxation time $\tau=1/\gamma$.

Second, for the steady state of constant or slowly changing fields and currents ($\mathrm{d}\mathbf{J}/\mathrm{d}t=0$), and invoking Ohm's law (1.30), it follows from (1.47) that

$$\sigma=\frac{Ne^2}{m}\tau \tag{1.48}$$

Intuitively, decay time of a transient current is proportional to the conductivity: The lower the resistance, the longer the stray current persists.

Following the same steps as in Sect. 1.4.1, we assume harmonic time dependence, $\exp(-i\omega t)$, in (1.47), and obtain the expression for the source term.

$$\mathbf{J}=\frac{\sigma}{1-i\omega\tau}\mathbf{E} \tag{1.49}$$

Substituting into wave equation (1.36), and postulating a plane-wave solution, we obtain the following condition for the wave number, the analog of (1.16) and (1.41)

$$k^2=\frac{\omega^2}{c^2}+\frac{i\omega\mu_0\sigma}{1-i\omega\tau} \tag{1.50}$$

Physical effects represented by the two terms in the dispersion equation (1.50) dominate at high and low frequencies, respectively. They are roughly equal at the so-called plasma[3] frequency:

[3] Plasma is a generic term for a gaseous (low-density) assembly of charged particles. Free electrons in a metal are well described as a plasma, which is held in place by the overall positive charge anchored to the lattice of metal atoms.

$$\omega_P = \left(\frac{\mu_0 \sigma \, c^2}{\tau} \right)^{1/2} \tag{1.51}$$

For $\omega \to \infty$, first term in (1.50) dominates, and yields $k = \omega/c$, the wave number of the freely propagating wave. For $\omega \to 0$, (1.50) yields

$$k^2 = i \omega \mu_0 \sigma. \tag{1.52}$$

By writing the complex wave number as $k = \beta + i\alpha$, and taking the square root in (1.52), we obtain the same value for the propagation and attenuation components:

$$\beta = \alpha = \left(\frac{\omega \mu_0 \sigma}{2} \right)^{1/2} \tag{1.53}$$

The inverse of the attenuation coefficient is the distance which the wave travels within the conductor before it is attenuated by the factor e^{-1}; this is the "skin depth" of metals, well known in microwave engineering:

$$\delta_S = \left(\frac{2}{\omega \mu_0 \sigma} \right)^{1/2} \tag{1.54}$$

To summarize, conductors are transparent to the electromagnetic wave above the plasma frequency. We see from (1.49) that the interior current $\mathbf{J}$ diminishes in size when the oscillation period becomes much shorter than the relaxation time: the electrons can not "keep up" with the field, owing to frictional drag.

Below the plasma frequency, conductors are opaque. The interior current rapidly dissipates the wave's energy as seen from the large imaginary component of the wave number (1.53), and the wave does not penetrate the conductor much beyond the skin depth.

We conclude this discussion on wave propagation in materials with a few words on reflection. Reflectance is defined as the fraction of energy bounced back from the material surface. It can be deduced from the field boundary conditions at the surface (tangential components of the fields must be continuous there), and after some algebra, the following simple formula can be obtained for the wave perpendicular to the surface:

$$R = \left| \frac{1 - n}{1 + n} \right|^2 \tag{1.55}$$

This formula is valid for the complex (dissipative) refraction index as well, and reflectance can be expressed in terms of the real and imaginary parts of the index:

$$R = \frac{(1 - \nu)^2 + \kappa^2}{(1 + \nu)^2 + \kappa^2} \tag{1.56}$$

For dielectrics, apart from resonances, ν is close to one, κ is close to zero, and the reflectance tends to be a few percent. For conductors below plasma frequency, how-

ever, ν and κ are similar in value, and larger than one: Reflectance has a value close to one. Conductors, therefore, reflect most of the radiated energy, and what enters the material is dissipated within the skin depth. In addition, plasma frequencies of metals lie almost without exception in the ultraviolet region (ca. 10^{15} Hz); therefore, for the purposes of radio communication all metals are very good reflectors.

1.5 Basics of Antennas

Having covered some fundamentals of radio physics, we now touch upon topics closer to engineering. Antennas straddle the boundary: They transfer the energy between moving charges (currents) and propagating electromagnetic waves. They differ from the usual lumped components and transmission lines, because their electric properties are not amenable to simplifying single-feature descriptions: They are neither pure capacitors, nor inductors, nor waveguides. Antennas come in a wide variety of designs, and describing them correctly usually requires invoking the formalism of electrodynamics at a fairly fundamental level. They are the least intuitive and most fascinating elements in the signal path of radio communication.

Directional antennas play a large role in selective use of the radio medium, such as long-distance point-to-point links, and in radiolocation. We review here the fundamentals of the electromagnetic theory of antennas, and devote some attention to the design principles used in the directional varieties.

Generally speaking, the radiating antenna is a current of known shape in space and behavior in time, typically confined to a metallic conductor. The current induces an electromagnetic wave, which propagates into infinite distance. Incidentally, the analysis is usually easier for a radiating antenna, but we see that the results are applicable to the receiving antenna as well (Sect. 10.2). To obtain the field around the antenna, one must solve (1.31)–(1.34) for the fields, given the radiator's current. We outline the procedure below, but the reader is referred to standard texts for a detailed discussion (Balanis 1997, Elliott 2003).

The crucial simplifying insight into antenna analysis lies in the fact that the curl of a vector field, any field, has no sources:

$$\nabla \cdot (\nabla \times \mathbf{V}) = 0 \tag{1.57}$$

This is somewhat plausible from the physical interpretations of curl and divergence, (1.5) and (1.6), but it is obvious algebraically, since the vector product is perpendicular to both of its components. We can therefore express the magnetic field $\mathbf{H}$ or $\mathbf{B}$ as the curl of a magnetic vector potential $\mathbf{A}$:

$$\mathbf{H} = \nabla \times \mathbf{A} \tag{1.58}$$

The field $\mathbf{A}$ has no physical meaning, but it serves to simplify the equations, much like expressing the static electric field as the gradient of the electrostatic potential. The equation describing the potential $\mathbf{A}$ turns out to be:

$$\nabla^2 \mathbf{A} - \frac{1}{c^2}\frac{\partial^2 \mathbf{A}}{\partial t^2} = -\frac{1}{c}\mathbf{J} \tag{1.59}$$

which is an inhomogeneous wave equation, with the driving source $\mathbf{J}$. Assuming harmonic time dependence of currents and fields, and skipping numerous mathematical steps, we state the solution of the Equation (1.59):

$$\mathbf{A}(\mathbf{x}) = \frac{1}{c}\int \mathbf{J}(\mathbf{x}')\frac{\exp(ik|\mathbf{x}-\mathbf{x}'|)}{|\mathbf{x}-\mathbf{x}'|}\,d^3x' \tag{1.60}$$

The integration in (1.60) is nominally over all of space, but is in fact limited to the conducting lines or surfaces, where $\mathbf{J} \neq 0$. Magnetic and electric fields follow from (1.58) and (1.2), respectively; the latter equation simplifies to

$$\nabla \times \mathbf{H} = -i\omega\varepsilon_0 \mathbf{E} \tag{1.61}$$

for harmonic time dependence.

Cumbersome appearance of the integral in (1.60) hints at extensive numerical computations required for theoretical design analysis of all but the very simple antennas. Such analysis is not our purpose here; instead, we use (1.60) to gain general insight into the shape of the antenna's field in two limiting regions (Jackson 1998).

In the regions in which $k\,|\mathbf{x}-\mathbf{x}'| \ll 1$, that is, when the dimensions of the antenna, and the distance from it, are shorter than the wavelength, the exponential factor in (1.60) can be approximated with the value of one. Equation (1.60) reduces to

$$\mathbf{A}(\mathbf{x}) = \frac{1}{c}\int \mathbf{J}(\mathbf{x}')\frac{d^3x'}{|\mathbf{x}-\mathbf{x}'|} \tag{1.62}$$

The field oscillates in time, in synchrony with the source current, and its value is determined by the overall shape of the antenna, but nothing in this expression indicates a propagating wave. This is the static or near-field region around the antenna.

At distances that are large compared with the overall dimensions of the antenna, it is reasonable to try to decouple the observation point's coordinates $\mathbf{x}$ from the integration coordinates $\mathbf{x}'$. We do this by using two (slightly inconsistent) approximations. Let us first define $r = |\mathbf{x}|$ as the distance of the observation point from an origin selected close to the antenna, such that $|\mathbf{x}| \gg |\mathbf{x}'|$. We now assert that

$$\left|\mathbf{x}-\mathbf{x}'\right| \approx r - \mathbf{n}\cdot\mathbf{x}' \tag{1.63}$$

in the exponential; here $\mathbf{n} = \mathbf{x}/|\mathbf{x}|$, the unit vector in the direction of the observation point. We also assert that $|\mathbf{x}-\mathbf{x}'| \approx r$ in the denominator (see Fig. 1.2), and with these simplifications (1.60) becomes

$$\mathbf{A}(\mathbf{x}) = \frac{1}{c}\frac{\exp(ikr)}{r}\int \mathbf{J}(\mathbf{x}')\exp(-ik\,\mathbf{n}\cdot\mathbf{x}')\,d^3x' \tag{1.64}$$

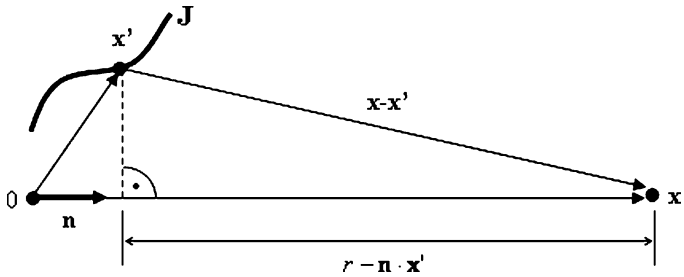

Fig. 1.2 Geometry of the vector-potential integration, and the far-field approximation

This potential represents an outgoing spherical wave, decreasing in amplitude as $1/r$. The amplitude of the wave in direction $\mathbf{n}$ (the amplitude's angular dependence) is determined by the integral over the shape of the antenna, which is not a function of r. Taking curls of $\mathbf{A}$ according to (1.58) and (1.61), yields the fields $\mathbf{H}$ and $\mathbf{E}$, whose amplitudes have the same asymptotic behavior as $\mathbf{A}$; this is because

$$\frac{\mathrm{d}^n}{\mathrm{d}r^n}\frac{\exp(ikr)}{r} \sim \frac{\exp(ikr)}{r} \quad r \to \infty, \quad n = 1, 2, 3 \ldots \tag{1.65}$$

Furthermore, $\mathbf{E}$ and $\mathbf{H}$ are perpendicular to the direction of r; the power flux vector $\mathbf{S}$ points in the radial direction, and its magnitude decreases as $1/r^2$, as expected. This is the radiation region, or far-field region, of the antenna, characterized by the radial outflow of energy, and by spherical waves whose angular distribution is independent of the distance r. Angular dependence of the power flux in the far field is usually referred to as the antenna's radiation pattern, or lobe pattern, as we shall see below.

1.5.1 Antenna Arrays

Directional antennas come in a multitude of designs, but there are only two physical principles that allow some control over the direction of the electromagnetic wave: Interference among the waves emitted by multiple sources, and reflection from conducting surfaces. We shall look at the interference first.

Let us calculate the current integral in (1.64), call it $\mathbf{A}(\mathbf{n})$, for a pair of identical radiating elements, $\mathbf{J}_1$ and $\mathbf{J}_2$, the latter being translated by a vector $\mathbf{d}$:

$$\mathbf{A}(\mathbf{n}) = \int [\mathbf{J}_1(\mathbf{x}) + \mathbf{J}_2(\mathbf{x}+\mathbf{d})]\, e^{-ik\mathbf{n}\cdot\mathbf{x}}\, \mathrm{d}^3x \tag{1.66}$$

We integrate separately over $\mathbf{J}_1$ and $\mathbf{J}_2$, and translate the origin of the integration variable in the second integral, so that $\mathbf{x}' = \mathbf{x} + \mathbf{d}$:

$$\mathbf{A}(\mathbf{n}) = \int \mathbf{J}_1(\mathbf{x}) e^{-ik\mathbf{n}\cdot\mathbf{x}}\, \mathrm{d}^3x + \int \mathbf{J}_2(\mathbf{x}') e^{-ik\mathbf{n}\cdot\mathbf{x}'}\, \mathrm{d}^3x' \cdot\, e^{ik\mathbf{n}\cdot\mathbf{d}} \tag{1.67}$$

Since $\mathbf{J}_1$ and $\mathbf{J}_2$ are identical in shape and merely displaced, it follows that $\mathbf{J}_1(\mathbf{x}) = \mathbf{J}_2(\mathbf{x}')$, and

$$\mathbf{A}(\mathbf{n}) = \int \mathbf{J}_1(\mathbf{x}) e^{-ik\mathbf{n}\cdot\mathbf{x}}\, \mathrm{d}^3x \cdot \left[1 + e^{ik\mathbf{n}\cdot\mathbf{d}}\right] \tag{1.68}$$

For a linear array of N identical elements, the above expression becomes

$$\mathbf{A}(\mathbf{n}) = \mathbf{A}_0(\mathbf{n}) \cdot \left[1 + e^{ik\mathbf{n}\cdot\mathbf{d}} + e^{2ik\mathbf{n}\cdot\mathbf{d}} + \cdots + e^{(N-1)ik\mathbf{n}\cdot\mathbf{d}}\right] \tag{1.69}$$

The integral $\mathbf{A}_0(\mathbf{n})$ in (1.69) carries the effects of the radiator's shape, common to all radiating elements. The bracketed term, called the array factor, expresses the phase differences, which are due to the spatial arrangement of the radiators.

With the radiator's shape factored out in $\mathbf{A}_0(\mathbf{n})$, the array factor is accurately depicted as an arrangement of isotropic point sources that are mutually in phase (see Fig. 1.3), and in which the phase difference between waves from adjacent sources is due only to the different distances from the sources:

$$\psi = kd \cos\theta \tag{1.70}$$

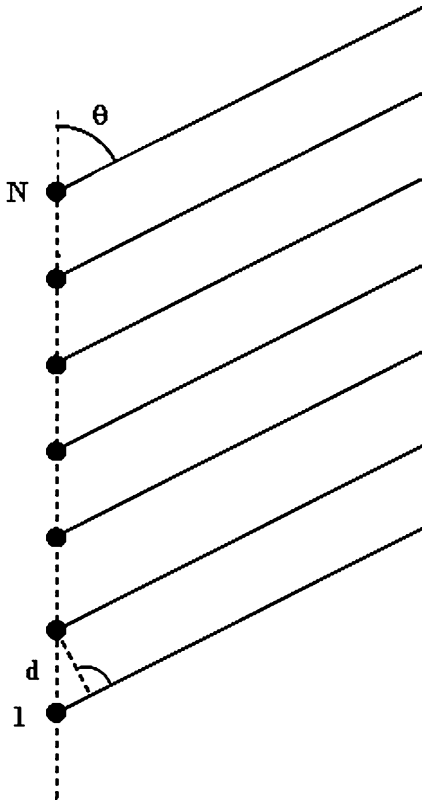

Fig. 1.3 Array factor as an array of identical point sources. Phase delays between adjacent sources equal $kd\cos\theta$

Waves from adjacent sources are mutually reinforced when the difference in the distance traveled equals a multiple of the wavelength, or

$$\psi = kd\cos\theta = 2\pi n, \quad n = 0, \pm 1, \pm 2 \ldots$$

The array factor of a linear array can be summed explicitly, and after moving the coordinate origin to the center point, expressed in the familiar form:

$$\text{array factor} = \frac{\sin\left(\frac{N}{2}\psi\right)}{\sin\left(\frac{1}{2}\psi\right)} \tag{1.71}$$

This factor is shown as a function of θ in Fig. 1.4, and of course, it has the main maximum for $\psi = 2\pi \cdot 0$. Its first null, defining the main lobe, appears at the angle

$$\theta_0 = \sin^{-1}\frac{2\pi}{Ndk} \tag{1.72}$$

measured from the direction of the main lobe's maximum.

Another useful description of the antenna's radiation pattern is the -3 dB half-width. This is the direction in which the radiated power [square of the expression in (1.71)] drops to one half, or -3 dB, of the power in the main direction. Measured from the main lobe's direction, this angle is given by

$$\theta_{3dB} = \sin^{-1}\frac{2.781}{Ndk} \tag{1.73}$$

for our linear array of point sources (see e.g., Balanis 1997).

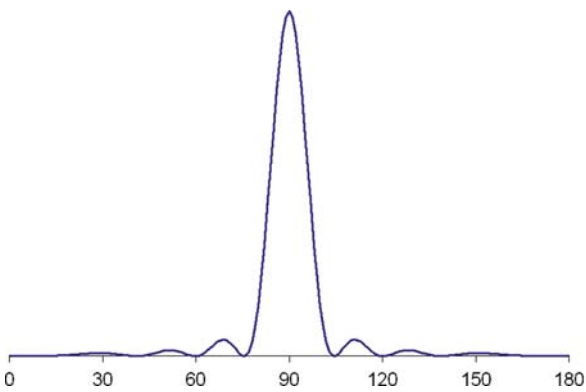

Fig. 1.4 Radiated power of an array of point sources, as a function of angle θ. The maximum is in the direction perpendicular to the array, where the waves from all sources are in phase. The array has 25 elements, at distances such that $kd = 1$

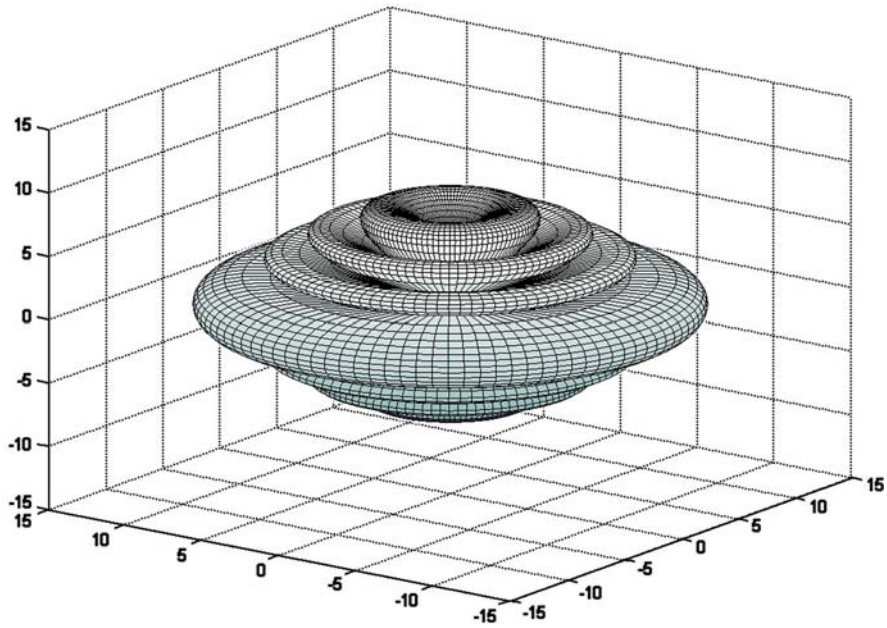

Fig. 1.5 Spatial power radiation of a linear array of isotropic point sources, on the logarithmic scale. The array is placed vertically in the center of the pattern and has the same geometry as that in Fig. 1.4

The reader will notice that the array factor, as discussed so far, is formally identical to the optical diffraction grating. Both consist of point sources, whose phase differences are due to the spatial arrangements, and which produce wave reinforcement in certain directions and cancellations in others (see Fig. 1.5). However, arrays of radio antennas need not be linear: Circular and other arrangements are also used. More importantly, controlled phase difference β can be introduced between elements via the feeding circuitry: In a linear array, we can modify the geometrically defined phase difference of (1.70), to make it equal $\psi = kd\cos\theta + \beta$. This phase difference determines the direction of constructive wave interference, and is the basis of beam-steering techniques for antenna arrays.

An interesting variation on the antenna array is the helical antenna. We refrain from the full mathematical treatment of the helix (see Elliott 2003), but, in a rough approximation, one can regard the corresponding current elements on each turn as the elements of an array (see Fig. 1.6). The signal source drives a propagating current wave in the helix – we assume that the helix is long enough to radiate away all the power, and that there is no significant reflection of the current wave back from the far end. With proper choice of the step and diameter of the helix, the current's phase difference β between adjoining turns can be made such that $\cos\theta = 1$ at the main diffraction maximum. Waves emitted by the adjacent current elements then reinforce each other in the direction of the axis of the helix; a reflective ground plane is placed at the feed end, and the antenna exhibits a nicely formed main lobe in the end-fire direction.

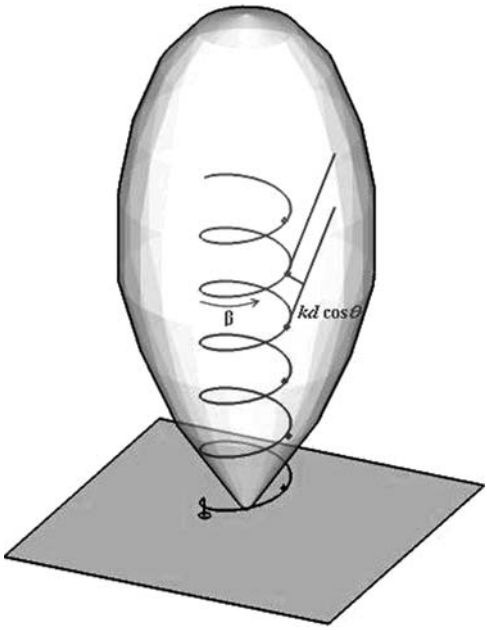

Fig. 1.6 Helical antenna, showing the geometric phase delay $k\,d\,\cos\,\theta$ and the current phase delay β

One final note on the topics in this section: A truly isotropic source of electromagnetic waves cannot exist. If it did, its power flux on the surface of a large sphere (in the far field) would be strictly radial, and of constant magnitude on the whole surface. Because of Fig. 1.1, each field would also be constant in magnitude, and tangential everywhere on the sphere. This is topologically equivalent to combing the entire surface of a furry sphere: it cannot be done without creating swirls or breaks. But swirls and breaks are discontinuities in the field (infinite spatial derivatives) and are not physically possible; therefore, an isotropic source is not possible either.

1.5.2 Reflector Antennas

Reflector antennas are readily understood in terms of geometrical optics: A conductive surface acts as a mirror, and reflects the "rays" emanating from a wide-directional antenna in the desired, narrower direction. The paragon of all reflector antennas is of course the paraboloid dish, since we know from elementary geometry that all rays emanating from the focus of the parabola are reflected in the same direction. Conversely, a plane wave traveling along the direction of the axis converges upon reflection on a single point, the focus. The "beam" of such an idealized antenna is infinitely narrow.

Calculating the radiation pattern of an actual reflector antenna is quite challenging, because it depends on the pattern of the feed antenna, as well as on the shape and size of the reflecting mirror(s). For the purposes of radiolocation, reflectors are more interesting as image forming receivers, and a paraboloid reflector is indeed analogous to a telescope, an instrument that maps the delocalized, space-filling plane wave of light into an illuminated dot in the focal plane. Direction of the wave (relative to the instrument) can then be deduced from the position of that dot.

Departing from geometrical optics, we can envision the reflection of the plane wave as re-emission of spherical waves from every point on the surface of the mirror. Constructive interference between these re-emissions forms a wave that converges on the focus. Because the wavelength is not infinitely small, relative to the size of the mirror, we expect to see a diffraction pattern in the focal plane, rather than a pure dot of illumination. In fact, diffraction patterns caused by apertures are well known (Fowles 1989, Born and Wolf 2006), and the illumination pattern caused by the circular opening of radius R is plotted in Fig. 1.7. The abscissa in this plot is the angle away from the axis, represented in the form $kR \sin \theta$. As we should expect, this pattern is similar to the beam pattern of the array antenna, since both have their origins in the same physics.

The angular diameter of the central illuminated area is given by

$$\theta_D = 2.44 \frac{\lambda}{D} \tag{1.74}$$

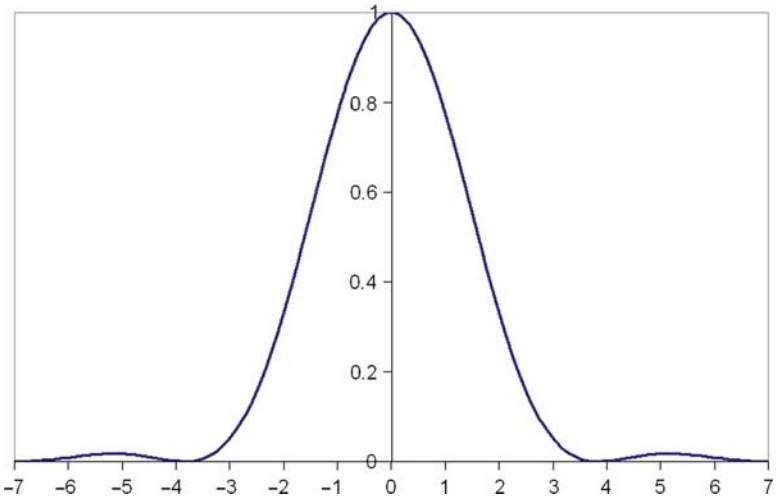

Fig. 1.7 Illumination pattern caused by the circular aperture. This is the plot of the function $[2J_1(\rho)/\rho]^2$ where J_1 is the Bessel function of the first kind, and $\rho = kR \sin \theta$

where D is the diameter of the aperture (mirror), and λ is the wavelength. This is the image size, or the angle which the paraboloid dish will sweep while detecting one point source, and is obviously a measure of the resolution of the antenna (resolution is usually defined as the minimum separation between two still discernible sources, and is half of θ_D).

We see that the resolution is better for larger dishes and for shorter wavelengths. To put things into perspective, let us compare a 1.5-m dish antenna at the 2.4 GHz microwave frequency ($\lambda = 12$ cm), with a modest amateur telescope with an aperture of 10 cm, in visible light ($\lambda = 5.5 \times 10^{-7}$ m). The angular image size turns out to be ca. 11° for the microwave antenna, and 0.0008° for the optical telescope!

1.5.3 Patch Antennas

The last antenna design that we discuss is the patch (or microstrip) antenna. This type of antenna consists of an area of conductor, placed above a larger ground plane, and separated from it by a dielectric layer. It is obvious that patch antennas can be fabricated using the printed circuit board technology, and they are often integrated into the circuit boards of inexpensive wireless electronics. Being flat, they are frequently used where space is at a premium, such as in avionics applications. Their main drawback is that they are relatively inefficient radiators, compared to other types of antennas, and that they have a narrow bandwidth around their resonant frequency.

The first question is why patch antennas work at all. Two metal plates separated by a thin dielectric layer form a capacitor, and why should a capacitor radiate? The answer is that the radiation from a patch antenna is due entirely to fringe effects, the noncontainment of the field at the edges of the patch. That explains the inherent inefficiency: Most of the electromagnetic energy is trapped in standing waves within the patch, and only a small fraction radiates out.

Nevertheless, practical advantage of the compact design is considerable, and patch antennas enjoy great popularity in microwave communication. To explain how these antennas work, let us assume that the patch is rectangular, and therefore forms a rectangular resonant cavity with the ground plane (see Fig. 1.8). Without going too deeply into the theory of resonators, we can assert on intuitive grounds that the main mode of oscillation (the one with lowest frequency) forms a standing wave along the largest dimension L, and that the electric field is directed from one conductor to the other. Figure 1.8 shows the electric field in the slots (open sides) between the two conductors. Fields in the shorter slots (W) are out of phase, but since the directions of radiation are opposite, the two slots act like two magnetic source currents in phase, and reinforce each other's emissions in the direction perpendicular to the patch. Radiation from the longer slots cancels out. The rectangular patch is, in effect, a two-element array with a mirror plane, the geometry of its radiators held in place by the standing wave inside a resonant cavity.

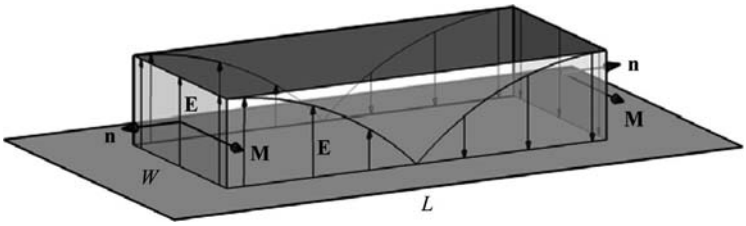

Fig. 1.8 Rectangular patch antenna. Radiating slots can be described by "magnetic source currents," which are defined as $\mathbf{M} = \mathbf{E} \times \mathbf{n}$ and are parallel and in phase for both slots W. The antenna has a broad lobe facing upward. Due to the boundary conditions, electric field forms a cosine wave, and $\mathbf{M}$ cancels out along the slots L. The magnetic field forms a sine wave in the L direction and is parallel to $\mathbf{M}$

Similarly, a circular patch can be described as a circular radiating slot. It is particularly interesting that two oscillation modes of the circular patch, at the same frequency and spatially orthogonal to each other, can be excited with two antenna feeds that are out of phase by a quarter-period in time, thus creating circular polarization.

Quantitative analysis of patch antennas requires description of the fringing effects either by semi-empirical formulas or by complete electromagnetic field simulations. We refer the interested reader to (Balanis 1997) for a detailed discussion, including a semi-empirical analysis of the rectangular patch.

References

Balanis, C.A. Antenna Theory, Analysis and Design, 2nd edition, Wiley, New York (1997)
Born, M., Wolf, E. Principles of Optics, 7th edition, Cambridge University Press, New York (2006)
Byron, F.W., Fuller, R.W. Mathematics of Classical and Quantum Physics, Dover, New York (1992)
Elliott, R.S. Antenna Theory and Design, Revised Edition, Wiley, New York (2003)
Fowles, G.R. Introduction to Modern Optics, 2nd edition, Dover, New York (1989)
Jackson, J.D. Classical Electrodynamics, 3rd edition, Wiley, New York (1998)
Ma, M.T. Theory and Application of Antenna Arrays, Wiley, New York (1974)
Morse, P.M., Feshbach, H. Methods of Theoretical Physics, McGraw-Hill, New York (1953)
Pozar, D.M. Microwave Engineering, Wiley, New York (1998)
Russer, P. Electromagnetics, Microwave Circuit and Antenna Design for Communications Engineering, 2nd edition, Artech, Boston (2006)

Chapter 2
Radiolocation with Multiple Directional Antennas

2.1 Introduction

All radiolocation methods are concerned with the same problem: determining the direction of the Poynting vector of a radio wave at the point of observation. In this chapter, we will delve into our chosen method of amplitude-based radiolocation, but let us first draw some useful analogies with optical detectors and optical image formation.

A radiolocation method that relies on the amplitude, or power measurements, must involve at least one directional detector. Sensitivity of the detector has a maximum in a known direction, relative to the instrument, and if a maximum reading is obtained, the instrument's maximum must be aligned with the direction of the wave. This implies that we must measure the power in a number of directions in the vicinity of the putative maximum.

In the approximation of geometrical optics, where the diffraction phenomena are neglected, formation of an image can be viewed as just such a many-directional measurement of amplitude. For every direction within its field of vision, a telescope, or even a pinhole camera, has a preferentially sensitive spot in its image plane. When a wave arrives from that direction, it causes an infinitely narrow maximum readout at the right place in the image plane. We know that it is a maximum, because the adjacent directions are also measured, and yield zero intensity: what we see (what the film or the CCD array sees) is a bright spot, the image of a distant point source.

As we have discussed in Sects. 1.5.1 to 1.5.3, geometrical optics approximation does not hold in radio communication by far. The wavelength is typically not small relative to the antenna size, and directional antennas have prominent wide lobes, due to wave diffraction. Since a simple one-feed radio antenna is intrinsically a single sensor, finite width of its sensing field is of some practical benefit for detecting the presence of a source in the first place. Multiple measurements that are required to find a maximum are traditionally accomplished by swinging the antenna around until the strongest signal is received. Some radiolocation instruments operate by aligning their antenna's null direction with the source: the principle is the same, and we need not discuss it further.

D. Antolovic, *Radiolocation in Ubiquitous Wireless Communication*,
DOI 10.1007/978-1-4419-1632-7_2,

Since the speed of such methods is limited by the mechanical motion of the antenna, they must rely on the source transmitting long enough, and with constant amplitude, to complete the radiolocation. We would, of course, prefer to do all the directional measurements at once, i.e., form an image of the source. We know that the image will be broadened by diffraction, and that it must be detected by a spatial array of sensors, to establish the location of the maximum intensity.

There are two plausible options for image formation. The first alternative is an aperture instrument, such as a reflector telescope; reflecting mirrors are widely used in rotating radar antennas, but we contemplate here forming an image with a stationary instrument. Referring back to the example in Sect. 1.5.2, our 1.5-m dish at 2.4 GHz forms an image of a distant point source, that has an angular size of 11°. At the focal distance of, let us say 1 m, that is a 19 cm image diameter. The image area will have to be covered with sensors whose sizes and mutual distances are small relative to the wavelength. This means that the sensors (miniature antennas) will have very low gain, and will have to be supplemented with low-noise amplifiers.

Furthermore, aperture instruments tend to have a narrow vision field, and wide-angle coverage is usually desired in radiolocation. It is difficult to construct wide-angle aperture instruments, especially in the reflector design, where the image plane obstructs the aperture to begin with. A radio lens could be used instead since the refractor design can have a wider field than the reflector, but radio lenses are bulky and expensive.

The second alternative, the one that we pursue in our implementation, performs multiple simultaneous directional measurements with an array of directional sensors, and is somewhat akin to the compound eyes of insects. There is no optical image, in the usual sense of a converging front of mutually reinforcing waves; instead, the direction of the incoming wave is reconstructed (in circuitry or in code) from the slight differences in the way in which it affects the adjacent sensors.

Figure 2.1 shows the principle of operation of the compound eye, in the framework of geometrical optics. The parallel rays, which stand for a plane wave, impinge upon a curved array of eyelets, causing the adjacent ones to produce simultaneous outputs, which either differ in intensity or are different discrete signals, as shown in the figure – the eyelets have some directional sensitivity. The schematic "cross-wiring" behind the array combines the related outputs and reconstructs the source's direction.

One may ask whether the cross-wiring is even necessary: why not simply have each sensor give a yes/no response to a wave in its own direction? Non-overlapping sensors would lead to blind spots, and sensors without angular sensitivity could not discern the directions that lie between them, even if their fields overlapped. Overlapping sensors with angular sensitivity can be combined to confirm and reinforce each other's findings, as indicated in Fig. 2.1; this improves the accuracy and alleviates aliasing, by better estimating the direction of a wave that is not directly aligned with *any* sensor. This is the level of sensory complexity that we have implemented in radiolocation.

Compound eyes have no limitation on the extent of their field of vision, and can in principle cover all 4π steradians – in fact, the enormous compound eyes of

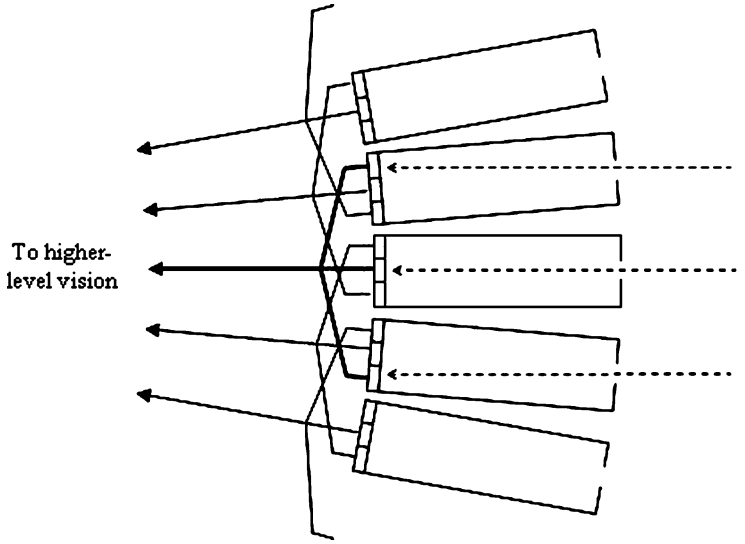

Fig. 2.1 Operation of the compound eye. Direction-sensitive outputs from adjacent sensors are combined to construct an image of the source of the wave

some dragonflies come close to doing just that. Furthermore, unlike the reflectors and refractors, which invariably suffer distortions in directions away from their optical axis, accuracy of the compound eyes is uniform throughout the vision field. Analogously, a suitable multi-element antenna can perform radiolocation equally accurately in all directions, without ever having to move.

However, individual sensors of this compound antenna must be directional and, unlike the sensors in the image plane of the telescope, must therefore be comparable in size to the wavelength. This places a practical limit on the number of elements, and limits the density of the coverage of directions. Reconstructing the direction of the wave from the sensors' inputs (the "wiring") is a computational problem, which we address in this chapter.

2.2 Rotated Lobe Theorem

Now, that we have decided upon the array of directional sensors, or the "compound eye" as our imaging architecture, we see the importance of the shape of the sensor's reception pattern. The lobe of an individual antenna must be wide enough to overlap with the neighbors' lobes, and furthermore, the lobe itself represents a form of directional sensitivity because the signal diminishes as the direction of the incoming wave deviates from the antenna's main direction. In order to show how to combine these sensory outputs and reconstruct the wave's direction, we first state one or two mathematical results, which are as useful as they are elementary.

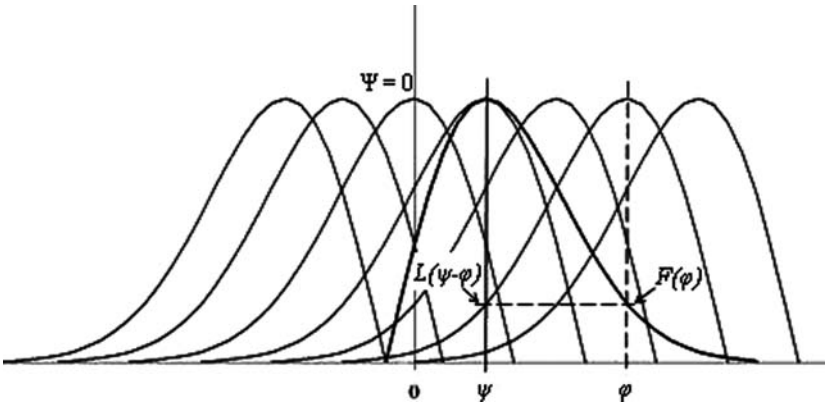

Fig. 2.2 Family of translated functions $L(\varphi - \Psi)$. Function $F(\varphi)$ represents values of L's at ψ, translated back to their centerpoints

Suppose that we have a family of arbitrary functions of an independent variable φ, such that the functions are obtained from one another by translation, that is, they all have the form $L(\varphi - \Psi)$, where Ψ is the parameter of translation (Fig. 2.2). For the sake of clarity, we depict the functions as having maxima at $\varphi = \Psi$, so that we can think of them as "centered" at Ψ, but this is not necessary: the function's shape is arbitrary, and so is the choice of origin, Ψ.

We now choose a particular value of the independent variable, we can call it ψ, and we define a function $F(\varphi)$, *which for any* φ *equals the value at* ψ *of an* L *centered at* φ. In symbols:

$$F(\varphi) = L(\psi - \varphi) \tag{2.1}$$

i.e., the function F is identical to a function L centered at the point ψ, and reflected with respect to that point. Figure 2.2 shows the family of L's, the function F, and the points representing left and right sides of (2.1).

We apply this to our array of sensors, which we here assume to be in the form of a planar ring of directional antennas. Variable φ becomes the angle around the ring, with appropriate periodic conditions. Antenna lobes are identical in shape, and are centered at discrete angles Ψ_i, i.e., they have the form $L(\varphi - \Psi_i)$. The angle of the incoming wave is ψ.

Signal strengths, which the wave induces on the antennas, are equal to $L(\psi - \Psi_i)$, by the definition of the antenna reception pattern. If we associate these values with directions Ψ_i, we see that they lie on a function, *which at any* Ψ_i *equals the value at* ψ *of an* L *centered at* Ψ_i. Because of (2.1), they lie on $L(\psi - \varphi)$, and we can formulate the main result of this section as follows:

The pattern of discrete signal strengths, induced in the identical antennas by a plane wave, lies on a lobe that is centered at the direction of the wave, and is a mirror image of the antenna lobe. ■

Lobes of most antennas are even around their central direction, and the mirror-image detail will not be very significant. Such a configuration is shown

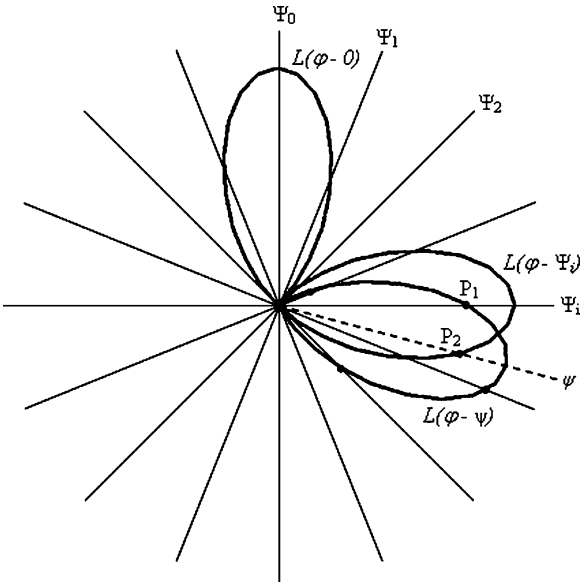

Fig. 2.3 Rotated-lobe theorem for a planar ring of even lobes, represented in polar coordinates

in Fig. 2.3, and we see immediately, from the symmetry of the problem that the value of a lobe centered at wave's direction ψ, taken in the direction Ψ_i of the i-th antenna, equals the signal of the wave coming from direction ψ, as detected on the i-th antenna.

These conclusions are also valid for spatial arrangements of axially symmetrical lobes in arbitrary directions. This is true because the directions ψ and Ψ_i in Fig. 2.3 can be any two directions in space, regardless of the directions of the other antennas: a wave induces signals in antennas that, when plotted on the antennas' directions, lie on a lobe of the same shape, pointing toward the source of the wave. It is immediately obvious that this conclusion does *not* hold if the lobe shape is not axially symmetrical.

2.3 Reconstruction of the Wave's Direction

We see from the previous section how the inputs from adjacent sensors combine to form the image of the source: the antennas' signal strengths delineate a known shape, an antenna lobe rotated toward the source. Obviously, the angular sensitivity and overlap of the antennas are crucial: they yield a source direction that is determined, redundantly and robustly, by the signals of several (all) antennas that contribute points along the rotated lobe.

Next required step is conceptually straightforward: we define a measure of error between the set of sensory inputs and the antennas' lobe function, and vary the

rotation angle until we obtain an optimal fit. We should point out, however, that we assumed in Sect. 2.2 that the function F (the fit of sensory inputs) is normalized to the same scale as the antenna lobes.[1] This simplified the formulas, but of course the shape delineated by the sensors' inputs changes in scale with the power carried by the wave, and can be treated only as proportional to the antenna lobe. In practice, the optimization must involve at least two variational parameters: the rotation angle and the scale.

It is clear that the direction finding relies on numerical computation of some intensity. Since there are stringent time requirements that need to be observed, implementation of the algorithm will be helped by an understanding of its underlying mathematical features. In the next two sections, we investigate the effects that the error function and the lobe shape have on this computation.

2.3.1 Variational Error

We will consider here the rotational and scaling variations, and we will assume that the shapes of the antenna (template) and input lobes are identical save for rotation and scaling. This is the ideal limit, in which inputs from an infinitely dense array of sensors fully delineate a shape identical to that of an antenna lobe. Real inputs are somewhat sparse – that is why pattern matching has to be performed in the first place – but it is nevertheless instructive to investigate how the match of two identical lobes behaves under variations.

Let us establish some conventions and nomenclature. The original (unvaried) lobe shape is defined by a function $L(\varphi)$ of the angle in polar coordinates (we can safely keep the problem two-dimensional). The variation is represented by small increments in angle and scale, ρ and σ, and the varied lobe has the form $(1 + \sigma)L(\varphi - \rho)$.

We represent the *variational error* as an integral over the angle:

$$\Delta(\rho, \sigma) = \int \mathrm{d}\varphi \; \delta(\rho, \sigma, \varphi) \tag{2.2}$$

and we call the function $\delta(\rho, \sigma, \varphi)$ the *error function*. The error function is the contribution to the error at each point along the lobe (all $\mathrm{d}\varphi$ integrals in this chapter are taken over the full circle).

We are further interested in the second-order behavior of the error around the minimum, and we define the *error curvatures* as the second derivatives of the error with respect to the variational parameters:

$$\Delta''(\rho) = \left[\frac{\partial^2}{\partial \rho^2} \Delta(\rho, \sigma) \right]_{\text{at } \rho, \sigma = 0} \tag{2.3}$$

[1] The convention for representing antenna patterns is that the maximum in the main direction equals one on the linear scale, or zero on the decibel scale.

$$\Delta''(\sigma) = \left[\frac{\partial^2}{\partial\sigma^2}\Delta(\rho,\sigma)\right]_{\text{at } \rho,\sigma=0} \tag{2.4}$$

We also make use of the *curvature integrand*, which is the integrand of the integral form of Δ'':

$$\Delta'' = \int \mathrm{d}\varphi\ \delta''(0,0,\varphi) \tag{2.5}$$

There are many ways to define the variational error. Intuitively, the error must increase in magnitude with the absolute value of variational parameters, and it must have a minimum for $\rho, \sigma = 0$. In order for the numerics of iterative optimization to work well, the variational error should also be continuous, and have continuous first and second derivatives, at the minimum point. It is also plausible (although not strictly necessary) that $\Delta(\rho,\sigma)$ should be an even function of ρ because for identical lobe shapes, the same error results from varying the angle in either direction (see Fig. 2.4).

Our variational error will be calculated by numerical integration of a suitable error function, along all points on a lobe; furthermore, it will be recalculated for every iteration step. Obviously, it is critical that we select a simple and fast error function, even at the expense of some physical plausibility: wrong choice may result in an unacceptably slow calculation.

The intuitive first choice of error, the area between displaced lobes, is actually a poor error measure. Since the area is an oriented quantity, it will be either identically zero (for even lobes), or have a zero (a cross-over point) rather than a minimum, at $\rho = 0$.

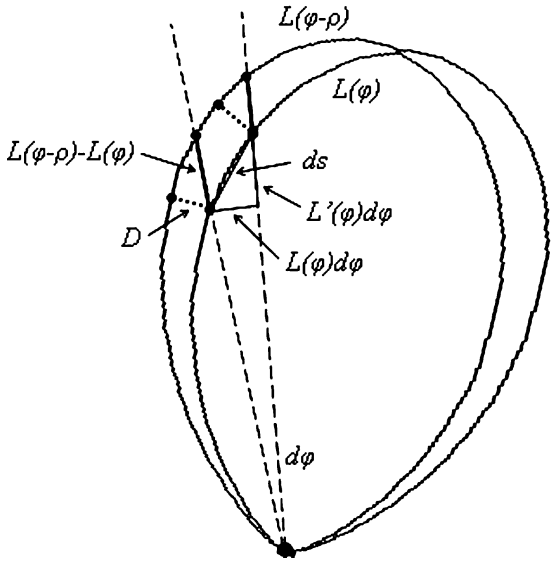

Fig. 2.4 Line element of the curve $[\varphi, L(\varphi)]$ in polar coordinates. Error function $[L(\varphi) - L(\varphi - \rho)]^2$ is integrated over the shape of the lobe

An appealing error measure is the square of the distance between curves, labeled D in Fig. 2.4. It is a slowly changing function along the lobe, and its integral over the length of the lobe curve provides a variational error that is well balanced over all parts of the lobe. The problem with this error measure is that it involves calculating distances along directions perpendicular to the lobe curve, and therefore requires long divisions.

Instead, we will use as our error function the square of the difference of original and varied lobes

$$\delta(\rho, \sigma, \varphi) = [L(\varphi) - (1 + \sigma)L(\varphi - \rho)]^2 \tag{2.6}$$

and integrate it over the curve defined by the lobe shape $L(\varphi)$:

$$\Delta(\rho, \sigma) = \int \mathrm{d}s\, [L(\varphi) - (1 + \sigma)L(\varphi - \rho)]^2 \tag{2.7}$$

Here, ds is the line element of the curve, and as shown in Fig. 2.4,

$$\mathrm{d}s^2 = (L(\varphi)\, d\varphi)^2 + \left(L'(\varphi)\, d\varphi\right)^2 \tag{2.8}$$

The terms on the right-hand side represent respectively the angular and radial component of the line element ds. The variational error becomes:

$$\Delta(\rho, \sigma) = \int \mathrm{d}\varphi\ [L(\varphi) - (1 + \sigma)L(\varphi - \rho)]^2 \sqrt{(L(\varphi))^2 + (L'(\varphi))^2} \tag{2.9}$$

We are furthermore interested in the error curvatures, starting with $\partial^2\Delta\big/\partial\rho^2$ at $\rho, \sigma = 0$. By expanding $L(\varphi - \rho)$ in a Taylor series in ρ, we obtain the expression

$$L(\varphi) - L(\varphi - \rho) \approx L'(\varphi) \cdot \rho \tag{2.10}$$

and the variational error in (2.9) can be written as

$$\Delta(\rho) \approx \rho^2 \cdot \int \mathrm{d}\varphi \left[L'(\varphi)\right]^2 \sqrt{(L(\varphi))^2 + (L'(\varphi))^2} \tag{2.11}$$

This is the quadratic term in the Taylor expansion of $\Delta(\rho)$ in powers of ρ, and it follows that

$$\Delta''(\rho)\big|_{\rho,\sigma=0} = 2 \int \mathrm{d}s \left[L'(\varphi)\right]^2 > 0 \tag{2.12}$$

We obtain $\partial^2\Delta\big/\partial\sigma^2$ by straightforward differentiation of (2.7) with respect to σ:

$$\Delta''(\sigma)\big|_{\rho,\sigma=0} = 2 \int \mathrm{d}s\, [L(\varphi)]^2 > 0 \tag{2.13}$$

And likewise the mixed second derivative:

$$\Delta''(\rho,\sigma)\big|_{\rho,\sigma=0} = -2\int \mathrm{d}s\; L(\varphi)L'(\varphi) \tag{2.14}$$

For even (symmetric) lobes, $\Delta''(\rho,\sigma) \equiv 0$, and the Hessian matrix of Δ is diagonal and positive definite at $\rho,\sigma = 0$, because of (2.12) and (2.13):

$$\det\begin{bmatrix} \Delta''(\rho) & \Delta''(\rho,\sigma) \\ \Delta''(\rho,\sigma) & \Delta''(\sigma) \end{bmatrix} > 0 \tag{2.15}$$

We can now summarize the insights that we have gained so far into the topography of the error minimum: In the vicinity of the minimum, the surface of the variational error is an elliptical paraboloid. Its vertical cross sections are parabolas, and any of its horizontal cross sections is an ellipse whose axes coincide with the coordinates ρ and σ for symmetric lobes. Parabolic cross sections in the directions of the coordinate axes are described by the error curvatures.

Integrals in (2.9) and (2.11) account properly for the geometric relationship between infinitesimal elements $\mathrm{d}\varphi$ and $\mathrm{d}s$, but the presence of the square root will slow down the numerical integration considerably. Notice that we can neglect one or the other term under the square root, or even dispense with the line integral altogether and pretend that φ is a Cartesian variable: in each case, we will still have a measure of variational error with continuous derivatives and a zero-valued minimum in the right place, at $\rho,\sigma = 0$.

These simplifications lead to the following formulas for error curvatures in the angular, radial and Cartesian approximations:

$$\Delta''_A(\rho) = 2\int \mathrm{d}\varphi \left[L'(\varphi)\right]^2 L(\varphi) \tag{2.16}$$

$$\Delta''_R(\rho) = 2\int \mathrm{d}\varphi \left[L'(\varphi)\right]^3 \tag{2.17}$$

$$\Delta''_C(\rho) = 2\int \mathrm{d}\varphi \left[L'(\varphi)\right]^2 \tag{2.18}$$

$$\Delta''_A(\sigma) = 2\int \mathrm{d}\varphi \left[L(\varphi)\right]^3 \tag{2.19}$$

$$\Delta''_R(\sigma) = 2\int \mathrm{d}\varphi \left[L(\varphi)\right]^2 L'(\varphi) \tag{2.20}$$

$$\Delta''_C(\sigma) = 2\int \mathrm{d}\varphi \left[L(\varphi)\right]^2 \tag{2.21}$$

Since the lobe is plotted as radial distance, Δ and Δ'' have the nominal dimension of cubed length ℓ^3. Cartesian approximation arbitrarily replaces integration over the line element with integration over the angle, and therefore has the dimension ℓ^2. Comparing quantities of different dimensions is not physically meaningful,

but we are interested only in the accuracy of computational approximations, and the physical dimension of the error is not significant [see (2.23) and (2.24)].

2.3.2 Numerical Significance of the Lobe's Shape

We have seen in Sect. 1.5 that the radiation/reception patterns of antennas are determined by such wave diffraction and reflection as can be reasonably implemented in an antenna design. In this section we investigate the way the shape of the antenna's lobe affects the speed and accuracy of the computation that yields the source's direction. To do this, we look at the effects that the lobe's shape has on the topography of the variational error around its minimum.

In order to perform the numerical analyses of this and the following sections, we use a convenient model lobe pattern, that of the linear array of point sources from Sect. 1.5.1. The pattern of this antenna array is described by the analytical formulas (1.70) and (1.71); we use only the plane cross section of the pattern, and suppress for simplicity all but the main lobe on one side of the array. We can think of this as an array with a mirror behind it, and with somewhat unnaturally weak side lobes. We can readily adjust the width of the lobe by changing the array parameters. This model pattern is in fact fairly similar to the pattern of the helical antenna used in our experimental prototype; this similarity should not surprise us too much because, as we have mentioned in Sect. 1.5.1, the helix is really an elegant variant of the linear array, and the physics behind its directional pattern is essentially the same.

2.3.2.1 Rotational Variation

Figure 2.5 shows the plot of a 150°-wide lobe, as a function of φ, as well as the plots of the curvature integrand $\delta''(\rho)$ and its approximations. It is obvious, and physically plausible, that the contribution to $\Delta''(\rho)$ comes mainly from the steep sides of the lobe, since they are perpendicular to the rotation, and sweep a larger variational error than the flat top; this is reflected in $[L'(\varphi)]^2$, the main term in $\delta''(\rho)$.

