

Radio Wave Propagation

John A. Richards

Radio Wave Propagation

An Introduction for the Non-Specialist

 Springer

John A. Richards
The Australian National University
Res. School of Information Sciences &
Engineering (RSISE)
Dept. Information Engineering
Canberra ACT 0200, Australia
John.Richards@anu.edu.au

ISBN: 978-3-540-77124-1

e-ISBN: 978-3-540-77125-8

Library of Congress Control Number: 2007940900

© 2008 Springer-Verlag Berlin Heidelberg

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: WMXDesign GmbH, Heidelberg

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

Understanding the propagation of radio waves in the vicinity of the earth's surface can be quite complex, especially if detailed theoretical knowledge is required. The transmission path is complicated by atmospheric, tropospheric and ionospheric effects, and the earth's surface itself, and other obstacles, can interact with the passage of radiation between a transmitter and receiver. Time of day and season of the year can also be important.

A full treatment of these aspects usually requires a detailed understanding of electromagnetic theory and Maxwell's celebrated equations and yet many practitioners, even electrical engineers, may not have that background in sufficient depth. Nevertheless, with the proliferation of wireless applications, particularly in the VHF and UHF ranges, there is often the need for the non-specialist to gain a working knowledge of the properties of radio waves and how they are affected by factors such as those outlined above. That is the purpose of this book. It treats the essential elements of radio wave propagation without requiring recourse to advanced electromagnetic concepts and equations; however it provides sufficient detail to allow those concerned with wireless systems to acquire quickly a practical working knowledge of the important concepts.

The treatment commences with an analysis of how *energy* (and power) is conveyed in free space, taking essentially a radiative transfer approach and thus avoiding the need to understand electric and magnetic field propagation at the outset. It then examines in some detail how the proximity of the earth and the atmosphere cause the radiation travelling from a transmitter to a receiver to follow one or more of three mechanisms – the surface, sky and space waves. Most attention is given to the space wave since it is the mechanism most commonly encountered in contemporary applications.

Radio wave propagation is placed in a practical context by considering the design aspects of communications systems at microwave frequencies. That requires an understanding of noise and its importance in systems design.

We take the unusual step of including a fuller consideration of the electromagnetic properties of materials late in the book rather than as an introductory chapter as

found in more theoretical treatments. It is placed here so that the contexts in which the knowledge of material properties is important have already been established.

The material is based on a single semester overview course suitable for later year undergraduate students in engineering or science.

Canberra, Australia 2007

John A. Richards

Contents

1	Fundamental Concepts: Propagation in Free Space	1
1.1	Free Space Versus Guided Propagation of Radio Waves	1
1.2	The Concept of Power Density	1
1.3	Electric and Magnetic Field Components	3
1.4	Velocity of Propagation and Frequency-Wavelength Relationship ..	5
1.5	Friis' Radiation Formula	6
1.6	Band Designation for Radio Waves	7
1.7	Radio Wave Propagation Near the Earth	8
1.8	Allocating the Radio Spectrum	10
	Problems	11
2	The Surface Wave	13
2.1	Ideal Surface Wave Field Strength	13
2.2	The Case of a Real Earth	14
2.3	Incorporating Earth Curvature and Atmospheric Refraction	17
2.4	Propagation at Very Low Frequencies	18
	Problems	22
3	The Sky Wave	23
3.1	The Ionosphere	23
3.2	The Refractive Index of an Ionospheric Layer	26
3.3	Refraction of a Radio Wave in the Ionosphere	27
3.4	The Set of Critical Frequencies	30
3.5	The Virtual Height of an Ionospheric Layer	31
3.6	Maximum Usable Frequency and Skip Distance	35
3.7	Range of the Sky Wave	36
	Problems	37
4	The Space Wave	39
4.1	The Received Field Strength	39
4.2	Effect of Earth Curvature on Space Wave Propagation	43

4.3	Diffraction	45
4.4	Refraction of the Space Wave	48
4.5	Effect of Rainfall on Space Wave Propagation	53
4.6	Atmospheric Attenuation	54
	Problems	55
5	Noise	57
5.1	What is Noise?	57
5.2	Sources of Noise	58
5.3	The Concept of Noise Temperature	59
5.4	The Noise Temperature of a Two Port	59
5.5	Noise Figure	61
5.6	Relationship Between Noise Figure and Output Signal to Noise Ratio	61
5.7	The Noise Properties of a Passive Two Port	62
5.8	Cascaded Two Ports: Friis' Noise Formula	63
	Problems	66
6	Examples of Microwave Systems	69
6.1	The Design of Open Microwave Repeater Systems	69
6.1.1	The Need for Transmission Diversity	72
6.1.2	The Use of Passive Reflectors	73
6.2	Propagation Aspects of Satellite Communication Systems	75
6.2.1	The Geostationary Orbit	75
6.2.2	Link Power Calculations	77
6.3	The Propagation Aspects of Cellular Radio	79
6.4	Mobile Wireless Systems	81
	Problems	84
7	The Effect of Materials on Propagation	87
7.1	Background	87
7.2	Propagation in Homogeneous Media	88
7.3	Frequency Dependence of Material Properties	92
7.4	Interactions with Ideal Interfaces	96
7.5	Reflection from Rough Surfaces	99
7.6	Transmission Through Media	101
7.7	Propagation in Tunnels	102
	Problems	104
A	A Simple Introduction to Antennas	107
A.1	Introduction: Radiation Resistance and Radiation Patterns	107
A.2	The Directivity and Gain of an Antenna	110
A.3	The Aperture of an Antenna	112

A.4 Radiated Fields	112
A.5 Some Typical Antennas	115
A.6 Baluns	117
B The Use of Decibels in Communications Engineering	119
C The Dielectric Constant of an Ionospheric Layer	121
Index	125

Chapter 1

Fundamental Concepts: Propagation in Free Space

1.1 Free Space Versus Guided Propagation of Radio Waves

Radio waves can travel between two points either by propagating in free space or by being guided in a medium such as a coaxial cable, waveguide or optical fibre. In the former, the spectrum available must be shared with all users. To ensure compatible operation, allocation of the free space radio spectrum is subject to regulation by internationally agreed charters, and directional antennas are often employed to minimise interference of services operating on similar frequencies in close geographical proximity.

In principle, if radiation is carried inside a guiding medium there is not likely to be any interference with other users. Instead, the full range of frequencies able to be supported by the medium is available to the user. Often the spectrum again is regulated, but now according to the specifications of a service provider rather than through treaty.

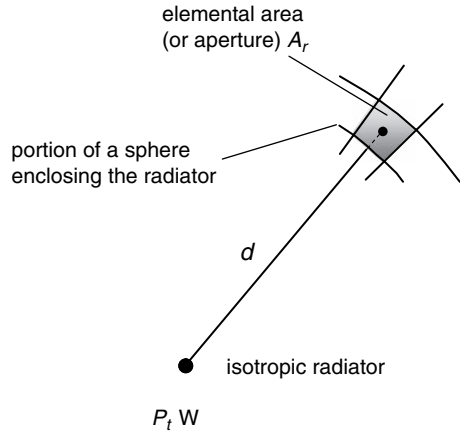
This book is concerned with radiation in free space. It examines, with a minimum of complex mathematical theory, the mechanisms by which propagation can take place between points in space, and places the material in the context of telecommunications systems. We commence with an understanding of how energy radiates.

1.2 The Concept of Power Density

A detailed understanding of methods of radio wave propagation would ordinarily commence with a treatment of Maxwell's equations and their combination into the wave equation. However, a good practical appreciation can be obtained by starting with a simple understanding of how power propagates outwards from a source of energy, such as a radio transmitter.

Consider a point source of energy, such as that depicted in Fig. 1.1. It could be a source of light, heat, sound or electrical energy. Often we characterise it by the

Fig. 1.1 The concept of an isotropic radiator and spherical propagation



power it radiates – i.e. the energy per second emanating from it – which is measured in watts (W). If the source radiates uniformly in all directions it is called *isotropic*, or an *isotropic radiator*. A point source must be isotropic since there is nothing to bias the flow of energy in any particular direction.

The energy from an isotropic source propagates outward in a spherical fashion. If we placed ourselves at a given distance d from the radiator and enclosed it by a sphere of that radius then we could intercept all of the power emanating from the source. While it originated from a point source isotropic radiator, the power is now smeared or distributed over the whole surface area of the sphere. It is convenient now to define the *power density* over the surface of the sphere as the power transmitted divided by the surface area of the sphere. It has units of watts per square metre, and is given by

$$p = \frac{P_t}{4\pi d^2} \quad \text{Wm}^{-2} \quad (1.1)$$

Note that this is the classical *inverse square law* found in many other fields of physics.

Rather than collect the outgoing power density over the full surface area of the sphere, we could instead intercept only that portion over the small cross-sectional area shown in Fig. 1.1. Then we will be able to extract

$$P_r = pA_r \quad \text{W} \quad (1.2)$$

watts of power from the outgoing wavefront. We have used subscripts r in this last equation to imply “received”. The intersecting cross-section A_r is generally referred to as an *aperture*, as though it were a hole through which power is received.

The point source isotropic radiator of Fig. 1.1 is fictional. Real radiators are designed to focus their transmitted power in preferred directions as depicted in Fig. 1.2. We now introduce a definition to continue with the discussion of power density: the *gain* of the radiator G_t is a measure of how much more power density

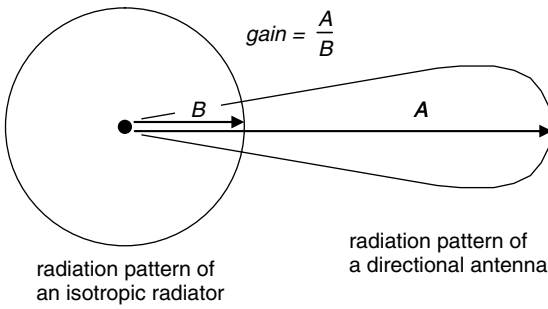


Fig. 1.2 The concept of antenna gain, based on how much more power density can be radiated in a preferred direction compared with an isotropic source

the real radiator is able to transmit in the preferred direction (usually by employing an antenna) than can the equivalent isotropic source. Thus the power density at distance d from the transmitter, and the power received, are now given respectively by

$$p_r = \frac{P_t G_t}{4\pi d^2} \quad \text{Wm}^{-2} \quad (1.3)$$

$$P_r = \frac{P_t G_t A_r}{4\pi d^2} \quad \text{W} \quad (1.4)$$

1.3 Electric and Magnetic Field Components

We have described the propagation of the radio wave so far in terms of the power density conveyed. In reality however it travels as the combination of electric and magnetic field vectors as illustrated in Fig. 1.3. Both fields are at right angles to the direction of propagation and at right angles also to each other.¹ The wave is therefore referred to as transverse electromagnetic (TEM). When referred to the earth's surface two orientations are defined – by reference to the orientation of the electric field vector – as noted in Fig. 1.3. They allow us to describe the radiation as *horizontally* or *vertically* polarised.²

It is important to recognise that the vectors shown illustrate the plane in which the respective fields oscillate (at the frequency of the transmitted radiation). Strictly

¹ This is only true in the case of free space propagation and well away from the transmitting antenna. Inside guiding media such as a waveguide there can be components of electric and magnetic field in the direction of propagation as there will also be in the very near vicinity of an antenna in free space.

² The radiation can also be elliptically or, as a special case, circularly polarised. In these cases there are both vertical and horizontal components with different magnitudes and with a relative phase difference.

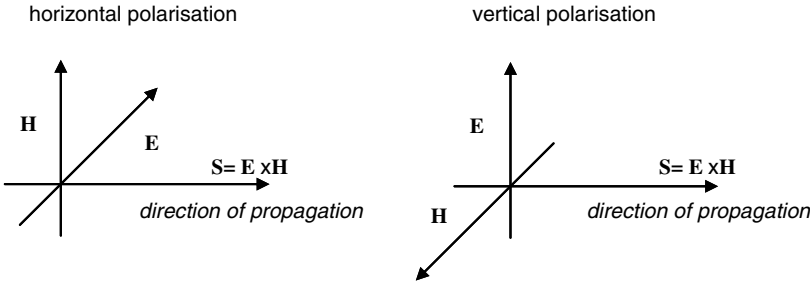


Fig. 1.3 Transverse electric and magnetic field components, and the Poynting vector, for a travelling wave

they are vector-phasors in that they contain information on the geometric orientation of the field, its magnitude and its relative phase angle.

The vector cross product of the electric and magnetic field vectors defines a new vector that points in the direction of propagation. Called the Poynting Vector, \mathbf{S} , it has units of watts per square metre, since electric and magnetic fields have units respectively of volts per metre (Vm^{-1}) and amps per metre (Am^{-1}). Thus the Poynting Vector has units of power density; even though a vector quantity, its magnitude is precisely power density. From the definition of vector cross product,³ the magnitude of the Poynting vector is the product of the magnitudes of the electric and magnetic fields. Thus we have another expression for power density, in addition to that used in (1.1) and (1.3), viz:

$$p = |\mathbf{S}| = |\mathbf{E}| |\mathbf{H}| \quad \text{Wm}^{-2} \quad (1.5)$$

In free space it can be shown⁴ that the electric and magnetic field intensities are related by

$$|\mathbf{E}| = \eta |\mathbf{H}| = 120\pi |\mathbf{H}| \approx 377 |\mathbf{H}| \quad (1.6)$$

in which $\eta = 120\pi = 377\Omega$ is the impedance of free space.

From (1.3), (1.5) and (1.6) we can see that the electric field strength created at the distance d from the transmitter is

$$|\mathbf{E}| = \frac{\sqrt{30P_t G_r}}{d} \quad \text{Vm}^{-1} \quad (1.7)$$

which shows that the field strength follows an *inverse distance law*, whereas we saw that power density follows an inverse square law.

³ For a very good vector treatment of electromagnetic propagation, including the fundamental concepts from vector algebra see J.D. Kraus, *Electromagnetism*, 5th ed., N.Y., McGraw-Hill, 1995.

⁴ Ibid.

1.4 Velocity of Propagation and Frequency-Wavelength Relationship

From Maxwell's equations we can show that the velocity of radio waves in a medium is given by

$$v = \frac{1}{\sqrt{\mu\epsilon}} \quad (1.8)$$

where μ and ϵ are the absolute permeability and permittivity of the medium respectively; they are two of the medium's electromagnetic properties. In free space

$$\begin{aligned} \mu &= \mu_o = 400\pi \text{ nHm}^{-1} \\ \epsilon &= \epsilon_o = 8.85 \text{ pFm}^{-1} \end{aligned}$$

which, when substituted into (1.8), give the velocity of radio waves (and light) in free space as

$$v = c = 299.8 \approx 300 \text{ Mms}^{-1}$$

In a medium with *relative permeability* μ_r and *relative permittivity* ϵ_r (also called *dielectric constant*) the absolute permeability and permittivity of the medium are

$$\begin{aligned} \mu &= \mu_r \mu_o \text{ Hm}^{-1} \\ \epsilon &= \epsilon_r \epsilon_o \text{ Fm}^{-1} \end{aligned}$$

Most media in which we are interested are non-magnetic, so that $\mu_r = 1$.

In a medium with dielectric constant ϵ_r the velocity of the radio waves is

$$v = \frac{1}{\sqrt{\mu_o \epsilon_r \epsilon_o}} = \frac{c}{\sqrt{\epsilon_r}} = \frac{c}{n} \quad (1.9)$$

where $n = \sqrt{\epsilon_r}$ is, by definition, the *refractive index* of the medium.

Since, for any wave motion, the wavelength and frequency are related by velocity according to

$$v = f\lambda$$

a very useful relationship can be derived for radio waves in free space. Based on the value for the velocity of light, we can see that

$$f(\text{MHz}) = \frac{300}{\lambda(\text{m})} \quad (1.10)$$

This is one of the most useful and important expressions in telecommunications and in the study of propagation of radio waves.

1.5 Friis' Radiation Formula

Imagine we want to transmit a signal between two points spaced d apart, well away from any effect of the earth's surface. The transmitting antenna is characterised by a gain G_t ; the receiving antenna can be described by an aperture A_r . If the transmitter delivers a power of P_t watts to the transmitting antenna then the power density at the receiving antenna is

$$p = \frac{G_t P_t}{4\pi d^2} \quad \text{Wm}^{-2}$$

from which the receiving antenna extracts a power (delivered at its terminals) of

$$P_r = pA_r = \frac{A_r G_t P_t}{4\pi d^2} \quad \text{W} \quad (1.11)$$

There is a relationship between the gain of an antenna when used for transmission and the aperture of the same antenna when used for reception. As seen in Appendix A, the aperture of the receiving antenna can be written

$$A_r = \frac{\lambda^2 G_r}{4\pi} \quad \text{m}^2$$

which, when substituted into (1.11), gives

$$P_r = \frac{G_r G_t P_t \lambda^2}{(4\pi d)^2} = G_r G_t P_t \left(\frac{\lambda}{4\pi d} \right)^2 \quad \text{W} \quad (1.12)$$

This last expression is known as *Friis' Radiation Formula*.

It is convenient now to take $10\log_{10}$ of (1.12) to give

$$10\log_{10} P_r = 10\log_{10} G_r + 10\log_{10} G_t + 10\log_{10} P_t + 20\log_{10} \frac{\lambda}{4\pi d}$$

or

$$10\log_{10} P_r = 10\log_{10} G_r + 10\log_{10} G_t + 10\log_{10} P_t - 20\log_{10} \frac{4\pi d}{\lambda} \quad (1.13)$$

Now what do expressions like $10\log_{10} G_t$ mean? We can write this expression as

$$10\log_{10} \frac{G_t}{1}$$

The "1" in the denominator can be regarded as the gain of an isotropic radiator (since the isotropic source radiates uniformly in all directions, and gain is defined in relation to isotropic behaviour). Thus, the last expression can be viewed as the gain of the transmitting antenna, expressed in decibels with respect to an isotropic radiator. This is written

$$10\log_{10}\frac{G_t}{1} = G_t \text{ dBi} .$$

Similarly $10\log_{10}P_r$ is the received power expressed in decibels with respect to 1 watt. This is written P_r dBW (see Appendix B). With these decibel substitutions (1.13) can be re-expressed

$$P_r \text{ dBW} = P_t \text{ dBW} + G_r \text{ dBi} + G_t \text{ dBi} - 20\log_{10}\frac{4\pi d}{\lambda}$$

In many communications systems power levels are more likely to be in the range of milliwatts or smaller. It is convenient therefore to re-write the last expression in relation to a reference of 1 milliwatt rather than 1 watt. As noted in Appendix B that only requires a re-expression of the power units based on a 1 milliwatt reference:

$$P_r \text{ dBm} = P_t \text{ dBm} + G_r \text{ dBi} + G_t \text{ dBi} - 20\log_{10}\frac{4\pi d}{\lambda} \quad (1.14)$$

The last term in (1.14) is called the *free space path loss* (FSPL), expressed as decibels. It represents the loss of signal reaching the receiver as a result of the spreading of the transmitted beam in an inverse square law fashion.

It is convenient now to express (1.14) in words, since it gives us a general means for treating communications systems when the quantities are expressed in decibel form:

$$\textit{received power} = \textit{transmitted power} + \textit{antenna gains} - \textit{losses} \quad (1.15)$$

While there is only one loss term in (1.14) – the free space path loss – the word equation of (1.15) allows all relevant losses to be accounted for. We will see in later chapters that the set of losses include

free space path loss
diffraction loss
refraction loss (fading)
rainfall loss
atmospheric absorption loss

Therefore (1.15) can be applied quite generally and logically when used in decibel form by noting that all losses present can be collected (and added) in the last term.

1.6 Band Designation for Radio Waves

Table 1.1 shows the nomenclature used when referring to radio waves in different bands of frequency. As seen, the names are referenced to the so-called “medium frequency” band, which was the frequency and wavelength range most used in the early days of radio. Note how (1.10) is used to go between frequency and wavelength in the table.

Table 1.1 Frequency and wavelength band designators

Band	Frequency Range	Wavelength Range
ELF	30–300 Hz	10–1 Mm
ULF	300 Hz–3 kHz	1 Mm–100 km
VLF	3–30 kHz	100–10 km
LF	30–300 kHz	10–1 km
MF	300 kHz–3 MHz	1 km–100 m
HF	3–30 MHz	100–10 m
VHF	30–300 MHz	10–1 m
UHF	300 MHz–3 GHz	1 m–100 mm
SHF	3–30 GHz	100–10 mm
EHF	30–300 GHz	10–1 mm

Table 1.2 Microwave band designators

Band	Frequency Range	Wavelength Range
P	300 MHz–3 GHz	1 m–300 mm
L	1–2 GHz	300–150 mm
S	2–4 GHz	150–75 mm
C	4–8 GHz	75–37.5 mm
X	8–12.5 GHz	37.5–24 mm
Ku	12.5–18 GHz	24–16.7 mm
K	18–26.5 GHz	16.7–11 mm
Ka	26.5–40 GHz	11–7.5 mm

In the microwave range an alternative letter designation is often used, especially when dealing with radar systems and with propagation in waveguides. Table 1.2 shows one of several definitions in common use.

1.7 Radio Wave Propagation Near the Earth

Consider an isotropic radiator on top of a mast as shown in Fig. 1.4. Being isotropic, the radiation will propagate outwards in all directions; the question is how does the signal reach a receiving antenna mounted on another mast situated some distance away? Figure 1.5a suggests that the transmission path can be the ray drawn as a straight line between the transmitter and receiver. There can be a second path as a result of reflection from the surface of the earth (or from other objects such as buildings and vehicles). Clearly the receiver and transmitter need to be in sight of each other for those rays to carry the signal. The two rays are referred to respectively as the *direct* and (ground) *reflected* components of what is collectively called the *space wave*.

Above the earth's surface is a series of charged layers called the ionosphere. If conditions are right one or some of the rays in Fig. 1.4 that travel upward can be

Fig. 1.4 How does the radiation from an isotropic source reach a receiver?

spherical propagation from an isotropic radiator

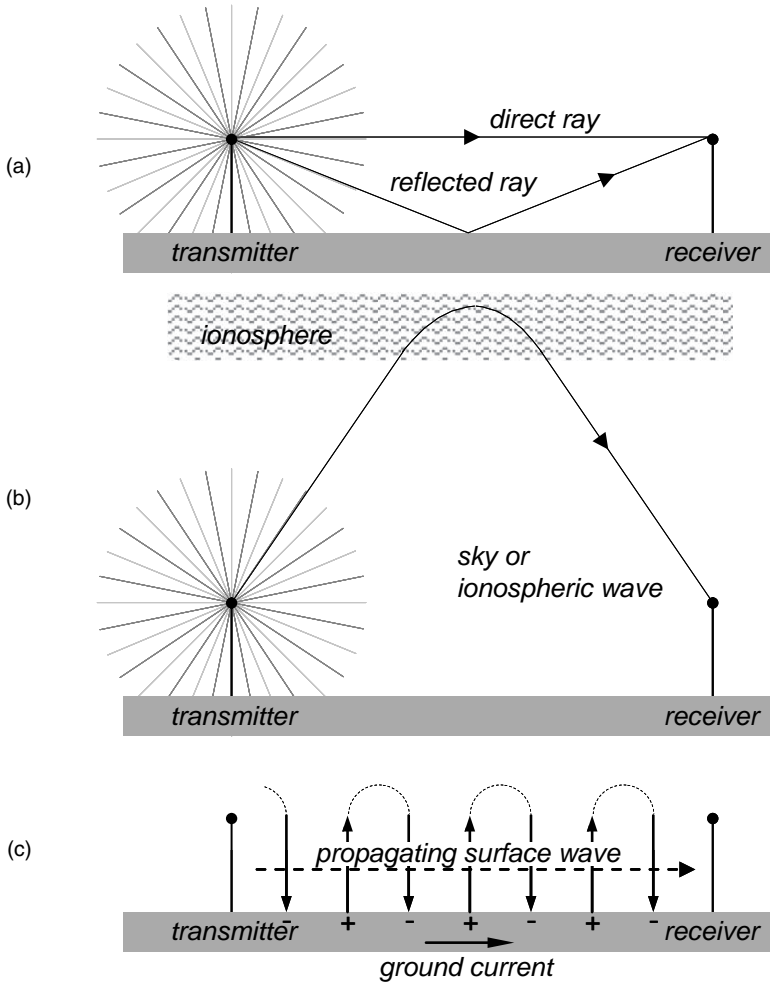
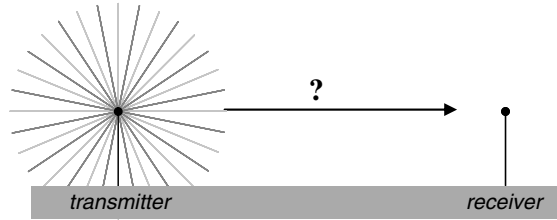


Fig. 1.5 Three modes of propagation of radiation between a transmitter and receiver (a) the space wave (b) the sky wave and (c) the surface wave

Table 1.3 Summary of the dominant free space radio wave propagation mechanisms (following E.V.D. Glazier and H.R.L. Lamont, *Transmission and Propagation*, HMSO, London, 1958)

Frequency Range	Dominant Mechanism	Some Typical Services
< 500kHz	Surface wave	Submarine communication, global navigational aids
500 kHz–1.5 MHz	Surface wave for short distances Sky wave for long distances	AM broadcast radio, air navigational aids
1.5 MHz–30 MHz	Sky wave	Short wave (including amateur) communication over long distances
> 30MHz	Space wave	Television, FM radio, air navigational aids, GPS, wireless LAN and internet, WiFi, mobile (cell) phones

refracted back to the earth's surface and thus the receiver. Shown in Fig. 1.5b, that ray is called the *sky* or *ionospheric* wave.

There is a third propagation mechanism that is less obvious when considering a ray view of the situation, as we have for the space and sky waves. Known as the *surface wave*, this mechanism transports the signal as a field that terminates on charges (and thus current) which travel in the surface layer of the earth, as illustrated in Fig. 1.5c.

In principle, all three propagation mechanisms – the space, sky and surface waves – can exist simultaneously. We will see, however, that each mechanism has its own preferred ranges of frequency as summarised in Table 1.3. In the following chapters we will examine each mechanism in some detail and thereby confirm the summary of Table 1.3. The table also shows some typical services that use each frequency range and are thus affected by those propagation mechanisms. Not surprisingly, the space wave carries many more services than the others because of the vast portion of the spectrum served at VHF and beyond and because of the very large bandwidths available. Line of sight propagation however is a significant limitation on space wave services; that can be overcome by the use of terrestrial repeaters or through the use of communications satellites, both of which are examined in Chap. 6.

While the illustrations in Figs. 1.4 and 1.5 have been based on isotropic transmitters, the same mechanisms will occur with directional antennas. In practice the antennas chosen would be designed to optimise the propagation mechanism most appropriate to the frequency range being used.

1.8 Allocating the Radio Spectrum

Because every user who radiates in a given geographical region is essentially using the same transmission medium – i.e. free space – it is necessary to have regulations in place to avoid services interfering with each other. Most often those regulations

relate to frequency usage, but they can also cover direction of radiation and transmitter power levels.

Each country has in place a regulatory regime including a policy for allocation of the electromagnetic spectrum over the various radiated services. That policy is derived from international agreements over use of the spectrum, determined from time to time at meetings of the International Telecommunication Union (ITU).⁵ The ITU and counterpart agencies in various countries publish charts of agreed usage of the spectrum, from the lowest usable frequencies up to the highest frequencies technically feasible at a given time. Those charts are quite complex given the range of competing services for different parts of the spectrum, particularly at VHF and above where there is sufficient spectrum available to permit a very large number of uses with good bandwidths. It is important that those services are line-of-sight limited thus permitting substantial duplication of use in different geographical regions – that is often referred to as frequency re-use.

Problems

1.1. A particular portable telephone system operates with a carrier frequency of 2.4 GHz. Plot a graph of free space path loss at that frequency in dB for distances between 100 m and 10 km.

1.2. A plane wave with peak electric field strength of 10 Vm^{-1} is travelling in polystyrene which has $\mu_r = 1$ and $\epsilon_r = 2.7$. Assuming the polystyrene is lossless determine the velocity of the wave and the average Poynting vector.

1.3. A satellite carries a transmitter that radiates 100 W isotropically. What is the electric field strength at a receiver sitting on the earth's surface if the satellite is at an altitude of (i) 1000 km, (ii) 12,000 km, (iii) 36,000 km?

1.4. A particular satellite is in an orbit 12 Mm above the earth's surface. It carries a transmitter that radiates 100 W at 20 MHz on an antenna with gain 20 dBi (i.e. with respect to an isotropic antenna). What is the field strength at the earth's surface?

1.5. A geostationary communications satellite at an altitude of 35,870 km above the earth's surface receives transmissions from a satellite that orbits the earth above the equator at an altitude of 1000 km. Assume the period of the orbiting satellite is 100 min. If the orbiting satellite uplinks to the geostationary satellite using an isotropic antenna and a transmitter power of 10 W, plot a graph of power density at the geostationary satellite as a function of the orbital position of the lower satellite

⁵ The specific roles of the ITU in relation to the radio spectrum are to set the allocation of bands in the spectrum, to allocate radio frequencies and register radio frequency assignments and associated orbital positions of geostationary satellites, to coordinate efforts to eliminate harmful radio interference and to improve the use of radio frequencies and of the geostationary-satellite orbit for radio communication services. See www.itu.int/ITU-R.

while it is in view. Using the graph what is the maximum variation in received power density and field strength seen at the geostationary satellite?

1.6. A radar operates with a peak transmitter power of 1 kW. Pulses are radiated outwards using an antenna with gain 10 dBi. Calculate the power received back at the radar antenna terminals from a target located 15 km away; the target can be viewed as appearing like an aperture of 7 m^2 which then re-radiates (scatters) the incident power density isotropically. Assume the same antenna at the radar set is used for transmitting and receiving. (Characterising a radar target by an aperture which then re-radiates isotropically is widely used in radar engineering and is referred to as the *radar cross-section* of the target.)

1.7. Suppose the target in question 1.6 is now replaced by a simple communications receiver and transmitter arrangement (i.e. a transponder) that, upon receipt of a pulse from the radar set, automatically transmits a reply back to the radar using a transmitter power of $P_s \text{ W}$ on an isotropic antenna. What should be the value of P_s to deliver the same power density at the radar set as the passive target arrangement of question 1.6? If the receiver on the target (which may be an aircraft) requires a power of 10 pW at its antenna terminals, what power needs to be transmitted from the radar set on the ground? Such an arrangement, in which the target carries a transponder as part of the radar detection process, is referred to as *secondary radar*, in contrast to *primary radar* that is used to detect passive targets. While requiring less transmitter power, secondary radars also have the advantage of allowing the target, in its response to being interrogated, to transmit back to the radar set other information such as altitude, bearing, identification codes and even distress or emergency information.

1.8. In the expression for received power in a communications link shown in (1.11) the product $G_t P_t$ is often referred to as the *Effective Isotropically Radiated Power* (EIRP). Why? Why it is a useful quantity?

1.9. You have been asked to do a preliminary system level design for a communications satellite receiver. The satellite is in geostationary orbit above the equator and the receiver is to be located on the same longitude as the satellite but at a latitude of 40 degrees south. The system is to operate at 12 GHz. The receiving antenna is a parabolic dish of 3.5 m diameter. It has an efficiency of 0.7. The receiver requires a power from the receiving antenna terminals of -105 dBm .

What EIRP is required at the satellite?

Chapter 2

The Surface Wave

2.1 Ideal Surface Wave Field Strength

Table 1.3 notes that the surface wave (sometimes also called the ground wave) is dominant below about 500 kHz. We will see why later. For the moment note that at those frequencies we are talking about wavelengths that are longer than 600 m! What does that suggest for the sorts of antenna that might be used for radiating and receiving the surface wave? A half wave dipole (Appendix A) would be 300 m long. If it were to be used as a horizontally polarised radiator then it would need to be placed several hundred metres above the earth in order to avoid the earth acting as a short circuit. Clearly that is impracticable; so vertical polarisation is used and a quarter wave, or even short, monopole is generally employed as the radiator, as depicted in Fig. 2.1.

Recall from (1.7) that the electric field strength at a distance d from a source with antenna gain G_t is given by

$$|\mathbf{E}| = \frac{\sqrt{30P_t G_t}}{d} .$$

This expression applies also to the propagation of a surface wave over a perfectly conducting earth, since the field above the earth is mirrored in the earth itself, giving the appearance of free space.

For a short vertical monopole $G_t = 3$ with respect to isotropic (Appendix A) so that for the surface wave propagating over an ideally conducting earth the field strength is

$$|\mathbf{E}| = \frac{\sqrt{90P_t}}{d} .$$

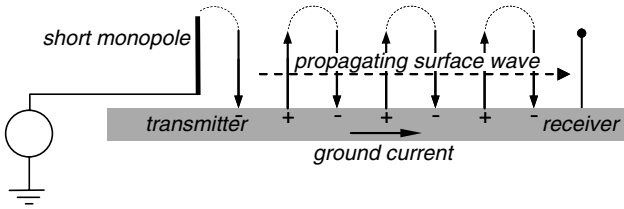


Fig. 2.1 Launch of the vertically polarised surface wave using a short vertical monopole

2.2 The Case of a Real Earth

To account for the fact that the earth is not an ideal conductor this expression is modified by the inclusion of an *attenuation factor* A to give

$$|\mathbf{E}| = \sqrt{90P_t} \frac{A}{d} \quad (2.1)$$

The factor A is quite complicated to determine. It's derivation is discussed in Rohan,¹ who outlines several approaches, one of which expresses A as

$$A = A_0 - (A_0 - A_{90}) \sin b \quad (2.2)$$

in which

$$A_0 = \frac{2 + 0.33p}{2 + p + 0.6p^2} \quad (2.3a)$$

and

$$A_{90} = \frac{2 + 170p}{2 + 210p + 310p^2} \quad (2.3b)$$

p is a normalised distance from the transmitter and b is a phase angle, defined, for vertical polarisation, as

$$p = \frac{\pi d \cos b}{\lambda x} \quad (2.3c)$$

with

$$b = \tan^{-1} \frac{\epsilon_r + 1}{x} \quad (2.3d)$$

and

$$x = 60\sigma\lambda \quad (2.3e)$$

¹ P. Rohan, *Introduction to Electromagnetic Wave Propagation*, Artech, Boston, 1991.

where σ is earth conductivity (in Sm^{-1}), ϵ_r is earth dielectric constant and λ is the operating wavelength. By employing the definitions of \tan and \cos , (2.3c) can be written, using (2.3d), as

$$p = \frac{\pi d \cos b}{\lambda x} = \frac{\pi d}{\lambda} \frac{1}{\sqrt{(\epsilon_r + 1)^2 + x^2}} \quad (2.3f)$$

Rohan gives alternatives to (2.3c–e) for the (seldom used) case of horizontal polarisation.

The attenuation factor of (2.2) is plotted as a function of p and b in Fig. 2.2. It is helpful in interpreting this figure if we simplify (2.3c) or (2.3f). For a highly conducting earth x in (2.3d) will be large compared with $\epsilon_r + 1$ so that b will approach zero. As seen from (2.3c) p is then well approximated by

$$p \approx \frac{\pi d}{60c^2\sigma} f^2 \quad (2.4a)$$

where c is the velocity of light in free space. This same approximation can be seen from (2.3f).

