

1

Basic Antenna and Propagation Theory

1.1 Introduction

This chapter explains the principles of antenna theory and propagation qualitatively, without going into the complex mathematical equations that are usually found in textbooks and reference books dealing with these subjects. Although this results in explanations of a simplistic nature, it enables the reader with a basic physics background to understand electromagnetic (EM) theory.

EM waves are transverse waves, unlike sound and ultrasonic waves, which are longitudinal waves that require a medium. By analogy transverse waves are like the waves one would obtain by moving a rope up and down to transmit a sine wave, whereas a longitudinal wave is like a series of train wagons being shunted along, so that each wagon moves horizontally back and forth whilst the wave also moves horizontally along the whole train.

Sonic waves cannot be transmitted in a vacuum, whereas EM waves do not require a medium and can be transmitted in a vacuum, such as deep space. This is why we can see the stars but do not hear the sound of meteors, and so on.

The EM spectrum extends from direct current (DC) that has no/zero frequency to cosmic radiation.

Above DC, we commonly encounter low frequencies from 3 Hz up to around 300 Hz used for communications with submarines.

Alternating current (AC) frequencies are used to transmit mains power. In Europe 50 Hz is used, whereas in North America and some other countries 60 Hz is more common.

The mains power on aircraft is usually 400 Hz. There are various reasons for the choice of this frequency for powering aircraft systems, one of them being that this frequency was selected as a compromise between weight, size and efficiency of the aircraft power units.

Above these frequencies, there are many applications such as communications with mines, broadcasting, and so on, until the frequency used in aircraft systems which extends into the microwave and millimetre wave region. The details of the aircraft systems are covered in Chapter 2.

We then have the higher frequencies used for radio astronomy before the infrared (IR) region where we have the heating effect. Then we move into the optical or visible region of the spectrum which is actually a very small range of frequencies.

Above these frequencies, we have the ultraviolet (UV) region which causes sunburn at its high frequency end. Much of the UV light is absorbed by the atmosphere, and thus the higher levels are only encountered in hills and mountains. It is worth noting that photochromic lenses darken on exposure to UV light and therefore do so much more quickly at higher altitudes, whereas they take a lot longer to darken in a car, where most of the UV is absorbed by glass. UV is not absorbed by plastics and thus can penetrate the Plexiglas of aircraft windows/portholes. EM waves up to these frequencies are non-ionizing.

Above ultraviolet the EM waves are ionizing and penetrate materials such as tissue, in the case of X-rays, and even concrete, in the case of gamma rays. Then we have cosmic rays that originate in space.

All these forms of radiation are in the EM spectrum; however, when engineers refer to Electromagnetics (in the case of antennas) or electromagnetic health (EMH) they usually are only referring to the EM spectrum from about 1 kHz to about 300 GHz.

The term antenna is used as a generic term for wire antennas such as dipoles as well as for aperture antennas such as horns, reflectors, and so on.

In some cases however, the term ‘antenna’ is restricted to aperture antennas in the upper radio frequencies (RFs) and microwave regions – above about 500 MHz – and the term aerial is used at the lower frequencies.

In guided circuits, such as wire and printed circuit tracks that have resistive and reactive components (inductors and capacitors), the electric and magnetic fields are 90° out of phase, but in free space the electric and magnetic fields are in phase.

1.2 Characteristics of Electromagnetic Waves

In order to understand the propagation of EM fields, all the properties of radiation have to be considered. The main phenomena that affect the propagation are reflection, refraction and diffraction.

At a macroscopic level we can think of EM radiation as travelling in straight lines, but at a microscopic level we have to consider the wave properties of the radiation as well.

If we think of the waves as spherically emanating from the source (like the waves obtained by throwing a pebble into the water), then the spherical shells are the wavefronts and the rays are the radii from the source and therefore perpendicular to the wavefronts. In Figure 1.1 the wavefronts and rays are shown in two dimensions for clarity.

Reflection and refraction can be explained by the rectilinear propagation (light travelling in straight lines) of EM waves, but diffraction can only be explained by the wave theory. Geometric optics deals with rectilinear propagation, whereas physical optics deals with light as waves. Figure 1.2 shows the rays produced in a plane normal to the cylinder axis, for an antenna located off the body of the cylinder. The stronger the colour, the greater is the intensity of the ray.

Figure 1.3 is a graphic indication of the rays reflected and diffracted off an airframe.

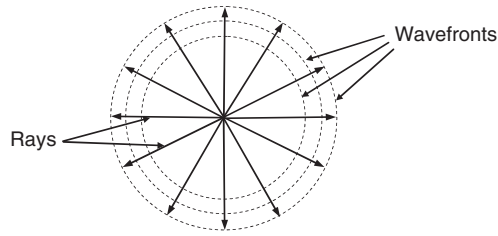


Figure 1.1 Relationship between wavefronts and rays.

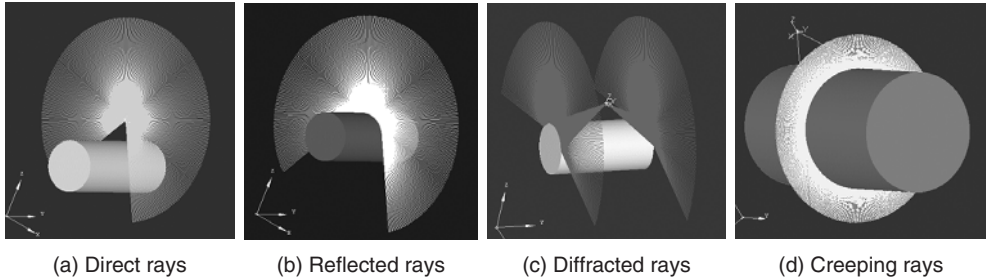


Figure 1.2 Graphic illustration of the direct, reflected, diffracted and creeping rays obtained for an antenna located above the surface of a cylinder. Figure 4 of [1]. Reproduced by kind permission of EADS. See Plate 1 for the colour figure.

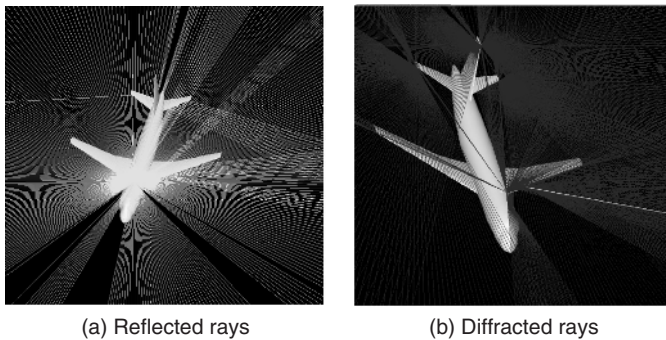


Figure 1.3 Rays obtained using the EADS ASERIS-HF GTD code. Figure 27 of [2]. Reproduced by kind permission of EADS. See Plate 2 for the colour figure.

1.2.1 Reflection

Reflection is explained by Snell’s first law, which states that if an incident ray strikes a planar reflecting surface, it is reflected at the same angle to the normal as the incident angle and that the incident, normal and reflected rays are all in the same plane. Snell’s laws are named after the Dutch mathematician Willebrord Snellius (1580–1626).

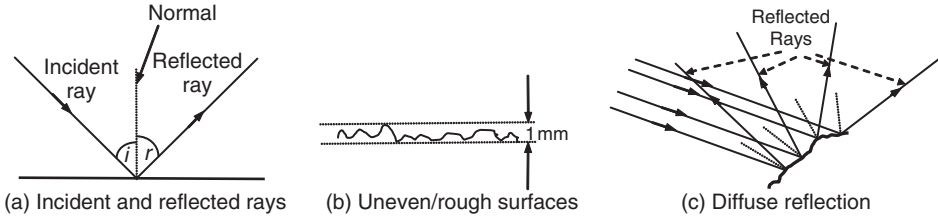


Figure 1.4 Snell's law, uneven surfaces and diffuse reflection for uneven surfaces.

The incident angle i is the angle between the incident ray and the normal to the surface, and the reflected angle r is the angle between the reflected ray and the normal to the surface, as shown in Figure 1.4a.

In the case of light waves, we can understand this phenomenon quite easily since we encounter it when looking in a plane mirror. However, if we consider a surface that only partially reflects the light, for instance parts of the dashboard reflected in the windscreen, then we are more likely to understand the EM waves reflected by surfaces that do not reflect most of the EM wave. In the case of the surface of an aircraft the complex reflected waves are akin to the images seen in the 'crazy' mirrors of a fairground.

In the case of a flat, perfectly conducting surface, the reflected ray would be similar to that obtained from a mirror, and in RF terminology this is called specular reflection.

If the surface is uneven, as in the case of most airframes, then there will be diffuse reflection, with waves scattered in a number of directions, in the same way as a sheet of white paper would diffuse the light falling on it.

When we say that a surface is flat or uneven, we mean relative to the wavelength of the radiation falling on it. For instance, if a surface roughness varies by 1 mm as shown in Figure 1.4b, this surface would be smooth/flat to a EM wave of 500 MHz that has a wavelength of 60 cm, whereas it is rough to visible light of frequency 450 THz (450×10^{12}) that has a wavelength of $0.7 \mu\text{m}$.

This unevenness also explains diffuse reflection obtained, for instance, when light falls on a sheet of white paper. A parallel beam of light falling on an arbitrarily uneven surface is shown in Figure 1.4c, and we can see that the reflected rays are scattered in several directions, known as diffuse reflection.

Apart from the direct wave that is propagated into free space, the first order reflected wave (i.e. the first reflected wave that is the result of a direct wave striking a reflecting surface) contributes the greatest to the resultant in the far field. If the first order reflected wave is in phase with the direct wave, then the resultant would be a maximum, whereas if the first order reflected wave is in antiphase (180° out of phase) with the direct wave then the resultant would be a minimum.

1.2.2 Refraction

In real life we encounter the phenomenon of refraction when looking into a swimming pool where the floor of the pool appears less deep than it actually is. Refraction is the bending of rays when an incident ray strikes a medium with a different refractive index. In

general, the denser media have higher refractive indices. Spectacle lenses that are thinner usually have higher refractive indices, to attain the same power (dioptr).