We see from Fig. 2.5 that the angular component δ''_A is small, when compared with the radial component δ''_R, and shifted toward the center of the lobe, relative to δ''; it is obvious from Fig. 2.4 that it should be so, and it follows that the radial form is a good approximation of the variational error $\Delta(\rho)$. These trends can only be more pronounced in narrower lobes.

Since the approximate forms δ''_C and δ''_R differ only in one power of $L'(\varphi)$ [see (2.17) and (2.18)], they peak at the same angle, but the Cartesian form of the error curvature is somewhat smaller. Nevertheless, this indicates that $\Delta_C(\rho)$ is also a plausible form of the variational error, which has the advantage of being very simple computationally. We will return to this point later.

Figure 2.6 (Antolovic 2009) shows the dependence of the rotational error curvature Δ'' on the width of the lobe. Apart from the expected increase in all variants

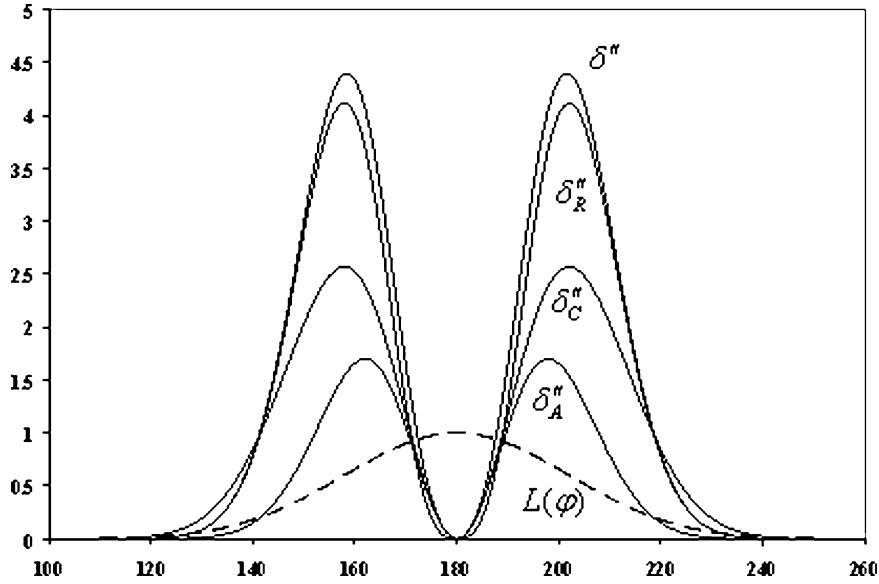

Fig. 2.5 Plot of the antenna lobe, 150° wide null-to-null (25° − 3dB half-width), and of the rotational curvature integrands, as functions of the angle. Angular approximation δ''_A shifts the emphasis toward the center of the lobe; radial approximation δ''_R follows the exact value δ'' closely

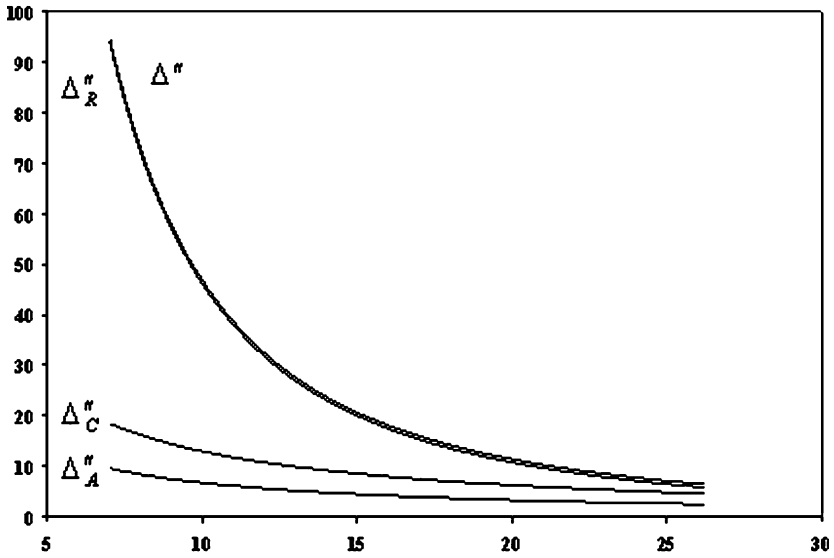

Fig. 2.6 Plots of the rotational error curvature, as functions of the −3dB lobe width in degrees. Radial approximation Δ''_R follows Δ'' closely for all lobe widths (© IEEE 2009, reproduced with permission)

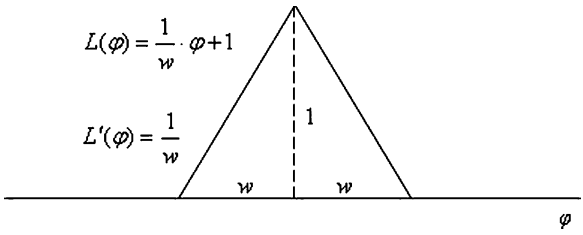

Fig. 2.7 Triangle approximation of the narrow lobe

of Δ'' for the narrower lobe, we see that Δ'' and Δ''_R remain consistently very close together. The variant Δ''_A remains small, although it reaches a relative magnitude of $0.4\,\Delta''$ as the lobe width approaches 180°. It is also apparent that the asymptotic behavior of Δ'' and Δ''_R differs from that of Δ''_C and Δ''_A, for decreasing lobe width. This difference is due to different powers of $L'(\varphi)$ in δ'', and we can easily explain it with a simple analysis: if we approximate each side of a narrow lobe with linear functions on the intervals $[-w, 0]$ and $[0, w]$ where w is the lobe's half-width, as shown in Fig. 2.7, and integrate over the intervals, it turns out that

$$\begin{aligned} \Delta''_R &\approx w^{-2} \\ \Delta''_C &\approx w^{-1} \\ \Delta''_A &\approx w^{-1} \end{aligned} \tag{2.22}$$

We will leave it to the reader to verify these results.

2.3.2.2 Scaling Variation

The picture is somewhat more problematic for the scaling variation. The curvature integrand, shown in Fig. 2.8, has a maximum around the transition point between the lobe's steep side and the flat top – this is where the integrand factors in (2.13) are both sizable. None of the approximations are very close, though, and δ''_R breaks down completely at the center, because the radial component of ds vanishes there.

Figure 2.9 shows that the scaling error curvature is much smaller than the rotational one, and that it grows larger for wider lobes, as expected. Cartesian and angular approximations also improve for wider lobes. Naturally, the optimization involves both ρ and σ, and we can have only one form of variational error. Fortunately, the radiolocation application has no requirements for the accuracy of the scaling, and the errors in the two parameters are not strongly linked. We will discuss this point in the next section.

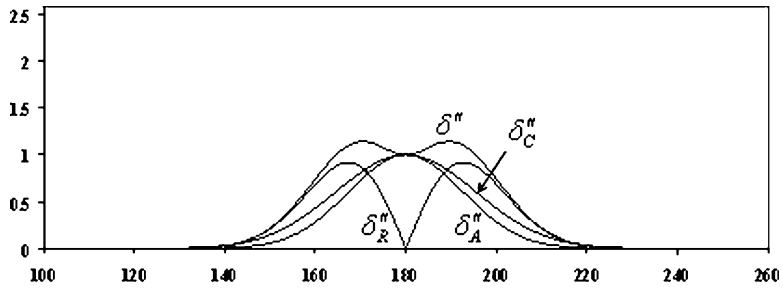

Fig. 2.8 Plot of the scaling curvature integrands, as functions of the angle, for the same lobe as in Fig. 2.5

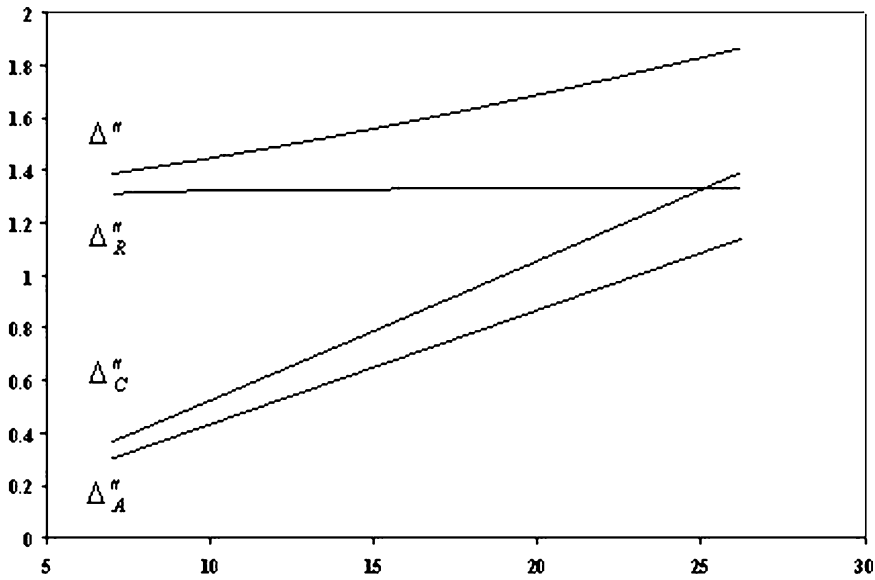

Fig. 2.9 Plots of the scaling error curvature, as functions of the −3dB lobe width in degrees. Notice that the vertical scale is much finer than in Fig. 2.6

2.3.3 The Optimization Algorithm

As we already said, the pattern matching that yields the source's bearing, and which follows from the rotated-lobe theorem, is a process of iterative minimization of the error function. The main variational parameter is of course the angle of rotation, but as we observed in Sect. 2.3, we must also allow for a variation in the scale of the input lobe. We can, of course, normalize the inputs to a fixed maximum value, and such normalization yields a good initial value of the scaling parameter, but this is not strictly correct, except in the case of exact alignment of the source with one antenna direction. Hence, there is a need for simultaneous optimization of direction and scale.

In our prototype implementation, we allowed for a third variational parameter, the base of the logarithmic conversion scale. This was driven by an engineering consideration: the measurement circuit AD8362, described in Sect. 5.2, reports its results roughly on the decibel scale, with some variation. Rather than calibrating the individual parts, we allowed for the base variation. This variation increases the computing time, and could well be eliminated in favor of a one-time calibration. Such calibration incurs practical costs also, and the selection of one course of action over another will depend on practical considerations of a specific implementation.

We should note here that the principles discussed so far do not depend on the choice of scale in which antenna data are expressed, and we switch freely between linear and decibel scales, as convenient. We always perform the minimization calculations in the linear scale, for an empirical reason: the logarithmic scale is biased in favor of lower magnitudes. This makes it a good visualization tool, but it gives the secondary lobes and their inevitable inaccuracies too great a say in determining the bearing.

In minimizing the error function, we utilized the widely used Nelder–Mead algorithm; it is one of many minimum-finding algorithms in multiple dimensions. For N variational parameters, the algorithm calculates the values of the error function on the vertices of a polyhedron in the $N + 1$ dimensional space. On the basis of some reasonable heuristic assumptions about the topography of the (continuous and differentiable) error function, it changes the size and shape of the polyhedron iteratively, until it settles on a vanishingly small polyhedron around the minimum. Because of the crawling appearance of the polyhedron iterations, Nelder–Mead optimization has earned the nickname "amoeba algorithm." The algorithm is reliable and fast, if given a good initial guess for the iterations, but like all such procedures, it is vulnerable to getting trapped in local minima, or to spending inordinate amounts of time crawling around in some irrelevant flat area of the error function. For a good description of the Nelder–Mead algorithm, we refer the reader to (Lagarias et al. 1998).

As we discussed in Sect. 2.3.1, choice of the form of variational error is important for the speed of the bearing calculation. Let us now estimate how accurately we must minimize the variational error. Suppose that we can only iterate the error down to a nonzero lower limit $\Delta_{\min}$; this limit defines an (approximately) elliptical horizontal cross section of the surface $\Delta(\rho, \sigma)$ around $\rho, \sigma = 0$, and the iterations terminate at a point within that ellipse (see Fig. 2.10). Vertical cross section of the surface Δ in the ρ-plane is a parabola described by:

$$\Delta(\rho) = \frac{\Delta''(\rho)}{2}\rho^2 \tag{2.23}$$

where $\Delta''(\rho)$ is the error curvature, defined in (2.3). For the error to be $\Delta_{\min}$, the rotational variation ρ must be within

$$\rho_{\max} = \sqrt{\frac{2}{\Delta''}\Delta_{\min}} \tag{2.24}$$

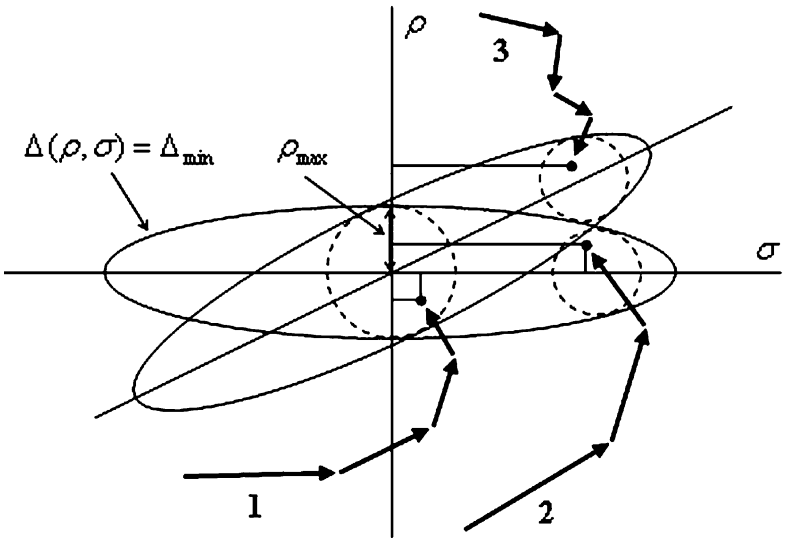

Fig. 2.10 Examples of iteration convergence in the ρ, σ plane

of the value zero. And conversely, if we wish to know the angle to an accuracy within $\rho_{\max}$, we must bring the variational error down to $\Delta_{\min}$.

We see from Figs. 2.6 and 2.9 that the scaling curvature is smaller than the rotational one, and the ellipse around the minimum will be elongated in the σ direction. If the iterations terminate in a point close to the origin, as in the Case 1 in Fig. 2.10, ρ and σ are comparable and small in size, and there is nothing to worry about. However, the iterations can terminate at a point along the minimum "valley" at which σ is sizable, as shown in Case 2. Diagonality of the Hess matrix in (2.15) insures that the valley is aligned with the σ-axis, and that the rotational variation, which is the important one, is still within $\rho_{\max}$. In the case of non-symmetric lobes, however, the minimum ellipse can be tilted, and the rotational parameter rendered inaccurate even though the error is below $\Delta_{\min}$ (Case 3 in Fig. 2.10).

The above discussion amounts to saying that for nonsymmetrical lobes, inaccurate scaling biases the optimization of the rotation, which is plausible. Of course, the value of $\rho_{\max}$ can be chosen small enough so that the entire ellipse falls within the desired angle accuracy, but at the price of more iteration steps and longer computing time.

2.3.4 Aliasing, or Too Few Antennas

Looking at Fig. 2.6, one might get the impression that the antenna lobe should simply be made as narrow as possible, because the rotational error minimum grows unambiguously "stiffer" for narrower lobes. There are limitations, however, to decreasing the lobe width. First, the more directional an antenna is, the larger it

must be, relative to the wavelength. For example, narrow-beam linear arrays must be long, utilizing the mutual reinforcement of many wave sources to restrict the spreading of radiation; this is reflected in (1.71) and (1.72) for the array's radiation pattern. The same can be said of reflectors, helices and various multi-element Yagi-Uda (see e.g., Elliott 2003) antennas as well. Trying to make the individual elements excessively directional threatens to drive the overall size of a "compound eye" antenna out of all proportions.

Second, narrow beam diminishes the overlap between antenna elements, reduces the number of adjacent elements that contribute to the rotated lobe and leads eventually to blind spots between them. Even without blind spots, we expect narrow lobes to lead to aliasing phenomena, and in this section we will describe and quantify this aliasing.

Figure 2.11 shows the result of a direction determination with simulated inputs: we plot the calculated source direction as a function of the "physical" source direction – ideally, this plot should be a straight line. The grid in Fig. 2.11 corresponds to the directions of antenna elements, and we see that each element causes a small wave in the plot. Rather than running at steady 45°, the line is a little flatter in the vicinity of the antenna, and compensates for that by making slight jumps between two antennas. Figure 2.12 shows this pattern in detail around the direction of 180°. It is as if the bearing calculation were attracted to the antenna directions.

Sequences of sketches in Figs. 2.13 and 2.14 illustrate what is going on. These are optimized matches between the known antenna element's lobe (dotted curve), and a lobe shape defined by the inputs (solid curve and discrete points along it); lobes are shown as functions of the angle in Cartesian coordinates. The input lobe

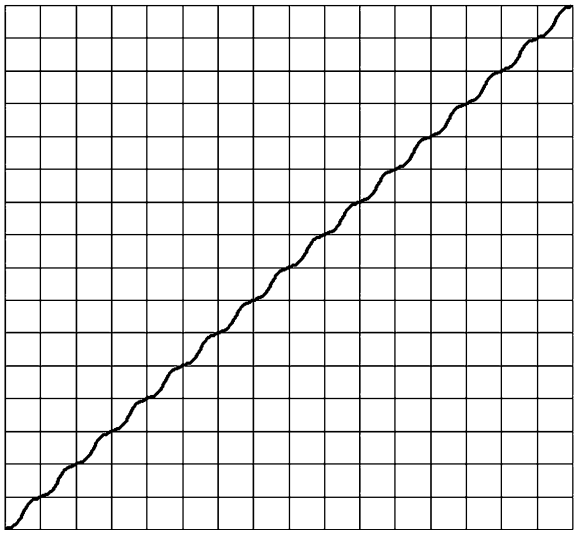

Fig. 2.11 Aliasing pattern of a 83° null-to-null lobe. Calculated bearing of the source is plotted with respect to its (simulated) physical bearing; angle ranges are 0–360° in both coordinates

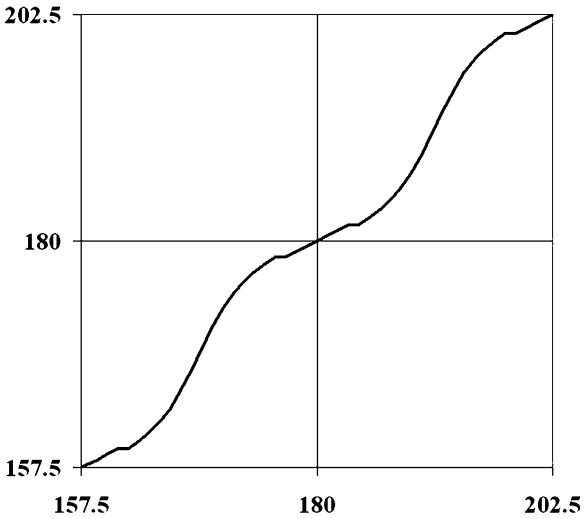

Fig. 2.12 Aliasing pattern in Fig. 2.11, magnified around 180°

(solid curve) is a cubic interpolation of the actual inputs. Physical direction of the simulated source is indicated by the vertical line, and the source moves through the sequence, from full alignment with the element at 180°, to the gap between elements, and onto the next antenna element.

Figure 2.13 shows the behavior of a narrow lobe: as the source moves to the right, there is not enough points on the right side of the lobe for a smooth hand-over, and the optimizing algorithm attempts to minimize the error by hanging onto the largest available input – hence the attraction to the physical direction of the antenna elements. Finally, it makes a jump toward the next element, and we see in the fourth slide that the algorithm distorts the interpolated input lobe quite desperately as it tries to cross the gap between the elements.

Figure 2.14 shows a wider lobe, making a much smoother transition through the same range of directions. There are always at least two physical points along each side of the lobe (possibly counting the top point once for each side), and even though the top of the input lobe follows the nearest element, the sides of the lobes are firmly lined up, and the aliasing is negligible. This gives rise to an intuitive criterion for adequate lobe width: the −3dB half-width (the inclination away from the main direction, at which the power is cut in half), should be at least as large as the angle between elements. When that is the case, there are always at least two physical points along the sides, and the direction of the rotated lobe is firmly defined.

For our model lobe, that of the linear array, the −3dB half-width is given by an analytical formula (1.73), and in the rest of this section we describe the lobe in terms of that half-width.

We quantify the aliasing as the maximum extent to which the curve in Fig. 2.11 deviates from the ideal 45-degrees line, within the entire 0–360° range. We perform these calculations for a continuous range of lobe widths, for different counts of

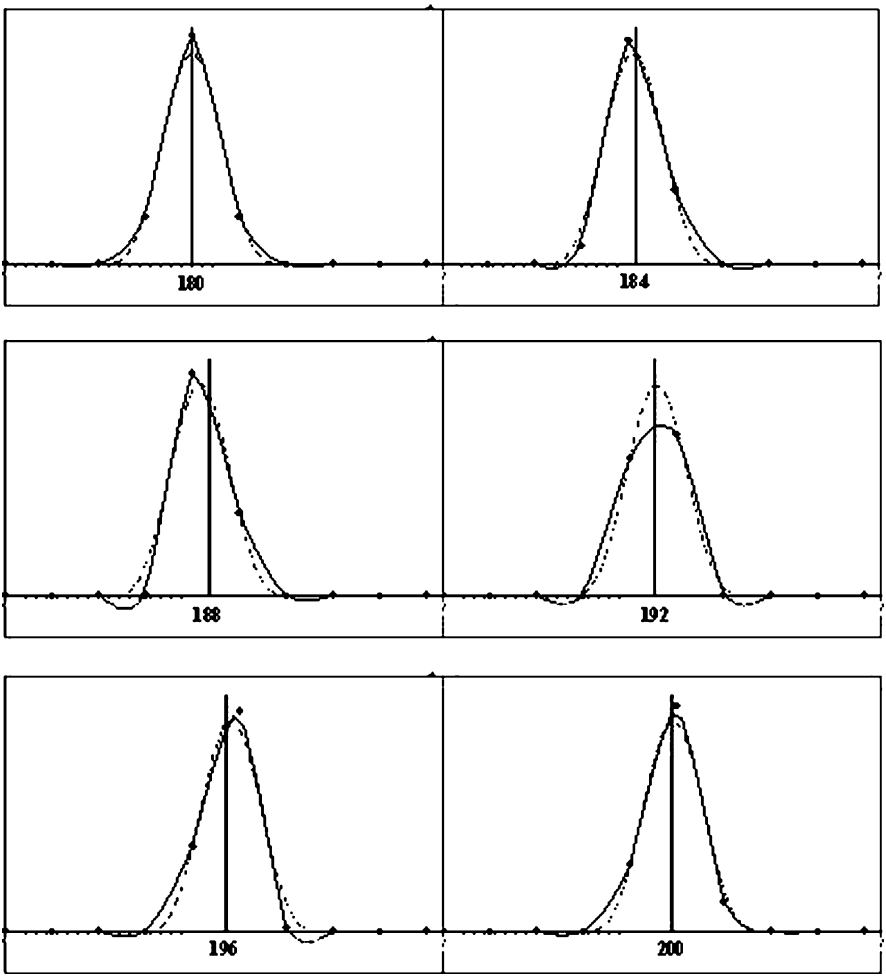

Fig. 2.13 Optimized overlap between rotated antenna lobe (*dotted*) and interpolated inputs (*solid curve*). Simulated input moves from 180° to 200°. Antenna lobe's −3dB half-width is 16°

antenna elements, and also for different interpolation schemes. Some typical results are shown in Fig. 2.15 (Antolovic 2009).

The aliasing amplitude shows a sharply defined threshold, a barrier in lobe width, below which the bearing calculation quickly becomes useless. Cubic and fifth-order interpolations change the shape of the barrier slightly, in the range of widths where the aliasing error is already growing large. It is clear from the discussion of Figs. 2.13 and 2.14 that increasing the interpolation order can only do so much, because for narrow lobes the input information about the lobe shape simply is not there.

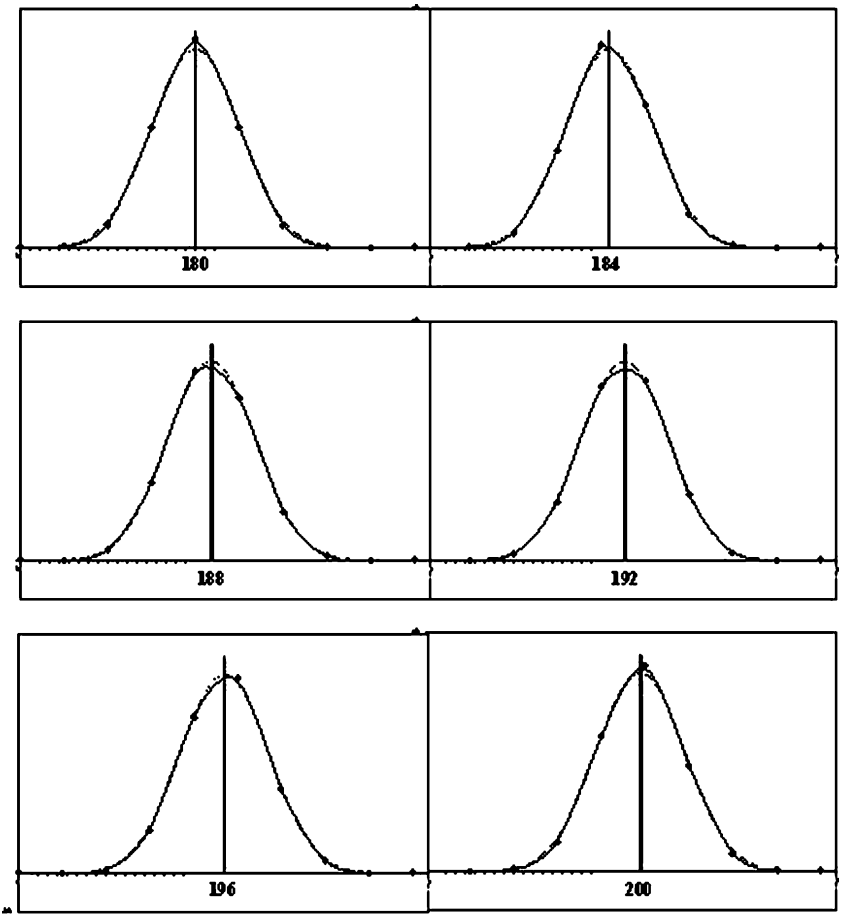

Fig. 2.14 Optimized overlap between rotated antenna lobe (*dotted curve*) and interpolated inputs (*solid curve*). Simulated input moves from 180° to 200°. Antenna lobe's −3dB half-width is 26.2°

Plotting the location of the aliasing barrier with respect to the inter-element angle produces a straight line, and the barrier location is always a few degrees below the inter-element angle itself (see Fig. 2.16). This is of course consistent with the fact that the barrier is simply caused by the inadequate overlap of the antenna elements, but more importantly, it confirms that our intuitive criterion – having the lobe half-width as large as the inter-element angle – is a valid, even somewhat conservative, criterion for avoiding the aliasing.

We conclude this section with the observation that, since the aliasing pattern in Figs. 2.11 and 2.12 is specific to the antenna, it could be measured, tabulated and compensated for. Such compensation works for aliasing amplitudes within a degree or so. However, flatness of the curve around antenna directions (Fig. 2.12) means a loss of algorithmic accuracy in these regions, i.e., the calculated numbers do not distinguish between physical directions very well. The jumps between the antenna

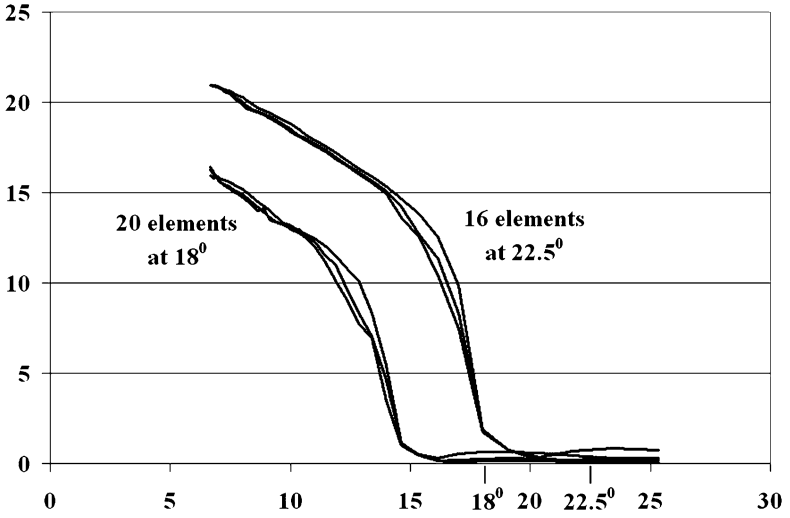

Fig. 2.15 Peak-to-peak amplitude of the aliasing (in degrees of deviation), plotted as the function of lobe's −3dB half-width (deg.), for two antenna arrangements. Sets of curves correspond to linear, cubic and 5-th degree interpolations, from right to left (© IEEE 2009, reproduced with permission)

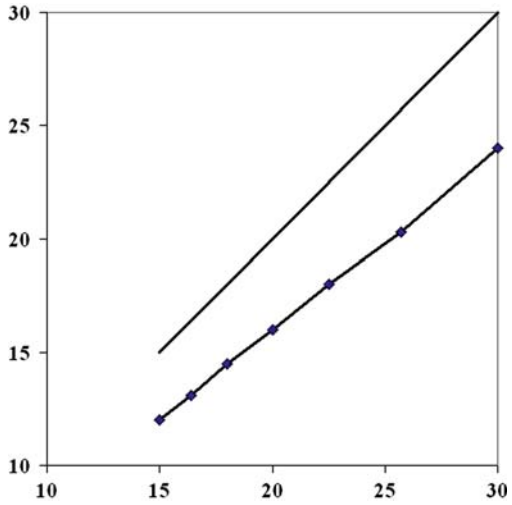

Fig. 2.16 Position of the aliasing barrier (*lower plot*), as a function of the inter-element angle. This figure shows that keeping the lobe half-width above the inter-element angle (*upper plot*) is more than sufficient to avoid aliasing. Both coordinates are measured in degrees

directions, on the other hand, amplify the measurement errors of the input, resulting in a bias toward one element or the other. It is best to minimize the aliasing through the physical design of the antenna, as much as the engineering considerations allow.

2.3.5 *Sources Above and Below the Antenna Plane*

We have carried out the analysis of the ring antenna in this chapter in terms of lobe geometry restricted to the plane of the ring. This is an appropriate way to investigate the antenna's properties, and the antenna is in fact designed for radiolocation in a plane, but in practice, directions of the sources will deviate from the ring's plane. In this section we estimate the robustness of our radiolocation algorithm under such variations.

The bearing of the source, as obtained by plane radiolocation, is in fact only its azimuth ψ in the ring's plane, and the source also has an elevation angle ε, above (below) the plane. Planar radiolocation can not determine that elevation, but it is important to estimate how accurately it can reproduce the azimuth for increasing elevation angles. Similar to the aliasing analysis in Sect. 2.3.4, we determine the amplitude of the discrepancy between physical and radiolocated azimuth, for all directions around the ring, and as a function of the increasing elevation angle. We carry out this investigation in simulation, and we limit it to rotationally (axially) symmetrical antenna lobes.

For an elevated source, the planar rotated lobe $L(\varphi - \psi)$ in Fig. 2.3 is replaced with a shape given by the intersection of the three-dimensional lobe facing the direction of the azimuth ψ, with a vertical cone of the angle $\pi/2 - \varepsilon$. As the elevation increases, the intersection of the cone and the main lobe becomes smaller and narrower, and it disappears when ε reaches the first null. In fact, the radiolocated azimuth exhibits an aliasing pattern, similar to that in Fig. 2.11, for borderline values of ε, due to the distortions of the lobe shape. Furthermore, the cone's intersection with the *secondary* lobe, which is toroidal in shape, does not decrease with increasing elevation, thus contributing to the algorithm's breakdown. We illustrate the above observations in Fig. 2.17, for a model lobe.

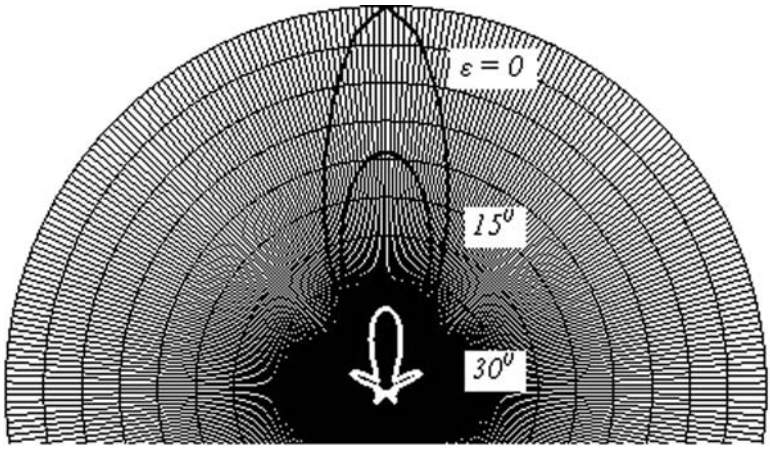

Fig. 2.17 Intersection of the axially symmetrical, three-dimensional lobe with a vertical cone of the angle $\pi/2 - \varepsilon$. The main lobe grows smaller and narrower for higher elevations, but the side lobes do not diminish in size

Figure 2.18 shows the magnitude of the discrepancy, plotted against the elevation angle and the lobe's 3dB half-width; the lobe half-widths in this figure are large enough to avoid the effects of inter-element aliasing, such as shown in Fig. 2.15. We see that there is a threshold elevation, beyond which the radiolocation breaks down rapidly, and this threshold is lower for narrower lobes, as one would expect. Significantly, Fig. 2.18 also shows that the accuracy of radiolocation remains intact for a significant range of elevations above and below the plane. For the lobe of the prototype's helical antenna (Figs. 2.20 and 4.4, lobe half-width of 20°), this result is shown in Fig. 2.19: the aliasing amplitude of ca. 2°, consistent with experimental measurements shown in Fig. 2.21, remains essentially constant for elevation angles up to ca. 27°, followed by a rapid breakdown.

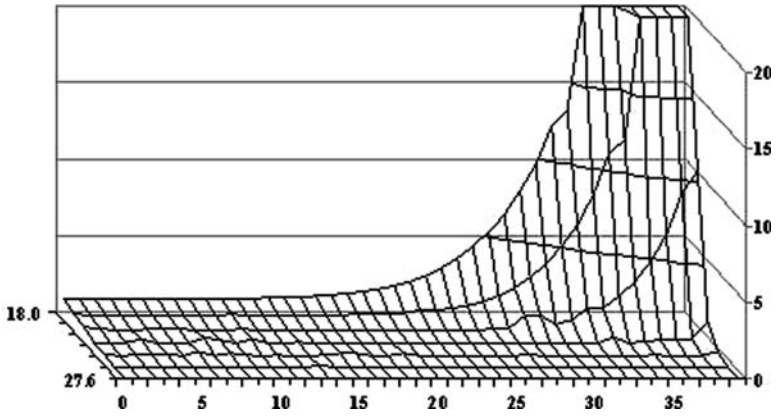

Fig. 2.18 Amplitude of the discrepancy between physical and radiolocated azimuths, as function of elevation (*long axis*) and lobe's 3dB half-width (*short axis*). The model lobe shape is that of the linear array, described in Sect. 2.3.2; all quantities are in degrees

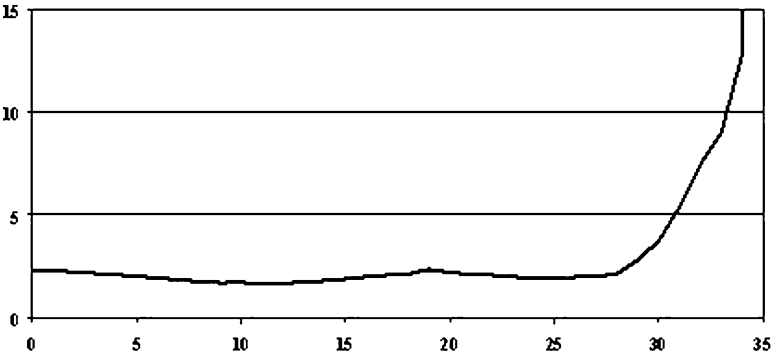

Fig. 2.19 Amplitude (maximum value over all directions) of the discrepancy for the prototype helical antenna, as shown in Fig. 2.20, with lobe half-width of 20°. Units of both coordinates are degrees

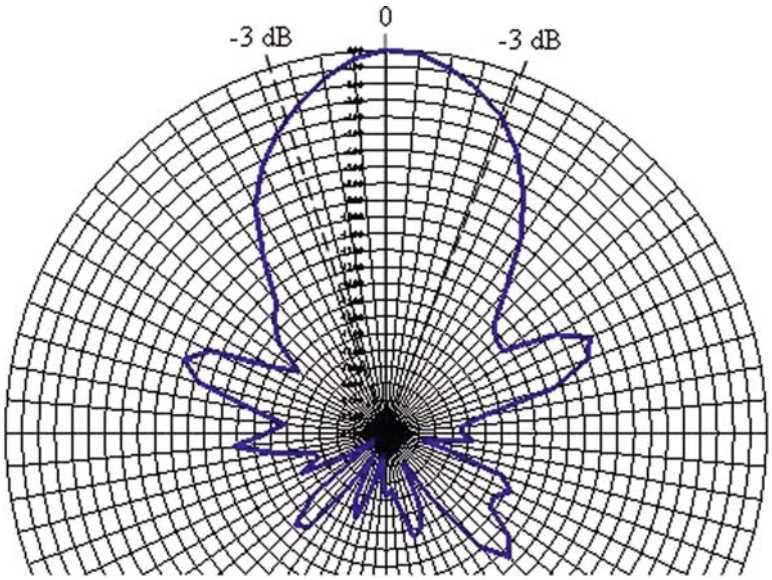

Fig. 2.20 Measured radiation pattern of the prototype helix described in Sect. 4.2.1, on the decibel scale. Angular intervals correspond to 5°

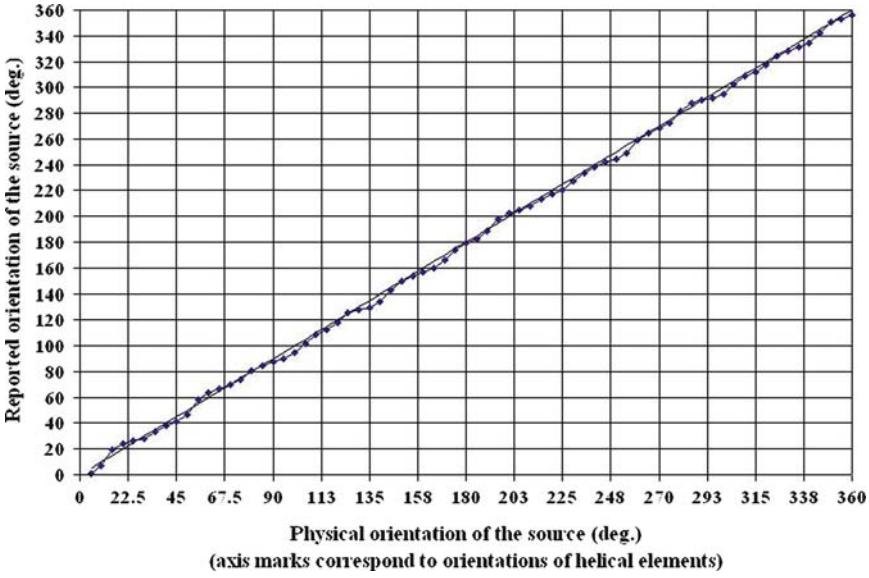

Fig. 2.21 Experimental test of the prototype radiolocation system. Test source was a 6 cm dipole, at the distance of 12 ft, and the operating frequency was 2.5 GHz. Data points are averages of 100 packets of unmodulated carrier, with $\sigma \approx 1$ degree. (© ACM 2006, reproduced with permission)

We can explain these results in qualitative terms, by observing that the cone's intersection of the main lobe is well approximated by a scaled-down planar intersection, for small elevation angles. That scaling is compensated for by the optimization algorithm (Sect. 2.3.3), and the bearing calculation remains unaffected. Stable range of elevations is determined by the shape of the lobe, and for lobes studied here it is close to the lobe's half-width, or equivalently, to the inter-element angle. For a fundamentally planar optimization algorithm, this range is quite satisfactory.

2.3.6 *A design Example*

Let us now use the results of this chapter to work through a design example. We will set the number of antenna elements to 16, arranged in a ring at equal angles of 22.5°. We next want to decide on the width of the element lobes, and we know from calculations illustrated in Figs. 2.15 and 2.16 that the lobe half-width of 22.5° will not give rise to aliasing. Half-width of 18° is about as narrow as we would venture, and we will consider both of these cases, and a wider lobe of half-width of 26°.

Next, we need to decide on the form of the error function. Figure 2.6 and (2.16–2.18) make it fairly clear that only Δ_C and Δ_R are worth considering: the full form Δ is computationally intensive, and is approximated very well by the radial form Δ_R; the angular approximation Δ_A, on the other hand, is neither computationally simpler nor more sensitive than Δ_C.

We set the desired angle accuracy $\rho_{\max}$ at a reasonable 0.1° ($1.7 \cdot 10^{-3}$ radians). After calculating the error curvature, we can determine the necessary accuracy of the variational error $\Delta_{\min}$, by means of (2.23). Using all of this, we perform simulated bearing calculations to determine the average processing time t_{avg}, average number of iterations I_{avg}, and the aliasing amplitude. We summarize the results in Table 2.1; the averages were taken over the entire circle of source directions.

Table 2.1 Design calculations for a 16-element compound antenna with linear-array lobes

−3dB half-width	26°	22.5°	18°
Null-to-null width	172°	120°	89°
Radial approx. Δ_R	$\Delta''_R = 5.8$ $\Delta_{\min} = 8.4 \cdot 10^{-6}$ $t_{\text{avg}} = 0.72$ ms $I_{\text{avg}} = 27$	$\Delta''_R = 8.8$ $\Delta_{\min} = 1.3 \cdot 10^{-5}$ $t_{\text{avg}} = 0.64$ ms $I_{\text{avg}} = 25$	$\Delta''_R = 14$ $\Delta_{\min} = 2.1 \cdot 10^{-5}$ $t_{\text{avg}} = 0.58$ ms $I_{\text{avg}} = 23$
Cart. approx. Δ_C	$\Delta''_C = 4.5$ $\Delta_{\min} = 6.5 \cdot 10^{-6}$ $t_{\text{avg}} = 0.83$ ms $I_{\text{avg}} = 31(*)$	$\Delta''_C = 5.4$ $\Delta_{\min} = 8 \cdot 10^{-6}$ $t_{\text{avg}} = 0.61$ ms $I_{\text{avg}} = 25$	$\Delta''_C = 6.95$ $\Delta_{\min} = 1.1 \cdot 10^{-5}$ $t_{\text{avg}} = 0.58$ ms $I_{\text{avg}} = 23$
Aliasing amplitude	0.21°	0.27°	1.7°

We see that the radial and Cartesian approximations differ only slightly in the overall computing time; time and iteration counts are shorter for the narrower lobe because of its larger Δ''. For the lobe, half-width of 26°, in Cartesian approximation, iterations grow long and fail to converge for some bearing angles because of the low error curvature Δ''_C (calculation marked with asterisk in Table 2.1).

The $\Delta_{\min}$ accuracy is achievable in single precision in all cases, which is advantageous for the speed of the calculation as well. The reader will appreciate our concern for the time budget of the calculations, since the shortest processing time of 0.58 ms translates into handling ca. 1,700 packets per second. Timing calculations in this example were carried out on a 2.0 GHz processor, which, as of this writing, is still a somewhat high clock speed for embedded applications; cutting the clock speed in half would yield the processing rate of ca. 860 packets per second, not an unthinkably high rate of wireless traffic.

The aliasing amplitude is somewhat problematic for the narrow lobe, and the null-to-null width between 100 and 120° may be preferable. To conclude, a 16-element compound antenna with elements 100° wide, supported by a single-precision bearing calculation, using the Cartesian error function, appears to be a reasonable design.

2.4 Implementation of a Compound Antenna

In the subsequent chapters, we will describe the physical implementation of a multi-antenna radiolocator system in considerable detail. In this section, we want to describe its "compound eye" antenna, in the light of the discussion laid out in the earlier sections of this chapter.

The antenna is a 16-element ring, consisting of helices described in Sect. 4.2.1, and shown in Fig. 4.4. The helical elements were designed according to standard engineering formulas for helical antennas (see e.g., ARRL 2000), and Fig. 2.20 shows their experimentally determined radiation pattern, on the decibel scale.

The −3dB half-width of this element is ca. 20°, slightly on the narrow side, according to our analysis. In the course of design work, we prototyped and tested a helix with a somewhat wider lobe; this alternative antenna had a troubling propensity to switch from its proper end-fire mode into a wide-radiating, dipole-like mode. It is known that helical antennas have this second, broadside mode (see Balanis 1997), which is more stable for shorter, wide-lobed helices. This placed a practical limit on the width of the lobe that we could achieve with the helical design.

Figure 2.21 (from Antolovic and Wallace 2006) shows the result of experimental testing of the radiolocation system. We see a slight aliasing pattern, which is within the range that can be compensated numerically; however, elements with somewhat broader lobes should be used in commercial devices of this type.

References

Antolovic, D., Wallace, S. Single-Packet Radiolocation of 802.11 Wireless Sources, Using an Array of Stationary Antennas and High-Speed RF Multiplexing, ACM Proceedings of Wireless Internet Conference (WICON), Boston, MA (August 2006)

Antolovic, D. Numerical Investigation of Algorithms for Multi-Antenna Radiolocation, Proceedings of 2009 IEEE International Conference on Portable Information Devices, Anchorage, AK (September 2009)

ARRL Antenna Book, American Radio Relay League, Newington, CT (2000)

Balanis, C.A. Antenna Theory, Analysis and Design, 2nd edition, Wiley, New York (1997)

Elliott, R.S. Antenna Theory and Design, revised edition, Wiley, New York (2003)

Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E. Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. SIAM J. Optim., 9(1), 112–147 (1998)

Chapter 3
Forming the Radio Image with Multiple Antennas

3.1 Introduction

In the preceding chapter, we discussed image-formation options at radio wavelengths, and we committed ourselves to an array of directional antennas, a "compound eye" architecture. We have seen that the directional sensitivity of individual sensors is essential, and that their fields must be somewhat overlapped. We have also seen (in Sect. 2.2) that every distant radio source produces sensors' outputs that lie on a lobe shape pointing toward that source; we refer to this shape as the rotated lobe.

The crux of radiolocating the sources lies in accurately determining the direction of the rotated lobe, on the basis of somewhat sparse data points, which are the antenna signals. Optimizing the fit between the antenna signals and the rotated lobe synthesizes the "image" of the source, i.e., its direction, and we see that the overlap and directional sensitivity of the individual sensors are essential for that synthesis. We entrusted the optimization to a straightforward iterative minimization algorithm, and while this approach was successfully implemented and proved reliable in a networking device, we explore in this chapter an alternative method of image synthesis.

The iterative optimization can be implemented only in sequential code, although the calculation of the variational error integral (2.2) can be carried out in parallel. The algorithm is nevertheless of variable duration, and is susceptible to the usual optimization derailments, which we discussed briefly in Sect. 2.3.3. Furthermore, it would be difficult to extend this iterative optimization to detect multiple sources simultaneously. For example, independently optimizing the rotated lobes for just two sources doubles the number of variational parameters, and much more than doubles the duration and likely computational difficulties.

One alternative is to subtract the signal of a found source from the total input, then repeat the optimization in order to find the next one; this algorithm is more tractable, but the computing time nevertheless grows proportionally to the number of sources, and the approach has an ad hoc air about it. We would prefer uniform processing of all antenna inputs, resulting in a signal with prominent peaks in the sources' directions. Simple thresholding and maximum-finding will then locate all of the sources in that signal at once.

D. Antolovic, *Radiolocation in Ubiquitous Wireless Communication*,
DOI 10.1007/978-1-4419-1632-7_3,

Of course, in strictly networking applications, the protocols limit the communication to one interlocutor at a time, on any given radio channel, and the issue of simultaneous radiolocation does not arise. The issue does arise, however, in broader radio surveillance, and from a theoretical perspective, we would like to understand the limits of what can be achieved with a "compound eye" sensor array. Here we will investigate an approach that can detect multiple sources, and which can be implemented as a computing process of fixed duration.

3.1.1 Note on Coherent Sources

Let us point out here that a multi-antenna method based on signal strength, i.e., on the power delivered to antennas, can distinguish only between noncoherent sources. This is because two coherent plane waves, traveling in different directions, fill the space with an advancing interference pattern. Signals received on an antenna add as phasors, and the power delivered is, in general, not equal to the sum of powers of the individual waves. Furthermore, the relative phase of the two waves changes from one antenna element to another, due to the waves' interference pattern, and the directional distinction between the two waves is lost.

If the phase between the two sources changes sufficiently rapidly during the power measurement, the detected signal strengths are additive on the average, and the sources can be distinguished. We will see in Chap. 5 that the measurement by an advanced analog power meter takes ca. 4 μs to complete; at the 2.4 GHz carrier frequency, that amounts to 9,600 periods. Two physically distinct sources, separated by a distance distinguishable by radiolocation, are very unlikely to maintain coherent phase over so many periods. The only realistic interference that can arise is from multi-path reception of several waves from the same source.

We may add that while compound eyes (spatially distributed arrays of directional sensors) are vulnerable to this type of interference, aperture instruments are not. The same linear additivity of waves, which gives rise to the interference, also allows the camera to collimate the two wavefronts onto two separate points (detectors) in the image plane, and their relative phase becomes immaterial. This is why our eyes can see an object and its image in the mirror at the same time, and see them as two separate things!

3.2 Representing the Antenna Signal in a Set of Basis Functions

We can envision the antennas' signal strengths as lying on a continuous surface, a function of all directions in space. We described this function for a single source as the rotated lobe in Chap. 2, but the function can be more complex. We know its actual values only in the directions of the physical antenna elements, but if we had infinitely many elements, densely covering all directions, we would know it completely.

Let us call our function $\Lambda(\mathbf{n})$, where $\mathbf{n}$ is the unit vector in a direction in space, which is in practice determined by the angular coordinates θ, φ. We will attempt to represent $\Lambda(\mathbf{n})$ as a linear combination over a family of known functions, which is called a *basis set*. If any arbitrary function can be expressed as a linear combination of the functions from a basis set, possibly using infinitely many basis functions, the basis set is *complete*. There are infinitely many complete basis sets of functions on any domain of continuous variables, but the expansion is likely to be useful only if the choice of basis reflects some inner structure of the problem that we are applying it to. For the interested reader, almost any introductory text on quantum mechanics (e.g., Messiah 1999) will offer a detailed, albeit usually somewhat idiosyncratic, coverage of these concepts (see also Byron and Fuller 1992).

Let us now choose as our basis set a collection of identical lobes $L(\mathbf{m})$, pointing in all directions $\mathbf{m}$; we assume the lobes to be axially symmetrical, for reasons explained in Sect. 2.2. It is intuitively obvious that this basis set must be complete enough to represent our function $\Lambda(\mathbf{n})$ because for a single source in a particular direction $\mathbf{m}_i$,

$$\Lambda(\mathbf{n}) \equiv L(\mathbf{m}_i) \tag{3.1}$$

In (3.1) we assume that L's have the shape of the lobes of the physical antenna elements; that is a convenient condition, which we will maintain, but it is not necessary for the completeness of the basis. We can also discern that the basis set must cover the directions densely (infinitely densely, ideally) because lobes in directions far from any $\mathbf{m}_i$ need not be linear combinations of those pointing in directions $\mathbf{m}_i$.

We write the basis-set expansion as:

$$\Lambda(\mathbf{n}) = \sum_{i=0}^{n-1} a_i \, L(\angle(\mathbf{n}, \mathbf{m}_i)) = \sum_{i=0}^{n-1} a_i \, L_i(\mathbf{n}) \tag{3.2}$$

Here, $L_i(\mathbf{n})$ is the abbreviation for the value of the lobe function centered at $\mathbf{m}_i$, in the direction $\mathbf{n}$ (see Fig. 3.1). We truncated the expansion to a finite number of terms, as must be the case in practice, but conceptually this is an infinite series, and can even be an integral over continuous parameters, as we shall see. This expression is not limited to single-source Λ functions, and in the case of multiple sources, we expect the expansion coefficients a_i to be sizable for the terms close to the directions of the sources. The trick is to find the expansion coefficients.