For a poorly conducting (dielectric) earth x in (2.3e) will be small compared with $\epsilon_r + 1$ so that the square root term in (2.3f) approaches $\epsilon_r + 1$, leaving

$$p \approx \frac{\pi d}{c(\epsilon_r + 1)} f \quad (2.4b)$$

As seen from both of these approximations, p increases with frequency for both conducting and dielectric earths, and is inversely proportional to conductivity for conducting earths. Consequently, p is small when the frequency is small, leading to an attenuation factor close to unity as seen from Fig. 2.2; this is the case also for

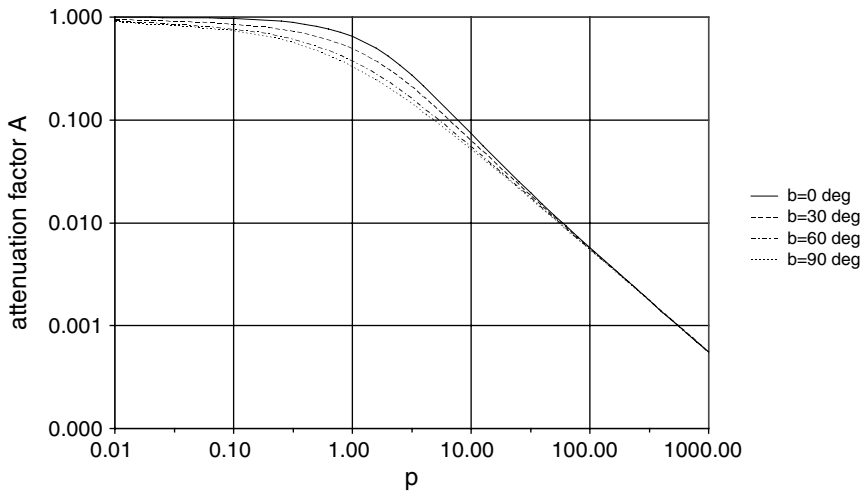


Fig. 2.2 The attenuation factor of (2.2) for vertical polarisation

highly conducting earths such as sea water. As frequency rises, or the conductivity of the earth becomes poorer, p becomes larger leading to a decrease in A as seen in Fig. 2.2. As a result, from (2.1), field strength decreases with rising frequency and falling conductivity.

Field strength curves for the surface wave at several frequencies are plotted in Fig. 2.3 covering the extreme cases of propagation over sea water and propagation over a poorly conducting soil. As expected, at a given frequency a greater range is possible over ocean than over land. It is important to note that these curves ignore

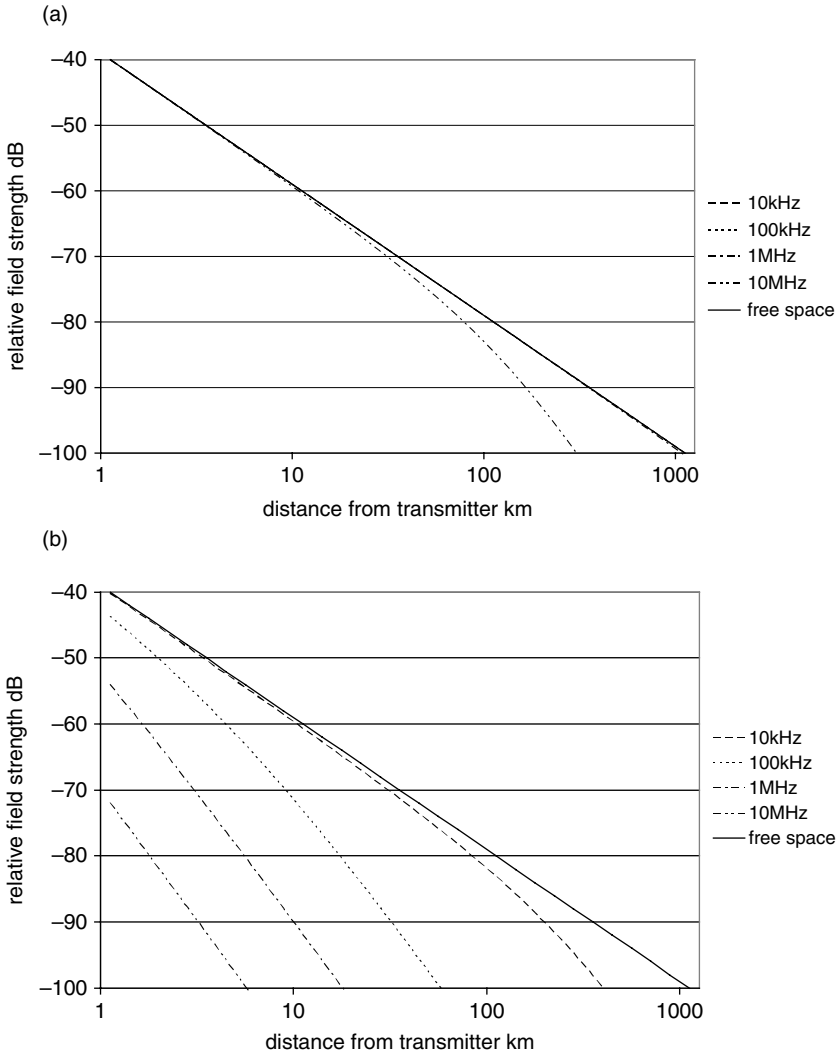


Fig. 2.3 Relative field strength curves for (a) sea water with $\epsilon_r = 80$, $\sigma = 4 \text{ Sm}^{-1}$ and (b) a poor earth with $\epsilon_r = 5$, $\sigma = 0.00001 \text{ Sm}^{-1}$; the effect of earth curvature has not been taken into account

the effect of earth curvature on propagation, which acts to reduce field strengths further at larger distances. Atmospheric refraction can also be important; it is usually accounted for by a $4/3$ adjustment to the earth's radius in any computations. Chapter 4 shows how this figure arises.

The surface currents in the earth that support the surface wave as depicted in Figs. 1.5c and 2.1 generate energy loss because of the finite conductivity of the earth, which is why the field strength decreases as discussed above. Sometimes effort is made to minimise those losses by burying conducting elements in the ground near the transmitter where the current densities are highest. Losses take energy out of the wave; that downwards flow, along with the forward propagation of energy in the wave itself, results in a tilted wavefront. When resolved into vertical and horizontal components the wave is seen to have a small field component in the direction of propagation (although not at the surface of the earth itself).

2.3 Incorporating Earth Curvature and Atmospheric Refraction

Earth curvature and atmospheric refraction are incorporated in the plots of Fig. 2.4; to gauge their significance an example is included in which both effects are ignored.

While reasonable propagation paths (say no more than 30 dB below free space propagation out to about 500 km) can be obtained over the ocean for frequencies up to about 10 MHz, the same performance can only be achieved over good conducting

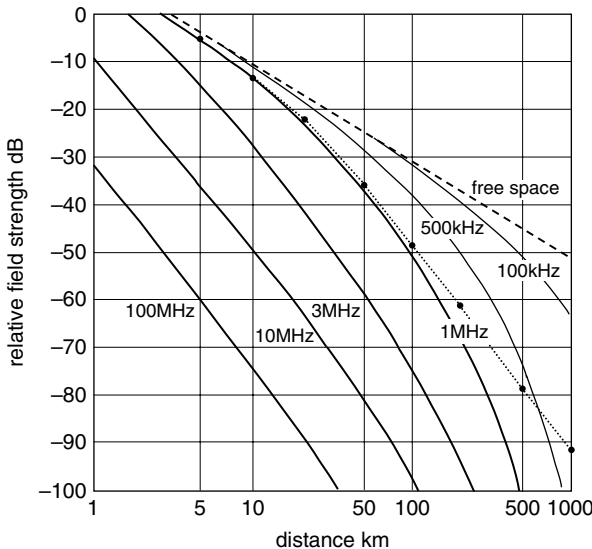


Fig. 2.4 Relative field strength curves for a good earth with $\epsilon_r = 15$, $\sigma = 0.01 \text{ Sm}^{-1}$, including the effect of earth curvature – two curves are shown for 1 MHz with the dotted one ignoring those effects; adapted from E.V.D. Glazier and H.R.L. Lamont, *Transmission and Propagation*, HMSO, London, 1958

earths for frequencies up to about 500 kHz or so. For poor earths the range is more limited, as can be seen in Fig. 3.18 of Rohan.²

At low frequencies the curvature of the earth can cause measurable diffraction of the surface wave;³ beyond about 1 MHz diffraction is negligible.

2.4 Propagation at Very Low Frequencies

Even though the bandwidths are small, some services rely on propagation in the VLF, ULF and ELF ranges, particularly for underwater and global operation. Table 1.1 reminds us how long the wavelengths are in those bands. As with simple surface wave propagation above, such extremely long wavelengths imply that vertical polarisation is the only viable mode, since horizontally polarised antennas cannot be placed far enough above the earth's surface to avoid significant short circuiting of their signal. The only available propagation mechanism is the surface wave. However, the wavelength is so large that the ionosphere plays an important part in propagation in those bands. The ionosphere is a region of charged particles of the order of 100 km above the earth's surface; we treat it in some detail in Chap. 3. It is sufficient for our purposes here though simply to note that it represents a conducting layer in the sky, capable of terminating electric field vectors. The height of the ionosphere is generally less than a wavelength above the earth at such low frequencies. Therefore the electric field lines terminate on surface charges on the earth and largely in the so-called D region of the ionosphere, as depicted in Fig. 2.5. This gives what is called a waveguide – i.e. a pair of conducting boundaries capable of supporting a wave between them as it propagates.

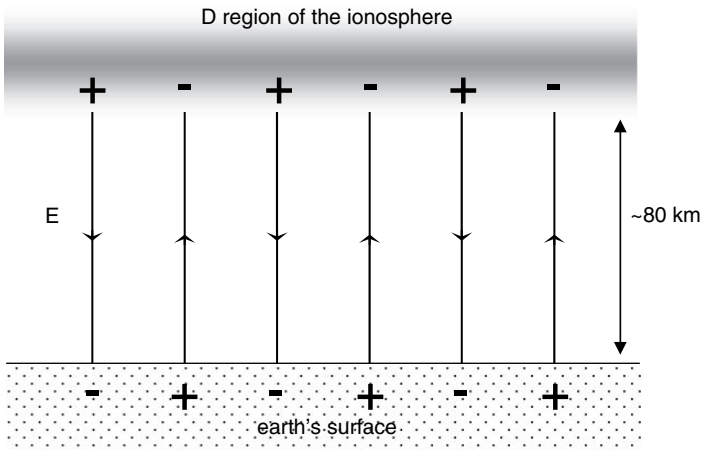


Fig. 2.5 Earth-ionosphere waveguide formed at VLF and lower frequencies

² *ibid.*

³ For a brief discussion of diffraction see Sect. 4.3.

Figure 2.5 shows only a localised region of the earth and ionosphere. In reality, of course, they are idealistically co-spherical as shown in Fig. 2.6, so that the waveguide formed is more in the nature of a spherical cavity, the theory of which is well known but beyond the scope of this treatment. In the figure we have indicated typical conductivities of the surfaces involved in supporting the very long wavelength field lines.

The magnitude E of the electric field vector $|\mathbf{E}|$ for a travelling wave can be represented as (Chap. 7)

$$E = E_0 e^{j\omega t - \gamma z}$$

in which γ is the *propagation constant*, given by

$$\gamma^2 = j\omega\mu\sigma - \omega^2\mu\epsilon \quad (2.5)$$

The constants σ , μ and ϵ are the conductivity, permeability and permittivity of the medium through which the wave propagates. Equation (2.5) can be re-written

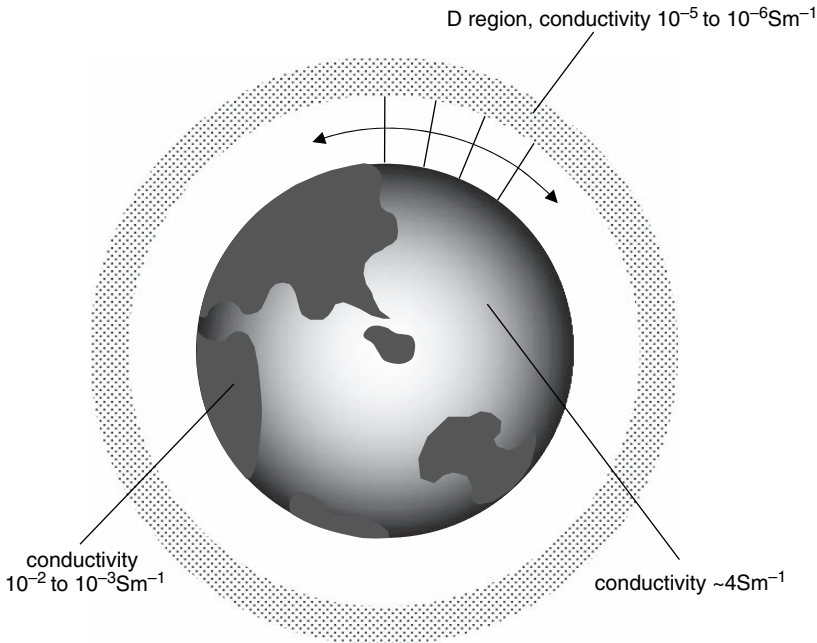


Fig. 2.6 The earth-ionosphere spherical waveguide (or cavity) assuming, for illustration, that the D region is always present

$$\begin{aligned}
\gamma^2 &= -\omega^2 \mu \varepsilon \left(1 - \frac{j\omega \mu \sigma}{\omega^2 \mu \varepsilon} \right) \\
&= -\omega^2 \mu \varepsilon_0 \varepsilon_r \left(1 - \frac{j\sigma}{\omega \varepsilon_0 \varepsilon_r} \right) \\
&= -\omega^2 \mu \varepsilon_0 \left(\varepsilon_r - \frac{j\sigma}{\omega \varepsilon_0} \right) = -\omega^2 \mu \varepsilon_0 \varepsilon_r^*
\end{aligned}$$

in which

$$\varepsilon_r^* = \varepsilon_r - \frac{j\sigma}{\omega \varepsilon_0} \quad (2.6)$$

is said to be the *complex dielectric constant* of the medium.

Note that for lossless media, $\sigma = 0$ so that $\varepsilon_r^* = \varepsilon_r$ and $\gamma^2 = -\omega^2 \mu \varepsilon$, giving

$$\gamma = j\beta \text{ in which } \beta = j\omega\sqrt{\mu\varepsilon}, \varepsilon = \varepsilon_r \varepsilon_0$$

For highly conducting media in which the second term of (2.6) dominates, we have

$$\varepsilon_r^* = -\frac{j\sigma}{\omega \varepsilon_0}$$

For seawater at 1 kHz, $\varepsilon_r \approx 80$, $\sigma = 4 \text{ Sm}^{-1}$ and $\varepsilon_0 = 8.85 \text{ pFm}^{-1}$ that will be the case, so that

$$|\varepsilon_r^*| = \frac{4}{2000\pi \times 8.85 \times 10^{-12}} \approx 10^8$$

Since refractive index is seen in (1.9) to be the square root of dielectric constant, the refractive index of sea water at these very low frequencies is of the order of 10^4 ! A wave propagating in the air above the surface of the ocean, on encountering the sea surface, will be very strongly refractive as depicted in Fig. 2.7. A consequence of this is that the vertically polarised wave in the air becomes horizontally polarised in the water. Therefore, to receive a signal in sea water horizontally polarised antennas are required, as they will also be if transmission occurs underwater for reception on land via the air path.

We should now consider the effect of the sea water on the wave propagating below the surface. From (2.5), we see that for highly conducting, non-magnetic media

$$\gamma^2 = j\omega\mu_o\sigma$$

Thus

$$\gamma = (1 + j)\sqrt{\frac{\omega\mu_o\sigma}{2}}$$

so that in sea water the wave propagates according to

$$E = E_o e^{j\omega t - \gamma z} = E_o e^{j\omega t} e^{-az} e^{-jaz}$$

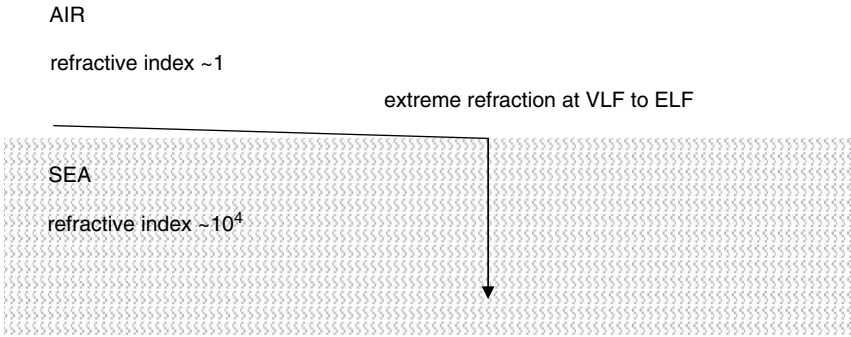


Fig. 2.7 Propagation into sea water at extremely low frequencies

in which

$$a = \sqrt{\frac{\omega\mu_0\sigma}{2}}$$

The term e^{-az} indicates attenuation of the signal and the term e^{-jaz} denotes phase change during propagation. The depth at which the signal has diminished to $1/e$ of its surface value – the so called *skin depth* – is given by

$$z_{skin} = \alpha^{-1} = \sqrt{\frac{2}{\omega\mu_0\sigma}}$$

At one skin depth the signal is attenuated by $20\log e^{-1} = 8.7\text{ dB}$! For each additional depth equivalent to the skin depth there is a further 8.7 dB attenuation. Table 2.1 shows the skin depth of sea water at a number of very low frequencies.

Is this substantial level of attenuation a problem? For underwater reception it may not be such a difficulty since noise from the surface is attenuated to the same level as the signal; in other words the signal to noise ratio at the receiver is the same as that at the surface. Reliable reception depends therefore on the equivalent input noise temperature of the receiver as will be appreciated from the material of Chap. 5. Transmission however is a different matter. The signal is significantly attenuated before reaching the surface; it then encounters atmospheric noise (and indeed any other source) in the air, leading to a poor signal to noise ratio. The usual practice

Table 2.1 The skin depth of sea water

Frequency	Skin Depth
10 kHz	2.5 m
1 kHz	8 m
100 Hz	25 m

therefore is that underwater vessels such as submarines can receive at depth if their receivers are low noise, but need to come to the surface, or at least have an antenna above the surface, for transmission.

In addition to the attenuation in the sea water there is also a very major coupling loss over the water-air interface and usually a significant antenna loss. These can be of the order of 60 dB, so that shore transmitting stations generally need to be very high power.

Problems

2.1. A particular AM broadcast radio station operates on a carrier frequency of 1.2 MHz and radiates 10 kW on an antenna of gain 3 with respect to isotropic. It is located near the coast. Determine the field strength of the station's transmission 45 km off shore based on the assumption of surface wave propagation.

2.2. A two way radio system involves a base station and a number of vehicle-mounted transceivers. If the base station radiates 100 W at 45 MHz using a 30 m monopole antenna placed on reflecting surface, and the vehicles carry monopole antennas of approximate height 1.5 m above the ground, find the signal strength at a range of 5 km assuming the earth is perfectly reflecting. Repeat the exercise for a very dry earth.

2.3. An AM radio station broadcasts 5 kW on 666 kHz using a quarter wave monopole antenna. What signal strength is received at 25 km range over sea, a poor earth and a highly conducting earth. Make reasonable assumptions for the properties of the earth in each case.

2.4. If sea water has a dielectric constant of about 81 and a conductivity of about 4 Sm^{-1} , verify by reference to (2.5) that it acts as a conductor at 1 MHz. What is its wave impedance at 1 MHz? An aircraft flying over the sea transmits a signal vertically downwards at 1 MHz giving a power density at the sea surface of 1 mWm^{-2} . What is the electric field strength at the sea surface (on the air side) and 1 m below the sea surface? The sub-surface field will depend on the transmission coefficient at the air-sea boundary, defined in (7.8b).

2.5. Learning from Problem 2.3, can you devise a guideline from (2.5) that can be used to determine if a material is predominantly conducting or predominantly dielectric? Is that a function of frequency?

2.6. The signal strength at 30 km from a transmitter radiating at 800 kHz over land with dielectric constant 4 is about 20 dB lower than the free space value. What is the approximate conductivity of the ground?

2.7. Describe what the antenna on a submarine might look like if it were to receive a signal at 12 kHz a few metres below the surface of the ocean.

Chapter 3

The Sky Wave

The sky or ionospheric wave depends on the ionosphere for its propagation path. It is important therefore to understand some of the properties of that very interesting part of the earth's atmosphere before analysing the nature of sky wave propagation.

3.1 The Ionosphere

The ionosphere is a region of the earth's upper atmosphere that is partially ionised by incident sunlight. It occupies altitudes from about 80 km to 400 km, beyond which there is not sufficient atmospheric density to develop significant levels of ionisation. Below 80 km atmospheric density is so high that recombination reactions dominate over dissociation, ensuring very little ionisation. This situation is depicted in Fig. 3.1.

Because of variations in atmospheric constituents and the wavelength range of the incident sunlight, the region of ionisation occurs in a number of *layers* as illustrated in the daytime ionosphere of Fig. 3.2a. The lowermost layer is usually not distinct and so is generally referred to as the *D region*. In the evening solar radiation is, of course, absent. There is still some residual ionisation because recombination rates are not sufficiently rapid to deplete the ionisation completely. As a result, a typical night time ionosphere appears as shown in Fig. 3.2b. While some variation with height is still apparent, only two layers are generally recognised as forming the ionosphere at night.

In the plots of Fig. 3.2 the degree of ionisation is represented in terms of the density of free electrons. Ion density could just as easily be used but it is the electrons that actually affect the propagation of radio waves.

Since the strength of solar radiation varies with season and time of day the layers show considerable changes with time. Figure 3.3 depicts seasonal and diurnal dependence of the layers. Note that the altitudes of the layers are described as *virtual heights* which can be quite different from their actual heights. That concept

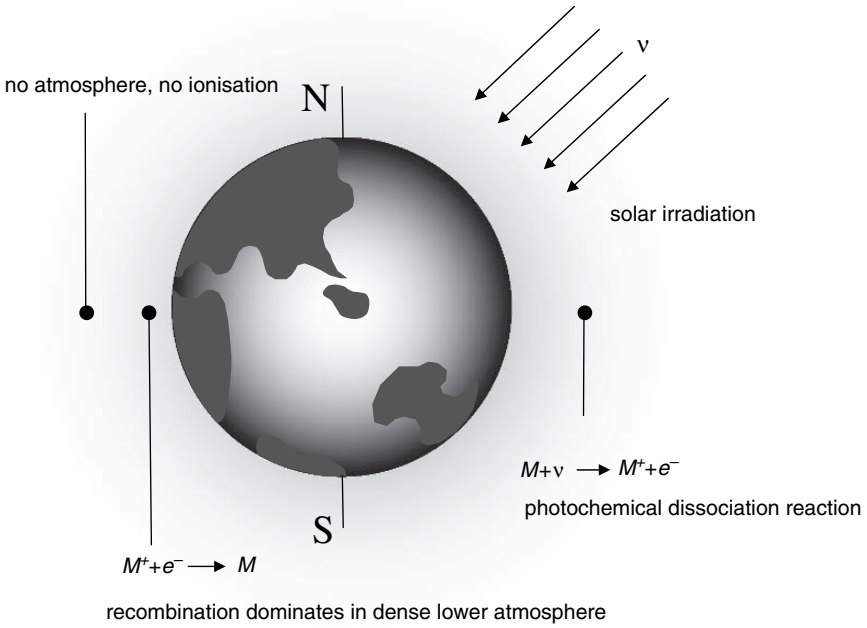


Fig. 3.1 Formation of the ionosphere through photochemical ionisation of atmospheric constituents by incident sunlight

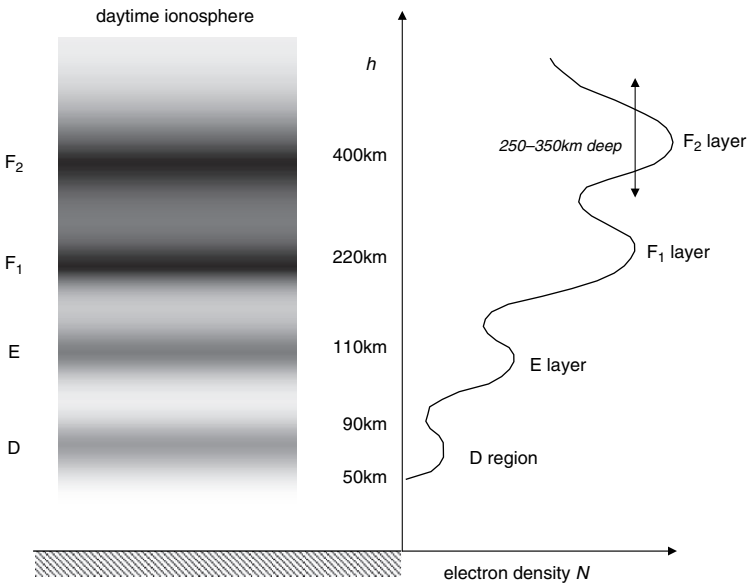


Fig. 3.2a Typical day-time ionosphere, showing the separation into layers; darker shading represents higher electron densities

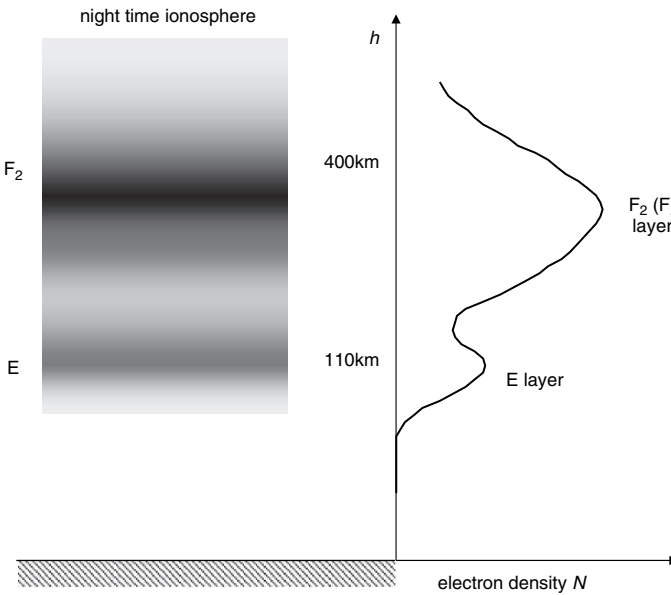


Fig. 3.2b Typical night time ionosphere. Darker shading represents higher electron densities

will become clearer later as will the property of critical frequency, which is directly related to the electron density maximum of a layer.

Solar strength also varies with latitude so that there will be geographical variations in the degree and height of ionisation. Sometimes there is also a drifting cloud of electrons both day and night between 90 km and 130 km that is unpredictable, called a *sporadic E layer*.

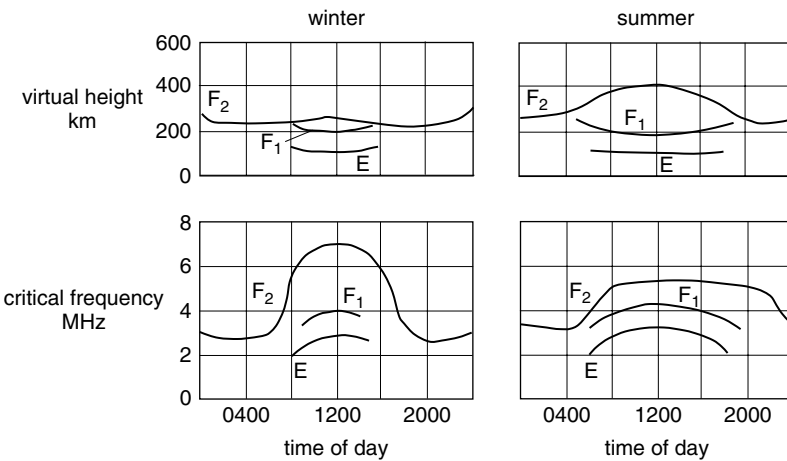


Fig. 3.3 Indicative seasonal and diurnal variations of the ionosphere; adapted from F.E. Terman, *Electronic and Radio Engineering*, 4th ed., McGraw-Hill, Tokyo, 1955, Fig. 22–20

Table 3.1 Overview of the characteristics of ionospheric layers

Layer	Height	Characteristics
D	This is a diffuse region between about 50 km and 100 km	It only occurs in daylight. It has low ionisation and thus large neutral density; the electron density is not high enough to influence the propagation of a radio wave. It can be strongly attenuating.
E	This layer reaches a maximum density between 100 km and 150 km virtual height	It remains weakly ionised at night and has a relatively constant height during the day. The electron density is lower in winter because of the lower solar irradiation.
F ₁	This layer reaches a maximum density at about 200 km virtual height	It exists only during the daytime.
F ₂	This layer reaches a maximum density between 200 km and 400 km virtual height depending on the season	It shows considerably more variation than the others because it is uppermost and most strongly ionised, reaching its highest altitude in the summer months. There is a single F ₂ (or F) layer in the evening.

Anything that affects the strength of the solar radiation reaching the earth can influence the degree of ionisation of the earth's atmosphere. Variations in the sun's own energy output as a result of sun spot activity and solar flares are important in this regard. The number of sun spots vary with an approximate 11 year cycle.

Table 3.1 summarises the characteristics of the layers. Importantly, note that the D region really does not contribute to the ionosphere as a propagating medium; instead, when present, it acts as an attenuator as discussed in Problem 3.2.

3.2 The Refractive Index of an Ionospheric Layer

As might be expected, when an electromagnetic ray enters a layer in the ionosphere the free electrons move in response to the oscillating electric field vector. The free ions also respond but far less than the electrons because of their much higher masses. We will see in the following that the interaction of the ray and the ionised layer leads to refraction of the wave. As a result, it is necessary to have some idea of the refractive index of an ionised medium. Appendix C shows that a region with ionisation density of N electrons per cubic metre has a refractive index of

$$n = \sqrt{1 - \frac{81N}{f^2}} \quad (3.1)$$

where f is the frequency of the incident radiation. Two points are immediately noteworthy about this expression. Unlike our experience with many dielectric media, the

refractive index (and thus dielectric constant, which is the square of the refractive index) is a strong function of the frequency of the radiation being refracted. Secondly, the refractive index of the ionised region has an upper bound of unity, and falls below unity when the electron density increases from zero (i.e. the case of no ionisation). Note that it is also feasible for the refractive index to become imaginary with the right combination of frequency and electron density.

3.3 Refraction of a Radio Wave in the Ionosphere

In order to analyse the effect a layer has on the passage of a radio wave we segment a layer of continuously varying electron density (and thus refractive index) into a number of slices, within each of which we assume the properties are constant, as noted in Fig. 3.4.

Suppose we now focus on a ray entering a layer of the ionosphere from below, where it encounters an electron density that increases from zero to a maximum and then falls again, as suggested by the plots in Fig. 3.2. Figure 3.5 shows part of that layer from the lowest, where the ray enters, up to a region of higher (but not yet necessarily the maximum) electron density, and thus lower refractive index. Because the refractive index falls with height the electromagnetic ray will refract away from the normal as it progresses from slice to slice as shown. Suppose the

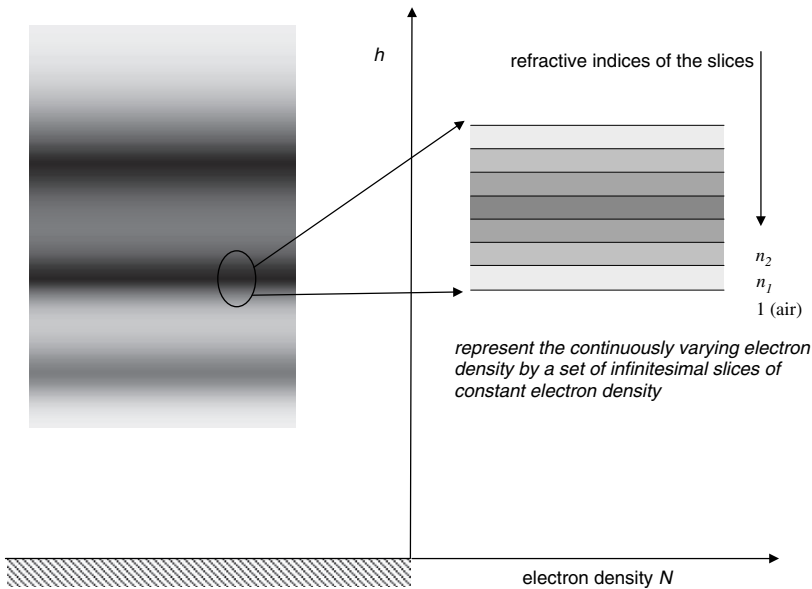


Fig. 3.4 Representing a region of continuously varying ionisation by a set of strips within each of which the electromagnetic properties are assumed constant

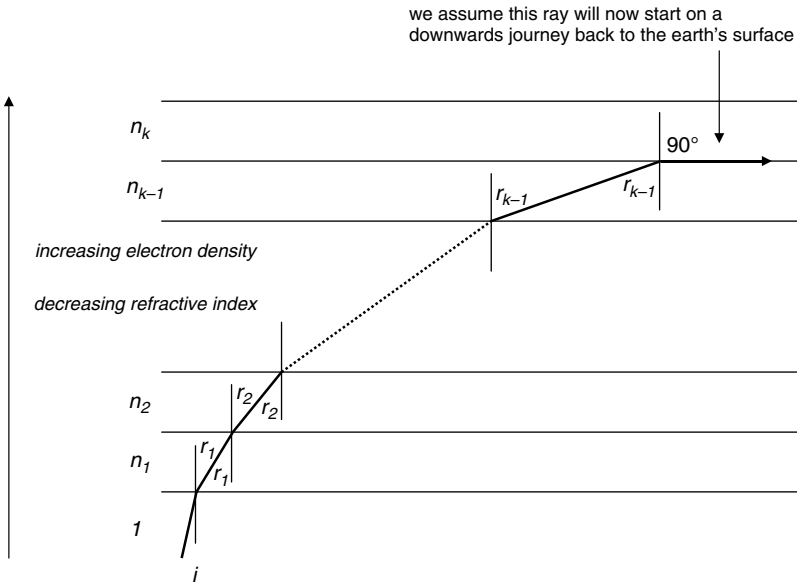


Fig. 3.5 Refraction of a radio wave in an ionospheric layer and the 90° assumption of return to the earth’s surface

ray is refracted so that ultimately the angle of refraction is 90°; in other words it then ends up (momentarily) travelling horizontally. We make the assumption that it will then refract back downwards again; thus the 90° condition is considered as that necessary for sufficient refraction having occurred that the ray will return to the earth’s surface.

Application of Snell’s Law of Refraction to the geometry of Fig. 3.5 shows

$$\frac{\sin i}{\sin r_1} = \frac{n_1}{1}, \frac{\sin r_1}{\sin r_2} = \frac{n_2}{n_1}, \dots, \frac{\sin r_{k-1}}{\sin 90} = \frac{n_k}{n_{k-1}}$$

so that

$$\frac{\sin i}{\sin r_1} \times \frac{\sin r_1}{\sin r_2} \times \dots \times \frac{\sin r_{k-1}}{\sin 90} = \frac{n_k}{1}$$

which gives as the condition that the ray be returned to the earth’s surface that

$$\sin i = n_k \tag{3.2}$$

Therefore for the ray to be returned there must be a region in the layer where the electron density gives rise to a refractive index numerically equal to the sine of the angle of incidence at the point of entrance to the ionosphere. Note that if $i = 90^\circ$ – a near horizontal ray at the layer – $\sin i = 1$. Thus, only a very weak level of ionisation is needed for return of the ray to the earth’s surface. As the angle of incidence is decreased from 90° the ray must penetrate further into the layer to encounter the

necessary refractive index for return. If the electron density is not sufficiently large to generate the required low refractive index, then the ray will pass through the layer to become an *escape ray*, as illustrated in Fig. 3.6. Clearly the likelihood of a ray being an escape ray depends on the angle of incidence. Also note that it is possible that there will be a region out from the transmitter in which there may be no return signal because of the escape ray condition.