Refraction is explained by Snell’s second law, which states that when an incident ray enters a more optically dense medium (i.e. with a higher refractive index), it is bent towards the incident ray. Conversely, if it enters a less dense medium, it is bent away from the incident ray, as shown in Figure 1.5.

Snell’s law is expressed as

$$\mu_1 \sin i = \mu_2 \sin r, \tag{1.1}$$

where

μ_1 is the refractive index of the first medium

i is the angle of incidence at the interface between the first and second medium

μ_2 is the refractive index of the second medium

r is the angle of refraction.

For instance, if the first medium has a refractive index of 1 (like air), the second medium has a refractive index of 1.4 and the angle of incidence is 36° , using Equation 1.1 the refracted angle is 24.8° .

However, if the refractive indices were reversed and the wave was travelling from a medium with a refractive index of 1.4 to one with a refractive index of 1, then the refracted angle would be 55.4° , as shown in Table 1.1.

If the second medium is in the form of a plate where the top and lower surfaces are parallel, then the ray emerging from the plate will be parallel to the incident ray but displaced from it by a distance that is directly proportional to the thickness of the plate.

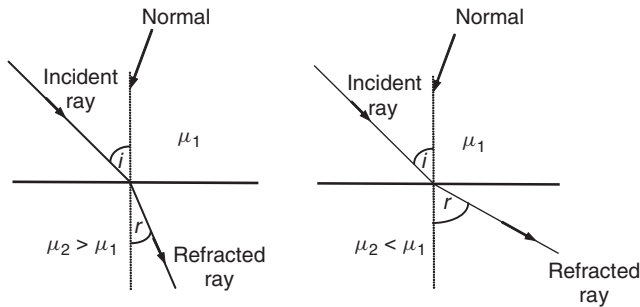


Figure 1.5 Snell’s law of refraction.

Table 1.1 Change of refracted angle with refractive index.

μ_1	μ_2	Angle i	Angle r
1	1.4	36	24.8
1.4	1	36	55.4

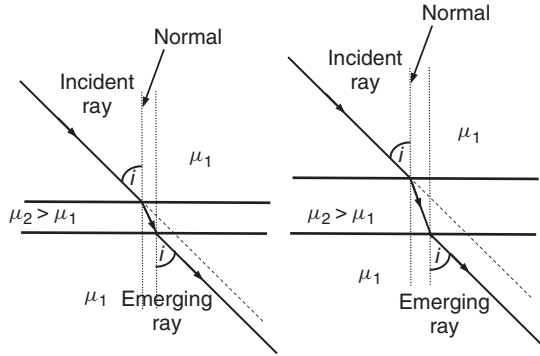


Figure 1.6 The effect of increasing the thickness of the plate.

In the case of aircraft antennas the materials that would subject the EM waves to refraction are composite fibreglass dielectric surfaces like those used in radar domes (radomes) that protect the antennas whilst still allowing the radiation to pass through them.

In Figure 1.6 we can see that as the thickness of the plate is increased the displacement of the emerging ray is also increased. The thickness that we have to consider is the electrical thickness, that is, the thickness in terms of wavelength. Thus if we keep the physical thickness the same but double the frequency the wavelength will have halved, and so that is tantamount to doubling the thickness of the plate. The displacement will be *double* the amount for the higher frequency (assuming that the refractive index has also doubled with frequency). This accounts for the fact that the fibreglass covering used for low frequency antennas such as ‘blades’ can be quite thick, but in the case of the nose cones the radomes have to be thin. If the radomes have to be strong enough to withstand birdstrike, for instance, then the radomes are made of several layers, where the effect of some of the layers compensates for the adverse effect of other layers. These are known as sandwich radomes.

We must also consider the effect of increasing the refractive index. If the refractive index is increased, the refracted ray is bent more and thus the emerging ray is more displaced compared to the incident ray. This can be seen in Figure 1.7. When the refractive index μ_2 of the plate is increased to a larger value μ_3 the emerging ray is displaced to a larger extent.

It should also be noted that the refractive index of the same material varies with frequency, and in general it increases with increasing frequency.

1.2.2.1 Total Internal Reflection

We have seen how a ray entering a medium of lower refractive index is bent away from the normal according to Snell’s law. At a certain angle of incidence i_c known as the critical angle, the angle of refraction is 90° so that the ray travels along the interface between the two media, as shown in Figure 1.8.

At angles of incidence greater than i_c , the ray is refracted back into the first medium. Because the ray now behaves like a reflected ray, with the angle of reflection equal to the angle of incidence, this effect is known as total internal reflection.

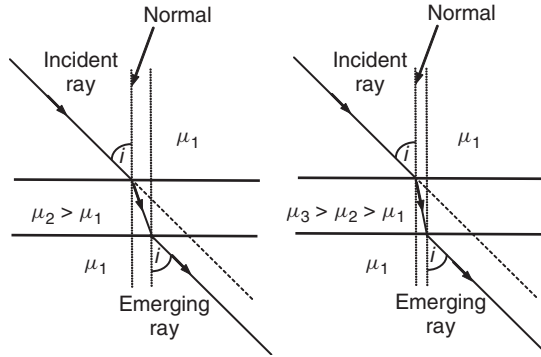


Figure 1.7 The effect of increasing the refractive index of the plate.

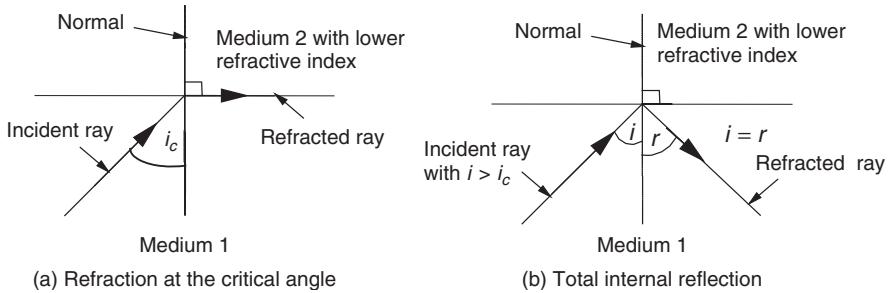


Figure 1.8 The effect of increasing the angle of incidence for a wave travelling to a medium of lower refractive index.

In the earth’s atmosphere the refractive index decreases with height above the ground. If we think of the atmosphere as layers of air with decreasing density (and hence decreasing refractive index) we can see from Figure 1.9 that as the altitude is increased, a wave travelling upwards is gradually bent further away from the normal, so that it eventually undergoes total internal reflection and is reflected back towards the earth. Rays that are

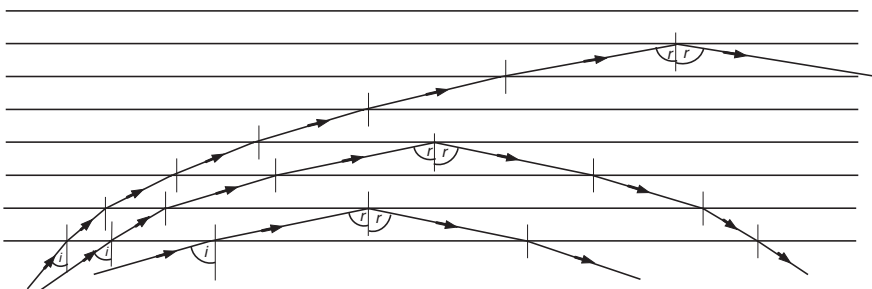


Figure 1.9 Total internal reflection for rays incident at different angles.

incident at large angles undergo total internal reflection at lower altitudes than those that are incident at small angles.

1.2.3 Diffraction

Diffraction is the bending of light waves when they strike an edge. In Isaac Newton's time, when light was considered as travelling in straight lines, this was called rectilinear propagation. However, when it was discovered that the edges of shadows were not sharp, Newton put forward his corpuscular theory to explain this, by postulating that light consisted of particles like miniature golf balls and when they passed over an object, some fell into the shadow region and blurred the edge of the shadow. In Figure 1.10 the corpuscles are shown striking the edge of a plate and most fall to the right of the extended line joining the source to the edge of the plate, and some fall in the semi-lit region, but none fall in the shadow region. Only the corpuscles near the edge are shown for clarity.

Later, when interference fringes were discovered, the corpuscular theory could not explain how combinations of particles could form light as well as dark areas. This could only be explained by considering light as waves. The wave theory explained this phenomenon by showing that the two waves could either cancel or sum, depending on their relative phases.

The wave theory explained the propagation of light until the discovery of the photoelectric effect. It was discovered that EM radiation falling on certain surfaces released

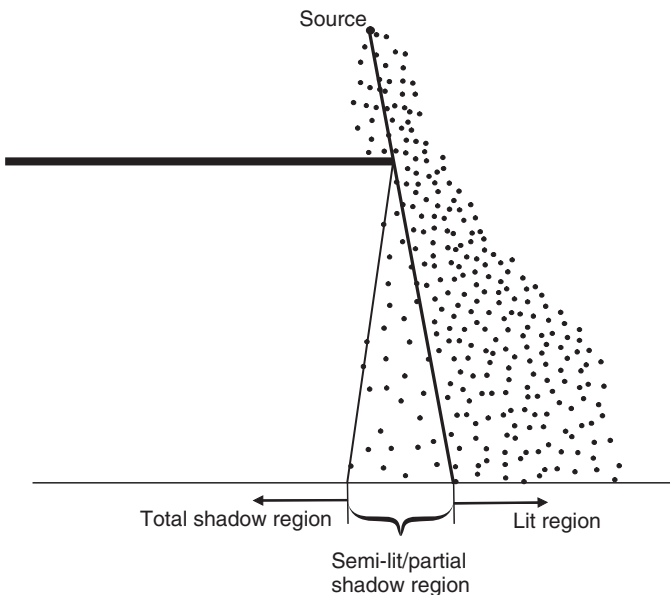


Figure 1.10 Newton's corpuscular theory of light.

particles of a different frequency from those surfaces. However, when the frequency of the irradiating radiation was decreased the photoelectric effect did not occur. The wave theory of light could not explain this, and thus it was concluded that the EM radiation consisted of quanta and the energy of each quantum was h/λ (where h is Planck's constant and λ is the wavelength). At low frequencies the wavelength is large and thus the energy of each quantum is small.