Assuming for a moment that we actually know the function Λ, we will choose the coefficients so as to minimize the sum-of-squares deviation E of the expansion from the function itself:

$$E = \int \mathrm{d}\mathbf{n} \left[\Lambda(\mathbf{n}) - \sum_i a_i \, L_i(\mathbf{n}) \right]^2 \tag{3.3}$$

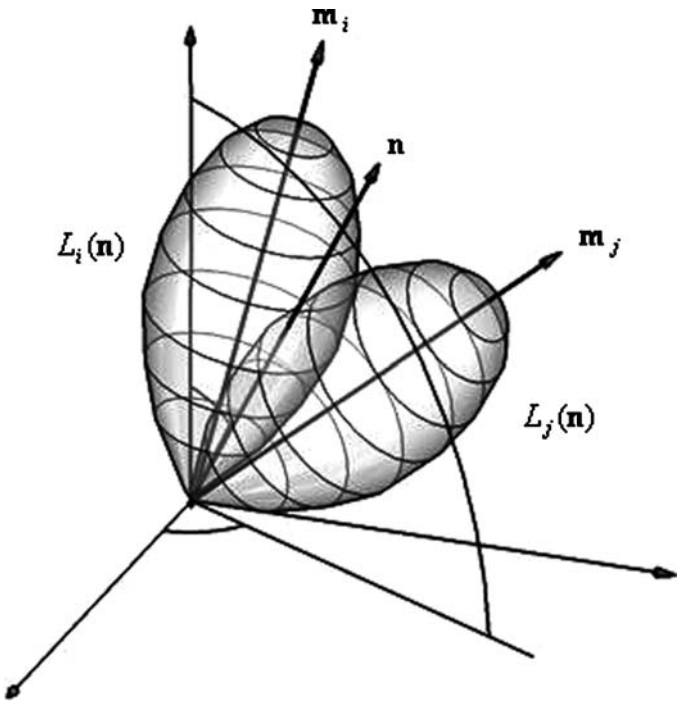

Fig. 3.1 Basis functions in arbitrary directions in space

Derivatives of E with respect to all coefficients must vanish for the optimal choice of coefficients:

$$\frac{\partial E}{\partial a_j} = (-2)\int \mathrm{d}\mathbf{n}\left[\Lambda(\mathbf{n}) - \sum_i a_i\, L_i(\mathbf{n})\right] L_j(\mathbf{n}) = 0; \quad \forall\, j \tag{3.4}$$

$$\int \mathrm{d}\mathbf{n}\left[\Lambda(\mathbf{n}) - \sum_i a_i\, L_i(\mathbf{n})\right] L_j(\mathbf{n}) = 0; \quad \forall\, j \tag{3.5}$$

And, by interchanging summation and integration,

$$\int \mathrm{d}\mathbf{n}\,\Lambda(\mathbf{n}) L_j(\mathbf{n}) = \sum_i a_i \int \mathrm{d}\mathbf{n}\; L_i(\mathbf{n}) L_j(\mathbf{n}) \tag{3.6}$$

This is a set of linear equations in the expansion coefficients, which can be written in matrix form

$$\mathbf{\Lambda} = \mathbf{Sa} \tag{3.7}$$

We introduced here the (symmetric) *matrix of overlap integrals* of the basis functions:

$$S_{ij} = S_{ji} = \int \mathrm{d}\mathbf{n}\; L_i(\mathbf{n}) L_j(\mathbf{n}) \tag{3.8}$$

and the *input overlap vector*:

$$\Lambda_j = \int \mathrm{d}\mathbf{n}\, \Lambda(\mathbf{n}) L_j(\mathbf{n}) \tag{3.9}$$

Finally, if we take second derivatives of E, we obtain from (3.4)

$$\frac{\partial^2 E}{\partial a_j \partial a_k} = 2S_{jk} \tag{3.10}$$

Overlap integrals S_{ij} have all the properties of scalar products of the basis functions L_i; in fact, (3.8) can be interpreted as a (continuous) sum of products of vector components $L_i(\mathbf{n})$ in a basis of delta functions placed at every $\mathbf{n}$. The matrix $\mathbf{S}$ is therefore the so-called *Gram* matrix (matrix of scalar products) of the basis L_i, and for a set of linearly independent vectors the Gram matrix is positive definite; see, for example, Gantmacher (1959) for the details of the proof.[1] It follows that the matrix of second derivatives in (3.10), the *Hessian* of E, is also positive definite. Therefore, if there is a unique solution of the linear system, (3.7), it will yield expansion coefficients at the true minimum of E, and will give the best approximation of the function Λ for a chosen set of basis functions. This is true even if the expansion in (3.2) is truncated, i.e. for an incomplete, finite basis set. For a complete finite basis, the expansion coefficients satisfy (3.5) by definition, and (3.7) follows automatically.

On the face of it, we have outlined a promising approach to forming multisource images of incoming waves. Once the basis set is chosen, the overlap matrix in (3.8) is fixed: it can be inverted, and $\mathbf{S}^{-1}$ tabulated; the input overlap vector is an integral over the known values of Λ, possibly with some interpolation, and the requisite values of L under the integral in (3.9) can also be tabulated. We solve the system (3.7) as $\mathbf{a} = \mathbf{S}^{-1}\mathbf{\Lambda}$, by a single matrix multiplication of fixed size, and the components of the solution vector $\mathbf{a}$ are proportional to the radiated power arriving from the corresponding directions. We can speak of the vector $\mathbf{a}$ as the image formed by the compound antenna.

The difficulty lies in the numerical properties of the system (3.7). In order to be complete, our basis set must contain many functions of identical shape, which differ only slightly from each other in their direction. Although the equations in the system (3.7) are nominally independent as long as no two basis functions are

[1] In short, the determinant of the Gram matrix of n linearly independent vectors is positive. But any subset of independent vectors must also be linearly independent; therefore all principal minors of the above Gram matrix also have positive determinants. It follows from an algebraic theorem (Sylvester) that the Gram matrix is positive definite, i.e., it has all positive eigenvalues.

identical, the system is numerically close to being singular because of the nature of the basis set. As an example, for a modest basis set of 36 lobes, 100° wide null-to-null, and arranged evenly in a circle, $\det \mathbf{S} \approx 3 \cdot 10^{-40}$. No physical quantity in this arrangement calls for numbers of such magnitude; rather, the system of equations is almost singular because of the close similarity of the basis functions. Generic equation-solving algorithms fail to solve (3.7): they run into numerical overflows, and lose accuracy by calculating small differences of large terms. We must approach the image formation with a bit more subtlety, and in the sections that follow, we will see that the very symmetry of the vision field opens an elegant way to not only calculating the image, but to understanding its structure as well.

3.3 Image Formation in Circular Geometry

Let us consider the overlap matrix of a basis set of identically shaped lobes, evenly spaced around a planar ring. The overlap between any two lobes (the matrix element S_{ij}) depends only on the magnitude of the angle between the lobes, not on their absolute positions. First row (column) of $\mathbf{S}$ contains the overlaps of the (arbitrarily chosen) zero-indexed element with itself and all the others (the self-overlap S_{00} will be the largest of them). Because the overlap depends only on the angular difference, all the successive rows (columns) must be successive circular shifts of the first one, and $\mathbf{S}$ is what is known as a *circulant* matrix. This is really a considerable simplification – a circulant matrix is fully determined by a single column – and it is brought about in this case by a particular symmetry of the arrangement, the invariance of the overlap under a discrete set of rotations.

As a remarkable consequence of this simplification, all circulant matrices of order n are diagonalized by the same unitary transformation. The transformation is, in matrix form:

$$\mathbf{U} = n^{-1/2}[\exp(-2\pi imk/n); \quad m,k = 0,1,\ldots,n-1] \tag{3.11}$$

This is the normalized matrix of the powers of complex roots of unity, i.e., the matrix representation of the discrete Fourier transform (DFT). It is easily shown that $\mathbf{U}$ is unitary, i.e.,

$$\mathbf{U}\mathbf{U}^* = \mathbf{U}^*\mathbf{U} = \mathbf{I} \tag{3.12}$$

where the star indicates the Hermite adjoint (conjugate transpose) of a matrix. The diagonal decomposition of a circulant matrix $\mathbf{C}$, determined by its first column $\mathbf{c}$, is as follows:

$$\mathbf{C} = \mathbf{U}^* \mathrm{diag}(n^{1/2}\mathbf{U}\mathbf{c})\,\mathbf{U} \tag{3.13}$$

Eigenvectors of all circulant matrices (of order n) are the same, and the eigenvalues are given by the unnormalized DFT of the matrix' first column (expression $\mathrm{diag}(\mathbf{x})$ represents a diagonal matrix with components of the vector $\mathbf{x}$ strung along

the diagonal). Algebraic proof of this result is not difficult, and can be found in (Gray 2006; Davis 1979), but for our purposes the following reasoning by analogy will prove more useful:

Let us suppose that the overlap **S** is not a discrete matrix, but an operator acting upon functions on the unit circle. The equivalent of the matrix-vector multiplication is the convolution of the operator with a function:

$$[\mathbf{Sa}](\varphi) = \int \mathrm{d}\varphi'\, S(\varphi' - \varphi)\, a(\varphi') \tag{3.14}$$

We can think of this convolution as the matrix product, represented in the infinitely dense basis of delta functions covering all directions on the unit circle. Crucially, the operator S depends on the difference of angles only.

A broad class[2] of operators can be diagonalized by convolution with an appropriate unitary transformation U, over both of its variables. This is equivalent to changing the representation basis to that of the operator's eigenvectors, in which basis the operator is diagonal. We will tentatively parameterize this basis with a parameter k, and write the double convolution as:

$$S_F(k, k') = \int \mathrm{d}\varphi\, \mathrm{d}\varphi'\, U^*(\varphi, k)\, S(\varphi' - \varphi)\, U(\varphi', k') \tag{3.15}$$

We would like to separate the integration variables, and use the unitary property of U to evaluate at least one integral. The operator S cannot be factorized, but substituting $\varphi'' = \varphi' - \varphi$ gives:

$$S_F(k, k') = \int \mathrm{d}\varphi\, \mathrm{d}\varphi''\, U^*(\varphi, k)\, S(\varphi'')\, U(\varphi + \varphi'', k') \tag{3.16}$$

This move has transferred the factorization requirement from S onto U, and there is indeed an operator, unitary on the circle, that factorizes under the sum in its parameter:

$$U(\varphi, k) = \frac{1}{\sqrt{2\pi}} \exp(-ik\varphi) \tag{3.17}$$

This is the operator of the Fourier transform; by substituting into (3.16), we obtain

$$S_F(k, k') = \int \mathrm{d}\varphi\; U^*(\varphi, k)\, U(\varphi, k') \int \mathrm{d}\varphi''\, S(\varphi'') e^{-i\, k' \varphi''} \tag{3.18}$$

and by using the unitary property of U:

$$S_F(k, k') = \delta(k - k') \int \mathrm{d}\varphi''\, S(\varphi'') e^{-ik' \varphi''} = \delta(k - k')\, S_F(k') \tag{3.19}$$

[2] The so-called *normal* operators. An operator is normal if it commutes with its adjoint (see e.g., Byron and Fuller 1992). Since S is real and symmetric, it is certainly normal.

The operator S has been represented in the basis of harmonic functions on the unit circle, and in that basis it is diagonal. Its diagonal elements comprise the (unnormalized) Fourier transform of its values for one variable held constant ($\varphi = 0$), i.e., the Fourier transform of one column of its matrix representation.

Periodic condition on the circle requires that $k = 0, 1, 2, \ldots$ We can replace the integral in (3.19) with a summation over n equally spaced angles

$$\varphi_m = 2\pi\, m/n; \quad m = 0, 1, 2, \ldots, n-1 \tag{3.20}$$

and correspondingly replace the normalization constant in (3.17) with $n^{-1/2}$; also, on a discrete set of angles, period of the harmonic functions is limited by $k \leq n-1$ (higher frequencies are aliased). Introducing this discretization into (3.19) yields the results contained in (3.11) and (3.13). ■

We should point out that the above derivation shows the close connection between the rotational invariance of the operator S and the form of its diagonalization transformation; the latter must be factorizable under the sum of angles.

Returning now to the solution of (3.7), we introduce the label $\mathbf{s}$ for the first row (column) of the overlap matrix $\mathbf{S}$, and we also introduce the vector:

$$\boldsymbol{\psi} = n^{1/2}\mathbf{U}\mathbf{s} \tag{3.21}$$

Applying the diagonalization (3.13) to the overlap matrix, we obtain the solution of (3.7), after a few elementary steps:

$$\mathbf{a} = \mathbf{S}^{-1}\boldsymbol{\Lambda} = \mathbf{U}^{*}\,(\operatorname{diag}\boldsymbol{\psi})^{-1}\,(\mathbf{U}\boldsymbol{\Lambda}) \tag{3.22}$$

The result in (3.22) is easily understood. The vector $\mathbf{Sa}$ in (3.7) consists of scalar products of $\mathbf{a}$ with increasingly (circularly) shifted vector $\mathbf{s}$; these products form the circular convolution, $\mathbf{s} \circ \mathbf{a}$. Fourier transform carries the convolution (labeled $\circ$) of functions into ordinary, term-by-term product (labeled $\cdot$) of the functions' transforms, and we can write (3.7) as follows:

$$\mathbf{s} \circ \mathbf{a} = \boldsymbol{\Lambda} \;\overset{F}{\longrightarrow}\; \mathbf{s}_F \cdot \mathbf{a}_F = \boldsymbol{\Lambda}_F \tag{3.23}$$

Here $\boldsymbol{\Lambda}_F = \mathbf{U}\boldsymbol{\Lambda}$ and $\mathbf{s}_F = \boldsymbol{\psi}$, and we solve the transformed equation for the spectrum of the image:

$$\mathbf{a}_F = \frac{1}{\mathbf{s}_F}\,\boldsymbol{\Lambda}_F = (\operatorname{diag}\boldsymbol{\psi})^{-1}\,(\mathbf{U}\boldsymbol{\Lambda}) \tag{3.24}$$

The expression $(\operatorname{diag}\boldsymbol{\psi})^{-1}\,(\mathbf{U}\boldsymbol{\Lambda})$ is the image spectrum, expressed as term-wise division of two transforms; (3.22) transforms the image spectrum back into a function in ordinary space, and we obtain the coefficients of the basis-set expansion, i.e., the image proper.

The topic of Fourier transform is covered in countless books, and from various perspectives (e.g., Morse and Feshbach 1953; Byron and Fuller 1992; Hamming 1998; etc.). For a practical and thorough engineering-level account of the discrete Fourier transform, see (Smith 1997). Tables of better-known transforms can be found in formula handbooks (e.g., Gradshteyn and Ryzhik 2007).

3.3.1 Image Resolution

Once we have ventured into detecting multiple sources, we must address the question of resolution: how close can the sources' directions be, and still be seen as separate sources by the detector? We mentioned in Sect. 1.5.2 that the half-size of the image of a point source is the usual measure of resolution, so let us determine what the formalism that we are developing can tell us about image size.

As we know, a point source gives rise to a rotated lobe oriented in its direction, and the rotated lobe in turn gives rise to the input overlap vector $\mathbf{\Lambda}$ in (3.7) and (3.9). In the ideal case, the rotated lobe is identical in shape to a basis function (scaled according to source's strength), and $\mathbf{\Lambda}$ consists of almost the same overlaps as the vector $\mathbf{s}$, the first column of the overlap matrix $\mathbf{S}$. However, vector $\mathbf{s}$ contains the overlaps with the basis function in zero-th direction, and in the ring geometry $\mathbf{\Lambda}$ is circularly shifted relative to it. In fact, it is obvious that if $\mathbf{\Lambda}$ is equal to, say, i-th column of the circulant matrix $\mathbf{S}$, (3.7) is solved by the vector $\mathbf{a} = [0\ 0\ \cdots\ a_i\ \cdots\ 0\ 0]^T$. Ideally, the image of a point source is a delta function, as it should be.

We may ask naively whether (3.7) could be solved simply by matching $\mathbf{\Lambda}$ with columns of $\mathbf{S}$ until we obtained the best fit, thus bypassing the numerical problems of inverting $\mathbf{S}$. The answer is: yes, but only for the single source, and even then the overlap vectors will not match perfectly, and we will be thrown back at a discretized (aliased) version of pattern matching. Should we follow that route, the full-blown iterative method described in Chap. 2 will be a better choice.

We saw in (3.24) that the image spectrum is given by this ratio:

$$\mathbf{a}_F(k) = \frac{\mathbf{\Lambda}_F(k)}{\mathbf{s}_F(k)} \tag{3.25}$$

Each point source makes its own scaled and rotated contribution to $\mathbf{\Lambda}$, because of the rotated lobe theorem; let us call these contributions $a_j \boldsymbol{\lambda}(\varphi - \varphi_j)$, where $\boldsymbol{\lambda}$ is a rotated lobe, and the index j counts the sources.[3] Using the linearity of the Fourier transform, and the fact that shifts in coordinate space transform into phase rotations, we obtain

$$\mathbf{\Lambda}_F(k) = \sum_j a_j \exp(-i\varphi_j k)\boldsymbol{\lambda}_F(k) \tag{3.26}$$

[3] In our discretized picture, the direction of j-th source is numbered as the m_j direction on the circle, and $\varphi_j = 2\pi\, m_j/n$.

But $\boldsymbol{\lambda}_F(k) = n^{-1/2}\mathbf{s}_F(k)$, (because each source induces a lobe of the same shape as the basis lobe) and we obtain this expression for the image spectrum:

$$\mathbf{a}_F(k) = \frac{\boldsymbol{\Lambda}_F(k)}{\mathbf{s}_F(k)} = n^{-1/2}\sum_j a_j \exp(-i\varphi_j k) \tag{3.27}$$

The image spectrum is a phasor sum, made up of strengths and bearings of individual sources. The beauty of the Fourier transform is that the back-transformation, which consists of convolutions with waves of increasing k (and which can be performed in parallel), reconstructs all the unknown strengths and bearings from the spectrum at once. This is the crux of the multi-source detection.

The sum in (3.27) is independent of the lobe shape: if ideal precision applied, any antenna lobe would give rise to a perfectly crisp image, made up of delta functions at correct bearings. However, this sum can be calculated only as a ratio of quantities that decrease rapidly with k, and which soon become too small to calculate with any precision. This is, of course, the root cause of the numerical difficulties in solving (3.7): some diagonal elements of $\mathbf{S}$ effectively vanish, even though vanishing quantities appear only in finite ratios, as shown in (3.27), and the system nominally possesses finite solutions. This numerical behavior does depend, and quite strongly, on the shape of the lobe function. In a complete basis set, Fourier convolution theorem reduces (3.25) further, to the ratio of spectra of the input function and the lobe. This simplification is of marginal numerical significance.

The more rapidly a function changes in the coordinate space, the higher the frequencies that must be present in its spectrum. Since the main lobes of directional antennas are typically rounded-out shapes, and we assume them to be normalized in height, this means that the width of a lobe's Fourier transform changes roughly inversely with the width of the lobe itself.

In order to illustrate the reasons for image broadening, we look at a model lobe, described by the Gaussian distribution with variance σ. This is not a realistic antenna pattern, but it has the advantage that its Fourier transform is also a Gaussian, in the k-coordinate, with variance $1/\sigma$. We will now show that, for the Gaussian lobe shape, the relationship between image width and lobe width is strictly linear.

First, we note that the overlap of two Gaussian lobes also has Gaussian dependence on the angle between the lobes. If the lobe function is

$$L(\varphi) = \exp[-\varphi^2/(2\sigma_L^2)] \tag{3.28}$$

then the overlap is

$$\begin{aligned} S_{i0} = S(\Psi_i) &= \int \mathrm{d}\varphi\, L(\varphi - \Psi_i)\, L(\varphi) \\ &= \left(\int \mathrm{d}\varphi\, \exp\left[-\varphi^2/\sigma_L^2\right]\right) \cdot \exp\left[-\Psi_i^2/(2\,(\sqrt{2}\sigma_L)^2)\right] \\ &= C \cdot \exp\left[-\Psi_i^2/(2\sigma_S^2)\right] \end{aligned} \tag{3.29}$$

We obtained this result by changing the integration variable, $\varphi = \varphi' + \Psi_i/2$, and rearranging the exponents. Here $C \approx \sqrt{\pi}\,\sigma_L$ for lobes much narrower than the full circle, but the important point is that the overlap is a Gaussian in Ψ_i, with variance

$$\sigma_S = \sqrt{2}\,\sigma_L \tag{3.30}$$

Next we observe that the Fourier transform of a Gaussian is again a Gaussian, with inverse value of the variance, i.e., the width of the spectrum is the exact inverse of the width of the original function. This is a well-known property of the Gaussian, and we will state it without proof (see e.g., Byron and Fuller 1992):

$$\exp[-\varphi^2/(2\sigma^2)] \xrightarrow{F} \exp[-k^2/(2\sigma_F^2)] \tag{3.31}$$

where, applied to the transform of the overlap, the variance of the spectrum is:

$$\sigma_F = 1/\sigma_S \tag{3.32}$$

For a single source, the ratio in (3.27) is that of two Gaussians with the same variance σ_F; the spectrum phasor has only one term and constant amplitude for all k; this is, of course, the spectral representation of a delta function at the angular position of the source. This ratio is, however, computable only within the non-vanishing footprint of the Gaussians $\mathbf{\Lambda}_F(k)$ and $\mathbf{s}_F(k)$, centered at $k = 0$; the width of the footprint we can measure in variance units: $w \cdot \sigma_F$.

If we simply throw away the uncomputable components, the image spectrum $\mathbf{a}_F(k)$ (also labeled $\mathbf{Ua}$) becomes a rectangular box of width $w\sigma_F$. Its inverse Fourier transform is no longer the delta, but the sinc[4] function, whose central lobe has the half-width to the first null:

$$\varphi_0 = 2\pi/(w\sigma_F) \tag{3.33}$$

This is the half-width of the image of a single source, or the customary measure of resolution (Sect. 1.5.2). As expected, the image of the point source has been broadened because of the numerical loss of spectrum at high frequencies.

In order to express the lobe width as our customary −3 dB half-width, we use the following formula, which applies to Gaussian lobes:

$$\text{lobe h.w.} = 1.177\,\sigma_L \tag{3.34}$$

[4] Definition of the sinc function is: $\text{sinc}(x) = \sin(x)/x$. The sinc and the rectangular box constitute a pair of well-known mutual Fourier transforms, and for the infinitely wide box the sinc becomes infinitely narrow: the delta function.

We can now gather the proportionalities in (3.30) and (3.32)–(3.34) to obtain the proportionality relationship between image resolution and the −3 dB half-width of the Gaussian lobe:

$$\varphi_0 = \frac{2\pi\sqrt{2}}{w \cdot 1.177} \cdot (\text{lobe h.w.}) \tag{3.35}$$

The image resolution is a direct function of the available spectrum: the narrower the lobe, the wider its computable spectral footprint, and the finer the resolution. Likewise, for a given lobe width, the larger the computed fraction of the footprint (larger w), the better is the resolution. The above discussion of spectral transformations is illustrated in Fig. 3.2.

Figure 3.3 shows the image half-width as a function of lobe half-width, for the Gaussian lobe and for the model linear-array lobe introduced in Sect. 2.3.2. The dependence is linear for the Gaussian lobe, as predicted, and the resolution is somewhat more favorable for the linear-array lobe since this shape does not have the broad tails of the Gaussian distribution.

The question of the spectral cut-off is somewhat delicate. If the cut-off is too low, useful information is lost, resulting in image broadening; if it is too high, some of the image spectrum will have more or less random magnitudes and phases, resulting in undesirable artifacts in the image. We can view this cut-off as a low-pass filter, and the choice of the filtering window is a heuristic one. A simple and workable criterion is to retain those parts of the image spectrum whose magnitude is smaller than the magnitude of the corresponding component of ψ, i.e.,

$$|(\mathbf{U\Lambda})_i| < |\boldsymbol{\psi}_i|^2 \tag{3.36}$$

3.3.2 Aliasing Again

So far, our discussion in this chapter assumed that the rotated lobe was known for all angles, i.e., that the vectors $\mathbf{s}$ and $\boldsymbol{\Lambda}$ were calculated with equal accuracy. In reality, the function $\Lambda(\mathbf{n})$ in (3.2) is known only from measurements in the discrete set of directions corresponding to the physical antennas, and the integration in (3.9) requires interpolation of $\Lambda(\mathbf{n})$. There are two consequences to this.

First, as we would expect, finite angle between antenna elements places a lower limit on the acceptable lobe width: for too narrow a lobe, integration over measurement points in (3.9) becomes meaningless. As in the pattern-matching algorithm in the previous chapter, there is a rapid rise in the aliasing error when the lobe's −3 dB half-width drops below a critical value, which is close to the inter-element angle. Figure 3.4 shows this effect for a 16-antenna ring of our two types of model lobes, Gaussian and linear array, with fifth-order polynomial interpolation of the antenna input.

Second consequence of the inputs' interpolation is an additional image broadening, as can be best seen from its effect on the spectra. Figure 3.5 shows the spectra for a 16-antenna configuration and an admittedly much too narrow lobe with the

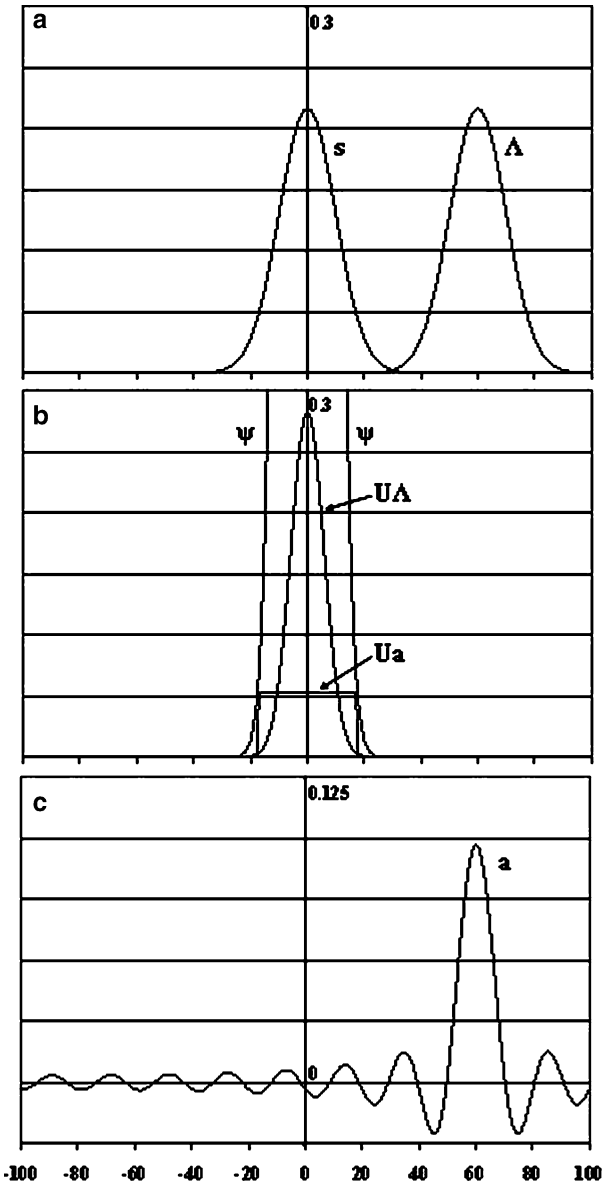

Fig. 3.2 Illustration of image formation for a source of unit strength at 60°. Basis functions are Gaussians with $\sigma = 7°$, $n = 360$, and the input function is ideal (not interpolated). (**a**) Basis and input overlaps are Gaussians with $\sigma = 9.9°$, identical in shape. (**b**) Spectral amplitudes of the overlaps ($\boldsymbol{\psi}$ is un-normalized DFT, and therefore larger than $\mathbf{U\Lambda}$), and their ratio **Ua**, truncated to the rectangular box of height $n^{-1/2} = 0.053$. (**c**) Image of the source is a sinc function centered at 60°, ca. 20° wide null-to-null

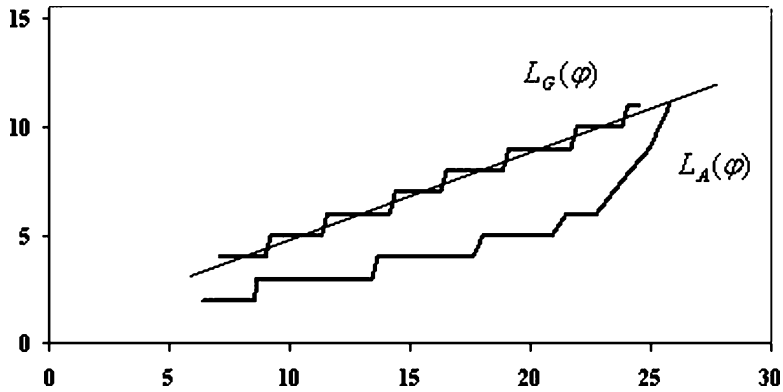

Fig. 3.3 Image size of a point source, as function of lobe width, for Gaussian and linear array lobes, at ideal input precision. Image size is discretized to one-degree accuracy, but the trends are apparent, and a straight-line fit is drawn for the Gaussian lobe. Both coordinates represent −3 dB half-widths, measured in degrees (© IEEE 2009, reproduced with permission)

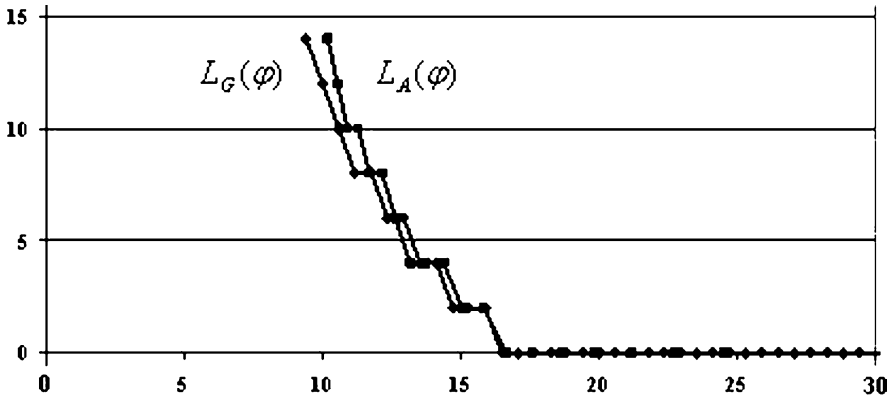

Fig. 3.4 Amplitude of the aliasing, as a function of lobe's −3 dB half-width, for a 16-element antenna ring, with 5-th order interpolation of the inputs. Both coordinates are in degrees

10.6° half-width: for this lobe width the aliasing is very large, as seen in Fig. 3.4, but the spectral effect is also prominent. In the case of the ideally precise input function $\Lambda(\mathbf{n})$, magnitudes of $\mathbf{U\Lambda}$ and $\boldsymbol{\psi}$ are proportional (at unit source strength, $\boldsymbol{\psi}$ is larger, being the unnormalized transform).

The interpolation makes the sides of the input lobe somewhat less steep, and loses the higher-frequency components. For a fifth- order polynomial interpolation of the input, $\mathbf{U\Lambda}$ becomes perceptibly narrower around zero frequency, and is no longer proportional to $\boldsymbol{\psi}$. This distortion and narrowing of the spectrum consequently broadens the image of the source.

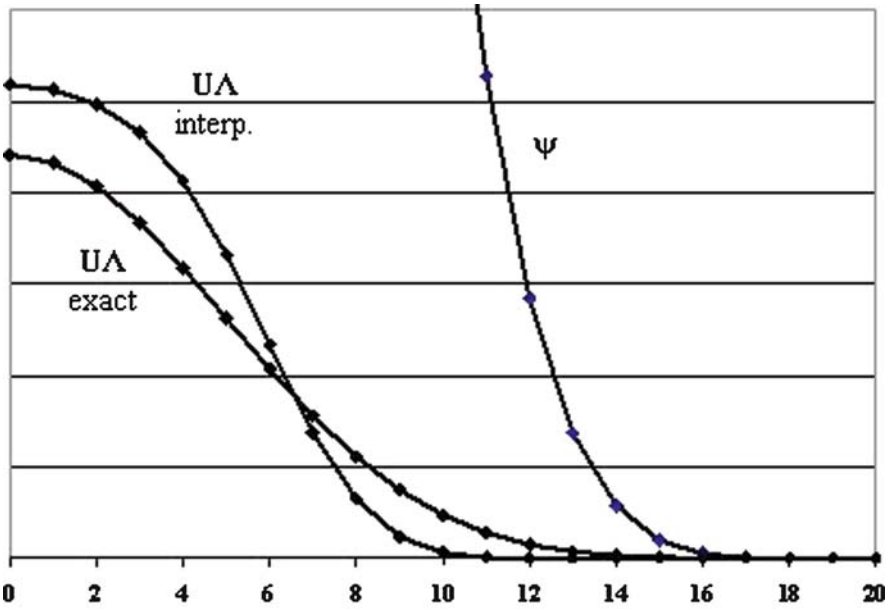

Fig. 3.5 Spectra of the basis overlap and input overlap vectors of a single source, for the basis functions $L_A(\varphi)$, at -3 dB half-width of 10.6° and $n = 360$. Only the magnitudes of the spectra are shown, on a relative scale. Abscissa is the component index (frequency). ()

3.3.3 Radio Image on the Circle

Let us now look at an example of a radio image obtained with our ring of antennas. This simulated example detects three sources at once, one at a sizable angular distance from others, and two closer together. Figure 3.6 shows the interpolated input function $\Lambda(\varphi)$, as obtained by the antennas; it also shows its basis-set approximation, i.e., the right-hand side of (3.2), with the coefficients a_i obtained from (3.22).

We see that the input function has only two peaks. The rotated lobes in the directions of the sources are added, and the sum of two identical lobes can have a minimum between the sources only if the sources are separated by at least twice the -3 dB half-width of the lobe. Since the basis lobe in this example has a -3 dB half-width of 23°, and the angular separation between the two nearby sources is 35°, the input function can no longer resolve them into separate maxima. The only (somewhat) visible indication that there are two sources hidden inside the higher peak is its slightly larger half-width.

The basis-set expansion follows the input function very accurately: the error is 1% of the value of the highest peak. The array of the expansion coefficients a_i, the image itself, is shown in Fig. 3.7. Clearly, the information about the source directions is contained in the array a_i in a better form, as the two adjacent sources are resolved into two maxima. The reason for that lies in the nature of the input function.

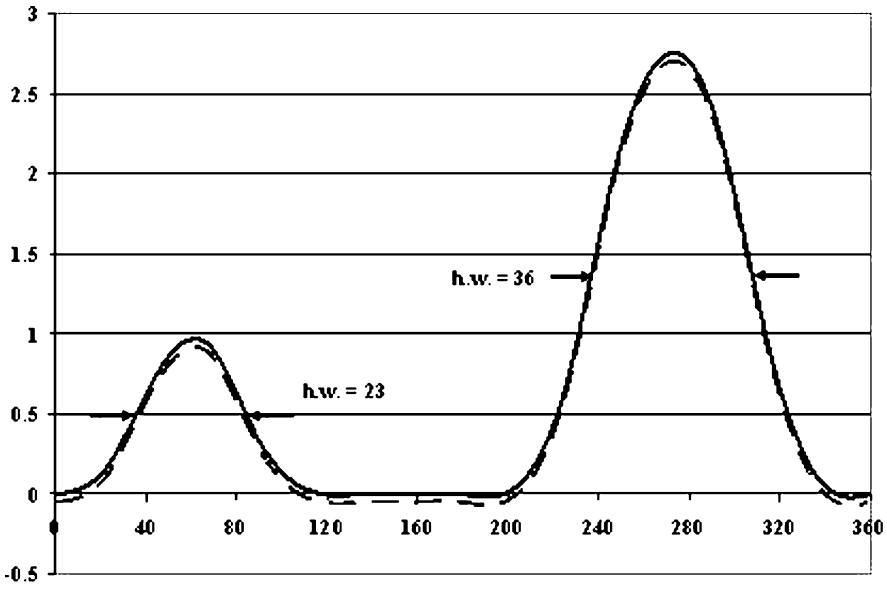

Fig. 3.6 Interpolated input function for a ring of 16 linear- array antennas, with −3 dB h.w. of 23°. Simulated input contains three sources, at directions 60°, 255° and 290°, and relative intensities of 1, 2 and 2. *Solid line* is the input function, and the *dotted line* is its basis-set expansion, offset slightly downwards for visibility ($n = 360$)

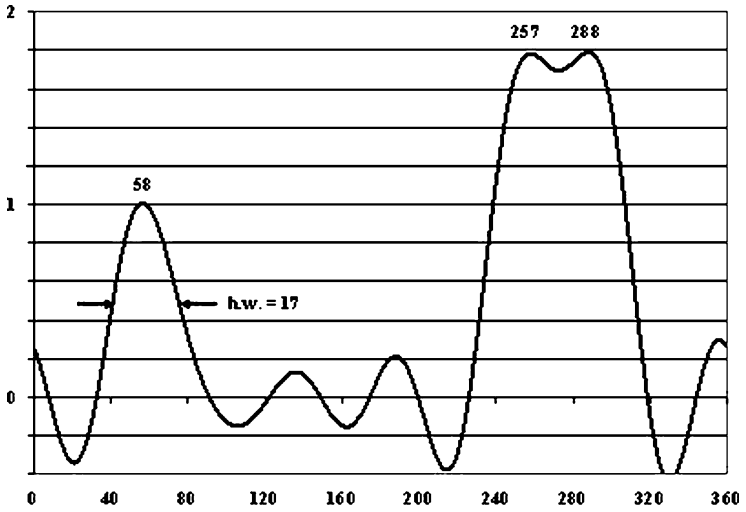

Fig. 3.7 Radio image of three sources, obtained by a ring of 16 linear-array antennas, with −3 dB h.w. of 23°. Simulated input contains three sources, at directions 60°, 255° and 290°, and relative intensities of 1, 2 and 2. Source images (*peaks*) are labeled with their angular locations ($n = 360$)

Input function obtained from the antennas is a sum of broadly overlapping, nonorthogonal rotated lobes, and as long as our antenna array operates outside of the approximation of geometrical optics, this state of affairs is unavoidable. Nonorthogonality of the constituent features of the input function precludes the use of sliding convolution mask techniques to detect these features. To illustrate this, let us suppose that the input function contains two features, of such kind that their spatially displaced instances are orthogonal under convolution:

$$\Lambda(\varphi) = f(\varphi - \varphi_1) + f(\varphi - \varphi_2) \tag{3.37}$$

Convolution of the input function with a sliding version of the feature itself picks out the locations of the two features in the input:

$$\int \mathrm{d}\varphi \; \Lambda(\varphi) f(\varphi - \psi) = \delta(\psi - \varphi_1) + \delta(\psi - \varphi_2) \tag{3.38}$$

This is, broadly speaking, the principle behind the sliding convolution masks used in image processing to detect peaks, edges etc.: the convolution has a high value at the desired feature, and nowhere else. Because the constituent features in our case are not localized, or orthogonal under displacement, the output of a sliding convolution is invariably a smudge that does not resolve nearby features.

The basis functions chosen to represent the input function carry directional information, and we have seen that the expansion obtained by error minimization in (3.3)–(3.10), is quite accurate. To understand how well the expansion reproduces the directions, we look at the expansion in a small set of broad lobes, a set for which the (3.7) can still be solved exactly, i.e., the image can be calculated using (3.22) without spectral truncation although just barely so. We show the image of four sources in Fig. 3.8, as a bar diagram that emphasizes the contributions of individual basis functions, and does not create the somewhat misleading impression of continuity in the basis index.

Images of the two sources collinear with basis functions are unequivocal delta functions, while the other two are predominantly described by combinations of two basis functions adjacent to the source's direction. A pattern of alternating contributions, decreasing with angular distance, is centered around the latter images: these are the higher-order corrections to images of noncollinear sources.

The image resolution is on the order of one to two times the angle between basis functions, and can in principle be made arbitrarily small by increasing the size of the basis set. This is the main benefit of the basis-set approach: resolution of the image a_i is not limited by the spatial overlap of the lobes, and we see that even in this crude example, it is already much better than the lobe's half-width of 36°. Of course, increasing the size of the basis set soon leads to spectral truncation and line broadening, and *that* effect is greater for broader basis lobes, as we discussed in Sect. 3.3.1.

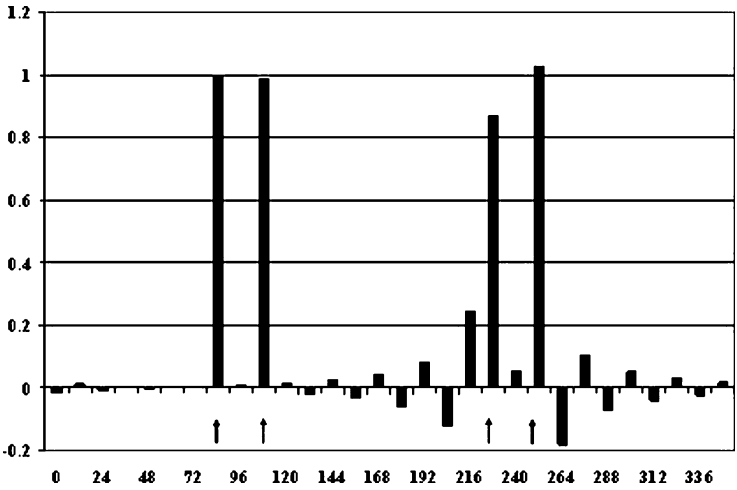

Fig. 3.8 Image of four sources, in the basis of 30 lobes, at 12° separation, with −3 dB h.w. of 36°. Sources are of unit intensity, and located at 84°, 108°, 226° and 250° (*vertical arrows*). Ideal input function is assumed, and the image is calculated without spectral cut-off

3.3.4 Peak Interactions

There is one more phenomenon in radio imaging that we want to discuss: images of independent sources do not behave fully independently of each other. To illustrate this, let us consider two sources of unit strength, one placed (conveniently) at the zero bearing, the other at some angle φ. We know from (3.27) that the spectrum of their image has the form (omitting the normalization constant)

$$\mathbf{a}_F(k) = 1 + \exp(-i\varphi k) \tag{3.39}$$

The full-spectrum inverse DFT convolves the spectrum with a series of waves of increasing frequency, and the orthogonality of the harmonic waves picks out the correct amplitude and location of the two sources, as we have seen in the example in Fig. 3.8.

Spectral cut-off complicates this picture. In Fig. 3.9, we plot the magnitude of the spectrum (3.39), which has the form

$$|\mathbf{a}_F(k)| = \sqrt{2(1 + \cos\varphi k)} \tag{3.40}$$

We see that the loss of high frequencies due to spectral cut-off varies with the angle φ, and is quite prominent around the half-width of the lobe function. This loss of higher frequencies broadens the sources' images and reduces their peak heights. We plot the peak height as a function of the angle between sources in Fig. 3.10:

At $\varphi = 0$ there is only one source, of doubled strength. As the angle reaches the −3 dB half-width (23°), the source images separate, and the peak broadens and loses height. The heights of the two distinct peaks remain equal, and oscillate slightly

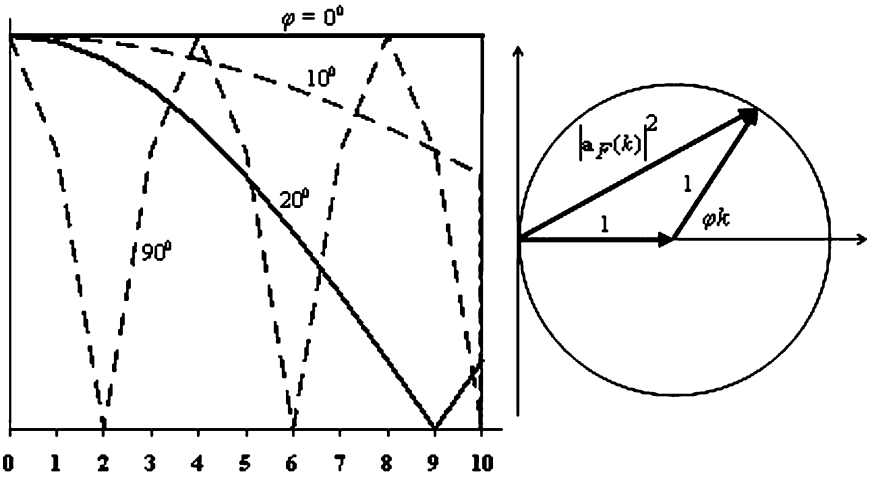

Fig. 3.9 *Left*: Magnitude spectrum of the image of two unit-strength sources, separated by the angle φ. Basis functions have the 23° half-width, $n = 360$, and the input function is of ideal precision. Abscissa is the component index (frequency), and the spectrum is truncated at $k = 10$. *Right*: phasor sum of the two sources, generating the spectrum in this image

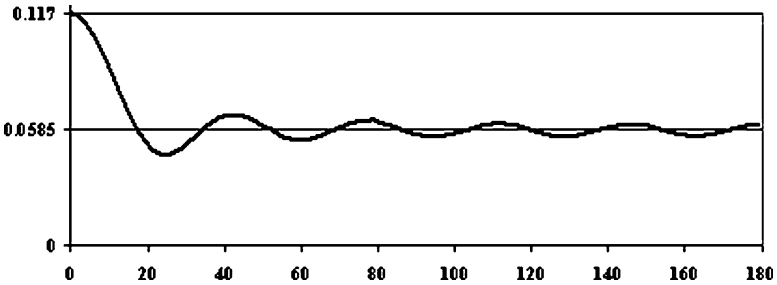

Fig. 3.10 Peak heights in the two-source image described in Fig. 3.9, as a function of the sources' angular separation (deg). Peak height is on a relative scale

around the single-source intensity. Their locations also oscillate, out of phase with each other, around their nominal positions, by a degree or two. Figure 3.9 indicates that, as the spectrum fills with periods of the function (3.40) for large values of φ, these wobbles must become less significant. They vanish in the limit of the full-spectrum calculation.

3.4 Image Formation in Spherical Geometry

In this section, we broaden the basis-set expansion methodology to spherical geometry. This is obviously desirable for practical reasons, but we must not expect facile analogies between the ring geometry and the sphere – the sphere is

substantially different. To begin with, the sphere does not offer an obvious set of discrete points, such as the equally spaced points on the ring (3.20), on which the basis-set formalism can be handled by finite algebra. It will be easier to treat our basis-set expansion as continuous.

Again, we assume that the antennas' signal strengths are a known function of all directions in space, $\mathcal{A}(\mathbf{n})$, the *input function*. As before, we seek to expand it in a basis of (axially symmetrical) lobes, a basis which now consists of a lobe in every direction in space: $L(\mathbf{n}, \mathbf{m})$ is a lobe centered on direction $\mathbf{m}$ and evaluated in direction $\mathbf{n}$ (see Fig 3.1). Rather than optimizing discrete expansion coefficients, as in (3.3)–(3.7), we postulate a complete basis set for which the following continuous expansion is valid

$$\mathcal{A}(\mathbf{n}) = \int \mathrm{d}\mathbf{m}\, a(\mathbf{m}) L(\mathbf{n}, \mathbf{m}) \tag{3.41}$$

In what follows, integrals over direction unit vectors such as $\mathbf{n}$ and $\mathbf{m}$ run over the entire surface of the unit sphere. This equation, the continuous equivalent of (3.2), is an integral representation of the input function, featuring unknown expansion coefficients $a(\mathbf{m})$. We seek an algorithm that will allow us to calculate $a(\mathbf{m})$ expeditiously and in a numerically stable manner.

We convolve the (3.41) with a lobe function in direction $\mathbf{r}$, and obtain

$$\int \mathrm{d}\mathbf{n}\, \mathcal{A}(\mathbf{n}) L(\mathbf{n}, \mathbf{r}) = \int \mathrm{d}\mathbf{m}\, a(\mathbf{m}) \int \mathrm{d}\mathbf{n}\, L(\mathbf{n}, \mathbf{r}) L(\mathbf{n}, \mathbf{m}) \tag{3.42}$$

Let us call the integral on the left-hand side of (3.42) the *input overlap* $\mathbf{\Lambda}(\mathbf{r})$, and also define the *basis overlap*:

$$S(\mathbf{r} \cdot \mathbf{m}) \equiv S(\angle(\mathbf{r}, \mathbf{m})) = \int \mathrm{d}\mathbf{n}\, L(\mathbf{n}, \mathbf{r}) L(\mathbf{n}, \mathbf{m}) \tag{3.43}$$

The basis overlap, crucially, depends again only on the angle between the two lobe directions. We write (3.42), the integral equation for $a(\mathbf{m})$, as follows:

$$\mathbf{\Lambda}(\mathbf{r}) = \int \mathrm{d}\mathbf{m}\, S(\mathbf{r} \cdot \mathbf{m}) a(\mathbf{m}) \tag{3.44}$$

The analogy with the linear system of (3.7) is obvious, and we will seek the solution in the same way, by attempting to diagonalize the overlap operator S. As in (3.15), we surmise that there must be a unitary operator, $U(\mathbf{m}, k)$, which diagonalizes the operator S by a transformation from the basis of spatial directions $\mathbf{m}$ into some new basis, tentatively parameterized by k:

$$S_T(k, k') = \int \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{m}\, U^*(\mathbf{r}, k)\, S(\mathbf{r} \cdot \mathbf{m})\, U(\mathbf{m}, k') = S_T(k)\, \delta(k - k') \tag{3.45}$$

The subscript "T" indicates a quantity transformed into the new basis. The transformation's unitary condition in the new basis is:

$$\int \mathrm{d}k\, U(\mathbf{m},k)\, U^*(\boldsymbol{\mu},k) = \delta(\mathbf{m}-\boldsymbol{\mu}) \tag{3.46}$$

In order to solve (3.44), we convolve it with $U^*(\mathbf{r},k)$, and insert a delta function under the integration sign, $\int \mathrm{d}\boldsymbol{\mu}\,\delta(\mathbf{m}-\boldsymbol{\mu})$, into the right-hand side:

$$\begin{aligned}\Lambda_T(k) &\equiv \int \mathrm{d}\mathbf{r}\, U^*(\mathbf{r},k)\, \boldsymbol{\Lambda}(\mathbf{r}) \\ &= \int \mathrm{d}\mathbf{r} \int \mathrm{d}\mathbf{m} \int \mathrm{d}\boldsymbol{\mu}\, U^*(\mathbf{r},k)\, S(\mathbf{r}\cdot\mathbf{m})\, \delta(\mathbf{m}-\boldsymbol{\mu})\, a(\boldsymbol{\mu})\end{aligned} \tag{3.47}$$

We now replace the delta with the integral in (3.46), and rearrange the integrations:

$$\Lambda_T(k) = \int \mathrm{d}k' \left[\int \mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{m}\, U^*(\mathbf{r},k)\, S(\mathbf{r}\cdot\mathbf{m})\, U(\mathbf{m},k')\right]\left[\int \mathrm{d}\boldsymbol{\mu}\, U^*(\boldsymbol{\mu},k')\, a(\boldsymbol{\mu})\right] \tag{3.48}$$

Substituting (3.45) in (3.48), we obtain the solution to (3.44) in the new basis:

$$\Lambda_T(k) = \int \mathrm{d}k'\, S_T(k')\, \delta(k-k')\, a_T(k') = S_T(k)\, a_T(k) \tag{3.49}$$

or, in complete analogy with (3.24), we write the transformed "spectrum" of the spherical image:

$$a_T(k) = \frac{1}{S_T(k)} \cdot \Lambda_T(k) \tag{3.50}$$

All that remains is to find the diagonalizing transformation U in (3.45). In order to simplify the subsequent formulas, we change the notation slightly in (3.45)

$$S_T(k,k') = \int \mathrm{d}\mathbf{n}\,\mathrm{d}\mathbf{n}'\, U^*(\mathbf{n},k)\, S(\mathbf{n}\cdot\mathbf{n}')\, U(\mathbf{n}',k') \tag{3.51}$$

and we see from Fig. 3.11 that S can be expressed as a function of a single angle θ'', if we rotate the coordinate system so that the new z-axis z'' coincides with $\mathbf{n}$. This is analogous to the reasoning that led to (3.16), and we would like to factor this coordinate rotation out of $U(\mathbf{n}',k')$. We rotate the xyz system around z, with the first Euler angle[5] equal φ, thus bringing the x-axis into the z-$\mathbf{n}$ plane, and making the y-axis perpendicular to it. Next, we tilt the z-axis to coincide with $\mathbf{n}$, by rotating around y''-axis by the second Euler angle, which equals θ. Third Euler angle can be left zero since the orientation of x- and y-axis does not matter. We have thus obtained the coordinate system $x''\ y''\ z''$.

[5] See Sect. 10.8 for the convention that we use for Euler angles.

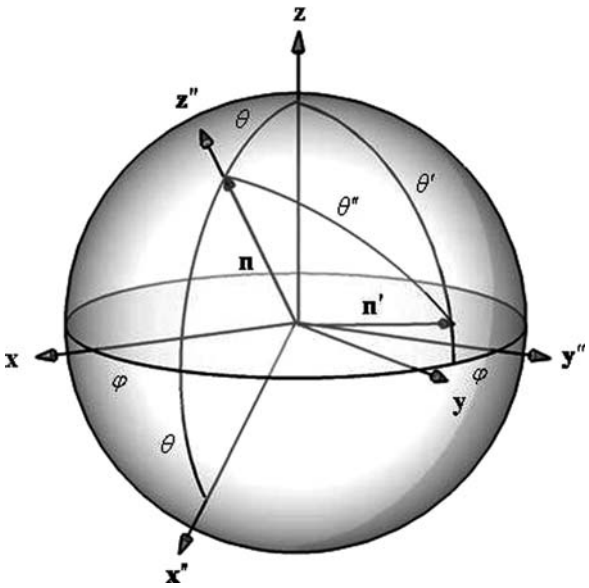

Fig. 3.11 Transformation of the coordinate system x, y, z into the doubly primed coordinate system, in which the operator $S(\mathbf{n} \cdot \mathbf{n}')$ is expressed only in terms of θ'', the inclination from the z″-axis. The transformation consists of two rotations: by φ around the z-axis, and by θ around the y″-axis. This aligns z″ with the unit vector $\mathbf{n}$

For the diagonalization transformation, we will attempt to use spherical harmonics since they are unitary and complete on the unit sphere, and have well-known rotation properties [Sect. 10.8, (10.52)–(10.54)]. Changing the parameter k into the two integer indices of spherical harmonics, we write (3.51) as

$$S_T(l, m; l', m') = \int \mathrm{d}\mathbf{n}\, \mathrm{d}\mathbf{n}'\, Y^*_{l\,m}(\mathbf{n})\, S(\mathbf{n} \cdot \mathbf{n}')\, Y_{l'\,m'}(\mathbf{n}') \tag{3.52}$$

Rotating the integration coordinate system of $\mathbf{n}'$ to $x''\ y''\ z''$ and using the rotation formula for spherical harmonics, (10.54),

$$S_T(l, m; l', m') = \sum_{\sigma=-m}^{m} \int \mathrm{d}\mathbf{n}\, Y^*_{l\,m}(\mathbf{n})\; D^{m'\,\sigma}_{l'}(\varphi, \theta, 0) \int \mathrm{d}\mathbf{n}''\, S(\theta'')\, Y_{l'\,\sigma}(\mathbf{n}'') \tag{3.53}$$

Here, $D^{m\,\sigma}_{l'}$ is the rotation matrix of the spherical harmonics. Since S does not depend on φ, only the term $\sigma = 0$ survives the integration over φ in the second integral, and we can write:

$$S_T(l, m; l', m') = \int \mathrm{d}\mathbf{n}\, Y^*_{l\,m}(\mathbf{n})\; D^{m'\,0}_{l'}(\varphi, \theta, 0) \int \mathrm{d}\mathbf{n}''\, S(\theta'')\, Y_{l'\,0}(\mathbf{n}'') \tag{3.54}$$

Equation (10.56) shows that the surviving element of the rotation matrix is a spherical harmonic of the rotation angles, which factors the unitary transformation $Y_{l'm'}(\mathbf{n}')$ under composite rotation θ, θ'', in analogy with (3.18):

$$S_T(l,m;l',m') = \sqrt{\frac{4\pi}{2l'+1}} \int \mathrm{d}\mathbf{n}\; Y^*_{lm}(\mathbf{n})\; Y_{l'm'}(\mathbf{n}) \int \mathrm{d}n''\; S(\theta'')\; Y_{l'0}(\mathbf{n}'') \tag{3.55}$$

Unitary property of spherical harmonics (10.52) accomplishes the diagonalization of S, analogously to (3.19) for the circular case:

$$\begin{aligned} S_T(l,m;l',m') &= \delta(l-l')\,\delta(m-m')\,\sqrt{\frac{4\pi}{2l+1}} \int \mathrm{d}\mathbf{n}''\; S(\theta'')\; Y_{l0}(\mathbf{n}'') \\ &= \delta(l-l')\,\delta(m-m')\; S_T(l,m) \end{aligned} \tag{3.56}$$

This is the discrete (and infinite) spectrum of S in the basis of spherical harmonics, enumerated by the indices l and m. ■

We see again that the rotational invariance of S is essential in finding its diagonalization transformation.[6] The spectrum S_T is real, as it must be, since S is a Hermitian operator (S is real and symmetric, and therefore Hermitian).