The question that now arises is: can circumstances exist under which there is no escape ray and for which receivers even adjacent to the transmitter can receive the sky wave? That would require the signal to be returned at vertical incidence and, thus, from (3.1) and (3.2)

$$\sin i = n = \sqrt{1 - \frac{81N}{f^2}} = 0$$

which in turn requires there to be, somewhere in the layer, a region of electron density such that

$$f = 9\sqrt{N} \quad (3.3)$$

This very interesting equation says we will always get return at vertical incidence (and indeed at all angles of incidence) provided we can find an electron density somewhere in a layer that is related to our operating frequency in this manner. Since

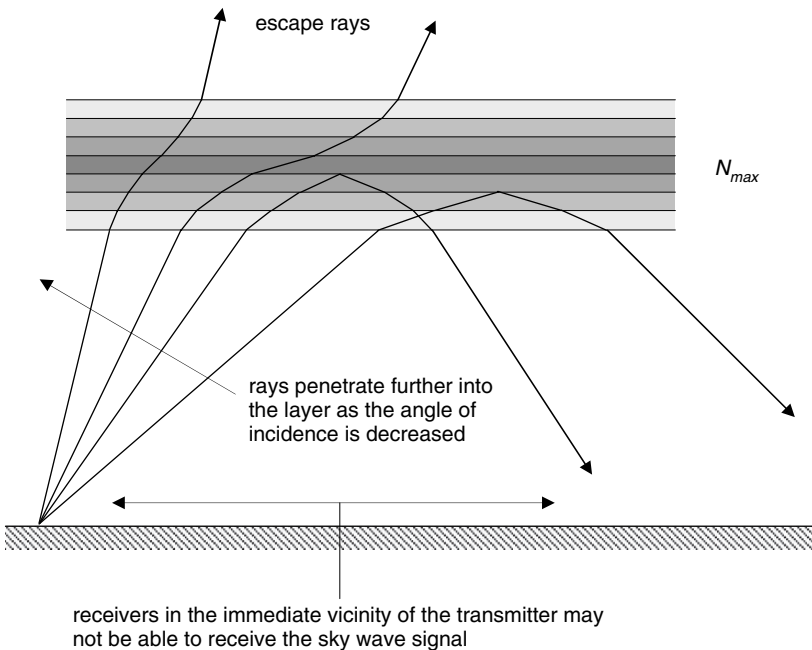


Fig. 3.6 The concept of an escape ray and its dependence on the angle of incidence

electron density increases with penetration into a layer (until the electron density maximum is reached) that means that signals with lower frequencies will be returned from lower in the layer. As frequency is increased the return of the signal will come from higher up in the layer until ultimately, at the electron density maximum, we have the highest frequency that will be returned for that layer. That frequency is called the *critical frequency* of the layer, and is given by

$$f_o = 9\sqrt{N_{\max}} \quad (3.4)$$

where N_{\max} is the electron density maximum for the layer. Note from (3.1) and (3.2) that whenever the signal is returned at vertical incidence (including the limiting case of the critical frequency) the region of electron density encountered has a refractive index of zero at the operating (or critical) frequency.

All radio waves at vertical incidence with a frequency above the critical frequency will be escape rays. Note that the critical frequency is not the highest frequency that will be returned by the layer; rather, it is the highest frequency at vertical incidence. At larger angles of incidence (3.2) shows that signals with higher frequencies can be returned to the earth's surface.

3.4 The Set of Critical Frequencies

Each layer of the ionosphere has its own critical frequency because it has its own electron density maximum; and because the maxima of the layers get progressively larger with altitude as seen in Fig. 3.2 it is possible to think of a simple experiment that will reveal the critical frequencies of each of the layers, and their heights.

If we fire a pulse of electromagnetic energy vertically to the ionosphere, starting with a low (carrier) frequency, the pulse will “reflect” from the lower reaches of the E layer (ignoring the D region since it does not have enough ionisation in general to cause refraction). If we increase the frequency the pulse will be returned from further up in the E layer until the critical frequency of the E layer is encountered. A further increase in frequency will see the pulse “punch” through the E layer and start reflecting from the lower reaches of the F_1 layer, where the electron density is comparable to that of the E layer maximum.

Increasing the frequency more will cause the pulse to travel further into the F_1 layer before reflection, until the F_1 layer critical frequency is reached. The pulse then will punch through the F_1 layer and start reflecting from the F_2 layer where the electron density is comparable to that of the F_1 maximum. Further increase in frequency will see the pulse reflecting from further up in the F_2 layer until its critical frequency is reached, following which the pulse then passes through the F_2 layer with no further prospect of being returned to the earth. Figure 3.7 shows this process diagrammatically. The experiment is referred to as ionospheric sounding and the instrumentation used to conduct it is called as *ionosonde*. The record is usually called an *ionogram*.

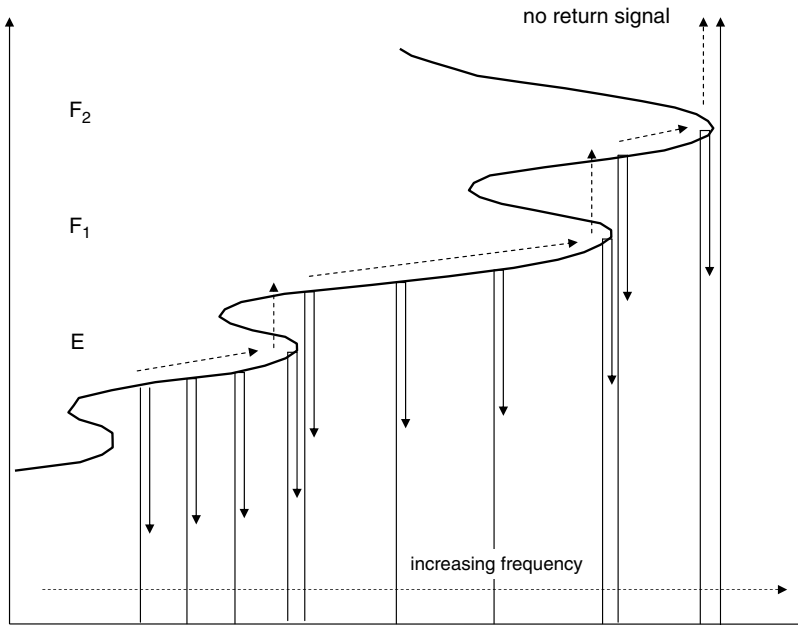


Fig. 3.7 Characterising an ionosonde experiment

In Fig. 3.8 we see the expected results and a stylised set of what the actual results might look like. The latter show considerable departure from the expected values in the vicinity of the critical frequencies and, at the critical frequencies, the layer height tends to infinity. Height measurements are indicated as *virtual* height. We now need to understand all of those properties. That requires an understanding of the difference between phase and group velocities, which is developed in the next section.

3.5 The Virtual Height of an Ionospheric Layer

To find out how fast a travelling wave moves (such as that on the surface of the ocean, or an electromagnetic wave), all we need to do is lock on to a point of constant phase and see how fast that point moves as developed in Sect. 7.2.

A continuous sinusoidal waveform, represented by $\cos \theta$, will be travelling if

$$\theta = \omega t - \beta z \quad (3.5)$$

in which z is the spatial coordinate and β is the phase constant of the wave, given by

$$\beta = \omega \sqrt{\mu \epsilon} \quad (3.6)$$

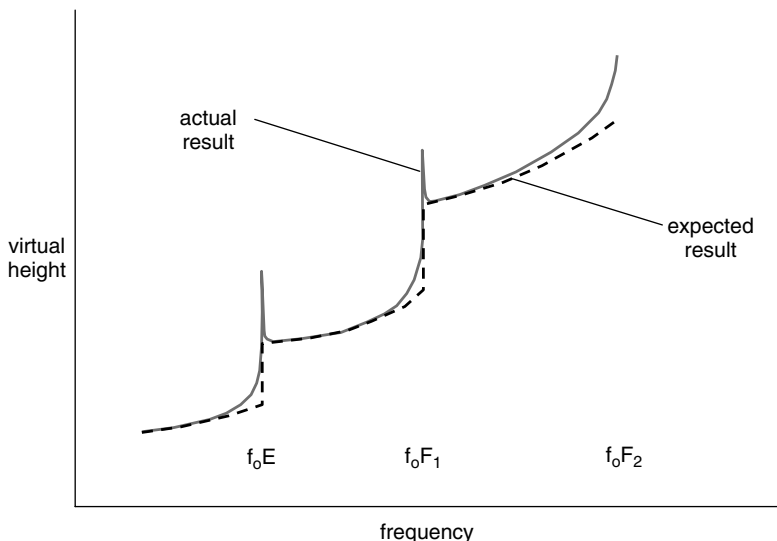


Fig. 3.8 Virtual height as a function of frequency from a hypothetical ionosonde experiment

By putting $\theta = \text{constant}$ in (3.5) we can see that the velocity of the sinusoid is

$$v_{\text{phase}} = \frac{\partial z}{\partial t} = \frac{\omega}{\beta} = \frac{1}{\sqrt{\mu\epsilon}} \quad (3.7)$$

That velocity is referred to as the *phase velocity* since it is the speed with which a point of constant phase is seen to move.

A continuous sinusoidal wave carries no information; neither does it have a time reference point or marker that would enable it to be a useful signal in an ionosonde experiment. We need to modulate the sinusoid to make it useful for time delay experiments of that type. Of course, the pulse modulation used in an ionosonde is imposed exactly for that purpose. By seeing how long the pulse takes from transmission to reception we can gauge the height of the layer. Interestingly, the pulse (i.e. the modulation) does not necessarily travel at the phase velocity. In order to use a modulated waveform for ionosonde purposes we need to determine the velocity of the modulation envelope.

Handling pulse modulation is unnecessarily complicated; we can derive the necessary theory by considering one of the simplest of all modulations – double side band suppressed carrier (DSBSC). A DSBSC signal consists of just two side bands, above and below the carrier frequency. It can be written, if travelling, as

$$\cos(\omega_1 t - \beta_1 z) + \cos(\omega_2 t - \beta_2 z)$$

Note that the phase constants will be different for the upper and lower sidebands because they are frequency dependent, as noted from (3.6). The last expression can be re-written as

$$2 \cos(\omega_0 t - \beta_0 z) \cos(\Delta \omega t - \Delta \beta z) \quad (3.8)$$

where

$$\begin{aligned}\omega_o &= \frac{\omega_1 + \omega_2}{2}, & \beta_o &= \frac{\beta_1 + \beta_2}{2} \\ \Delta\omega &= \frac{\omega_1 - \omega_2}{2}, & \Delta\beta &= \frac{\beta_1 - \beta_2}{2}\end{aligned}$$

The first term in (3.8) is the (suppressed) carrier and the second is the modulation which, recall, carries the information. Note that the modulation frequency is $\Delta\omega$ and its phase is $\Delta\beta$. By choosing a point of constant phase on the modulation envelope we can see that the speed with which the modulation travels is given by

$$v_{\text{group}} = \frac{\partial z}{\partial t} = \frac{\Delta\omega}{\Delta\beta} \rightarrow \frac{\partial\omega}{\partial\beta} \quad (3.9)$$

We refer to this as *group velocity* since, in a sense, it represents the speed of the packet, or group, of waves travelling together that make up the signal.

Substituting (3.1) into (3.6), we have for an ionospheric layer

$$\beta = \omega\sqrt{\mu_o\epsilon_o\epsilon_r} = \frac{\omega}{c}n = \frac{\omega}{c}\sqrt{1 - \frac{81N}{f^2}} = \frac{1}{c}\sqrt{\omega^2 - (2\pi)^2 81N} \quad (3.10)$$

where n is the refractive index of the medium. Thus

$$\frac{\partial\beta}{\partial\omega} = \frac{1}{2} \frac{1}{c} 2\omega (\omega^2 - (2\pi)^2 81N)^{-1/2}$$

so that

$$\frac{\partial\omega}{\partial\beta} = \frac{c}{\omega} \sqrt{\omega^2 - (2\pi)^2 81N} = nc$$

Using $\mu = \mu_o$, $\epsilon = \epsilon_o\epsilon_r$ in (3.7) we therefore have, in summary,

$$v_{\text{group}} = nc \quad (3.11a)$$

$$v_{\text{phase}} = \frac{c}{n} \quad (3.11b)$$

and

$$v_{\text{group}}v_{\text{phase}} = c^2. \quad (3.11c)$$

Even though derived on the basis of DSBSC these results apply to any modulation. We can now interpret the results shown in Fig. 3.8. Note that as the frequency approaches the critical frequency of a layer, and the pulse penetrates further into the layer, a falling refractive index is encountered; from (3.11a) this is seen to cause the pulse modulation to drop below the speed of light, the free space value. At the critical frequency, which occurs at the electron density maximum, the refractive index is zero, implying that the pulse velocity falls to zero. Thus when trying to establish the height of an ionospheric layer using the delay between transmission and reception of a pulse, the height will be overestimated when assuming (as with radar in free

space) that the signal travels at the speed of light; that is why it is referred to as *virtual height*. At the critical frequency the height will appear to be infinite because the group velocity approaches zero.

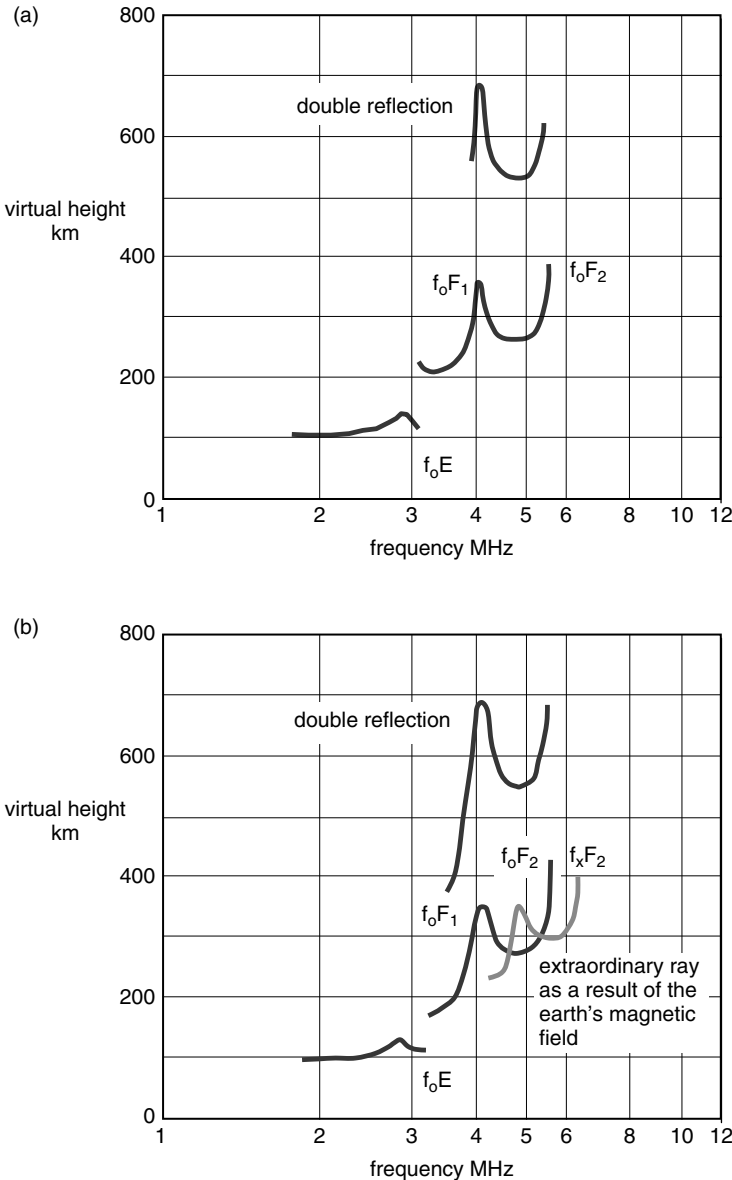


Fig. 3.9 Ionograms derived from recordings made by the Australian Ionospheric Prediction Service at (a) Sydney on 16 Sep 2007 and (b) Brisbane on 15 Sep 2007; note the extraordinary ray in (b)

Figure 3.9 shows ionograms that display those characteristics. Two further features should be noted. First, there is a second set of returns higher up, corresponding to the pulse reflecting from the ground on reception and then travelling the ionosphere-earth path a second time. Secondly, the curves in Fig. 3.9b are split in two. That is a result of the fact that the electrons in the ionospheric layer, while responding to the electric field vector of the passing radio wave, are also oscillating in the presence of the earth's magnetic field. The two separate traces are labelled O for "ordinary" and X for "extra-ordinary".

3.6 Maximum Usable Frequency and Skip Distance

The critical frequency of a layer is the highest frequency that will be returned at vertical incidence; at larger incidence angles higher frequencies can be returned but, as we saw in Fig. 3.6, the penalty is a zone out from the transmitter, known as the *skip distance*, within which a signal cannot be received. Nevertheless, it is useful now to determine an expression for the highest frequency for return at a given angle of incidence: that is called the *maximum usable frequency* (MUF).

Equation (3.2) gives the condition for the return to the earth's surface of a sky wave. Together with (3.1) it gives

$$\sin i = \sqrt{1 - \frac{81N}{f^2}}$$

If N is the electron density maximum this last expression can be written from (3.4) as

$$\sin i = \sqrt{1 - \frac{f_o^2}{f^2}}$$

so that the maximum usable frequency is

$$f = f_o \sec i \quad (\text{MUF}) \quad (3.12a)$$

For ranges beyond about 1000 km the curvatures of the earth and ionosphere need to be accounted for. That is done with a correction factor k :

$$f = kf_o \sec i = k \sec i f_o = \text{MUF factor} \times f_o \quad (3.12b)$$

Tables and maps of MUF factors are available, sometimes represented in the form of T factors, which give MUF values accounting for range of mechanisms that can affect the degree of ionisation.¹

¹ See the web sites for services that monitor ionospheric conditions continually, such as the Australian Ionospheric Prediction service www.ips.gov.au.

All of this assumes that the layer in the ionosphere is reasonably static. Because of the strong time dependence of the ionosphere, the refraction is really quite dynamic so that any attempt to operate up to the calculated maximum usable frequency is likely to cause signal loss from time to time when the layer critical frequency falls. In practice operation is usually restricted to about 85% of the calculated MUF. That is often called the *optimum working frequency* (OWF).

3.7 Range of the Sky Wave

Figure 3.10 shows the geometry of sky wave propagation, in which the limitations imposed by earth curvature are evident.

It is readily shown that

$$\theta = \cos^{-1} \frac{a}{a+h}$$

so that

$$d = 2a \cos^{-1} \frac{a}{a+h} \quad (3.13)$$

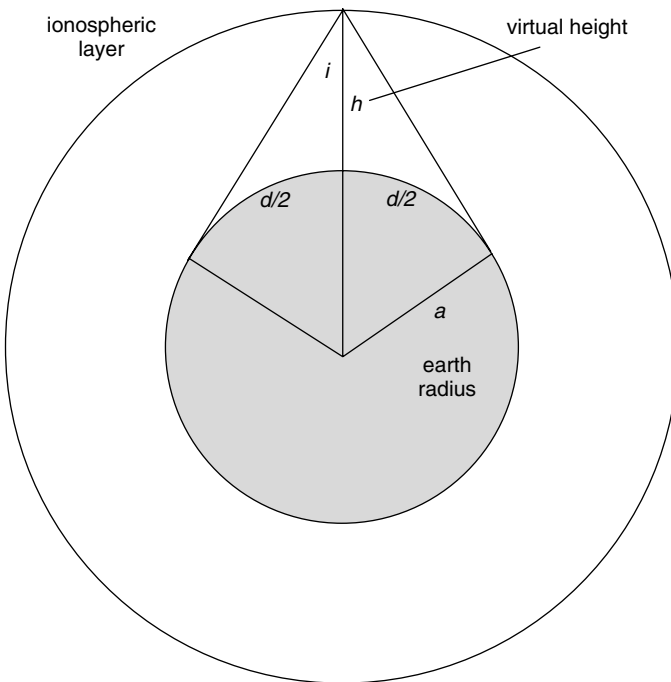


Fig. 3.10 The path of the sky wave at maximum range

where a is the radius of the earth and h is the virtual height of the layer. For a typical value of, say, $h = 400$ km we see that the maximum range is approximately $d = 4500$ km. In principle that requires the signal to be launched and received parallel to the ground – clearly a high loss strategy. In practice the transmitting antenna would need to angle the beam upwards slightly, thus reducing the range calculation. Nevertheless, it is possible to use sky wave propagation over very long distances because multiple hops are also possible. The downwards signal in Fig. 3.10 can reflect from the earth and travel to the ionosphere for a second refraction back to the earth. In principle that extends the range to about 9000 km. Compare that with the earth's circumference of about 40,000 km.

Problems

3.1. Refer to the satellite arrangements described in Problem 1.3.

- (a) Suppose the altitude of the satellite is to be checked using a pulse communications system in which a pulse is transmitted from a ground station at 16 MHz. On reception of the pulse at the satellite, and after an electronic delay of $10\mu\text{s}$ in the satellite's electronic circuits, the satellite responds by transmitting a pulse on a 20 MHz carrier. In the absence of an ionosphere what would be the round trip delay time between ground transmission and reception?
- (b) Consider now the effect of the ionosphere on the measurement in part (a). Suppose the ionosphere consists of a single layer between 300 km and 350 km of constant electron density of 1.4×10^{12} electrons per cubic metre. First verify

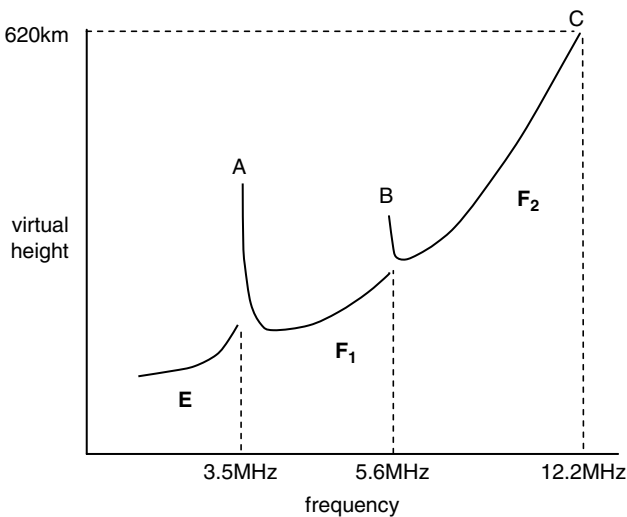


Fig. 3.11 Results of a hypothetical ionosonde test

that the transmitted and received signals will propagate through the ionosphere. By finding the appropriate group velocities determine the (erroneous) height that will be indicated by the test in (a).

3.2. It is often possible to receive a sky wave at considerable distances (1000s of km) in the evening when solar illumination is low. The same sky wave cannot be received during the day time. By reference to the properties of the D region, describe why that is the case.

3.3. For a particular radio service operating at 4.5 MHz suppose the F2 layer refracts the wave from a virtual height of approximately 400 km. By reference to the material of Fig. 2.4, and making any reasonable assumptions, compute the range at which the sky wave will have about the same field strength as the surface wave. What is the phase difference between those two waves on reception at that range? By how much would the height of the F2 layer have to vary to cause considerable fading and strengthening of a receiver placed at that location?

3.4. Suppose the results shown in Fig. 3.11 have been recorded at a specific location. What was the maximum electron density of the F₁ layer at the time the graph was recorded? What is the minimum frequency that could be used for reliable communication with a satellite? To answer this you could assume that the actual height of a layer is approximately its virtual height and that the radius of the earth is 6.37 Mm.

3.5. Using expressions for the phase constant of a propagating wave and the dielectric constant of an ionospheric layer plot a graph of ω (the operating frequency) versus β (the phase constant) for the layer. Use the graph to verify that the phase velocity is infinite at the critical frequency and that the group velocity approaches zero at the critical frequency. When the wave passes through the layer well above the critical frequency use the graph to show that the phase and group velocities are approximately the same.

3.6. Can an ionosonde be carried on a satellite orbiting at an altitude of 1500 km to diagnose the structure and properties of the earth's ionosphere?

Chapter 4

The Space Wave

Because of the very large number of services available above 30 MHz, the space wave is arguably the most important wave propagation mechanism we need to consider. Although moderately simple in many respects, the space wave can be annoying because of its line of sight limitations and the phenomenon known as multi-path.

4.1 The Received Field Strength

In principle, well removed from the earth's surface, the field strength of the space wave is given by the inverse distance law of (1.7). However, most services operate in the vicinity of the earth and propagation paths other than the direct ray from the transmitting to the receiving antennas are possible, most notably involving reflection from the ground, as shown in Fig. 4.1. As expected, the difference in path lengths between the direct and ground reflected rays will lead to a phase difference on reception; that will cause interference between the two rays. In addition, there is a further phase change introduced into the ground reflected ray at the point of reflection that adds to the interference. We will now analyse this situation to derive an expression for space wave field strength.

Consider the redrawn geometry of Fig. 4.2, from which we can see that the path lengths for the two rays are given by the following expressions, noting the simplifications possible since generally $d \gg h_t, h_r$.

$$d_d = \sqrt{(h_t - h_r)^2 + d^2} \approx d \left\{ 1 + \frac{1}{2} \left(\frac{h_t - h_r}{d} \right)^2 \right\} \quad (4.1a)$$

$$d_r = \sqrt{(h_t + h_r)^2 + d^2} \approx d \left\{ 1 + \frac{1}{2} \left(\frac{h_t + h_r}{d} \right)^2 \right\} \quad (4.1b)$$

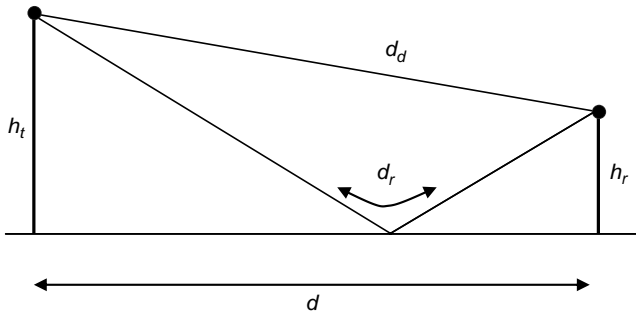


Fig. 4.1 The geometry of simple space wave propagation, involving a pair of interfering rays

From (4.1) the difference in path lengths is

$$\Delta d = d_r - d_d = \frac{2h_t h_r}{d}$$

from which the phase difference (calculated as 2π times the fractional difference in wavelength) is seen to be

$$\phi = \frac{\Delta d}{\lambda} \cdot 2\pi = \frac{4\pi h_t h_r}{\lambda d}$$

This is one element of the phase delay of the reflected ray compared with the direct ray. As noted above, there is a further phase delay introduced at the point of reflection with the ground. The electric field just after reflection compared with that just before is described by the *reflection coefficient* $\rho = |\rho|e^{j\psi}$, a complex quantity that describes a change in amplitude and phase (see Sect. 7.4). For glancing

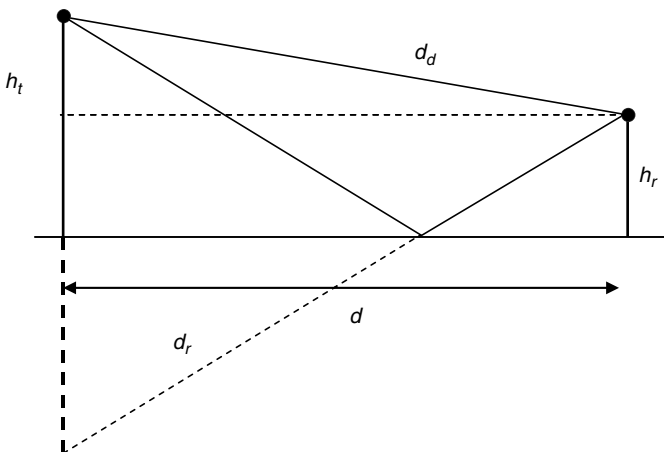


Fig. 4.2 Geometric construction for path length calculations

incidence – the case most commonly encountered – $|\rho| \approx 1$ and $\psi \approx 180^\circ$ so that $\rho = -1$. Thus the scalar magnitude of the vector field at the receiving antenna, as the result of both rays, is

$$E = E_d + E_r = E_d (1 + \rho e^{-j\phi}) \approx E_d (1 - e^{-j\phi})$$

This last approximation assumes that the magnitude of the reflected ray at the receiver is the same as the magnitude of the direct ray at the receiver. Because of their slightly different path lengths the magnitudes will be different as a result of the inverse distance law by which free space electric fields propagate. However, unlike the case with phase, the amplitude difference can generally be ignored.

From Euler's Theorem we have

$$E = E_d(1 - \cos \phi + j \sin \phi)$$

so that the amplitude of the received field strength is given by

$$\begin{aligned} |E|^2 &= |E_d|^2 \{ (1 - \cos \phi)^2 + \sin^2 \phi \} \\ &= |E_d|^2 \{ 1 - 2\cos \phi + \cos^2 \phi + \sin^2 \phi \} \\ &= 2 |E_d|^2 (1 - \cos \phi) \\ &= 4 |E_d|^2 \sin^2 \frac{\phi}{2} \end{aligned}$$

Thus

$$|E| = 2 |E_d| \sin \frac{\phi}{2} = 2 |E_d| \sin \frac{2\pi h_r h_t}{\lambda d} \quad (4.2)$$

Several important observations can be made about this last expression. First, it is clear that the field strength has maxima and minima that are determined by, and change with, variations in the heights of the transmitting and receiving antennas and the distance between the antennas. It varies also with changes in wavelength. Anyone who has tried to adjust a television receiving antenna will have experienced how even small changes in position can affect the quality of the received signal. The same is true of obtaining a good signal on a mobile telephone: moving just a small distance can improve reception. What is small? At 900 MHz the wavelength is approximately 33 cm so movement within that range can change the received signal significantly as we will see later.

The second point to note about (4.2) is that the maximum received field strength is twice that of the direct ray – i.e. twice the expected free space value. That represents the situation when the reflected ray positively interferes with – i.e. reinforces – the direct ray. Thirdly, the field strength at the earth's surface is zero, seen by placing the receiving antenna height to zero in (4.2). That is the case of complete negative interference between the direct and reflected rays – i.e. cancellation.

Note that at very long wavelengths $\phi \approx 0$ in (4.2) so that $|E| \approx 0$. This is because the phase difference resulting from the difference in path lengths of the direct and

reflected rays is negligible, leaving the phase reversal on ground reflection as the only difference in phase.

If we assume the direct path length is approximately the distance between the transmitter and receiver, we can substitute (1.7) into (4.1) to give a more complete expression for the field strength

$$|E| = 2 \frac{\sqrt{30P_t G_t}}{d} \sin \frac{2\pi h_r h_t}{\lambda d} \tag{4.3}$$

which at large distances from the transmitter – i.e. $d \gg h_t, h_r$ – reduces to

$$|E| = 4\pi \sqrt{30P_t G_t} \frac{h_r h_t}{\lambda d^2} \tag{4.4}$$

showing that at large ranges the space wave field strength falls away twice as quickly as that for free space propagation.

Figure 4.3 shows a typical space wave field strength curve that exhibits all of the properties just discussed. Note that the nulls in the figure seem not to go to zero. That is just the result of the samples of range chosen for the computation. In principle they do go to zero.

The theory and example above have been based on an ideal reflection at the earth’s surface, with the assumption that the reflection coefficient is -1 . A real earth will not necessarily behave that way, and the reflection coefficient is likely to lead to a reduction in the amplitude of the reflected ray and a phase change less than 180° .

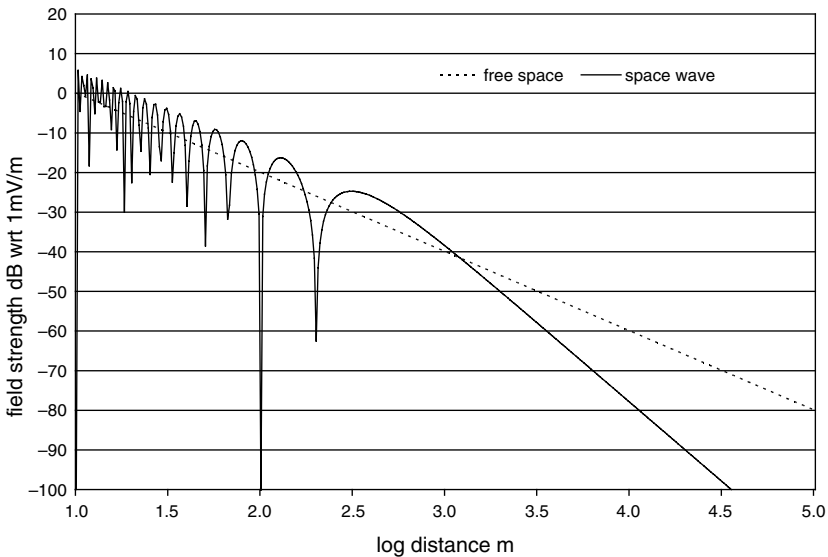


Fig. 4.3 A typical field strength curve for space wave propagation, based on $h_t = 10\text{m}$, $h_r = 2\text{m}$, $\lambda = 0.2\text{m}$, $P_t = 2\text{W}$, $G_t = 1.67 \equiv 2.23\text{dB}$ i; the deep minima theoretically go to zero ($-\infty\text{dB}$)

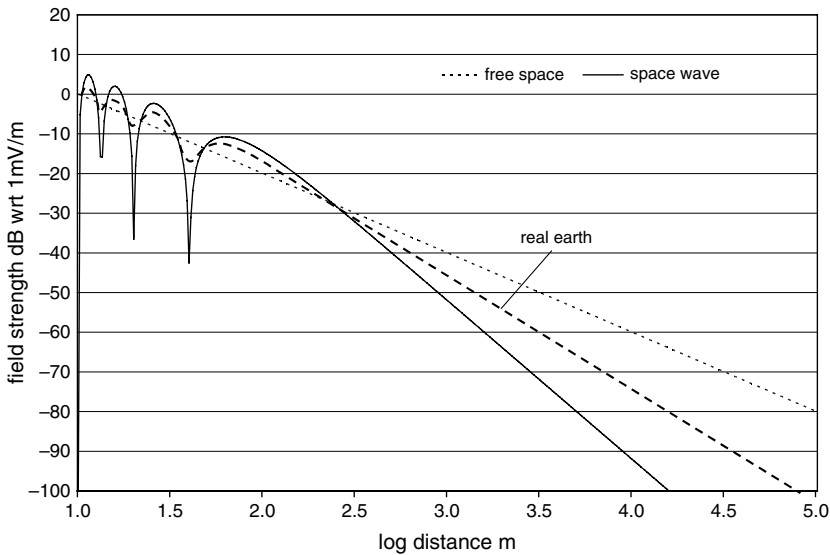


Fig. 4.4 Space wave field strength with ideal and real earths based on $h_t = 10\text{m}$, $h_r = 2\text{m}$, $\lambda = 1\text{m}$, $P_t = 2\text{W}$, $G_t = 1.67 \equiv 2.23\text{dBi}$

As a consequence the interference between the direct and ground reflected rays at the receiving antenna will not lead to complete reinforcement and cancellation, giving a received field strength curve more like that depicted in Fig. 4.4. The nulls are partially filled in, the maxima are less than twice their free space value and the fall off at large ranges is somewhere between inverse distance (theoretical free space) and inverse square (theoretical space wave).

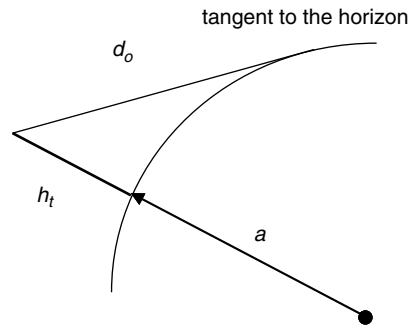
4.2 Effect of Earth Curvature on Space Wave Propagation

Because it is a line of sight mechanism the space wave is clearly affected by the curvature of the earth; principally its range is limited unless the signal is artificially carried beyond the horizon by means of terrestrial repeaters or through the use of communications satellites. We will consider both of those situations in Chap. 6, but for now we need to concentrate on the primary earth curvature problem.