If light did not consist of quanta then the photoelectric effect would occur by just increasing the intensity of the radiation, regardless of the frequency.

However, in order to still explain the interference fringes it was concluded that light had a dual nature, consisting of waves and quanta.

Under the wave theory of light, when light falls on a surface, secondary sources of light or wavelets, are formed (commonly called Huygens sources). Each of these secondary sources emit spherically, which accounts for the fact that an illuminated edge can be seen at angles away from the normal, whereas the original source cannot be seen at these angles.

This phenomenon that results in the 'bending' of light rays at an edge is known as diffraction and applies to all forms of EM radiation, such as RF, IR, UV, and so on.

A good example of diffraction can be seen with the appearance of the diamond rings when the sun goes into and comes out of a total eclipse.

1.2.3.1 Creeping Waves

EM waves also tend to follow a curved surface and 'creep' along its surface. This effect is due to the EM wave travelling along smooth surfaces, and the diffracted rays being emitted tangential to the surface into the surrounding space. These creeping waves account for the fact that when, for instance, a monopole is installed on the top of a cylinder a significant part of the radiation occurs in the lower hemisphere.

Creeping waves that travel around the earth are also sometimes called ground waves or Norton waves. They are guided along the ground in the same way as an EM wave is guided along a transmission line. As they travel over the ground they undergo attenuation. The energy lost from the wavefront into the ground is replaced by energy in parts of the wavefront higher up above the ground. This results in a continuous movement of energy towards the ground.

Creeping waves are particularly pertinent to aircraft, since most fuselages are of circular or elliptical cross-section. The waves travel around the curved surface, taking a number of paths known as geodesics. A geodesic is defined as the shortest route between two points on a mathematically derived surface that includes the points, and in the case of the earth's surface, this would be a segment of a great circle cut. A great circle has the same circumference as the equator. Geodesics are sometimes defined as curves whose tangent vectors remain parallel if they are transported along it.

In the case of a monopole on a cylinder, if we consider the circular cross section, there are two geodesics between diametrically opposite points. These travel along the circumference of the circular cross-section, one in each direction. However, in the case of elliptical cross-sectional cylinders, there may be as many as six geodesics.

1.3 Interaction between Two Waves

Because light consists of waves, we have to consider the phase as well as the amplitude. If we consider the interaction between two waves, we have to consider their relative amplitudes and their relative phases. The interaction and the resultant produced can be depicted by looking at the waveforms or considering the waves as vectors using phasors. The resultant of the phasors can be obtained by drawing the individual phasors to scale or by calculation using simple trigonometry.

When we look at the waveforms in the time domain, the resultant is also in the time domain, so we can see the actual waveform of the resultant. However, when we use phasors we get the amplitude and the phase of the resultant, but we do not see its actual waveform. Also, phasors can only be used for waves of the same frequency, whereas time domain addition can be used for waveforms of different frequencies.

1.3.1 Waveforms in the Time Domain

If the two waves are in phase and of equal amplitude, as shown in Figure 1.11a,b, and if we add the magnitudes at each phase angle together, we will get a resultant wave of double the amplitude and in phase with both the waves, as shown in Figure 1.11c. This is known as constructive interference. The same effect occurs in the case of sound waves in organ pipes and is known as resonance in that case.

If the two waves are 180° out of phase (i.e. in antiphase) and of equal amplitude, as shown in Figure 1.12a,b, the resultant is a wave of zero amplitude. This is known as destructive interference, and the resultant is shown in Figure 1.12c.

If the two waves are in antiphase (180° out of phase) and the second wave has half the amplitude of the first one, as shown in Figure 1.13a,b, then the resultant wave will have half the amplitude of first wave (i.e. the same amplitude as the second wave) and be in phase with the first wave, as shown in Figure 1.13c.

If the two waves have equal amplitude and are out of phase by 90° , as shown in Figure 1.14a,b, then the resultant wave will have 1.41 times the amplitude of either wave and be 45° out of phase with each wave, as shown in Figure 1.14c.

If the second wave, as shown in Figure 1.15b, is 1.93 times the amplitude of the first wave and 135° out of phase with it, the resultant is a wave of amplitude 1.41 times

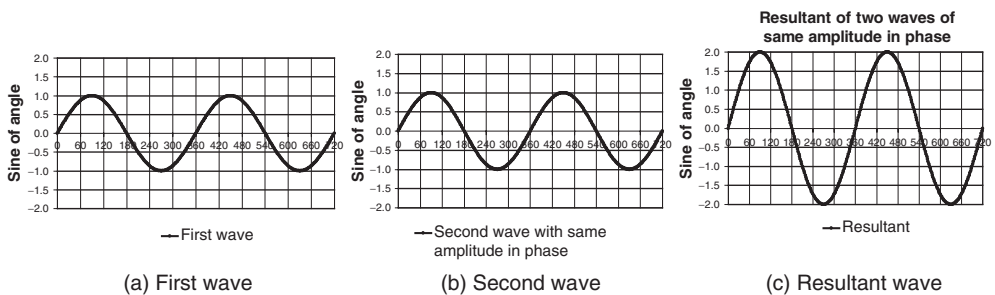


Figure 1.11 Resultant of two waves that are in phase and of equal amplitudes. The resultant is a wave of double the amplitude and in phase with both the waves.

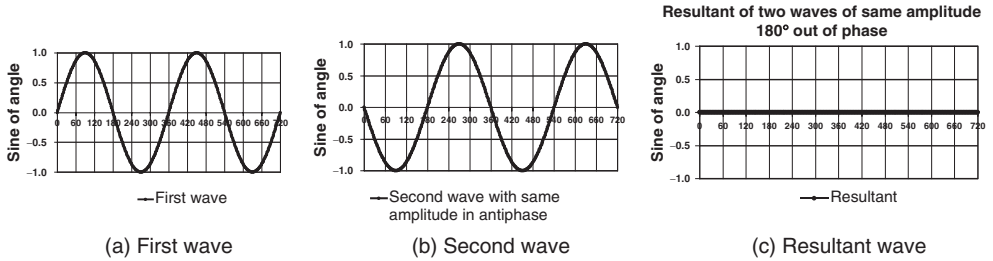


Figure 1.12 Resultant of two waves that are 180° out of phase and of equal amplitudes. The resultant is a wave of zero amplitude.

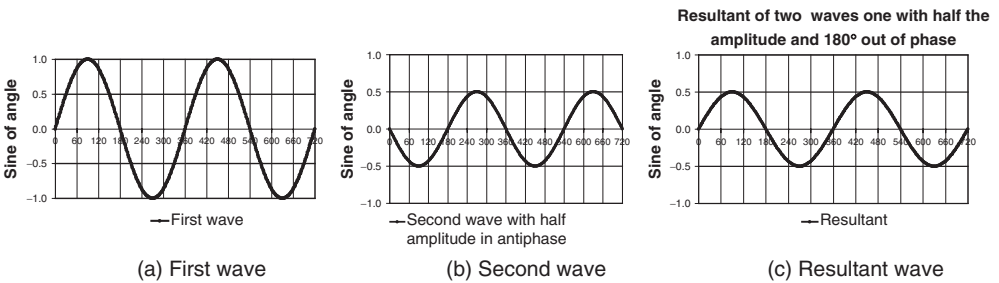


Figure 1.13 Resultant of two waves that are 180° out of phase and where the amplitude of the second wave is half that of the first wave. The resultant is a wave in phase with the first wave and of half the amplitude of the first wave.

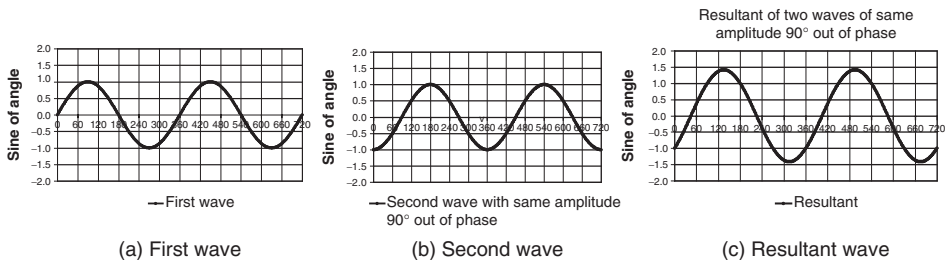


Figure 1.14 Resultant of two waves of equal amplitude that are 90° out of phase. The resultant is a wave in phase with the first wave and of 1.41 times the amplitude of the first wave.

the amplitude of the first wave and about 105° out of phase with it. This can be seen in Figure 1.15c. Note that this is the same resultant (in amplitude) as obtained in Figure 1.14, although the second waves are not the same in either amplitude or phase.

The same occurs in the case of the radiation pattern of an antenna. It gives the resultant of the interaction(s) of usually two or more waves, but we cannot identify the source(s) of the individual waves that give us the resultant and more than one combination can have the same resultant. Thus we do not have an unique solution in these cases.

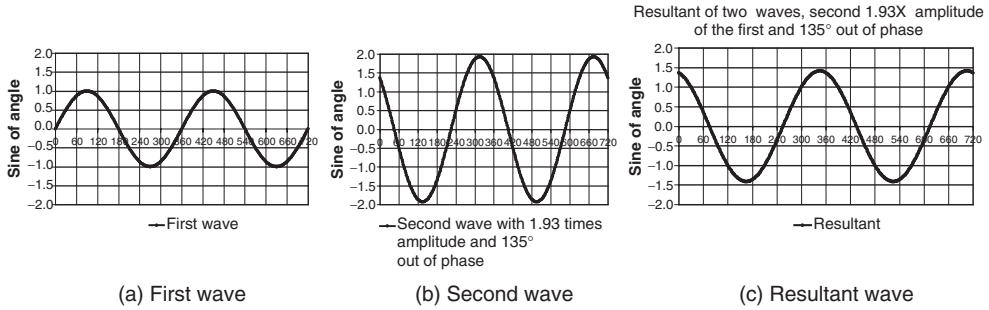


Figure 1.15 Resultant of two waves, the second being 1.93 times the amplitude of the first wave and 135° out of phase with it. The resultant is a wave of amplitude 1.41 times the amplitude of the first wave and about 105° out of phase with it.