3.4.1 Radio Image on the Sphere

Returning to (3.50), the image spectrum $a_T(k)$ is given as the ratio of input overlap and basis overlap spectra, which we write using the spherical-harmonic indices:

$$a_T(l,m) = \frac{1}{S_T(l,m)} \cdot \Lambda_T(l,m) \tag{3.57}$$

Here

$$\Lambda_T(l,m) = \int \mathrm{d}\mathbf{n}\, Y^*_{lm}(\mathbf{n})\, \mathbf{\Lambda}(\mathbf{n}) \tag{3.58}$$

$$a_T(l,m) = \int \mathrm{d}\mathbf{n}\, Y^*_{lm}(\mathbf{n})\, a(\mathbf{n}) \tag{3.59}$$

By using the definition of the input overlap in (3.42) and reversing the order of integrations, (3.58) can also be written as

[6] These results can also be obtained by invoking the multipole expansion, which is the equivalent of the Fourier series, taken over spherical harmonics (see Jackson 1998). We prefer to use the matrix/operator formalism uniformly throughout this chapter, since it treats the circular and spherical geometries as similarly as possible, and it highlights nicely the role of symmetry in obtaining image spectra.

$$\Lambda_T(l,m) = \int \mathrm{d}\mathbf{m}\, \Lambda(\mathbf{m}) \int \mathrm{d}\mathbf{n}\, Y^*_{l\,m}(\mathbf{n})\, L(\mathbf{m},\mathbf{n}) \tag{3.60}$$

i.e., as convolution of the input function with the lobe's dipole moment.

Because $\mathbf{\Lambda}(\mathbf{n})$ is a real quantity, $\Lambda_T(l,m)$ satisfies the same complex-conjugation symmetry as the spherical harmonics (10.51), and since S_T is real, that symmetry applies to the image spectrum as well:

$$a_T(l,-m) = (-1)^m a^*_T(l,m) \tag{3.61}$$

We recover the image by back-transformation into the basis of spatial directions

$$a(\mathbf{n}) = \sum_{l,m} Y_{l\,m}(\mathbf{n})\, a_T(l,m) \tag{3.62}$$

Equation (3.62) derives from (3.59) and (10.53). Because of the symmetry relations (10.51) and (3.61), terms (l,m) and $(l,-m)$ in the sum in (3.62) are complex conjugates of each other, and the image $a(\mathbf{n})$ is a real quantity.

For the special case of a single point source of unit strength, in the direction $\mathbf{N}$, the input function consists of a single rotated lobe directed toward $\mathbf{N}$, and the image has the simple form $a(\mathbf{n}) = \delta(\mathbf{n}-\mathbf{N})$. According to (3.59), the image spectrum of the single-source image must be:

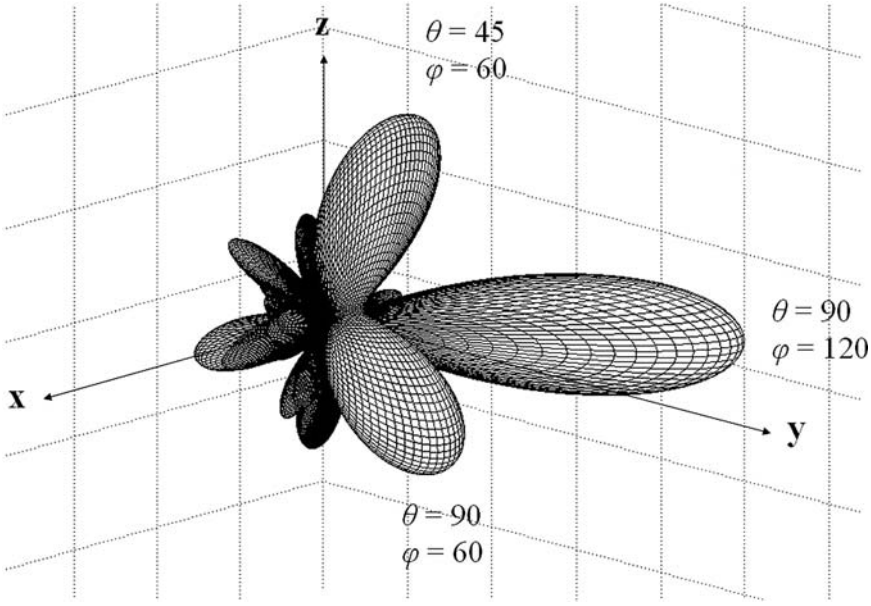

Fig. 3.12 Spherical image of three unit-strength sources, in the indicated directions. Peaks in the directions of the sources have −3 dB half-widths of 15–16°, comparable to those in the circular image, Fig. 3.7. The simulation assumes ideal input function (without interpolation), the basis lobes have −3 dB half-width of 23°, and are distributed at one-degree intervals in both θ and φ. Multipole spectrum was truncated at $l = 8$

$$a_T(l,m) = Y^*_{l\,m}(\mathbf{N}) \tag{3.63}$$

We leave it as an exercise for the reader to show that (3.57), in conjunction with (3.56) and (3.58), when applied to the single-source image, indeed reproduces the spectrum in (3.63). [Hint: re-visit the derivation of (3.56)]

For multiple sources with strengths a_j and directions $\mathbf{N}_j$, image spectrum has the form

$$a_T(l,m) = \sum_j a_j Y^*_{l\,m}(\mathbf{N}_j) \tag{3.64}$$

This equation is analogous to (3.27) for the circular geometry, and the back-transformation (3.62) reconstructs the sources' strengths and bearings by virtue of (10.53). As in the circular case, the ideal spectrum yields a crisp image, independent of the shape of the basis functions. Of course, the spectrum must be obtained by means of (3.57), as a ratio of quantities whose magnitude decreases with the increase in spherical indices (equivalent of higher wave frequency in the Fourier transform). That means that the spectrum cannot be calculated accurately beyond some index l, and must be truncated.

Figure 3.12 shows a simulated spherical image, obtained by (3.57)–(3.62). The image peaks are again narrower than the basis lobes, and can be made arbitrarily narrow in principle. Spectral truncation leads to background oscillations, which can be seen on the back side of the image, as well as to peak interactions, in very close analogy with the circular geometry.

References

Antolovic, D. An Algorithm for Simultaneous Radiolocation of Multiple Sources, Proceedings of 2009 IEEE International Conference on Portable Information Devices, Anchorage, AK, September 2009, (2009)

Byron, F.W., Fuller, R.W. Mathematics of Classical and Quantum Physics, Dover, New York (1992)

Davis, P.J. Circulant Matrices, Wiley-Interscience, New York (1979)

Gantmacher, F.R. The Theory of Matrices, Chelsea, New York (1959)

Golub, G.H., Van Loan, C.F. Matrix Computations, 3rd edition, Johns Hopkins University Press, MD (1996)

Gradshteyn, I.S., Ryzhik, I.M. Table of Integrals, Series and Products, 7th edition, Academic Press, New York (2007)

Gray, R.M. Toeplitz and Circulant Matrices: A Review, now Publishers, The Netherlands, http://www- ee.stanford.edu/∼gray/toeplitz.pdf (2006)

Hamming, R.W. Digital Filters, 3rd edition, Dover, New York (1998)

Jackson, J.D. Classical Electrodynamics, 3rd edition, Wiley, New York (1998)

Messiah, A. Quantum Mechanics, Dover, New York (1999)

Morse, P.M., Feshbach, H. Methods of Theoretical Physics, McGraw- Hill, New York (1953)

Smith, S.W. The Scientist and Engineer's Guide to Digital Signal Processing, California Technical Publishing, CA (1997)

Chapter 4
Radiolocator Design: High-Frequency Front End

4.1 Design Requirements and General Architecture

We have seen in Chaps. 2 and 3 that the direction of an incoming wave can be reconstructed from the interaction of that wave with an assembly of directional antennas of known characteristics. In this chapter, we begin to look at the engineering aspects of collecting data from the antennas, and, in the spirit of good design practice, we will outline the system requirements and consider some plausible architectures.

The overarching requirement in our entire radiolocation work is that the direction of the incoming wave be obtained for every wireless packet, and that the direction be assigned to that packet before the next one arrives, i.e., at the rate of wireless traffic. Two requirements follow:

(1) Signal strengths on all antennas must be captured during the passage of the packet. Antenna signals must be captured consistently, that is, they must all be measured at the same (average) carrier amplitude.
(2) Logical content of the packet must be decoded as usual, i.e., networking communication must not be impeded by radiolocation.

Broadly speaking, there are two possible architectures for data collection: parallel and serial. With regard to the first requirement, parallel capture of the signal strengths is faster, its time limited effectively to the time required for one power measurement. This affords flexibility in the timing of radiolocation within the passage of the packet. On the down side, parallel architecture leads to greater system sprawl, and because of the replication of components, to more stringent requirements for uniformity and more difficult calibration.

Serial architecture is physically more compact, and, since only one measurement path is involved, it is more easily calibrated. The uniformity requirements are much less severe. On the other hand, the overall measurement time is proportional to the number of antenna elements, and substantial intervals of constant amplitude in the packet's waveform are required. Beyond a certain number of antennas, and for amplitude-modulated waveforms, serial architecture may not be feasible.

D. Antolovic, *Radiolocation in Ubiquitous Wireless Communication*,
DOI 10.1007/978-1-4419-1632-7_4,

Regarding the second requirement, we should make it clear that both radiolocation and content-decoding use the baseband signal; we discuss this point further in Sect. 4.1.1. Serial architecture does not allow the same baseband signal to be used for both radiolocation and data decoding because sequential switching of antennas distorts the signal to such an extent that data loss is inevitable (see Sect. 6.2). This architecture, therefore, requires a separate, independent path for data communication.

In the parallel architecture, there are multiple basebands, all containing the same data. While these basebands are all valid and potentially decodable, their amplitudes vary greatly. In this architecture, we face the issue of selecting, on a per-packet basis, the highest-amplitude baseband to extract the logical content from.

Let us make all of this more concrete by means of diagrams. Figure 4.1 shows a fully expanded parallel architecture, and we see that the sprawl is substantial. In the lower, radiolocation branch, tuning, and measurements are done in parallel, and the digitized results are gathered serially from analog-to-digital converters (ADCs) connected to a bus; most ADCs will latch the digital result until they are instructed

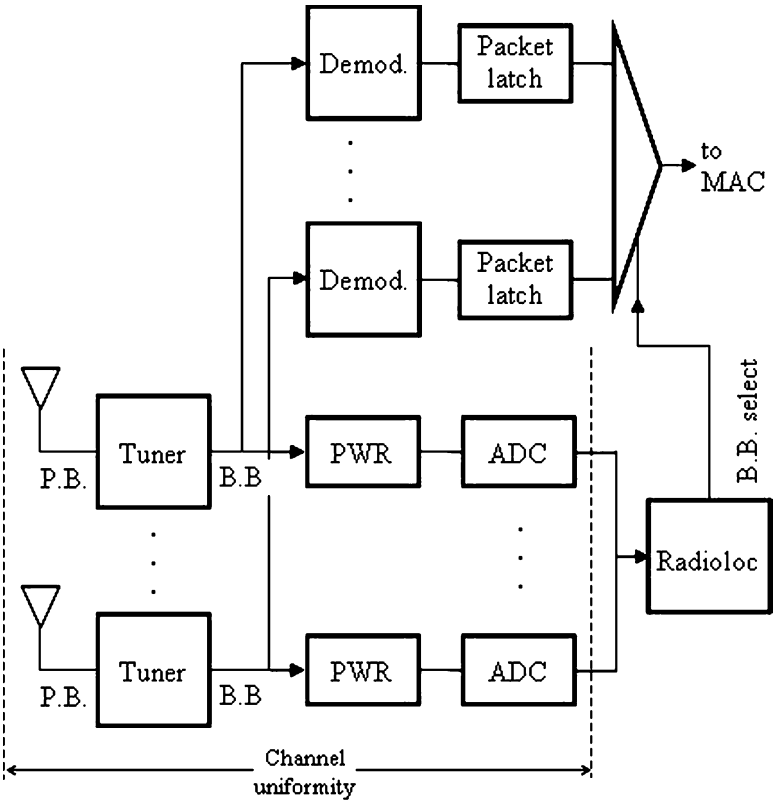

Fig. 4.1 Parallel radiolocator architecture

to perform the next conversion. Multiplexing the analog outputs from the power meters to a single ADC would be possible, but probably not very cost-effective.

In the upper, communication branch, every baseband signal is demodulated, and the packet payload is stored. Even though the packets are nominally identical, we can expect the error rate to be higher for low amplitudes. We can use the radiolocation data to select the version of the packet that was received at the highest signal strength.

It would be desirable to know the basebands' strengths early, so as to select the right one for demodulation and eliminate multiple demodulators. Unfortunately, demodulators rely on the very early baud intervals of the waveform to synchronize with the packet transmission, and power meters are slow compared to baud rates.[1] In addition, digitization takes some time as well. Therefore, by the time the baseband strengths are known, it will be too late to switch basebands without compromising the demodulation. The only way to reduce component replication in the communication branch is to have a separate, omnidirectional antenna, followed by the normal receiver path.

Serial architecture (Fig. 4.2) is considerably simpler, but, as we said before, its baseband is absolutely unsuitable for demodulation, and a separate communication path is essential. We will discuss the integration of these two paths at length in Chap. 6.

In this discussion on architecture, our envisioned architectural elements so far have been physical parts. It is possible to implement parts of this system in software, as the so-called software-defined radio (SDR); so let us discuss the ramifications of that choice.

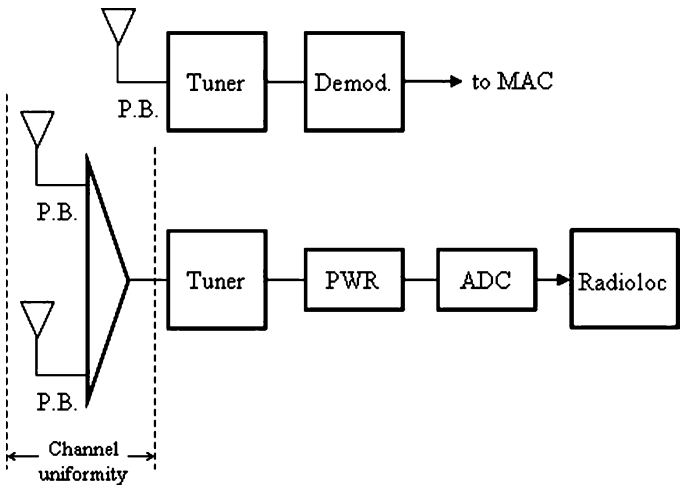

Fig. 4.2 Serial radiolocator architecture

[1] For example, in the 802.11b protocol, the physical preamble (the very first bits of a packet) runs at 1 Mbaud (1 μs per bit), while a reasonably fast power meter settles in 6–7 μs.

The passbands of interest lie in the ranges of frequencies between 2 and 5 GHz; these waveforms are clearly not digitizable, due to impractically high sampling rates that are required, and due to the volume of data that would be generated. The digitization can only occur after the (analog) step of downconversion, and whether the subsequent steps of tuning and amplification are implemented in hardware or software is immaterial for our discussion. If the baseband signals in Fig. 4.1 are digital data streams, some additional flexibility is gained. Signal strength of each baseband will still have to be measured (calculated) independently, but multiple baseband processing in the communication path will be avoided. This is simply due to the fact that digital basebands can be delayed in storage long enough to detect the strongest one, and then proceed with demodulating that baseband only. Unfortunately, the SDR approach does nothing to solve the problem of baseband distortions in the serial architecture.

As is always the case with SDR, one must be mindful of the data volume. Assuming the baud rate of 10 Mbit/s, sampling rate of 5–10 times the baud rate, and 16 antennas, all of that yields 800–1,600 Mbit/s of digitized basebands to process. It is obvious that an SDR approach would also require rather specialized hardware for its implementation.

In selecting an architecture for the radiolocator for the 802.11b standard, we tried and ruled out the parallel option as too bulky, costly, and difficult to calibrate. In addition, there is no amplitude modulation in the 802.11b standard, so that there is ample time for serial sampling even in the shortest packets. SDR implementation has no distinct advantage in the serial architecture, and the required SDR data rates were too high for off-the-shelf communication links like RS232 and USB in any case. We settled on a hardware-based serial architecture, whose implementation we shall describe in this and the following chapter.

4.1.1 Radiolocation and the Receiver's Signal Path

Our discussion in the previous section has quickly led to a choice of radiolocator architecture, but it will be an instructive detour to take a look at the typical radio receiver's signal path from the viewpoint of radiolocation. To that end, we show in Fig. 4.3 a fairly generic two-stage heterodyne receiver, and we discuss the suitability of its signals for radiolocation measurements.

First, we know from Chap. 2 that the antenna is the radiolocator's sensor, and its characteristics must be fully known. Any unintended reception, such as from a second antenna or an exposed trace, will degrade the radiolocation accuracy. The receiver is allowed to see only one antenna at a time, that antenna must be well-characterized, and the antenna-to-receiver circuitry must be protected from RF interferences.

In our methodology, radiolocation measurements are consistently performed on the baseband signal. Even though it is feasible to measure the power of passband-frequency signals, tuning to the desired radio channel occurs immediately after the

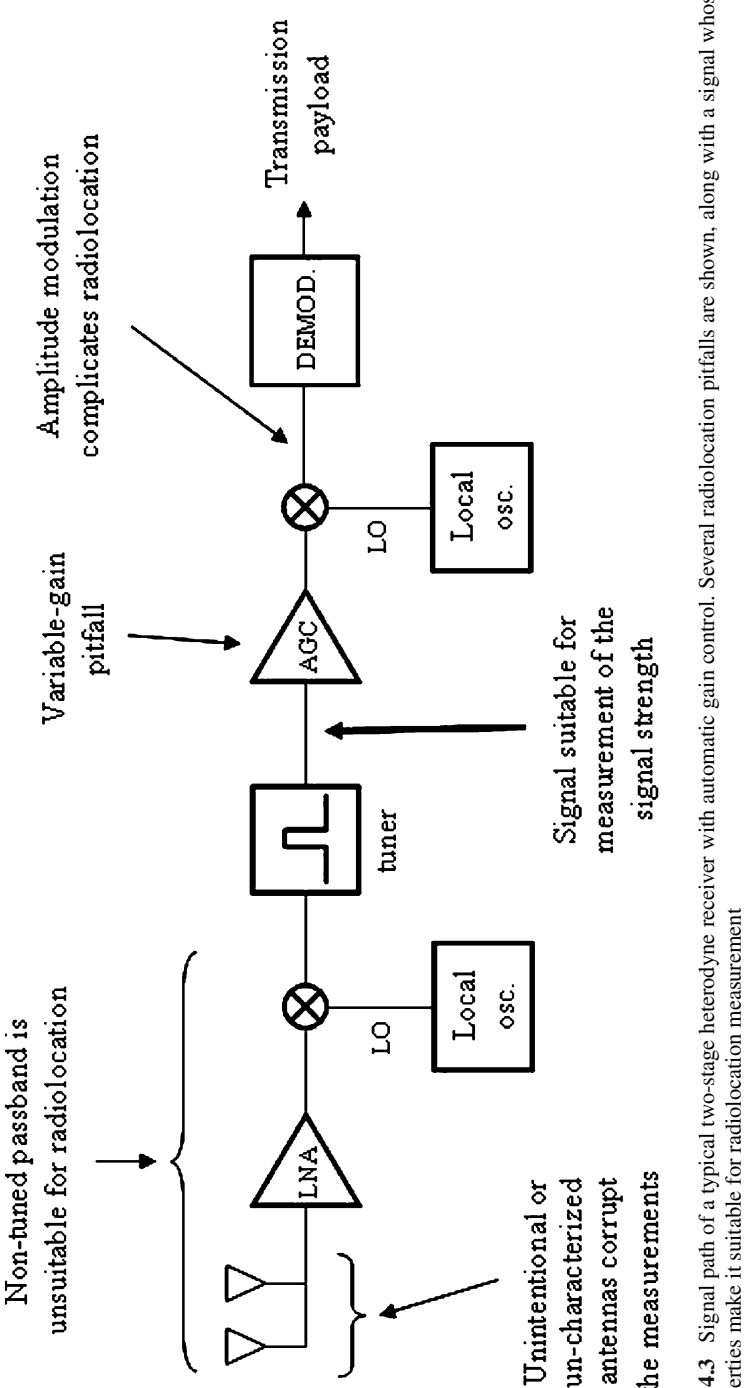

Fig. 4.3 Signal path of a typical two-stage heterodyne receiver with automatic gain control. Several radiolocation pitfalls are shown, along with a signal whose properties make it suitable for radiolocation measurement

downconversion, as shown in Fig. 4.3; as we discuss in Sect. 10.3, direct tuning of the passband places impractical demands on the quality of the tuning filter. Consequently, the passband of most receivers is wide open, receiving the signals from any and all sources, and radiolocation would be meaningless at that stage. In some encoding schemes (CDMA), radio transmissions are distinguished from one another by different chipping sequences, rather than by the passband frequency alone, and the concept of "tuning" acquires a somewhat different meaning; we discuss this detail in Sect. 9.7.

An important engineering requirement, specific to radiolocation, is that the gain along the signal path be held constant. This requirement is obvious for amplitude-based radiolocation methods, but it is easily forgotten that most commercial receivers include the automatic gain control (AGC). This feature preferentially enhances the signals from weak transmitters, and is very desirable in the communication pathway; but in a radiolocator, it blots out the crucial variations in the strength of the baseband signal, and diminishes the accuracy of radiolocation.

Finally, if amplitude modulation is present, it can corrupt the radiolocator's measurements, and it requires design considerations that circumvent the problem. We will return to this issue in the chapters that follow.

4.2 Front End of the Serial Architecture

As the Fig. 4.2 shows, front-end parallelism in the serial architecture extends from the antennas to the root of the antenna selector. Unlike in the parallel architecture, this path is short, and contains no active (gain-generating) components. Main things to consider will be the uniformity of the channels (in terms of geometry and impedance), the channel losses, and the isolation between channels. We start with the antennas themselves.

4.2.1 Directional Antenna Elements

We chose the helical design for the directional antenna elements, mostly for the practical reasons of manufacturability and cost. The prototype helical element has six turns, with the turn circumference equal to the wavelength λ, and the step equal to $\lambda/4$; a square ground plane with side length of λ is attached at the feed end (Fig. 4.4). Here $\lambda = 12\,\text{cm}$, the wavelength of the 802.11b band. The established design formulas for the helical antenna (see American Radio Relay League 2000) yield a gain[2] of 13 dBi at the boresight, and the null-to-null lobe width of 94°; the

[2] Antennas have no active gain. What is meant by antenna gain is the maximum power per solid angle (i.e., radiation in the direction of the main lobe), relative to that of an isotropic radiator radiating the same total power. dBi stands for "decibels over isotropic."

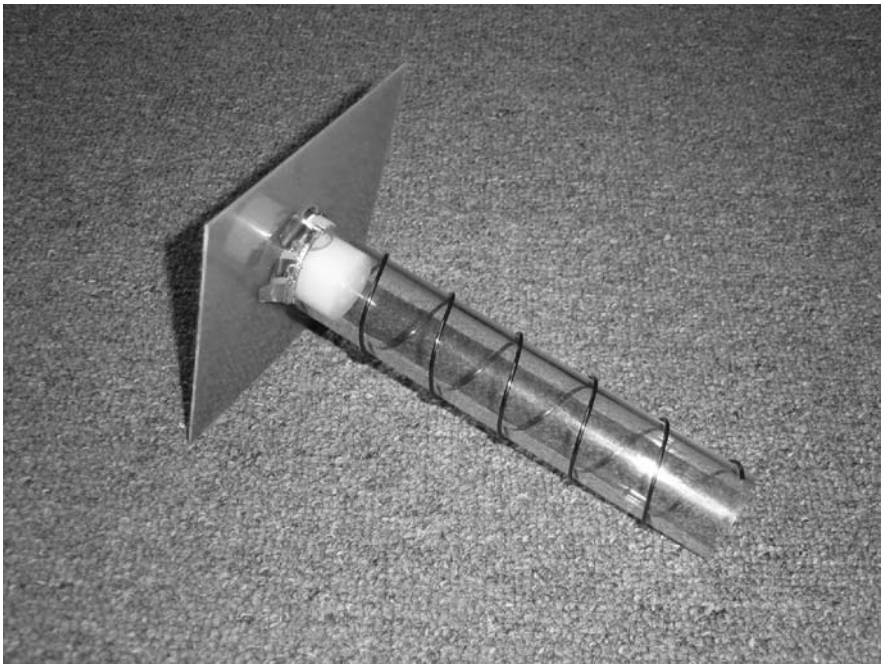

Fig. 4.4 Directional antenna element: six-turn helix with ground plane and quarter-wave transformer flaps

radiation pattern of this antenna is shown in Fig. 2.20. The inherent impedance at the feed of the helix is ca. 140 Ω, and our antennas are equipped with copper flaps which serve as a quarter-wave transformer, to bring the impedance to the standard 50 Ω. By adjusting the flaps, we can reduce the reflected power by −20 dB relative to input, fairly uniformly for all elements.

The lobe width of this antenna element matches well with our analysis in Chap. 2, but the helical design is by no means essential: patches or horns of similar lobe width would be just as suitable (Antolovic 2009). Our helical antennas do not have a significant noise floor, or any variation in it, and the variations in the gain were controlled by uniform fabrication and tuning. The antennas were calibrated for the remaining small variations in gain, which were tabulated and compensated for in software.

4.2.2 Design of the Radio-Frequency Multiplexer

It is fair to say that the passband-frequency multiplexer, which samples the antennas during the passage of a packet, constitutes the most crucial component of the serial architecture. The multiplexer is where the parallel channels are brought together, and this must be done in a manner that is compatible with the quantitative nature

of the task – we emphasize again that radiolocation is a matter of quantitative measurements, not of communication. Let us spell out the design requirements of the multiplexer:

(1) The device must conduct the passband-frequency signal with only modest losses. This is to avoid having to place active elements (amplifiers) in the parallel paths, since these could introduce non-uniform noise floors.
(2) Channels (that is, channel losses) must be uniform enough for the variations to be calibrated and compensated reasonably easily downstream.
(3) Cross-channel isolation must have at least the same decibel span as the dynamic range of the rest of the instrument. If that is not the case, cross-channel leaks will be added to the valid signal at high signal intensities, and this will corrupt the measurement results.
(4) Switching times should be reasonably fast compared to the sampling times, so as not to slow down the sampling process.
(5) Ports of the device must have the standard 50 Ω impedances to match the other components of the system.

Two plausible multiplexer topologies are conceivable: binary tree of switches, and a bus (Fig. 4.5). We have chosen the binary tree, because the path geometry can be made more uniform than in the bus topology, at least if we are restricted to two dimensions. Also, direct merging of traces in the bus topology leads to impedance discontinuities and reflection losses. The disadvantage of the tree topology is that each path leads through several switches, increasing the loss.

Board layout of our 16-fold multiplexer is shown in Fig. 4.6: it is a binary tree with four layers of switches. The obvious benefit of this geometric arrangement is that all the paths through the multiplexer are equally long, keeping the path losses comparable. A further benefit is that the points at which the impedance is discontinuous, that is the chip pads and cable connectors, have similar shape and surroundings, and, therefore, lead to comparable losses.

The board has two RF layers: top layer with traces and a coplanar reference plane surrounding them, and the second layer, which is an uninterrupted RF reference plane; the dielectric is FR-4. Via fields and fences surround the traces and part mounts, forming mutually isolated waveguides. Cross-section of the waveguide is shown in Fig. 4.7: this board structure is usually referred to as the coplanar waveguide with ground (CWG), and it behaves somewhat like a hollow metallic waveguide.

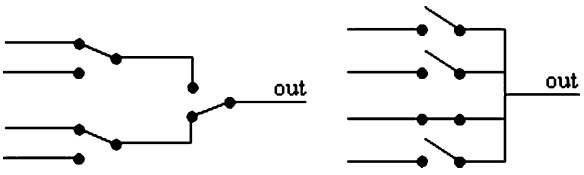

Fig. 4.5 Plausible multiplexing topologies: tree (*left*) and bus (*right*)

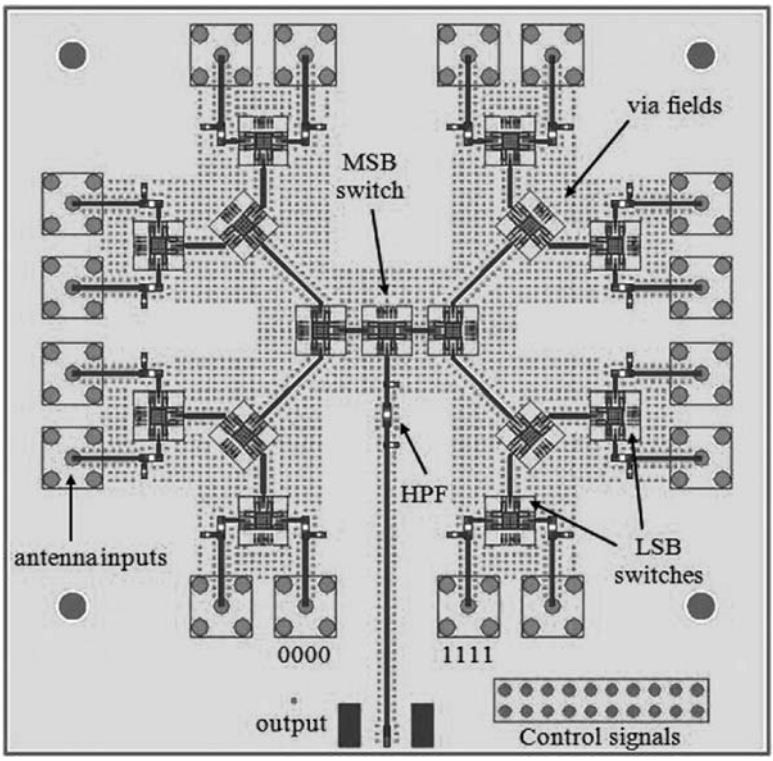

Fig. 4.6 Layout of the multiplexer circuit (© IEEE 2007, reproduced with permission.)

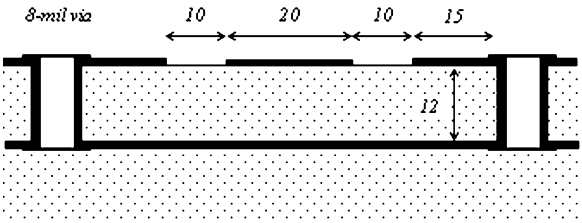

Fig. 4.7 Perpendicular cross-section of the RF waveguide. The structure is made of copper traces and the FR-4 dielectric. Dimensions are in thousandths of an inch (mils)

The multiplexer is constructed from double-throw RF switches, parts PE4257 by Peregrine Semiconductor Corp. [2005]. We chose these switches because their bandwidth exceeds 3 GHz, they have good isolation when open, small insertion loss, and an adequate switching speed. Furthermore, unlike many RF switches of their vintage, they work well with a single +3.3 V power supply, which eliminates the need for a bipolar power supply to run the switches.

The circuit has a few auxiliary features worth mentioning. Antenna inputs are coupled to the first layer of switches by 100 pF capacitors, and to the ground by RF chokes; all switch-control lines have chokes on them to prevent RF leakage into the digital circuitry.

Low-frequency voltage changes on the digital control lines of the switches leak into the RF waveguides through the switches. To keep this switching noise from corrupting the RF signal, output from the final layer of the multiplexer passes through a high-pass L-C-L Butterworth filter (see e.g., Besser and Gilmore 2003; Horowitz and Hill 1989). For the RF band and the switching speeds of interest, the adequate cut-off frequency of the filter is ca. 1.5 GHz.

In the next few subsections, we present a quantitative analysis of this design.[3] While every design is unique, and our results will not be readily transferred elsewhere, we find it worthwhile to illustrate the level of quantitative analysis that we found was necessary for building a radiolocation instrument.

4.2.2.1 Trace Impedance and Reflections

Characteristics of the CWG are well-studied. We designed the waveguide with the geometry shown in Fig. 4.7 to conform to the standard impedance, using known analytical formulas for CWG (Wadell 1991). We then verified the design by modeling it with Mentor Graphics' HyperLynx software (Mentor Graphics); Hyper-Lynx calculates the cross-sectional distribution of the electromagnetic field around the trace, from which it obtains the impedance, attenuation etc. Finally, we compared these results to the analytical results for the microstrip of analogous dimensions (IPC Association 2003; Bogatin 2004). The comparison is shown in Table 4.1, indicating that our waveguide design conforms well to the standard 50 Ω value.

We also wanted to estimate the losses that are caused by the impedance discontinuities. These discontinuities partially reflect the electromagnetic wave back to its source, and although undesirable, they are inevitably associated with any irregularity in the geometry of a transmission line: pads and chip contacts, turns in the traces etc. Section 10.4 offers a summary of the concepts relevant to the transmission lines.

We performed return loss measurements on all channels of the multiplexer, by using a standard set-up of a signal generator, a return-loss bridge, and a spectrum analyzer (see Fig. 4.8). At the measurement frequency of 2.462 GHz, and with its

Table 4.1 Modeled impedance of the waveguide traces

HyperLynx model	46.7 Ω
CWG analytical formula	49.8 Ω
Microstrip analytical formula	52.8 Ω

[3] Experimental results in the following three sections are from (Antolovic and Wallace 2007), © IEEE 2007, reproduced with permission.

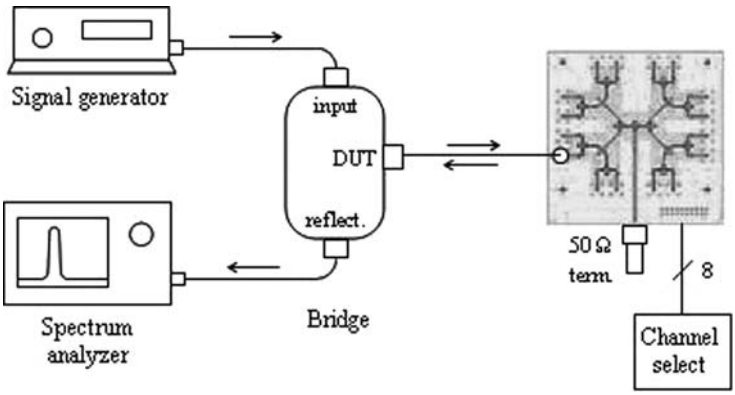

Fig. 4.8 Experimental set-up for the measurement of return loss in the RF multiplexer

output terminated in 50 Ω, multiplexer's attenuation of the reflected power ranged between −9.6 and −8.5 dB, relative to the input power. On the linear scale, that amounts to an 11–13% of power reflected back from the multiplexer inputs.

This average return loss of 12% translates into a reflection coefficient $\Gamma = 0.34$, or the SWR number of 2. Quite understandably, a path through the multiplexer is not the best of transmission lines, due to its various impedance discontinuities, but the retention of 88% of the input power means that the reflection attenuates the signal power by only 0.5 dB, which is entirely acceptable.

4.2.2.2 Path Attenuation

In this section, we look at the dissipative losses along the signal's path through the multiplexer. First, we performed isolation and attenuation measurements on the switch PE4257, to verify the manufacturer's specifications. We used the manufacturer's evaluation board (Peregrine Semiconductor Corp. 2005), and the results in Table 4.2 show a good agreement with the specifications.

We then assessed the losses in RF waveguide traces in two ways. We modeled the waveguide structures in Fig. 4.7 with HyperLynx, and compared the results to those of the microstrip configuration. The CWG is rather similar to the microstrip, and there exists a body of well-known formulas for the analysis of the latter (see Bogatin 2004). Table 4.3 shows the attenuation per unit length of trace, divided into conductive and dielectric components, as well as other board losses, and a comparison with the measured overall path loss. We see that the resistive loss in the conductor is relatively small, even though the skin depth for copper at 2.4 GHz is only 1.6 μm. Dielectric losses are greater, which can be attributed to the loss tangent 0.02 of the dielectric FR-4.

Overall, agreement between modeling, analytical formulas, and measurements in Table 4.3 is good, and the discrepancy can be attributed to experimental error. These results justify the use of the inexpensive FR-4 as the board's dielectric substrate,

Table 4.2 Properties of the switch PE4257, at 2.5 GHz

Isolation, input-to-output, measured	−45 dB
Isolation, input-to-input, measured	−49 dB
Attenuation, measured	−1.3 dB
Attenuation, from data sheet	−1.1 dB

Table 4.3 Modeled and measured attenuation in waveguide traces, at 2.5 GHz

Method:	Conductive (dB/in)	Dielectric (dB/in)	Full trace (dB)	Switches (dB)	Return loss (dB)	Full path (dB)
HyperLynx	0.076	0.16	0.74	5.2	0.5	6.44
Microstrip formulas	0.037	0.24	0.87			6.57
Measured						6.3–6.8

Total trace length is 3.14 in. in every branch of the multiplexer

even though much better high-frequency dielectrics are available (Konsowski and Gipprich 2000). At the total trace length of 3.14 in. in every branch, trace attenuation is much smaller than the attenuation due to the four ICs along the path through the multiplexer. This is, of course, a fairly large prototype design, and miniaturization in commercial applications can only improve this trade-off.

We may add here that the coaxial cables connecting the antenna elements with the multiplexer (Belden 8240, 24 in. long) introduce an attenuation of ca. 0.5–0.7 dB.

4.2.2.3 Crosstalk

We measured the crosstalk between the channels by applying a 2.462 GHz, 1-mW carrier signal to one input, selecting other channels in turn, and measuring the leakage of the aggressor signal into the selected channel on the multiplexer output. The number of possible cross-measurements is large (256 combinations), but the typical result is ca. 30 dB cross-channel attenuation for channels separated by a single open switch, and 40 dB or more otherwise.

We recall from Table 4.2, that a single open switch introduces isolation of −45 dB, which is more than the isolation seen on the board. The increased leakage is most likely due to the crosstalk between traces, which is also consistent with greater (but not doubled) isolation for the distance of two open switches – the aggressor signal is simply kept further away from the selected channel.

The crosstalk measurements were supplemented with HyperLynx' crosstalk analysis of the board's layout. Setting the aggressor voltage on a RF trace to 10 V, and rise time to 0.1 ns,[4] yields induced voltages below 10 mV on any other trace, RF, or digital. This amounts to a coupling below 0.1%, or −30 dB, for RF signals, consistent with the measurements above.

[4] Frequency spectrum of a signal with rise time of 0.1 ns has significant components up to 5 GHz (see Johnson and Graham 1993).

In the radiolocator, the multiplexer stage is followed by a homodyne receiver (MAX2820 by Maxim, see Sect. 4.3), whose linear dynamic range was found to be ca. 35 dB in laboratory tests. The level of crosstalk in the multiplexer is, therefore, too low to cause distortions in radiolocation results.

4.2.2.4 Wave Propagation in RF Traces

Because of the presence of two dielectrics, FR-4 and air, the coplanar waveguide (CWG) does not allow pure transverse electromagnetic (TEM) propagation; its main propagation mode is sometimes described as quasi-TEM. TEM mode has the appealing property of allowing wave propagation at any frequency; it resembles wave propagation in free space in that respect (see Pozar 1998, Sects. 3.8–3.11, also Sect. 10.4).

For the waveguide geometry in Fig. 4.7, analytical description (Wadell 1991) yields an effective dielectric constant[5] $\varepsilon_{\text{eff}} = 3.31$, which is closer in value to $\varepsilon \approx 4.5$ for FR-4 than it is to $\varepsilon = 1$ for air. This indicates that the field energy is greater in the dielectric than in the air above, and that the main propagation mode resembles TEM reasonably well.

Our waveguide does also allow undesirable radiative modes and standing waves. The latter can exist only if the largest dimension of the guide is at least equal one half-wavelength. The long output trace of the multiplexer, which would be difficult to avoid in the fully planar layout, comes close to the half-wavelength in FR-4. This trace could support a mode that would radiate outward through the slots in the guide.

In order to verify whether this is happening, we can view the output trace approximately as a very short rectangular waveguide opening upward, or as a leaky resonant cavity. In either case, the lowest frequency forms a standing half-wave along the length of the trace. The electric and magnetic fields must have the characteristic sine/cosine variation along the trace, and the power flux from the slots has the shape $P_{10} \approx \sin^2(\pi x/l)$, where l is the length and x is the position along the trace (see Pozar 1998, Sect. 3.3).

We measured the fields along the output trace (from the connector to the high-pass filter), by sliding small probes across the soldermask surface between the waveguide slots. We measured the H field with a wire loop 3 mm in diameter, and the E field with a Hertz dipole 8 mm long. Result for H_y (magnetic field perpendicular to the board) is shown in Fig. 4.9, as power measured by a loop probe parallel to the board. The expected power distribution for this field in the radiative mode is $P \approx \cos^2(\pi x/l)$, but the measured distribution is flat within the experimental error. Fields H_x (parallel to the board) and E_y (perpendicular), which are strong in the TEM mode, show a slight standing-wave distribution along the trace, corresponding to SWR = 1.1. We conclude that this trace operates in a clean (quasi) TEM mode.

[5] Effective dielectric constant is a simplification used in analyzing structures containing multiple dielectrics; it replaces the spatial distribution of actual dielectrics with a uniform medium of an average dielectric constant, ε_{eff} (see Wadell 1991).

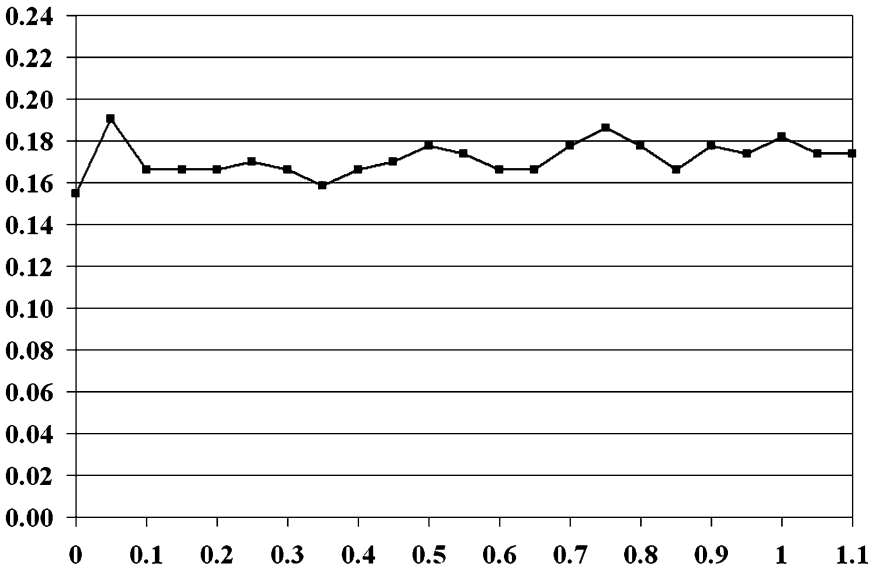

Fig. 4.9 Detected radiated power along the output trace of the RF multiplexer. Abscissa is the distance (in.) from the output connector in Fig. 4.6; ordinate shows the probe's power output (μW), i.e., the squared magnetic field strength on a relative scale. Test source was a 2.462 GHz, 1 mW signal (© IEEE 2007, reproduced with permission.)

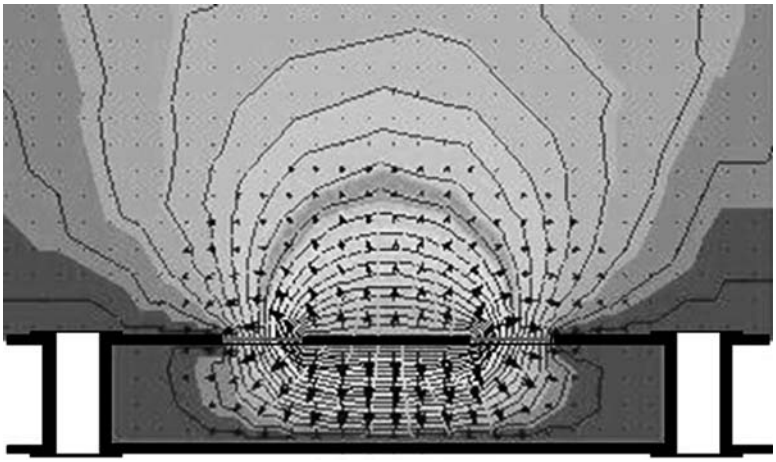

Fig. 4.10 Simulated cross-sectional distribution of the transverse electric field (*arrows*) and potential (*solid lines* and *colors*) around the waveguide trace (© IEEE 2007, reproduced with permission.)

Finally, we show the cross-sectional distribution of the amplitude of the transverse electric field and potential, which, in the TEM propagation mode, is equivalent to a two-dimensional electrostatic field (Fig. 4.10, generated with QuickField

software[6]). We see that the waveguide functions mostly like a microstrip, with high field intensity between the trace and the Layer 2 ground. The coplanar ground plane limits the lateral extent of the field, and the field in the air diminishes rapidly beyond the height of ca. three dielectric thicknesses. This field distribution does not, in itself, confirm the propagation mode, since the field simulation is two-dimensional. It is plausible, however, in the light of the above measurements.

4.3 Radiolocator's Tuner

As was discussed in Sect. 4.1, serial radiolocator architecture needs a downconversion and tuning stage to limit the received signal to the desired frequency channel. For our radiolocator prototype, we chose the integrated 802.11b transceiver MAX2820, by Maxim Integrated Products [2003]. This device has the architecture of a standard homodyne transceiver (we use the receiver side only), operating at passband frequencies of 2.4–2.5 GHz, and at the channel bandwidth of ca. 20 MHz.

The transceiver has an internal phase-locked loop, which generates the mixing frequency in the passband range, and which uses an external sinusoidal oscillator as reference; the mixing frequency is controlled digitally. In fact, the device has a digital control interface consisting of control signals and configuration registers; the latter are accessed through a synchronous serial link. In the course of the operation, the radiolocator's control circuits use this interface to change the wireless channels, and to turn the tuner on or off as needed.

The gain of the baseband amplifier is controlled by an external analog voltage, designed to be set by an external AGC circuit in standard applications. In the radiolocator, this gain is constant, and can be set manually by a potentiometer.

We investigated the dynamic range of the radiolocator's receiver section in considerable detail; a typical plot of the baseband versus passband power is shown in Fig. 4.11. The curves exhibit the usual noise floor and saturation, with reasonably good linearity in between. As expected, the noise floor rises with the gain of the tuner, restricting the dynamic range somewhat.

The reference oscillator for the tuner is the part MAX2620 (see Maxim Integrated Products 2002), driven by a quartz crystal with the resonant frequency of 22.11 MHz. Some care was needed in designing the board layout of the oscillator circuit: the oscillator has short feed lines to the tuner, it has its own reference plane, and its power supply and ground are isolated from the rest of the board with appropriate RF chokes and filters (see also Chap. 8 for additional layout details).

[6] QuickField is a two-dimensional field-solver program by Tera Analysis. See http://www.quickfield.com/.

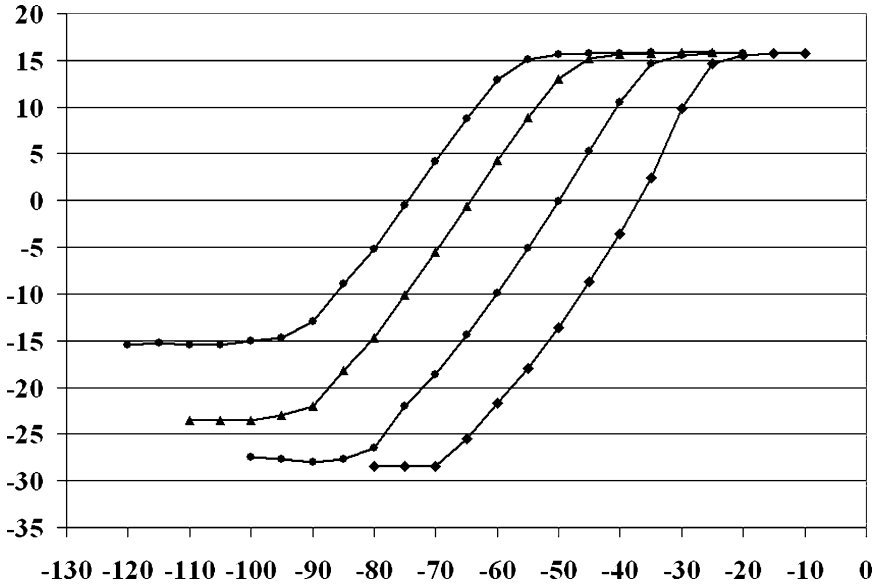

Fig. 4.11 Dynamic response of the radiolocator's tuner circuit. Baseband output power at 1 MHz is plotted as a function of input power at 2.463 GHz, in dBm units. Curves correspond to different settings of the gain control, from low (*right*) to high (*left*); input range −80 to −35 dBm (second curve from the right) was most useful in prototype tests

References

Antolovic, D., Wallace, S. Design of a Radio-Frequency Multiplexer, Used in Radiolocation of 802.11 Wireless Sources, Proceedings of IEEE International Conference on Portable Information Devices, Orlando, FL, March 2007, (2007)

Antolovic, D. Numerical Investigation of Algorithms for Multi-Antenna Radiolocation, Proceedings of 2009 IEEE International Conference on Portable Information Devices, Anchorage, AK, September 2009, (2009)

American Radio Relay League, ARRL Antenna Book, American Radio Relay League, Newington, CT (2000)

Besser, L., Gilmore, R. Practical RF Circuit Design for Modern Wireless Systems, Artech, Boston (2003)

Bogatin, E. Signal Integrity Simplified, Prentice Hall, NJ (2004)

Horowitz, P., Hill, W. The Art of Electronics, Second Edition, Cambridge University Press, Cambridge (1989)

IPC Association. IPC-2251, Design Guide for the Packaging of High-Speed Electronic Circuits, IPC Association, Northbrook, IL (2003)

Johnson, H., Graham, M. High-Speed Digital Design, Prentice Hall, NJ (1993)

Konsowski, S.G., Gipprich, J.W. Substrates for RF and Microwave Systems, in High Performance Printed Circuit Boards, C.A. Harper ed., McGraw-Hill, Boston, MA (2000)

Maxim Integrated Products. Data sheet for MAX2620, http://www.maxim-ic.com/ (2002)

Maxim Integrated Products. Data sheet for MAX2820, http://www.maxim-ic.com/ (2003)

Mentor Graphics, http://www.mentor.com/products/pcb/expedition/analysis_verification/hyperlynx/index.cfm
Peregrine Semiconductor Corp. Data sheet for PE4257, http://www.peregrine-semi.com/ (2005)
Pozar, D.M. Microwave Engineering, Wiley, New York (1998)
Wadell, B.C. Transmission Line Design Handbook, Artech, Boston (1991)

Chapter 5
Radiolocator Design: Power Measurement and Digital Data Path

5.1 Design Requirements

In this chapter, we continue to follow the radiolocation signal through the data-collection stages: to power measurement and digitization of the power level, then onto gathering, formatting, and communication of radiolocation data to the supporting systems.

We are now at the stage where the signal has been downconverted to baseband frequency, and filtered to the desired radio channel. Provided that the received signal belongs to a valid wireless packet, a topic which we will take up in Chap. 6, we can be sure that we are detecting the transmission from one networking source, and one source only, and that radiolocation on the basis of this signal will give meaningful results.

Our power-measurement stage has, in its general design, a fairly typical real-time data-collection architecture: a sensor, in this case an integrated circuit measuring the average power of an AC signal, produces an analog reading of the sensory input after a fixed, known measurement time. The reading is digitized, also in real time, and raw data are grouped into meaningful units, in this case, readings of the same packet from all antennas, and sent off for higher-level processing. This higher-level step is the calculation of the packet's direction of arrival (bearing); we discussed the algorithm in Chap. 2, and we will return to this algorithm's timing aspects in Chap. 6.

Looking back at Figs. 4.1 and 4.2, the power measurement step is essentially the same for serial and parallel architectures. As we have discussed in Sect. 4.1, a crucial difference is that multiple paths in the parallel architecture require extensive calibration to achieve path uniformity. Another issue is the time required for the power meter to listen to the signal before yielding a stable reading. Shorter time is of course preferable, but the power measurement is inherently an averaging process that can not be made arbitrarily short. We found the time requirements of the power meter to be the main time-limiting requirement in our serial architecture.

The input range of the power measurement should match the dynamic range of the tuner's output; it is more important, however, to cover the low end of the range than the high end, since few transmitters are ever detected at the saturation level of

D. Antolovic, *Radiolocation in Ubiquitous Wireless Communication*,
DOI 10.1007/978-1-4419-1632-7_5,

a receiver. If they are, they are likely to be prominent enough in the environment to be known nuisances, rather than sources in need of localization!

More significantly, dynamic ranges of the meter and the analog-to-digital converter (ADC) need to be matched. Unless the parts form a chipset, and are meant to be used together, their voltage ranges will be fairly unrelated, and the usable overlap between the meter's output and the ADC's input may be small. Furthermore, power characteristics of the two parts must also be matched.