The most obvious limitation is to constrain the range, as illustrated in Fig. 4.5. Assuming that the maximum range is determined by the tangent to the earth, as indicated, then the maximum range will be

$$d_o = \sqrt{(a + h_t)^2 - a^2} = \sqrt{2ah_t + h_t^2} \approx \sqrt{2ah_t} \quad (4.5)$$

Fig. 4.5 Limiting geometry for the space wave



where $a = 6370$ km is the radius of the earth. If there is a receiving antenna of height h_r located beyond the horizon, and it can just see the transmitter across the tangent to the horizon, then the total range will be

$$d_o \approx \sqrt{2ah_t} + \sqrt{2ah_r}$$

The presence of the atmosphere acts to refract the ray towards the ground. We will see later that the effect of that is to increase the range by about 15%. To account for refraction an effective earth radius can be used in (4.5), with value

$$a' = \frac{4}{3}a \approx 8500 \text{ km} \tag{4.6}$$

There are some interesting effects of earth curvature on the received signal strength that largely cancel, as illustrated in Fig. 4.6. First, by drawing a tangent plane to the earth we can see that one impact is to reduce the effective antenna heights, as noted in Fig. 4.6a. From (4.4) that suggests a loss of signal strength at large ranges. However, the curvature of the earth also causes the reflected ray to be diverging as seen in Fig. 4.6b. Thus the level of interference with the direct

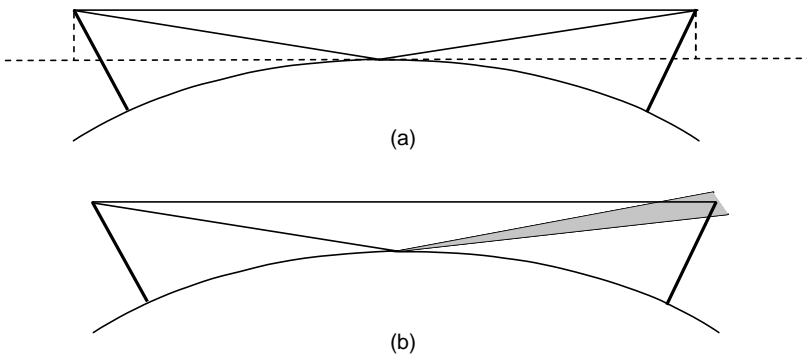


Fig. 4.6 Effect of earth curvature on the received signal strength (a) showing reduced effective antenna heights and (b) divergence of the ground reflected ray

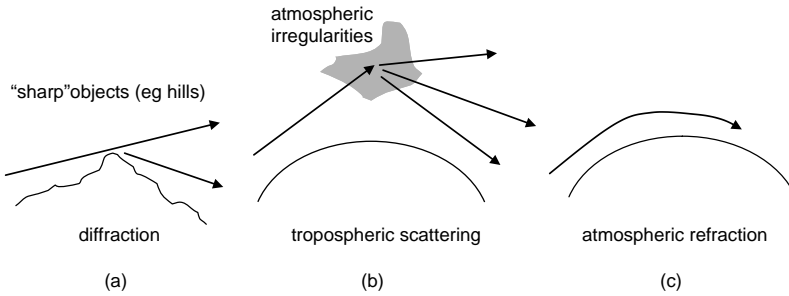


Fig. 4.7 Mechanisms for getting energy into a shadow zone, including around the curvature of the earth

ray will be reduced, leading to improved signal strengths at large distances. Since these two observations lead to opposite impacts they are generally assumed to offset each other.

Even though, from a practical point of view, one of the more significant problems with earth curvature is the likelihood of having the receiver in a *shadow zone* and thus, in principle, unable to receive any signal, there are several mechanisms by which energy can get into a shadow region. These apply even if the shadow is caused by obstacles such as hills. Figure 4.7 illustrates three common mechanisms. One is diffraction, in which the presence of sharp protrusions into the direct beam cause a deflection of some of the energy around the protrusion as seen in Fig. 4.7a. Another is scattering from refractive index irregularities in the high atmosphere (the troposphere); with sufficiently directive antennas and high transmitter powers it is possible to have energy scattered into a shadow zone or, more particularly, around the curvature of the earth as depicted in Fig. 4.7b. This mechanism is known as *troposcattering*. The third mechanism is atmospheric refraction as illustrated in Fig. 4.7c. We will now have a look at diffraction and refraction in more detail.

4.3 Diffraction

In preparing for this discussion it is important to recognise that when we draw rays to represent the passage of a radio wave, as in the previous diagrams, we are really just indicating the centroid of the flow of energy from transmitter to receiver. In reality the wave has a wavefront as implied by the outwardly propagating spherical wave of Figs. 1.1 and 1.2. At the position of the receiver we generally regard the wavefront as being plane, and transverse to the propagation direction. When that wavefront encounters an obstacle, such as depicted in Fig. 4.7a, it will be diffracted, a phenomenon we can induce from Huygens' principle. That principle says essentially that every point on a wavefront can be regarded as a isotropic radiator, so that the future behaviour of the wavefront can be synthesised from the interference of the fields from these "secondary" radiators.

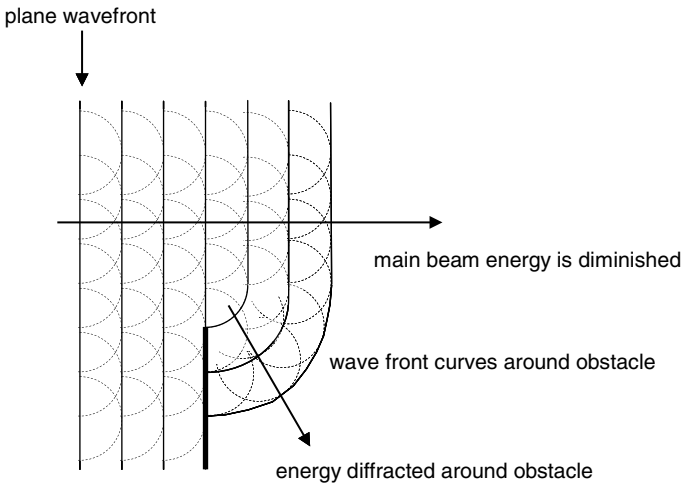


Fig. 4.8 Diffraction of energy from a wavefront into the shadow zone behind an obstacle; the dotted circles indicate the hypothetical initiation of spherical waves from each point on a wavefront, according to Huygens' principle

Figure 4.8 depicts how such a model will cause energy to be diffracted away from the primary direction of propagation by an obstacle in the path of the wave. As expected, energy travelling forward in the intended direction will be diminished by energy being diffracted around the object into the shadow zone.

Although we commenced this discussion in relation to propagating energy into a shadow zone, loss resulting from diffraction is particularly important to take in to consideration in line of sight communication links if an object such as a building, tree or sharp geographical feature is close to the centroid ray of the link. Accounting for that loss was anticipated in (1.15).

Figure 4.9 shows the geometry of an idealised situation involving diffraction (loss) in which a sharp object protrudes into the propagation path at the position shown. The question is, how far should the ray clear the object in order to minimise the loss to the forward propagating signal?

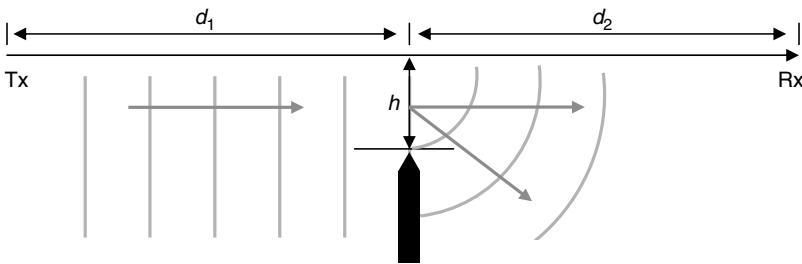


Fig. 4.9 Geometry for computing diffraction loss

The loss of signal in the situation shown in Fig. 4.9 is described by a diffraction gain, which is derived from the geometrical theory of optics. It is expressed as the ratio of the power density in the forward direction after the obstacle to the forward travelling power density prior to the obstacle being encountered. For a knife edge, it can be written¹

$$G(u) = \frac{1}{2} \left[(C(u) + 0.5)^2 + (S(u) + 0.5)^2 \right] \quad (4.7a)$$

in which $C(u)$ and $S(u)$ are the Fresnel cosine and sine integrals and the parameter u is defined by

$$u = \sqrt{2h^2(d_1 + d_2)/\lambda d_1 d_2} \quad (4.7b)$$

Equation (4.7) is plotted in Fig. 4.10 from which it can be seen that if u is greater than about 1.0–1.5 then loss is minimised (gain is approximately 0 dB). If for convenience we choose $u = \sqrt{2}$ then, from (4.7b), to minimise diffraction loss we need

$$h^2(d_1 + d_2)/\lambda d_1 d_2 \geq 1$$

or

$$\lambda d_1 d_2 / (d_1 + d_2) \leq h^2$$

The equality in this last expression defines the equation of an ellipse (or ellipsoid in three dimensions). With the antennas as foci, it is described by the locus of a point that is one half wavelength greater than the distance between the antennas and that just touches the object, as shown in Fig. 4.11. Objects that lie outside that ellipse

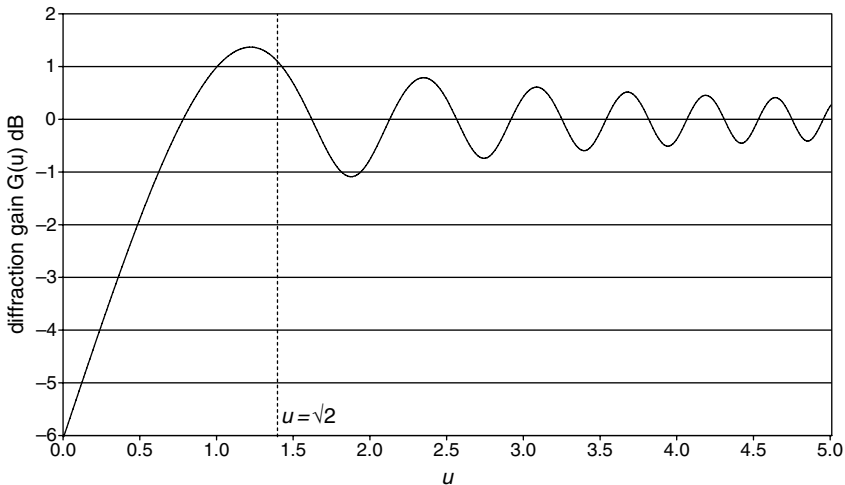


Fig. 4.10 Diffraction gain as a function of the parameter u defined in (4.7)

¹ K.F. Sander, *Microwave Components and Systems*, Addison-Wesley, Wokingham, 1987.

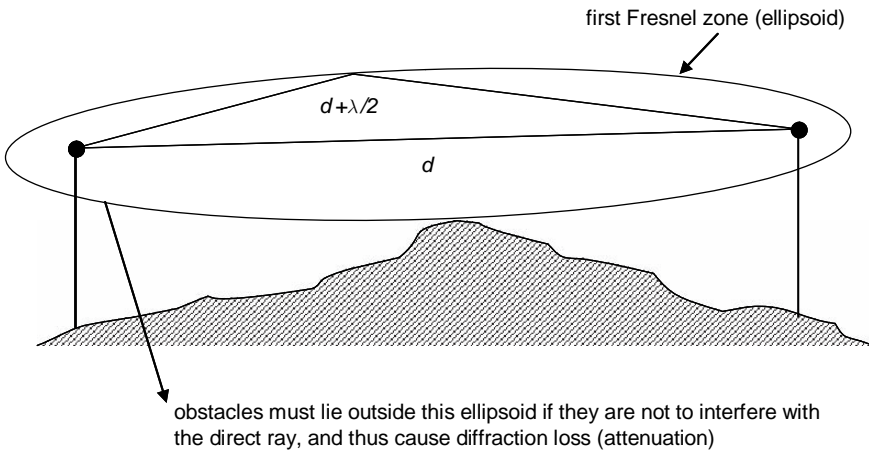


Fig. 4.11 The first Fresnel zone, as an ellipse with the transmitting and receiving antennas as foci

are said to lie outside the first Fresnel zone and thus are said to have first Fresnel zone clearance in our communications system. They are then assumed to have little effect on the forward energy carried by the ray.

4.4 Refraction of the Space Wave

A major consideration in space wave propagation is refraction by the atmosphere. Not only does it lead to improved range, as noted earlier, but extreme cases of refraction can lead to anomalous propagation that sees the wave travelling over very large distances. To understand those effects we have to start with an understanding of the vertical refractive index profile of the atmosphere and how it influences a ray.

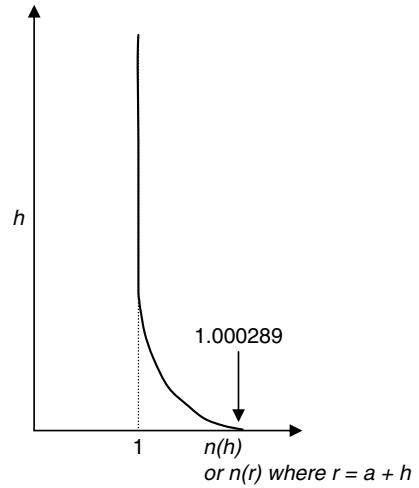
Figure. 4.12 shows a standard atmospheric refractive index profile, described by the equation

$$n(h) = 1 + 289 \times 10^{-6} e^{-0.136 \times 10^{-3} h} \quad (4.8)$$

where h is the height above the earth's surface. It decreases asymptotically to 1 as height increases (1 being the free space value in the absence of an atmosphere) and increases as we move towards the earth's surface as a result of increasing atmospheric density.

By using only the first two terms of the polynomial (Taylor) series expansion of e , (4.8) can be approximated as

Fig. 4.12 Standard profile of atmospheric refractive index with height



$$\begin{aligned}
 n(h) &\approx 1.000289 - 0.136 \times 289 \times 10^{-9}h \\
 &= 1.000289(1 - 39.3 \times 10^{-9}h) \\
 &= n_o(1 - kh)
 \end{aligned}
 \tag{4.9}$$

This is a linear approximation for use near the earth's surface; it shows that at the surface the refractive index of a standard atmosphere is 1.000289, which seems hardly enough to cause meaningful refraction. But, as we shall see, it does.

We will use Snell's Law of Refraction to understand how the wave is refracted by an atmosphere with the atmospheric profile of Fig. 4.12. That requires the variable atmosphere to be striated into a number of thin slices of constant properties, so we can simulate and analyse refraction in a piecewise linear fashion. However, it is important to recognise that the atmosphere is a spherical shell-like medium around the earth. Thus the slices we form must be thin spherical slices as shown in Fig. 4.13. The refractive indices of the slices are represented by n_s , etc, while the heights of the slices are indicated by radius r_s from the earth's centre.

Applying Snell's Law at the first interface in Fig.4.13 gives

$$n_s \sin \theta_s = n_{s+1} \sin \theta'_s$$

whereas applying the sine rule in the triangle formed between the two slices and the earth's centre gives

$$\frac{\sin(180 - \theta'_s)}{r_{s+1}} = \frac{\sin \theta_{s+1}}{r_s}$$

Combining these last two expressions gives

$$n_s r_s \sin \theta_s = n_{s+1} r_{s+1} \sin \theta_{s+1}$$

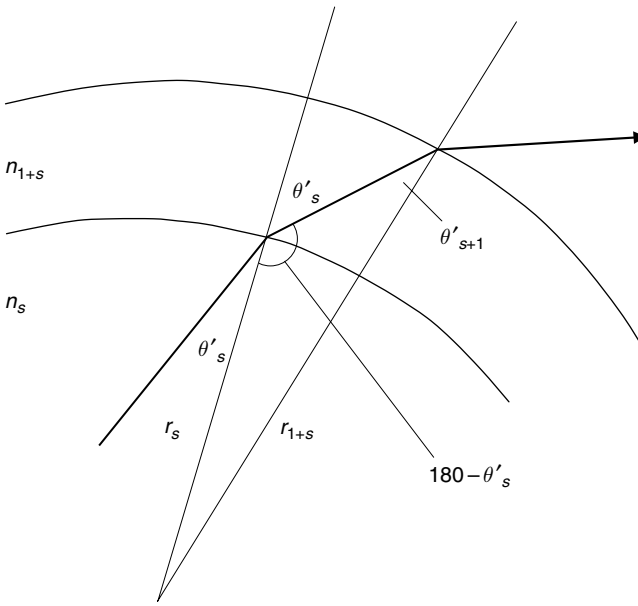


Fig. 4.13 Geometry for applying Snell's Law to understand space wave refraction

This last expression shows that the triple product of refractive index, radius and the sine of the angle of incidence is the same at each interface. Thus we can equate the product at a general radius r with the particular case of the earth's surface with $r = a$:

$$n(r)r \sin \theta = n_o a \sin \theta_o \quad (4.10)$$

where θ_o is the angle with which the space wave is launched with respect to the vertical; $\theta_o = 90^\circ$ means it is radiated horizontally, while $\theta_o = 0^\circ$ means it is radiated straight up. Note that n_o is given in (4.9).

Using the linear approximation to the refractive index profile from (4.9), and noting $r = a + h$, (4.10) becomes

$$n_o(1 - kh)(a + h) \sin \theta = n_o a \sin \theta_o \quad (4.11a)$$

therefore

$$n_o(a + h - kha - kh^2) \sin \theta = n_o a \sin \theta_o \quad (4.11b)$$

The squared term in h is generally small and can be neglected; dividing through by a then gives

$$n_o \left\{ 1 + h \left(\frac{1}{a} - k \right) \right\} \sin \theta = n_o \sin \theta_o \quad (4.12)$$

We now define

$$\frac{1}{a_e} = \frac{1}{a} - k \tag{4.13}$$

which substituted into (4.12) gives

$$n_o \{ a_e + h \} \sin \theta = n_o a_e \sin \theta_o \tag{4.14}$$

Now compare (4.14) with the general expression in (4.10). By equating

$$\begin{aligned} n_o & \text{ with } n(r) \\ a_e + h & \text{ with } r \\ a_e & \text{ with } a \end{aligned}$$

we can interpret (4.14) as describing propagation over a spherical earth, but with a uniform refractive index n_o and an effective earth radius a_e . From (4.13), using the parameters for a standard atmosphere in (4.9), we see

$$a_e = \frac{a}{1 - ak} = \frac{a}{1 - 0.25} = \frac{4}{3}a$$

Thus, the effect of refraction can be accounted for by adjusting the earth radius in this manner. Applying that in the range equation of (4.5) shows

$$d_o = \sqrt{2a_e h_t} = \sqrt{\frac{4}{3}} \sqrt{2ah_t} = 1.155 \sqrt{2ah_t}$$

This implies that the effect of the atmospheric refraction is to increase the range of the space wave by about 15%. This is depicted in Fig. 4.14.

We now return to (4.10) and forego the assumption that the squared term in h , that led us to (4.14), is negligible. Dividing throughout by a and noting $r = a + h$, (4.10) can be written

$$n(h) \left(1 + \frac{h}{a} \right) \sin \theta = n_o \sin \theta_o$$

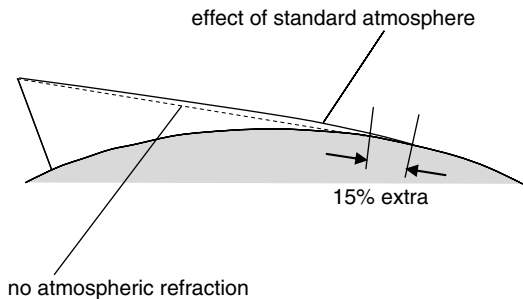


Fig. 4.14 Increased range of the space wave resulting from refraction in a standard atmosphere

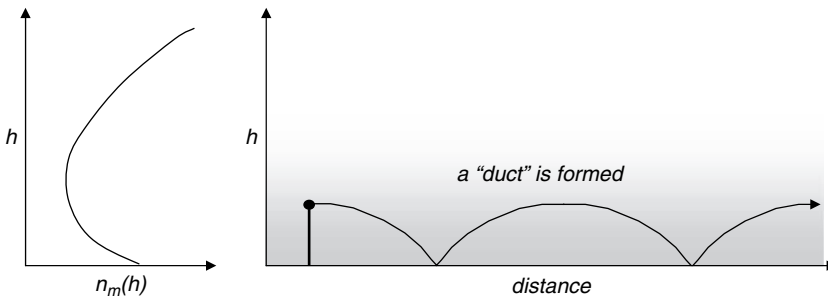


Fig. 4.15 (a) Unusual refractive index profile leading to (b) ducted propagation of the space wave

This can be interpreted as Snell's law describing refraction above a flat earth but with a modified refractive index of

$$n_m(h) = n(h) \left(1 + \frac{h}{a}\right) \approx n_o(1 - kh) \left(1 + \frac{h}{a}\right) \quad (4.15)$$

where $n(h)$ is the actual refractive index variation with height above the earth's surface. The only real unknown in (4.15) is k . While it is a parameter that usually comes from approximating the refractive index for a standard atmosphere near the earth's surface, we can get some idea of how unusual space wave behaviour can occur by taking a little license with its value. For a standard atmosphere its value is 39.3×10^{-9} ; for higher atmospheric moisture contents it can be larger. Suppose, as an illustration, it took the value $k = 1/a = 1.57 \times 10^{-9}$. Then from (4.15)

$$n_m(h) = n_o \left(1 - \frac{h}{a}\right) \left(1 + \frac{h}{a}\right) = n_o \left[1 - \left(\frac{h}{a}\right)^2\right]$$

which, for h small, is a constant, n_o . Thus, under these conditions there is no (or little) atmospheric refraction above an equivalent flat earth.

Usually $n_m(h)$ increases with height, as seen in (4.15). That means there is refraction away from the (flat earth) leading to a range increase around the spherical earth. Under extreme atmospheric conditions the modified refractive index profile, however, can first decrease near the earth's surface and then increase again, as illustrated in Fig. 4.15a. The consequence of that is to give very strong refraction towards the earth's surface as depicted in Fig. 4.15b. Reflection from the surface can then occur, as shown, resulting in refraction back to the surface again, and so on, as illustrated. The wave can then travel very long distances, effectively trapped in a "duct" formed by the surface of the earth and the strong modified refractive index profile. It is of course possible to launch the ray at such an angle upwards that it is not able to be refracted sufficiently back to the earth and is lost; nevertheless, launching the ray at appropriate angles will cause it to be trapped and propagate over anomalously long paths.

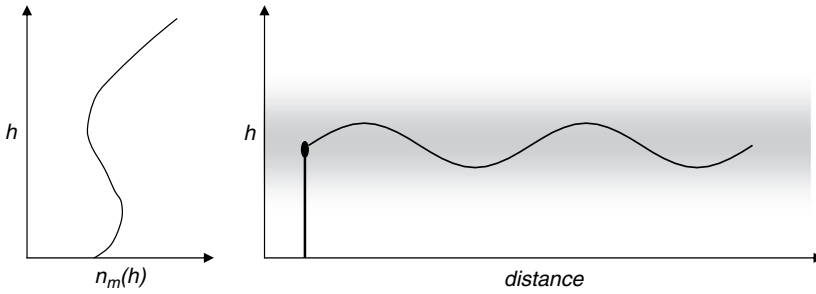


Fig. 4.16 The formation of an elevated duct

Sometime the atmospheric conditions are so strange that an elevated duct is created as illustrated in Fig. 4.16. Provided the ray is launched into the duct it can also be carried exceptionally long distances. Typically ducts of this type occur over the ocean and exist several tens of metres above the surface.

4.5 Effect of Rainfall on Space Wave Propagation

Rainfall can be a problem for space wave propagation for frequencies as low as 4 GHz. Principally, rainfall causes attenuation of the transmitted signal through absorption. Non-spherical raindrops can also cause polarisation rotation during propagation. In a linearly polarised system that leads to additional loss with the

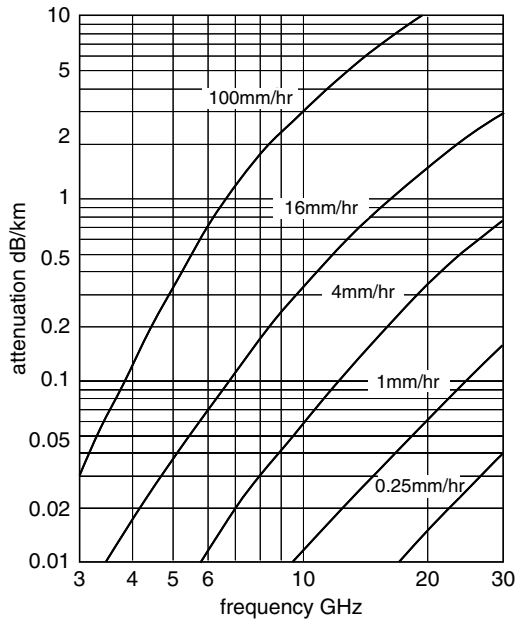


Fig. 4.17 Attenuation of the space wave resulting from rainfall; adapted from Fig. 14.38 of E.V.D. Glazier and H.R.L. Lamont, *Transmission and Propagation*, HMSO, London, 1958

designed polarisation, as well as cross talk, if orthogonal polarisations are used to double system capacity. Figure 4.17 shows a theoretical relationship between attenuation, frequency and rainfall rate.

4.6 Atmospheric Attenuation

The atmosphere does not significantly affect the propagation of the space wave for frequencies lower than about 12 GHz. Above that, however, oxygen and water vapour content cause attenuation, as illustrated in the composite graph of Fig. 4.18. The water vapour contribution to the curve will be dependent on relative humidity; the monotonic background is illustrative of signal loss through scattering of the signal from its nominal ray path by atmospheric constituents. The attenuation

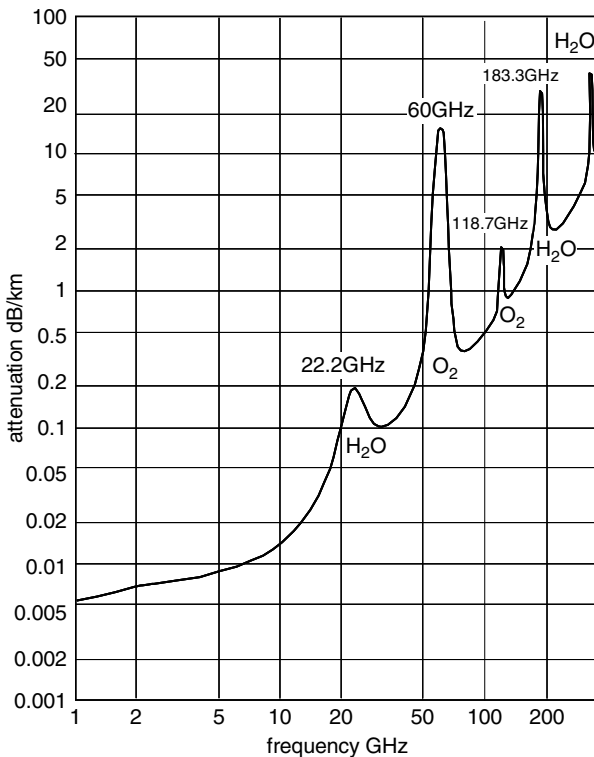


Fig. 4.18 Attenuation of the space wave by atmospheric constituents at a pressure of 1 atmosphere, a temperature of 15 K and a water vapour density of 7.5 gm^{-3} ; based on Fig. 5. of ITU Recommendation ITU-R P.676-7, 2007, which shows separately the contributions of water vapour and oxygen (see www.itu.int/ITU-R)

peaks shown are the result of absorption at specific resonance wavelengths at which the incident radiation excites vibrational and rotational modes in the constituent molecules.

Problems

4.1. The received field strength of the space wave can vary considerably with conditions along the propagation path. In particular the free space wavelength can change with variations in atmospheric refractive index since the refractive index changes along the path with temperature and time of day. As a result, even with a carefully designed communications link, with parameters chosen such that the received field strength is maximised, one can find that the system parameters change, causing the received signal to fall into one of the nulls associated with destructive interference of the direct and ground reflected rays. Discuss how a practical system might be designed that will cope with the variations to ensure reliable reception of the space wave.

4.2. Demonstrate that the space wave is not a viable mechanism for propagation at 100 kHz.

4.3. Why is the space wave the only reliable mechanism for communications above about 30 MHz? You will need to consider some typical field strengths in answering this question.

4.4. Discuss the method by which you think communications might take place at 15 kHz. Is there any significant space wave? What about a sky wave?

4.5. The emergency services department in a particular country used to use HF as their communications medium for keeping in touch with each of their members during an emergency. In an upgrade they changed over to a UHF system operating at 400 MHz. To their concern, however, they discovered they could not communicate easily with colleagues in a small boat about 300 m offshore. Outline what is happening here, and why the HF system seemed more reliable at shorter and longer distances, including over water.

4.6. Suppose you are in a moving vehicle listening to FM radio. The signal breaks in and out. What might be happening?

4.7. A particular mobile service operates at 3 GHz. The antenna on the vehicle can be assumed to be 2 m above the ground. The base station transmitting to the vehicle is about 10 m above the ground. Assume isotropic transmitting and receiving antennas. If the vehicle is travelling at 60 km/hr describe how the signal at the terminals of the receiving antenna changes with time if the vehicle travels directly towards the transmitter over a distance of 500 m to 50 m. Repeat the exercise if the vehicle is travelling at 45° to the direct line to the transmitter, again over the same change of range.

- 4.8.** The derivation in Sect. 4.1 has been based on a single ground reflected ray interfering with the direct ray. Repeat the exercise but with two reflected rays, one of which might come from a nearby building. Assume the reflection coefficients for both reflected rays are -1 .
- 4.9.** A particular air traffic control radar installation operates at 1 GHz with an antenna that provides a beam that is narrow in the horizontal (azimuthal) direction and broad vertically – in other words it looks like a vertical fan. The lower extremity of the beam reaches the ground. Suppose it is located near the coast. Can you foresee any aircraft altitudes that might make the aircraft difficult to detect with the radar?
- 4.10.** Revisit the situation of Problem 4.9 and show that the blind spots identified can actually be characterised in terms of a modification of the vertical pattern of the antenna.
- 4.11.** An instrument carried on a satellite detects atmospheric water vapour content by measuring radiation propagating upwards to the satellite at 22.2 GHz and 183.3 GHz. Discuss the basis of this technique, noting that thermal equilibrium requires an absorbing medium also to be a radiator.
- 4.12.** How can VHF television be received in a shadow zone, such as behind a mountain?
- 4.13.** A UHF communications link over water uses parallel polarisation. If the ray reflected from the water undergoes glancing incidence describe the time behaviour of the received field if the water level varies with tide. When the reflected wave is incident onto the water at 83.7° the signal reaching the receiver consists just of the direct ray. Why? Assume the dielectric constant of the water is 81 and note (7.12).

Chapter 5

Noise

At first sight it might appear strange including a coverage of noise in a book on radio wave propagation. It is, however, the level of noise in a communications system that determines the power density and field strength needed at the receiver and, consequently, the power levels that have to be transmitted. In other words, in order to put the propagation material into practice it is necessary to understand noise as the property that relates circuit quantities (transmitted and received power, operating frequency, antenna characteristics) to propagation quantities (free space path loss and path characteristics such as atmospheric constituents and rainfall, obstacles in the path, and multi-path behaviour). We will demonstrate that relationship with some system examples in Chap. 6.

5.1 What is Noise?

Noise can come from several sources and often is random in nature; that is the type of noise of most interest to us in this introductory treatment. Whenever temperature is above absolute zero, noise will be present in any physical system.

In general we talk of noise as any “signal” or disturbance that is unwanted. Sometimes that includes cross-talk – a signal intended for another receiver that we happen to receive in full or part, and which interferes with the reception of the signal in which we are interested. In this coverage we are not so much interested in cross-talk as in the random noise that can interfere with our ability to receive an intelligible message.

Usually noise adds to the transmitted signal and we receive both together.¹ In this combination it is the *signal to noise ratio* in which we are interested and which

¹ Noise can also occur multiplicatively, which means its magnitude goes up and down with the signal level.

determines the performance of a telecommunications system. Signal to noise ratio (SNR) is usually defined by the ratio of the root mean square power of the signal and the root mean square power of the noise. Most frequently it is expressed in dB.

5.2 Sources of Noise

Both active and passive devices generate noise. In the case of a piece of conductor, or a circuit component such as a resistor, electron motion through the crystal lattice gives rise to what is called *thermal* or *Johnson* noise. For active devices such as transistors, carrier collisions with crystal lattice sites generate *shot* noise which, as the name suggests, is more impulsive in nature than the random time variation of thermal noise.

A particularly important source is environmental noise, which is the radiation given off by a black body at non-zero temperature; it is described by Planck's law of radiation. Both the sun and the earth are strong emitters of radiation. Figure 5.1 shows their outputs in the optical wavelength range; however their tails extend out to the radio part of the spectrum. The atmosphere itself is also a weak noise emitter because of radiation from its constituent molecules.

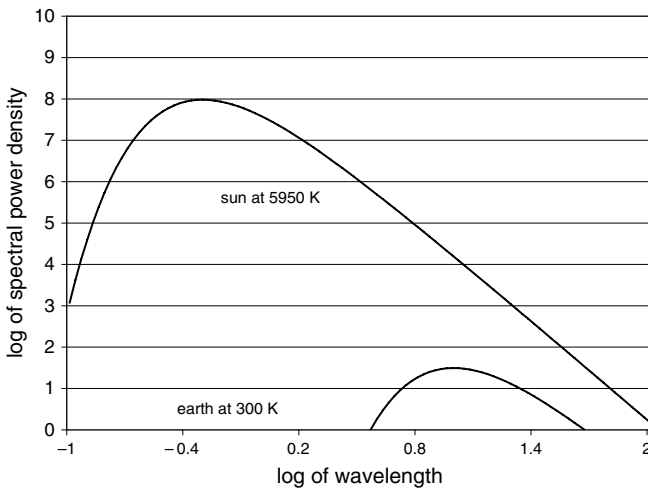


Fig. 5.1 Levels of radiation from the sun and earth acting as black body radiators according to Planck's law; when intercepted by an antenna this radiation constitutes noise, although note that the solar curve has to be reduced to account for the earth-sun distance

5.3 The Concept of Noise Temperature

To a very good approximation the available noise power from a thermal source, such as a conductor or resistor, at temperature T_o over a bandwidth B Hz is given by²

$$N_a = kT_oB \quad \text{W} \quad (5.1)$$

where k is Boltzmann's constant ($1.38 \times 10^{-23} \text{ JK}^{-1}$). If we could measure the noise power coming from the source N_a then we could infer its temperature from (5.1). Although this result relates strictly only to thermal sources it is applied more generally to all sources of noise as though they were thermal in nature. The temperature then inferred may not be the strict *physical* temperature T_o , but instead is called the *noise temperature* T_s defined as:

$$T_s = \frac{kN_a}{B} \quad \text{K} \quad (5.2)$$

5.4 The Noise Temperature of a Two Port

Devices such as amplifiers, attenuators and filters are two ports, in that they have an input and an output. Lengths of cable, waveguides and optical fibres are also two ports for the same reason. As a signal travels through a two port, noise is added to it because of the noise sources within the two port. Those sources would be the collection of active and passive devices in the case of an amplifier and the conducting material itself in the case of a coaxial cable.

Consider a matched two port driven from a noiseless source, as shown in Fig. 5.2a. The fact that the source is noiseless is indicated by the zero noise temperature ascribed to it. Generally the two port will be characterised by a bandwidth (called here the noise bandwidth B_n), a gain (which would be less than one in the case of a piece of cable) and by the amount of noise power it adds to the signal in transmission though the two port. We describe the latter by the noise power we can measure at the output of the two port, in the absence of source noise power.