1.3.2 Phasors

Instead of plotting the waves in the time domain to obtain the resultants between the waves, we can use phasors. Phasors are vectors that are used to represent the waves, the amplitude being represented by the magnitude or length and the phase being represented by the direction of the vector. The resultant of two vectors can be derived by addition using the triangle of forces. The phasors can be drawn to scale with the angles measured using a protractor, or trigonometry can be used to calculate the resultant.

1.3.2.1 Phasors Drawn to Scale

The first vector is drawn to a suitable length and the second one is drawn to the same scale at an angle to represent the phase between the two. The resultant is obtained by joining the line between the start of the first vector and the end of the second vector. For a description of vector addition see [3], p. 52.

The five cases shown in Figure 1.11–1.15 are depicted as phasors in Figure 1.16. Note that the first and second vector are drawn cyclically (i.e. both are drawn in an anticlockwise direction) and the resultant drawn in the opposite or clockwise direction. If we consider case (e) of Figure 1.16 that corresponds to the addition of the waves shown in Figure 1.15, we can see that the angles are drawn as mathematical angles, that is, anticlockwise from the first quadrant. The following steps describe in detail the process for case (e):

1. Draw a horizontal line to a convenient length with an arrow at the right-hand end. This is used as the scale unit 1 for the first phasor.
2. Using a protractor mark out 135° from the arrow end of the first phasor.
3. Mark out the length of the second phasor to be 1.93 times the length of the first phasor, with an arrow at its top end.
4. Join the start of the first line and the arrow end of the second phasor.
5. The new line represents the resultant phasor.

The resultant is of length (and hence amplitude) 1.41 times the first vector and its phase is 105° compared with the first vector.

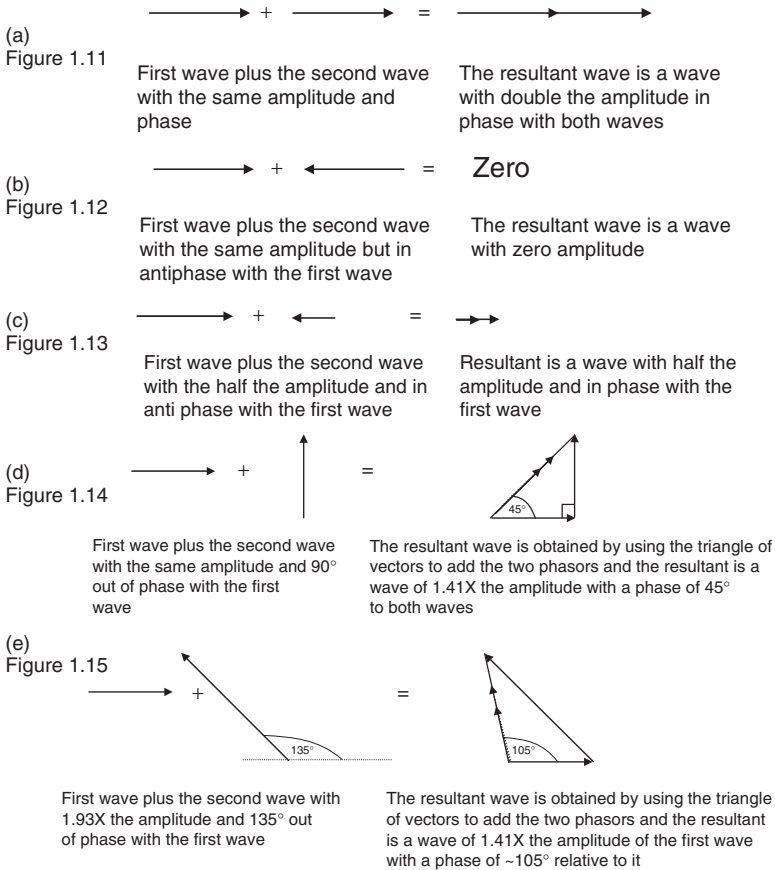


Figure 1.16 Interaction between two waves drawn as phasors to scale.

1.3.2.2 Phasors Used in Calculations

Simple trigonometry can be used to obtain the resultants of Section 1.3.2.1 above.

The cosine and sine formulas (see [3], pp. 39–42) are used in this case. In the triangle the sides are denoted a , b and c , and the angles opposite these sides are denoted A , B and C , respectively. Referring to Figure 1.17a:

1. The first phasor corresponds to side b and the angle opposite it is angle B .
2. The second phasor corresponds to side c and the angle opposite it is angle C .
3. The resultant phasor corresponds to side a and the angle opposite it is angle A .
4. The angle ϕ is the phase difference between the first and second waves. Note that angle A in the cosine formula is $180 - \phi$.
5. The angle C gives us the phase between the resultant and the first wave.

Thus we need to find the values of length a and angle C to find the magnitude and phase of the resultant, respectively.

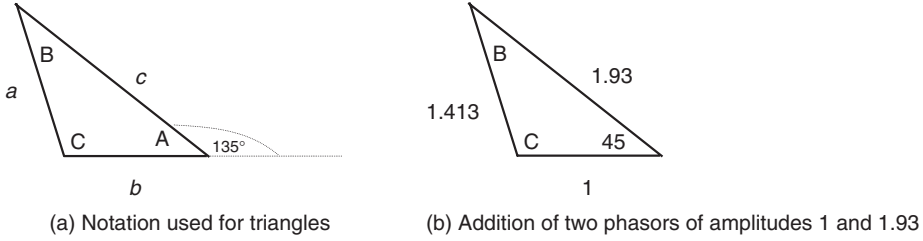


Figure 1.17 Interaction between two waves using trigonometry to calculate the resultant amplitude and phase.

The cosine formula can be used for any type of triangle – acute, right-angled or obtuse. In the case of a right-angled triangle, the cosine of angle $A (=90^\circ)$ is zero and the formula reduces to Pythagoras' theorem. The cosine formula is given by

$$a^2 = b^2 + c^2 - 2bc \cos A. \quad (1.2)$$

In the case of (e) in Figure 1.16, we have $b = 1$, $c = 1.93$, angle ϕ is 135° so angle A is 45° . Using these values in Equation 1.2 we get

$$a^2 = 1 + 1.93^2 - 2 \times 1 \times 1.93 \cos(45^\circ).$$

Thus $a^2 = 1.995$, giving $a = 1.41$, which is the amplitude of the resultant.

The sine formula is now used to calculate the angle C . The sine formula states that

$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C}. \quad (1.3)$$

We only need to use

$$\frac{a}{\sin A} = \frac{c}{\sin C},$$

that is,

$$\sin C = \frac{c \sin A}{a} = \frac{1.93 \sin 45^\circ}{1.41}$$

so that angle $C \approx 75^\circ$ or 105° . In order to decide which is the correct angle, we need to look at the value of the third angle B .

The exterior angle 135° is equal to the sum of the interior angles B and angle C . Thus if angle C is 75° then angle B is 60° , giving $b/\sin B = 1.1547$. If angle C is 105° then angle B is 30° , giving $b/\sin B = 2$. Since $b/\sin B = a/\sin A$, and since $a/\sin A$ is 1.9954, the second result for $b/\sin B$ is the correct one, and angle B must be 30° . Therefore the correct value for angle C is 105° . Thus we get the same value as that obtained by drawing the phasors to scale in Section 1.3.2.1 above.

Using trigonometry gives us a more accurate result for the phase and magnitude of the resultant, but it can be seen that it has to be used carefully, since more than one value could result in the case of the phase angle.

1.4 Polarization

In the case of light waves, the polarization is random. Polarizing filters work on the principle of eliminating the polarization in all planes except one. Thus if we were to put one polarizing filter (a piece of Polaroid) behind the other and then rotate one, there would be positions where there is maximum light passing through both and a minimum (darkness) at right angles to the maximum position. We can think of the polarizing filter as a set of vertical slots so that only light with the plane of polarization parallel to these slots can pass through the slots. When the two lots of slots are at right angles to each other no light can get through.

In Section 1.2.1 we considered light as rays and we ignored the polarization. However, the polarization is altered when a wave is reflected off a surface. This is easier to see if we look at light diffusely reflected off a flat surface – not a mirror. If we look at light reflected at different angles through Polaroid filter, at one particular angle (depending on the surface) only plane polarized light would be reflected, and this is known as the Brewster angle. The reflected wave does not have any polarization in the plane containing the incident, normal and reflected waves. The Brewster angle was named after Sir David Brewster (1781–1868). The angle between the refracted and reflected rays is 90° at the Brewster angle. We can check that it is plane polarized light by using a second polaroid filter and rotating it relative to the first until it is perpendicular to first, at which point no light would be visible.

In the case of RF waves the polarization could be linear, elliptical or circular, although theoretically all polarizations could be considered to be variations of elliptical polarization.

The plane wave EM field in free space consists of a magnetic field at right angles to an electric field. Both fields vary sinusoidally in space as well as time. Thus if we were to consider the field at a fixed point in space it would vary in amplitude sinusoidally and if we were to capture the field at a fixed point in time, the fields would vary as a sine wave in space. Each of the electric and magnetic fields is represented by a vector and has an amplitude as well as a direction.

Figure 1.18 shows the orientations of the electric and magnetic field for a plane wave. The electric field varies in the yz plane, and the magnetic field varies in the xz plane, whilst the wave travels along the z axis.

When we refer to polarization, by default this refers to the polarization of the electric field vector, rather than the magnetic field vector. The polarization of an EM signal refers to direction of the electric field vector during at least one full cycle.

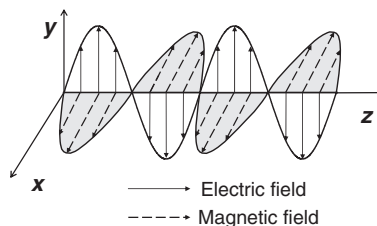


Figure 1.18 Orientation of the electric and magnetic fields for a plane wave.