We will discuss all of this in more specific terms in the course of this chapter, as we continue to describe our prototype radiolocator.

Precision of the measurement result can be expressed as the number of stable output bits from the ADC. Obviously, we should not allow inadequate precision of the converter to limit the accuracy of the bearing result, but there are other precision-limiting factors in the data path, and we should choose the converter's precision accordingly. In our prototype, one such factor was the accuracy to which the antenna's lobes were known.

We offer the schematic overview of the prototype radiolocator in Fig. 5.1 (Antolovic and Wallace 2006), and in the sections that follow, we will discuss the stages of its data path in detail.

5.2 Power Meter at the Heart of Radiolocation

Power meter used in our prototype is the part AD8362, by Analog Devices. We chose this part primarily for its input dynamic range of ca. 60 dBm, covering the range of tuner's baseband rather well. In comparison, a related part, AD8361, has a range of only about 30 dBm. The chosen meter is also tolerant of waveform crest factors[1] up to 14 dB, and operates at the supply voltage of 5 V, which turned out to be convenient for practical design reasons. Its input is differential, like the baseband output of the tuner, and there is no need for conversion between one-sided and differential signal forms.

This part measures the mean-square amplitude of the input waveform, i.e., the average power, and expresses the result as a dc voltage proportional to the logarithm of the root-mean-square input power. Figure 5.2 shows the output voltage, as a fairly tightly linear function of the input power in dBm (Pilotte 2005). Basic structure of the device is shown in the manufacturer's diagram in Fig. 5.3: details of the design are obviously proprietary, but the principle of operation is easily understood.

The device operates as an automatic gain control (AGC) loop: it subjects the input signal to an adjustable gain, calculates the property of interest, and compares the result to an internal reference. It uses the error signal (the measure of deviation from the reference) in a damped negative feedback, adjusting the gain of the input

[1] Crest factor of a waveform is defined as the ratio of the peak to the rms value, and serves as a measure of "spiked-ness" of the waveform.

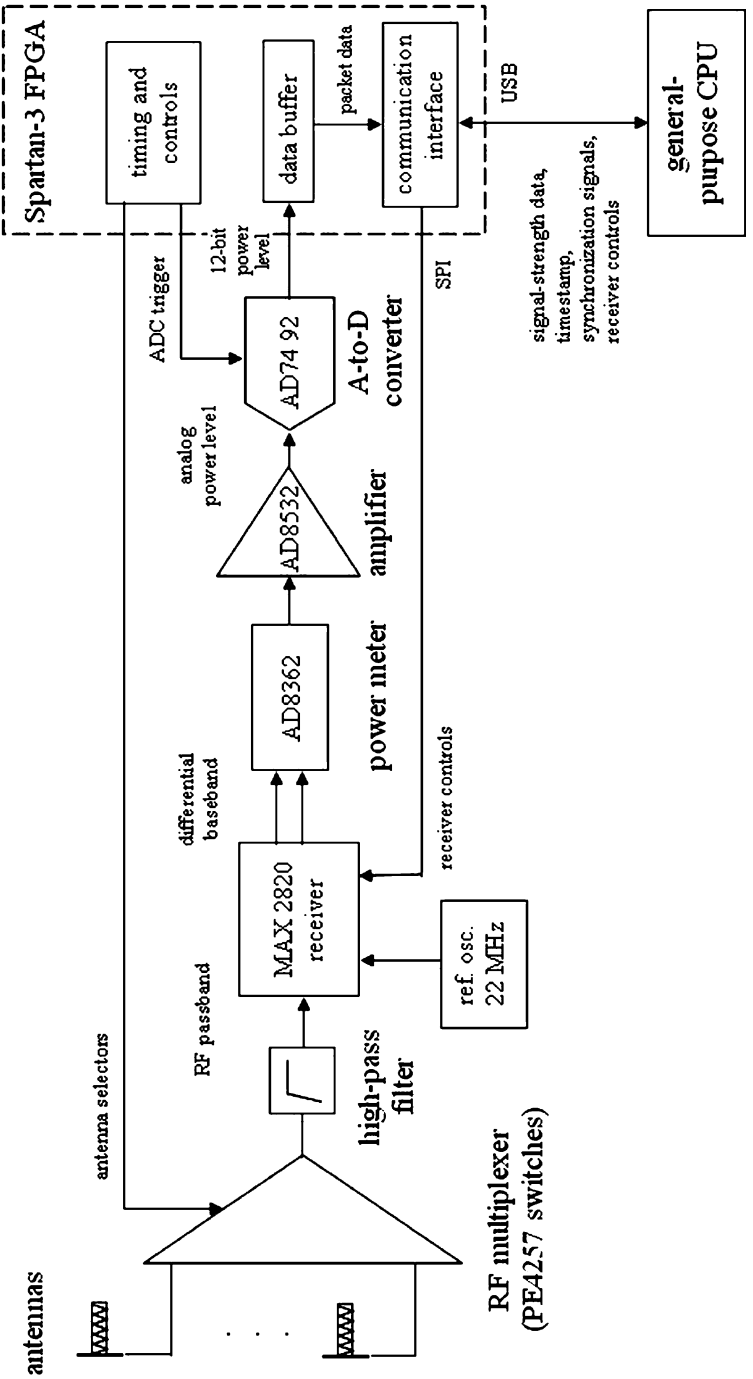

Fig. 5.1 Block diagram of the radiolocator prototype (© 2006 ACM, reproduced with permission)

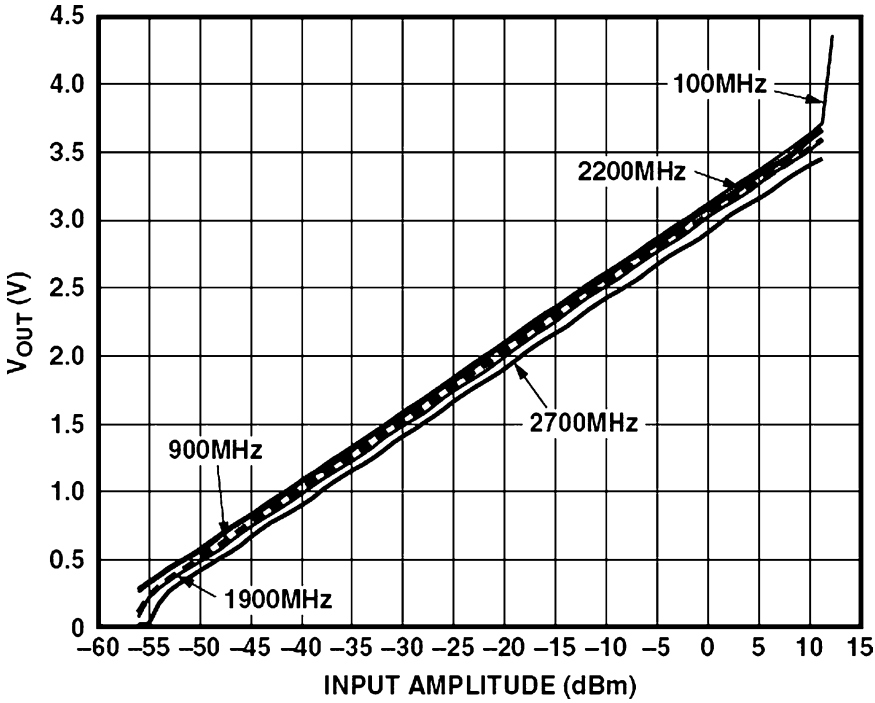

Fig. 5.2 AD8362 log slope for various frequencies (© Analog Devices Inc. 2005, reproduced with permission)

amplifier, in order to drive the error signal to zero. Once the loop stabilizes, the gain control signal serves as the measure of the input's property.

In the part AD8362, a variable-gain input amplifier, labeled VGA in the diagram Fig. 5.3, sends the amplified input signal to an analog squaring circuit, whose output is a current, I_{SQU}. This current is proportional to the square of the input's amplitude, and, therefore, to the input power. An accurate internal reference voltage (V_{REF}) is applied to the pin VTGT, and translated by a second, identical squaring circuit into a target, or reference current, I_{TGT}. The two squares serve as a current's source and a sink, respectively, and the resulting current difference, the error signal, is integrated (averaged) by charging the capacitor C_{LPF}. The voltage on this capacitor, V_{OUT}, is the averaged error signal, which is fed back into the input V_{SET}, adjusting the gain of the input amplifier VGA. V_{OUT} is also the meter's output, which reports the power of the input signal, as shown in Fig. 5.2. The logarithmic power scale derives from the exponential characteristic of the VGA, i.e., the amplifier's gain is an exponential function of V_{SET}.

This level of description is sufficient for our discussion. For a detailed (and authoritative) account of the meter's operation and characteristics, we refer the reader to the manufacturer's documentation (Analog Devices Inc. 2003).

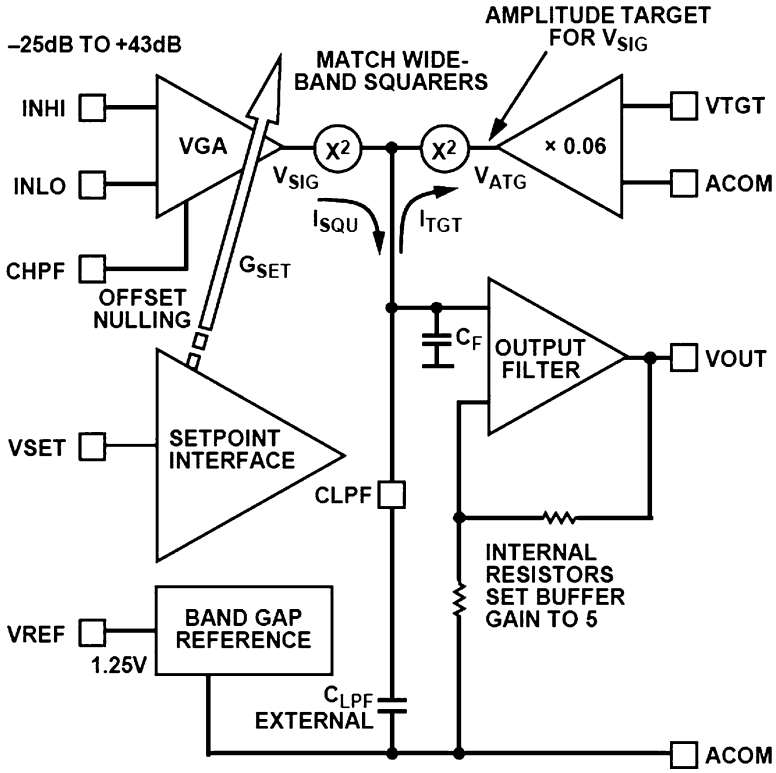

Fig. 5.3 Basic structure of the AD8362 (Data sheet for AD8362, Rev. D; © Analog Devices Inc. 2003, reproduced with permission)

Power-meter's main engineering trade-off concerns the measurement time. The 802.11b modulation (DBPSK and DQPSK) does not, in principle, alter the carrier's amplitude, only its phase, and the signal's power is constant in the first approximation. It is obvious, however, that the modulation phase shifts must have some effect on the signal power, an effect which we would expect to average out over several baud periods. We explore this point in additional detail in Sect. 10.5.

As can be seen from Fig. 5.3, the integrating capacitor C_{LPF}, which is external to the power meter chip, determines the meter's settling time. Very small (pF) values of C_{LPF} lead to excessive modulation ripple on the output, and large ones (in the μF range) result in settling times that are impractically long. Some trial and error led us to adopt the value $C_{\mathrm{LPF}} = 0.47\,\mathrm{nF}$; this yields rise times of ca. 4 μs, fall times of ca. 6.5 μs, and a modulation ripple in the power reading, which did not exceed 100 mV, from peak to peak.

The meter does not indicate the completion of its measurement: there is no "done" signal, and the completion is for the circuit designer to decide upon. We found that performing a reading of the meter's output at the very end of the antenna sampling period is adequate in all cases.

5.3 Digitization of the Power Measurements

For the digitization stage of the prototype radiolocator, we chose the analog-to-digital converter AD7492 by Analog Devices. This part converts its input voltage, ranging from zero to 2.5 V, to a 12-bit number. The choice of the ADC is not critical, but this part has several convenient features. It is a sample-and-hold ADC, implementing the successive approximation algorithm (see e.g., Horowitz and Hill 1989), with internal clocking. This makes the digital interface very simple: the falling edge of a sampling clock initiates the conversion, and at the end of conversion, the ADC de-asserts a "busy" output signal, indicating that the digital value is ready. That value is latched on the output pins until the next conversion. The part has a special power supply, dedicated to feeding its digital outputs, so that their voltage can be matched to the voltage of the data-receiving circuits (in this case, the 3.3 V inputs to an FPGA).

The AD7492 has a short acquisition time of ca. 120 ns. This is important because, as we have seen in Sect. 5.2, much of the sampling time of a single antenna element is taken by the settling of the power meter. The acquisition, during which the ADC input is actually read, happens at the end of that interval, just before the antennas are switched over. The data conversion time is 880 ns, a time interval which is completed several comfortable microseconds before the next antenna sampling is initiated (see Analog Devices 2001).

There were two mismatches on the interface between the power meter and the ADC that needed to be resolved. First, the tuner's output range of -30 to $+10$ dBm (see Fig. 4.11) is mapped into the power meter's output from 1.5 to 3.5 V (Fig. 5.2). Since the ADC's input range is zero to 2.5 V, direct overlap is only 1 V wide, which amounts to a significant loss of range. These two ranges must be linearly mapped onto each other.

Second, the meter's output can not charge the sampling capacitor of the ADC directly – doing so results in ADC's input voltage decaying over tens of microseconds, yielding meaningless conversion results. One may say that the meter's output has too high source impedance, and a conversion interface is needed between these two ICs.

Mapping the two ranges is easily accomplished with an operational amplifier, configured as inverting amplifier with bias (see the schematic in Fig. 5.4). The voltages in this circuit are governed by the equation

$$y = -x \cdot G + V_{\mathrm{B}}(1 + G), \tag{5.1}$$

where the gain and the bias of the first stage are given by:

$$G = \frac{R_{\mathrm{L}}}{R_{\mathrm{IN}}} \tag{5.2}$$

$$V_{\mathrm{B}} = V_{\mathrm{CC}} \cdot \frac{R_2}{R_1 + R_2} \tag{5.3}$$

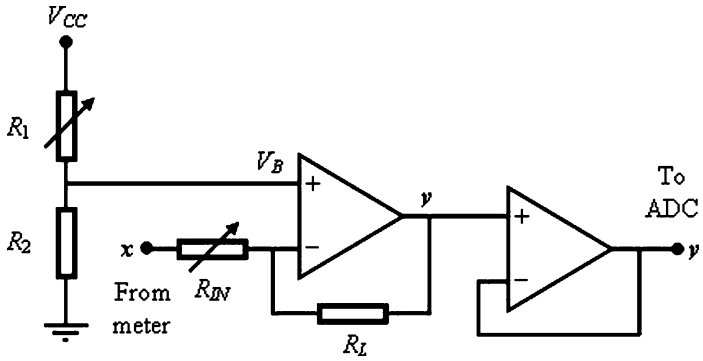

Fig. 5.4 Amplifier stages matching the dynamic ranges of the power meter and the ADC

From these equations, and the above-quoted ranges for the meter's output x, and the ADC's input y, appropriate resistor values can be easily derived; we made R_1 and R_{IN} variable, so that the mapping can be adjusted as needed. This mapping inverts the high and low values, a small drawback which is easily corrected by subsequent processing in code. Because the ADC's range is lower than the meter's, non-inverted mapping would require a negative power supply for the op-amp, an engineering headache that we tried to avoid.

The second op-amp in this circuit is a voltage follower, whose only role is to fully isolate the meter's and ADC's currents, so as to prevent unwanted discharge of the ADC's sampling capacitor; this second stage proved essential for our choice of components. This entire circuit was implemented with a single IC, the dual, single-supply op-amp AD8532. Finally, for the theory of operational amplifiers, we refer the reader to standard textbooks (e.g., Horowitz and Hill 1989; Jung 1997).

Our tests indicated that the 12-bit ADC yields 8 bits of accuracy reliably, possibly up to 10. For reasons of design simplicity, especially with respect to communication outside the radiolocator board itself, we limited the accuracy to eight bits. An erroneous ninth bit amounts to 0.2% of error in the maximum signal; compared to other sources of error in the system, that is very acceptable. For example, measurements of the antenna radiation pattern involved a turntable that could be adjusted to (realistically) 0.5°, or an error of 0.14%; the spectrum analyzer used for these measurements was accurate to ca. 0.1 dBm, which in a 20 dBm signal amounts to a 0.5% error, and so on. Doubling the number of data bytes that would have had to be moved around, in order to accommodate another bit or two in accuracy, was not a very cost-effective proposition.

5.4 Data Collection Cycle

Timing diagrams in Fig. 5.5 summarize the process of antenna switching and sampling. Shaded area in the first line is the physical preamble of a wireless packet; in the 802.11b standard, this preamble is 144 μs long, counting from the very start

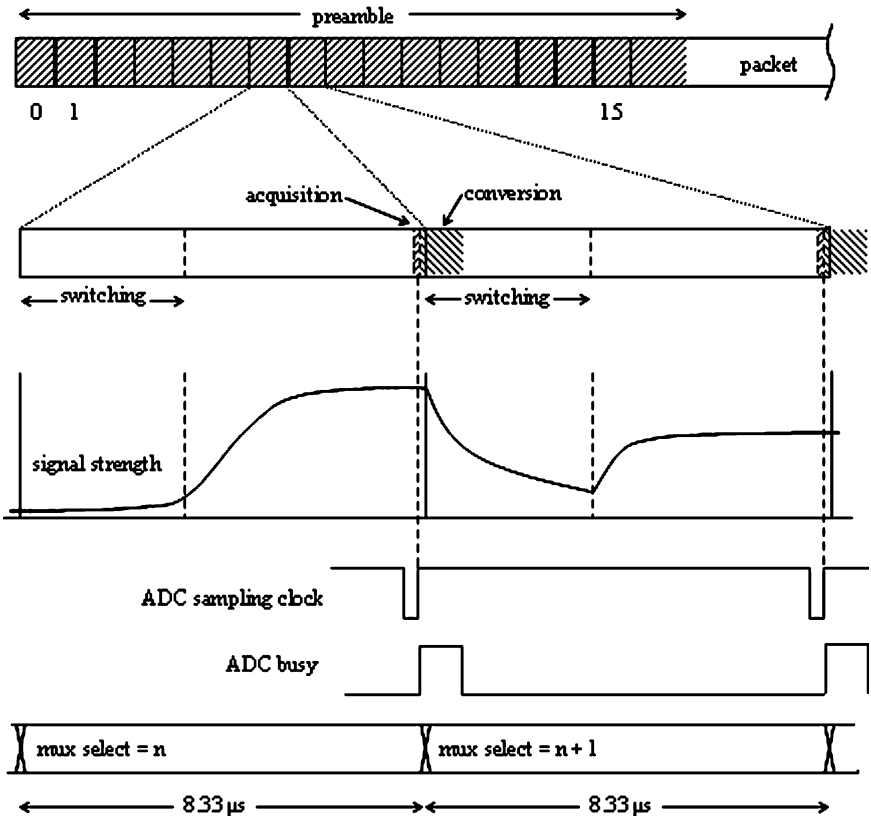

Fig. 5.5 Timing diagram of the data collection cycle

of the carrier signal, and consists of synchronization sequences. Since the presence of the preamble is guaranteed, we fit the sampling of all antennas into its duration; this is in fact somewhat more restrictive than necessary, because even very short management packets in 802.11b are not shorter than ca. 300 µs, and the signal's amplitude is constant throughout.

The 16 sampling intervals are overlaid on the packet's preamble. Antenna switching is done at 120 KHz, or at 8.33 µs per antenna, and all antennas are sampled in 133 µs. We show two adjacent sampling intervals in the second line of Fig. 5.5. We should note here that the time intervals in Fig. 5.5 are drawn roughly to scale, with the time scale being obviously denser in the first line of the diagram.

The initial 3.4 µs of every sampling interval are taken up by RF switching. This is the period after the instantaneous (6 ns) switch-over of the multiplexer's control signals, during which the RF switches fully close and open. During that period, the signal strength of the passband drops rapidly from its current level, then rises to the level corresponding to the next selected antenna.

In our design of the power-meter circuit (see Sect. 5.2), 4 μs is an adequate time to settle the rising output of the meter. We show a typical shape of that output in the third line of the diagram: since the amplitude of the RF signal is generally low during the switching period, most of the meter's rise happens afterwards.

On the falling side, about 6.5 μs is enough time for the meter to zero out after any power measurement. Since the RF signal is low during switching, the switching time counts as downward settling time for the meter: whether the next power level is higher or lower than the previous one, there is enough time in the sampling interval to settle the measurement.

The ADC's acquisition interval of 120 ns is placed close to the end of the sampling interval. This is the time immediately preceding the rise of the sampling clock, during which the ADC's input must remain steady, to allow the charge (i.e., voltage) on the sampling capacitor to stabilize. The rising edge of the sampling clock occurs some 50 to 100 ns before the next switching of the multiplexer, and the conversion interval spills over into the next sampling interval, which is immaterial. We see from the second line of Fig. 5.5 that the data conversion, the digitization, ends long before the next sampling occurs.

Digitized signal strengths are transferred from the ADC, via a 12-bit bus, into a Spartan-3 FPGA (made by Xilinx), running at the clock rate of 50 MHz. This FPGA (Field-Programmable Gate Array) is the central controller of the radiolocation device. Use of FPGA-implemented circuits as controllers in real-time data collection has become fairly common since configurable chips were first introduced, and with good reason. Short of using custom-designed and custom-manufactured integrated circuits, FPGAs are ideal for handling large numbers of coordinated signals in real time. In our device, eight multiplexer controls must be generated in synchrony with the ADC sampling clock, and the resulting data must be collected with reliable timing. Even if we had chosen a microcontroller fast enough not to introduce excessive delays into the sequence of events, the large number of control signals to maintain would have strained a microcontroller-based design.

The FPGA-implemented circuit, with its many general-purpose pins, and as many concurrent processes as we choose to design, with its outputs' strict adherence to the clock, and absence of an instruction cycle, takes this coordination task in stride. Apart from coordinating the hard real-time data collection described in Fig. 5.5, the circuit also handles the off-board communication, and implements a small instruction set that allows outside processes to change radio channels, switch between the reception and transmission modes etc. We will return to this controller circuit repeatedly throughout our discussion.

Data formatting is straightforward: 16 bytes of signal strengths (radiolocation data) are followed by an 8-byte timestamp. This timestamp is under the control of the data communication path, as we shall see in Chap. 6, and is used to synchronize the radiolocation data in the non-real-time stages of processing. Boundaries between packets are marked by 16-bit synchronization words, which indicate to the receiving CPU the beginning of a new set of packet data. Optionally, the CPU can request that the radiolocator send a sync word alone, for example, in order to re-synchronize the communication.

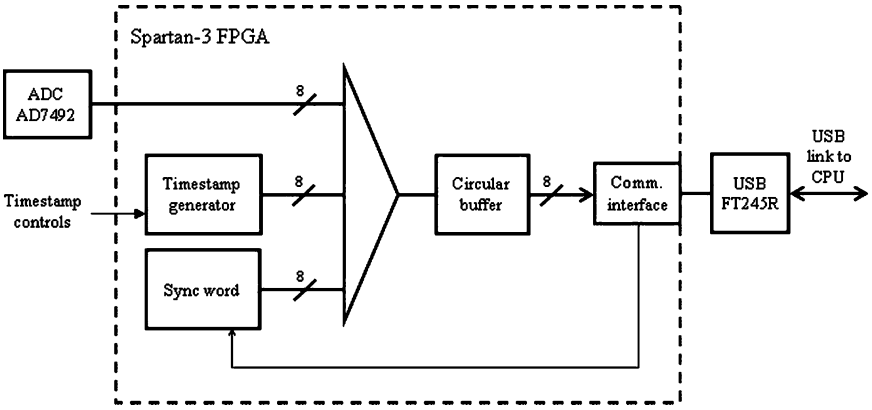

Fig. 5.6 Digital part of the radiolocation data path

Radiolocation data, timestamps, and synchronization words are inserted in sequence into a circular buffer (FIFO), which is emptied into the radiolocator's USB link. The same USB connection is used by the outside CPU to issue instructions to the radiolocation device. Figure 5.6 shows a (very simplified) digital path of the radiolocation data.

References

Antolovic, D., Wallace, S. Single-Packet Radiolocation of 802.11 Wireless Sources, Using an Array of Stationary Antennas and High-Speed RF Multiplexing, ACM Proceedings of Wireless Internet Conference (WICON), Boston, MA (August 2006)

Analog Devices. Data sheet for AD7492, http://www.analog.com/ (2001)

Analog Devices. Data sheet for AD8362, Rev. D, http://www.analog.com/ (2003)

Horowitz, P., Hill, W. The Art of Electronics, second edition, Cambridge University Press, Cambridge (1989)

Jung, W.G. IC Op-Amp Cookbook, third edition, Prentice Hall, NJ (1997)

Pilotte, M. Application Note AN-691 – Operation of RF Detector Products at Low Frequency, http://www.analog.com/, Analog Devices (2005)

Chapter 6
Application to Wireless Networking: Tracking Sources in Real Time

6.1 Introduction

In this chapter, we begin to make the connection between the radiolocation proper, that we have discussed so far, and the data communication within which the radiolocation is meant to function. In other words, we want to start using our radiolocator in real wireless networking.

We have said this already, but let us repeat the fundamental requirements: Radiolocation in digital communication should be done on a per-packet basis, and it should be done at the pace of the communication itself, i.e., in real time. Also, by radiolocation, we consistently mean location of sources which, although legitimate interlocutors within some wireless standard, do not cooperate by informing us of their location; the only thing that they reveal about their position is the electromagnetic wave that they transmit. True radiolocation is, therefore, an out-of-channel process, independent of the content of the communication, and relying entirely on the properties of the radio wave. This is fundamentally different from either the GPS, which can be viewed as a form of range-finding with the help of collaborative beacons (satellites) (see e.g. Kaplan and Hegarty 2006), or from any schemes in which the transmitter embeds its own location data into the transmissions. The latter typically amount to broadcasting the location obtained from one's own GPS receiver.

Unlike analog radio communication, digital communication is sporadic and interrupted, with data bundled into carrier bursts (packets) of finite duration, typically under a millisecond. Requirement that each and every packet be separately radiolocated is important for security applications: If a source does not want to be located, it will naturally seek to send few packets, and send them sporadically, at irregular intervals. Radiolocation methods that rely on continuous traffic to reconstruct the source's location will be easily foiled by such simple stratagems.

The real-time requirement is not essential for security applications, as long as the location is obtained quickly enough for security and management measures that rely upon it to be effective. Real time is essential, however, if we wish to use the location information to improve the communication. A typical wireless network node communicates with several mobile interlocutors, in no particular order; therefore, it must

D. Antolovic, *Radiolocation in Ubiquitous Wireless Communication*,
DOI 10.1007/978-1-4419-1632-7_6,

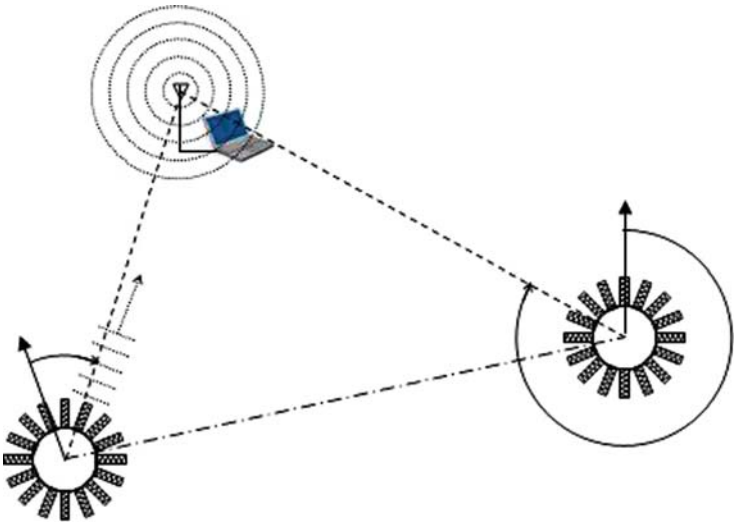

Fig. 6.1 A simple triangulation arrangement with two radiolocation stations

have current positional information to steer each of its transmissions independently. Of course, one could make a distinction between fast- and slow-moving sources, and choose the relevant time scale accordingly, but obtaining the location information from every packet is the most reliable approach that the digital communication allows, and it is reasonable to aim for that.

Figure 6.1 shows a simple triangulation arrangement, with two radiolocation stations and a mobile source. In practice, we would want a more redundant arrangement of radiolocators for overall accuracy, and also to eliminate the monocular direction that is collinear with the radiolocation stations.

In triangulation applications, multiple radiolocators must share their data with a central processing station. That communication should not be of the wireless type; at the very least, it should not occur in the channel being monitored! Triangulation also depends on the accurate geographic locations of the radiolocating stations, that is, on their position and orientation – all of this is readily determined for stationary devices, such as wireless access points.

Geometry requirements are less stringent for adaptive communication. In this application, all that is needed is the bearing of the source, relative to the transmitter, to steer the radio beam accordingly.

Let us add in passing that it is entirely futile to attempt to determine the distance of a source on the basis of its signal's strength at the receiver. Even though the intensity of the signal in the far field in empty space decreases with the square of the distance, that does not help us unless we know the source's actual strength, just as the light of a boat's lantern in the night tells us nothing about how far the boat is. The lantern could be bright and distant, or dim and near, and when a patch of fog

pulls in (i.e., when the radio signal is attenuated by material objects), the observed light from the boat tells us even less.

6.2 Radiolocation Baseband

As we discussed in the introduction to Chap. 4, parallel and serial antenna-sampling architectures entail two very different quandaries in the processing of the available baseband(s). In the parallel architecture, one must quickly find the baseband of adequate strength, or else store everything until the strongest one is selected; in serial sampling, the resulting baseband is simply unsuitable for demodulation, and a second, clean baseband must be obtained.

We have since then committed ourselves to the serial architecture, and in Fig. 6.2 (Antolovic 2008), we show an example of the serial radiolocation baseband, measured on the outputs of the radiolocator's tuner, MAX2820. Comparison with the timing diagram, Fig. 5.5, is helpful in understanding what is happening:

The ADC sampling clock runs in 16-fold bursts, one tick per antenna element. Not shown in the diagram, the multiplexer selector counts from 0 to 15 in that time period, sampling the antennas. We see that the baseband's amplitude within a single antenna scan varies greatly between front and back elements: in this example, the test source was facing antenna elements 12 and 13. This is as it should be – we reconstruct the direction of the incoming wave on the basis of that variation in amplitude.

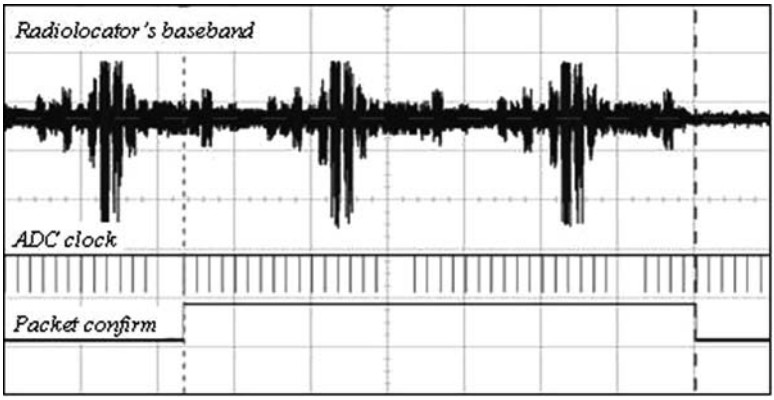

Fig. 6.2 Plot of the radiolocator baseband and auxiliary signals. *Long vertical bars* show the large baseband amplitude corresponding to antenna elements facing the source; *gaps* between bars are due to antenna switching. Baseband strength of each element is sampled on the falling edge of ADC clock; 16-fold bursts of the ADC clock correspond to full scans of the ring of antenna elements. Horizontal time divisions are 50 μs. Plot was generated on an Agilent 54642D oscilloscope. (© IEEE 2008, reproduced with permission)

We also see fairly sharp dips in amplitude between the antenna sampling intervals. As we discussed in Sect. 5.4, there is a switch-over period of ca. 3.4 μs, during which the RF signal on the multiplexer's output essentially vanishes. Taken together, these two distortions make the demodulation of the radiolocation baseband unfeasible.

Figure 6.2 also shows an important packet-validation signal. We will discuss that signal in the next few sections of this chapter.

6.3 Integration of Two Data Paths

In Fig. 6.3, we bisect the now familiar two-path serial architecture from Fig. 4.2 yet again. This time, we distinguish the early real-time stages from the more relaxed later ones, which, in our architecture, means separating analog signal processing and ASIC digital stages from stages implemented in code. We recall that we have set aside the software-radio option in Sect. 4.1, mainly because it demanded high data-transfer rates; therefore, the boundary in Fig. 6.3 is really a boundary between hardware and software implementations.

We have already covered the timing of the radiolocation process in detail in Sect. 5.4; without going into the details of demodulation, we can state in general terms that the left side of Fig. 6.3 comprises processes whose rates are relatively constant, and which do not lag behind the input data rate. Physical passage of the wireless packet is essentially the slowest, rate-setting process, and the outputs of the top and bottom stages on the left side have a firm mutual time relationship.

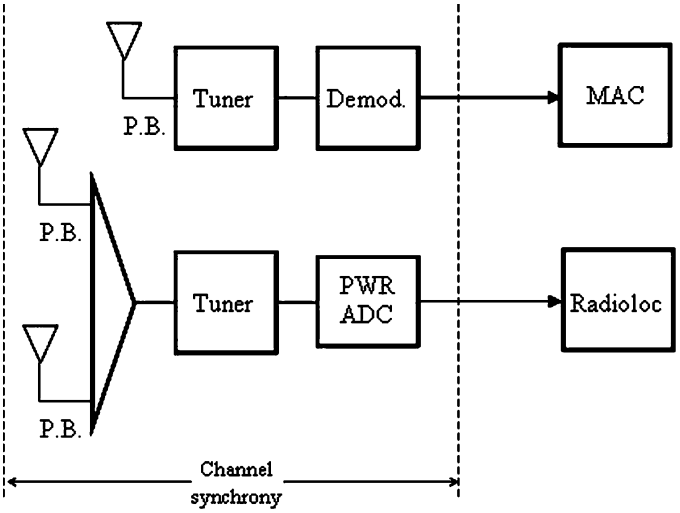

Fig. 6.3 Timing of radiolocation and data processing: real time vs. best effort

What comes out of these stages does so at fixed time intervals, and pertains to the radiowave burst that is currently sweeping over the antennas.

By contrast, the right side consists of coded processes whose duration is fairly strongly dependent on input, whether it is the processing of the packet's logical content, or the iterative calculation of its bearing. There is currently a strong tendency toward using operating systems even in embedded computing; this has many engineering advantages, and the right-hand side of both paths is likely to be implemented under an operating system. While some embedded operating systems provide a semblance of time-guaranteed response, all of this makes the timing on the right side somewhat unpredictable. A mechanism is needed that ensures that the radiolocation result is associated with the correct data payload at the output of the right-hand processing stages.

The simplest architectural solution is to join the corresponding payload and radiolocation data on the left side, and pass them to the right side as a single data structure. This data structure passes through the MAC and bearing calculations either in sequence, or as shared memory that is accessed by two dedicated processors. In either case, the hardware on the right-hand side must carry enough computing capacity to handle the computationally intensive bearing calculation in synchrony with wireless traffic.

This unified-hardware architecture is not the easiest to implement from the engineering standpoint: in effect, it requires integrating a number-crunching 0.5–1.0 GHz CPU with radio hardware. There are good practical reasons to consider a more modular approach, in which the two processes on the right-hand side are implemented independently, and data flows are coordinated by labeling.

There are two plausible choices of labels: first possibility is to extract some identifying information from the data packet in real time, and label the radiolocation data with it, before we enter the best-effort phase. Alternatively, we can label both units of data (packet and radiolocation) with the same label, such as a timestamp, in the real time phase. In either case, the two data streams are re-matched (integrated) after the MAC and radiolocation stages.

6.3.1 Internal Label

Conceptually, this labeling approach amounts to a limited, real-time pre-parsing of the packet, a step preferably implemented in hardware. The difficulty lies in finding a good internal label within the packet's structure. One logical choice is the transmitter's address: we see from Fig. 6.4 that an eight-byte transmitter address is present in (almost) every MAC-layer frame of the 802.11 standard, in the same place relative to the packet's beginning. There are a few types of frames for which this is not the case, but they can be detected by parsing the Frame Control field, which specifies the frame type among other things. The reader will at this point benefit from reading Sect. 10.6, which provides a brief overview of the relevant details of the 802.11b standard. This choice of label gives us the bearing of a source

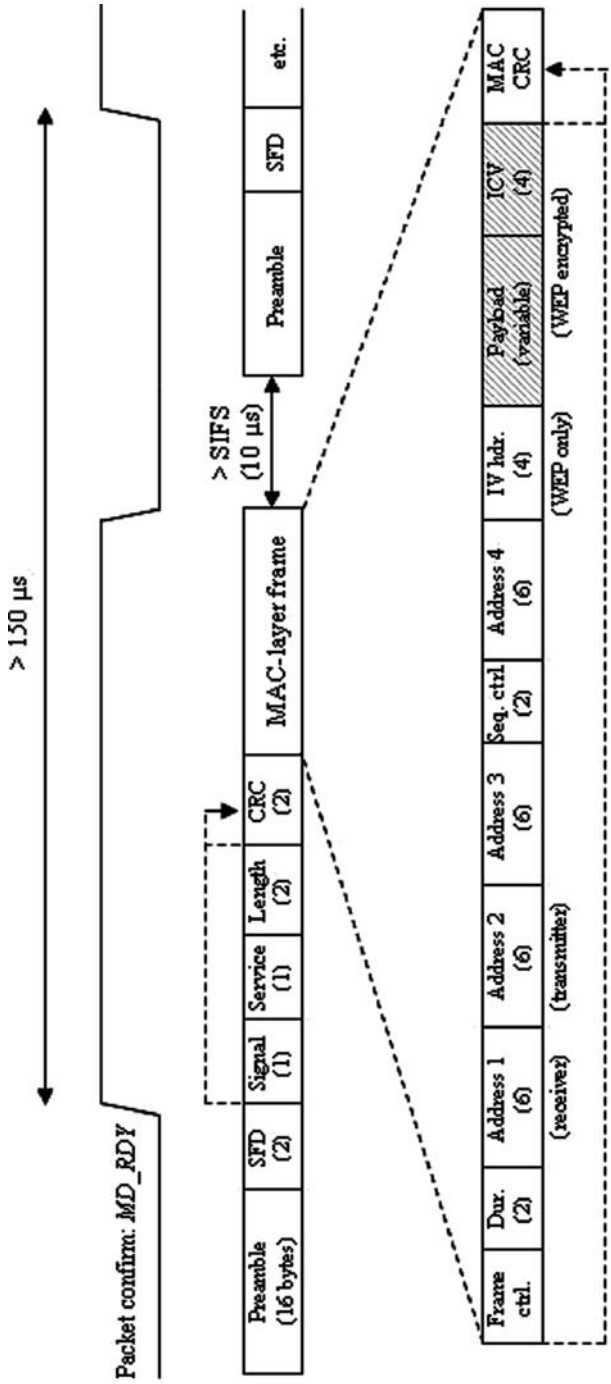

Fig. 6.4 Structure of the PHY/MAC data frame

whose address corresponds to the address in the current packet, but not necessarily the bearing of the packet itself – that bearing may have actually been obtained from another packet. This correspondence is adequate for the purpose of above-the-board communication, but it may not be adequate for security applications: if a malevolent source falsifies the address of a currently present legitimate one, it may well avoid (or at least complicate) radiolocation. One of the timing fields within the data could be used as a label, with the same caveat that this method also relies on the integrity of the sender, and that the label could be malformed, intentionally or not.

In addition, internal labels are standard-specific, and typically require several standard-specific parsing steps for proper disambiguation. An FPGA implementation of this parsing, using some hardware-definition language, is not substantially more onerous than traditional processor programming; however, the approach is inelegant, and lacks generality.

6.3.2 External Label

A label originating outside of the data stream, such as a timestamp or a packet counter, has the advantage of being easy to generate, and of being fully under the control of the radiolocator device; it has the disadvantage of adding volume to the communication data stream. We assume that the radiolocation data packets will always be labeled, since they do not contain anything that could be used as an intrinsic label – short of appending the entire radiolocation packets directly to the communication stream, which really pertains to a shared-memory architecture rather than to a labeling-based one. Figure 6.5 summarizes the labeling alternatives.

Of the external labels, the timestamp is somewhat cleaner than the packet counter, because its presence and validity are independent of the data flow: in the case of the packet counter, missing data packets lead to either duplicated or missing labels on the radiolocation packets, and special provisions have to be made to disambiguate them. Drawing the conclusion from the above arguments, we chose to implement the external labeling scheme, in the form of a timestamp.

We now present the architecture of the two-path networking and radiolocation device, in Fig. 6.6 (Antolovic 2008). The architecture is determined by the serial sampling of antenna signals, and the two paths are integrated by a timestamping scheme. We discuss the specifics of the architecture in the following two sections.

6.4 The Communication Data Path

We used for the communication path of our prototype an existing 802.11b transceiver, the CalRadio, developed by the company Cal-(IT)2, and by University of California at San Diego (Jow et al. 2007). We will discuss this device in greater detail in Chap. 8, but here we touch upon those aspects of the data reception path that are important for the integration with the radiolocator.

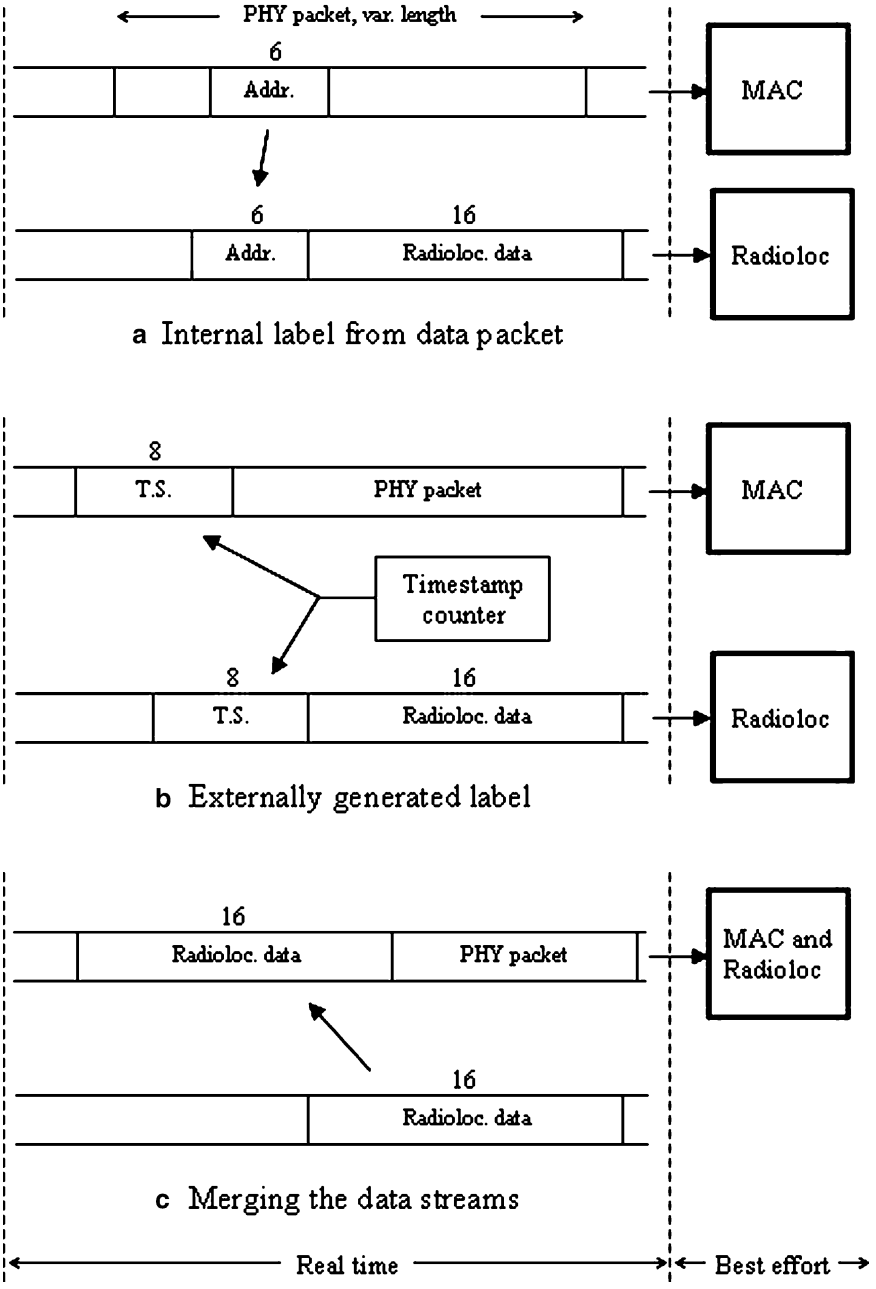

Fig. 6.5 Labeling alternatives and the integration of networking and radiolocation data streams

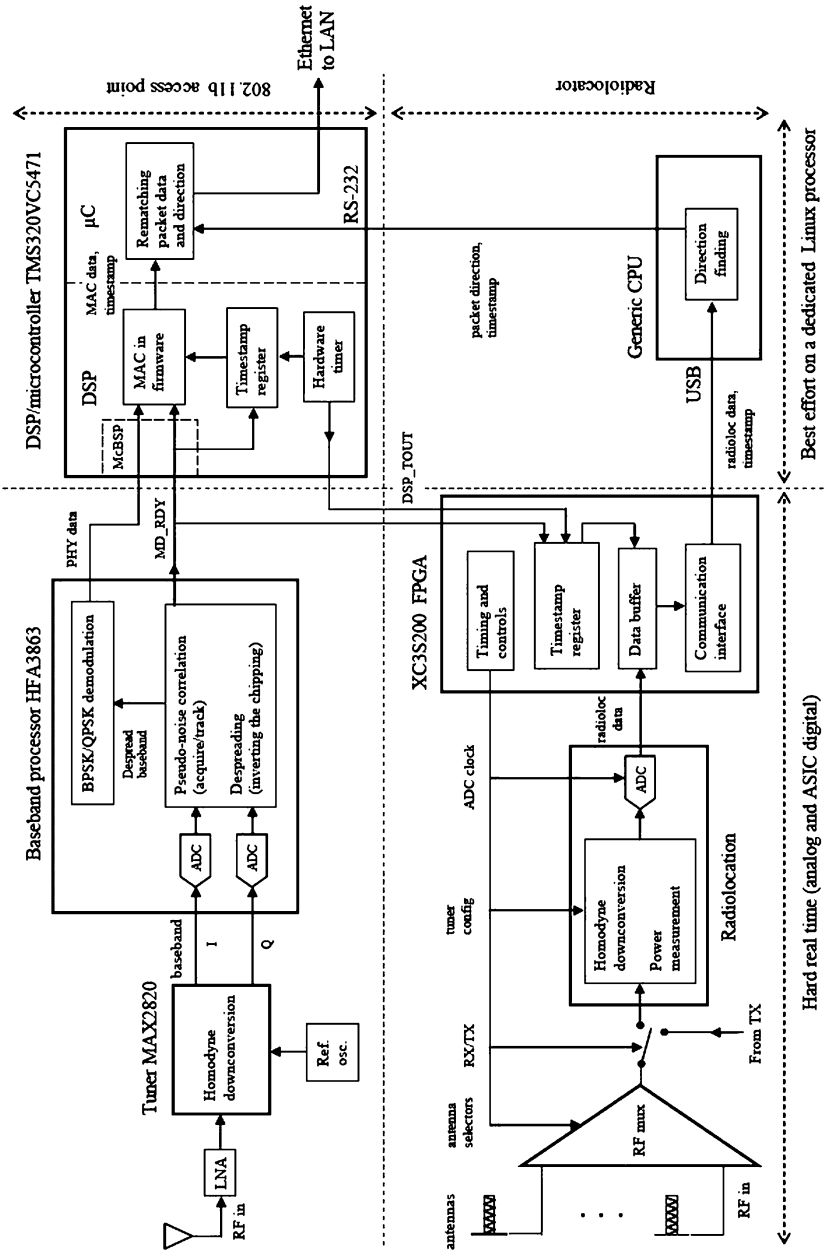

Fig. 6.6 Architecture of the access point with radiolocation (receiver only). (© IEEE 2008, reproduced with permission)

The *tuner* is a fully analog homodyne downconverter, MAX2820 by Maxim Integrated Products (Maxim 2003); the same part is used in the radiolocator circuit (see Sect. 4.3). The tuner converts the signal to baseband, filters its spectrum to the specified 802.11b channel, and amplifies the signal to the power level required for the demodulation stage. The tuner is configurable via a serial communication line, under the control of the board's microprocessor.

Baseband processor is the part HFA3860, by Intersil Corp. (Intersil 2000). This part performs the standard DSSS baseband operations: it correlates the baseband input with a pseudo-noise sequence, despreads the signal, and demodulates it in order to extract the data. There are potentially three modulation schemes present in an 802.11b packet, and we again refer the reader to Sect. 10.6 for an overview of the DSSS/CDMA baseband processing, and to (Gast 2005; Sklar 2001; Razavi 1998) for more complete accounts of these topics.

Physical-layer header starts with the Start Frame Delimiter and ends with the Cyclical Redundancy Check. Baseband processor can be configured to key on either of these fields, and assert its packet confirmation signal (signal MD_RDY in Fig. 6.4) to indicate a valid packet header.

The baseband processor sends demodulated packet data to the MAC processor. This data transfer is synchronous serial, driven by the packet confirmation signal, and has fixed time relationship to the data's reception by the tuner. All of the baseband processing is implemented in hardware, mostly digital, as we can surmise from the input ADCs in the functional diagram of the part HFA3860 in Fig. 6.6. The baseband processor performs no networking functions, and is not programmable in the general sense. Therefore, it can not parse or label the data that it sends to the MAC processor.

MAC processor. The chip controlling the communication path is TMS320VC 5471, by Texas Instruments (Texas Instruments 2000 and 2002). This is a dual-core DSP/microprocessor, whose digital signal processing part is dedicated to running the code implementation of the MAC layer. Details of the MAC layer are of no importance here, but as we have said in Sect. 6.3, the timing issue is crucial.

The MAC-layer code enters different branches, depending on the type and content of the received frame, and there is no firm time relationship between the packet confirmation signal and the processed frame at the MAC's output. This is the stage at which real-time processing becomes intrinsically unfeasible. Data payload is transferred to the microprocessor half, an ARM7 architecture (see for example Wolf 2005), and the subsequent processing happens under an ordinary operating system, μClinux. None of this affords any reliable timing, and that is how we arrived at the need for a labeling scheme to rejoin the two data streams.

6.5 The Timestamp

The crux of the timestamping scheme is to present the same time counter to two or more processes, which use it to label the data in their respective data streams, in real time, as the data become available. One approach to this is to have a

central timestamp register, incremented by a clock; the processes read the register as needed, with provisions made to avoid write/read collisions.

In our two-processor architecture, this would require communicating the timestamp value over a bus, raising all manners of timing and bus-sharing issues, unless we had a dedicated communication line for the timestamp. The alternative that we chose has two separate timestamp registers, which are incremented by the same clock, and reset by the same reset signal. Apart from the fact that the write/read protection must be implemented for multiple registers, this architecture works equally well as the centralized one, and it has the great advantage that only two simple signals have to be shared among the processes: the clock and the reset.

It is of no particular importance where the clock and reset signals originate. In our case, the existing hardware made it easier to implement the timestamp clock in the data-communication processor. However, the clock's period must be short enough to avoid aliasing in the labeling of the packets, and, since the timestamp is not implemented in dedicated hardware, long enough not to soak up too much processor time just by incrementing the timestamp.

We derive the design requirement for the clock's period from the 802.11b standard, which specifies a physical-layer preamble of 144 μs, and allows the shortest interval between transmissions of 10 μs (this is the interval SIFS in Fig. 6.4; see also Gast 2005). Assuming the worst case of zero payload, which is quite unrealistic, the smallest possible interval between the starting points of two consecutive packets, marked by rising edges of the packet confirmation signal (MD_RDY signal in Fig. 6.4), would be around 150 μs. As Fig. 6.2 shows, actual length of the packet confirmation pulse for a beacon frame is ca. 350 μs, and the consecutive packets can be reliably marked by a timestamp with a period of 50 μs or less.

The length of the timestamp is a compromise between the need for a long rollover time on one hand, and not burdening the communication with too much timestamp data on the other. We settled upon the length of eight bytes, which provides a rollover period of 8.5 billion days at the clock period of 50 μs. The communication processor has the clock speed of 100 MHz, and maintaining this timestamp places no significant burden upon it. As we shall see in Chap. 8, communication channels within the prototype were able to handle the data volume as well.

6.6 Test of the Radiolocator Access Point

In this section, we discuss the experiments that demonstrated the radiolocation of true, decoded 802.11b packets; this was the goal of the architecture described in this chapter. The setting in which we conducted our tests was a wide and relatively uncluttered interior, similar to the environment in which the results in Figs. 2.20 and 2.21 were obtained. Experimental results in this section are from (Antolovic 2008).

We used as test sources two unmodified commercial access points, which were configured to merely broadcast beacon frames on the same Wi-Fi channel (channel 11, 2,462 MHz), and do nothing else. The radio environment was not entirely silent,

but channel 11 was relatively quiet at the time. The test access points were made by different manufacturers; they used different omnidirectional antennas, and transmitted different numbers of beacons per second. They were also close enough to detect each other's signals and avoid transmission collisions, as required by the 802.11 standard. This arrangement was realistic and similar to what one would expect to find in an open office area. Figure 6.7 shows the layout in space, and Table 6.1 summarizes the configuration of the experiment, and lists the average bearings of the detected sources.