It will be convenient in the following analysis to refer the noise measured at the output back to the input of the two port (N_i) as though it had come from an external source, as shown in Fig. 5.2b, and the two port itself were noiseless. We now assume that that noise component can be modelled as an equivalent thermal noise source of temperature T_e so that

$$N_i = \frac{N_a}{g} = kT_eB_n$$

² This seems a strange result since it depends only on the temperature of the object and the bandwidth over which the noise is observed. It comes from a statistical mechanics analysis of the random motion of electrons in a conductor.

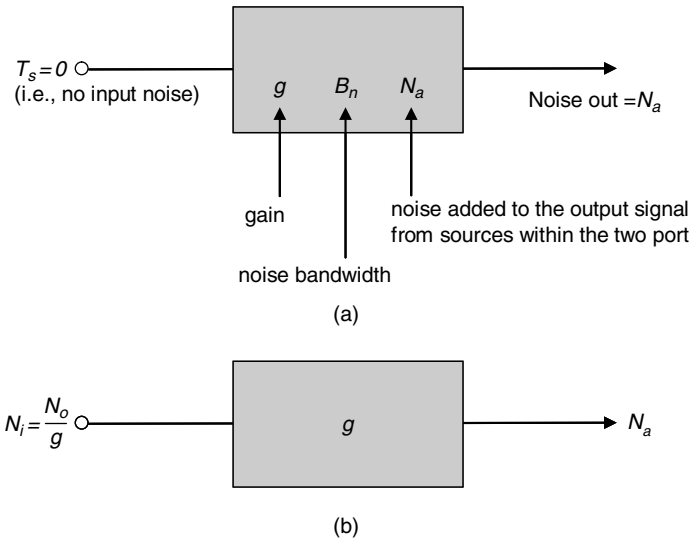


Fig. 5.2 (a) Two port being driven from a noiseless source and (b) referring the two port noise back to an equivalent input

giving as the *equivalent input noise temperature* of the two port

$$T_e = \frac{N_a}{gk B_n} \tag{5.3}$$

The input noise temperature is a property of the two port. It can be added to the actual source noise temperature to allow the noise at the output to be computed simply, consisting of the component that came from the source and that generated in the two port itself. Thus, using the nomenclature of Fig. 5.3

$$\begin{aligned} N_{ao} &= N_a + gkT_s B_n \\ &= gkT_e B_n + gkT_s B_n \\ &= gk(T_e + T_s) B_n \end{aligned}$$

Although it is not often used we could refer the noise of a two port to its output rather than its input and define an equivalent output noise temperature which recognises that the equivalent input noise has been enhanced by the gain of the two port.

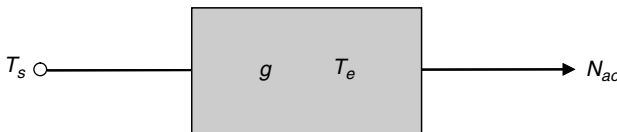


Fig. 5.3 Noisy two port driven from a noisy source

From (5.3) the equivalent output noise temperature is

$$T_{eo} = g \frac{N_a}{gkB_n} = \frac{N_a}{kB_n}$$

5.5 Noise Figure

Noise temperature is a useful concept when the equivalent noise temperature is lower than the ambient temperature. An equivalent measure of the noise added to a signal in transmission through a two port, which is more usable when the noise is high, is the noise figure F , defined by

$$F = \frac{\text{actual output noise power}}{\text{output noise power if the two port were noiseless}}$$

This is based on the assumption that the two port is impedance matched and fed by a thermal noise source at temperature T_o , taken in practice to be 290 K.

Using the nomenclature in Fig. 5.3, but with $T_s = T_o$, the definition of noise figure can be written as

$$F = \frac{N_{ao}}{gkT_oB_n} = \frac{gk(T_e + T_o)B_n}{gkT_oB_n}$$

Thus

$$F = 1 + \frac{T_e}{T_o} \quad (5.4)$$

5.6 Relationship Between Noise Figure and Output Signal to Noise Ratio

Consider the typical transmission of a signal through a two port such that noise enters the two port from the signal source and is also added to the signal by the two port itself. Let the noise from the source³ be characterised by the source temperature T_s , and that added by the two port be described by the equivalent input noise temperature T_e , as shown in Fig. 5.4. If the input signal power is P_i , then the output signal and noise levels respectively are

$$\begin{aligned} P_o &= gP_i \\ N_{ao} &= gk(T_s + T_e)B_n \end{aligned} \quad (5.5)$$

³ Strictly only that component of any source noise that passes through the noise bandwidth of the two port.

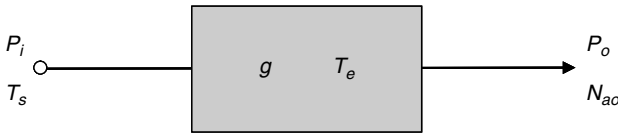


Fig. 5.4 Signal and noise in a two port

Therefore the input and output signal to noise ratios are

$$\begin{aligned} SNR_i &= \frac{P_i}{kT_s B_n} \\ SNR_o &= \frac{gP_i}{gk(T_s + T_e)B_n} = \frac{P_i}{k(T_s + T_e)B_n} \end{aligned} \quad (5.6)$$

As we will see (5.6) is important in its own right but here it allows us to build a bridge between noise figure and signal to noise ratio. The ratio of the output and input signal to noise ratios is

$$\frac{SNR_o}{SNR_i} = \frac{T_s}{T_s + T_e} = \frac{1}{1 + T_e/T_s} \quad (5.7)$$

If, in (5.7), we assume that the noise source is thermal, so that $T_s = T_o$, that equation can be written

$$\frac{SNR_o}{SNR_i} = \frac{1}{1 + T_e/T_o} \equiv \frac{1}{F} \quad (5.8)$$

where F is the noise figure of the two port. Note that (5.8) can be expressed in the decibel form

$$SNR_o = SNR_i - F \quad \text{dB} \quad (5.9)$$

which shows explicitly that the noise figure of a two port acts to degrade the signal to noise ratio of a signal in passage though the two port (on the assumption it is driven from a thermal noise source).

5.7 The Noise Properties of a Passive Two Port

We now derive quite a remarkable result for passive two ports such as attenuators, filters and lengths of cable. To do so we assume that the two port is impedance matched and that the input noise is described by the ambient temperature T_o – in other words it is fed from a matched, thermal source. The two representations in Fig. 5.5 are equivalent: Fig. 5.5a shows the two port as noisy whereas Fig. 5.5b

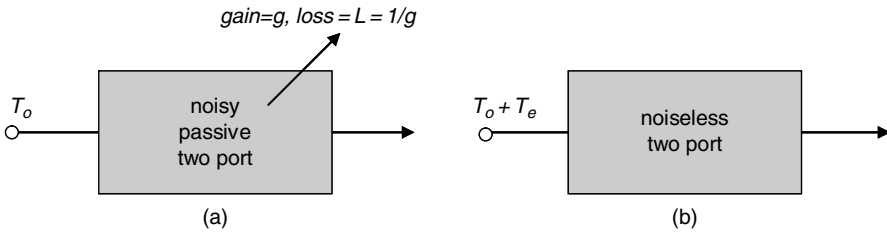


Fig. 5.5 Representation of a real passive two port by its noiseless equivalent

shows the noise to have been translated to an equivalent input noise, with the two port regarded as noiseless.

If all the components in the two port are considered to be at the physical temperature T_o , which is reasonable for a passive device, then the output noise for Fig. 5.5a can be expressed, using (5.1), as

$$N_{ao} = kT_o B_n \quad (5.10)$$

Considering Fig. 5.5b we also see that the noise output power of the two port can be expressed by (5.5). Thus equating (5.5) and (5.10) we see

$$kT_o B_n = gk(T_o + T_e)B_n = \frac{k(T_o + T_e)B_n}{L}$$

where $L = 1/g$ is the “loss” of the two port. From this last result we see that the equivalent input noise temperature of the two port is given by the simple expression

$$T_e = (L - 1)T_o \quad (5.11)$$

and, in view of (5.8), its noise figure is given by

$$F = L \quad (5.12)$$

The last is a remarkable expression. It says that the noise figure of a lossy two port is numerically equal to its attenuation loss. Thus a 10 dB ($L = 10$) attenuator will have $F = 10$, $T_e = 9T_o$ and will degrade signal to noise ratio by 10 dB. Also a 10 m length of coaxial cable with a loss of 0.05 dB/m will have a noise figure of 0.5 dB (loss of 1.12) and a noise temperature of $0.12 \times T_o = 34.8$ K for $T_o = 290$ K.

5.8 Cascaded Two Ports: Friis' Noise Formula

Most communications systems consist of a cascade of active and passive two ports, each with their own noise properties. It is useful, particularly for system design purposes, to develop an expression for the composite input noise temperature and

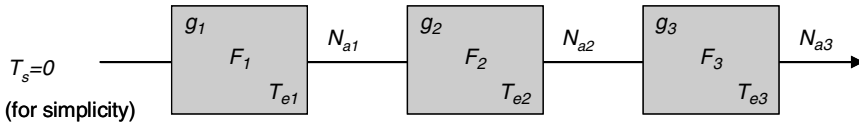


Fig. 5.6 A cascade of two ports

noise figure. Consider the cascaded set of two ports shown in Fig. 5.6. We assume for the moment that there is no source noise, thus $T_s = 0$.

The noise output powers for the first and second stages are

$$\begin{aligned} N_{a1} &= g_1 k T_{e1} B_n \\ N_{a2} &= g_2 g_1 k T_{e1} B_n + g_2 k T_{e2} B_n \\ &= k(g_1 g_2 T_{e1} + g_2 T_{e2}) B_n \end{aligned}$$

The bracketed term in the last expression is the equivalent output noise temperature of the first two stages. To find the equivalent input noise temperature we divide by the gain to that point, thus

$$T_e = \frac{g_1 g_2 T_{e1} + g_2 T_{e2}}{g_1 g_2} = T_{e1} + \frac{T_{e2}}{g_1}$$

Continuing this analysis over several two ports we find

$$T_e = T_{e1} + \frac{T_{e2}}{g_1} + \frac{T_{e3}}{g_1 g_2} + \dots \quad (5.13)$$

This is known as Friis' formula.⁴ It is a very telling result. It emphasises that the first stage is the most important in determining the overall noise properties of the system (i.e. the cascade of two ports). To keep the equivalent input noise temperature of the system as low as possible we need to have a low noise first stage with a high gain; the gain is important since it features as a divisor in the second and all subsequent terms in (5.13). If, as can often be the case, the first stage is a length of lossy transmission line then (5.13) shows that the noise properties of the subsequent stages will be enhanced because the gain of the first stage will be less than unity.

We can recast (5.13) in terms of noise figure:

$$F = F_1 + \frac{F_2 - 1}{g_1} + \frac{F_3 - 1}{g_1 g_2} + \dots \quad (5.14)$$

It is instructive to consider a simple application of (5.13) based on the example of Fig. 5.7. Note that there are two options for the connection from the antenna. One places the low noise amplifier (LNA) after the waveguide from the antenna, and the other puts the LNA right at the antenna terminals, with the waveguide carrying the signal to the travelling wave tube (TWT). Note that the antenna acts as the signal

⁴ Not to be confused with Friis' radiation formula in (1.12).

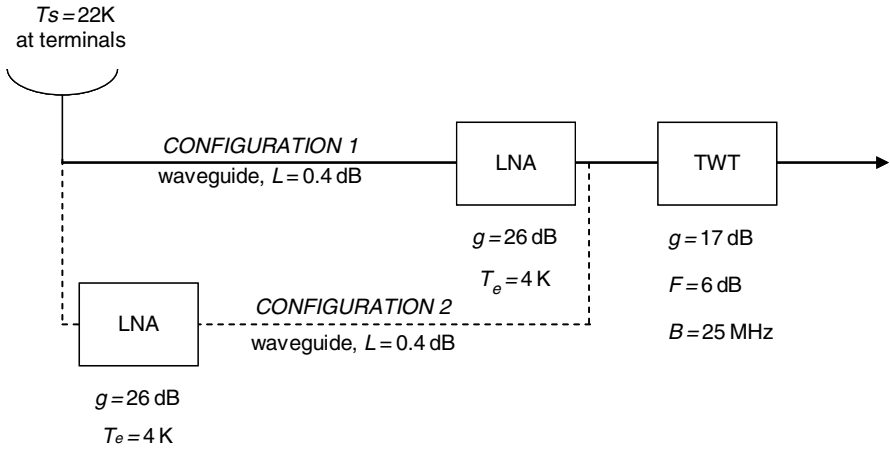


Fig. 5.7 Three cascaded two ports (cable, low noise amplifier LNA and travelling wave tube TWT), in two configurations

(and noise) source and it does not explicitly enter in to the calculations to follow. We assume that the received power at the antenna terminals is $P_s = 2\text{ pW}$, and we need to compute the output signal to noise ratio of the cascade.

Consider Configuration 1. The equivalent input noise temperature at the input to the waveguide from the antenna is

$$T_e = T_{ewg} + \frac{T_{eLNA}}{g_{wg}} + \frac{T_{eTWT}}{g_{wg}g_{LNA}} + \dots$$

Now

$$T_{ewg} = T_o(L - 1) = 290(1.1 - 1) = 29\text{K} \text{ since } L = \frac{1}{g} = 0.4\text{dB} \equiv 1.1$$

$$T_{eTWT} = T_o(F - 1) = 290 \times 3 = 870\text{K} \text{ since } F = 6\text{dB} \equiv 4$$

$$g_{wg} = \frac{1}{1.1} \quad g_{LNA} = 26\text{dB} \equiv 398 \quad g_{TWT} = 17\text{dB} \equiv 50$$

Thus

$$T_e = 29 + \frac{4}{1/1.1} + \frac{870}{1/1.1 \times 398} = 36\text{K}$$

so that from (5.6) the output signal to noise ratio is

$$SNR_o = \frac{P_i}{k(T_s + T_e)B_n} = \frac{2 \times 10^{-12}}{1.37 \times 10^{-23} \times (22 + 36) \times 25 \times 10^6} = 100 \equiv 20\text{dB}$$

If we now look at Configuration 2, then

$$T_e = T_{eLNA} + \frac{T_{ewg}}{g_{LNA}} + \frac{T_{eTWT}}{g_{wg}g_{LNA}} = 4 + \frac{29}{398} + \frac{870}{1/1.1 \times 398} = 6.5\text{K}$$

so that

$$SNR_o = \frac{P_i}{k(T_s + T_e)B_n} = \frac{2 \times 10^{-12}}{1.37 \times 10^{-23} \times (22 + 6.5) \times 25 \times 10^6} = 200 \equiv 23\text{dB}$$

Thus we see that by putting the first stage of amplification at the antenna terminals we improve the signal to noise ratio at the output by 3 dB for this example. In general, it is better to avoid a passive two port, such as a cable or waveguide, as the first stage since it degrades the overall noise performance. It is preferable, instead, to have amplification at the antenna terminals for low noise performance. Sometimes such an amplifier is called a (mast) head amplifier.

Problems

5.1. By reference to Fig. 5.1 discuss how the temperature of the ocean's surface can be determined by measuring upwelling microwave radiation.

5.2. Using (5.3) determine the equivalent output noise temperature of a two port. Is that a less useful concept than equivalent input noise temperature? How is noise figure related to equivalent output noise temperature?

5.3. A particular communications system has 2 identical receiver stages connected in cascade, each with gain 7 dB and noise figure 4 dB. Compute the noise figure of the combination assuming the stages are impedance matched. Repeat for 3 identical stages with the same properties. What about the hypothetical case of a large number of stages? What does that imply in terms of the properties of the first stage?

5.4. Represent a real noisy resistor by a Thévenin equivalent circuit of a noiseless resistor and an equivalent noise voltage source. From (5.1) we know the available noise power from the noisy resistor; that is the power that can be delivered to a matched load. For simplicity assume the matched load is noiseless and thus demonstrate that the mean square open circuit noise voltage generated by the resistor is $4kTBR$ volts. What is the mean square noise voltage across the matched load?

5.5. An amplifier is matched at input and output at 50Ω . There is no input signal, but the equivalent source impedance is assumed to be at 290 K. If the amplifier has a bandwidth of 15 kHz, a gain of 35 dB and the measured noise voltage delivered across the 50Ω load is $125 \mu\text{V}$, what is the noise figure of the amplifier?

5.6. It is easier to read under bright light than under poorly lit conditions. By making any plausible assumptions describe why.

5.7. A trunk coaxial cable communications channel consists of long lengths of cable connected in series by repeater amplifiers. Assume all cables have a loss of L , and the amplifiers correspondingly have a gain of $g = 1/L$ to compensate for the cable loss. Each amplifier has a noise figure F . Determine an expression for the total noise figure of the system, assuming the first component is one of the cable lengths and, for simplicity $L, F \gg 1$. Is there a simple relationship between system noise figure and the number of repeaters?

5.8. An ideal capacitor which, by definition, will be noiseless since it has no mobile charge carriers is connected in parallel with a real (noisy) resistor. Determine an expression for the mean square noise voltage developed across the capacitor.

Chapter 6

Examples of Microwave Systems

This chapter presents examples of radiated communications services that depend on the material developed in earlier chapters. Since many modern services use VHF and higher, the emphasis is on space wave – i.e. line-of-sight – services.

6.1 The Design of Open Microwave Repeater Systems

Although now only a relatively small part of the set of radio wave communications systems encountered in practice, it is instructive to consider how a broadband radiated system operating at microwave frequencies might be designed. While useful in itself, it will also help us understand some of the propagation and signal aspects of satellite communications, and of wide band wireless applications more generally.

Carriers at VHF and above are necessary to support the bandwidths required for high speed digital and video applications. Typically, bandwidths of 10 MHz and above would demand carriers of about 1 GHz so that several services could be carried simultaneously. At those frequencies unguided propagation will be line-of-sight space wave. Thus long transmission paths will require means by which the signal can be carried around the curvature of the earth. There are three simple means for doing that. First, we can use terrestrial repeaters, which is largely the focus of this material. Generally those repeaters would be sited on mountains or other regions of high relief to maximise range and to clear topographic features as depicted in Fig. 6.1. The alternative is to use coaxial cables or optical fibres for that purpose. While they have the benefit of larger potential bandwidths and relative freedom from interference, they present installation challenges owing to the need to secure land easements for laying the fibres or cables. In contrast, radiated systems have more flexibility.

A second method for overcoming the limitations of earth curvature would be to use the mechanism of troposcatter as depicted in Fig. 4.7. Thirdly, we could use communications satellites; that is the topic of Sect. 6.2.

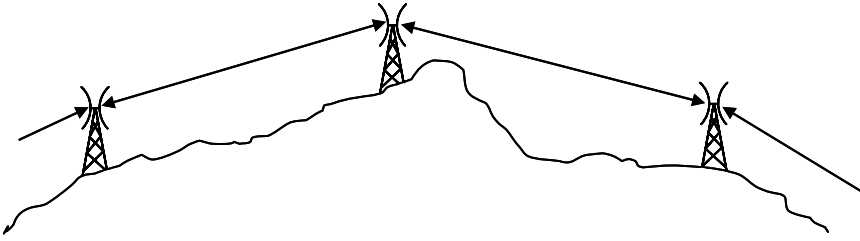


Fig. 6.1 The use of microwave repeaters to clear topographic features and to cope with earth curvature

Consider the simple single link of Fig. 6.2. Essentially, what we want to know is how much power do we need to transmit from one repeater to the next. Based on the example of Fig. 5.7, it is clear that that will depend on the signal to noise ratio we must have at the output of any RF amplifier in the receiver of the next repeater, so that the signal can be successfully demodulated or detected. Figure 6.3 shows a simplified version of the RF stages in the receiver along with some example system parameters.

In designing such a link, we start with the signal to noise ratio we need at the RF output and then work back to the necessary transmitter power from the previous repeater. That requires a knowledge of the free space distance between the transmitter and receiver.

There are several ways for determining the signal to noise ratio required at the output of the RF stage; for illustration assume that the carrier is frequency modulated. In order to capitalise on the signal to noise ratio improvement possible with FM on demodulation, it is necessary to ensure that the signal to noise ratio before demodulation exceeds the so-called FM threshold, as illustrated in Fig. 6.4, otherwise we may even degrade the demodulated SNR. We might assess therefore that a pre-detection signal to noise ratio of 20 dB is needed in order comfortably to exceed the threshold.

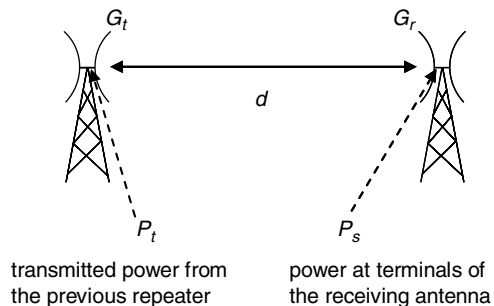


Fig. 6.2 Single link in a microwave repeater system, with symbols describing the transmission from left to right

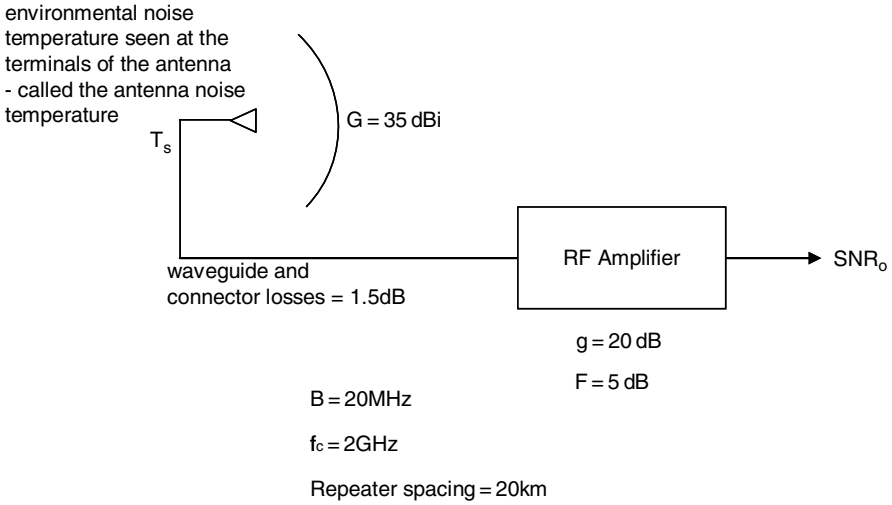


Fig. 6.3 Simplified RF receiver stage, with some typical parameter values to be used to demonstrated how a microwave link design might be carried out

Next we need to consider what fade margins, if any, need to be included in order to ensure that the received signal to noise ratio remains above 25 dB in the face of rainfall, diffraction loss, atmospheric refraction and absorption. Without going into the detail here, but referring to the material provided in Chap. 4, we assume a fade margin of 25 dB is considered satisfactory. We now have to determine the equivalent input noise temperature of the receiver and provided a figure for the antenna noise temperature. We will then be in the position to compute the necessary transmitter power from the previous stage.

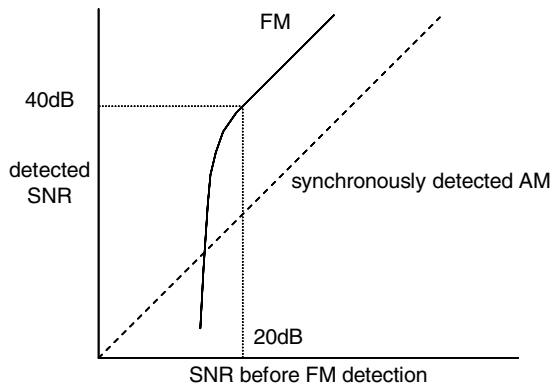


Fig. 6.4 Relationship between the pre- and post-detection signal to noise ratios for an FM system, showing the improvement possible when operating above the FM threshold

Noting

$$T_e = T_{e_{wg}} + \frac{T_{e_{RFamp}}}{g_{wg}}$$

$$T_{e_{wg}} = T_o(L - 1) = 290(1.41 - 1) = 119\text{K}$$

since $L = 1/g_{wg} = 1.5\text{dB} \equiv 1.41$ and $g_{wg} = 0.714$

$$T_{e_{RFamp}} = T_o(F - 1) = 290(3.16 - 1) = 626\text{K} \text{ since } F = 5\text{dB} \equiv 3.16$$

we find

$$T_e = 119 + \frac{626}{0.714} = 995\text{K}.$$

A value for the antenna noise temperature, at its terminals, can generally be found from tables or charts that express the temperature in terms of the operating carrier frequency and the direction in the sky to which the antenna is pointed. It will be highest for antennas pointed horizontally since it will then contain a component of terrestrial noise in addition to atmospheric and galactic noise. A typical value might be 200 K for horizontally pointing antennas and 4 K for those that look vertically upwards, but avoiding the sun. Using $T_s = 200\text{K}$, from (5.6) we find

$$P_s = SNR_o k(T_s + T_e) B_n$$

$$= 31620 \times 1.38 \times 10^{-23} \times (995 + 200) \times 20 \times 10^6 = 10.43\text{nW} \equiv -49.8\text{dBm}$$

Using (1.14) we can now determine the power P_t required for transmission from the previous stage. For a repeater spacing of 20 km, and a carrier frequency of 2 GHz (wavelength of 0.15 m) the free space path loss is 124.5 dB. If the transmit and receive antennas are identical with gains of 35 dBi then the transmitter power required is

$$P_t = 4.7\text{dBm} \equiv 3\text{mW}$$

6.1.1 The Need for Transmission Diversity

The received signal can undergo a time varying fade as the result of several effects. Figure 6.5 suggests, for example, that atmospheric refraction (most likely near sunset) and multi-paths involving moving reflectors, such as a water surface subject to tidal or other height variations, can cause variations in received signal strength.

In an endeavour to offset fading, a technique known as *diversity* is employed. Three types of diversity are possible:

Frequency diversity, in which simultaneous transmission on two different carrier frequencies is employed, on the basis that it would be unlikely that both carriers would fade together and, in the case of multi-path, that the positions of the

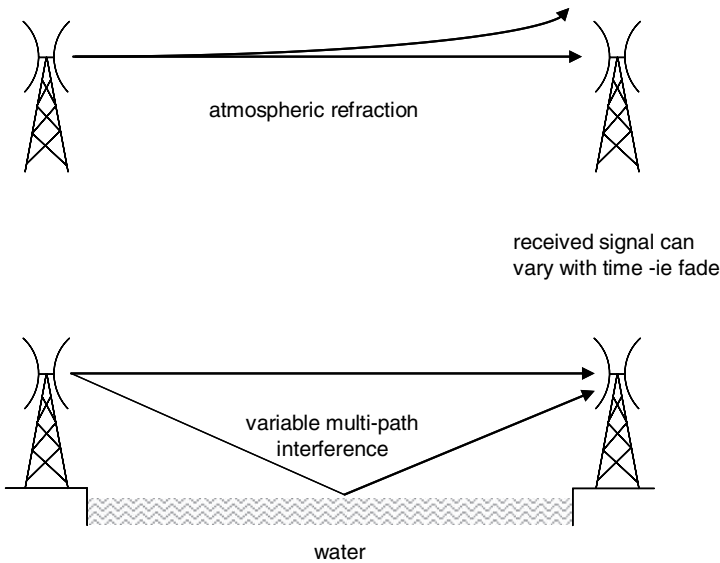


Fig. 6.5 Situations giving rise to fading in the received signal

nulls and maxima seen in Fig. 4.3 would be different for different wavelengths. Clearly, this is wasteful of bandwidth.

Time diversity, in which the signal is repeated; clearly that halves the available data rate.

Space diversity, in which two different receiving antenna heights are used to allow one to have a high likelihood of receiving a strong signal while the other is in a fade, as illustrated in Fig. 6.6. Since that does not affect data rate and requires only additional receiving hardware it is the diversity solution most often adopted in practice.

If fading were the result of simple multi-path variations as depicted in Fig. 6.5 then one way of selecting the height difference between the two antennas in Fig. 6.6 might be to ensure one antenna is in a field strength maximum if the other is in a null. Referring to (4.2) this would require

$$\frac{2\pi(h_r + l)h_t}{\lambda d} - \frac{2\pi h_r h_t}{\lambda d} = \frac{\pi}{2}$$

or $l = \frac{\lambda d}{4h_t} = 7.5\text{m}$ if $d = 20\text{km}$, $h_t = 100\text{m}$ and $\lambda = 0.15\text{m}$.

6.1.2 The Use of Passive Reflectors

When repeater sites are not economically viable for the provision of power, or access is difficult in general, consideration can be given to the use of passive reflectors instead of active repeaters. They usually consist of flat metallic plates that are

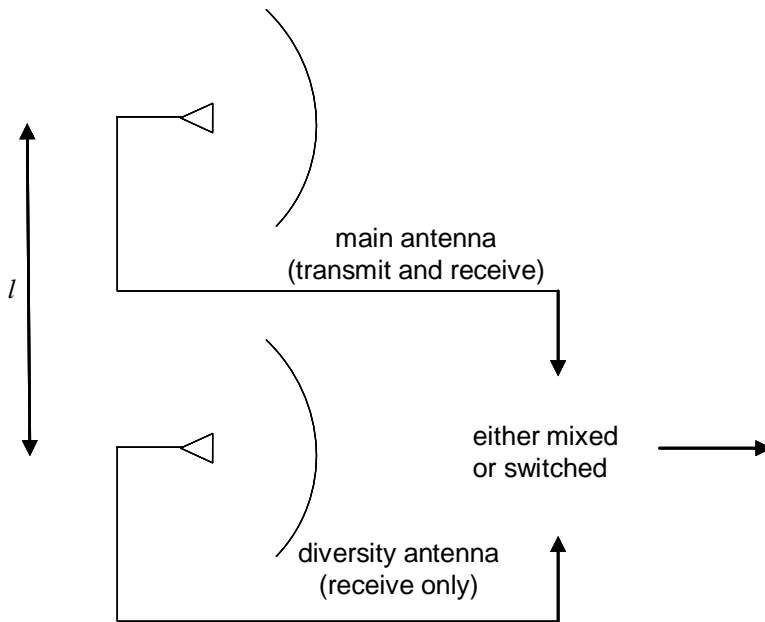


Fig. 6.6 The use of two receiving antennas to provide space diversity

employed to alter the direction of a wavefront, as depicted in Fig. 6.7. Their influence can be incorporated in link power calculations by using their bi-static radar cross section as an equivalent aperture, as seen in the following. The reflected wave can then be treated as though it were isotropic.

The aperture (radar cross section) of a flat conducting plate of dimensions $a \times b$ is given as

$$A(\theta) = \frac{4\pi(ab)^2}{\lambda^2} \cos^2 \theta \text{ m}^2$$

in which θ is the incidence angle measured with respect to the normal to the plate.

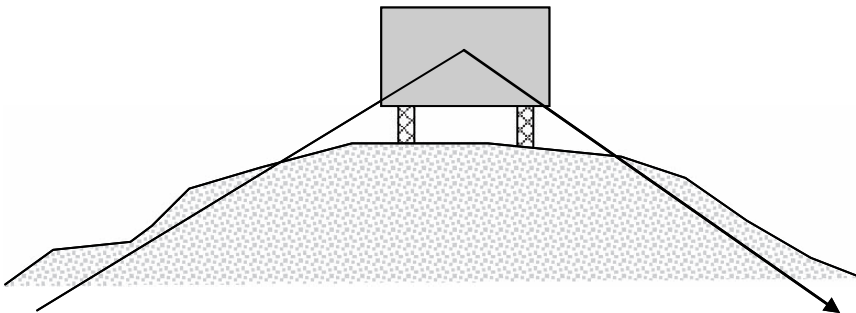
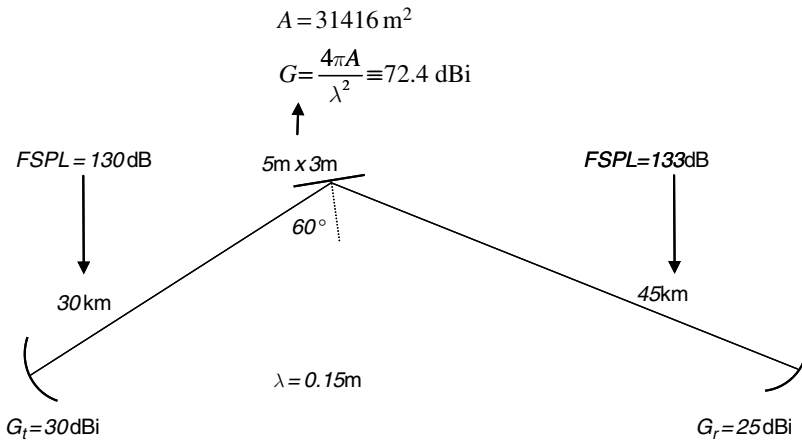


Fig. 6.7 Use of a flat plate passive reflector



$$P_r = P_t + 30\text{dBi} + 25\text{dBi} + 72.4\text{dBi} - 130\text{dB} - 133\text{dB} = P_t - 135.6\text{dB}$$

compare with a FSPL of 136dB for a single hop of same distance, giving $P_r = P_t - 81\text{dB}$

Fig. 6.8 Link path calculations involving passive reflectors

Figure 6.8 shows a typical set of calculations for a link involving a passive reflector. Even though the reflector is very useful for circumventing line of sight obstacles, and is simple and relatively inexpensive to install and operate, it can be seen that the additional free space path loss encountered leads to a considerably reduced received power (in this case by 55 dB) or a much increased transmitter power.

6.2 Propagation Aspects of Satellite Communication Systems

6.2.1 The Geostationary Orbit

Clearly a very effective means for propagating around the earth's curvature at VHF and above is to place a repeater in satellite orbit. If the satellite can be made to appear stationary above the earth then, for all practical purposes, it functions in a manner similar to a terrestrial microwave repeater.

To appear stationary the satellite has to be in orbit above the equator, move in the same direction as the earth's rotation and have an orbital period of 24 hours. The satellite and earth then rotate together. If the satellite orbit were not equatorial then, when viewed upwards from the earth, it would appear to oscillate either side of the equator with a 24 hour period.

How high does a satellite need to be above the earth in order to have a 24 hour period? The answer to this question is given from the expression (derived from Kepler's Laws of planetary motion)

$$T = \frac{2\pi}{\mu} \sqrt{r^3}$$

in which T is orbital period in seconds, r is orbital radius in metres and μ is the earth gravitational constant, given by

$$\mu = 3.986 \times 10^{14} \text{m}^3 \text{s}^{-2}$$

Noting that $r = a + h$, where a is earth radius (6.37 Mm) and h is the height of the satellite above the earth's surface, then a period of 24 hours requires $h = 35.87$ Mm. Therefore the satellite would need to be inserted into an orbit 35,870 km above the equator travelling in the same direction as the earth's rotation. Such an orbit is called *geostationary* or sometimes *geosynchronous*. The most efficient launch strategy in terms of minimising booster rocket energy, and thus fuel, is to launch from an equatorial or near equatorial site in an easterly direction.

Figure 6.9 shows that a satellite in geostationary orbit subtends an angle of 163° at the earth's surface. Clearly, a minimum of three satellites is needed for full equatorial coverage ($3 \times 163 = 489^\circ$). Note that the same geometry demonstrates that there is a latitude limitation on transmission to and from a geostationary satellite. Satellites are not visible beyond 81.5° latitude north or south. Even then ground based antennas at those latitudes would have to face horizontally; that would give an unacceptable level of ground noise and attenuation, and obstacles would be a problem. As a result, a minimum acceptable latitude is 75° for communicating via geostationary satellites. At higher latitudes, near-polar orbiting satellites would normally be used.