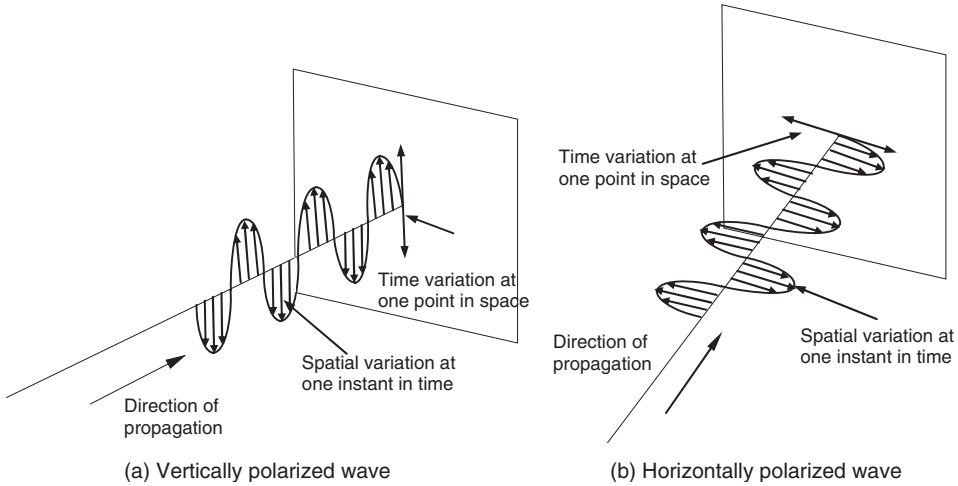


Figure 1.19 Linearly polarized waves.

1.4.1 Linear Polarization

The wave could be linearly polarized in any plane. In the case of a linearly polarized plane wave, the wave progresses in the direction of propagation and the electric field would appear as shown in Figure 1.19. If we were to look at the wave at one point in space, the electric field would move up and down (i.e. vary sinusoidally with time) in the case of the vertically polarized wave as shown on the square. The total excursion of the electric field would be from the positive maximum (amplitude of the sine wave) down to zero and then down to the negative maximum, and back again to the positive maximum over one period of time – that is, when the wave has completed one cycle. Thus one period is the time taken to complete one total excursion and is the reciprocal of the frequency. In the spatial domain the distance for one complete cycle gives us the wavelength.

The other most common types of linear polarizations used for RF propagation are left slant and right slant polarizations. The right slant is usually defined as making an angle of $+45^\circ$ (clockwise like navigational angles) with respect to the vertical when looking from the transmitted radiation, as shown in Figure 1.20a. The left slant is usually defined as making an angle of -45° with respect to the vertical when looking from the direction of the transmitted radiation, as shown in Figure 1.20b.

1.4.2 Circular and Elliptical Polarization

Circular polarization is a form of elliptical polarization. In the case of elliptical polarization, the magnitude as well as the direction of the electric field will vary during a cycle, whereas in the case of circular polarization, the magnitude of the electric field vector remains constant.

In the case of elliptical polarization, the tip of the electric field vector would appear like a helix with an elliptical cross-section in space if we were to freeze it at one instant in time. In addition, if we were to look at the electric field vector at one point in space,

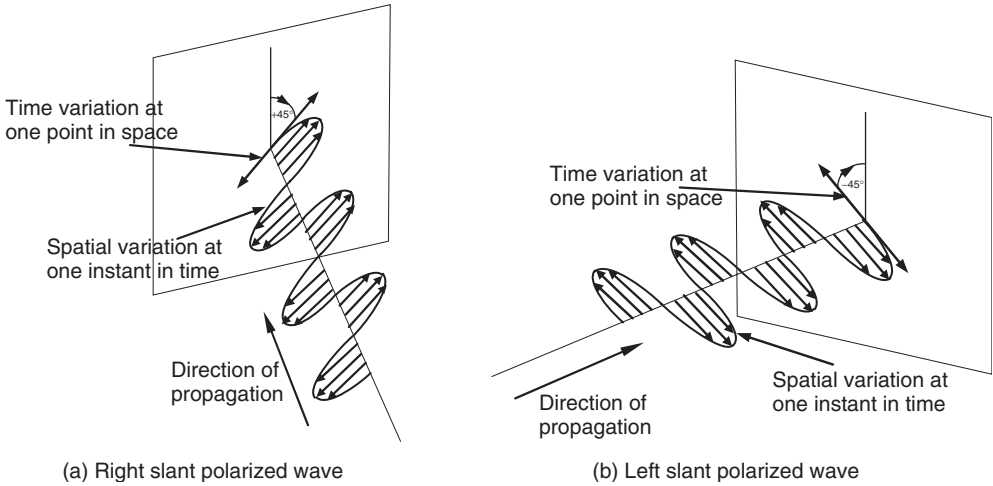


Figure 1.20 Linearly polarized slant waves.

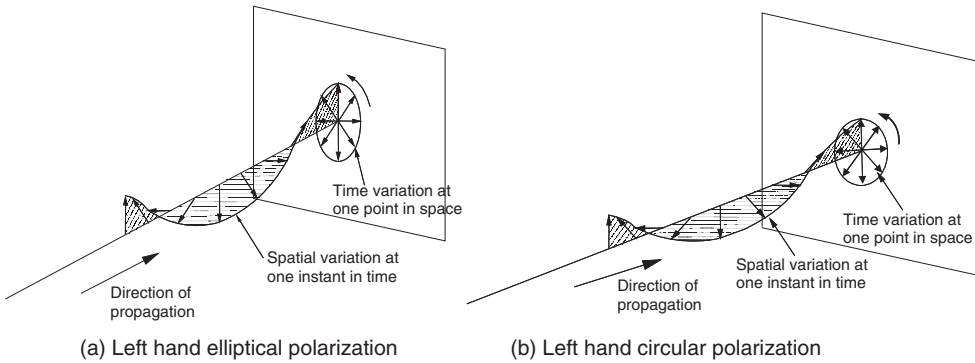


Figure 1.21 Elliptical and circular polarization.

it would appear like an elliptical disc over one period or cycle in time, as shown in the square of Figure 1.21a.

In the case of circular polarization, although the magnitude of the electric field vector remains constant, its direction changes as the wave propagates in space so that its tip appears like a helix at one instant in time; and at one point in space, a circular disc is swept out, as shown Figure 1.21b. The direction in which the vector rotates determines the hand of polarization. Of course, the direction has to be defined in relation to the viewpoint, since the same vector would appear as either clockwise or anticlockwise, depending on the direction from which it is viewed.

The Institute of Electrical and Electronics Engineers (IEEE) defines right hand circular polarization (RHCP) as a clockwise rotation of the electric field vector when looking along the direction of propagation from the transmit antenna, and left hand circular polarization

(LHCP) as an anticlockwise rotation. Thus in Figure 1.21 we have left hand elliptical and circular polarization, since the electric field vectors are rotating anticlockwise in both cases.

In the EM field the ellipticity e of circular polarization is defined as the ratio of the minor to major perpendicular components of the electric fields E_{minor} and E_{major} , and is a measure of the polarization purity or the cross-polar discrimination. E_{minor} and E_{major} are the semi-major and semi-minor axes of the ellipse as shown in Figure 1.22. The ellipticity is given by

$$\rho_e = \frac{E_{\text{minor}}}{E_{\text{major}}} \quad (1.4)$$

The ellipticity in dB, $\rho_{e(\text{dB})}$, is given by

$$\rho_e \text{ (dB)} = 20 \log \left(\frac{E_{\text{minor}}}{E_{\text{major}}} \right) \quad (1.5)$$

For circular polarization the ellipticity is one, that is, the two electric field components are equal. In dB the ellipticity would be 0dB for circular polarization. If, in the case of Figure 1.22a, we take the vertical component as E_{major} and the horizontal component as E_{minor} , then when E_{minor} is zero, the ellipticity is zero and we have linear vertical polarization. Similarly, when E_{major} is zero in Figure 1.22b, the ellipticity is infinite and we have linear horizontal polarization.

1.4.2.1 Tilt Angle

There is a difference between the tilt angle and tilt axis. The tilt angle is defined as the angle of the major axis, whereas the tilt axis is the angle of the polarization ellipse as defined in Section 1.4.2.2. The plane of polarization is the plane of the ellipse (or circle), that is, the time variation at a point in space, and the tilt angle is defined as the angle α that the major axis makes with a reference direction such as the horizontal. The angle α is measured looking from the direction of propagation and is measured clockwise from a reference axis (such as the horizontal) to the major axis of the ellipse as shown in Figure 1.22c. However, sometimes the angle is measured anticlockwise from the horizontal, giving an acute angle α' as shown in Figure 1.22c.

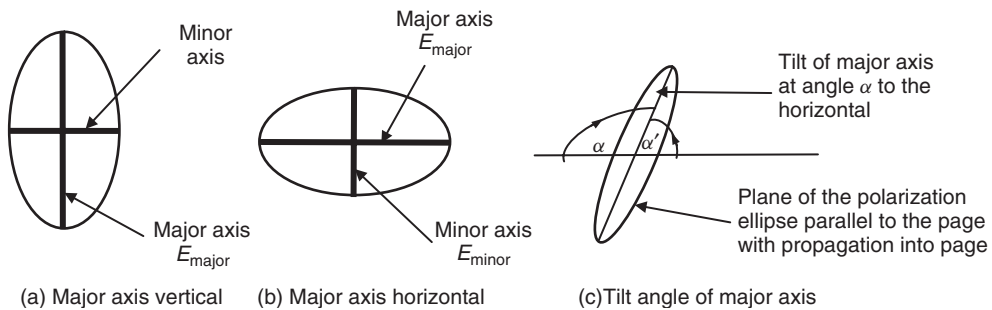


Figure 1.22 The electric fields for elliptical polarization.