The radiolocation device listened passively and recorded the received packets, with their source MAC addresses, and with bearings associated with each packet. Table 6.2 gives a brief contiguous segment from a longer experimental run: each line corresponds to a packet, and the packets are displayed in the order in which they came out of the MAC-layer processing. In Table 6.2, we see two sources, transmitting different numbers of packets per second (the source "77:48" was roughly twice as talkative as the source "9f:00"), and we see the source's bearing associated with each packet. The table also shows the timestamps given to the data streams in the two data paths described above. These timestamp values differ by the same amount within each line (16 ticks), and could be used to reliably re-associate the two data streams. The difference in timestamps stems from the details of the de-

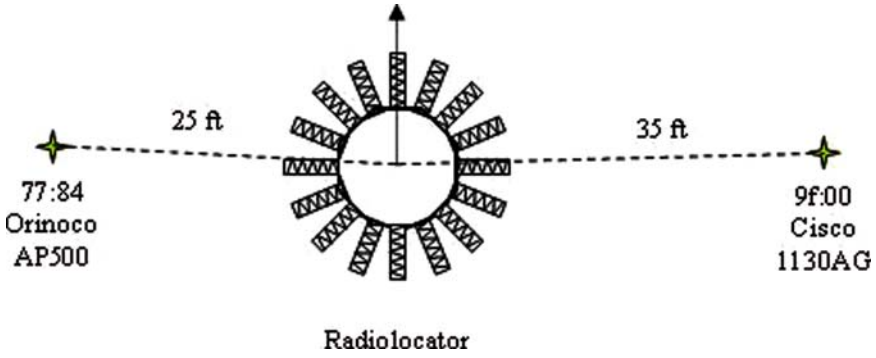

Fig. 6.7 Layout of the functionality test: the two access points were transmitting beacon frames only, and the radiolocator listened passively to the wireless traffic

Table 6.1 Summary of the functionality test of the radiolocator (© IEEE 2008, reproduced with permission)

Source MAC and model	Dist. (ft)	Pack/sec.	Avg. bearing (deg.)	σ (deg.)	Packet count
00022d0f7784 Orinoco AP500	25	10	272.87	0.3	432
001244b09f00 Cisco 1130AG	35	5	89.02	0.94	205

Duration of test: 55 s, total of 637 packets received
Bearings are referenced clockwise, from the direction of zero-th antenna element
Bearing averages and deviations pertain to the entire run

Table 6.2 Functionality test of the radiolocator access point: packets are shown in the sequence in which they are received, along with their respective bearings (© IEEE 2008, reproduced with permission)

Dest. (bcast)	Source MAC	Packet stamp	Bearing (deg.)	Bearing stamp
ff...ff	00022d0f7784	1086505	273.01	1086521
ff...ff	001244b09f00	1090221	89.59	1090237
ff...ff	00022d0f7784	1090600	272.88	1090616
ff...ff	00022d0f7784	1092648	272.54	1092664
ff...ff	001244b09f00	1094317	92.33	1094333
ff...ff	00022d0f7784	1094695	272.99	1094711
ff...ff	00022d0f7784	1096743	273.19	1096759
ff...ff	001244b09f00	1098412	90.39	1098428
ff...ff	00022d0f7784	1098790	272.88	1098806
ff...ff	00022d0f7784	1100838	272.72	1100854
ff...ff	001244b09f00	1102507	89.55	1102523
ff...ff	00022d0f7784	1102886	273.11	1102902
ff...ff	00022d0f7784	1104933	272.99	1104949
ff...ff	001244b09f00	1106602	89.98	1106618
ff...ff	00022d0f7784	1106981	273.04	1106997

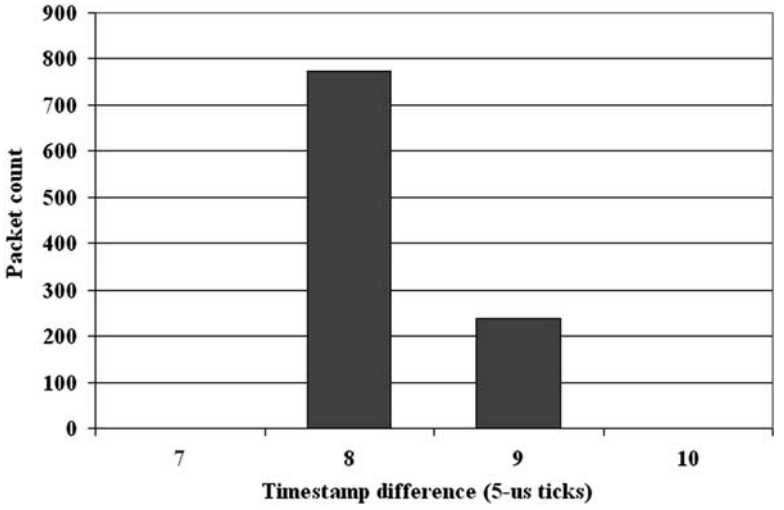

Fig. 6.8 Statistical distribution of timestamp differences in a test run of 1,000 packets. Single test source was used (Orinoco AP500), and the timestamp period was set to 5 μs. Median value differs from that in Table 6.2, due to small changes in the startup code. (© IEEE 2008, reproduced with permission)

vice's boot-up sequence, and it does not matter much, as long as it is known and unchanging.

As part of these tests, we also estimated the accuracy of the entire timestamp-matching process. For this process to work, the variation in the timestamps' difference must be small, compared to the time scale of the communication, i.e., the timestamping must be precise relative to packet durations and inter-packet intervals.

We see from Fig. 6.8 that this is indeed the case. We shortened the timestamp interval to 5 μs, to obtain as accurate a measurement as feasible, without overwhelming the processor, and we measured the distribution of the difference between the two timestamps. The results show that the overall variation falls within 10 μs, which is more than adequate to reliably associate packets with the bearings of their sources, even in heavy wireless traffic.

References

Antolovic, D. Architecture of a 802.11b Access Point with Single-Packet Radiolocation, Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), Las Vegas NV, March-April 2008 (2008)

Gast, M.S. 802.11 Wireless Networks, 2nd edition, O'Reilly, CA (2005)

Intersil Corporation. Data sheet for HFA3863 (2000)

Jow, A., Schurgers, C., Palmer, D. CalRadio: A Portable, Flexible 802.11 Wireless Research Platform, Proceedings of MobiEval'07, San Juan, Puerto Rico, June 2007; also http://calradio.calit2.net/ (2007)

Kaplan, E.D., Hegarty, C.J., eds., Understanding GPS: principles and applications, 2nd edition, Artech (2006)

Maxim Integrated Products. Data sheet for MAX2820. http://www.maxim-ic.com/ (2003)

Razavi, B. RF Microelectronics, Prentice Hall, NJ (1998)

Sklar, B. Digital Communications, 2nd edition, Prentice Hall, NJ (2001)

Texas Instruments. TMS320C54x DSP Functional Overview (2000)

Texas Instruments. TMS320VC5471 Fixed-Point DSP Data Manual (2002)

Wolf, W. Computers as Components, Elsevier, Amsterdam (2005)

Chapter 7
Application to Wireless Networking: Adaptive Response

7.1 Introduction

In Chap. 6 we resolved the central issue of radiolocation in wireless networking, that of assigning directional information to individual packets. In this chapter, we build on this central functionality to give our wireless architecture the ability to direct its response toward the interlocutor.

Directing our own transmissions in the desired direction is conceptually trivial, compared with the task of determining the unknown direction of a radio wave sent by somebody else. Nevertheless, the process must be fast enough to switch direction between any two transmissions, in order to be of use in wireless networking – a rotating antenna is just as unacceptable here as in radiolocation. In addition, we would like to integrate directional transmission and radiolocation in the same RF front end and antenna complex to keep the size and cost of the device within reason. In this chapter, we revisit the radiolocator's RF section, discuss the design requirements for the directional response, describe the architecture of a transceiver with radiolocation and directional response, and give the results of functionality tests of the device.

7.2 Circular Phased Array

There are two ways to change the beam's direction in stationary antennas: In discrete steps, by switching the signal from one directional antenna to another, or continuously by controlling the phase difference between array's elements, thereby changing the direction of the interference maximum.

We touched upon the linear phased arrays in Sect. 1.5.1, where we explained that by inserting the same phase delay between pairs of adjacent elements we can adjust the direction in which the waves from all elements reinforce each other. We pointed out that the helical antenna is a clever form of a phased array as well.

Since the radiolocation antenna consists of a ring of directional elements, we discuss here the interference of waves in a circular arrangement of sources. Let us begin with a set of isotropic point sources, equal in intensity and phase, same as

D. Antolovic, *Radiolocation in Ubiquitous Wireless Communication*,
DOI 10.1007/978-1-4419-1632-7_7,

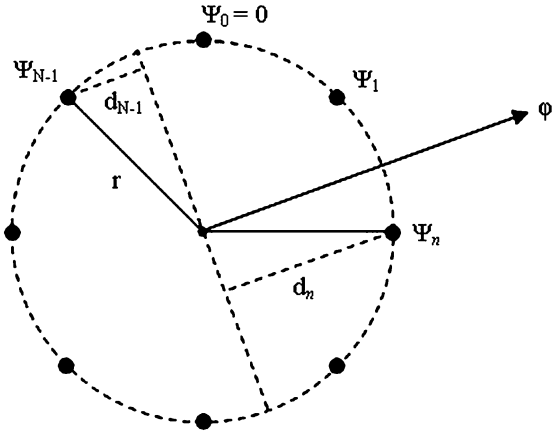

Fig. 7.1 Ring of point sources contributing to radiation in direction φ. The phase of each wave, relative to the wavefront through the center, is $k\,d_n$

the sources in the linear array in Sect. 1.5.1. Figure 7.1 shows the ring of sources, located at angles Ψ_n. The phase of a wave in direction φ, originating at the source at Ψ_n, and relative to a plane wavefront passing through the center of the circle, is $k\,d_n$, where $k = 2\pi/\lambda$ is the wave number. The array factor for the circular array can be written as

$$\text{array factor} = \sum_{n=0}^{N-1} e^{ikr\cos(\varphi-\Psi_n)} \tag{7.1}$$

This array factor reflects only the phase differences due to the geometry of the array. For evenly spaced sources, we can see from the symmetry of the arrangement that the array factor is periodic with N periods around the circle and an even function around its maxima (or minima). For an even-numbered N, this is also a strictly real function, because for every term $e^{ikr\cos(\varphi-\Psi_n)}$ in the sum in (7.1) there is a complex conjugate term $e^{ikr\cos(\varphi-\Psi_n+\pi)} = e^{-ikr\cos(\varphi-\Psi_n)}$.

This array factor has N preferred directions around the circle, in which waves reinforce each other. In order to form a single beam in a particular direction ψ, we add a phase shift

$$p_n = -kr\cos(\psi - \Psi_n) \tag{7.2}$$

to every source. This additional phase shift exactly cancels the geometrical phase shift in the direction $\varphi = \psi$, so that all sources contribute reinforcing waves with phase zero in that direction. We can now replace the isotropic sources with directional elements with (amplitude) lobes $L_A(\varphi)$, and write the expression for the normalized lobe function (the far-field *amplitude*, normalized to unity and to unit distance) for the planar ring of directional antennas:

$$\Lambda(\varphi,\psi) = \frac{1}{N}\sum_{n=0}^{N-1} L_A(\varphi-\Psi_n)\,e^{ikr[\cos(\varphi-\Psi_n)-\cos(\psi-\Psi_n)]} \tag{7.3}$$

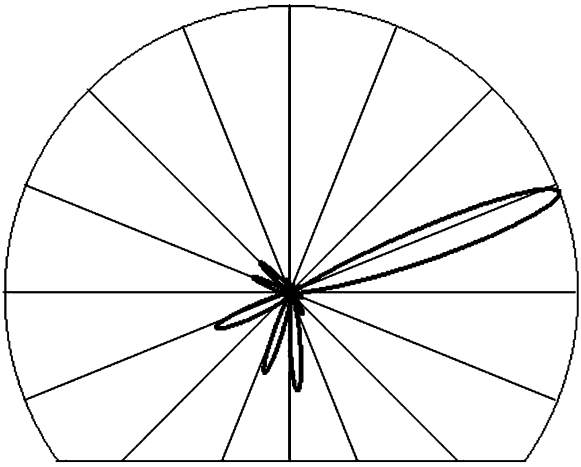

Fig. 7.2 Simulated beam of the phased array of helices, on the linear scale. Direction of the beam is 70° from the vertical, and the phases are given by (7.2)

We should point out an implicit approximation in this formula: Individual directional elements are not point sources fitted with lobes, but extended shapes in space. The radius r is only a judiciously chosen average, and a more detailed analysis would involve integrals over antenna currents, as outlined in Sect. 1.5. Numerical investigation shows, however, that the shape of the array's main lobe is insensitive to variations in r that are small compared to the wavelength.

Furthermore, the circular array has a full radiation distribution in space, while we have restricted our discussion to the ring's plane for simplicity. For a detailed account of circular arrays, we refer the reader to (Balanis 1997; Elliott 2003; Ma 1974).

Figure 7.2 shows the plot of the calculated function $|\Lambda(\varphi, \psi)|^2$ for our ring of 16 helices, with lobes of helices as given in Fig. 2.20. We see that the ring can function as a phased array with a moveable beam, although it has somewhat prominent side/back lobes. This is not too surprising, since the elements are directional: Elements facing away from the beam contribute to the beam with constructive interference of their weak back-end radiation, and the elements facing the beam reduce the side/back lobes through destructive interference from their own, equally weak, back-end radiation.

7.2.1 Phase Shifting

In (7.2) we wrote down, somewhat cavalierly, the magnitude of the phase shift that has to be applied to each antenna channel to form the beam in certain direction. This phase shift varies from channel to channel, and is, in fact, not a particularly

straightforward function of the beam direction. Not very many options are available for implementing such a set of phase-shifted transmissions. We could generate a number of mutually phase-shifted carriers, using a ring of coupled oscillators, but that architecture would take us too far afield from the range of RF parts currently available in commercial wireless networking. A more realistic alternative is to use ready-made phase shifters.

Physically, there are only a few ways to shift the phase of a harmonic signal. Placing magnetic cores in hollow waveguides is perhaps the oldest phase-shifting technique, but this is bulky and hardly applicable in our context. Controlled changing of the length of the signal path or controlled loading of the line (to achieve a phase shift through impedance discontinuity) is also used, and shifters based on these principles are implemented commercially, as in-cable attachments or as board-mounted devices. The amount of phase shift is controlled either digitally or by an analog voltage level.[1] In either case, we would need 16 phase shifters and the appropriate control circuitry to steer the antenna beam according to (7.2).

The purpose of this section is not to discuss phase shifting in depth, only to assess its ramifications for adaptive-response architecture. We refer the reader to (Pozar 1998) for a more detailed discussion of the subject.

7.2.2 Simultaneous Use of Multiple Antennas

This section is also a good place to discuss the use of several antennas at once, either for transmission or reception. Our communication device must have an omnidirectional receiving mode in which it listens for signals coming from unknown directions, and it should also be able to broadcast omnidirectionally, at least in order to send out beacons to potential interlocutors whose location is not yet known. The design question arises whether the radiolocator ring could be used for that purpose, or a separate omnidirectional antenna is needed.

Simultaneous use of multiple antennas is often advocated as beneficial "antenna diversity," but let us take a closer look. To start with, feeding two antennas from the same transmitter is not the same as powering two light bulbs from the same battery. When a transmission line of impedance Z_0 divides into two lines with the same impedance, the wave incident upon the junction sees the impedance change from Z_0 to $Z_0/2$, and the magnitude of the reflection coefficient is $|\Gamma| = 1/3$ (see 10.41). The junction leads to reflection losses, even if all other impedances are perfectly matched.

[1] See, for example, analog-voltage controlled shifters, part series RVPT, and digitally controlled shifters, part series RFPSHT, by RF-Lambda Inc.

This reflection loss can be eliminated with the Wilkinson divider, a simple, passive three-port device. With matched impedances on all three ports, the divider splits the transmitter's signal losslessly into two signals of half the original power (−3 dB split). The downside is that the divider is lossy in the reverse direction, and cannot be made lossless.

In the receiving direction, a radio wave impinges on two antennas, resulting in two signals on the feed lines. Again, if these lines are simply joined, the two signals will not add up happily on the way to the receiver: They will bounce off the discontinuity and leak from one antenna into the other. Wilkinson divider fully isolates the antennas from each other, and it will join their signals losslessly, but only in the rare event that the two signals are identical. Let us consider, on the contrary, the extreme case of only one antenna receiving a signal: The divider will cut its power by 3 dB on the way to the receiver, just as it did in the transmit direction, only this time half of the power is dissipated in the divider, not diverted to another port. We refer the reader to Sect. 10.7 for further discussion of the Wilkinson divider.

In addition to the above issues of signal distribution and rejoining, multiple antennas always constitute an array whose radiation/reception pattern is determined by wave interference. This pattern will have preferred directions. For example, it is easy to see from the formulas for the linear array (Sect. 1.5.1) that just two sources, placed half a wavelength apart, interfere destructively and form a null in the direction collinear with the sources, even though the sources themselves are nicely isotropic. In practice, these sources could be two "diverse" antennas, positioned injudiciously on a device and producing an unintended signal cancellation.

In the case of our prototype radiolocation ring, formula in (7.3) – with the ψ-term omitted from the exponent – yields the radiation/reception pattern. Figure 7.3 shows an example of the power distribution $|\Lambda(\varphi)|^2$: As expected, the pattern exhibits 16 periods around the circle, but no nulls (the ring of point sources has sharp nulls for certain radius values). The ratio of maximum to minimum power varies somewhat with the choice of radius in (7.3), but it stays between 1.5 and 2 for realistic radius values.

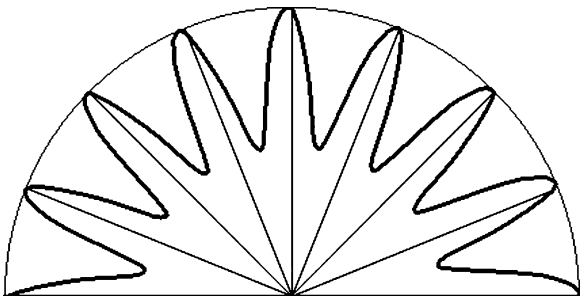

Fig. 7.3 Top half of the radiation pattern of the ring of helices, with all antennas in phase. The pattern exhibits 16 periods around the circle

7.3 Design Requirements of the Adaptive Response

Having revisited some aspects of the RF front end, let us now address the design requirements for an adaptive communication device. We start by spelling out the minimum modes of operation that it must have:

- Radiolocation/reception; this is the coordinated two-path processing described in Chap. 6. This mode is obviously the core functionality that makes the adaptive communication possible.
- Omnidirectional transmission; this is needed in order to broadcast signals, such as beacon frames in the 802.11 standard that let potential interlocutors know of the existence of the wireless network.
- Adaptive transmission; this is the transmission directed toward an interlocutor in the known direction. This is the core functionality for improving the quality of service.

One can also envision an adaptive reception mode, in which the device listens in a preferred direction only. Such a mode would be useful in a device that needs to autonomously make itself part of a WDS network (see Sect. 10.6.1): Once it has detected a WDS interlocutor, it limits all communication to its direction.

There are also design constraints dictated by size and cost. As a practical matter, our device can have only one antenna of the "compound eye" type that we adopted in this work. Nothing prevents multiple antenna arrays in principle, but as we discussed in Chaps. 1 and 2, the size of directional antennas is always limited from below by the wavelength, and miniaturization entails a loss of directionality and gain. One compound antenna is the best that we can expect.

Following closely on this point is the question of beam control. For a compound eye antenna there are two options already mentioned in Sect. 7.2: Phased array or switched array. The phased array can direct the beam in any desired direction with only minor changes in the beam shape, while the switched array is restricted to a discrete set of beam directions. However, as discussed in Sect. 7.2.1, the phased array requires more complex support circuitry: A set of phase shifters and a control circuit that generates correct phase shift for each antenna element. The switched array requires no more than an RF multiplexer.

The choice between these methods will be influenced by the specifics of the application: We would expect long-distance communication to benefit from a narrow and precise beam, and the phased array is therefore appropriate for larger communication units, covering a wide area. For smaller access point in an interior setting, switched arrays should be adequate. Either option is feasible in our architecture, but the considerations of complexity and cost made us choose the switched array for implementation in our prototype.

There are other, less critical, requirements balancing circuitry sprawl, complexity, and cost. For example, since the RF multiplexer must be present for the radiolocation mode, and the device must never receive and transmit simultaneously (!), it is desirable to use the same multiplexer for the adaptive transmission.

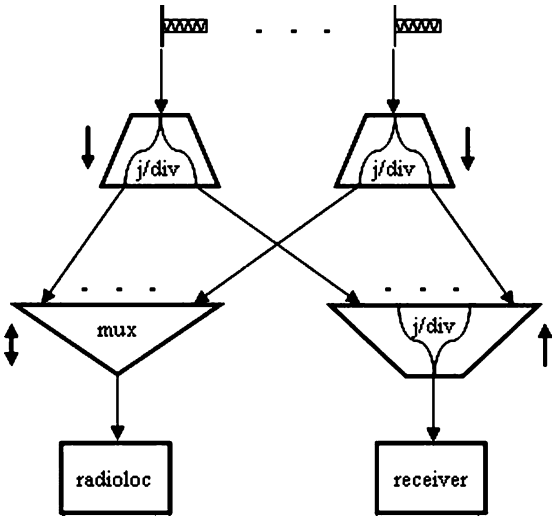

Fig. 7.4 Radiolocation and communication paths fed by the same antenna array. Notice that the manifold Wilkinson on the right operates in the potentially lossy direction, as a signal joiner

Finally, there is a design constraint that we pointed out in Sect. 7.2.2, that of the combined use of all antenna elements in omnidirectional modes. Radiolocation needs omnidirectional reception to supply valid packet data to the communication path – it has to be omnidirectional, of course, because the source's bearing is not known prior to radiolocation. Figure 7.4 shows what this implies for the signal path from the antenna array to the receivers.

The signal from each antenna element must be divided, to feed both radiolocation and communication paths. An array of Wilkinsons will accomplish this, because they divide the signal evenly and losslessly, and the isolation of the dividers' output ports prevents most of the radiolocator's choppy signals from leaking into the communication path. Isolation of 20–30 dB between ports is common in commercial parts.

As we know from Sect. 6.2, antenna signals for the same packet vary widely in amplitude, and we do not know in advance which ones are strong enough for reliable demodulation. Short of demodulating every antenna signal separately, we may consider recombining them with a tree of Wilkinsons, working as joiners. Such a tree can be easily implemented, and integrated parts exist that form ready-made Wilkinson trees. For example, the four-output joiner/divider SBD-4–25 (by Mini Circuits) works well in the dividing direction, distributing the signal losslessly, and we use it for omnidirectional transmission mode (see Mini Circuits 2005). However, in the receiving mode the joiners are lossy in messy ways, the magnitude of the loss depending on the relative amplitudes and phases of the antenna signals. In addition, the antenna array has the inevitable uneven angular distribution due to phasor addition, as discussed in Sect. 7.2.2 and shown in Fig. 7.3.

These complications need not be fatal to the functionality of the design, but they were not worth the additional array of dividers in Fig. 7.4, and other circuit

complexities. We decided to use a separate omnidirectional antenna for omni reception in our prototype. To summarize, then, we implemented the device's operational modes as follows:

- Radiolocation: Antenna ring and RF multiplexer in the radiolocation path; auxiliary omni antenna, followed by a low-noise amplifier (LNA) in the communication path.
- Omni transmission: 16-fold divider tree feeding the antenna ring.
- Adaptive transmission: Antenna ring and RF multiplexer, functioning in the reverse direction as antenna selector. The switches comprising the multiplexer are entirely indifferent to the signal's direction, as one would expect.

Apart from the multiplexer and the divider, there is a layer of switches immediately following the antenna ring, and a few additional routing switches, which connect the receivers and transmitters to the right antennas. The topology of the RF switchboard is shown in Fig. 7.5; the switchboard is controlled by the FPGA circuit which translates the mode selection into appropriate throws of the switches.

7.4 Overview of the Adaptive-Response Architecture

Figure 7.6 gives the high-level overview of the entire communication device with radiolocation and adaptive response (Antolovic 2008b). We recognize the familiar radiolocator/receiver, along with the ASIC controller and the supporting CPU, as a subset of the architecture. These subsystems were depicted in greater detail in Fig. 6.6, and we meet again the "packet ready" and timestamp signals from Chap. 6.

The novel aspect of the architecture at this point is the adaptive response. When the transceiver is ready to send a packet, it first sends an instruction to the controller to place the antenna switchboard in the transmit mode; the controller circuit recognizes a small set of two-byte instructions, which it receives via the radiolocator board's USB port.

It is by means of these instructions that the transceiver changes the mode between packets, turns the radiolocator and the low-noise amplifier off during transmission, and selects the best antenna for adaptive response. Upon completing the instruction, the controller raises a confirmation signal ("board ready"); this synchronization signal allows the start of the transmission cycle. The controller and the transceiver are mutually asynchronous subsystems, and if a race condition allowed the transmission to start while the receiving circuits were still turned on, damage to the front-end receiving electronics would almost certainly occur.

The controller's instruction set is also used for the initial configuration of the radiolocator's tuner (Maxim 2003), and for manual channel switching, if desired.

In our prototype, radiolocator board and the CalRadio transceiver (Jow et al. 2007) are integrated at the level of individual devices, rather than by combining the circuitry on a common board. This engineering expediency dictates the somewhat circuitous path of the control instructions, as shown in Fig. 7.6, and requires some

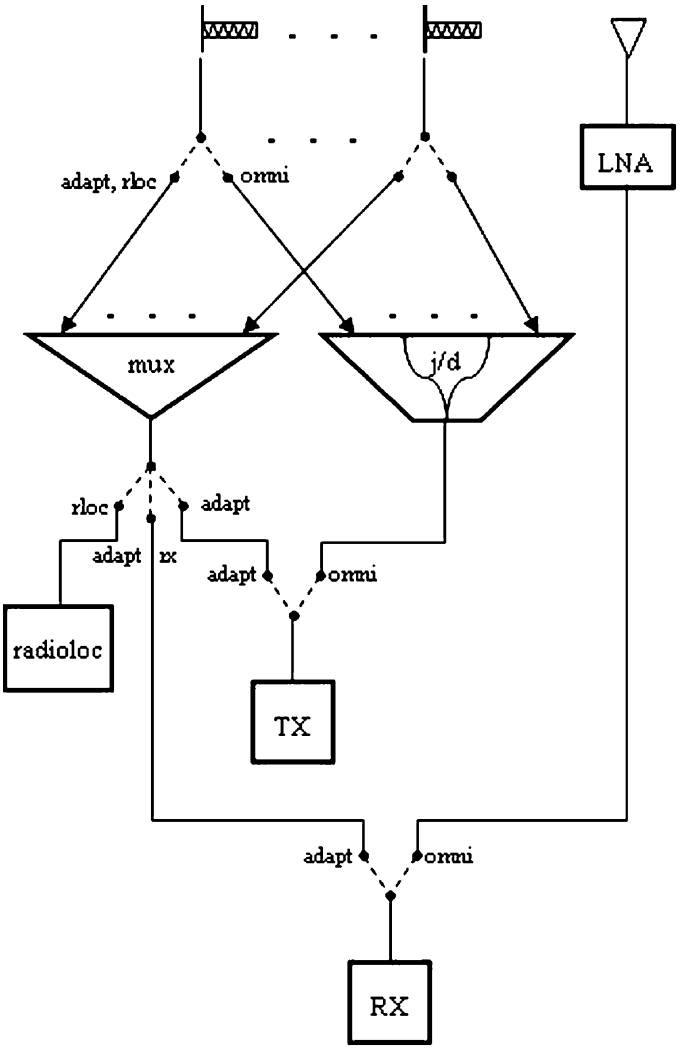

Fig. 7.5 Schematic of the RF switchboard

unconventional code features in the transceiver's firmware, such as direct kernel-space communication of a driver module with the serial port.

The master process of the reception/transmission cycle is the Logical Link Control (LLC) layer, which is implemented in the transceiver's firmware and runs under μClinux. By the time this network layer is reached, radiolocation information is assigned to the received 802.11b packets, and it is here that the response is initiated.

During the reception periods, the reception data path and the radiolocator run in parallel, as we have described in Chap. 6. As radiolocation bearings are associated

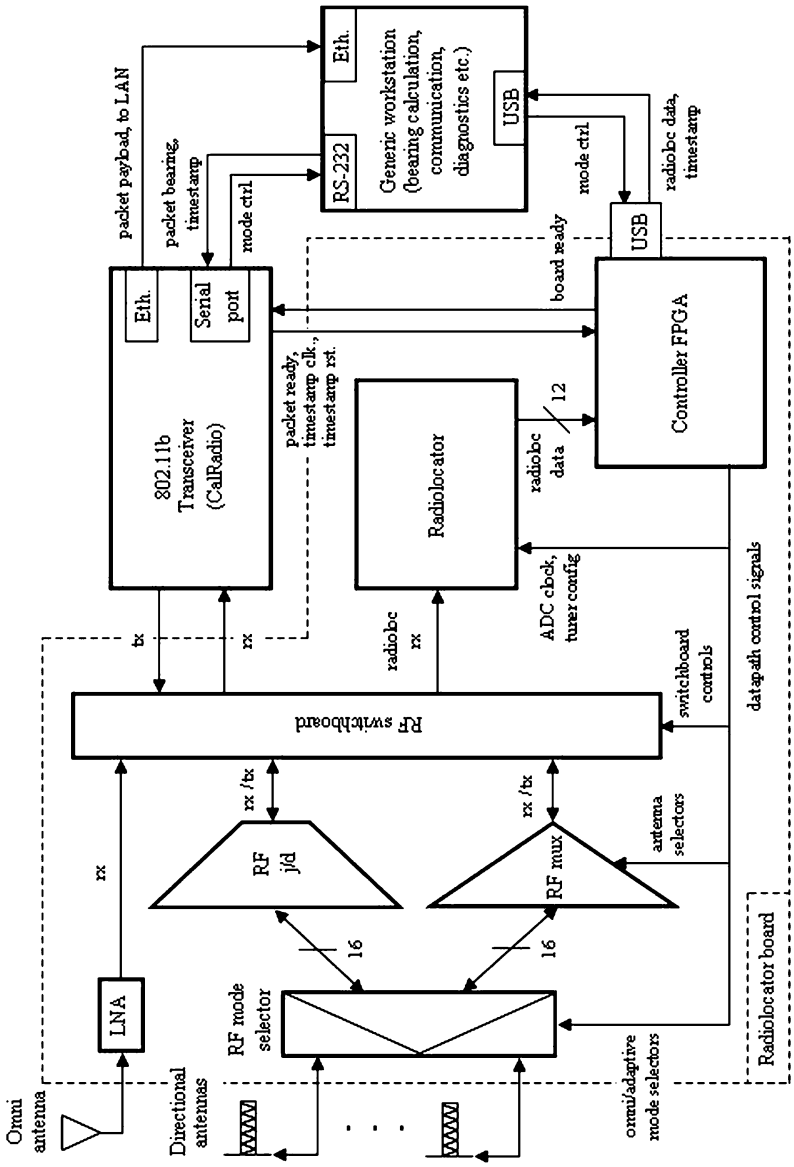

Fig. 7.6 Architecture of the adaptive access point, showing the main components: radiolocator, 802.11b transceiver, and the supporting CPU. (© IEEE 2008, reproduced with permission)

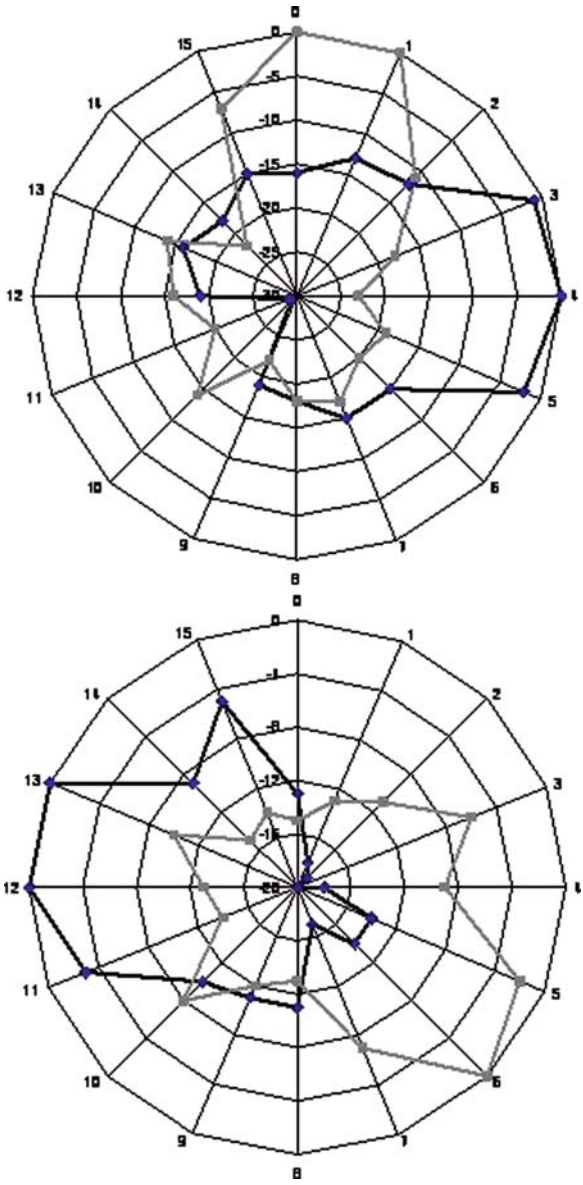

Fig. 7.7 Test of the autonomous directional response. The prototype was pinged by a Linksys wrt54gl wireless router, from the approximate directions of antenna elements 0, 4, 6 and 12, at distances ranging from 45 to 68 ft. Plot shows the signal strength of the prototype's ping response, measured around the compound antenna, in dB relative to the maximum. As the test source moves, the system autonomously selects the best antenna for its response. Signal strengths were measured with AirMagnet wireless diagnostic tool, running on a Hewlett-Packard Pocket PC. ()

with incoming packets, the firmware maintains a list of recent source addresses and their corresponding bearings. The list is updated by every received packet, and contains bearings that are as up-to-date as the sources' activity allows.

Upon transmission, the destination address is checked against this list, the directional antenna closest to the destination's bearing is selected, and the switchboard placed into adaptive transmit mode. The transmitter then sends the packet through the selected antenna. Broadcast packets and transmissions to sources of unknown direction are sent in the omnidirectional mode, through all directional antennas.

7.5 Test of the Adaptive Directional Response

We conducted functionality tests of the adaptive response outdoors, at communication distances ranging from 40 to 70 ft. The test source was a wholly standard 802.11b router, Linksys wrt54gl, running dd-wrt firmware. Our prototype device established logical association with the test source, and exchanged ping queries and replies with it; ping exchanges were performed with either device as the querying source. Experimental results in this section are from (Antolovic 2008a and 2008b).

While the prototype and the router engaged in continuous ping exchanges, we measured the signal strength of the prototype's transmissions on a circle of points around it; we used AirMagnet's Handheld Wireless LAN Analyzer, a Wi-Fi diagnostic tool. Results are shown in Fig. 7.7; even though the signal strengths were reported rather imprecisely, we see that, as the test point moved around, the radiation pattern changed to follow it: Our device was continuously radiolocating its interlocutor, and was autonomously switching the transmission antennas to follow it.

References

Antolovic, D., Demonstration of Fast, Single-Packet Radiolocation, Applied to 802.11 Wireless Networking; prototype demonstration at Fifth IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV (January 2008a)

Antolovic, D., Directional Radio Response of a 802.11b Device Guided by Radiolocation, Proceedings of 2nd IEEE International Interdisciplinary Conference on Portable Information Devices, Garmisch-Partenkirchen, Germany (August 2008b)

Balanis, C.A.: Antenna Theory, Analysis and Design, 2nd edition, Wiley, New York (1997)

Elliott, R.S.: Antenna Theory and Design, Revised Edition, Wiley, New York (2003)

Jow, A., Schurgers, C., Palmer, D., CalRadio: A Portable, Flexible 802.11 Wireless Research Platform, Proceedings of MobiEval'07, San Juan, Puerto Rico (June 2007); also http://calradio.calit2.net/

Ma, M.T.: Theory and Application of Antenna Arrays, Wiley, New York (1974)

Data sheet for MAX2820, Maxim Integrated Products (2003). http://www.maxim-ic.com/

Mini Circuits application note AN-10–006, Understanding Power Splitters (2005). http://www.minicircuits.com

Pozar, D.M.: Microwave Engineering, Wiley, New York (1998)

Chapter 8
Engineering Aspects of the Transceiver Design

8.1 Introduction

In this chapter, we gather together various engineering details of our radiolocation transceiver. One may be tempted to think of the present chapter as an implementation blueprint, but we should have no such pretentions. Every implementation of a technology goes through its own development cycles, addressing its own specific requirements of functionality, size, marketability and cost. Every new device has its own set of noise problems, electromagnetic interferences, integration issues and design errors that must be dealt with, and those can never be fully forestalled with written descriptions.

The aim of the discussion in Sect. 8.2 is to recognize yet again that radiolocation involves quantitative measurements. We have already touched upon the quantitative evaluation of crosstalk, losses, settling times etc. in Chaps. 4 and 5; here we present a general description of the instrument, in the hope that it will be helpful to the readers in avoiding their own specific pitfalls.[1] This is also a good place to offer a summary of some practicalities, such as the breakdown of the radiolocator-transceiver into functional subsystems, communication between the subsystems, and other details that will be helpful in understanding this instrument.

Section 8.3 offers an engineering overview of a major subsystem, which is CalRadio, the open-design 802.11b transceiver that provides the wireless networking functionality to the radiolocator. Section 8.4 briefly describes the physical implementation of the instrument.

[1] Electronics literature is replete with varied and sometimes contradictory guidelines for electronic design, each guideline supported by its advocate's extensive experience. We do not wish to claim any universal validity for our approach; we are merely satisfied that it resulted in an acceptably accurate device. Reliable evaluation of design alternatives must entail detailed testing and E&M simulations, and whether such efforts are justified depends on the circumstances of the project.

D. Antolovic, *Radiolocation in Ubiquitous Wireless Communication*,
DOI 10.1007/978-1-4419-1632-7_8,

8.2 Radiolocator Board

The board has four layers. The guiding design principle has been the separation of analog and digital signals throughout, with particular attention paid to the radiolocation path from the antennas through the switchboard, to the tuner, power meter, and the ADC. This is the circuitry that performs quantitative measurements, and we made every effort to protect it from digital noise. All RF traces are implemented as two-layer waveguides, of the type described in Chap. 4 (see Fig. 4.7), and digital traces never intersect the waveguides in their lower reference plane. For instance, it can be seen from Fig. 8.2 that ten switchboard control signals (eight multiplexer controls and two receive/transmit controls on the rear side of the board) form a wide loop around the RF divider's section and fan out between shielded RF traces. These signals control the RF switches on the top side of the board, and the multiplexer waveguides are shielded from them by ground planes in the inner layers.

Importance of the regular geometry of the radiolocator's RF section was already discussed in Chap. 4, and all paths from antenna connectors to the tuner are of equal length (naturally, coaxial antenna cables are equal in length, too). Similar regularity was adhered to in the RF divider, although with less precision, since the transmission path is used for communication only and has no quantitative aspect.

The board has three grounds: digital, analog, and a separate analog ground for the tuner's oscillator circuit. The two main grounds meet in the power supply section, and are mutually AC-isolated with choke coils. Power and ground of the oscillator derive from the analog power/ground system, but are AC-isolated from it, and the oscillator section does not physically overlap with any of the other layers of circuitry.

The board has four digital power supplies, at 5, 3.3, 2.5 and 1.2 V; the last two are required by the internal logic of the FPGA alone. These voltage levels were obtained by cascading linear regulators.

There are four analog power supplies on the board, 5 and 2.7 V, two apiece. Each power supply is backed by its own linear regulator, and a set of tantalum bypass capacitors is placed next to each regulator. Two of these supplies are dedicated to the low-noise amplifier and the tuner's PLL, respectively, following the recommendations of the parts' manufacturers.

Overall, digital and analog power distributions (power and ground) are AC-isolated from each other with 330 μH choke coils. Lines from the power connector pass first through a 100 MHz common-mode choke, designed to fend off some of the commonly encountered power-line noise. The device is powered at 7 V DC, and draws a current of ca. 0.3 A.

8.2.1 Subsystems

The radiolocation device consists of the following subsystems: RF switchboard (encompassing the multiplexer and the divider), tuner with local oscillator, measurement unit (power meter and ADC), controller FPGA with the configuration PROM and JTAG loader, and the USB communication interface.

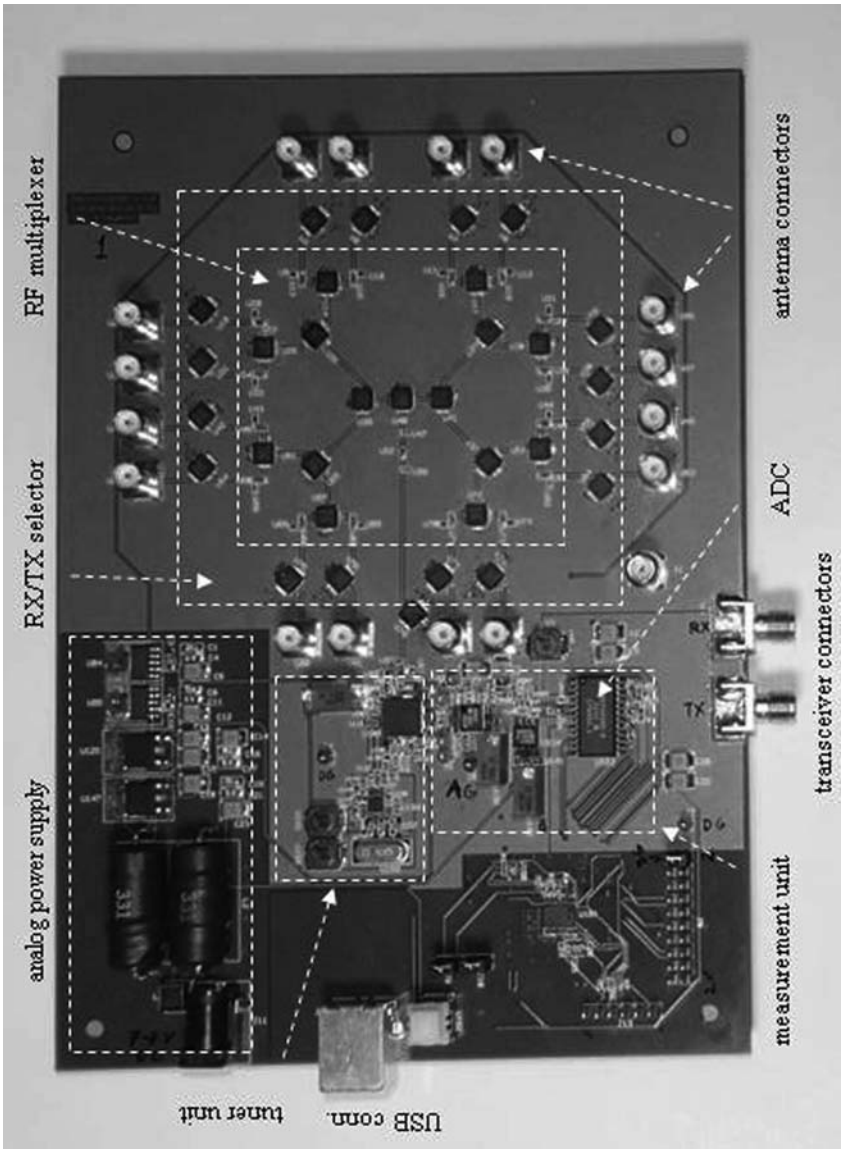

Fig. 8.1 Front side of the radiolocator board

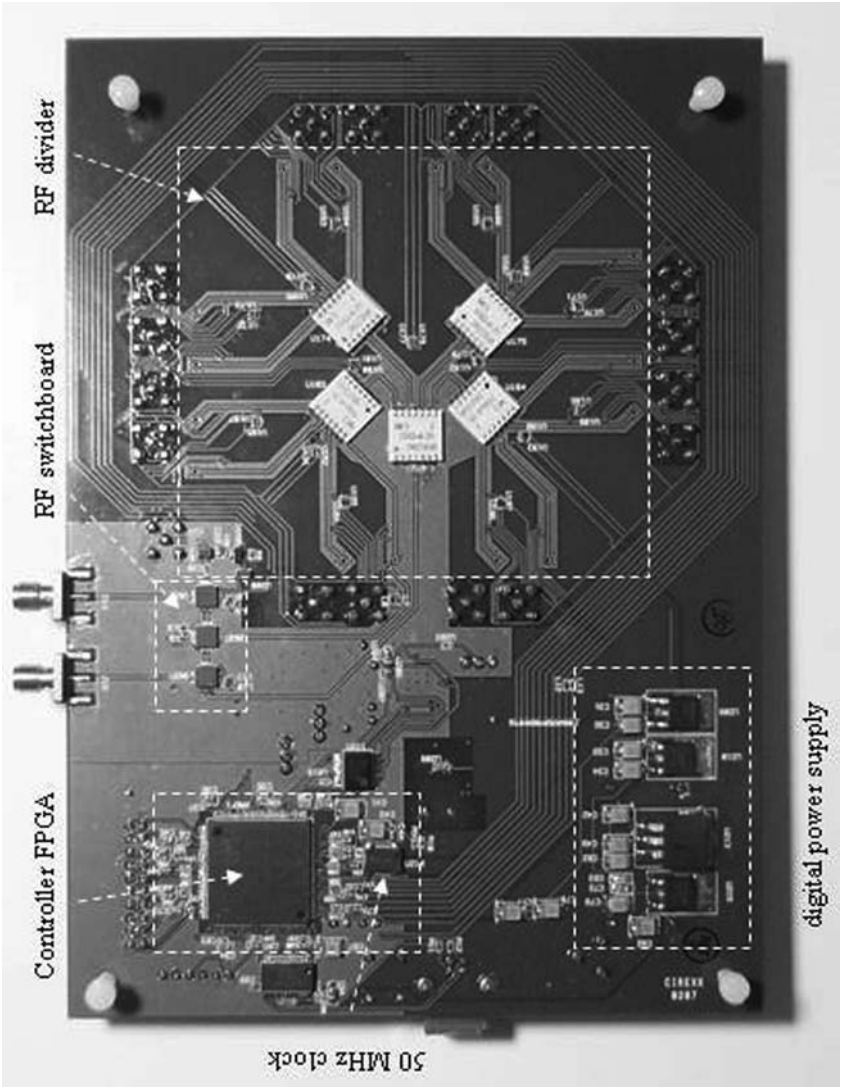

Fig. 8.2 Back side of the radiolocator board

8.2.1.1 Controller FPGA

The programmable chip, Spartan-3 XC3S200 by Xilinx (see Xilinx 2006), draws power at three voltage levels: 3.3, 2.5, and 1.2 V. This part is potentially the largest source of digital noise on the board; therefore, its supply pins are outfitted with a total of 45 ceramic and tantalum bypass capacitors, in the recommended value sequence of 10, 0.1, and 0.01 μF, and with smallest capacitors placed as close to the chip as possible. Vias connect the chip's ground pins directly to the underlying digital ground plane. The FPGA is clocked by a 50 MHz digital quartz oscillator, which is also placed as close to it as possible (see Fig. 8.2).

The controller subsystem includes a configuration PROM, from which the FPGA is configured on startup, and a JTAG loader. This circuitry is standard and is well described in the manufacturer's documentation (Xilinx 2004).

8.2.1.2 USB Communication Interface

The device's digital communication interface is implemented by the part FT245R by FTDI Ltd (see FTDI 2005). This part presents the USB 2.0 peripheral on the connector side, and a straightforward 8-bit bus on the board side. The bus runs into the controller FPGA, where it is met by the ASIC section that implements the bus interface and arbitrates the communication with other parts of the controller circuit.

On the other end of the USB communication line, the support workstation provides the required USB master. In our case, the workstation runs a device driver that presents the USB link to the operating system as a serial port, and the net result is a 0.5 Mbit/s serial communication link to the radiolocator board.

8.2.1.3 Radiolocation Tuner

The 802.11b tuner MAX2820 (see Maxim 2003), which we have already discussed in Chap. 4, is permanently configured for the receiving mode. The passband signal from the RF multiplexer is in one-sided form; it is converted into differential form by a balun transformer before entering the receiver. The tuner yields differential baseband, as in-phase and quadrature components; since the power meter is interested in amplitude only, we use only the in-phase component for power measurements.

The tuner generates its mixing frequency (2.4–2.5 GHz) by frequency multiplication, using an internal, programmable phase-locked loop, and an external sine-wave oscillator. We implemented the oscillator with a series-resonant quartz crystal with fundamental frequency 22.1184 MHz as reference, and the integrated oscillator circuit MAX2620. This latter part combines the equivalent of a bipolar (NPN) transistor providing positive feedback (negative resistance) for the oscillations, and an amplification stage for the generated sine signal. The configuration used is common collector, i.e., the quartz crystal is placed between the base and the emitter of the transistor (see the data sheet for MAX2620, Maxim 2002).

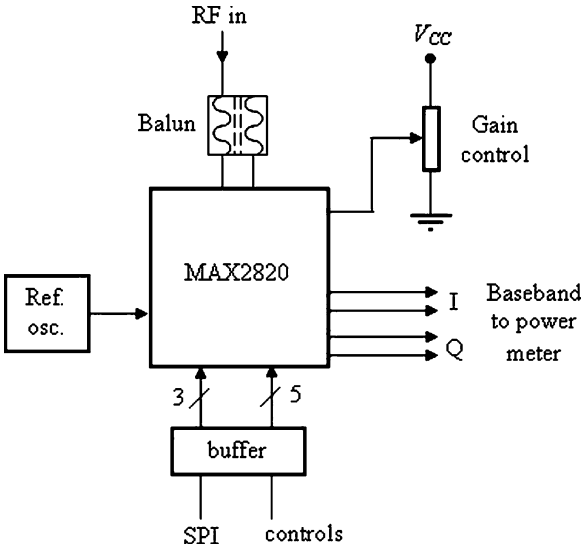

Fig. 8.3 Radiolocator's tuner and support circuitry

The oscillator is placed in close proximity of the tuner, with short, capacitively coupled feeds to it. As already mentioned, oscillator's power and ground are AC-isolated from the rest of the circuitry, and the layout of the tuner/oscillator circuit is shown in Fig. 8.4.

Figure 8.3 shows a very simplified schematic of the radiolocator's tuner section: the main part, surrounded by support systems (see also Fig 8.4). Voltage on the tuner's automatic gain control pin is controlled with a potentiometer, allowing for manual adjustments of the tuner's dynamic range. The tuner has a serial peripheral interface (SPI), which is used to write to the part's configuration registers; among other things, channels are changed by writing the frequency value to a configuration register. In addition, the tuner has several control pins, used for settings that may require rapid change: shutdown, receiver on/off, transmitter on/off, high-gain/low-gain mode. All of the tuner's controls are generated by the controller FPGA.

For a detailed example of tuner implementation, with guidelines for board design and a schematic of the supporting circuitry, see the manufacturer's description of the MAX2820 evaluation board (Maxim 2002).

8.2.1.4 Measurement Unit

The measurement unit consists of the power meter, a range-matching amplifier and an analog-to-digital converter (ADC). Since this subsystem is central to radiolocation, we have explained its functioning in detail in Chap. 5. The major engineering concern in this subsystem is containing the digital noise.

The ADC is, of course, the point at which digital switching noise is most likely to leak into analog circuitry. In fact, switching of the ADC's control signals is visible on the analog input, but the switching spikes are narrow, and they do not appear

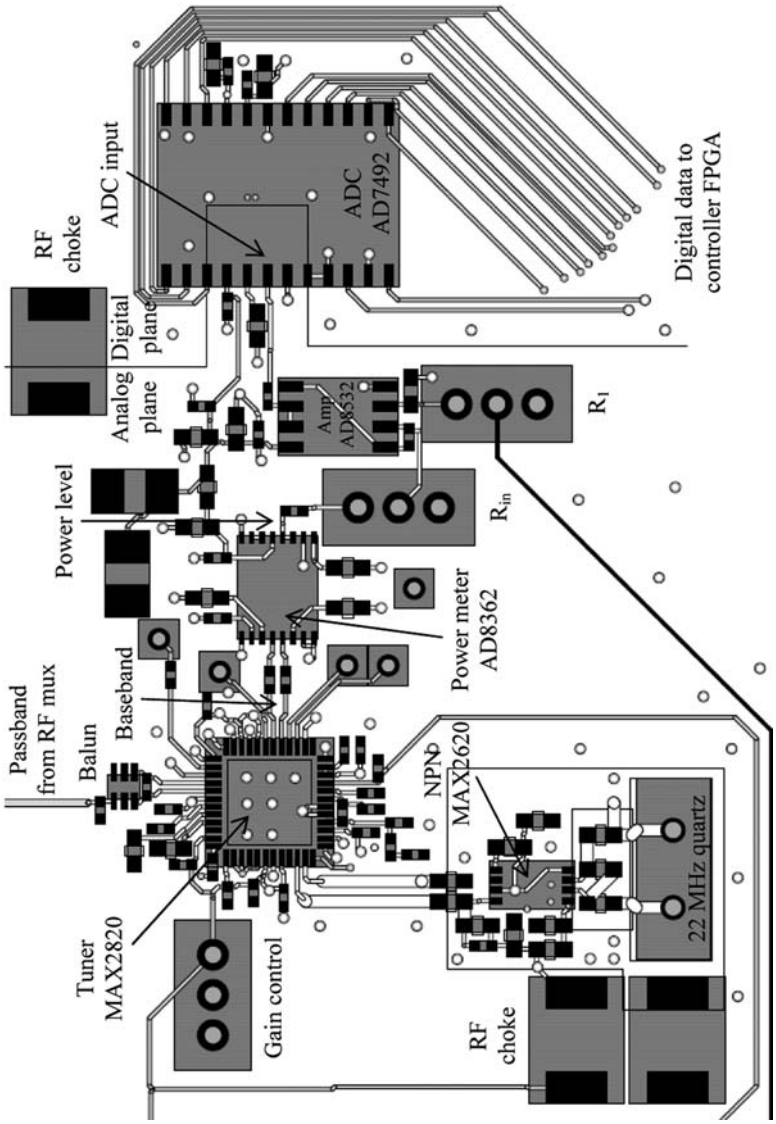

Fig. 8.4 Board layout of the tuner and power-measurement unit. Notice the sequestered power supply and ground plane of the 22 MHz reference oscillator, and the separation of digital and analog sides of the ADC. Tuner's digital controls are on the back side, isolated by inner-layer ground planes. As can be seen in Fig 8.1, traces in this area are surrounded with top-layer reference planes; these top-layer planes were removed from the figure for clarity

during the times at which the analog signal's value is stabilized and read in. To minimize the digital noise, certain precautionary measures were taken in the layout of the ADC circuit:

The ADC, part AD7492, uses 5 V analog and digital power supplies for the digital conversion process; in addition, it has a digital supply for its outputs, which is set to 3.3 V to match the controller FPGA's input voltage levels. All power supplies are provided with tantalum bypass capacitors (series of 1 nF, 0.1 μF, 10 μF, 47 μF), according to manufacturer's guidelines (see Analog Devices 2001).