Another consideration with communication via a geostationary satellite is that the long transmission paths can cause significant delays in transmission. The worst case is for high latitude operation for which the total (two way) time delay is 280 ms, found by dividing twice the path length by the velocity of light. For transmission involving two way information exchange, such as a telephone conversation, that will

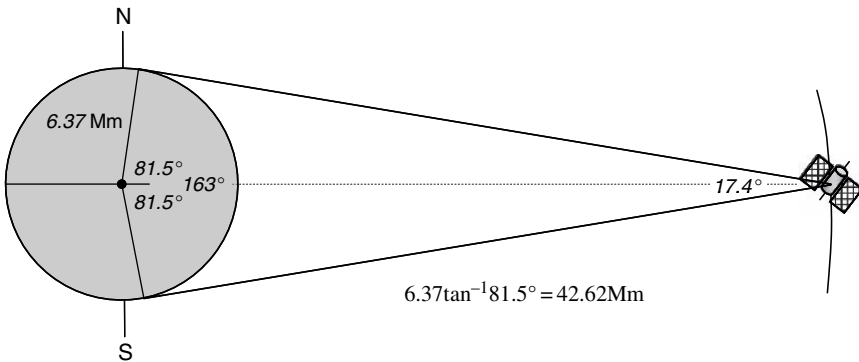


Fig. 6.9 Geometry of geostationary satellite communications

give a delay of about a half a second. If, for some reason, echoes occur in the system they will happen, therefore, a half a second behind the message that is transmitted.

6.2.2 Link Power Calculations

Figure 6.10 shows the essential signal and noise power considerations when a satellite takes the place of a terrestrial repeater in a microwave communications system. Two carrier to noise ratios are shown: that for the up-link, CNR_u , and that for the down-link, CNR_d .

For the up-link we can see

$$P_1 \text{ (dBm)} = P_t \text{ (dBm)} + G_t \text{ (dBi)} + G_{sr} \text{ (dBi)} - L_u \text{ (dB)}$$

$$= EIRP_t \text{ (dBm)} + G_{sr} \text{ (dBi)} - L_u \text{ (dB)}$$

and

$$N_1 = k(T_s + T_e)B = kT_e' B \equiv 10 \log(kT_e' B) \text{ dB}$$

so that

$$CNR_u \text{ (dB)} = EIRP_t \text{ (dBm)} + G_{sr} \text{ (dBi)} - L_u \text{ (dB)} - 10 \log(kT_e' B)$$

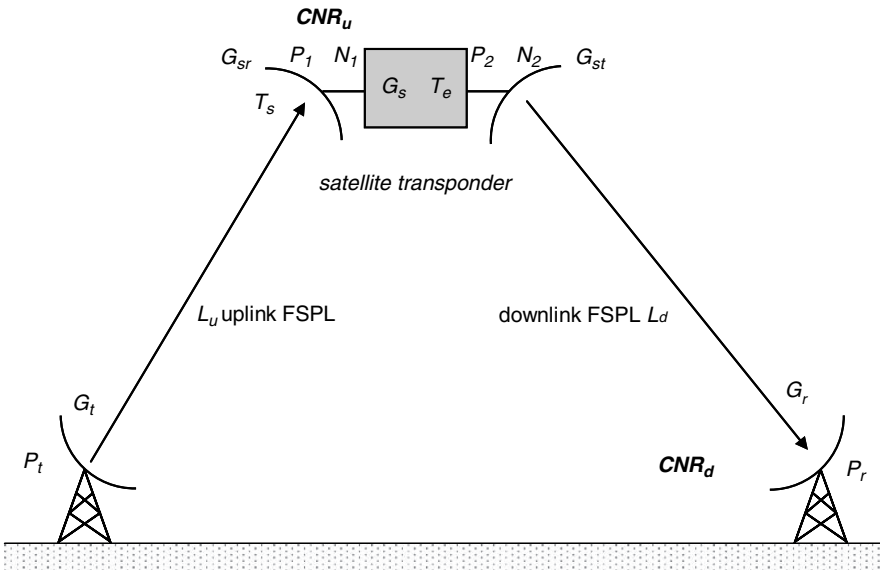


Fig. 6.10 Link power variables for satellite link

This equation can be re-written

$$\begin{aligned} CNR_u(\text{dB}) &= EIRP_t(\text{dBm}) + G_{sr}(\text{dBi}) - L_u(\text{dB}) - 10\log(T'_e) - 10\log kB \\ &= EIRP_t(\text{dBm}) - L_u(\text{dB}) - 10\log kB + 10\log(G_{sr}/T'_e) \\ &= EIRP_t(\text{dBm}) - L_u(\text{dB}) - 10\log kB + G_{sr}/T'_e(\text{dBK}^{-1}) \end{aligned}$$

The last term in this expression is the ratio of antenna gain to the total equivalent input noise temperature of the satellite receiver. It is referred to as the receiver *figure of merit* or *G-to-T ratio*. For a given transmitter power, bandwidth and altitude it is the receiver G-to-T ratio that determines the quality of the received signal.

If for example

$$\begin{array}{ll} P_t = 1\text{kW} = 30\text{dBW} & \text{carrier frequency} = 6\text{GHz} \\ G_t = 53\text{dBi}, & \text{bandwidth} = 35\text{MHz} \\ (\text{so that } EIRP_t = 83\text{dBW}) & T'_e = 580\text{K}, G_{sr} = 19\text{dB} \end{array}$$

then $L_u = 199\text{dB}$ for a distance of 35,870 km.

also $10\log kB = -153\text{dB}$

and $G_{sr}/T'_e = -8.6\text{dBK}^{-1}$

giving $CNR_u = 83 + 153 - 8.6 - 199 = 28.4\text{dB}$

which would probably be a bit small since little fade margin is possible.

Now consider the down link. The power and noise outputs from the satellite transponder at the terminals of its transmitting antenna are

$$\begin{aligned} P_2(\text{dBm}) &= G_s(\text{dB}) + P_1(\text{dBm}) \\ N_2(\text{dBm}) &= G_s(\text{dB}) + N_1(\text{dBm}) \end{aligned}$$

so that the power and noise at the terminals of the ground receiving antenna are

$$\begin{aligned} P_r(\text{dBm}) &= P_2(\text{dBm}) + G_{sr}(\text{dBi}) + G_r(\text{dBi}) - L_d(\text{dB}) \\ &= EIRP_2(\text{dBm}) + G_r(\text{dBi}) - L_d(\text{dB}) \end{aligned}$$

$$N_d(\text{dBm}) = N_2(\text{dBm}) + G_{sr}(\text{dBi}) + G_r(\text{dBi}) - L_d(\text{dB})$$

Further noise is added to the received signal as a result of the receive antenna noise temperature and the equivalent input noise temperature of the receiver itself, so that

$$N_r(\text{mW}) = N_d(\text{mW}) + kT_a B(\text{mW})$$

in which T_a accounts for the receiver and antenna noise. Since the noise components are additive the last expression is not in dB form. Thus the down-link carrier to noise ratio at the receiver is given by the non-dB expression

$$CNR_d = \frac{EIRP_2 G_r L_d^{-1}}{N_2 G_{st} G_r L_d^{-1} + kT_a B}$$

Inverting this expression gives

$$\begin{aligned} CNR_d^{-1} &= \frac{N_2}{P_2} + \frac{kT_a B}{EIRP_2 G_r L_d^{-1}} \\ &= \frac{G_s N_1}{G_s P_1} + \frac{kT_a B}{EIRP_2 G_r L_d^{-1}} \\ &= \frac{kT_e' B}{P_1} + \frac{kT_a B}{EIRP_2 G_r L_d^{-1}} \end{aligned}$$

The first term in this last expression is the reciprocal of the up-link carrier to noise ratio (i.e. the noise to carrier ratio) and the second accounts for the degradation in carrier to noise ratio caused by ground receiver noise. Clearly the degradation in carrier to noise ratio needs to be kept to a minimum. Writing the reciprocal of the last term as

$$\Psi = \frac{EIRP_2 G_r L_d^{-1}}{kT_a B}$$

we have

$$\Psi(\text{dB}) = EIRP_2(\text{dB}) + G_r / T_a(\text{dBK}^{-1}) - kB(\text{dB}) - L_d(\text{dB})$$

in which it is seen that maximising the receiver G-to-T ratio is again an important parameter to maintain good carrier to noise ratio on reception. In the down-link it is more important since the EIRP of the satellite will be limited because of weight restrictions.

6.3 The Propagation Aspects of Cellular Radio

For a mobile radio system, particularly carrying telephony and other consumer services, the major requirement is to support many users in a common geographical region without them interfering with each other. That involves what is sometimes called “frequency re-use” in which the same set of carrier frequencies is used but separated geographically so that likely interference is minimised. This can be achieved by arranging the spatial domain into a set of *cells*, nominally hexagonally shaped, each of which has its own available set of carrier frequencies. The set of cells forms a *cluster*. Clusters are repeated spatially and the carrier frequencies are re-used in each cluster. The transmitter power from a base station located at the centre of each cell is kept sufficiently small that the signal should not interfere with that from the “same” cell in an adjacent cluster.

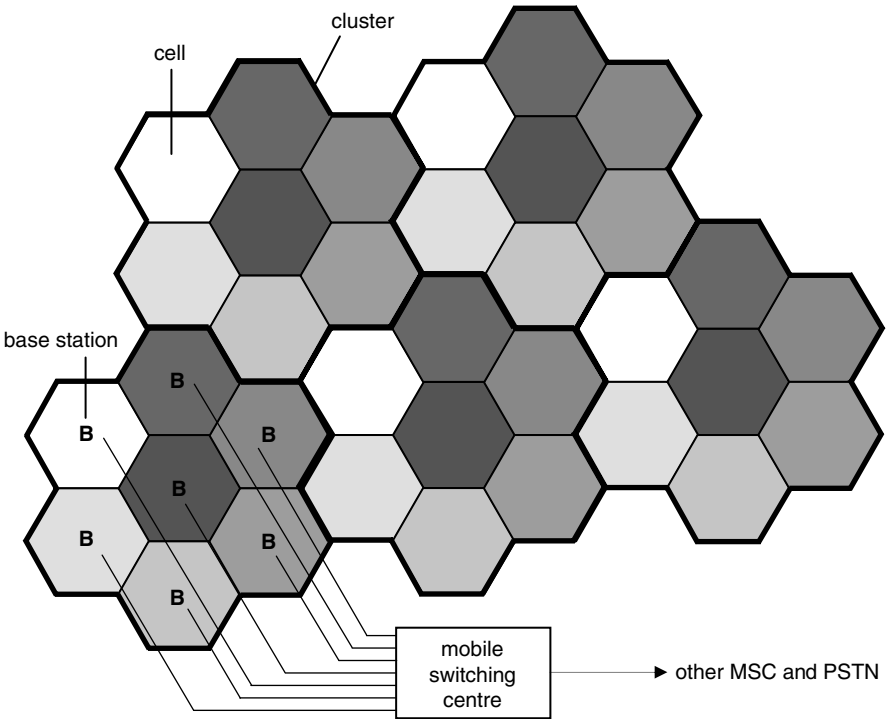


Fig. 6.11 Essential elements of a cellular telephone network

A repeating cluster of cells is shown in Fig. 6.11. Generally there are 4, 7 or 12 cells per cluster. The base station at the centre of each cell acts as a transmitter and a receiver for all users currently in the cell. Each base station is connected to a mobile switching centre (MSC) which is itself connected to other MSCs and the public switched telephone network (PSTN).

We can now calculate the level of carrier interference likely to be experienced as a result of frequency re-use. The worst case will be for a user at the edge of a cell receiving a signal from the same cell of an adjacent cluster (interference) when it is also receiving a signal from the base station in its own cell (signal). This is called *co-channel interference*. From the geometry identified in Fig. 6.12 we can see that the distances of the receiver to the intended and interfering base stations respectively can be shown for an N cell cluster to be

$$d_1 \approx r$$

$$d_2 \approx r \left(\sqrt{3N} - 1 \right) \approx 3.6r \quad \text{for } N = 7$$

The level of co-channel interference will depend upon the propagation mechanism that leads to the reduction of both the signal and the interference at the receiver.

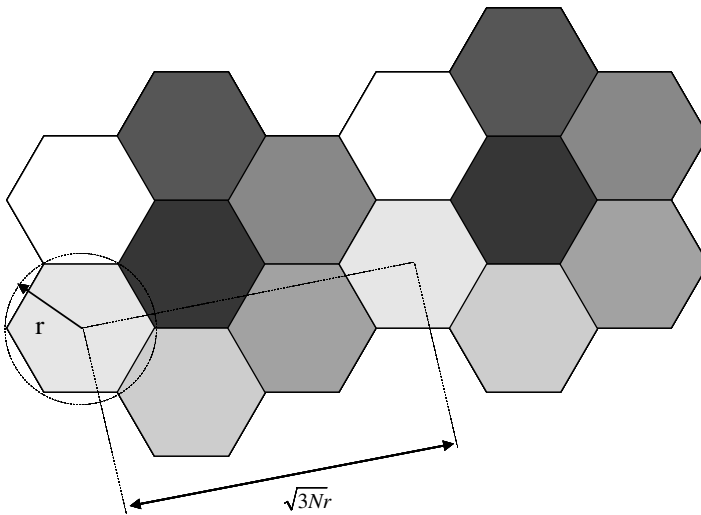


Fig. 6.12 Geometry for calculating co-channel interference

Mobile telephony services operate at carrier frequencies of the order of 1 GHz, so that the propagation mechanism is line-of-sight space wave; reflections from and diffraction around obstacles, including buildings in cities, will often allow the signal to be received in shadow zones.

If, for simplicity, we assume that the signal reaches the receiver from both base stations involving the classical case of a direct and single ground reflected ray then the received signal strength will fall off with the square of distance, and the received power by the fourth power of distance. Therefore the ratio of the power from the interferer (P_2) and the local base station (P_1) will be

$$\frac{P_2}{P_1} = \frac{r^4}{(3.6r)^4} = \frac{1}{168} \equiv -22\text{dB}$$

With the complex multi-path situation that might be experienced in an urban region the dependence of power on distance is taken to be between the inverse third and inverse fourth power. Also, the situation just considered ignores interference from similar cells in other clusters still; when combined these can degrade the co-channel interference by several dB.

6.4 Mobile Wireless Systems

A great number of modern wireless telecommunications systems are mobile, in the sense that they are either hand carried (as in mobile or cell phones, or personal data assistants) or they are mounted in vehicles (such as mobile phones, FM radios and

navigation systems). Most of those services operate in the VHF and UHF bands, typically with frequencies around 100 MHz for FM radio and 900 MHz to 2.4 GHz for cell phones, PDAs and Bluetooth-like operation. As a consequence, the propagation mechanism is the space wave and is subject to interference of direct and any (or often many) reflected paths that transfer energy from the transmitter to the receiver.

Figure 4.3 illustrates the type of interference that can occur and how it depends on distance between the transmitter and receiver. In a mobile system, of course, that distance will vary with time, so that the received field strength can vary profoundly with time, with the service likely to fade, both rapidly and with a longer term cycle.

The situation depicted in Fig. 4.3, is very idealised, in that it represent the case of a single ground reflected ray. Most mobile systems encounter many and varied reflections; sometimes reception may not involve a direct ray at all since the receiver may be out of line of sight of the transmitter. That is particularly the case with reception in urban regions; that reception is possible at all depends upon many reflected paths. A simple depiction of a mobile situation is shown in Fig. 6.13, in which both reflection and diffraction are shown as important receiving mechanisms.

To illustrate the complexity of the received signal in such situations Fig. 6.14 shows the received signal with one, two and three reflected rays, along with the direct path. As noted there is significant interference in all cases, making the situation in practice difficult to predict, especially when the number of pathways is large, unknown and time-varying.

In principle, the received signal will be a (phasor) sum of the form

$$\mathbf{E}_r = \sum_n E_n(d) e^{-j\phi_n(d)}$$

in which the phase angles ϕ_n are effectively randomly distributed and time varying (because d varies with motion); they also depend on the nature of the reflection coefficients at the surfaces encountered at each reflection (see Sect. 7.4). The individual ray amplitudes E_n are also random and again vary according to distance (and

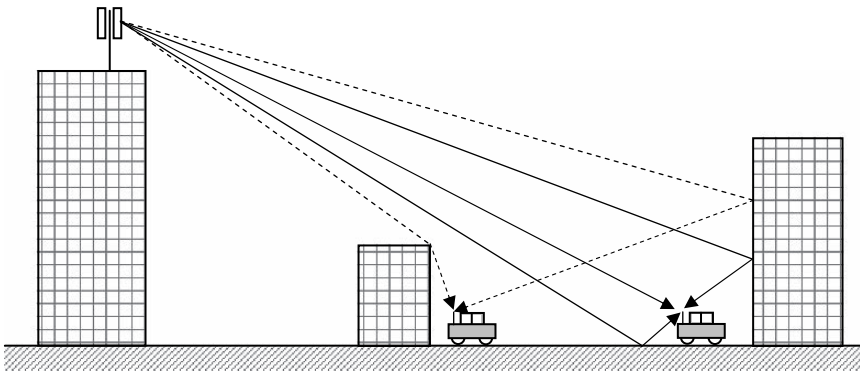


Fig. 6.13 Illustrations of typical propagation pathways in mobile telephony; in reality there can be reflections from many other buildings, other vehicles and even trees and people

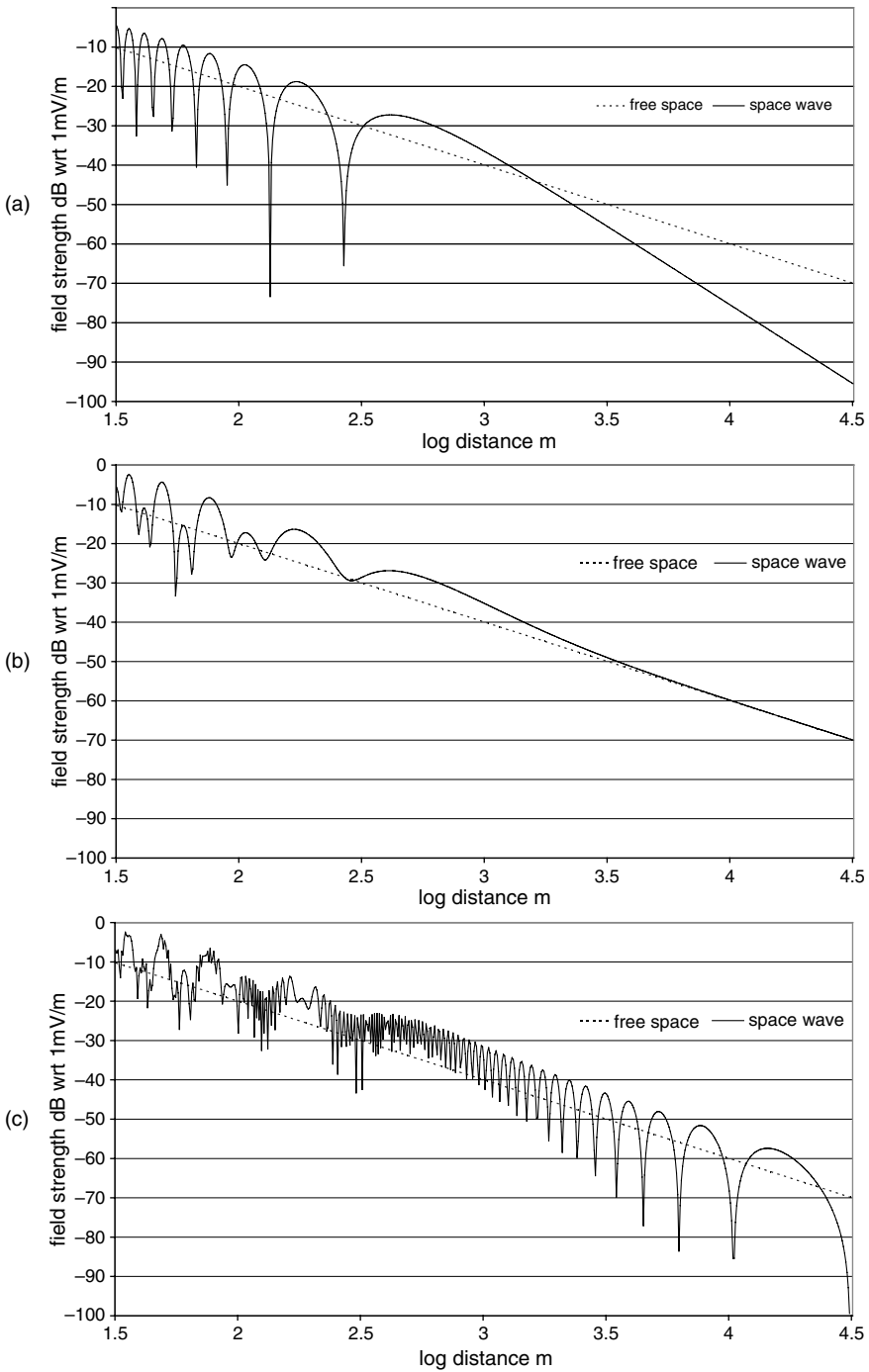


Fig. 6.14 (continued)

thus time in a mobile system) and with surface reflection coefficient. The number of interfering rays, n , can also vary as paths drop in and out. Because of these factors it is difficult to model any reasonable practical system. Consequently, the statistics of the received signal, particularly in an urban region, are usually measured by driving a vehicle-mounted receiver through a city and recording the received field strength. Signals received in the absence of a direct ray are observed to have a Rayleigh distribution whereas, if a direct path is present such that that ray dominates reception, the distribution changes to Rician.¹

A final practical point to note is that even though the average received signal strength seen in Figs. 4.3 and 6.13 falls in an inverse distance fashion with distance, the user of such a system is largely unaware of that fact because receiver systems incorporate a degree of automatic gain control (AGC). AGC adjusts the receiver output signal up and down with decreases and increases in RF reception level. On the average therefore the signal perceived by the user does not increase or decrease in level, but does suffer drop-out with the deep fades evident in the figures since the AGC does not respond to such short term and major drops in signal strength.

Problems

- 6.1.** Discuss the causes of fading in open microwave communications systems and means for minimising their effect on received signals.
- 6.2.** Why is G/T such an important figure of merit in a satellite communications system?
- 6.3.** At high latitudes geosynchronous communications satellites are of limited value. Why is that and what is the solution to this problem in order to provide reliable satellite communications?
- 6.4.** Assuming that the interfering signals add incoherently, so that their field strength magnitudes can simply be added, determine the co-channel interference in a 7 cell

Fig. 6.14 Received signal strength for a transmitting antenna of height 10 m, a receiving antenna of height 2 m, a transmitter power of 2 W, a transmitting antenna gain of 2.23 dBi and an isotropic receiving antenna, with an operating wavelength of 15 cm: (a) the situation with the direct ray and a single ground reflected ray; (b) the received signal when a second reflected path is added in which the effective transmitter height for that path is 9 m and the receiver height is 1 m – such as might happen with a second reflection from the top of a low structure in the path, such as another vehicle; (c) the received signal when a third reflecting path is added, such as from a nearby building in which the “equivalent” transmitter and receiver heights for the third path are 40 m and 30 m respectively

¹ A detailed discussion of the statistics of space wave signals can be found in H.L. Bertoni, *Radio Propagation for Modern Wireless Systems*, NJ, Prentice Hall, 2000.

cluster cellular radio system such as that shown in Fig. 6.12, but in which the interference comes from all nearest clusters. This requires you to work out how many clusters are immediately adjacent to the cluster of interest.

6.5. Passive reflectors in an open microwave repeater system are generally more effective when they are closer to one of the active repeaters. Why would that be the case?

6.6. Is diversity reception likely to be required over rough terrain between microwave repeaters?

6.7. What are the comparative advantages of microwave radio repeater systems and optical fibres for broadband trunk applications?

Chapter 7

The Effect of Materials on Propagation

7.1 Background

The mechanisms for propagation in the various frequency ranges outlined in the previous chapters, with a few exceptions, take little account of the properties of the various materials that interact with the passage of radio wave energy. We have looked at the effect of atmospheric refraction, attenuation resulting from atmospheric constituents and rainfall, and the presence of obstacles in the propagation path. It is now important to look at environmental effects in more detail particularly at the frequencies used by services such as wireless networking. In those cases, the fact that the services are often mobile and operate close to environmental features such as walls, trees and buildings means that some of the more quantitative results we examined, particularly for space wave propagation, need to be qualified by a knowledge of matter-wave interaction. It is the purpose of this chapter to look at the electrical properties of materials so that such an appreciation can be developed.

Essentially, there are two considerations: the material properties themselves and how those properties combine with the geometric nature of the domain to influence the propagation of radiation.

The three material properties of importance are:

conductivity	σ	Sm^{-1}
dielectric constant	ϵ_r	dimensionless
relative permeability	μ_r	dimensionless

Conductivity describes the tendency for a material to absorb energy from a propagating wave via conversion to heat. Dielectric constant describes how the electrical properties of a medium, represented by the electrons in orbit around their parent atoms and the rotation and vibration resulting from the polar nature of some molecules, influence the passage of a wave. In a similar manner relative permeability describes the influence of the medium on the wave resulting from its magnetic properties. Most media in which we are interested are non-magnetic so we can assume $\mu_r = 1$ in all cases; as a result we can focus our attention on conductivity and dielectric constant.

Conductivity and dielectric constant affect electromagnetic energy because of their influence on the propagation parameters of a medium and their discontinuities at interfaces. The medium's parameters are ideally summarised in the propagation constant. Interface effects are summarised in so-called reflection coefficients or, in some cases, by scattering coefficients if the interfaces are irregular on the scale of a wavelength. A medium's bulk properties are affected by inhomogeneities in dielectric constant and conductivity which, in turn, influence radio waves. We will consider the ideal situations first and then come back to the case of rough surfaces and inhomogeneous media.

7.2 Propagation in Homogeneous Media

The electric field of a wave travelling in a homogeneous medium can be represented as

$$E = E_o e^{j\omega t - \gamma z}$$

in which γ is the *propagation constant* given, for non-magnetic media, by

$$\gamma^2 = j\omega\mu_o\sigma - \omega^2\mu_o\epsilon \quad (7.1)$$

The accompanying magnetic field (see Fig. 1.3) has the same propagation constant. Since the square of the propagation constant is complex, the propagation constant itself, in general, is also complex:

$$\gamma = \alpha + j\beta \quad (7.2)$$

in which α is called the *attenuation constant*; it is usually expressed in dBm^{-1} although when calculated from (7.1) its raw units are nepers per metre (Npm^{-1}). We can convert Npm^{-1} to dBm^{-1} by

$$\text{dBm}^{-1} = 8.686 \text{Npm}^{-1}$$

β is the *phase constant* expressed in rad.m^{-1} . Using (7.2) the expression for the propagating electric field can be written

$$E = E_o e^{-\alpha z} e^{j(\omega t - \beta z)}$$

showing that a non-zero attenuation constant means that the strength of the field (and thus the available power density) diminishes with distance travelled in the medium. Likewise, the phase of the field changes with propagation in the medium.

We can use this last expression to demonstrate that the wave travels. First, assume it is lossless so that $\alpha = 0$. The phase term is

$$\theta = \omega t - \beta z$$

Suppose we now lock ourselves on to a point of constant phase and see if, and how quickly, we move much as a surf board rider, sitting at a particular position on a water wave, gets carried forward by the wave. Thus θ is constant and we have

$$z = \frac{\omega}{\beta}t - a$$

where a is a constant. Thus there is a linear relationship between position and time, linked by the (phase) velocity ω/β .

Return now to (7.1) and re-write it as

$$\begin{aligned} \gamma^2 &= -\omega^2 \mu_o \epsilon \left(1 - \frac{j\omega \mu_o \sigma}{\omega^2 \mu_o \epsilon} \right) \\ &= -\omega^2 \mu_o \epsilon_o \epsilon_r \left(1 - \frac{j\sigma}{\omega \epsilon_o \epsilon_r} \right) \\ &= -\omega^2 \mu_o \epsilon_o \left(\epsilon_r - \frac{j\sigma}{\omega \epsilon_o} \right) = -\omega^2 \mu_o \epsilon_o \epsilon_r^* \end{aligned} \quad (7.3)$$

in which ϵ_r^* is said to be the *complex dielectric constant* of the medium through which the wave travels. Conventionally, the complex dielectric constant is written

$$\epsilon_r^* = \epsilon_r' - j\epsilon_r''$$

in which it can be seen that its real part is the ordinary dielectric constant of the medium and its imaginary part is directly related to the medium's conductivity (and thus loss) through $\sigma/\omega\epsilon_o$.¹ If the conductivity of the medium is negligible, i.e. $\sigma \rightarrow 0$ then $\epsilon_r^* = \epsilon_r$ (real) and $\gamma^2 = -\omega^2 \mu \epsilon$, giving $\alpha = 0$ in (7.2) and

$$\beta = \omega \sqrt{\mu_o \epsilon}, \quad \epsilon = \epsilon_r \epsilon_o$$

showing that the wave will travel without attenuation in the medium and will incur only a phase change. The medium is then called *lossless*.

If we recall from (1.8) that $\sqrt{\mu_o \epsilon_o} = 1/c$, then

$$\beta = \frac{\omega}{c} \sqrt{\epsilon_r} = \frac{\omega n}{c} \quad (7.4)$$

where n is the refractive index of the lossless medium. Using the definition of phase velocity in (3.7) – which applies in general, and not just for the case of ionospheric propagation – then the phase velocity in a non-conducting (i.e. lossless) medium is

¹ The real situation is more subtle than that portrayed here because of the actual behaviour of dielectrics in response to an alternating electric field. The conductivity term used in our equations describes the ohmic losses of the medium as the result of lattice collisions of free carriers. There are also losses in the dielectric itself resulting from damping of the polarisation response of the material. That is discussed in Sect. 7.3 following.

$$v_{\text{phase}} = \frac{\omega}{\beta} = \frac{c}{n}$$

showing that the phase velocity of the radiation is slower than the speed of light if the refractive index of the medium is greater than unity. Similarly, using (3.9), the group velocity in the medium – i.e. the velocity with which any modulation of the radiation will travel – is given by

$$v_{\text{group}} = \frac{\partial \omega}{\partial \beta} = \frac{c}{n}$$

which is constant and equal in value to the phase velocity for this lossless medium. Therefore all frequency components of the signal carried by the radiation (i.e. any side bands and carrier) travel with the same velocity so that the signal stays intact – in other words there is no dispersion and thus no distortion of the signal.

Consider now the other extreme, in which the medium is highly conducting so that $\frac{j\sigma}{\omega\epsilon_0} \gg \epsilon_r$ in (7.3), giving

$$\epsilon_r^* = -\frac{j\sigma}{\omega\epsilon_0}$$

and

$$\gamma^2 = j\omega\mu_0\sigma$$

so that

$$\gamma = (1 + j)\sqrt{\frac{\omega\mu_0\sigma}{2}}$$

Thus the attenuation coefficient in highly conducting media (such as sea water) is

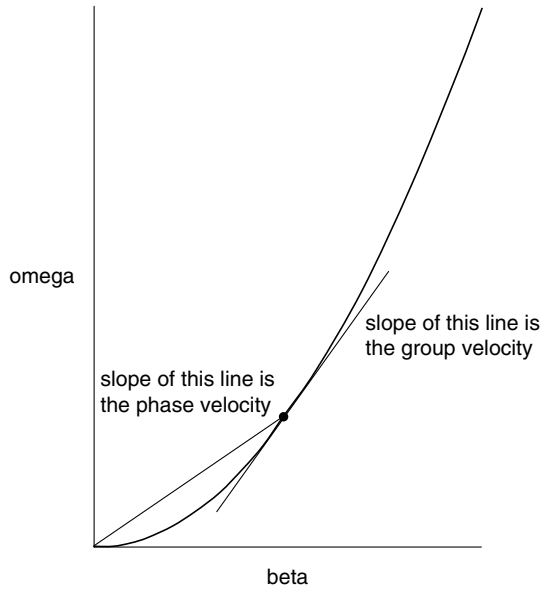
$$\alpha = \sqrt{\frac{\omega\mu_0\sigma}{2}} \quad (7.5)$$

which increases as the square root of operating frequency; as a consequence better transmission is possible at lower frequencies. Likewise the phase constant for highly conducting media is

$$\beta = \sqrt{\frac{\omega\mu_0\sigma}{2}} \quad (7.6)$$

Given that we do not now have a linear relationship between ω and β the group and phase velocities will be different and the medium will be dispersive. This can be observed conveniently by plotting the relationship in what is called an $\omega - \beta$ diagram, as shown in Fig. 7.1. From their definitions, we can see that phase velocity is given by the slope of the line drawn from the origin to a point on the curve at a frequency of operation, while group velocity is the instantaneous slope of the curve at that frequency.

Fig. 7.1 Omega-beta diagram for a conducting medium



We have looked at the extremes of ideally lossless media and highly conducting media in the cases just treated. Of course, more generally real media are both conducting and dielectric. To develop a concept of how their behaviours may be considered return to (7.1) and re-write it as

$$\gamma^2 = j\omega\mu_0\sigma - \omega^2\mu_0\varepsilon = -\omega^2\mu_0\varepsilon\left(1 - j\frac{\sigma}{\omega\varepsilon}\right)$$

from which we can see that if

$$\frac{\sigma}{\omega\varepsilon} \gg 1 \text{ a medium behaves like a conductor}$$

$$\frac{\sigma}{\omega\varepsilon} \ll 1 \text{ a medium behaves like a dielectric}$$

In between, say,

$0.01 < \frac{\sigma}{\omega\varepsilon} < 100$ a medium is said to be a quasi-conductor and analysis requires use of (7.1) in its full form.

Thus frequency determines whether a material acts like a conductor or a dielectric. Consider sea water as an example. If we assume it has a dielectric constant of 80, and a conductivity of 4Sm^{-1} , then at 1 MHz we have

$\frac{\sigma}{\omega\varepsilon} = 4/(2\pi \times 10^6 \times 80 \times 8.85 \times 10^{-12}) = 899$, which means it behaves as a conductor at MF and below.

However, at 100 GHz we have

$\frac{\sigma}{\omega\varepsilon} = 4/(2\pi \times 10^{11} \times 80 \times 8.85 \times 10^{-12}) = 0.009$, which means it behaves like a dielectric at EHF.

7.3 Frequency Dependence of Material Properties

It is important now to return to a consideration of the material properties σ and ϵ_r . The previous discussion treats them as though they are constants. Unfortunately, that is not the case; at minimum they are functions of frequency. Consider water, as an example. At microwave frequencies it has a dielectric constant of about 80, giving a refractive index of about 9. However, at light frequencies, its refractive index is about 1.33! To see why there is such a difference it is necessary to understand a little about the physical basis for permittivity (or dielectric constant, or refractive index).

In a conductor the electrons are only loosely bound to their parent atoms and can migrate through the material if an electric field is applied. Their collisions with atomic sites is what defines the resistivity, and thus the conductivity, of the material. Conductivity can be a strong function of temperature but is often regarded as not strongly dependent on frequency for the ranges of interest in radio wave applications.

In a perfect insulator or dielectric electron migration cannot occur (unless the field is so strong that dielectric breakdown is caused) although the electron clouds will be shifted about their respective nuclei as a result of the applied field. That is called polarisation. If the field is alternating, as in the case of a radio wave travelling through the medium, then the electron cloud, in principle, oscillates about the nucleus in synchronism with the field. In addition, if the medium is composed of polarised molecules then those molecules can vibrate and, if free as in the case of water, can even rotate in sympathy with the field. If they are rigidly bound, such as in ice crystals, then they are unable to rotate.

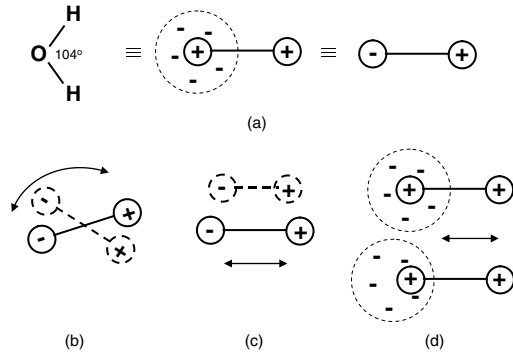
The responses of the molecules and atoms to the incident field gives rise to a field resulting from their own charge separations. Therefore the actual field in the dielectric is a combination of the applied and induced fields.² It is this influence on the applied field that leads to the concept of permittivity. Essentially, the higher the permittivity (and thus dielectric constant) the greater the effect the material has on the field, and properties such as its phase velocity. It is instructive to consider this further with water.