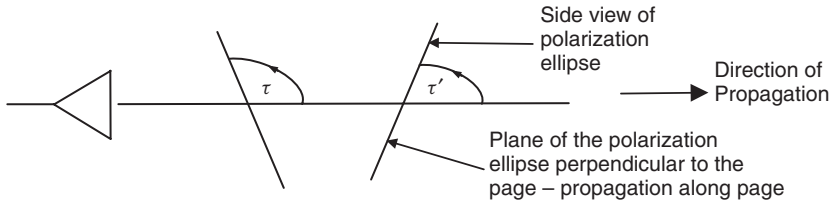


Figure 1.23 Tilt axis for elliptical/circular polarization.

1.4.2.2 Tilt Axis

The tilt axis τ is defined as the angle that the plane of polarization makes with the direction of propagation. This angle is measured from the axis and direction of propagation to the plane of the polarization ellipse. In Figure 1.23 the tilt axes are shown for two different cases, where the tilt axis is an obtuse angle τ and an acute angle τ' .

1.4.3 Axial Ratio

The axial ratio is sometimes defined by the reciprocal of the ellipticity ρ_e . However, the more commonly used definition of the axial ratio χ_r is given by

$$\chi_r = \frac{1 + \rho_e}{1 - \rho_e}, \quad (1.6)$$

where ρ_e is the ellipticity, and since

$$\rho_e = \frac{E_{\text{minor}}}{E_{\text{major}}},$$

we get

$$\chi = \frac{1 + E_{\text{minor}}/E_{\text{major}}}{1 - E_{\text{minor}}/E_{\text{major}}}. \quad (1.7)$$

The axial ratio χ_{dB} in dB is given by

$$\chi_{\text{dB}} = 20 \log \left(\frac{1 + E_{\text{minor}}/E_{\text{major}}}{1 - E_{\text{minor}}/E_{\text{major}}} \right). \quad (1.8)$$

or

$$\chi_{\text{dB}} = 20 \log \left(\frac{1 + \rho_e}{1 - \rho_e} \right) \quad (1.9)$$

The ellipticity in terms of the axial ratio can be derived since

$$\begin{aligned} \chi(1 - \rho_e) &= 1 + \rho_e \\ \chi - \chi\rho_e &= 1 + \rho_e \\ \rho_e &= \frac{\chi - 1}{\chi + 1}. \end{aligned} \quad (1.10)$$

Similarly, the ratio of the perpendicular components of the electric fields can be calculated if the axial ratio χ_{dB} is known, by using the formula

$$\rho_e = \frac{E_{\text{minor}}}{E_{\text{major}}} = \frac{10^{\chi_{\text{dB}}/20} - 1}{10^{\chi_{\text{dB}}/20} + 1}. \quad (1.11)$$

1.4.4 Measurement of Polarization Purity

An antenna designed to receive one type of linear polarization (or one hand of polarization) will in the ideal case not receive a signal with polarization that is orthogonal or of the opposite hand. However, a practical antenna will receive a signal that is orthogonal or of the opposite hand at a level equal to its cross-polar level. Typical cross-polar levels are -14 dB, although levels as low as -40 dB are possible. Section 9.13 (in Chapter 9) shows the level received by antennas of different polarizations.

The ability of an antenna to reject the cross-polar radiation is known as cross-polar discrimination (XPD). The ellipticity or polarization purity, which is a measure of XPD, can be measured when undertaking the measurement of the radiation patterns. The antenna under test (AUT) is illuminated by a rotating antenna, such as a horn, so that the polarization angle is continuously varied over 360° . This is known as the spinning technique. The typical radiation pattern is shown in Figure 1.24, and the axial ratio χ_r in dB is the distance between a trough and peak.

For instance, if the axial ratio on the radiation pattern is 0.2 dB, the linear value of χ_r is $10^{0.2/20} = 1.0233$. Then

$$\begin{aligned} \chi - \chi \frac{E_{\text{minor}}}{E_{\text{major}}} &= 1 + \frac{E_{\text{minor}}}{E_{\text{major}}} \\ \chi - 1 &= \chi \frac{E_{\text{minor}}}{E_{\text{major}}} + \frac{E_{\text{minor}}}{E_{\text{major}}} \\ \chi - 1 &= \frac{E_{\text{minor}}}{E_{\text{major}}} (\chi + 1) \\ \frac{E_{\text{minor}}}{E_{\text{major}}} &= \frac{\chi - 1}{\chi + 1} = \frac{1.0233 - 1}{1.0233 + 1} = 0.0114. \end{aligned}$$

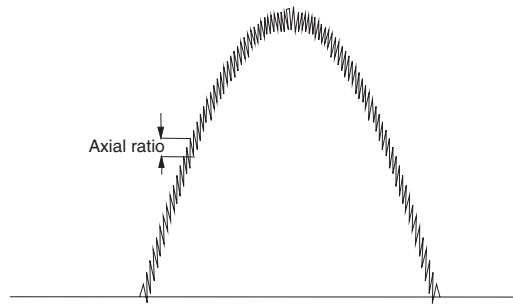


Figure 1.24 A typical radiation pattern obtained for investigation of the polarization purity of an antenna.

Polarization purity or XPD in dBs is $20 \log(0.0114) = -38.8$ dB. This means that the difference in reception between the copolar and cross-polar radiation is 38.8 dB. This is a very good level. Most blade antennas have XPDs of around -15 dBs.

Note, however, that the sense of polarization, tilt angle and tilt axis cannot be determined from this measurement.

1.5 Characteristics of an Antenna

A transmitting antenna transfers a guided EM wave from a transmission line into free space. In the case of an aircraft the antennas provide the interface between the systems inside the airframe and the outside world. Note that the plural of antenna (used in engineering) is *antennas* and not *antennae*.

It is important to know the exact spatial distribution of power (i.e. the radiation pattern in 3D space) provided by antenna in order to deduce the performance of systems connected to the antenna and to enable suitable measures to be taken for any deficiency in a particular direction. For instance, if an aircraft has to communicate with the ground it is important for a reasonable amount of power to be radiated/deflected towards the ground. On the other hand, if the prime aim of an antenna is satellite communications or navigation then the radiated power and/or the boresight of the transmit/receive antenna should be oriented towards the satellite. Antenna manufacturers provide the radiation pattern of the antenna on a standard ground plane. However, when an antenna is installed on a structure the spatial power dispersion is very different from that obtained when the antenna is on a standard ground plane.

An antenna can have the following characteristics, not all of which are meaningful to all antenna types:

1. radiation pattern
2. directivity, gain and efficiency
3. beamwidth and gain of the main lobe
4. position and magnitude of the lobes
5. bandwidth
6. polarization of electric field that it transmits or receives
7. handling power.

There are two principal planes in which the antenna characteristics are measured. These are known as the azimuth plane and elevation plane, and can be considered as the horizontal and vertical planes respectively for land-based antennas. The angles in the azimuth plane are conventionally denoted by the Greek letter phi (ϕ), and in the elevation plane they are denoted by the Greek letter theta (θ).

On an aircraft the azimuth plane is known as the yaw plane, and there are two elevation planes. The elevation plane that is transverse to the aircraft (i.e. wing to wing) is known as the roll plane, and the plane that is longitudinal to the aircraft (i.e. nose to tail) is known as the pitch plane, as shown in Figure 1.25.

In the yaw plane the elevation angle θ is zero and the azimuth angle ϕ varies from 0 to 180° .

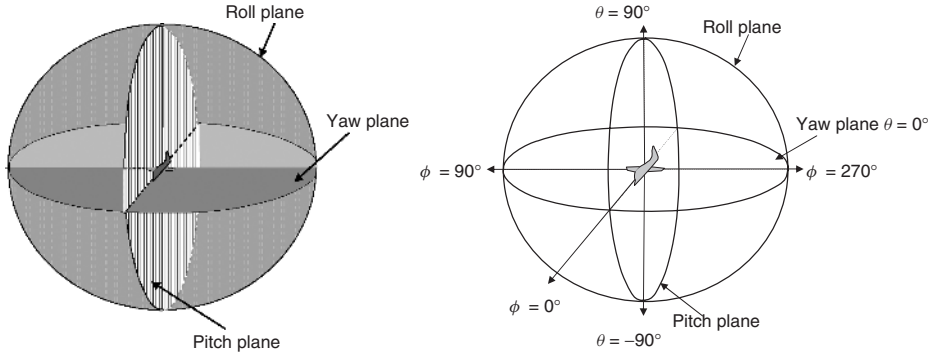


Figure 1.25 Radiation pattern cuts used for aircraft antennas.

In the roll plane the azimuth angle ϕ is 90° (or 270°) and the elevation angle θ varies over 360° , with angles between -90 and $+90^\circ$.

In the pitch plane, the azimuth angle ϕ is 0 and the elevation angle θ varies over 360° , with angles between -90 and $+90^\circ$.

For aircraft antennas, the angles are plotted as navigational angles also known as bearings and not as convention mathematical angles. Bearings go clockwise from 0° at the top/north, through to 90° , 180° , 270° and back to $0/360^\circ$, as shown in Figure 1.26a. Mathematical angles on the other hand start at 0° on the right and go anticlockwise to 90° at the top, then 180° to the left and then 270° at the bottom and back to $0/360^\circ$ as shown in Figure 1.26b.

1.5.1 Radiation Patterns

The line radiation pattern can be plotted using either rectangular/Cartesian or polar coordinates. Rectangular plots can be read more accurately (since the angular scale can be

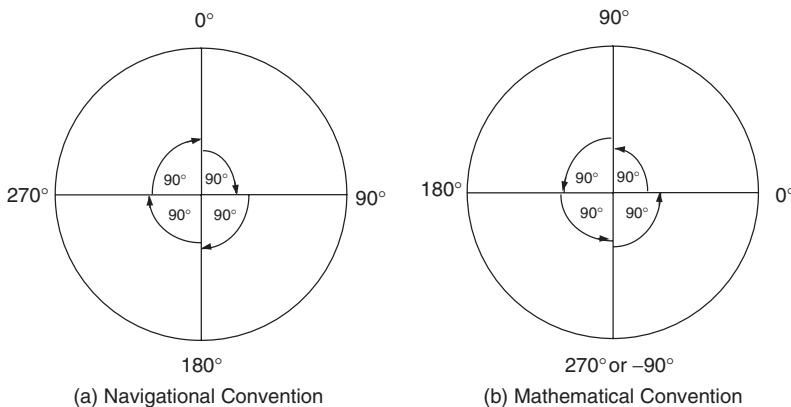


Figure 1.26 Navigational and mathematical convention for angles.

expanded), but polar plots give a more pictorial representation and are thus easier to visualize; rather like an analogue clock, or the plan position indicator (PPI) used in many radio assisted detection and ranging (radar) sets. Contour plots are another way of depicting the radiation from antennas.