The part straddles analog and digital reference planes. Digital signals are routed away from the part, and are shielded by the top-layer digital reference plane; analog power and input are physically separated from the digital side as much as possible, and are in turn shielded by the top-layer analog reference plane; see Fig. 8.4 for the layout of the ADC circuit.

This ADC part is adversely affected by imbalances in its analog and digital power voltages (or grounds). For that reason, analog and digital grounds were DC-shorted in the vicinity of the part with a 56 μH choke coil.

In Chap. 5 we have already described the two-stage amplifier, which matches the dynamic ranges of the power meter and the ADC, and provides a low-impedance source for the ADC. Equations (5.1) to (5.3) give the amplifier's bias and gain as functions of the resistor values. By iteratively adjusting the variable resistors R_1 and R_{IN}, the desired input dynamic range is easily mapped onto the 0–255 range of the ADC. In the prototype, the resistors have the values:

$$\begin{aligned} R_1 &= 500\ K \\ R_2 &= 100\ K \\ R_{\mathrm{IN}} &= 5\ K \\ R_L &= 2\ K \end{aligned}$$

8.3 CalRadio Transceiver

Figures 6.6 and 7.6 show an 802.11b transceiver as part of the radiolocation device, providing the communication functions. This transceiver is CalRadio, an open platform for wireless research and development, developed by Cal-$(IT)^2$ at University of California, San Diego. The transceiver implements the physical layer in hardware, and the MAC and higher layers on a dual microprocessor, as indicated in Fig. 6.6. This section describes CalRadio's main components, with emphasis on integration with the radiolocator.

8.3.1 DSP Hardware

The DSP half of CalRadio' microprocessor TMS320VC5471 is designed to run fast, compact and repetitive signal-processing code. Typically, it boots directly into

its functional code, and the boot-up is coordinated by the already running ARM half of the microprocessor. Main attraction of the DSP are its fast, synchronous serial ports, the McBSP's, backed by direct memory access (DMA) channels, which allow moving data packets between the baseband processor and DSP's memory fast and with a degree of control over the timing of the transfers. See Texas Instruments (2000, 2002) for an overview of the DSP.

The two processors do not share the memory space: they have separate data/address buses, and in this device, they utilize (quite reasonably) separate physical memories. The DSP and ARM communicate through a 16-Kbyte block of shared on-chip memory, and each side can raise an interrupt signal to the other, indicating that there is something new in the shared memory. In this arrangement, the Linux OS on the ARM processor sees the DSP as a peripheral, and interacts with it via a device driver of more or less standard Linux type.

This dual processor architecture makes it easy to perform fast low-level data transfers on the DSP side, and allows using the existing Linux networking stack and standard driver architecture on the ARM side. The price for this convenience lies in the fact that both the data stream and the coordinating instructions must cross a somewhat artificial asynchronous boundary (the two processor halves run on the same clock, after all!), by means of the cumbersome mailbox (shared-memory) mechanism.

Hardware interrupt requests (IRQs) drive the DSP data path described in the next section. Let us summarize them here:

IRQ 3, board ready – this is an external IRQ, raised by the radiolocator's controller; it indicates that a requested reconfiguration of the antenna switchboard has been completed.

- IRQ 6, timer – this is the timing signal of the DSP's hardware timer.
- IRQ 7, timestamp – input interrupt of the serial port McBSP0, raised for every word of available input.
- IRQ 12, ARM messages – raised by ARM to indicate new data in the mailbox.
- IRQ 15, DMA channel 4 – transfer of a block of data completed on the DMA.

8.3.2 DSP Data Path

Figure 8.5 offers an expanded, although still simplified, view into the upper right corner of Fig. 6.6, the DSP block of the receiver path. Demodulated baseband data arrive through the DSP's Multichannel Buffered Serial Port (McBSP). This is a clocked serial port, and the arriving data are made available to the internal bus in 16-bit words, in the port's output register, DRR (see Texas Instruments 2007, Ch. 2). We also see our familiar MD_RDY, the packet validation signal, which serves here to alert the port to the arriving data.

As new data become available in the DRR, an internal signal (REVT) alerts the DMA hardware subsystem to transfer the contents into a packet buffer in the regular memory. The DMA operates on blocks of data, and raises the interrupt IRQ 15 at the

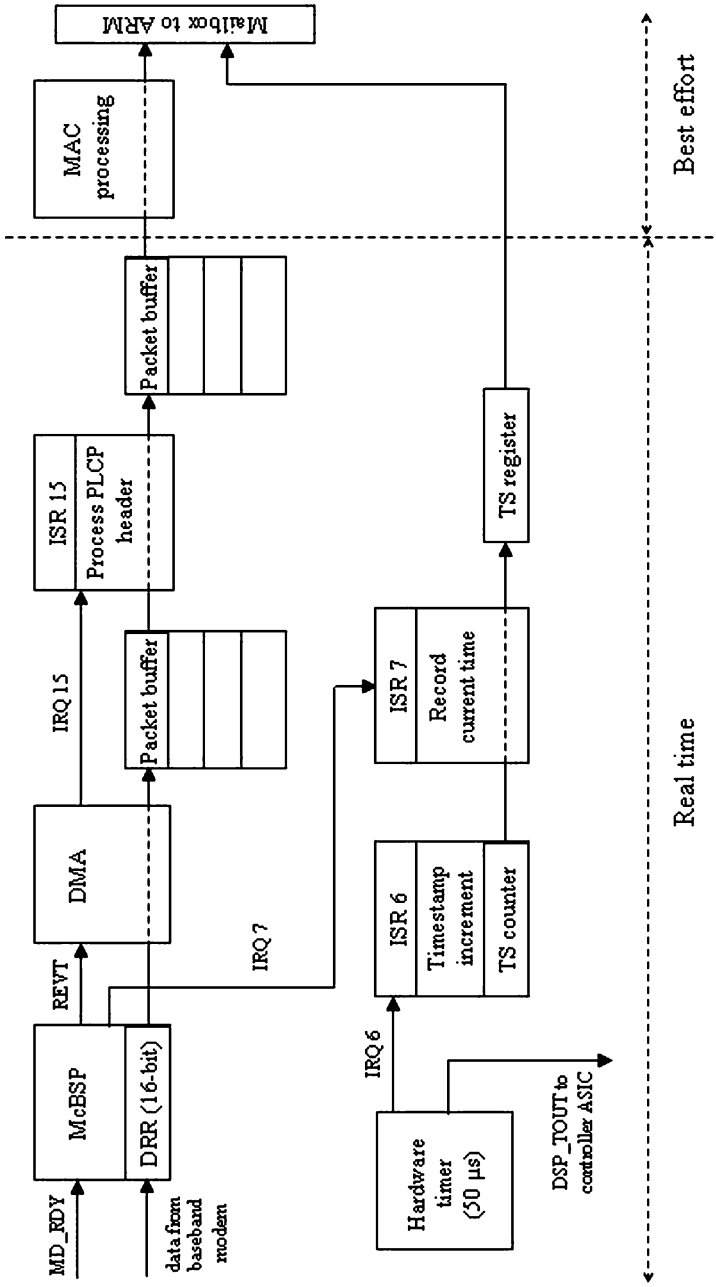

Fig. 8.5 DSP receiver data path

end of each transferred block (Texas Instruments 2007, Ch. 3). The interrupt service routine invoked by this interrupt, the ISR 15, performs some limited processing of the information in the PHY header, and hands the completed packet to the MAC process.

The MAC process is the background code running on the DSP, and as we discussed in Chap. 6, it is of variable (and hard to predict) duration – we draw the real-time boundary between ISR 15 and the MAC process in Fig. 8.5. Nevertheless, the DSP has no scheduler, and the delays in the background process can only be caused by the interrupt service routines; McBSP and the DMA are hardware subsystems, concurrent with the CPU.

The MAC process parses the frame types, implements the authentication/association protocol, and sends the data frames to the ARM processor, where the rest of the networking stack is implemented. It does so through the mailbox interface, the memory area shared by the two halves of the microprocessor. When data are ready in the mailbox, DSP raises an interrupt signal for the ARM processor, the ARM's IRQ 15.

In the lower portion of Fig. 8.5, the configurable hardware timer generates IRQ 6 pulses every 50 μs; these pulses are also present on a DSP's external pin (DSP_TOUT). Within the DSP, ISR 6 increments the timestamp counter, while the external signal increments the parallel timestamp counter in the controller ASIC.

When the serial port has ready data in the DRR, it raises the REVT and IRQ 7 signals simultaneously, one for the DMA and the other for the CPU. The ISR 7 records the value of the timestamp counter, at the time data became available, in the timestamp register. Since IRQ 7 occurs for every 16-bit word of data, only the packet's first IRQ 7 is made to record the timestamp.

Packet processing in the upper branch and the timestamp update in the lower branch are concurrent processes, started by the same REVT/IRQ7 signal. Processes in ISR 6 and ISR 7 are much shorter in duration, and of higher priority, than anything occurring in the upper branch; therefore, there is no danger of a race between these processes, a race that would result in a read/write collision on the timestamp register, or in stale timestamp being associated with packet data.

Interrupt service routines ISR 6 and ISR 7 could, in principle, yield a read/write collision on the timestamp counter. Fortunately, the interrupt management of the DSP does not allow ISRs to preempt one another: if the CPU is servicing an interrupt, incoming IRQs are held, and serviced according to their priority once the current ISR terminates. This feature protects the timestamp counter automatically, at the price of some jitter in the timing of ISR 6 and ISR 7. However, both of these routines have fixed execution times that are much shorter than the timer's 50 μs cycle, and the jitter does not cause any problems. For a full description of the DSP's interrupt system, see Texas Instruments (2001, Ch. 6). The transmit path is roughly the reverse of the above, excluding the timestamping. It is driven by IRQ 12, the mailbox interrupt, which hands over the LLC frame to the MAC. Ultimately, the DMA and the serial port ship the packet out to the baseband modem.

8.3.3 The ARM Processor and Data Path

The non-DSP half of TMS320VC5471 is an ARM7 processor architecture. The system that it embodies is best understood in terms of the data flow among the three peripherals that are available to this processor:

- The DSP itself is seen as a peripheral device, as discussed above; together with the RF transceiver and baseband modem, it appears to the ARM processor as something akin to a radio card.
- The 115 Kbit full-duplex RS-232 serial port is primarily the reception channel for the packets' bearings calculated by the supporting general CPU, as shown on Fig. 6.6. It also serves to transmit the (much less frequent) instructions to the radiolocation board.
- Ethernet interface (10/100 Mbit/s) is the device's wired networking port, as well as its diagnostic window.

The communication code runs under μClinux, entirely in the kernel space, as a kernel module. The module has four threads of execution, which we discuss below, and Figure 8.6 depicts the relationships among execution threads and data flows. This figure offers an expanded view of the ARM block in Fig. 6.6, and it also depicts the adaptive-response mechanism of Chap. 7.

Interrupt service routine associated with IRQ 15, the mailbox interrupt. This thread collects the incoming MAC-layer packets, and performs the LLC processing. Briefly, this means removing the header of the Logical Link Control layer and forming an 802.3 (Ethernet) packet, which under Linux is a data structure of type sk_buff (see, e.g., Rubini and Corbet 2001). The processed packets are written into one of a pair of circular buffers that are used to match packets with bearings.

Bearing thread, a scheduled thread that reads the serial port. Because of the operating system's restrictions, this thread is not an ISR for the serial port's interrupt, IRQ 6. Rather, it is a regularly scheduled thread that reads bytes directly from the data stream of the serial port, and relinquishes the CPU with sleep-waits when there are no data to read; buffering within the serial port is sufficient to prevent losses. Of course, this thread collects the timestamped direction data from the supporting CPU, and places these direction data into the second of the pair of circular buffers.

Matching thread is another scheduled thread. This tread repeatedly scans the circular buffers and extracts the oldest pair of data with matching timestamps. It updates the list of sources with the MAC address of the packet's source, and the bearing angle. After that, it hands the packet, which is by now in the 802.3 (Ethernet) form, to the Linux kernel call netif_rx(), which passes it up the networking stack.

Transmit thread; technically, this is a service routine that the kernel invokes when it has a 802.3 packet that is routed for transmission by the "radio card" interfaced by this kernel module; Linux name of that "soft" ISR is hard_start_xmit(). On the basis of the sources' list, the thread selects the transmission antenna nearest to the direction of the packet's destination and sends an instruction to that effect to the antenna switchboard via the serial port. Finally, it adds the 802.2 (LLC) header to the packet and sends it to the DSP's mailbox. The packet does not pass to the baseband modem

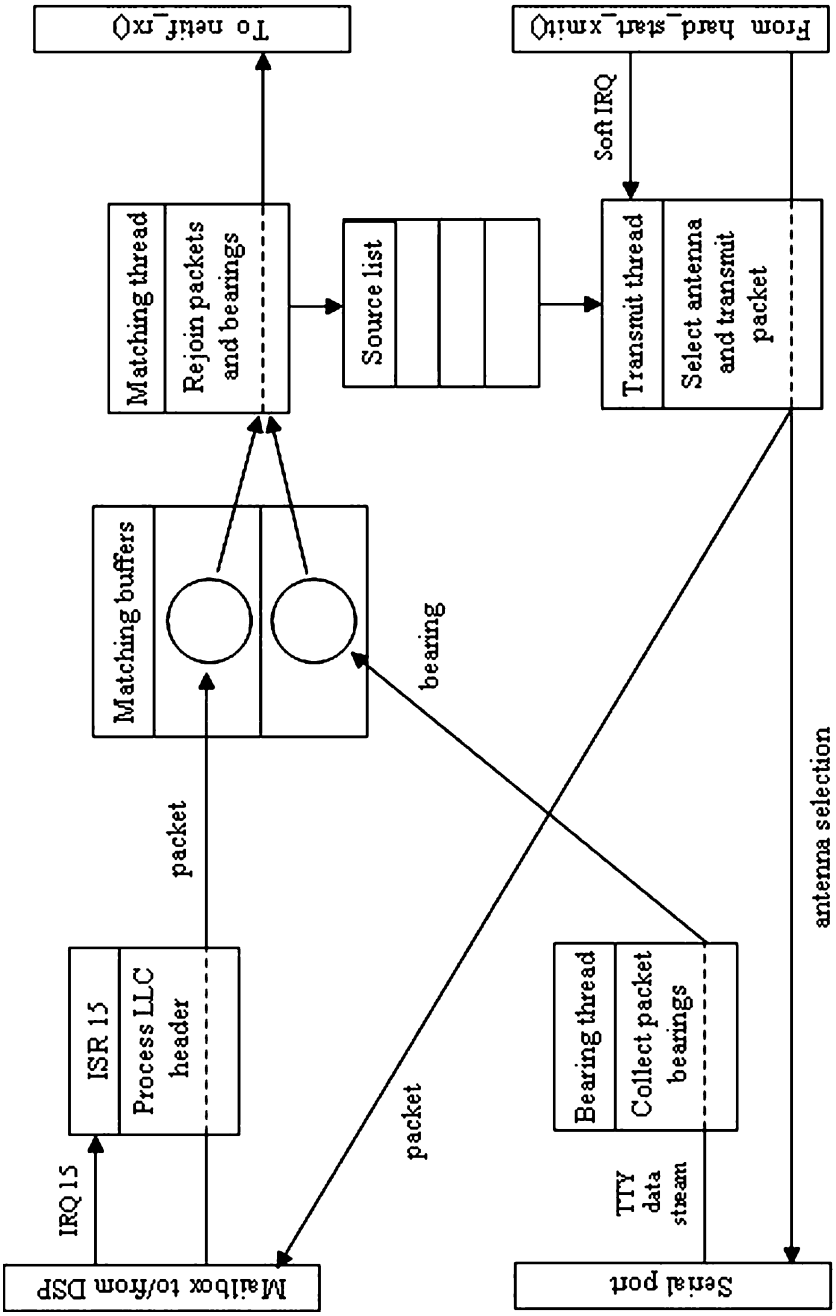

Fig. 8.6 ARM receiver/transmitter data path

until the DSP receives IRQ 3 ("board ready"), the confirmation signal from the switchboard that the antenna is selected and the switchboard configured for transmission.

Many good textbooks can be found on Linux kernel programming; we will merely point the reader to our preferred volume on device drivers (Rubini and Corbet 2001), and to a standard reference on the Linux kernel (Bovet and Cesati 2003).

8.3.4 Baseband and RF Sections

The baseband modem in CalRadio, the system that performs the translation between the analog baseband and digital data, is the part HFA3862 by Intersil Corporation (see Intersil Corp. 2000). This is a DSSS baseband processor that implements all the baseband functions of the 802.11b standard: spreading and despreading, the PHY-layer preamble, and all the required data rates and modulation schemes (see Sect. 10.6).

The baseband processor has three synchronous, half-duplex serial ports; two of these, transmit port and receive port, communicate with the (full-duplex) port McBSP0 of the DSP and transfer the digital data stream to and from the modem. On the analog side, the modem exchanges the baseband waveforms with the tuner, in-phase and quadrature, as four pairs of differential signals.

This part has a plethora of configuration registers (96 exactly), which are accessible via the third serial port, the control port. This port communicates with the port McBSP1 of the DSP, and that is how the configuration of the baseband processor is accomplished.

The tuner is MAX2820, the same part that is used in the radiolocator. We have already described this tuner in Sects. 4.3 and 8.2.1.3; its functioning and the surrounding circuitry are quite similar in CalRadio, except of course that its transmitter is enabled. Like the baseband processor, the tuner is configured by the DSP, through the port McBSP1; we should mention here that an Altera CPLD serves to route the DSP's port to either of these two peripherals.

We have limited this description of CalRadio to those aspects that are important for our integration into the adaptive radiolocator/transceiver. CalRadio is an appealing open-source, open-design platform for research and development, and we refer the reader to the developers' documentation for full details (Jow et al. 2007, UCSD Cal-$(IT)^2$ 2007).

8.4 The Laboratory Prototype

Figure 8.7 shows the laboratory prototype of the radiolocator/transceiver with adaptive response. The device's dominant feature is, of course, its "compound eye," the 16-fold ring of helical antennas. The compound antenna is sizable: 58 cm in

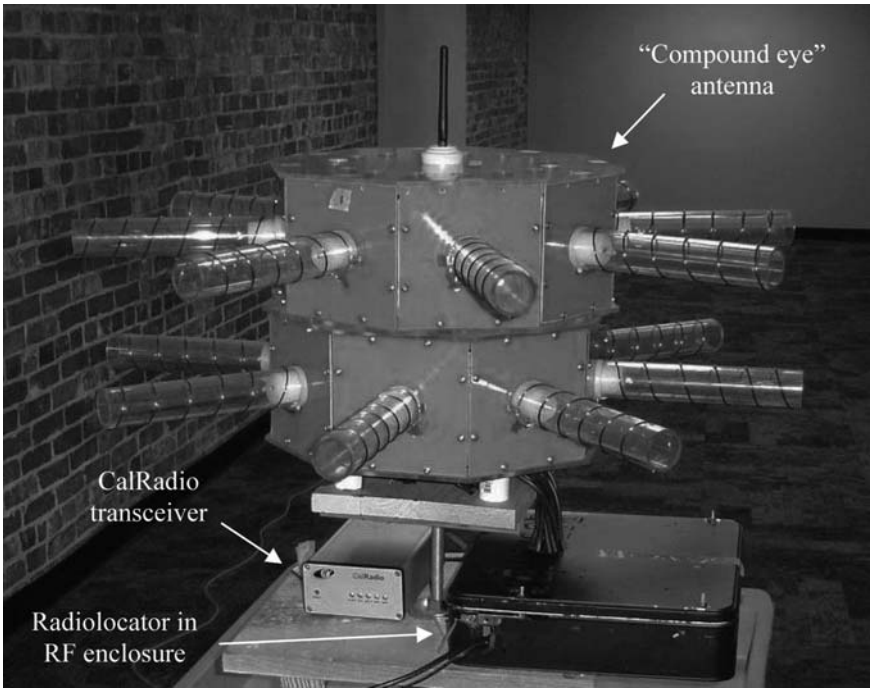

Fig. 8.7 Laboratory prototype of the radiolocator-transceiver with adaptive response

diameter, measured to the tips of the helices; and 28 cm in height. The helical element is easy to design and manufacture, and the axially symmetrical lobe and circular polarization have certain functional advantages, which we discussed in Chap. 2. This antenna design is suitable for a wide-area monitoring and communication installation. Use of the more compact patch antenna elements, and engineering trade-offs in gain and resolution, can reduce the size of the compound antenna, but, as we discussed in Chap. 1, dimensions of a directional antenna are fundamentally limited from below by the wavelength, which is 12 cm in this case (see also Sect. 9.8).

The radiolocator resides in a steel RF enclosure. This is a standard requirement for radiolocation devices, and in this case, the circuit proved quite sensitive to stray radio receptions. Even cable openings in the enclosure, when facing the source, caused distortions in radiolocation data. Therefore, the RF section of the radiolocation board itself has an additional grounded shield placed directly above it; this shield is within the RF enclosure, and is not visible in Fig. 8.7. Sixteen coaxial cables of equal length connect the compound antenna with the radiolocation board.

The transceiver has its own RF enclosure, and it causes no discernible RF interference or line noise in the radiolocator. These two subsystems have separate power supplies, which meet only at the 120 V AC outlet.

References

Data sheet for AD7492, http://www.analog.com/, Analog Devices (2001)
Bovet, D.P., Cesati, M., Understanding the Linux Kernel, 2nd Edition, O'Reilly (2003)
CalRadio 802.11b Development Platform, General Specifications, http://calradio.calit2.net/, UCSD Cal-$(IT)^2$ (2007)
Data sheet for FT245R, http://www.ftdichip.com/, Future Technology Devices International (2005)
Data sheet for HFA3863, File number 4856, Intersil Corp. (2000)
Jow, A., Schurgers, C., Palmer, D., CalRadio: A Portable, Flexible 802.11 Wireless Research Platform, Proceedings of MobiEval'07, San Juan, Puerto Rico (June 2007)
Data sheet for MAX2620, http://www.maxim-ic.com/, Maxim Integrated Products (2002)
Data sheet for MAX2820/MAX2821 evaluation kits, http://www.maxim-ic.com/, Maxim Integrated Products (2002)
Data sheet for MAX2820, http://www.maxim-ic.com/, Maxim Integrated Products (2003)
Rubini, A., Corbet, J., Linux Device Drivers, 2nd Edition, O'Reilly (2001)
TMS320C54x DSP, Functional Overview, data sheet SPRU307A, Texas Instruments (2000)
TMS320VC5471, Fixed-Point DSP Data Manual, data sheet SPRS180C, Texas Instruments (2002)
TMS320C54x DSP, Reference Set, v. 1: CPU and Peripherals, data sheet SPRU131G, Texas Instruments (2001)
TMS320C54x DSP, Reference Set, v. 5: Enhanced Peripherals, data sheet SPRU302B, Texas Instruments (2007)
Spartan-3 Starter Kit Board User Guide, UG130, http://www.xilinx.com, Xilinx (2004)
Spartan-3 FPGA Family: Complete Data Sheet, DS099, http://www.xilinx.com, Xilinx (2006)

Chapter 9
Wider Application of Radiolocation in Digital Wireless Communication

9.1 Introduction

In this final chapter, we look ahead at the application of our radiolocation methodology beyond the 802.11b framework, and in doing so, we retain our main objective of radiolocating every wireless packet in real time.

Any analysis of the incoming wave must take into account the fact that the wave is modulated. In radiolocation based on the signal strength, this means that amplitude changes must not corrupt the measurements required to determine the wave's direction. As we discuss in Sects. 5.4 and 10.6.2, packets in the 802.11b standard are purely phase-modulated, offering ample time to serially collect signal-strength data from multiple detectors (antenna elements); variations in the wave's amplitude are insignificant from one measurement to the next.

Other wireless standards include amplitude modulation, and finding a part of the packet's waveform that is suitable for power measurement becomes the central issue. There are two parts to this issue: first, the appropriate segment of the waveform must be detected in real time, and second, signal strengths from multiple detectors must be collected within it, without modulation interference. The more complex the waveform, and the larger the number of radiolocation antenna elements, the more stringent will be the requirements on the architecture of the radiolocator. In this chapter, we look at some of the prevalent wireless standards, and draw conclusions about feasible radiolocation strategies.

9.2 Frequency Hopping 802.11

In the 802.11 standards, frequency hopping (FH) is a less-used, somewhat obsolete form of spread-spectrum transmissions; the more common ones being DSSS, described in Sect. 10.6.2 and OFDM (Sect. 9.5 below). In the frequency-hopping protocol, the transmitter changes its carrier frequency ("hops") among frequency channels, in a particular predefined sequence, and the receiver follows. The 802.11 standard specifies 79 channels, 1 MHz wide, in the 2.4–2.5 GHz band, and it also prescribes the hopping sequences. Prescribed dwell time in any one channel is ca. 0.4 s.

D. Antolovic, *Radiolocation in Ubiquitous Wireless Communication*,
DOI 10.1007/978-1-4419-1632-7_9,

The modulation used in frequency-hopping 802.11 is GFSK (Gaussian frequency shift keying), that is, changing the carrier frequency, as in regular FSK, only somewhat gradually in time, in order to prevent excessive broadening of the modulated carrier's spectrum. The modulation alphabet consists of two or four discrete deviations from the frequency of the carrier, encoding one or two data bits at a time (Gast 2005).

This modulation does not affect the signal's power level much, and the dwell time in one channel is substantial; therefore, even a relaxed serial implementation of radiolocation per frequency hop can be easily implemented.

However, a monitored radio source hops among 1-MHz channels, and other sources traverse the same channels using different hopping sequences. A radiolocator monitoring a single channel cannot update the source's location per every transmission – it will receive an update only when the source hops into that channel again. For per-packet monitoring, the radiolocator must know the hopping sequence followed by the monitored source. The available sequences are of course publicly known, and FH access points broadcast their current hopping state in their beacon frames. Client nodes synchronize their hopping with the access point during the authentication/association steps (see Sect. 10.6.3).

The radiolocator must either capture the source's beacon frame (if the source sends any) or its association/authentication exchange with an AP, in order to discern the source's hopping sequence; or alternatively, radiolocator can monitor all the channels and deduce the source's sequence from a few of its hops.

This point is moot in radiolocator access points, which steer the beam towards their clients. The access point sets the hopping sequence for its own infrastructure network, and therefore knows the hopping state of its interlocutors.

9.3 Bluetooth

Frequency hopping is alive and well in the Bluetooth specification, the communication protocol that has come to be associated with short range, autonomous communication among portable devices. Bluetooth version 1.2 has been canonized as the IEEE standard 802.15.1.

Bluetooth uses the same 79 channels of the 2.4 GHz band as FH 802.11, albeit with a shorter dwell time of 625 μs, and it uses a combination of GFSK and PSK modulations. Hopping sequences are again coordinated by initial exchange of appropriate management packets between the interlocutors (Bray and Sturman 2001). Bluetooth packets have a guaranteed preamble and header, lasting ca. 120 μs, and the per-packet collection of radiolocation data is straightforward. As in FH 802.11, following the hopping sequence of the monitored source is the main issue.

Unlike 802.11, Bluetooth networks are organized as mastered buses, coordinated by master nodes, and these, in turn, can be organized into tree structures of larger networks; the network structure can change dynamically and autonomously. Since Bluetooth operates at short distances, and the devices are usually small in

size, equipping Bluetooth nodes with radiolocation and beam steering would be impractical. Monitoring Bluetooth networks by means of passive radiolocation could be of interest, however. Since the network configuration can change rapidly, the monitoring device should follow the association traffic of the master nodes, in order to build its local map of the current topology of the network, and to follow the hopping sequences of the monitored nodes.

9.4 802.11g

The last-mile wireless link, serving mobile devices and usually loosely labeled "wireless" or "Wi-Fi" is a moving target. In the ongoing quest for higher transfer rates, the field has seen several standards and a number of modifications within the standards, and the quest continues. However, much of the commercially available equipment currently covers the combination of 802.11a, -b and -g protocols.

The 802.11g operates in the same 2.4 GHz band as 802.11b, and it could be viewed as a hybrid, in that it uses the OFDM encoding (described below), but it also provides for a mode in which there is a preamble and header almost identical to that of the 802.11b standard (Gast 2005).

From the viewpoint of radiolocation, the backward-compatible preamble is BPSK-encoded, and can be used to collect radiolocation data. The short (72 μs) preamble is all that the standard requires, and if the data collection is serial, the design must strike the balance between the measurement speed and the number of antenna elements.

The so-called ERP-OFDM mode, which does not require the backward-intelligible header, is structurally the same as the 802.11a protocol, which we discuss in Sect. 9.6.

9.5 Orthogonal Frequency-Division Multiplexing

Orthogonal frequency-division multiplexing (OFDM) is an ingenuous and rather complex encoding scheme, designed to increase the utilization of a given band of the spectrum. It does that by modulating multiple (virtual) carriers in parallel, while using only one physical carrier for transmission. OFDM is used in 802.11a and 802.11g standards, and we describe it briefly here (see, e.g., Litwin and Pugel 2001).

Figure 9.1 shows how the actual (physical) radio signal is synthesized in the transmitter; the receiver is conceptually the same, with the function of every stage reversed. The input stream of n bits is broken up into N bit groups at a time, and the groups are encoded (in parallel) as distinct complex numbers Z_i. The numbers are represented by a discrete set of points in the complex plane, known as a

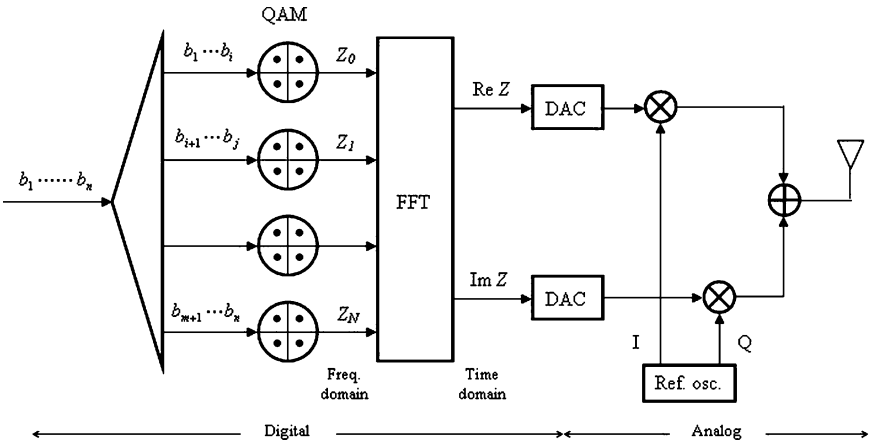

Fig. 9.1 OFDM transmitter. Sequence of input bits is broken into subsequences (groups), which are encoded in parallel, as complex numbers Z_i. Encoding need not be identical in all channels. Zs determine amplitude and phase of waves in a Fourier series (virtual carriers). Fourier waveform is synthesized, upconverted to passband, and transmitted. *DAC* digital-to-analog converter, *FFT* fast Fourier transform

constellation, and each point defines a unique amplitude and phase; this is known as the quadrature-amplitude modulation (QAM). For example, all possible sequences of six bits can be encoded by 64 points, usually selected in a somewhat symmetric pattern around the origin of the complex plane – this is called the 64 QAM.

These complex numbers, associated with groups of bits, are interpreted as the N coefficients of a Fourier series; they multiply N mutually orthogonal waves (virtual carriers) whose frequencies, ω_i, span the available bandwidth, and the time-domain waveform of the Fourier series is synthesized as:

$$Z(t) = \sum_{i=0}^{N-1} Z_i \exp(i\omega_i t) \tag{9.1}$$

The waveform (9.1), which persists for some fixed duration of time, encodes and transmits all of the n input bits at once. The low frequencies involved in baseband processing, and the considerable mathematical complexity of these calculations dictate that they be carried out in digital representation. An analog signal is synthesized from the digital Fourier waveform, upconverted to the passband, and sent to the transmitting antenna.

The receiver downconverts the passband, and, after tuning into the selected channel, performs the inverse Fourier transform on it. This yields the N complex coefficients Z_i, which are in turn decoded as bit groups, and the original stream

of n bits is recovered. We notice that the many Fourier coefficients, differing in phase, and changing in amplitude with the encoded bit stream, virtually guarantee the variable amplitude of the synthesized baseband signal.

9.6 802.11a

This protocol operates at 5 GHz, and specifies OFDM encoding throughout, with 52 virtual carriers covering each 20-MHz 802.11a channel. Even though some of the early fields in the packet specify BPSK modulation of the virtual carriers, the structure of the packet offers little by way of stretches of constant amplitude. Serial radiolocation architecture is all but precluded in this standard.[1]

Virtual carriers can be analyzed digitally, in order to extract a reliable measure of the signal strength. The standard specifies that, of the 52 carriers, 4 are pilot signals carrying a control sequence of bits in BPSK encoding. These constant-amplitude pilots offer a convenient means of determining the signal strength, except of course that they exist as individual signals only in the digital form, in the frequency domain of the Fourier transform. This requires parallel Fourier conversions of the signals from each antenna, since a serially chopped-up baseband cannot be reliably transformed into individual virtual carriers.

Analysis of the digitized signals also encounters a practical problem: digitization accuracy in the demodulation path is often sacrificed for sampling speed, which is more important of the two in tracking the shape of the baseband waveform. Just the opposite is true of radiolocation: relatively few samplings are needed per packet, but their accuracy must be good enough not to diminish the accuracy of the calculated bearing. The ADC's used will have to balance these requirements.

Within the parallel radiolocator architecture, amplitude modulation does not present an insurmountable problem, but a few caveats are in order. Power levels of the analog OFDM basebands from the individual antennas can be measured, and as long as they are measured within the *same* segment of the waveform, i.e., in the same time interval, amplitude variations will average out in the same way, and the power levels will reflect the varying antenna gains. Power readings will be different during a different time interval, but their ratios will remain the same, and that is all that is needed for amplitude-based radiolocation.

Nevertheless, composite signals like the OFDM baseband tend to have large amplitude variations (large crest factors), which increase the error in the power measurements. This can be compensated for by averaging the power reading over longer stretches of the packet, preferably once the power reading has been digitized by its own low-rate, high-accuracy ADC. Sketch of an appropriate parallel architecture is shown in Fig. 9.2.

[1] A related encoding called scalable orthogonal frequency-division multiple access (SOFDMA) is the basis of IEEE 802.16, or WiMax, a wireless standard that encompasses point-to-point as well as mobile devices, and operates at several microwave bands.

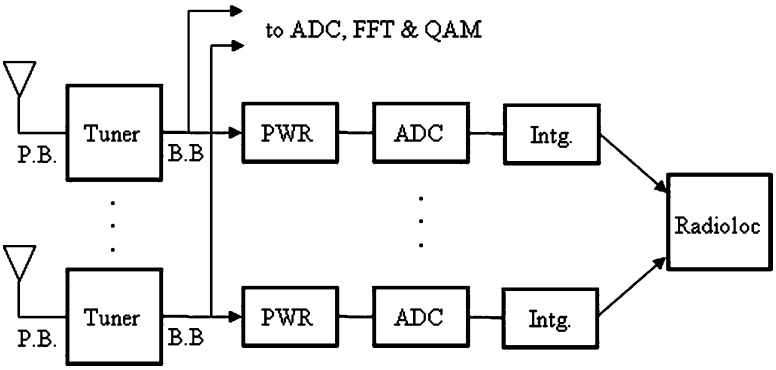

Fig. 9.2 Sketch of the architecture of the front-end parallel data collection for an OFDM radiolocator

9.7 Code-Division Multiple Access

The direct-sequence spread spectrum (DSSS) techniques broaden the spectrum of the transmission signal, by introducing an additional modulation with a pseudorandom sequence. This is usually referred to as chipping, and we describe the technique briefly in Sect. 10.6.2, in the context of the 802.11b standard.

In the narrow-band radio, a transmitter is identified by the frequency range of its spectrum, since, in the absence of interference or intentional jamming, it is the only transmitter in that range. DSSS techniques open up the possibility of multiple transmitters using the same broadened spectrum, isolated from one another not by the orthogonality of the sine waves in time (frequency separation), but by the orthogonality of their chipping sequences. This technique is referred to as code-division multiple access (CDMA) and it forms the basis of cellular (mobile) phone systems.

Radiolocation of CDMA sources is possible, but with an added requirement that we restrict the measurements of signal strength to that part of the signal which pertains to one orthogonal chipping sequence. We cannot separate the sources by frequency alone, since multiple sources can use the same frequency band simultaneously. Incidentally, this issue does not arise in 802.11b radiolocation, because Wi-Fi is not a CDMA scheme: every transmitter is assigned its own 20-MHz frequency channel, within which it practices its own spectrum spreading, using perhaps the CCK encoding for larger data throughput (see Sect. 10.6.2). Let us describe how CDMA transmission works; our notation follows that of the discussion given by Sklar (Sklar 2001).

The data stream and the chipping can be viewed as two distinct modulations imposed upon the carrier. In the PSK, which is the modulation scheme in CDMA, modulation amounts to multiplying the carrier with a function of the form $x(t) = \exp[i\varphi(t)]$, where $\varphi(t)$ is a step-wise time function with discrete values $\varphi_0, \varphi_1, \ldots, \varphi_N$, phase values encoding the bits of the transmitted sequence. In BPSK, the phases are selected as $\varphi \in \{0, \pi\}$, and $x(t) \in \{+1, -1\}$. We call the

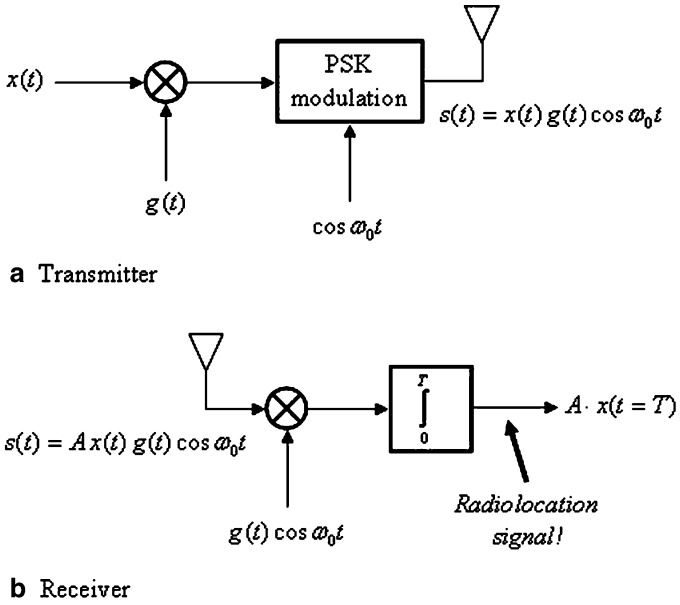

Fig. 9.3 CDMA transmission/reception path. The architecture of the receiver allows the capture of signal-strength data for a single CDMA channel

encoding of the data stream $x(t)$ and the (higher-baud) chipping sequence $g(t)$, and we assume that they are aligned at the symbol boundaries. The simple implementation is to digitally XOR the data and chipping streams, and impose the combined phase changes on the carrier, equivalent to the multiplication with $x(t)g(t)$ (see Table 10.1). Figure 9.3a shows this arrangement for the CDMA transmitter, where ω_0 is the carrier's frequency.

The receiver, for its part, detects a data stream waveform, modified by a chipping sequence $\bar{g}(t)$, potentially different from its own:

$$s(t) = A\, x(t)\, \bar{g}(t) \cos\, \omega_0 t. \tag{9.2}$$

The amplitude of this waveform is what is of interest for radiolocation, and we leave out the frequency tuning from this discussion for the sake of simplicity. Multiplying the input (9.2) with the locally generated waveform $g(t) \cos\, \omega_0 t$ and integrating over T, the time duration of one data symbol, yields:

$$z(T) = A \int_0^T \mathrm{d}t\, x(t)\, \bar{g}(t)\, g(t) \cos^2 \omega_0 t. \tag{9.3}$$

In this expression, $x(t)$ is constant over the time interval of the data symbol, and $g(t)$, $\bar{g}(t)$ are constant over the chipping intervals. We can break up the integration in (9.3) into a sum of integrals over the N chipping intervals per data symbol:

$$z(T) = A\,x(T) \sum_{i=0}^{N-1} \bar{g}_i g_i \int_0^{T_C} \mathrm{d}t\ \cos^2 \omega_0 t\,. \tag{9.4}$$

Integrals over the square of the carrier are positive, and roughly the same for all chipping periods because the carrier frequency is much higher than the chipping frequency.[2] The chipping sequences are mutually orthogonal by design:

$$\sum_{i=0}^{N-1} \bar{g}_i g_i = \begin{cases} N & \text{for} \quad \bar{g} = g \\ 0 & \text{otherwise} \end{cases}. \tag{9.5}$$

This orthogonality is, of course, what provides isolation between CDMA channels. The resulting data waveform is:

$$z(T) \approx A\,x(T) \tag{9.6}$$

for $\bar{g} = g$, and zero otherwise. The data waveform (9.6) is what carries the signal strength of a single transmitter, and that waveform must be used for radiolocation. Unlike the other wireless standards discussed here, we could not simply capture the strength of the baseband, because it potentially contains the mixture of signals from several transmitters. Figure 9.3b is the graphic representation of this CDMA receiver path.

It is possible to downconvert the waveform in (9.2) to baseband, by mixing it with cos $\omega_0 t$ alone, perform the PSK demodulation (detect all phase changes), and recover a digitized, chipped data stream. The original data stream can be recovered by XOR-ing with the digital chipping sequence, but this resulting data stream cannot be used for radiolocation: being expressed as digital bits, it had lost the signal-strength information before a particular CDMA channel was selected.

9.8 Summary

This account of wireless standards is by no means complete, and it would be impractical to make it so. Some standards (e.g., 802.11n) are still in a state of flux at the time of this writing, while others, such as WiMax, are merely complex in their

[2] It can be easily shown that $\int_{T_1}^{T_2} dt \cos^2 \omega\, t \ \to (T_2 - T_1)/2$ when $\omega \to \infty$, i.e. the square of the high-frequency carrier averages out to ca. 1/2 over any time interval that is long compared to the period of the carrier.

details, but introduce no new radiolocation issues. Our selection of standards and protocols in this chapter is meant to illustrate the engineering challenges ahead, and we can summarize, on the basis of our experience, that the following issues must be addressed in wireless radiolocation architectures:

1) The radiolocator must isolate (tune into) the signal of a single receiver, be it in the frequency domain or in the space of orthogonal pseudorandom sequences;
2) It must measure the signal strengths on all of its antenna elements on equal footing, either by performing serial measurement on waveforms of guaranteed constant amplitude, or by measuring all the strengths in parallel, within the same time interval;
3) Closely related to the above, it must complete the measurements on all antennas within one packet;
4) It must coordinate its finding (the bearing angle) with the actual digital communication, so as to assign that bearing to the right source.

Various architectures – parallel and serial – and combinations of analog and digital processing are at our disposal to address these requirements.

Finally, a word should be said about the antenna design. We see the "compound eye" design as highly promising for radiolocation of mobile communication devices, due to its wide and isotropic aperture, and its instantaneous collection of spatial data. The operating wavelength determines the size of the antenna element, and in the 12-cm range (2.4 GHz), commercial patch antennas are available that are ca. 8.5 cm in diameter, and have respectable main lobes with the −3 dB half-width of ca. 35 degrees. A compound antenna of the double-ring type, similar to that shown

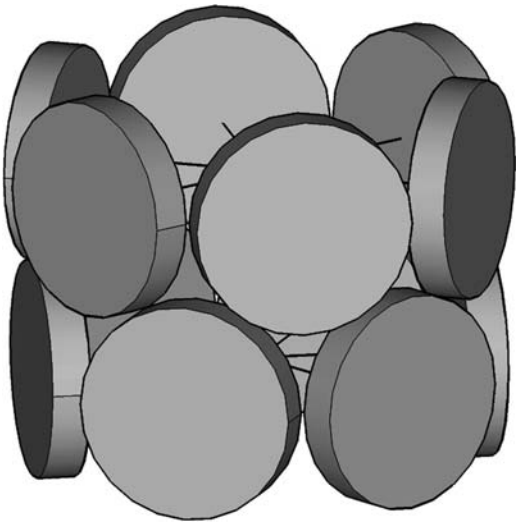

Fig. 9.4 Compact configuration of a double-ring compound antenna, made of 12 circular flat panel elements

in Fig. 8.7, and made of 12 such elements, will be ca. 19 cm in diameter, and 15 cm high. That is the size of a tea kettle (for a lack of better comparison), and with proper design, the antenna can enclose its front-end electronics, and form a very practically sized radiolocator access point (see Fig. 9.4).

Commercial patch antennas are available in the 5 GHz range that are even somewhat smaller, as we would expect. We give an analysis of the applicability of some 2.4 GHz commercial antenna models for radiolocation in (Antolovic 2009).

References

Antolovic, D.: Numerical Investigation of Algorithms for Multi-Antenna Radiolocation, Proceedings of 2009 IEEE International Conference on Portable Information Devices, Anchorage, AK (September 2009)

Bray, J., Sturman, C.: Bluetooth 1.1: Connect Without Cables, 2nd Edition, Prentice Hall (2001)

Gast, M.S.: 802.11 Wireless Networks, 2nd Edition, O'Reilly (2005)

Litwin, L., Pugel, M.: The Principles of OFDM, RF Signal Processing, www.rfdesign.com (2001)

Sklar, B.: Digital Communications, 2nd Edition, Prentice Hall (2001)

Chapter 10
Appendices

10.1 The Laplacian Operator

As we said in Sect. 1.1, Laplacian relates the value of a field at a point in space to the field's average value around that point. To see this, let us imagine a small cube with side a, centered around the given point; we will also make this point the coordinate origin, for simplicity. The average value of a scalar field φ over the cube is:

$$\overline{\varphi} = \frac{1}{a^3} \int_a \varphi(\mathbf{x}) \mathrm{d}x_1 \mathrm{d}x_2 \mathrm{d}x_3 \tag{10.1}$$

We expand the field in a Taylor series around the origin:

$$\varphi(\mathbf{x}) = \varphi(0) + \sum_i \left(\frac{\partial \varphi}{\partial x_i} \right)_0 x_i + \frac{1}{2} \sum_{i,j} \left(\frac{\partial^2 \varphi}{\partial x_i \partial x_j} \right)_0 x_i x_j + \cdots \tag{10.2}$$

and substitute the series into the integral in (10.1). Integration over any odd power of x_i yields zero; therefore, all linear terms vanish, as well as all quadratic terms for which $i \neq j$. Integration over zero-th power of x_i yields a, the dimension of the cube, and integration over x_i^2 yields factors proportional to a^3. The end result is:

$$\bar{\varphi} - \varphi(0) = \frac{a^2}{24} \left(\nabla^2 \varphi \right)_0 \tag{10.3}$$

Difference between the average and the central value of φ decreases with the size of the cube, as it must, but for a given cube size, it is proportional to the value of the field's Laplacian at the center.

Fields which satisfy the equation $\nabla^2 \varphi = 0$ are called *harmonic*. Such fields are always equal to their average, and obviously cannot have a true maximum or minimum, except at some boundary. A good example is the weightless soap membrane within a wire loop: it is under elastic tension only, its shape satisfies the harmonic equation, and cannot develop bulges, regardless of the shape of its boundary.

D. Antolovic, *Radiolocation in Ubiquitous Wireless Communication*,
DOI 10.1007/978-1-4419-1632-7_10,

10.2 Antenna Reciprocity

It is well known that the lobe pattern of an antenna has the same shape for both transmission and reception; this fact is often presented as obvious in engineering literature. It is indeed a very useful principle, which allows us to discuss the antenna theory, and measure the antennas' characteristics in whichever mode is more convenient: the results will always apply to the other mode as well.

In the radiation mode, the antenna's lobe pattern means power transmitted per unit solid angle, as a function of direction $\mathbf{n}$; it is proportional to $|\mathbf{A}(\mathbf{n})|^2$, where $\mathbf{A}(\mathbf{n})$ is the integral in (1.64) (Elliott 2003). In the case of reception, it means power delivered as alternating current on the antenna's terminals, in response to an incident plane wave from direction $\mathbf{n}$. For any two sets of measurements, the ratio of these two quantities is the same for all $\mathbf{n}$, although it is not immediately obvious that this must be so.

Let us discuss this topic in some detail in this section. We will derive a general theorem regarding oscillatory fields and currents, and use it to establish the reciprocal nature of an abstract model device. We will then describe the lobe measurement setup as such a device and prove the similarity of lobe patterns for transmission and reception.

10.2.1 Lorentz Reciprocity Theorem

Let us consider two sets of currents, $\mathbf{J}_1$ and $\mathbf{J}_2$, each separately giving rise to a pair of fields $\mathbf{E}_1$, $\mathbf{H}_1$ and $\mathbf{E}_2$, $\mathbf{H}_2$. The currents (and consequently the fields) in both sets are assumed to have harmonic time dependence $e^{i\omega t}$, with the same frequency. Maxwell's equations apply to each set, and they assume the form

$$\nabla \times \mathbf{E} = -i\omega\mu_0\mathbf{H} \tag{10.4}$$

$$\nabla \times \mathbf{H} = i\omega\varepsilon_0\mathbf{E} + \mathbf{J} \tag{10.5}$$

$$\nabla \cdot \mathbf{E} = 0 \tag{10.6}$$

$$\nabla \cdot \mathbf{H} = 0 \tag{10.7}$$

Consider now the formal expression:

$$\nabla \cdot (\mathbf{E}_1 \times \mathbf{H}_2) - \nabla \cdot (\mathbf{E}_2 \times \mathbf{H}_1) \tag{10.8}$$

These vector terms do not represent immediate physical quantities, because the two sets of fields need not be present simultaneously. As we shall see below, the reciprocity theorem is useful precisely when the fields do *not* appear simultaneously. Expanding the threefold products in (10.8), according to the vector identity

$$\nabla \cdot (\mathbf{a} \times \mathbf{b}) = \mathbf{b} \cdot (\nabla \times \mathbf{a}) - \mathbf{a} \cdot (\nabla \times \mathbf{b}) \tag{10.9}$$

and applying the $\nabla \times \mathbf{H}$ Maxwell equation, (10.5), we obtain

$$\nabla \cdot (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) = \mathbf{E}_1 \cdot \mathbf{J}_2 - \mathbf{E}_2 \cdot \mathbf{J}_1 \tag{10.10}$$

Formally again, scalar cross-coupling between the two sets of currents and fields forms the divergence of a vector cross- coupling between the two sets of fields. This is the differential form of the reciprocity theorem.

Remembering that divergence within a volume equals the field flux through the boundary surface of that volume, we rewrite (10.10) in the integral form, also known as the Lorentz' reciprocity theorem:

$$\oint_S (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathrm{d}\mathbf{S} = \int_V (\mathbf{E}_1 \cdot \mathbf{J}_2 - \mathbf{E}_2 \cdot \mathbf{J}_1)\mathrm{d}V \tag{10.11}$$

The integrations in (10.11) are over an arbitrary volume V and its bounding surface S. Applying this theorem to a volume excluding all sources, we obtain the following simplification:

$$\oint_S (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathrm{d}\mathbf{S} = 0 \tag{10.12}$$

This is the form of the reciprocity theorem that we will need in the arguments that follow.

10.2.2 Reciprocal Two-Port Device

Let us now envision a device of unspecified function, with two waveguides (ports) leading in and out of it. These ports can be true hollow waveguides, coaxial cables, or any other type of conduit, as long as we can assume a clean wave propagation mode along it. We also assume that the device has a linear response

$$\mathbf{V} = \mathbf{Z}\mathbf{I} \tag{10.13}$$

where $\mathbf{V}$ and $\mathbf{I}$ are column vectors of voltages and currents, with components counted over the ports, and $\mathbf{Z}$ is the device's impedance matrix. The discussion in this section applies to multiport devices, but two ports are all we need.

We can now envision two cross sections through the ports: port planes, which conceptually separate the device from the exterior, and two sources, a and b, placed outside the device, which generate input signals to the ports. We select an enclosing surface S, which consists of: (a) both port planes, (b) perfectly conducting boundaries of the device to which the electric field is perpendicular, and (c) an overall enclosing surface, chosen distant enough to assume that the fields on it are vanishingly small (see Fig. 10.1).