As depicted in Fig. 7.2, the water molecule has three responses to an incident alternating electric field: the electrons can oscillate about their nucleus, the molecule can flex or vibrate and the molecule can rotate. In principle, all three mechanisms are present together.

Imagine now that the frequency of the electric field is low. Then the water molecules can easily rotate with the alternations of the field vector and thus influence the wave. As frequency is increased the molecules keep up until such a frequency is encountered that the molecules' inertia will not allow them to rotate quickly enough to influence the field. However, the polar molecules still vibrate; their positive and negative ends are pushed in and pulled out in unison with the

² This total field is complicated to determine, but for which see J.R. Reitz and F.J. Milford, *Foundations of Electromagnetic Theory*, Addison-Wesley, Mass., 1962 or J.D. Kraus, *Electromagnetics*, 5th ed., McGraw-Hill, New York, 1999.

Fig. 7.2 Depicting the water molecule as polar as a result of the concentration of the electron orbitals towards the oxygen atom (a) and its rotational (b) vibrational (c) and electron displacement (d) responses to an applied electric field vector



applied field. This is a more agile response than rotation at lower frequencies and thus, while the molecules still have an influence on the field, it is not as great as the rotational effect. Ultimately, the frequency can be so high that the vibrations cannot keep up with the field. All that is left to respond is the clouds of electrons around each of the constituent atoms. Although not negligible, that is a very small effect in contrast to the other two. If the frequency is made extremely high – well out beyond visible light – even the clouds of electrons can't respond and thus the medium ceases having an effect on the wave. Then the dielectric behaviour is at its lowest – equivalent to that of a vacuum – so that the dielectric constant of the medium approaches unity.

There are losses associated with each the mechanisms in Fig. 7.2 that attempt to dampen the responses. Therefore, even though there may be no free electrons to give rise to the classical concept of conductivity, there are nevertheless losses that can be described by an equivalent conductivity for the dielectric. Generally this is written in the form of a complex permittivity in which the imaginary component summarises the lossiness of the pure dielectric material:

$$\epsilon^* = \epsilon' - j\epsilon''$$

Substituting this last expression into (7.1), which applies to a medium with both conducting and dielectric properties, we have

$$\gamma^2 = j\omega\mu_0(\sigma + \omega\epsilon'') - \omega^2\mu_0\epsilon'$$

which we can write

$$\gamma^2 = -\omega^2\mu_0 \left(\epsilon' - \frac{j(\sigma + \omega\epsilon'')}{\omega} \right) = -\omega^2\mu_0\epsilon^*$$

in which

$$\epsilon^* = \epsilon' - \frac{j(\sigma + \omega\epsilon'')}{\omega}$$

is the complex permittivity of the combined conductive and dielectric medium. Sometimes the *loss tangent* is defined

$$\tan \delta = \frac{\sigma + \omega \epsilon''}{\omega \epsilon'}$$

which for non-conducting media is

$$\tan \delta = \frac{\epsilon''}{\epsilon'}$$

This is a property of all dielectrics.

The mechanisms just described are summarised in Fig. 7.3. While we have discussed them in the context of water, they essentially apply to any material that exhibits a polar molecular form. Note that there are three distinct resonances corresponding to the cessations of each mechanism at which the ideal dielectric appears abnormally lossy.

We can now see why the dielectric constant of water can vary over such extreme values. At microwave frequencies we have not encountered even the first of the transitions in Fig. 7.3. The water molecules can keep up with the field and have a significant influence on it, leading to a high dielectric constant. At the frequency of visible light the rotational response has ceased and the dielectric constant is now dominated by the vibrational response of the molecule itself and the electron cloud; it is therefore much smaller.

In the natural environment water is interesting since its dielectric constant is so high. Table 7.1 shows indicative dielectric constants of a number of materials that might be encountered in free space radio wave propagation.³ The most notable feature is how low those values are compared with water (because of its highly polarised molecules). As a result, any moisture content will play a significant role on a material's interaction with electromagnetic energy. For example, Fig. 7.4 shows

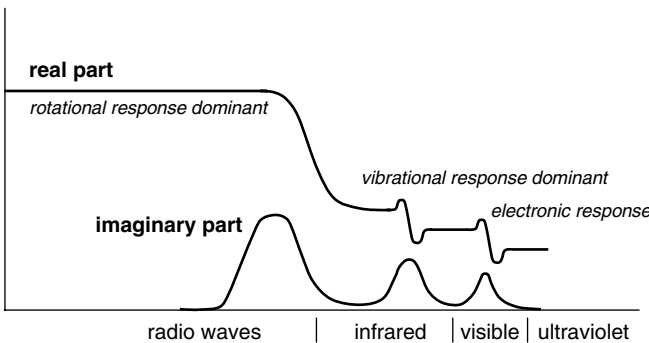


Fig. 7.3 Behaviour of the complex dielectric constant as a function of frequency

³ Many studies of the complex dielectric constants of a range of materials can be found through web searching, which should be used to find reasonably accurate values under those conditions and frequencies of interest.

Table 7.1 Typical dielectric constants for some non-conducting materials

Material	Dielectric Constant	Loss Tangent
Window Glass	4.0	0.0008 at 1 MHz 0.0012 at 100 MHz
Neoprene	5.7	0.095 at 1 MHz
Wood	1.8	0.027 at 3 GHz
Concrete	7	0.12 at 1 GHz
Plywood	2.7	0.18 at 2 GHz
Styrofoam	1.15	~ 0
Dry Brick	4	0.025 at 3 GHz
Polystyrene	2.55	0.0003 at 10 GHz
Dry Sand	2.5	0.004 at 10 GHz
Nylon	3.14	0.011 at 10 GHz
Distilled Water	80	0.2 at 10 GHz
Ice	3.2	0.00091 at 3 GHz
Snow	1.4	0.0018 at 3 GHz

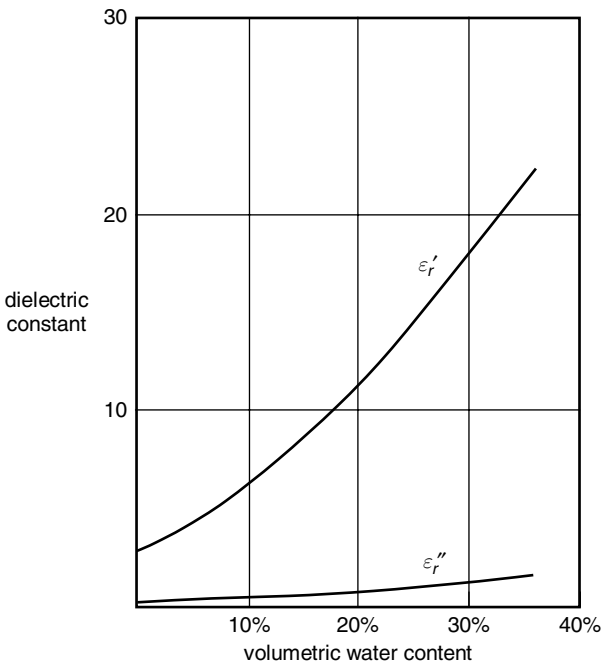


Fig. 7.4 The effect of moisture content on the complex dielectric constant of sand at 1.4 GHz. Based on J.R. Wang, The dielectric properties of soil-water mixtures at microwave frequencies, *Radio Science*, Vol. 15, No. 5, 1980, pp. 997–985

the striking effect of water content on the complex dielectric constant of sand. Any propagation path involving a soil surface, such as a ground reflected wave in space wave propagation, will be influenced by recent rainfall, and in the extreme when dry ground is replaced by flood waters.

Finally, note that the dielectric constants summarised in Table 7.1 are for the sorts of materials that might be encountered environmentally and are not indicative of the very high values for materials such as those used for dielectrics in capacitors.

7.4 Interactions with Ideal Interfaces

We now turn our attention to what happens when a wave encounters the interface between two media. In particular, we examine first an air-medium boundary for which part (or indeed all) of the wave may be reflected. Such a situation is encountered for the reflected pathways in space wave propagation.

Figure 7.5 shows the situation in outline, using rays to represent the passage of an electromagnetic field. The incident signal has a component that is transmitted into the medium (and refracted according to Snell's Law of Refraction) and a component that is reflected according to the Law of Reflection.

Assuming that the medium on one side of the interface is free space (as a good approximation to air) then the reflected wave travels without loss. However, the component transmitted into the medium will undergo attenuation (and most likely dispersion) as treated in the previous section.

At the interface, the transmitted and reflected fields are related to the incident fields through transmission and reflection coefficients τ and ρ :

$$\frac{E_{\text{transmitted}}}{E_{\text{incident}}} = \tau \qquad \frac{E_{\text{reflected}}}{E_{\text{incident}}} = \rho$$

Those coefficients are functions of the material properties either side of the interface. They are also functions of the polarisation of the radiation. In Chap. 1 we

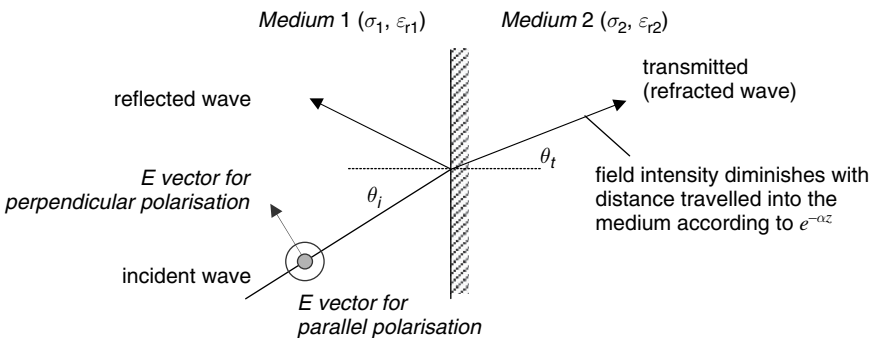


Fig. 7.5 The reflected and transmitted components of a wave incident on an interface

described polarisation as vertical or horizontal, by reference to the orientation of the electric field vector with the earth's surface. Polarisation can also be circular or elliptical, in which the field vectors rotate in transmission, but that situation is less common than the two linear forms of polarisation and so is not considered further here.

However, we do need to be a little more precise in how we describe linear polarisation. Recalling that the electric (and magnetic) field vector is orthogonal to the direction of propagation, the electric field will be (or can be resolved into components) either parallel to the interface or perpendicular to the incident ray, as shown in Fig. 7.5. We call these respectively *parallel* or *perpendicular* polarisation. If the interface were the earth's surface then parallel polarisation is the same as horizontal polarisation. However, vertical polarisation will only be the same as perpendicular polarisation when the ray is travelling parallel to the interface (the earth's surface).

For perpendicular polarisation the reflection and transmission coefficients are given by⁴

$$\rho_{\perp} = \frac{Z_2 \cos \theta_t - Z_1 \cos \theta_i}{Z_2 \cos \theta_i + Z_1 \cos \theta_t} \quad (7.7a)$$

$$\tau_{\perp} = 1 + \rho_{\perp} \quad (7.7b)$$

while for parallel polarisation they are given by

$$\rho_{//} = \frac{Z_2 \cos \theta_t - Z_1 \cos \theta_i}{Z_1 \cos \theta_i + Z_2 \cos \theta_t} \quad (7.8a)$$

$$\tau_{//} \frac{\cos \theta_t}{\cos \theta_i} = (1 + \rho_{//}) \quad (7.8b)$$

Z_1 and Z_2 are the wave impedances of the media, defined for non-magnetic materials by

$$Z = \frac{j\omega\mu_o}{\gamma} \quad (7.9)$$

For lossless media, in which $\sigma = 0$, (7.1) and (7.9) show

$$Z = \frac{j\omega\mu_o}{j\omega\sqrt{\mu_o\varepsilon}} = \sqrt{\frac{\mu_o}{\varepsilon}}$$

which in free space ($\varepsilon = \varepsilon_o$) has the value $Z_o = \sqrt{\frac{\mu_o}{\varepsilon_o}} = 377\Omega$.

Return now to the case of Fig. 7.5. If both media are non-conducting (i.e. lossless), medium 1 is free space with a unity dielectric constant, and medium 2 has a dielectric constant of ε_{2r} , then the reflection coefficients become

⁴ For details and derivations see J.D. Kraus, *Electromagnetics*, 5th ed., McGraw-Hill, New York, 1999.

$$\rho_{\perp} = \frac{\cos \theta_i - \sqrt{\epsilon_{r2} - \sin^2 \theta_i}}{\cos \theta_i + \sqrt{\epsilon_{r2} - \sin^2 \theta_i}} \tag{7.10}$$

$$\rho_{//} = \frac{-\epsilon_{r2} \cos \theta_i + \sqrt{\epsilon_{r2} - \sin^2 \theta_i}}{\epsilon_{r2} \cos \theta_i + \sqrt{\epsilon_{r2} - \sin^2 \theta_i}} \tag{7.11}$$

If, on the other hand, medium 2 is conducting so that $\gamma = (1 + j)\sqrt{\frac{\omega\mu_0\sigma}{2}}$ then

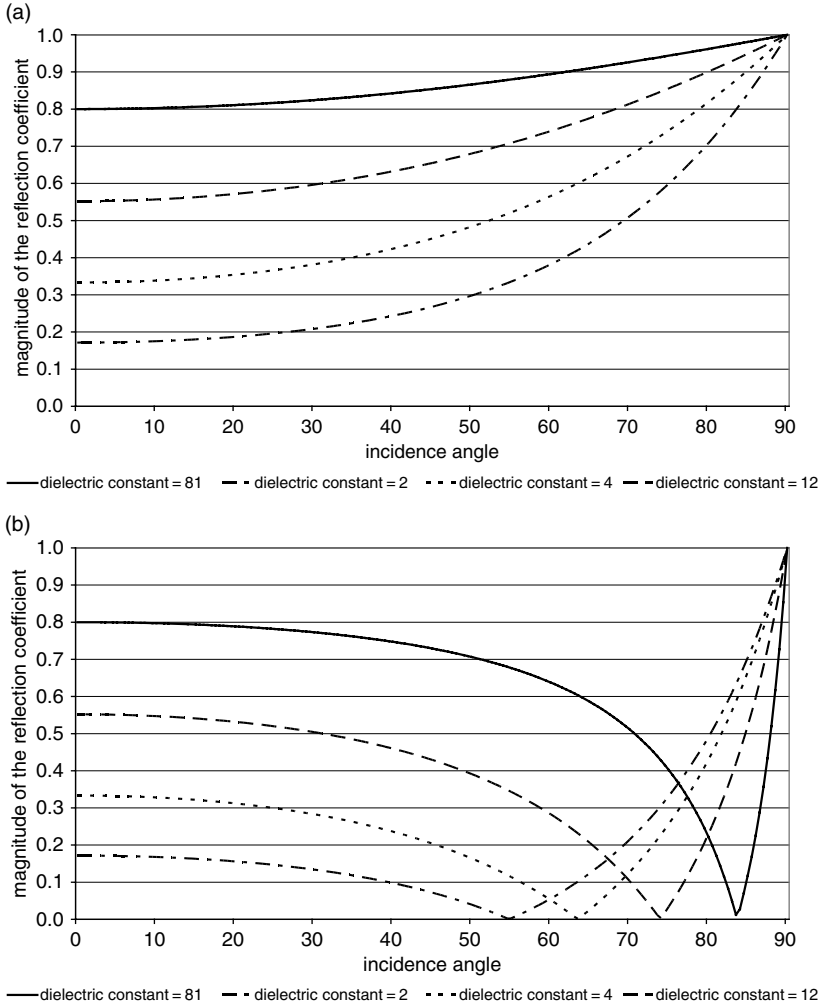


Fig. 7.6 Magnitudes of the perpendicular (a) and parallel (b) reflection coefficients at an interface between free space (nominally air) and media with the dielectric constants indicated

$$Z_2 = \frac{j\omega\mu_o}{\gamma} = (1 + j) \sqrt{\frac{\omega\mu_o}{2\sigma}}$$

If the medium is *highly* conducting, so that $\sigma \rightarrow \infty$, then $Z_2 \rightarrow 0$, which leads to

$$\rho_{\perp} = \rho_{//} = -1$$

Thus the whole field is reflected, as would be expected, with a 180° phase change.

Now return to an air to dielectric medium interface as described by (7.10) and (7.11). Unless $\epsilon_{2r} = 1$, in which case there is no interface, the perpendicular polarisation reflection coefficient cannot be zero. However, from (7.11) the parallel polarisation reflection coefficient can be zero if

$$-\epsilon_{r2} \cos \theta_i + \sqrt{\epsilon_{r2} - \sin^2 \theta_i} = 0$$

which requires

$$\theta_i = \sin^{-1} \sqrt{\frac{\epsilon_{2r}}{1 + \epsilon_{2r}}} \quad (7.12)$$

which is perfectly achievable. Thus there is a particular angle for which there is no reflection and all the field is transmitted across the boundary. That angle is called the *Brewster Angle*.

Figure 7.6 shows graphs of the reflection coefficients at an air-dielectric interface, illustrating their dependence on the dielectric constant of the medium and the angle of incidence. The Brewster Angle is evident in the parallel polarisation case.

7.5 Reflection from Rough Surfaces

The surfaces encountered in practice are not ideally smooth, as implied in the previous section. Rather they exhibit roughness on varying scales. Whether the roughness affects the reflection of an electric field according to the formulas given above depends upon the scales of the roughness compared with the wavelength of the radiation. It also depends upon the angle of incidence measured with respect to the surface normal θ_i with which the radiation encounters the surface. A surface is generally thought to act as if it were smooth if its vertical height variation h satisfies the Rayleigh criterion

$$h < \frac{\lambda}{8 \cos \theta_i} \quad (7.13)$$

Fortunately, many of the manufactured surfaces encountered by most radio wave services will appear moderately smooth under this criterion and thus the results of the previous section can be used in analysis. For example at 2.4 GHz the wavelength is 12.5 cm. The worst case is for vertical incidence for which the surface would need to be smoother than about 1.5 cm for the material of the previous section to apply.

Most 2.4 GHz services (Bluetooth, PDAs and similar) are short range and often indoors so such an assumption is plausible. At 1 GHz, which is typical of mobile telephony, the worst case (vertical incidence) is 3.75 cm. While many manufactured surfaces would exhibit variations of these orders over several tens of metres, it is unlikely that those variations would occur on the much shorter spatial scale over which the incident wavefront is assumed to be plane.

One significant exception to the above will be natural surfaces such as soils, grassland and other than placid water bodies, all of which are important when analysing the space wave with a ground reflected component.

The roughness of a surface influences the reflected wave in several ways. First, the reflected energy is no longer just in the direction described by the Law of Reflection (the so-called specular direction). Instead it tends to be scattered over a range of angles. If the surface is only slightly rough the reflection will be predominantly in the specular direction; as the scale of the roughness increases the energy increasingly scatters in all directions, including out of the plane of the geometry of Fig. 7.5. As a result the outgoing wave is referred to as a *scattered*, instead of a reflected, wave. Thus, the energy in the specular direction is reduced below the ideal value. This is particularly significant in the case of ground or other reflections for the space wave of Chap. 4. It also means that surfaces such as the walls of buildings, if sufficiently rough, will make signals available to receivers in a number of different directions. That is important in mobile telephony.

A second, less obvious, effect of surface roughness is that it depolarises the signal. That is, the scattered field can have a plane of polarisation different from that of the incident wave. Indeed the scattered wave will have both vertical and horizontal components.

Thirdly, if there is a significant transmitted wave – i.e. the rough interface is dielectric – then the transmitted wave will also be depolarised and scatter around the ray defined by Snell’s Law of Refraction.

Describing the relationship among incident, reflected and transmitted signals across rough interfaces is not simple because the scale and nature of the roughness are not known precisely. Instead, a number of approximate models are used. The first assumes that the surface is so rough that the “reflected” signal is scattered in all directions. Such a surface is often referred to as Lambertian, and is exemplified well by most natural surfaces at optical wavelengths – that is why we see most surfaces in a room as equally bright in all directions. Occasionally, there will be a specular (mirror-like) component, although for ideally rough surfaces that will not occur.

Lambertian scattering is best described by the scattering of power density rather than field, and can be represented by

$$p_s = p_i \frac{\rho^2}{\pi} \cos \theta_s \quad (7.14)$$

in which p_s and p_i are the scattered and incident power densities, ρ is the surface reflection coefficient and θ_s is the scattering angle at which the power density is of interest.

Specular reflection from very smooth surfaces, able to be characterised by the material of Sect. 7.4, and the Lambertian scattering just considered, can be regarded as the opposite extremes of scattering from real surfaces – i.e. the extremes of smoothness and roughness respectively. For intermediate scales of roughness the situation is much more complex to model, particularly if properties such as polarisation rotation are of interest. However, a very simple model sometimes used involves a straightforward modification of the specular reflection coefficient to make it exponentially decaying with scale of roughness, to give:

$$\rho_{eff} = \rho \exp[(-2\pi\sigma\lambda^{-1} \cos \theta_i)^2] \quad (7.15)$$

in which σ is the standard deviation of surface roughness. As expected this shows a reduction in the reflection coefficient at higher frequencies and larger scales of roughness. This is a well known model that is good for both horizontal and vertical polarisation; known as the Physical Optics model it's major limitation is a slightly inaccurate prediction of the Brewster angle for parallel polarisation.

Note that (7.15) describes the (reduced) reflection in the specular direction. It does not describe the accompanying field that is scattered diffusely about that direction, although we can say that the proportion of power density scattered away from the specular direction is $(1 - \rho_{eff}^2)$.

7.6 Transmission Through Media

Based on the representation in Fig. 7.5 it would seem that transmission through intervening media, such as a wall or forest, is simply a matter of determining how much field (or power density) crosses the interface and then noting how that diminishes with distance as a result of the attenuating properties of the medium. However the situation is complicated by the possibility of multiple reflections if a second interface is present, as it will be for a wall. The situation can be handled effectively using transmission line theory by regarding each of the three media as having their own characteristic impedances and computing the net signal that gets reflected and that which gets transmitted. Usually, unless a fairly precise analysis is required, we often just concern ourselves with the bulk attenuating properties of the medium, in the knowledge that the transmitted signal will be reduced further through reflection (or perhaps properly called *insertion*) loss. The loss of signal via reflection at the dielectric interface can be greater than that resulting from absorption in the medium, particularly if the medium has a reasonably high dielectric constant and low conductivity.

Table 7.2 gives the ohmic conductivities for a range of materials likely to be encountered in radio wave applications. To find the corresponding attenuation coefficient in the medium the material properties need to be substituted into (7.1) and the real part of the resulting complex expression found.

Table 7.2 Conductivities of common media at microwave frequencies

Material	Conductivity Sm^{-1}
Dry soil	0.00001
Sea water	4
Fresh water	0.01
Polystyrene	$\sim 10^{-16}$
Glass	$\sim 10^{-12}$

There will also be dielectric loss in natural media resulting from the imaginary part of the dielectric constant treated in Sect. 7.3. Assuming that the ohmic conductivity σ is negligible we can use the data from Fig. 7.4 and the definition of complex permittivity to show that the attenuation in dry sand at 1.4GHz is about 1.5 dBm^{-1} , whereas for sand with a 30% volumetric moisture content the value increases to about 45 dBm^{-1} .

Finally, a comment on transmission through heterogeneous dielectric media is important in the context of wireless communication at microwave frequencies. A forest essentially is a dielectric mixture of air (free space) and vegetative matter. The latter consists of trunks, branches, twigs and foliage. As a result, when a wave-front is incident on a forest stand there will be reflections at each air-leaf, air-branch, etc. interface resulting in scattering and thus loss of signal in the transmitted direction. In addition, the energy that does penetrate the leaves and branches (and trunks) themselves will be subject to attenuation through absorption. The situation is thus particularly complex to understand theoretically, although certain fairly crude models of the electromagnetic scattering properties of forests have been devised. It is perhaps best in practice to seek empirical models based on experiments.

The situation is further complicated, for dense forest stands, by the possibility for waves to be guided above the canopy, and for diffraction around sharp tree transitions to occur.

It can be assumed that losses in forests will be worse at higher frequencies.

7.7 Propagation in Tunnels

Unless special measures are taken to carry radio waves inside, signals can fade as a vehicle carrying a receiver moves into a tunnel, especially if it has highly conducting walls. Long wavelength signals are more likely to be affected; those with shorter wavelengths can often propagate effectively inside a tunnel. The effect is most stark in the contrast between broadcast AM reception, which fades badly in most tunnels, and FM radio reception which does not.

Analysis of the situation can be complex, particularly if the tunnel walls, roof and floor are irregular in shape and of mixed materials. In addition, the presence of other

(conducting) vehicles in the tunnel can complicate an understanding of what occurs. To a good first approximation, however, we can use classical waveguide theory to form an impression of what is happening. That is based on the assumption that the tunnel has ideally conducting walls and constant cross-sectional shape. Given that the walls in many tunnels would be steel reinforced we can make that assumption in most cases.

Understanding waveguide behaviour requires a background in Maxwell's equations and the wave equation and is beyond the current treatment. It is sufficient for our purposes to note that enclosed guiding media, such as those simple examples shown in Fig. 7.7, will only allow the passage of signals with wavelengths above a certain minimum, determined by the dimensions of the medium. The wave actually propagates in a number of modes, each of which requires the electric field vector of the wave to terminate on the walls of the medium. In order that the conducting walls not short circuit the positive and negative terminations of the field vectors the circumference of the medium must be of the order of a wavelength or greater. More specifically, the lowest wavelength that can be supported and thus propagate through the tunnel is given by

$$f_c = \frac{150}{b} \text{MHz for a tunnel of rectangular shape} \quad (7.16a)$$

$$f_c = \frac{87.9}{r} \text{MHz for a tunnel with circular cross section} \quad (7.16b)$$

Those formulas show that a rectangular tunnel of 10 m width will not support propagation below 15 MHz. A circular tunnel of 5 m radius will not support propagation below 17.6 MHz. It is thus clear why AM radio at a carrier frequency of about 1 MHz will not carry in a tunnel (with conducting walls) whereas FM, with a carrier close to 100 MHz, will propagate.

Waveguide theory has been developed extensively and it is possible to find the lowest operating frequency for guides with other regular cross-sectional shapes. It is also possible to analyse guides with walls that are not ideally conducting.

During the 1960s and 1970s there was enormous interest in waveguides as means for long distance transmission of broadband signals. That was just before the advent

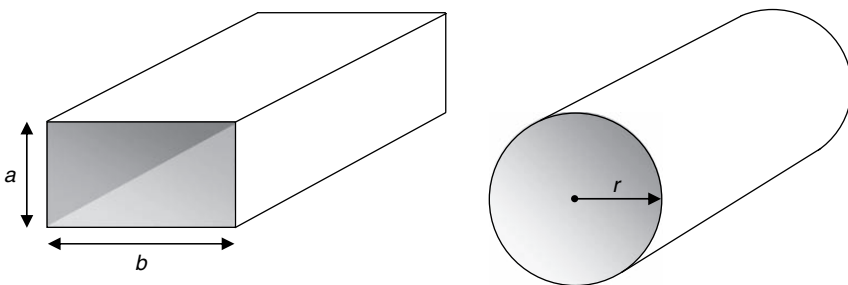


Fig. 7.7 Representation of tunnels as rectangular or circular waveguides

of the practical optical fibre which, as well as being able to operate over extensive bandwidths, is much simpler to manufacture and much less expensive than waveguide communications systems would have been. Although the use of waveguides as trunk communications channels disappeared they are still used as feeders of microwave energy, particularly at high power levels, in radar systems.

Even though a signal will not carry in a tunnel below the cut-off frequency there will be some energy penetration near the openings. Ignoring any reflections caused by the discontinuity from free space to the tunnel, a signal below the cut-off frequency, in attempting to penetrate the tunnel, will decay exponentially with distance along the tunnel. Such a wave is called *evanescent* and is described by the attenuation coefficient (for a waveguide of any regular cross section)

$$\alpha = \frac{2\pi}{\lambda_o} \sqrt{\left(\frac{\lambda_o}{\lambda_c}\right)^2 - 1} \quad \text{Npm}^{-1} \quad (7.17)$$

In this expression λ_o is the free space wavelength of the radiation and λ_c is the cut-off wavelength associated with the cut-off frequency of (7.16). For example, the 10 m wide tunnel considered above will at 1 MHz have an evanescent attenuation of 0.021 Npm^{-1} , or 0.18 dBm^{-1} . Thus the signal will halve at about 17 m from the entrance to the tunnel.

Problems

7.1. An aircraft flying over the sea transmits a signal vertically downwards at 1 MHz, leading to a power density just above the surface of 1 mWm^{-2} . What is the electric field strength just above the surface? By computing the transmission coefficient at the air sea interface and the attenuation constant in the sea water determine the electric field strength 1 m below the surface. Assume $\epsilon_r = 81$, $\sigma = 4 \text{ Sm}^{-1}$.

7.2. Horizontally polarised radiation at 1 GHz is incident from air onto a smooth dry soil surface at an incidence angle of 45° . Compute the change in reflection coefficient that would occur if the dry surface were flooded and thus just covered with water. Make reasonable assumptions about the dielectric constants of the surfaces, but neglect losses.

7.3. Is a slightly rough surface more likely to appear smooth at vertical incidence or at glancing incidence?

7.4. Imaging radar systems map the landscape by irradiating it with microwave radiation and measuring the strengths of the echoes received. At wavelengths of the order of 10 cm and longer forests have an interesting response in that most foliage is moderately transparent and the manner in which the radiation is scattered involves reflection from tree trunks onto the ground and then specular reflection from the ground back to the radar; together the trunk and ground act like a corner reflector. It

has been observed that floods under forest canopies can increase the received echo strength by about 6 dB. Why?

7.5. Waveguides such as those illustrated as tunnels in Fig. 7.7 support a number of so-called modes of propagation each of which has its own cut-off frequency similar to those of (7.16) which are for the lowest frequency propagating modes. For the lowest mode in a rectangular waveguide the relationship between the phase constant in the guide, the operating frequency and guide width b is

$$\beta = \sqrt{\omega^2 \mu \epsilon - \left(\frac{\pi}{b}\right)^2}$$

Construct an $\omega - \beta$ diagram for this mode that demonstrates its cut-off property and the fact that it is dispersive, assuming it is filled with free space.

7.6. Equations (7.10) and (7.11) are special cases of reflection coefficients for waves incident from free space onto a higher dielectric constant medium. The more general forms for reflection from interfaces between any two lossless dielectric media are

$$\rho_{\perp} = \frac{\cos \theta_i - \sqrt{\frac{\epsilon_{r2}}{\epsilon_{r1}} - \sin^2 \theta_i}}{\cos \theta_i + \sqrt{\frac{\epsilon_{r2}}{\epsilon_{r1}} - \sin^2 \theta_i}}$$

$$\rho_{//} = \frac{-\frac{\epsilon_{r2}}{\epsilon_{r1}} \cos \theta_i + \sqrt{\frac{\epsilon_{r2}}{\epsilon_{r1}} - \sin^2 \theta_i}}{\frac{\epsilon_{r2}}{\epsilon_{r1}} \cos \theta_i + \sqrt{\frac{\epsilon_{r2}}{\epsilon_{r1}} - \sin^2 \theta_i}}$$

Using these expressions demonstrated that total internal reflection is possible for a wave incident onto an interface when propagating from a higher to a lower dielectric constant medium.

7.7. Using the expression for β in Problem 7.5 demonstrate that a waveguide is non-dispersive when the frequency is well above cut-off.

7.8. What does the omega beta diagram for a lossless medium look like?

Appendix A

A Simple Introduction to Antennas

A.1 Introduction: Radiation Resistance and Radiation Patterns

An antenna is a transitional device, or transducer, that forms an interface for energy traveling between a circuit and free space, as depicted in Fig. A.1. It is reciprocal in the sense that it can transfer energy from the circuit to free space (transmission) and from free space to a circuit (reception). We can represent both the circuit and the antenna by their Thévenin equivalents as shown.

In engineering we represent the conversion of energy from electrical to some other non-recoverable form by a resistive load, since the resistor is an element that absorbs real power. We do the same for antennas; the radiation resistance R_r , shown in Fig. A.1 models how much power is taken from the transmitter circuit and is radiated non-recoverably into free space. Alternately it is the source resistance of the antenna when it receives a signal, in which case the antenna model will also include a generator to represent the source of energy it is providing to the receiver circuit.

A reactive component X_r of the antenna model is seen in Fig. A.1. Ideally that should be zero because it signifies that energy reflects back from the antenna to the circuit. In the design of an antenna, one object is to make its radiation impedance real because then there can be a smooth transition from the circuit to free space. Ideally the radiation resistance of the antenna should match the output resistance of the circuit, and the circuit's output reactance should be zero, so that maximum power transfer can occur without reflection. In practice it is often difficult to achieve such a match so tuning circuits are sometimes employed between the circuit and the antenna.

Now consider the construction of the antenna. If there were no antenna and the circuit terminated in an open circuited transmission line, then theoretically all the power from the transmitter would be reflected backwards along the line. If the end of the line were flared, or even terminated in a dipole arrangement as shown in Fig. A.2, then a sizable proportion of the energy traveling forward along the transmission line from the transmitter circuit will detach and radiate into free space.

The dipole arrangement shown in Fig. A.2 is a very common form of antenna, particularly if its length from tip to tip is equivalent to a half wavelength of the signal being radiated. We can deduce some of its properties qualitatively, particularly in relation to the directions in which it radiates. For example, if we walked around a

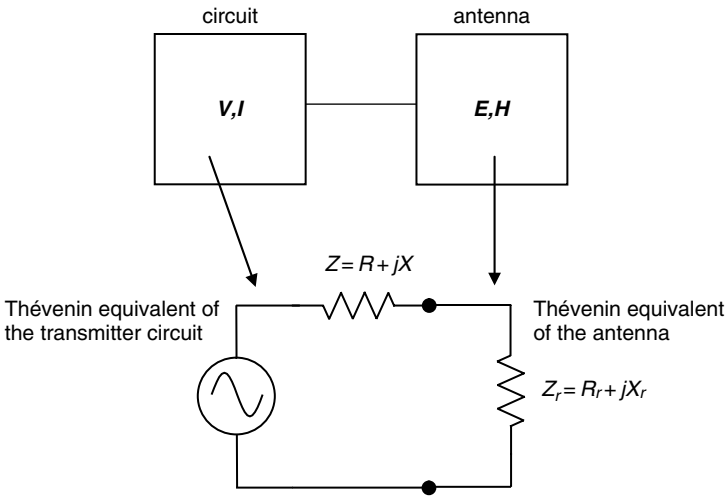


Fig. A.1 The antenna as an interface between a circuit and free space, along with their Thévenin equivalent circuits; the subscript r on the antenna model stands for “radiation”

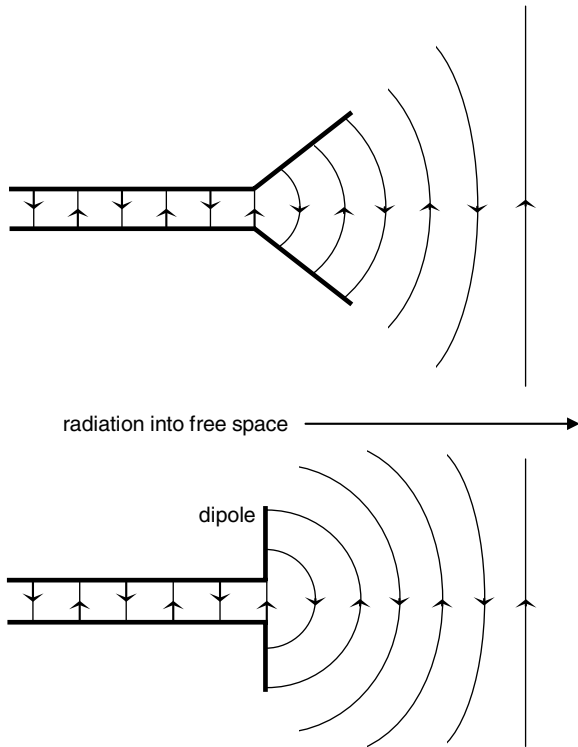
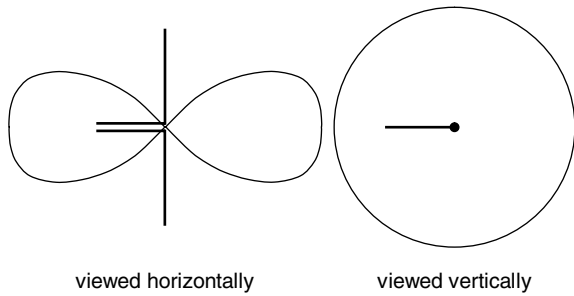


Fig. A.2 Use of a flared horn or a dipole as a transition from a circuit to free space

Fig. A.3 The radiation pattern of a dipole antenna



vertically deployed dipole antenna in the horizontal plane it would look no different when viewed from any angle. Therefore we would conclude that its radiating properties will not vary with angle around the dipole. If however we moved around it vertically its aspect would change from its appearance in Fig. A.2 to a single point when viewed directly from above or below. We can conclude therefore that its radiating properties will vary with vertical angle and, indeed, it may not radiate at all in the vertical direction. That is in fact the case for a dipole; it has a *radiation pattern* or *polar pattern* of the form of a doughnut as depicted in Fig. A.3.