1.5.1.1 Radiation Patterns of Different Types of Antennas

Antennas can be isotropic, omnidirectional or directional, depending on the directions in which they radiate.

An isotropic antenna radiates uniformly in all directions so that the radiated power at any point on a sphere (with the antenna at its centre) has the same magnitude as depicted in Figure 1.27. The darker colours indicate the higher powers found nearer the antenna. However, this cannot be realized in practice, and would require the antenna to be a point source. The nearest approximation to an isotropic antenna is a Hertzian dipole, which is a dipole that is very small in terms of wavelength.

Omnidirectional antennas such as monopoles, dipoles and biconicals radiate uniformly in one plane. Figure 1.28a shows the radiation from an ideal vertical dipole. In the vertical plane the cross-section of the radiation is in the shape of figure of eight on its side as shown in Figure 1.28b, and radiation is uniform in the horizontal plane as shown in Figure 1.28c.

A directional antenna is one that radiates most of its power in one particular direction. Examples of directional antennas are horns, reflector systems, log-periodics and Yagis. Figure 1.29 shows the radiation from a reflector antenna. For a circularly symmetrical reflector the uninstalled radiation pattern is also circularly symmetrical.

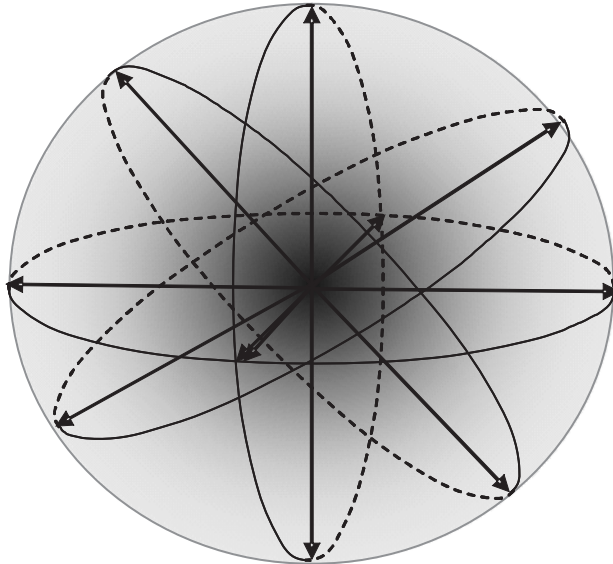


Figure 1.27 The radiation obtained from an isotropic antenna.

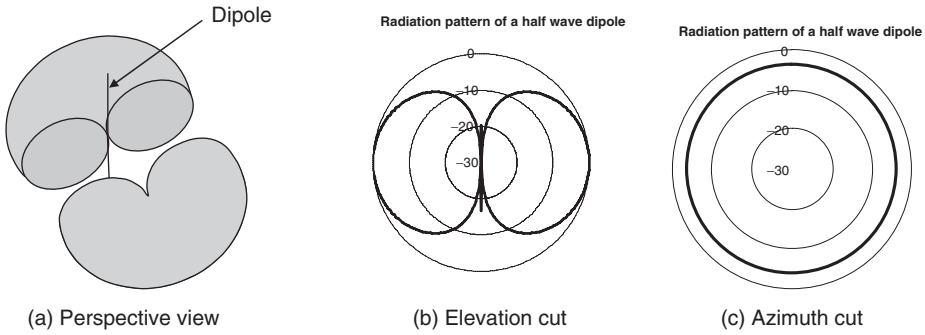


Figure 1.28 The radiation obtained from an idealized omnidirectional antenna.

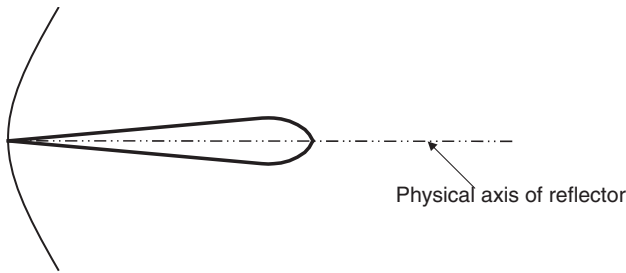


Figure 1.29 The radiation obtained from a directional antenna.

1.5.1.2 Representations of Radiation Patterns

The radiation pattern of an antenna can be presented as a 2D line graph or as a 3D representation, but shown in 2D. The line radiation patterns show the power in a specified plane, whereas the 3D representation could be shown as a perspective view, or as a contour plot projected onto a cylinder like Mercator’s projection of a map of the world.

Line Radiation Patterns

The line radiation pattern could be taken at different cuts, but the most common ones are great circle and conical cuts.

Great Circle Cuts

If we imagine a sphere with the antenna at its centre, circles cut through the centre of the sphere will have the same diameter as the sphere and are known as great circle cuts as shown in Figure 1.30. If the great circle is horizontal (goes through the equator) then in the case of an aircraft antenna at the centre of the sphere, that would be known as a yaw plane cut, as shown in Figure 1.30. Similarly, if the great circle is vertical (goes through the north and south poles) then, in the case of aircraft antennas, that would be known as a pitch or roll plane cut, depending on the orientation of the aircraft – see Figure 1.25.

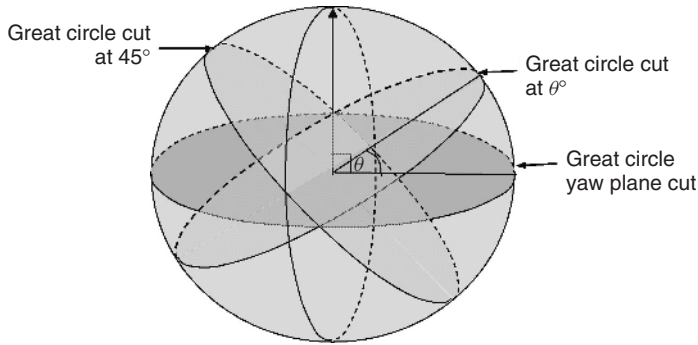


Figure 1.30 Great circle cuts.

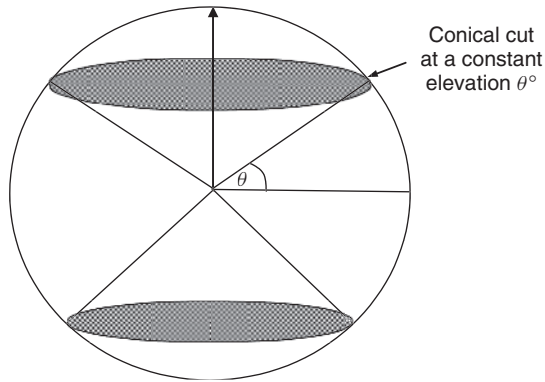


Figure 1.31 Conical cuts.

Conical Cuts

If we imagine cuts parallel to the azimuth plane of the antenna (the equator in the case of a globe), we would get circles of smaller diameter as we move away from the equator. The conical cuts are equivalent to the lines of latitude on the globe and the angle of elevation θ is constant for each cut as shown in Figure 1.31. Note that the cone semi-angle is in fact $90 - \theta$.

Rectangular/Cartesian Plots

Cartesian plots are named after the mathematician and philosopher, René Descartes. They are standard xy plots where the axes are plotted at right angles to each other. In a radiation plot the angle with respect to boresight (as defined in Section 1.5.3) is varied and the magnitude of the radiated power is measured. The x axis is used for the angle and the power radiated is plotted on the y axis. The x coordinate is known as the abscissa and the y coordinate is known as the ordinate. It is important to remember that the power radiated is measured in the far field. A typical rectangular plot of an antenna radiation pattern is shown in Figure 1.32. All values, whether negative or positive, are often shown without a sign. Thus the y axis in Figure 1.32, should be shown from 0 to -80 or -40 dB,

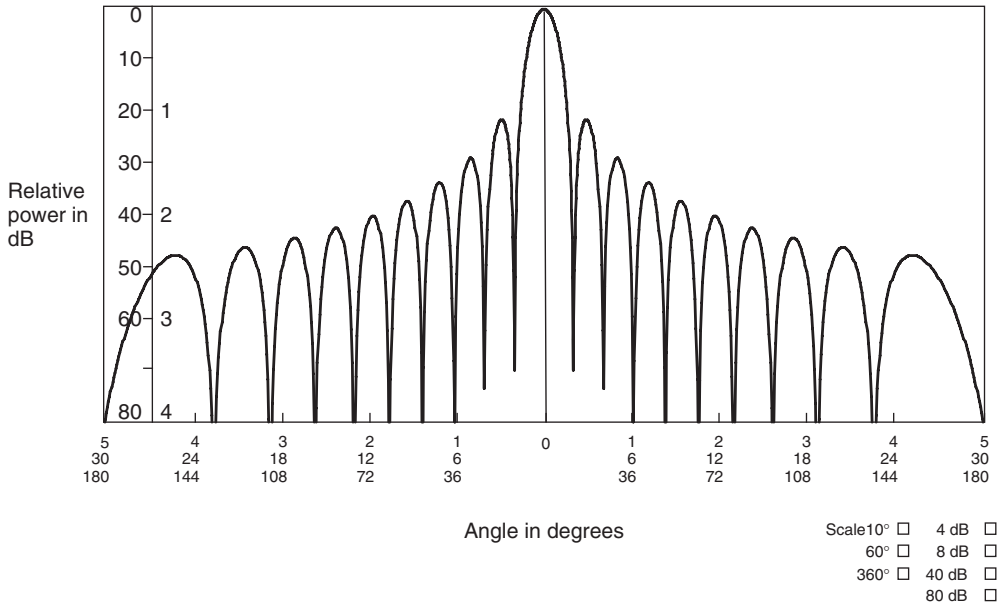


Figure 1.32 Rectangular plot of an antenna radiation pattern.

indicating that the maximum value is at boresight or 0° . The radiation patterns show the angles measured clockwise and anticlockwise from the boresight position, and in standard mathematical convention the x axis would be denoted by positive and negative signs, but on radiation patterns the signs of the angles are often omitted. The boxes on the lower right would be ticked/crossed to indicate the scale used for angles and relative power.