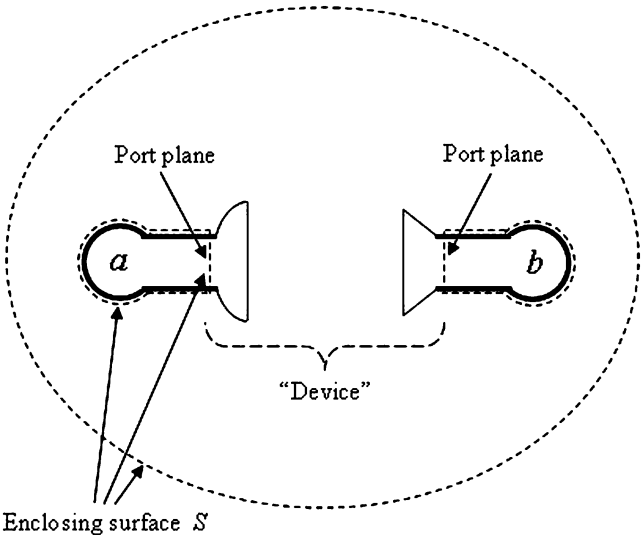

Fig. 10.1 Integration surface around the two-port device

We apply the sourceless reciprocity theorem, (10.12), to these sources and surfaces, since the sources a and b are outside S and the volume integral in (10.11) vanishes. The surface integrand vanishes on the conducting boundaries, since

$$(\mathbf{E} \times \mathbf{H}) \cdot \mathrm{d}\mathbf{S} = (\mathrm{d}\mathbf{S} \times \mathbf{E}) \cdot \mathbf{H} \equiv 0 \tag{10.14}$$

The surface integration therefore reduces to integrals over the port cross-sections $S1$ and $S2$:

$$\oint_{S1} (\mathbf{E}_a \times \mathbf{H}_b - \mathbf{E}_b \times \mathbf{H}_a) \cdot \mathrm{d}\mathbf{S} + \oint_{S2} (\mathbf{E}_a \times \mathbf{H}_b - \mathbf{E}_b \times \mathbf{H}_a) \cdot \mathrm{d}\mathbf{S} = 0 \tag{10.15}$$

According to the theory of waveguides (e.g., Pozar 1998), fields in a pure propagation mode can be written as

$$\mathbf{E} = V\mathbf{e} \tag{10.16}$$

$$\mathbf{H} = I\mathbf{h} \tag{10.17}$$

where $\mathbf{e}$ and $\mathbf{h}$ are two-dimensional vector fields describing the cross section, and V and I are effective voltages and currents, carried by the propagating wave. Transversal fields can be normalized, so that

$$\oint_{S} \mathbf{e} \times \mathbf{h} \cdot \mathrm{d}\mathbf{S} = 1 \tag{10.18}$$

over the cross section of the guide (the port plane), and the integral in (10.15) becomes, after representing V and I as column vectors again,

$$\mathbf{V}_a^T \mathbf{I}_b = \mathbf{V}_b^T \mathbf{I}_a \tag{10.19}$$

Casting (10.13) in the admittance form $\mathbf{I} = \mathbf{YV}$, where $\mathbf{Y} = \mathbf{Z}^{-1}$, and using it to eliminate the currents in (10.19), yields

$$\mathbf{V}_a^T \mathbf{Y} \mathbf{V}_b = \mathbf{V}_b^T \mathbf{Y} \mathbf{V}_a \tag{10.20}$$

A bit of matrix algebra shows that this relation will be true for arbitrary $\mathbf{V}$'s (and they *are* arbitrary) only if $\mathbf{V}^T = \mathbf{V}$, which also means $\mathbf{Z}^T = \mathbf{Z}$ or

$$Z_{12} = Z_{21} \tag{10.21}$$

that is, the impedance (admittance) matrix is symmetric. A signal propagates through the device from port 1 to port 2 with the same impedance as from port 2 to port 1.

10.2.3 Two-Antenna Measurement System

We now identify port 1 with the feed of the tested antenna, the one whose pattern we are interested in. Port 2 is the feed of a test antenna, which can be moved on a spherical surface around the antenna 1. We do not make assumptions about the type of either antenna at this point, and our two antennas form the two- port device whose Z_{12} changes with the position and orientation of antenna 2 (see Fig. 10.2).

Suppose that the source current a is active, i.e., current I_{a1} flows at the feed of antenna 1. Voltage measured at the "open" feed of antenna 2 (no current flowing in the receiver port) is, due to (10.13),

$$V_{a2} = Z_{12} I_{a1} \tag{10.22}$$

When the source current b is active, voltage measured at the feed of antenna 1 is

$$V_{b1} = Z_{21} I_{b2} \tag{10.23}$$

If measurements are performed for different positions of antenna 2, Z_{12} always equals Z_{21}, due to (10.21), although this impedance changes with the relative position of antenna 2; we also hold the source current constant. Therefore, measured voltages V_{a2} and V_{b1} are the same function of the direction θ, φ of the test antenna, multiplied by proportionality constant:

$$V_{a2}(\theta, \varphi) = \frac{I_{a1}}{I_{b2}} V_{b1}(\theta, \varphi) \tag{10.24}$$

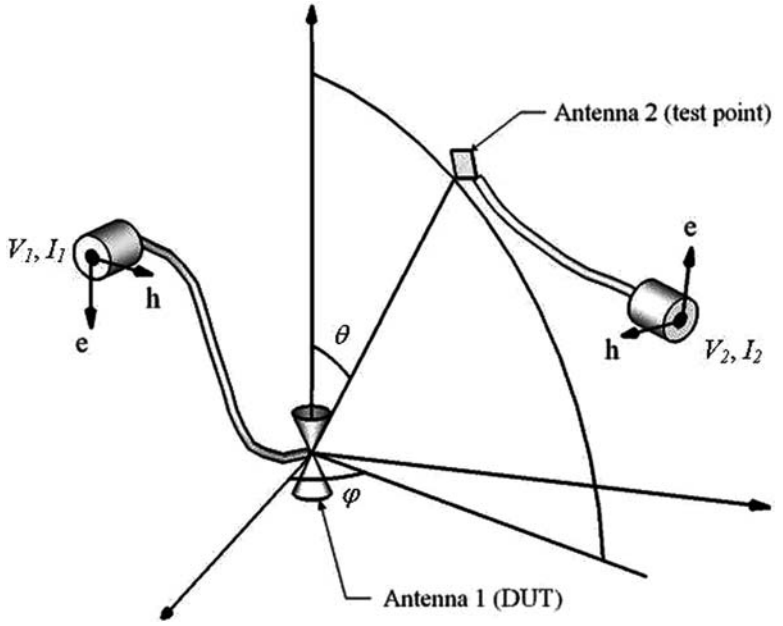

Fig. 10.2 Measurement of the radiation/absorption pattern of an antenna

For the sake of completeness, we also discuss the more realistic case of receivers attached to antenna feeds. The currents through the receiver ports are then $I_{a2} = -V_{a2}/Z_{r2}$ and $I_{b1} = -V_{b1}/Z_{r1}$, where Z_r are the receiver impedances. We use these equations and (10.13) to calculate V_{a2} and V_{b1}; we eliminate Z_{12}, Z_{21} from the resulting expressions by invoking (10.21), and obtain the formula:

$$\frac{V_{a2}(\theta,\varphi)}{I_{a1}}\left(1+\frac{Z_{22}}{Z_{r2}}\right)=\frac{V_{b1}(\theta,\varphi)}{I_{b2}}\left(1+\frac{Z_{11}}{Z_{r1}}\right) \tag{10.25}$$

We see that the ratio of received voltages V_{a2} and V_{b1} is again the same for all directions, and that (10.25) reduces to (10.24) for $Z_r \to \infty$. An equivalent result was derived in (Elliott 2003) without directly invoking device reciprocity.

Analogous argument applies to the power delivered at the feeds of the antennas, which is proportional to the square of the voltage: lobe pattern of the tested antenna is the same for both reception and transmission.

Finally, let us restate the obvious: the reciprocity argument is valid only if the spatial configuration of the two antennas is exactly identical in both modes. In practice, test antenna should approach an ideal isotropic receiver, or else precautions must be taken that its polarization does not distort the measurement results.

10.3 Fundamentals of Radio Communication

In this section, we review the fundamentals of transmitting a message by radio waves. It is not necessary to cover here the numerous and sophisticated elaborations on the basic principles, nor to go into the engineering details of implementation. We wish to explain the fundamental signal processing that comprises the radio technology, since the understanding of these principles is essential for understanding radiolocation. Naturally, our discussion will emphasize the radio reception over transmission somewhat.

The transport medium of a radio message is the monochromatic electromagnetic wave; it is aptly named the "carrier." Obviously, an endless uniform oscillation carries no message, so the message must be imprinted as some modification of the carrier waveform. Apart from the direction of propagation, a wave has three parameters: amplitude, frequency and phase. Encoding schemes can be devised that translate a message (analog or digital) into variations in these parameters; this process of encoding is referred to as modulation. Variation in amplitude or frequency is traditionally used for (analog) encoding of sound, while combined variations in amplitude and phase are used to encode digital data.

The Fourier transform (or the spectrum) of a monochromatic wave is of course a delta function at the frequency of the wave. For a modulated wave, the transform broadens into a shape of finite height and width, centered around the carrier frequency; its frequency range is often referred to as a "radio band" or the passband. This broadening is understandable, since a modulated wave can be seen as the carrier with waves of neighboring frequencies added to it (remember that the wave equation is linear!); these added waves comprise the modification of the carrier, which is the modulation. The broader the frequency range, the greater the amount of information that is impressed upon the carrier, and this is why the rate of data transfer is commonly called the "bandwidth."

The simplest (and oldest) encoding is the encoding of sound by amplitude modulation, in which the envelope (amplitude) of the wave follows the shape of the transmitted sound wave. In a very crude approximation, an AM radio station is an oscillator that generates the carrier wave, followed by an amplifier whose gain control is connected to a microphone, and which changes the carrier's amplitude in the rhythm of the sound.

All radio receivers are conceptually also very simple, as shown in Fig. 10.3: past the antenna, we first need a band-pass filter that limits the received signal to the

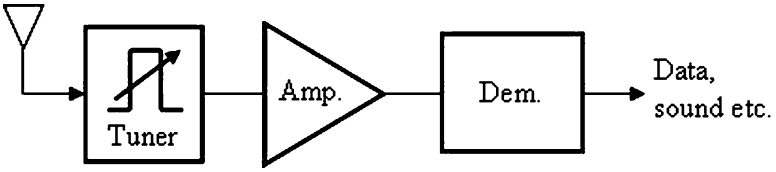

Fig. 10.3 Conceptual architecture of a receiver

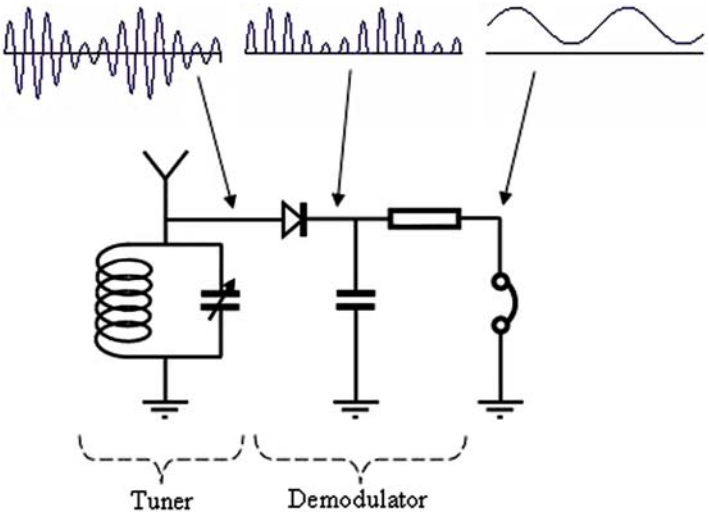

Fig. 10.4 The simplest AM receiver: a diode "detector"

band of interest. This filter is of course essential, because it selects our channel of communication out of the endless clutter of ambient radio waves; it is typically somewhat variable, and it is what is commonly referred to as the "tuner." Next we need a device that retrieves the encoded information; this is the demodulator, and its complexity will reflect the intricacy of the encoding scheme. Fortunately for us, mechanics of demodulation remains irrelevant in our discussion of radiolocation. Finally, we also need an amplifier somewhere in the receiving path, to boost the weak signal received from the antenna.

As an example, we show the schematic of the simplest AM receiver in Fig. 10.4. The resonant LC circuit serves as a filter that passes a narrow band of frequencies, and the central frequency of that band can be selected by using a variable capacitor. A rectifier diode and an RC low-pass filter serve to extract the audio-frequency envelope of the carrier. This audio signal is passed through the earphones, and with luck you may actually hear the local AM station!

Actual receivers are vastly more complex than this, for engineering rather that conceptual reasons. Primarily, at high carrier frequency it becomes difficult to build a tuning filter that will crisply select a narrow radio band. Coils and capacitors decrease in size and precision as the frequency increases, and the parasitic effects become significant. Making a high-quality filter tunable is an even greater challenge. The standard answer to this problem is heterodyning, or frequency downconversion, which we now describe.

We can view the modulated carrier received at the antenna terminals as a fairly general time function $f(t)$, where f is the electric current. The frequency spectrum (the radio band) of that carrier is described by its Fourier transform, which is the expansion into sine and cosine waves of all frequencies. In order to simplify the

formulas, we will use the complex Fourier transform, although the argument could be carried out in terms of real sine and cosine transforms just as well:

$$f(t) = \int_{-\infty}^{\infty} \mathrm{d}\omega\, F(\omega)\, \exp(i\omega t) \tag{10.26}$$

Here $F(\omega)$ is the coefficient of expansion in the ω-th wave, the complex Fourier transform of $f(t)$, a function of frequency that we refer to as the spectrum.

We now multiply (mix) the modulated carrier $f(t)$ with a cosine harmonic oscillation of frequency ω_0:

$$\begin{aligned} f(t)\, \cos\omega_0 t &= \int_{-\infty}^{\infty} \mathrm{d}\omega\, F(\omega)\, \exp(i\omega t)\, \cos\omega_0 t \\ &= \frac{1}{2} \int_{-\infty}^{\infty} \mathrm{d}\omega\, F(\omega)\, \exp[i(\omega - \omega_0)t] \\ &\quad + \frac{1}{2} \int_{-\infty}^{\infty} \mathrm{d}\omega\, F(\omega)\, \exp[i(\omega + \omega_0)t] \\ &= \frac{1}{2} \int_{-\infty}^{\infty} \mathrm{d}\omega\, [F(\omega + \omega_0) + F(\omega - \omega_0)]\, \exp(i\omega t) \end{aligned} \tag{10.27}$$

We used the Euler formula for the cosine, and changed the variables of integration in the last line. We see that mixing doubles the spectrum, forming two bands that are shifted down and up by ω_0, and have the same shape as the original band (see Fig. 10.5).

In other words, if the radio signal is multiplied by a harmonic oscillation of frequency ω_0, its spectrum is shifted by that same frequency, without change in shape – modulation and all. This is what the heterodyning technique consists of: the raw radio signal from the antenna is multiplied (mixed) with the tightly controlled sig-

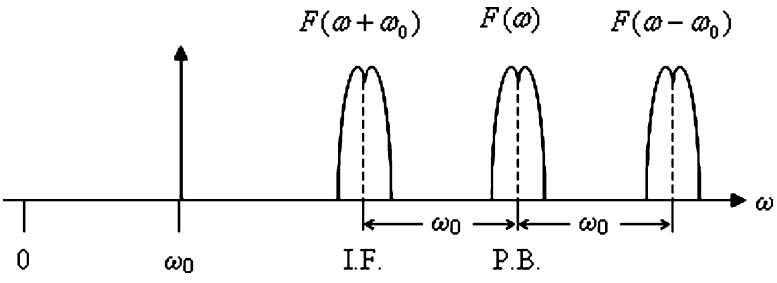

Fig. 10.5 Spectrum of the mixed (heterodyned) signal

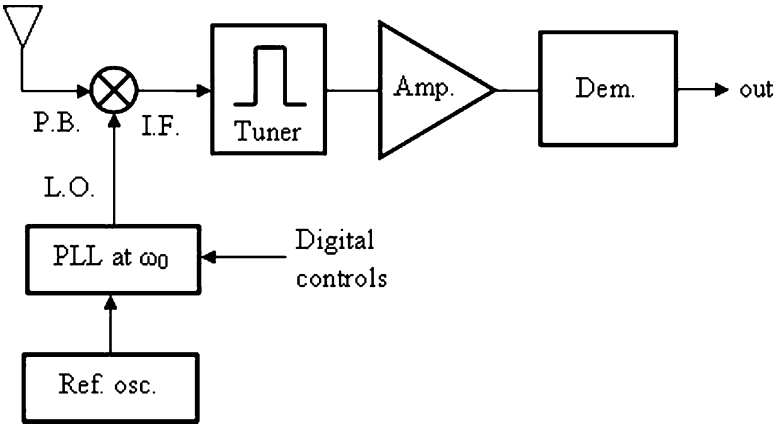

Fig. 10.6 Heterodyne receiver

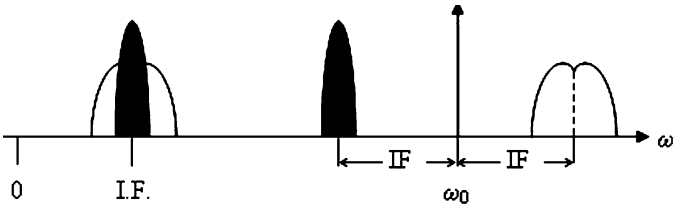

Fig. 10.7 The image problem

nal from a local oscillator, and is thereby brought down to a frequency at which a good tuning filter can be designed (see Fig. 10.6). The upper band is typically high enough to be reliably eliminated by the same filter.

Furthermore, the tuning filter need no longer be variable: it must have the bandwidth of the radio band, centered at the downconverted (intermediate) frequency, but the tuning itself is accomplished by varying the frequency of the local oscillator. The latter can be done very precisely, for example, by using a digitally controlled phase-locked loop (Razavi 1998; Horowitz and Hill 1989).

The downconversion does have a fundamental problem, though, which requires some remedial design features. The term $\cos(\omega - \omega_0)t$ brings the passband frequency ω down to intermediate frequency $\text{IF} = \omega - \omega_0$, but since the cosine is even, it does so on both sides of ω_0: passbands $\omega = \omega_0 + \text{IF}$ and $\omega = \omega_0 - \text{IF}$ are downconverted to the same intermediate frequency, leading to possible selection of two transmitters at once (Fig. 10.7). This is usually referred to as the "image problem."

A very popular receiver architecture in digital radio communication is the homodyne, also known as the ZIF (zero intermediate frequency) receiver. This is a variant of the heterodyne in which the local frequency is approximately equal to the passband, and the baseband resides around the zero frequency. Main advantage of

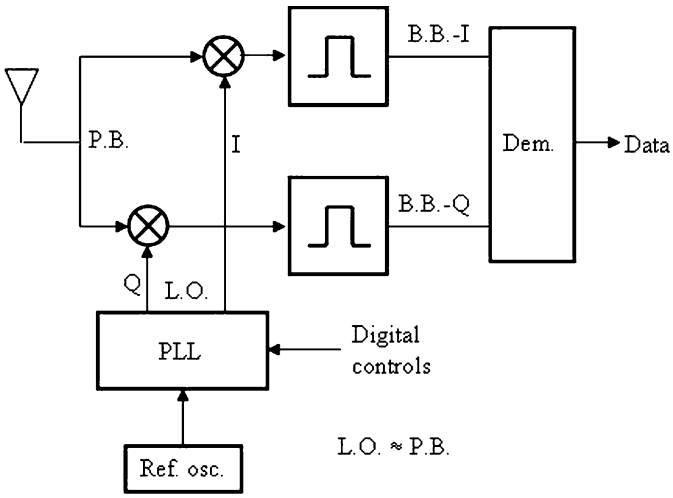

Fig. 10.8 Homodyne (ZIF) receiver

the homodyne architecture is that it does not give rise to the image problem. Simplified ZIF architecture is shown in Fig. 10.8: the passband is downconverted to zero frequency, by mixing with two oscillators, which are 90° out of phase. This allows the reconstruction of phase shifts in the phase-amplitude modulations used in data transmissions; the same technique is used in data heterodyne receivers, as a second downconversion stage.

10.4 Transmission Lines

Transmission lines guide the electromagnetic power over distance, in the form of oscillating fields. Unlike the lumped components, transmission lines are not small compared to the wavelength, and we cannot assume that the fields and currents present in them are uniform in space. Transmission lines are simpler structures than antennas, though: their fields can be described either in one dimension, as is the case with cables, or as products of simpler one- and two-dimensional fields, as is the case with waveguides and resonant cavities. In this section, we limit our discussion to the simplest one-dimensional description of a transmission line: we derive the wave equation for voltages and currents, describe the behavior at discontinuities, and discuss some experimental techniques.

10.4.1 Free Space in One Dimension

Intuitively, we think of "cables" as thin two-conductor structures, in which power travels along the cable's length, with negligible bouncing across the cross section;

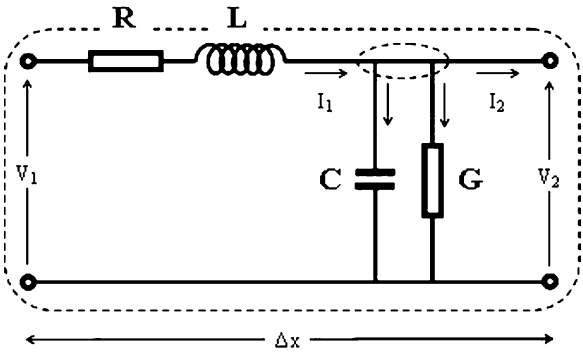

Fig. 10.9 Equivalent circuit of an infinitesimal element of a TEM transmission line

this is the transverse electromagnetic mode (TEM). The "cable" can be a pair of wires, a coaxial cable, a pair of adjacent traces on a board, or a trace paired with a reference plane.

Every cable offers certain dissipative resistance, R, in the wires, some leakage between the conductors, usually expressed as a conductance, G, inductance, L, due to the build-up of the magnetic field around and along the wires, and a capacitance C due to the build-up of the electric field between them. For uniform stretches of cable, these quantities are expressed per unit length, and we anticipate that, for a good cable, dissipative quantities R and G should be small: resistance along the cable is low, and the insulation between the wires is high.

We can represent a uniform, infinitesimally short line segment as the equivalent circuit shown in Fig. 10.9; we wish to determine how the voltage and current change from one end of the segment to the other, because of the presence of R, L, G and C. Voltages around the outer loop add to zero, according to Kirchhoff's voltage law, and the currents entering and leaving the marked junction add to zero according to the current law. Applying Ohm's law to resistances, and using the definitions of capacitance and inductance, we obtain the following equations from the Kirchhoff's laws:

$$V_1 - R\,\Delta x\, I_1 - L\,\Delta x\,\frac{\partial I_1}{\partial t} - V_2 = 0 \tag{10.28}$$

$$I_1 - G\,\Delta x\, V_2 - C\,\Delta x\,\frac{\partial V_2}{\partial t} - I_2 = 0 \tag{10.29}$$

We now divide each equation by Δx, and let $\Delta x \to 0$ to obtain the spatial derivatives of V and I:

$$\frac{\partial V}{\partial x} = -RI - L\,\frac{\partial I}{\partial t} \tag{10.30}$$

$$\frac{\partial I}{\partial x} = -GV - C\,\frac{\partial V}{\partial t} \tag{10.31}$$

where we have replaced the endpoint voltages and currents with interval averages V and I; we are justified in doing so in the limit $\Delta x \to 0$. Equations (10.30) and (10.31) are also known as the telegrapher's equations.

To understand the physics behind (10.30) and (10.31), let us assume first that R and G are negligible, i.e., the line is lossless. By differentiating one equation in time and the other in space, and by eliminating mixed terms, we obtain equations for V and I, in the following form:

$$\frac{\partial^2 U}{\partial x^2} = L\,C\,\frac{\partial^2 U}{\partial t^2} \tag{10.32}$$

This is of course the one-dimensional wave equation, analogous to (1.13), where the phase velocity of the propagating disturbance (the equivalent of c) equals:

$$u = \frac{1}{\sqrt{LC}} \tag{10.33}$$

We see that our lossless transmission line behaves like a one- dimensional "free space" in which wave propagation is determined by the reactances of the line. Inductance plays the role of μ_0, and the capacitance the role of ε_0, with corresponding dimensions as well. Equation (10.33) is the equivalent of (1.7) for the free space. One-dimensional waves for which $k = \omega\sqrt{LC}$ are solutions of (10.32), and the characteristic impedance of the line (ratio of voltage and current amplitudes) is easily derived, as the equivalent of (1.27):

$$Z_0 = \sqrt{\frac{L}{C}} \tag{10.34}$$

For a lossy line, we introduce harmonic time dependence $\exp(i\omega t)$ into (10.30) and (10.31), differentiate the equations in x and eliminate terms, to obtain

$$\frac{\partial^2 U}{\partial x^2} = \gamma^2 U \tag{10.35}$$

where

$$\gamma^2 = (R + i\omega L)(G + i\omega\, C) \tag{10.36}$$

is a complex propagation constant, and the waves are dampened due to the dissipation. The characteristic impedance of a lossy line is

$$Z_0 = \sqrt{\frac{R + i\omega L}{G + i\omega\, C}} \tag{10.37}$$

10.4.2 Impedance Discontinuities

The discussion in Sect. 10.4.1 applies to an infinite and uniform transmission line, but real transmission lines bend, go through connectors, and terminate in sources and loads. These discontinuities are accompanied by changes in the characteristic impedance Z_0, so let us discuss the passage of a wave across an impedance discontinuity.

Let us imagine that a transmission line has a discrete break, a boundary at which the impedance changes from Z_0 to Z_1, as shown in Fig. 10.10. An incident wave crosses the boundary and becomes the transmitted wave; whatever else may happen to its waveform, voltages and currents must remain continuous at the boundary, or we would see infinite electromagnetic fields there. But the boundary conditions and the impedance equations

$$V_i = V_t; \quad I_i = I_t \tag{10.38}$$

$$V_i = Z_0 I_i; \quad V_t = Z_1 I_t \tag{10.39}$$

can only be satisfied simultaneously if $Z_0 = Z_1$, which is a contradiction, i.e., the conditions are impossible. The situation is made physically possible by the presence of a reflected wave, propagating back towards the source of the incident wave. Now the boundary conditions become:

$$\begin{aligned} V_i + V_r &= V_t \\ I_i - I_r &= I_t \end{aligned} \tag{10.40}$$

The reflected current has a negative sign because of the wave's backward propagation. These conditions are compatible with $Z_0 \neq Z_1$, and we obtain the ratio of incident and reflected voltage amplitudes, which is called the reflection coefficient:

$$\Gamma = \frac{V_r}{V_i} = \frac{Z_1 - Z_0}{Z_1 + Z_0} \tag{10.41}$$

This picture is physically equivalent to the reflection of a wave in space from a boundary between two different media. The formula (10.41) is equivalent to (1.55) for the fraction of reflected power, and we see that the ratio of impedances Z_0/Z_1 corresponds to the index of refraction.

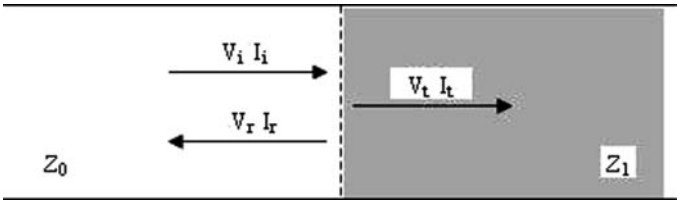

Fig. 10.10 Impedance discontinuity in a transmission line

A line with impedance Z_1, stretching from the boundary into infinity, is equivalent to a terminating load at that boundary, with the same impedance Z_1. The load and the unbounded line are both infinite signal sinks; therefore, the above discussion applies to terminating loads as well. If $Z_1 \to \infty$ (open end of line), $\Gamma = 1$; if $Z_1 = 0$ (shorted end of line), $\Gamma = -1$, that is, the wave reflects fully from both open and shorted line ends. The reflection is absent only if the line and the load are matched, $Z_0 = Z_1$.

The incident and reflected waves form a standing wave on the line between the source and the load. A traditional measure of the degree of reflection is the "standing wave ratio" or SWR (also voltage standing wave ratio, VSWR): this is the ratio of largest and smallest amplitude of the standing wave, and, for submicrowave frequencies, it used to be measured directly on long stretches of transmission lines. It can be shown that

$$\mathrm{SWR} = \frac{1 + |\Gamma|}{1 - |\Gamma|} \tag{10.42}$$

SWR equals one for matched impedances, and grows to infinity with the mismatch.

At microwave frequencies, the degree of reflection is measured much more conveniently by using a directional coupler, or return- loss bridge. We will not discuss couplers here, except to say that they effectively separate the reflected from the incoming wave, so that the power carried by the reflection can be measured (see Fig. 4.8 for a typical measurement setup). For a very detailed explanation of directional couplers, we direct the reader to (Pozar 1998).

10.5 Power Flux in the Modulated Signal

We discuss here the power flux in a modulated signal, a topic that is relevant for the design of the power-meter circuit; this section supplements Sect. 5.2.

Power flux in a signal is proportional to the square of the signal's voltage (or to the square of the magnitude of the electric field), although we are rarely interested in the instantaneous value of the power flux. More commonly, we wish to know the average flux, with the average taken over some physically meaningful time period.

It can be easily shown that, in a pure sine/cosine wave, the average power flux equals $(1/2) \cdot (V_{\max}^2/Z)$, where the flux is averaged over any time interval whose length is equal to a multiple of the half-period. In Figs. 10.11 to 10.14, average power flux is represented by a dash-dot line, and we assume for simplicity that the impedance of the medium $Z = 1$.

Modulation schemes that are relevant for our present discussion are BPSK and QPSK. In the binary phase-shift keying (BPSK), the carrier is modulated by introducing 180° phase shifts; regardless of the exact location of the phase shift in the period of the waveform, this shift always amounts to simply changing the sign of the wave from that point on. As we can see from Fig. 10.11, the instantaneous power flux remains the same as for the unmodulated wave, and so does the average.

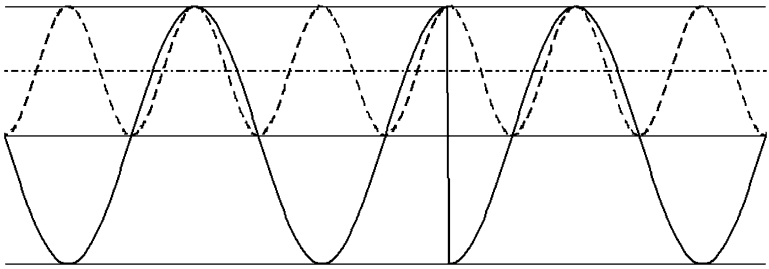

Fig. 10.11 BPSK phase change

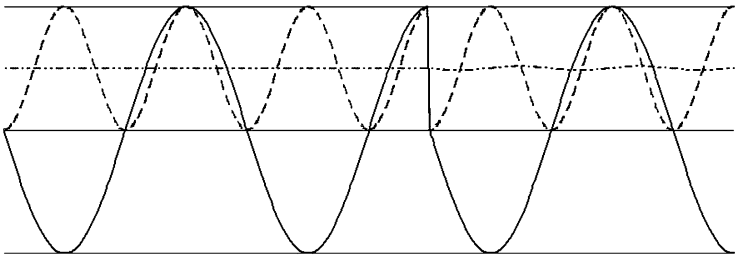

Fig. 10.12 QPSK phase change

Magnitude of the power flux in a perfectly modulated BPSK signal is not affected by the modulation at all, although the direction of the flux changes.

Such is not the case for QPSK and higher-order phase modulations. As shown in Fig. 10.12, the 90° phase shift introduces an anomaly into the instantaneous flux, which in turn causes ripples and level shifts in the power averaged over multiples of the half-period. This ripple, of course, becomes less prominent for longer averaging intervals.

Ideally, crisp phase modulation, such as shown in Figs. 10.11 and 10.12, introduces discontinuities into the monochromatic waveform, and discontinuities, in principle, carry infinitely high frequencies within them. Naturally, spectra of all realistic signals have an upper cut-off limit, and the waveforms of bandwidth-limited BPSK and QPSK signals are illustrated in Figs. 10.13 and 10.14: loss of high-frequency components smoothens out the phase discontinuities and turns them into amplitude dips; it also introduces slight amplitude changes along the rest of the wave. The resulting average power flux exhibits modulation ripples for both modulation schemes, even though ideal BPSK does not influence the magnitude of the power flux.

Exact shape of these modulation ripples depends on the location of the phase shift in the waveform, the length of the averaging interval, and on the frequency cut-off in the low-pass filtering. We are merely interested in explaining why they exist – a quantitative analysis would not be particularly useful. In the limit of instantaneous measurement, the power meter becomes unstable, and for very long averaging times, the ripples disappear. The practical engineering compromise lies somewhere in between, and is best determined empirically.

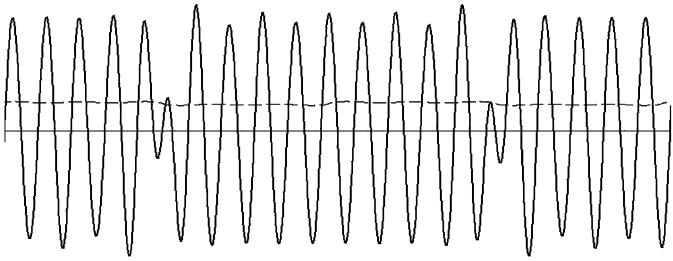

Fig. 10.13 Bandwidth-limited BPSK signal, showing baud boundaries

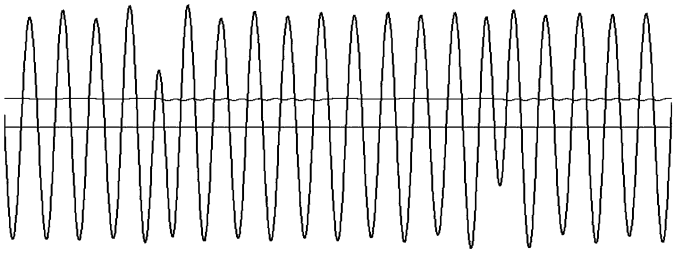

Fig. 10.14 Bandwidth-limited QPSK signal

10.6 Overview of the 802.11b Standard

This section provides a very brief overview of the conventions and protocols of the Wi-Fi communication. An exhaustive account of the 802.11 communication standards is given in the excellent book by M. Gast (2005), and for a broader discussion of topics relevant to digital communication, we refer the reader to the standard text by B. Sklar (2001).

10.6.1 Types of Networks

The 802.11 standards make provisions for three types of wireless networks:

Infrastructure networks, also called Infrastructure Basic Service Sets. These networks consist of an immobile access point, which is associated with a number of mobile stations; the access point is usually also connected to a conventional wired network. A wireless network of this type is identified by an address, usually the address of its access point, called the BSSID. Mobile stations join and leave the network dynamically, and remain aware of the BSSID for the duration of the association.

Ad hoc networks, or Independent Basic Service Sets, consist of mobile stations only, and are typically of limited duration. Their BSSID is a randomly chosen, temporary address.

Wireless bridges, or Wireless Distribution Systems, consist of wired networks linked through a pair of communicating access points. Since the association is static, no BSSID is needed.

10.6.2 Physical Layer

This part of the standard deals with the specifics of data communication via radio. As with all radio transmission, the fundamental means of transmitting information is by impressing changes upon the carrier wave. In the 802.11b standard, these changes are simple phase shifts. Prior to modulation, however, data bits are subjected to an additional transformation known as chipping: every data bit is bitwise added, modulo 2 (that is, XOR-ed), with the same sequence of chipping bits; there is always an integer number of chips per data bit, and the chip boundaries fall on the data boundaries. There are several reasons for introducing the chipping.

The chipping sequence has a higher baud rate than data (the 802.11 b standard specifies the basic data rate of 1 Mbit/second, and 11 chips per data bit), and therefore the modulated carrier's spectrum is widened according to that higher baud rate. The chipping sequence is redundant, in the sense that it adds no data content to the modulation stream, and it is a bit sequence from a collection of sequences that have certain desirable statistical properties. The chipping sequences are deterministic, but they have the statistical character of random noise, and they are mutually orthogonal in the sense that the scalar product of a sequence with itself yields the sequence length (a sizable number), while the cross-multiplication of sequences yields small, random-looking numbers, preferably between -1 and $+1$.

As the result, the signal spectrum is broad, and the signal looks like noise to a narrow-band receiver. The original application of this technique, called direct-sequence spread spectrum (DSSS), was in military communications, for the purpose of hiding the radio communication altogether, but the larger benefit lies in the fact that the spread spectrum overlaps weakly with narrow-band signals, and is therefore relatively immune to interferences from narrow-band transmitters.

The unchipped data stream is recovered by adding bitwise the chipped stream with the chipping sequence once more. We can easily see why this works: adding the same chipping bit to a data bit twice, modulo 2, either goes fully around the base-2 number circle or not at all, and leaves the original data bit unchanged. The sender's chipping sequence must be known to the receiver, and the receiver's application of the chipping must be synchronized with the chipping already on the data stream: multiplication with another (orthogonal) sequence, or with the correct sequence that is out of synchronization, yields random-looking chaff.

All of this opens the possibility of multiple communication channels within the same radio band, one for each orthogonal sequence. While this may seem paradoxical at first, it is perfectly legitimate, and is the basis of the Code Division Multiple Access (CDMA) technique, used in cellular phone systems (Sklar 2001). The chipping is redundant and merely widens the band, without adding new

information to it. Theoretical data capacity of this wider band is larger, but it is not fully used by any single chipping sequence. We reclaim that bandwidth by using the orthogonality of the pseudo-random chipping; we could have just as well subdivided the band into multiple narrow-frequency channels.

Another way to reclaim the bandwidth is to dispense with the full redundancy of the chipping, and use the whole set of orthogonal chipping sequences as a means to encode information. Since chipping has higher baud rate, this trick achieves higher data throughput in a single channel, but the signal retains its spread-spectrum character. This encoding scheme, called complementary code keying (CCK), is part of the 802.11 b standard, and is used for data transfer at rates of 5.5 and 11 Mbits/second (Pearson 2000).

We should add here that chipping does not amount to meaningful cryptographic encoding. Spread spectrum hides the transmission from narrow-band receivers by making it look like noise, but if such spread spectrum transmission is expected, it can be detected and deciphered.

As we already said, modulation is in the form of differential phase shifts (see Figs. 10.11 and 10.12 as examples). Keeping track of the absolute phase of the modulated signal would require continuous comparisons with an unmodulated reference wave, and it is much more convenient to assign data content to phase *changes* instead, for example, as shown in Table 10.1. We see that an absence of phase change at the expected time of change also has meaning; therefore, on the demodulating side, an initial synchronization interval is needed, to establish the baud boundaries. In addition, modulation data stream is subjected to a pseudo-random scrambling, which reduces the chances of long zero sequences. These zeros would translate into stretches of uniform carrier wave, leading to timing drift and consequently to erroneous demodulation. The scrambling is deterministic, and is readily reversed at the receiver.

Every 802.11 b packet starts with 16 bytes of DBPSK-modulated all 1's, which serve as the synchronization sequence (see Fig. 6.4). Following the sync sequence are four fields of header data, which guide the demodulation process:

Start frame delimiter (*SFD*) – this is a set sequence of bits, which unambiguously indicates the beginning of a frame. This field is necessary, since some of the early bits in the synchronization sequence may be missed, erroneously offsetting the interpretation of everything that follows.

Signal – this field indicates the transmission rate of the payload that follows the PHY header. In 802.11 b, the transmitter is allowed to choose from one of four rates: 1, 2, 5.5 or 11 Mbits/second, by using DBPSK or DQPSK, or the latter in

Table 10.1 Encoding data as phase shifts

DBPSK		DQPSK	
1-bit symbol	Phase shift	2-bit symbol	Phase shift
0	0	00	0
		01	90°
1	180°	11	180°
		10	270°

conjunction with CCK. We should add here that the PHY-layer fields must always be transmitted at the fixed rate of 1 Mbit/second, using DBPSK modulation.

Service – all zero, not currently used.

Length – number of microseconds required to transmit the frame, needed because the standard does not prescribe an end-of-frame field. This field is calculated for the transmitted packet, from the number of bits and the transmission rate, and the receiver assumes that the packet is completed after this time period.

The header is completed with a two-byte CRC of the signal, service and length fields. The demodulation circuit cannot assume that it is receiving a valid packet until at least the SFD field has been received; a safer course of action is to wait for a correct CRC of the header before announcing the reception of a valid 802.11 b packet.

10.6.3 Medium Access Control Layer

Medium access control (MAC) is a specialized sublayer of the Data Link layer, present in the 802 family of communication standards; we are, of course, interested in its 802.11 version. By the time a received wireless packet reaches the MAC layer, all the issues of tuning, downconversion and demodulation have been resolved, and the physics of radio is no longer visible to the network. However, the communication protocol itself is strongly shaped by the radio medium, and that is what the 802.11 MAC layer is set up to handle.

The layer distinguishes three types of communication packets (frames) and numerous frame subtypes within the main types: *Data frames* are of course the useful payload, and they contain encapsulated frames of the Logical Link Control (LLC) and higher network layers. *Control frames* are concerned with the coordination of data exchanges in the relatively lossy and unpredictable radio channel. *Management frames* coordinate the logical relationships among the network nodes or interlocutors. The control and management frames are specific to the MAC layer, and are not passed to the higher layers in the course of normal data communication.

It is a basic requirement of the 802.11 standard that every transmitter must wait for the clear medium, that is, for an absence of detectable transmissions, before it starts transmitting its own message; the standard provides a back-off protocol designed to avoid transmission collisions and consequent loss of data. Transmitters can reserve the medium for a period of time for uninterrupted transmission, using specialized control frames. Control frames are also used to coordinate fragmented transmission of large data packets, and for the acknowledgment of received data.

Radio communication has no routing, and no physical boundaries other than the detectable signal strength: the "connection" between two network interlocutors is a matter of mutual agreement, and of keeping track of each other's status. The 802.11 standard specifies three degrees of relationship between a wireless node and an infrastructure network:

Not authenticated – the node and the network have no logical relationship. In order to help set up relationships, existing networks periodically broadcast

beacon frames, announcing their presence. Individual nodes can also broadcast probe frames, to which compatible networks are obliged to respond with probe response frames.

Authenticated – the node sends an authentication request frame to the network, and receives authentication approval. Very little actual authentication is required by the standard: the node and the network merely establish a record of each other's presence and basic characteristics. Most implementations insert somewhat more meaningful, cryptographic authentication steps here.

Associated – similarly, association request and approval frames are exchanged. This step assigns (associates) the node to an access point, and the packets destined for that node are routed to, and transmitted by, that access point. As it moves around, the mobile node has the option of transferring its association from one AP to another. Once a node is associated with the network, it is allowed to exchange data frames with it. Either side can terminate the association and authentication.

Special subtypes of management frames handle the described protocol. We see from the above discussion that this multitude of frame types and subtypes is a consequence of the nature of radio, and of the mobility that the radio allows. We will leave the full description of frame types and formats to specialized books (see Gast 2005), but we wish to summarize here the addressing scheme, which is of significance for the discussion in Chap. 6.

Every MAC-layer frame starts with a two-byte Frame Control field, which directs the interpretation of the remainder of the frame. The Frame Control contains a two-bit code of the frame type, and a 4-bit code of the subtype. It contains specifications for interpreting the subsequent address fields; it also contains information pertaining to packet fragmentation, station power management, etc.

MAC-layer frames contain up to four 6-byte address fields; see the expanded data payload in Fig. 6.4 for an example. Generally, first field is the address of the receiving radio unit, and the second field is the address of the transmitting radio unit. In data frames, the third and fourth fields are used for the addresses of the data source and destination. In ad hoc networks, these are the same as the radio units, but in infrastructure networks and wireless bridges, they can be nodes on the wired network to and from which the radio units route the packets.

Management and control frames do not recognize sources and destinations different from the radio units, since they only coordinate the operation of communicating radio units. Management frames have three address fields, and they use the third field for the BSSID, the identifier of the infrastructure network. Control frames typically have two address fields: receiver and sender. However, since clear-to-send is always sent in response to a request-to-send, CTS contains only the receiver address: its sender address is the receiver address of the previous RTS frame. The same applies to the acknowledgment (ACK) frame, which is used to confirm the reception of a data packet.

10.7 Wilkinson Divider

The Wilkinson divider is shown in Fig. 10.15: it consists of two quarter-wave transmission lines and a dissipative resistor. The divider has three ports, all of which are matched to the feeding lines with impedance Z_0. The reference ground is not shown in the schematic for simplicity, and, for convenience of analysis, the schematic is drawn symmetrically with respect to the dividing dotted line: the feeding line into Port 1 is shown as two parallel lines with twice the impedance Z_0, and the resistive element is divided into two identical resistors in series, Z_0 each.

Looking into Port 1, we see that the feed impedance is matched by the two quarter-wave lines of impedance $\sqrt{2}Z_0$, each terminated in Z_0:

$$(Z_{\text{in}})_{\text{Port1}} = \frac{1}{2}\frac{\left(\sqrt{2}Z_0\right)^2}{Z_0} = Z_0 \tag{10.43}$$

Let us send a signal into Port 1: because of the symmetry of the circuit, it divides into two identical signals. Voltages on Ports 2 and 3 are identical, no current flows through the resistor, and the circuit's operation is nondissipative (lossless). The power of the output signals must add up to the power of the input signal, due to the conservation of energy, and it follows that

$$P_2^{\text{out}} = P_3^{\text{out}} = P_1^{\text{in}}/2 \tag{10.44}$$

$$V_2^{\text{out}} = V_3^{\text{out}} = V_1^{\text{in}}/\sqrt{2} \tag{10.45}$$

Conversely, if we send two identical signals into Ports 2 and 3, there is again no current through the resistor, the operation is lossless,

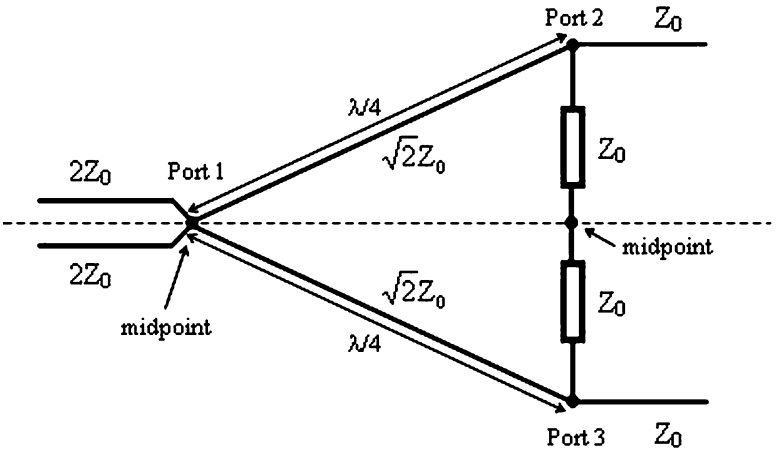

Fig. 10.15 Schematic of the Wilkinson divider

$$P_1^{\text{out}} = 2P_2^{\text{in}} = 2P_3^{\text{in}} \tag{10.46}$$

and the circuit is reciprocal. However, if the input signals differ in any way, the joining operation will incur losses in the resistor.

To see that the circuit isolates Ports 2 and 3 (as long as all ports are matched), let us send a signal into Port 2, and match/terminate the other two ports. To analyze this case, we will depict it as a sum of an even mode $\left(V_2^{\text{in}} = V_3^{\text{in}}\right)$ and an odd mode $\left(V_2^{\text{in}} = -V_3^{\text{in}}\right)$, and analyze the modes separately.

In the even mode, no current ever flows through the midpoints, and we can break the circuit in half, by disconnecting it at midpoints, and looking at the top half only – the two halves behave identically. The resistor now leads nowhere and is superfluous. Port 2 is therefore matched by the quarter-wave line terminated in $2Z_0$, as it should be, and a bit of analysis of the transmission line shows that

$$V_1^{\text{out,even}} = (-i\sqrt{2})V_2^{\text{in,even}} \tag{10.47}$$

In the odd mode, midpoints are always at zero voltage, and we can break the circuit by grounding them. The now shorted quarter-wave line has infinite impedance, and Port 2 is matched by the resistor of value Z_0. Taken together, the upper and lower half of the circuit set up a standing half-wave on the transmission lines, with a node at Port 1, and opposite-phase maxima at Ports 2 and 3, because $V_2^{\text{in,odd}} = -V_3^{\text{in,odd}}$.

Adding the two modes together, we obtain $V_2^{\text{in}} = 2V_2^{\text{in,even}}$ and $V_3 = 0$; Ports 2 and 3 are mutually isolated. We can visualize this as the signal traveling from Port 2 through the standing wave, with inversion of phase, and through the resistors without inversion, and canceling out at Port 3. In addition, $V_1^{\text{out}} = (-i/\sqrt{2})V_2^{\text{in}}$ and

$$P_1^{\text{out}} = P_2^{\text{in}}/2 \tag{10.48}$$

The signal into Port 2 does not see Port 3, and it comes out of Port 1 with half the power and a phase shift of $\pi/2$. The other half of its power is dissipated in the resistors.

The answer to the question whether a three-port reciprocal, matched divider could be made lossless, is no. A three-port device can be described by a 3×3 scattering matrix, whose elements are ratios of input and output voltage amplitudes, on all ports. If the ports are matched, there are no reflections from them, and the diagonal elements are zero. Reciprocity requires the matrix to be symmetric, and the losslessness (energy conservation) requires it to be unitary. A bit of algebra shows that it is impossible for a matrix of dimension three to satisfy all of these three requirements (although it can satisfy any two of them). For this and other topics on dividers, we refer the reader to the uncommonly detailed and rigorous exposition in (Pozar 1998).

10.8 Spherical Harmonics

Spherical harmonics are the solutions of Laplace's equation on the surface of the unit sphere. Like the sine and cosine on a circle, they describe the amplitude of standing waves on a spherical surface: ocean waves on a completely flooded planet, or the surface vibrations of a freely floating soap bubble, can be described by using these functions. They are closely related to the symmetry group of the sphere, and they turn up in quantum-mechanical description of the angular momentum. These functions are amply described in mathematical physics literature, and we will merely list their relevant properties here.

Spherical harmonics are usually defined as functions of the azimuth angle φ and the declination from the polar axis, θ. They exhibit circular harmonic behavior in φ, and in the θ-angle they behave as associated Legendre functions. The latter functions can be defined by means of a compact, albeit not always the most useful formula:

$$P_l^m(x) = \frac{(-1)^m}{2^l l!}(1-x^2)^{m/2}\frac{d^{l+m}}{dx^{l+m}}(x^2-1)^l \tag{10.49}$$

Spherical harmonics themselves are given by

$$Y_{lm}(\theta,\varphi) = \sqrt{\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}}P_l^m(\cos\theta)e^{im\varphi} \tag{10.50}$$

Indices l and m are integers, such that $l \geq 0$ and $m = -l, \ldots, 0, \ldots, +l$; we should not be surprised by the presence of integer indices, since these functions satisfy the wave equation and describe normal modes of oscillation on the sphere. There are m equally spaced null planes through the polar axis, and $l - |m|$ null cones in the range $0 \leq \theta \leq \pi$; for large values of l, and m close to $l/2$, the pattern of nulls resembles the system of meridians and parallels on the globe.

Functions Y_{lm} satisfy a complex-conjugation symmetry:

$$Y_{l-m}(\theta,\varphi) = (-1)^m Y_{lm}^*(\theta,\varphi) \tag{10.51}$$

They are unitary under integration over the unit sphere:

$$\int_0^{2\pi} d\varphi \int_0^{\pi} d\theta \sin\theta Y_{l'm'}^*(\theta,\varphi)Y_{lm}(\theta,\varphi) = \delta_{ll'}\delta_{mm'} \tag{10.52}$$

and they satisfy the completeness condition, which can also be viewed as a unitary property, this time under the summation over all indices:

$$\sum_{l=0}^{\infty}\sum_{m=-l}^{l} Y_{lm}^*(\theta',\varphi')Y_{lm}(\theta,\varphi) = \delta(\varphi-\varphi')\delta(\cos\theta-\cos\theta') \tag{10.53}$$

If the coordinate system is subjected to an arbitrary rotation, harmonics in the old (unprimed) coordinates can be expressed as linear combinations of harmonics defined with respect to the new (primed) coordinates:

$$Y_{lm}(\theta,\varphi) = \sum_{\sigma=-l}^{\sigma=+l} D_l^{m\sigma}(\alpha,\beta,\gamma) Y_{l\sigma}(\theta',\varphi') \tag{10.54}$$

We see that these linear combinations mix only the harmonics, Y_{lm}, within the same value of the index, l; for each l, the $2l + 1$ spherical harmonics form the basis of a (irreducible) representation of the rotation group. The rotation matrices $D_l^{m\sigma}$ (not to be confused with the rotation matrices for coordinate values) are functions of the Euler angles describing the coordinate rotation. For the $z - y - z$ definition of Euler angles ("rotate by α around the z-axis; by β around the new y-axis; by γ around the new z- axis"), the elements of rotation matrices are given by:

$$D_l^{m\sigma}(\alpha,\beta,\gamma) = \exp(im\alpha + i\sigma\gamma) d_l^{m\,\sigma}(\cos\beta) \tag{10.55}$$

The functions of the angle of tilt of the z-axis, $d_l^{m\sigma}(\cos\beta)$, are real-valued and have an explicit expression, which is a messy trigonometric formula in β; they also obey certain recurrence relations that are more convenient for actual calculations. We will not go any deeper into the topic of rotational matrices, except to state a useful special case:

$$d_l^{m0}(\cos\beta) = \sqrt{\frac{4\pi}{2l+1}} Y_{l\,m}(\beta, 0) \tag{10.56}$$

This element of the rotational matrix is essentially the spherical harmonic of the tilt angle, with zero azimuth.

Properties of Legendre functions are exhaustively covered in most reference volumes of mathematical formulas (see e.g., Abramowitz and Stegun 1972); see also the classic treatise by Hobson (1965). Good introductions to spherical harmonics, in the context of physical applications, can be found in (Jackson 1998) and (Messiah 1999), and their group-theoretical significance is discussed in (Talman 1968) and (Altmann 1986).

References

Abramowitz, M., Stegun, E.A. Handbook of Mathematical Functions, Dover, New York (1972)
Altmann, S.L. Rotations, Quaternions and Double Groups, Oxford University Press, Oxford (1986)
Elliott, R.S. Antenna Theory and Design, revised edition, Wiley, New York (2003)
Gast, M.S. 802.11 Wireless Networks, 2nd edition, O'Reilly, CA (2005)
Hobson, E.W. Spherical and Ellipsoidal Harmonics, 2nd reprint, Chelsea, New York (1965)
Horowitz, P., Hill, W. The Art of Electronics, 2nd edition, Cambridge University Press, Cambridge (1989)
Jackson, J.D. Classical Electrodynamics, 3rd edition, Wiley, New York (1998)

Messiah, A. Quantum Mechanics, Dover, New York (1999)
Pearson, B. Complementary Code Keying Made Simple, application note AN9850.1, Intersil Corporation, FL (2000)
Pozar, D.M. Microwave Engineering, Wiley, New York (1998)
Razavi, B. RF Microelectronics, Prentice Hall, NJ (1998)
Sklar, B. Digital Communications, 2nd edition, Prentice Hall, NJ (2001)
Talman, J.D. Special Functions: A Group Theoretic Approach, Benjamin, New York (1968)

Index