As would be imagined, the radiation pattern of an antenna can be quite complicated. That leads to a number of definitions that are helpful in describing and antenna’s properties. Figure A.4 shows a typical pattern in one plane, remembering that the full pattern will be a three dimensional figure. It is plotted in Cartesian rather than polar coordinates, which is often the case in practice.

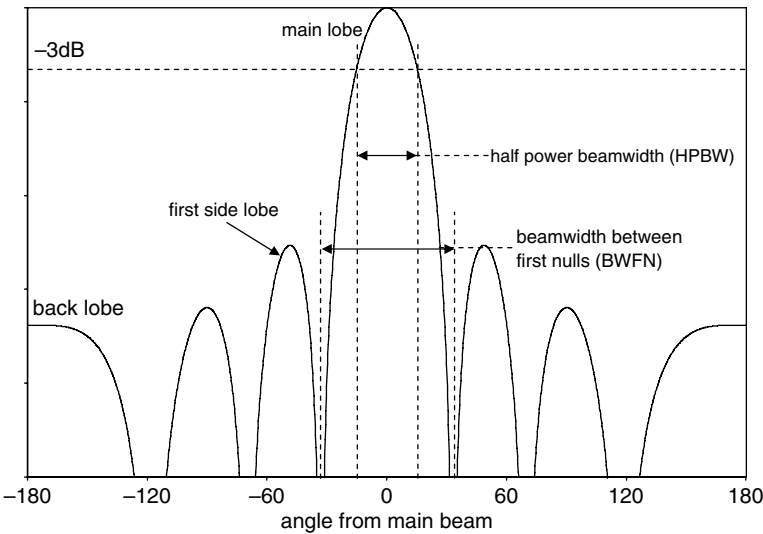


Fig. A.4 Radiation pattern of a fictional antenna plotted in Cartesian coordinates

A.2 The Directivity and Gain of an Antenna

It is now important to describe the antenna quantitatively. We commence with the definition of its *directivity* which, as the name implies, is a measure of how much it concentrates the energy in a certain direction. This is a three dimensional concept since an antenna, in principle, radiates in all directions. Geometric definitions for the derivation of an expression for directivity are given in Fig. A.5, from which it can be seen that the power available over an incremental area α at distance r from the antenna is given by

$$P = p\alpha = \frac{P_t \alpha}{4\pi r^2} = \frac{P_t \Omega}{4\pi}$$

giving as the power available per unit of solid angle – also known as *angular power density*

$$p = \frac{P}{\Omega} = \frac{P_t}{4\pi} = p(\theta, \phi) \text{ Wsr}^{-1} \tag{A.1}$$

Equation (A.1) is an algebraic expression for the three dimensional polar pattern. To proceed to a definition of directivity it is normalized with respect to its maximum to give

$$p_n(\theta, \phi) = \frac{p(\theta, \phi)}{p_{\max}(\theta, \phi)}$$

Integrating this dimensionless, normalized quantity over three dimensions we end up with the angular quantity

$$\Omega_A = \iint_{4\pi} p_n(\theta, \phi) d\theta d\phi \tag{A.2}$$

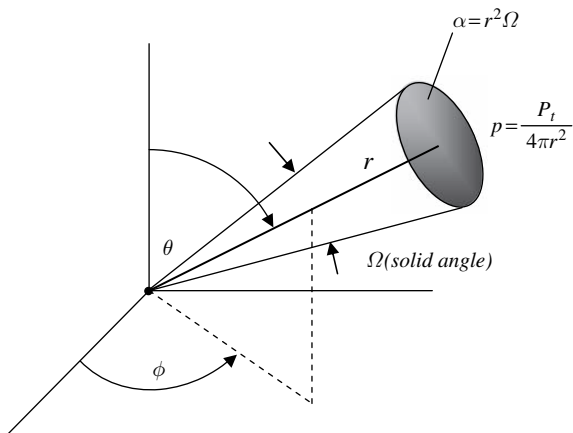


Fig. A.5 Coordinates and definitions for calculating the directivity of an antenna

which is called the solid beam angle of the antenna. It is the three dimensional angle through which all the power of the antenna would be transmitted if its polar pattern were uniform over that angle as depicted in Fig. A.6.

We now define the directivity of the antenna as

$$D = \frac{4\pi}{\Omega_A} \tag{A.3}$$

which can be approximated

$$D = \frac{4\pi}{\theta_{HP}\phi_{HP}} \text{ with angles in radians, or}$$

$$D = \frac{41253}{\theta_{HP}\phi_{HP}} \text{ with angles in degrees.}$$

The *gain*, G , of an antenna is closely related to its directivity. They differ only through the efficiency, k , of the antenna, which accounts for ohmic losses in the antenna material. Thus

$$G = kD \quad 0 < k < 1$$

Although (A.3) allows the directivity and thus gain of an antenna to be derived theoretically it is more usual to measure an antenna's gain. That is done relative to a reference antenna. One reference, even though not experimentally practical, is the isotropic radiator, which has a directivity and thus gain of unity (see Sect. 1.2). In principle, the gain determined from (A.3) is with respect to isotropic. It is more usually expressed in decibels with respect to isotropic, and written as

$$G = 10 \log \frac{4\pi}{\Omega_A} = 10 \log 4\pi - 10 \log(\theta_{HP}\phi_{HP}) = 11 - 10 \log(\theta_{HP}\phi_{HP}) \text{ dBi}$$

assuming the antenna is ideally efficient.

It is possible to calculate the gain of the dipole antenna when its length is equal to half a wavelength. It is, of course, also possible to construct a half wave dipole,

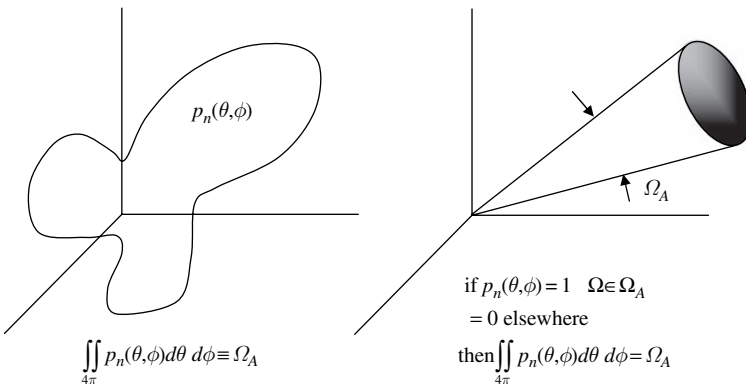


Fig. A.6 Demonstrating the definition of the angular beamwidth of an antenna

so that it can be used as a reference antenna when measuring the gain of another antenna. Such a measurement is undertaken by transmitting to the antenna of interest and measuring the received signal, and in fact the full radiation pattern. The experimental antenna is then replaced by the reference dipole and the measurement repeated so that the measurement of the experimental antenna can be normalized to the dipole. The measured gain is then expressed as dB with respect to the half wave dipole, written as $\text{dB}_{\lambda/2}$

The calculated gain of the dipole is 2.16 dBi; thus we have

$$G \text{ dBi} = G \text{ dB}_{\lambda/2} + 2.16\text{dB}$$

A.3 The Aperture of an Antenna

Derivation of expressions for the aperture of an antenna requires a field theory analysis. For some antennas, such as a parabolic dish reflector, the aperture concept is straightforward and, provided the dish diameter is much larger than a wavelength, the aperture is related directly to the area presented to an incoming wave front. For other antennas, such as linear structures and even an isotropic radiator, the aperture concept is less straightforward but can, if we know its gain, be determined from

$$A = \frac{\lambda^2 G}{4\pi} \text{ m}^2 \quad (\text{A.4})$$

If the antenna has a physical aperture, such as a parabolic reflector, we introduce the concept of aperture efficiency to account for the difference between the physical and electromagnetic apertures:

$$A = k_{\text{aperture}} A_{\text{physical}} \quad 0 < k_{\text{aperture}} < 1$$

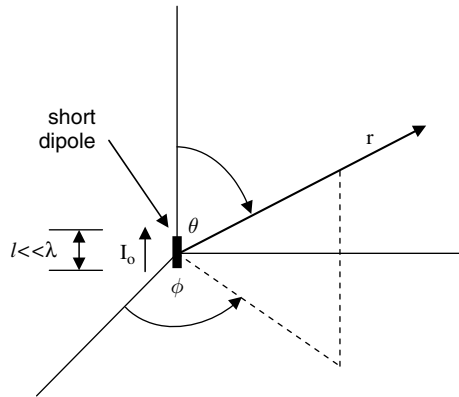
A.4 Radiated Fields

A treatment of the fields radiated by an antenna requires a field theory treatment and is beyond this coverage. It is, however, useful to examine well-known expressions for the fields produced by a so-called short dipole because the fields generated by other antennas can be derived from the short dipole results; it also allows us to understand the concept of near and far fields. Figure A.7 shows the geometry of a short dipole in which distance and direction out from the antenna is described by the radial coordinate r .

If the short dipole is carrying a sinusoidal current

$$I_0 e^{j\omega t}$$

Fig. A.7 The short dipole



and is so short that there is no distribution of current along its length at any time, then the complete set of fields generated about the short dipole is

$$E_r = \frac{I_0 l e^{j(\omega t - \beta r)} \cos \theta}{2\pi \epsilon_0} \left(\frac{1}{cr^2} + \frac{1}{j\omega r^3} \right) \tag{A.5a}$$

$$E_\theta = \frac{I_0 l e^{j(\omega t - \beta r)} \sin \theta}{4\pi \epsilon_0} \left(\frac{j\omega}{c^2 r} + \frac{1}{cr^2} + \frac{1}{j\omega r^3} \right) \tag{A.5b}$$

$$H_\phi = \frac{I_0 l e^{j(\omega t - \beta r)} \sin \theta}{4\pi} \left(\frac{j\omega}{cr} + \frac{1}{r^2} \right) \tag{A.5c}$$

In other words there are transverse components (θ, ϕ) of the magnetic and electric fields. There is also a radial electric field component (r) – i.e. in the direction of propagation. Note however it has a stronger inverse dependence on distance than the transverse components so that if the distance is sufficiently large it disappears and the transverse components themselves become just inverse distance dependent. This is demonstrated by letting r go large in (A.5a–c) to give

$$E_r = 0 \tag{A.6a}$$

$$E_\theta = \frac{j\omega I_0 l e^{j(\omega t - \beta r)} \sin \theta}{4\pi \epsilon_0 c^2 r} \tag{A.6b}$$

$$H_\phi = \frac{j\omega I_0 l e^{j(\omega t - \beta r)} \sin \theta}{4\pi cr} \tag{A.6c}$$

Thus for large distances the wave is TEM – i.e. transverse electromagnetic. Equations (A.6a–c) describe the so-called *far field* of the antenna. The far fields are inverse distance dependent and the treatment in this book, based on simple power and power density relationships, is valid. In contrast, closer to the antenna (A.5a–c) are needed to describe the field. That is called the *near field* of the antenna. The transition from near to far field is said to occur when the inverse distance terms in

(A.5b,c) are equal to the inverse distance squared terms, assuming that any inverse cubic terms are then negligible. Therefore the near field/far field transition is when

$$\left| \frac{\omega}{cr} \right| = \left| \frac{1}{r^2} \right|$$

which gives

$$r \approx \frac{\lambda}{6}$$

It is of interest to note from (A.6b,c) that in the far field

$$\left| \frac{E_\theta}{H_\phi} \right| = \frac{1}{\epsilon_0 c} = \frac{\sqrt{\mu_0 \epsilon_0}}{\epsilon_0} = Z_0$$

the free space impedance, as would be expected.

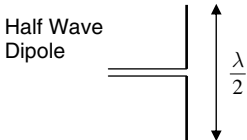
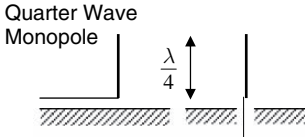
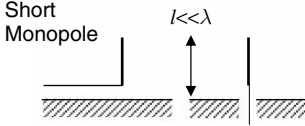
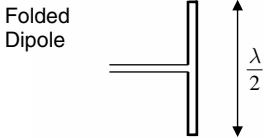
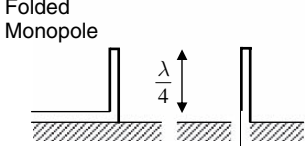
	Gain	Radiation Impedance
 <p>Half Wave Dipole</p>	1.64 (2.15dBi)	73 + j42.5 70 + j0 <i>if antenna slightly shorter</i>
 <p>Quarter Wave Monopole</p>	3.3 (5.19dBi)	36.5 + j21.3
 <p>Short Monopole</p>	3 (4.77dBi)	$10\pi \left(\frac{l}{\lambda}\right)^2 - jX$ (large)
 <p>Folded Dipole</p>	1.64 (2.15dBi)	292
 <p>Folded Monopole</p>	3.3 (5.19dBi)	146

Fig. A.8 Some common linear antennas

A.5 Some Typical Antennas

Figure A.8 shows a number of simple antennas and their characteristics. Figure A.9 shows two common, compound antennas that are built up from combinations of active antennas, of the types shown in Fig. A.8, and passive linear elements.

The folded antennas shown in Fig. A.8 tend to have slightly broader bandwidths than their unfolded counterparts and are often used, particularly the dipole, in more complex structures such as the Yagi-Uda array illustrated in Fig. A.9. The short monopole in Fig. A.8 is commonly used as an AM receiving antenna on motor vehicles.

The log periodic antenna shown in Fig. A.9 is used when operation is necessary over a wide band of frequencies. Although it is more complex in construction than the Yagi, its wide operating bandwidth makes it attractive in many applications.

Figure A.10 shows a number of aperture and slot antennas, along with a bi-cone. Aperture reflectors tend to be used when the wavelength is much smaller than the diameter of the reflector, so they behave somewhat similar to optical reflectors of the same type.

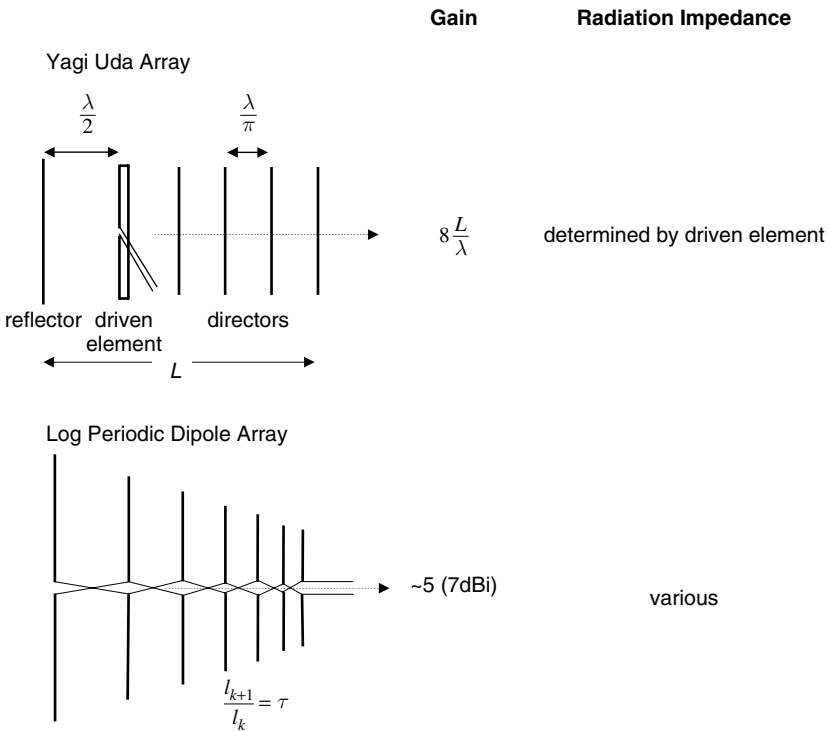


Fig. A.9 Some common compound antennas; the antenna lengths and spacings for the log periodic antenna are in the constant ratio τ as indicated for length

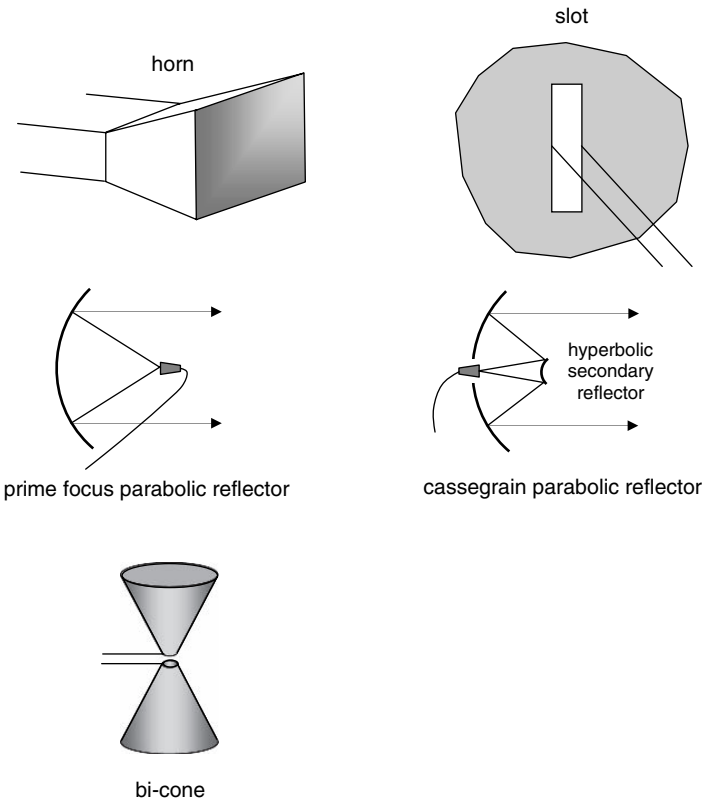


Fig. A.10 Aperture, slot and bi-conical antennas; the bi-cone is broad band and omni-directional in the horizontal plane

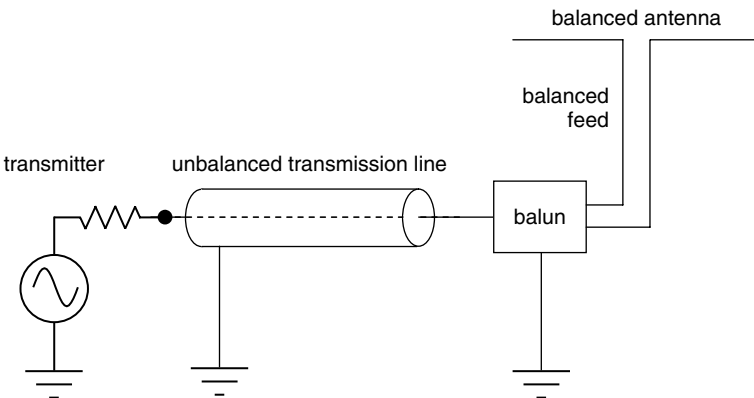


Fig. A.11 Demonstrating the use of a balun to provide the matched transition between an unbalance transmission line and a balanced antenna

A.6 Baluns

With the exception of monopoles, the other antennas in Figs. A.8 and A.9 require balanced feeds. In other words they need to be fed by transmission lines that have neither conductor at earth potential. Yet many of the feed lines in practice are coaxial cables that clearly have one of their conductors – the braid – at earth potential. Coaxial cables are also compatible with many transmitter output circuits that are also unbalanced as noted by the manner in which the Thévenin equivalent is depicted in Fig. A.1. To render the unbalanced transmission line compatible with a balanced antenna a device referred to as a balun is employed, as illustrated in Fig. A.11. Short for *balanced-unbalanced*, this device can be constructed in several forms, each of which not only has to perform the unbalanced-to-balanced transformation but also has to match impedances for maximum power flow and to minimise reflections.

There are many forms of balun, the simplest of which is a transformer. For narrow band operation a simple balun can be constructed from sections of transmission line.

Appendix B

The Use of Decibels in Communications Engineering

Logarithms have two major benefits: they readily summarise numbers that extend over a large range and they simplify multiplication. As a consequence, the decibel (dB), which is defined using base 10 logarithms, is widely used as a convenient measure in many branches of engineering, but especially in communications. Although it can be used with signals generally, it is principally defined in terms of power (or power density). More precisely, the decibel (dB) is defined on the basis of a *reference* power:

$$10 \log_{10} \frac{P}{P_{ref}} = x \text{ dB}$$

We say that P is x dB larger than P_{ref} . For example

If $P =$	then x is
$2P_{ref}$	3 dB wrt P_{ref}
$10P_{ref}$	10 dB wrt P_{ref}
$100P_{ref}$	20 dB wrt P_{ref}
$0.5P_{ref}$	-3 dB wrt P_{ref}
$0.1P_{ref}$	-10 dB wrt P_{ref}

Many factors have easily constructed dB equivalents. For example

$$200 = 2 \times 100 \rightarrow 3 \text{ dB} + 20 \text{ dB} = 23 \text{ dB},$$

as a result of the additive property of logarithms.

Similarly

$$\begin{aligned} 17 \text{ dB} &= 20 \text{ dB} - 3 \text{ dB} \rightarrow 100 \div 2 = 50, \\ 36 \text{ dB} &= 30 \text{ dB} + 6 \text{ dB} \rightarrow 1000 \times 4 = 4000. \end{aligned}$$

In telecommunications, two common values of P_{ref} are used. The dBs are then given special symbols that imply absolute, as against relative, quantities.

If $P_{\text{ref}} = 1 \text{ W}$, then we use dBW.

If $P_{\text{ref}} = 1 \text{ mW}$, then we use dBm.

Thus we can see the following equivalences:

$$\begin{array}{ll}
 17 \text{ dBm} \rightarrow 50 \text{ mW} & 23 \text{ dBW} \rightarrow 200 \text{ W} \\
 3 \text{ dBm} \rightarrow 2 \text{ mW} & 10 \text{ dBW} \rightarrow 10 \text{ W} \\
 30 \text{ dBm} \rightarrow 1 \text{ W} & 20 \text{ dBW} \rightarrow 100 \text{ W} \\
 0 \text{ dBm} \rightarrow 1 \text{ mW} & 0 \text{ dBW} \rightarrow 1 \text{ W} \\
 -20 \text{ dBm} \rightarrow 10 \mu\text{W} & -40 \text{ dBW} \rightarrow 100 \mu\text{W}
 \end{array}$$

Decibels can also be used with voltages, but the definition still rests upon power. For example

$$10 \log \frac{P}{P_{\text{ref}}} = 10 \log_{10} \frac{V^2}{V_{\text{ref}}^2} = 20 \log \frac{V}{V_{\text{ref}}} = x \text{ dB}$$

So that if $V = 2 V_{\text{ref}}$, then $x = 6 \text{ dB}$.

Appendix C

The Dielectric Constant of an Ionospheric Layer

Equation (3.1) notes that the refractive index of an ionospheric layer is given by

$$n = \sqrt{1 - \frac{81N}{f^2}}$$

We derive that expression below, following the approach of D.J. Angelakos and T.E. Everhart, *Microwave Communication*, McGraw-Hill, New York, 1968.

An ionised region of the atmosphere, such as one of the layers of the ionosphere, will be composed of free ions, electrons and neutral molecules. We assume that the ions, because of their mass, do not respond as well to the passage of an electromagnetic field as the electrons and thus have little effect on it. We will therefore concentrate our attention just on the free electrons, which we assume to be present with density N electrons per cubic metre.

We also assume that the earth's magnetic field has no effect, and that the collisions that occur between electrons and neutral atmospheric constituent molecules can be neglected. Those collisions are significant if we are interested in the attenuation of a wave in transmission through the atmosphere; we mention that below, after the derivation of refractive index.

The response of an individual electron of mass m to an applied electric field $|\mathbf{E}| = E_m \cos \omega t \text{ Vm}^{-1}$ (resulting from the passage of a radio wave) is given from Newton's law

$$F = ma$$

If the charge on the electron is e and its velocity is v then this last expression can be written

$$eE_m \cos \omega t = m \frac{dv}{dt}$$

which gives for the electron velocity

$$v = \frac{eE_m}{\omega m} \sin \omega t.$$

This movement of electrons gives rise to a conduction current described by the transverse areal current density

$$\begin{aligned}\mathbf{j}_{\text{cond}} &= veN \text{ Am}^{-2} \\ &= \frac{e^2 NE_m}{\omega m} \sin \omega t\end{aligned}\quad (\text{C.1})$$

There will also be a displacement areal current density as a result of the dielectric behaviour of the medium, found from

$$\mathbf{j}_{\text{dis}} = \frac{d\mathbf{D}}{dt} = \varepsilon \frac{d\mathbf{E}}{dt} \quad (\text{C.2})$$

in which \mathbf{D} is the electric displacement vector and ε is the permittivity of the medium.

For a plasma as dilute as an ionospheric layer $\varepsilon = \varepsilon_o$, so that

$$|\mathbf{j}_{\text{dis}}| = \varepsilon_o \frac{d|\mathbf{E}|}{dt} = -\varepsilon_o \omega E_m \sin \omega t$$

Thus the magnitude of the total current in the layer induced by the passage of a radio wave is

$$|\mathbf{j}_{\text{total}}| = |\mathbf{j}_{\text{cond}} + \mathbf{j}_{\text{dis}}| = \left(\frac{e^2 N}{\omega m} - \varepsilon_o \omega \right) E_m \sin \omega t \quad (\text{C.3})$$

Even though there are free electrons present, we now regard the layer as behaving entirely as a dielectric with permittivity ε – i.e. as though there were no free electrons. The displacement current density is then given just by (C.2). If we call that an *effective* displacement current density and equate it to the actual current density given by (C.3) we have using (C.1)

$$|\mathbf{j}_{\text{effective,dis}}| = -\varepsilon \omega E_m \sin \omega t = |\mathbf{j}_{\text{cond}} + \mathbf{j}_{\text{dis}}|$$

so that

$$-\varepsilon \omega = \left(\frac{e^2 N}{\omega m} - \varepsilon_o \omega \right)$$

or

$$\varepsilon = \varepsilon_o \left(1 - \frac{\omega_p^2}{\omega^2} \right) \quad (\text{C.4})$$

in which

$$\omega_p = \left(\frac{e^2 N}{m \varepsilon_o} \right)^{1/2} \quad (\text{C.5})$$

is called the plasma frequency of the region with electron density N .

Since

$$\begin{aligned}e &= 1.6 \times 10^{-19} \text{C} \\m &= 9.11 \times 10^{-31} \text{kg} \\ \epsilon_0 &= 8.85 \text{ pFm}^{-1}\end{aligned}$$

then

$$\omega_p^2 = 3175N$$

Thus (C.4) becomes

$$\epsilon = \epsilon_0 \left(1 - \frac{3175N}{\omega^2}\right) = \epsilon_0 \left(1 - \frac{81N}{f^2}\right) \quad (\text{C.6})$$

from which we recognise that the equivalent dielectric constant (or relative permittivity) of the region is

$$\epsilon_{\text{rel}} = \left(1 - \frac{81N}{f^2}\right)$$

Thus the refractive index of a region of the ionosphere of electron density N and frequency f is

$$n = \sqrt{1 - \frac{81N}{f^2}} \quad (\text{C.7})$$

Recall that this expression, and (C.6), was derived by ignoring losses resulting from electron-neutral collisions. If they were included Rohan¹ shows that (C.6) would be

$$\epsilon = \epsilon_0 \left(1 - \frac{e^2N}{m\epsilon_0(\omega^2 + \nu^2)}\right) \quad (\text{C.8})$$

in which ν is the collision frequency of the electrons and neutrals. While the collision frequency is very high in the lower atmosphere because of the neutral density, it is of the order of 1000 or less at the height of the upper ionospheric layers. As a consequence, at the sorts of frequencies normally associated with sky wave propagation $\omega^2 \gg \nu^2$, so that (C.8) reduces to (C.6).

Electron-neutral collisions give rise to losses in the ionosphere, particularly at the lower levels; their effect can be characterised by an equivalent conductivity from which an attenuation constant can be derived. Again, following Rohan, the conductivity of a region of ionisation is

$$\sigma = \frac{e^2N\nu}{m(\omega^2 + \nu^2)}$$

¹ P. Rohan, *Introduction to Electromagnetic Wave Propagation*, Artech, Boston, 1991.

The attenuation constant is then

$$\alpha = \frac{60\pi\sigma}{n} = \frac{60\pi e^2 N \nu}{nm(\omega^2 + \nu^2)}$$

in which n is refractive index. Thus the attenuation of a layer decreases with an increase in operating frequency.

To obtain an idea of the levels of attenuation likely to be encountered by a wave travelling through the D region (above its critical frequency) suppose we choose typical values of $\nu = 10^7 \text{ s}^{-1}$, $N = 10^8$ electrons m^{-3} and $f = 1 \text{ MHz}$ (ie the AM broadcast band). After substituting we have

$$\begin{aligned}\alpha &= 3.8 \times 10^{-5} \text{ Npm}^{-1} \\ &= 0.33 \text{ dBkm}^{-1}\end{aligned}$$

Thus if the layer were equivalently 25 km thick at that effective electron density then the total attenuation at 1 MHz would be 8.25 dB at vertical incidence and considerably more at oblique incidence. In the evening such a high level of attenuation does not occur because of the absence of the D region. As a consequence it is possible to receive distant AM stations in the evening that are not available during daylight hours.

Index

- Angular power density, 110
- Antenna
 - aperture, 2, 112
 - aperture efficiency, 112
 - back lobe, 109
 - beamwidth between first nulls, 109
 - bi-cone, 116
 - cassegrain reflector, 116
 - directivity, 110
 - folded dipole, 114
 - folded monopole, 114
 - gain, 2, 111
 - half power beamwidth, 109
 - half wave dipole, 107, 114
 - horn, 116
 - Log periodic array, 115
 - main lobe, 109
 - polar pattern, 109
 - prime focus parabolic reflector, 116
 - quarter wave monopole, 114
 - radiation impedance, 107
 - radiation pattern, 109
 - radiation resistance, 107
 - short dipole, 112
 - short monopole, 13, 14, 114
 - side lobe, 109
 - slot, 116
 - Yagi-Uda array, 115
- Atmosphere
 - modified refractive index, 52
 - refractive index, 48
 - standard, 48
- Atmospheric attenuation, 54
- Attenuation constant, 88
 - in waveguide, 103
- Automatic gain control, 84
- Balun, 117
- Band designators, 7
 - microwave, 8
- Brewster angle, 99
- Cellular radio systems, 79
- Co-channel interference, 80
- Conductivity, 15, 19
- Critical frequency, 25, 30
- Dielectric constant, 5, 88
 - complex, 20, 89
 - effect of moisture content, 94
- Diffraction, 45
 - gain, 47
- Direct ray, 8, 39
- Diversity, 72
 - antenna, 74
 - frequency, 72
 - space, 73
 - time, 73
- Ducting, 52
- Earth gravitational constant, 76
- Effective earth radius, 44, 51
- Effective isotropically radiated power, 12
- Electron density, 23
- Escape ray, 29
- Extra-ordinary ray, 35
- Fade margin, 71
- Far field, 113
- Free space path loss, 7
- Frequency re-use, 11, 79
- Fresnel zone, 48
- Friis' Noise Formula, 63

- G-to-T ratio, 78
- Geostationary orbit, 75
- Ground current, 9, 14
- Ground reflected ray, 8, 39
- Group velocity, 33, 90

- Huygens' principle, 45

- Impedance of free space, 4, 97
- Inverse distance law, 4
- Ionogram, 30
- Ionosonde, 30
- Ionosphere, 8
 - D region, 18, 24
 - daytime, 24
 - E layer, 24
 - F layer, 25
 - F1 layer, 24
 - F2 layer, 24
 - night time, 25
 - refractive index, 26
 - sporadic E layer, 25
- Ionospheric sounding, 30
- Ionospheric wave, 10, 23
- Isotropic radiator, 2

- Lambertian scattering, 100
- Loss
 - atmospheric absorption, 7
 - dielectric, 102
 - diffraction, 7, 46
 - free space path loss, 7
 - rainfall, 7
 - refraction, 7
- Loss tangent, 94

- Maximum usable frequency, 35
- MUF factor, 35
- Multi-path, 39, 72

- Near field, 113
- Neper, 88
- Noise, 57
 - additive, 57
 - atmospheric, 72
 - environmental, 58
 - galactic, 72
 - Johnson, 58
 - multiplicative, 57
 - shot, 58
 - thermal, 58
 - voltage, 66
- Noise bandwidth, 59
- Noise figure, 61

- Noise temperature, 59
 - cascaded two ports, 63
 - equivalent input noise temperature, 60
 - equivalent output noise temperature, 61
 - two port, 59

- Omega-beta diagram, 90
- Optimum working frequency, 36

- Passive reflectors, 73
- Permeability
 - absolute, 5
 - relative, 5, 87
- Permittivity
 - absolute, 5
 - complex, 94
 - relative, 5, 123
- Phase constant, 88
 - in waveguide, 105
- Phase velocity, 32, 89
- Plasma frequency, 122
- Polarisation
 - circular, 3
 - dielectric, 92
 - elliptical, 3
 - horizontal, 3
 - parallel, 97
 - perpendicular, 97
 - vertical, 3
- Power density, 2
- Poynting vector, 4
- Propagation constant, 19, 88

- Quasi-conductor, 91

- Radar, 12
 - primary, 12
 - secondary, 12
- Radar cross section, 12, 74
- Rainfall attenuation, 53
- Rayleigh criterion, 99
- Rayleigh distribution, 84
- Receiver figure of merit, 78
- Reflection coefficient, 40, 97, 98, 105
- Refraction
 - atmospheric, 17, 45, 48
- Refractive index, 5
- Rician distribution, 84

- Shadow zone, 45
- Signal to noise ratio, 57, 61
- Skin depth, 21

- Skip distance, 35
- Sky wave, 10, 23
 - range, 36
- Space wave, 8, 39
 - atmospheric attenuation, 54
 - ducting, 52
 - field strength, 41
 - rainfall effects, 53
 - range, 43
 - refraction, 48
- Surface wave, 10, 13
 - attenuation factor, 14
- T factors, 35
- Transmission coefficient, 97
- Troposcattering, 45
- Tunnels, 102
- Virtual height, 23, 31
- Wave impedance, 97
- Waveguide, 103
 - circular, 103
 - cut-off frequency, 104
 - evanescent attenuation, 104
 - rectangular, 103