Polar Plots

In a polar plot the angles are plotted radially from boresight and the intensity or power is plotted along the radius, as shown in Figure 1.33. This gives a pictorial representation of the radiation pattern of the antenna and is easier to visualize than the rectangular plots. However, since the scale of the angular positions cannot be increased (i.e. they can only be plotted to scale from 0 to 360°), the accuracy cannot be increased as in the case of rectangular plots. The level of the intensity or power, however, can be varied as in the case of rectangular plots. On polar plots each circle represents a particular level, where the power has the same magnitude at all angles and is shown relative to the power at boresight. Since the power is in general a maximum value at boresight, this level is often shown as 0 dB and the levels elsewhere will always be less than the power at boresight. Thus these should be shown as negative values. However, they are usually written without a sign, and should be assumed to be negative, contrary to standard arithmetic convention. Figure 1.33 shows levels from 0 to -25 dB in 5 dB steps. Some polar plots may have a level of $+3$ or $+10$ dB as the maximum, go through the 0 dB, and then down to -30 dB. In almost all cases the signs are omitted.

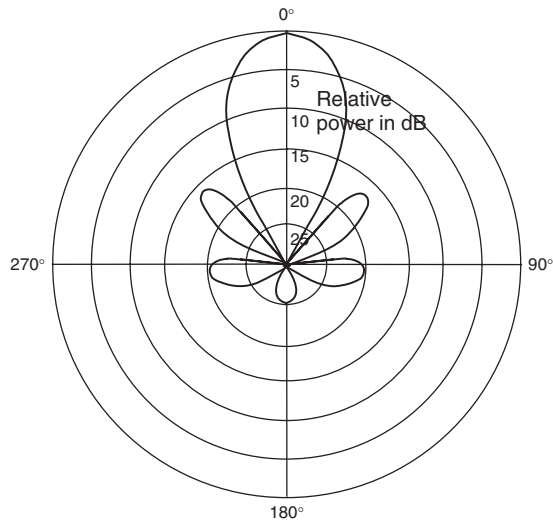


Figure 1.33 An idealized polar plot of a directional antenna.

Contour Plots

Contour plots can be plotted:

1. on a sphere and appear like a global map on a 2D surface
2. like Mercator's projection of a spherical surface onto a cylinder, like global maps
3. as line contour plots.

The contour plots are colour-coded, so that each colour represents a particular level of gain, and the levels go up in discrete steps. Thus, for instance, if the discrete steps are 2 dB, all levels between 2 and 4 dB are the same colour and the levels do not change continuously as in the case of line radiation patterns. Thus the gain at a particular angle cannot be read off accurately. However, they provide a good visual indication of the total spherical radiation pattern, and angular sectors of any 'dropouts' or nulls can be seen. These are particularly useful to systems engineers who have to evaluate the spatial coverage of avionics systems.

One version of Mercator's plots is the projection of the 3D spherical surface onto a cylinder which is then shown as a 2D rectangular surface, as if the cylinder has been opened out into a rectangle. These are the type of plots that we see in an atlas. They require some skill in deciphering the actual angles since, for instance, the whole of the top line represents just one angle of 90° elevation, as in some world maps where the north pole is represented by a line instead of a single point. A line through the centre/equator corresponds to the azimuth line radiation plot when plotted in polar or rectangular coordinates. The contour plot of Figure 1.34 is plotted as a Mercator's projection in Figure 1.35.

Instead of using solid colours for each discrete interval of power level, the contour plot can be presented as a contour line plot, where only the positions with the same power

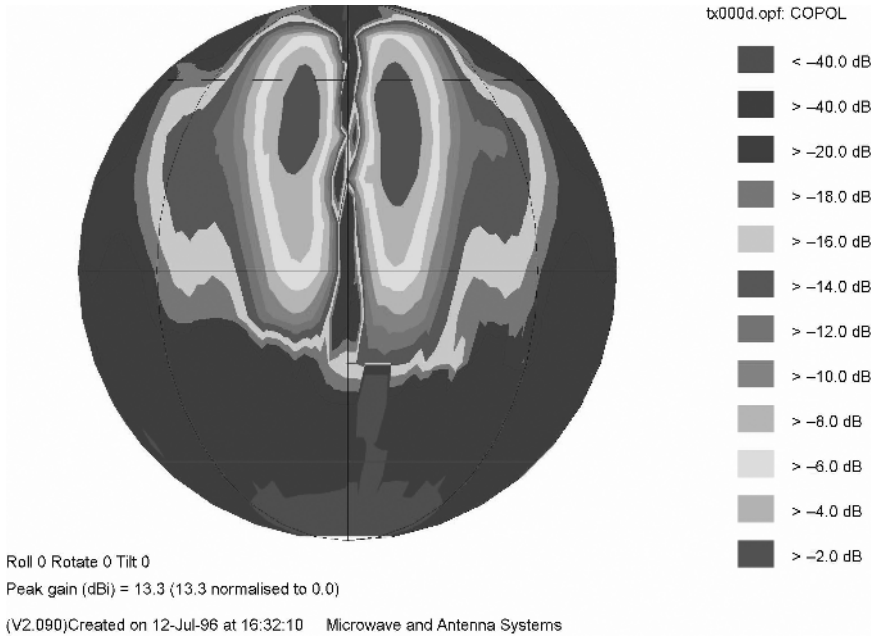


Figure 1.34 Contour plots on a spherical surface. Reproduced by kind permission of ASL [5]. See Plate 3 for the colour figure.

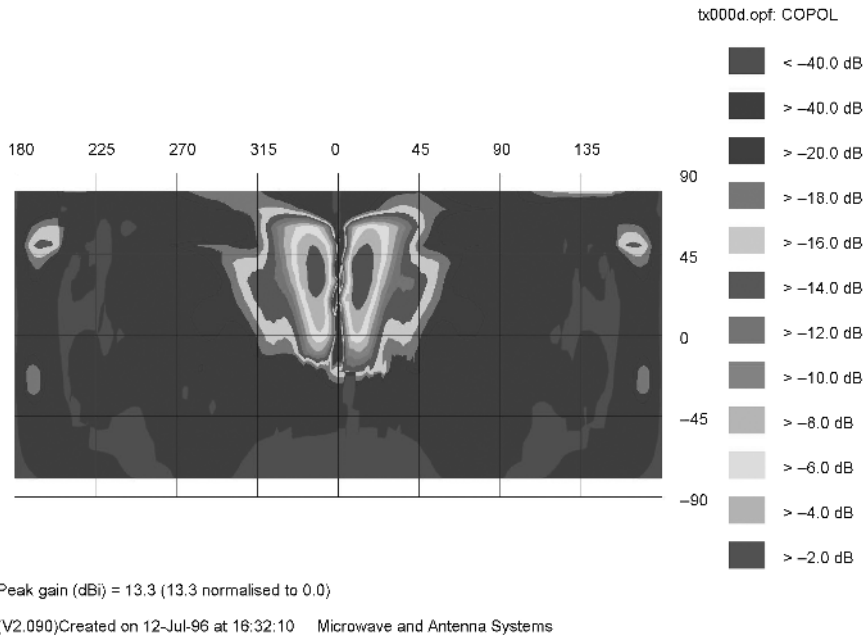


Figure 1.35 Mercator's projection of a contour plot. Reproduced by kind permission of ASL [5]. See Plate 4 for the colour figure.

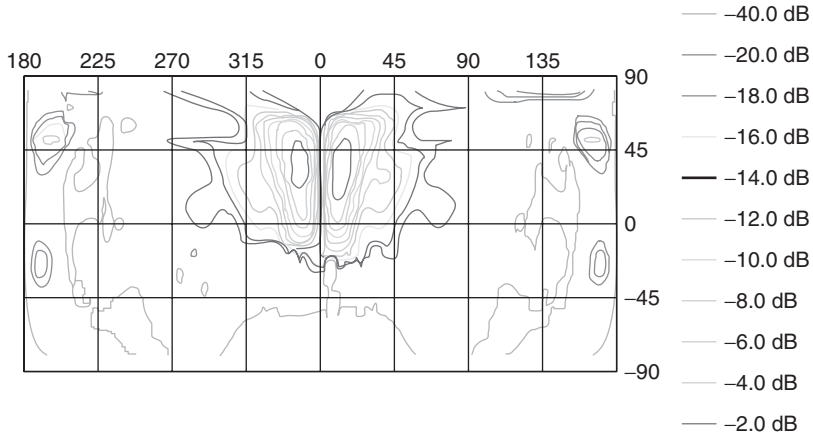


Figure 1.36 Mercator’s projection of the radiation pattern, showing lines of equal gain/power levels. Reproduced by kind permission of ASL [5]. See Plate 5 for the colour figure.

level are plotted as lines (sometimes in different colours) each representing different power levels – see Figure 1.36. Thus each line represents a single different power level and we would not be able to correlate this plot with a contour plot where a number of power levels have the same colour. This looks similar to the contour plots on an Ordnance survey map or a weather map with isobars showing high and low pressures of the atmosphere.

1.5.2 Directivity, Gain and Efficiency

In the case of antennas the gain is not the same as the gain of an amplifier. We can think of the gain of an antenna as its ability to focus the total power in a particular direction or angular sector. An isotropic antenna is analogous to a candle with its light radiating in all directions, whereas a directional antenna is analogous to a torch with its light focused in one particular direction.

One definition of the gain of an antenna relates the power radiated by the antenna to that radiated by an isotropic antenna (which radiates equally in all directions) and is quoted as a linear ratio or in decibels isotropic (dBi). When we say that the gain of an antenna is, for instance, 20 dBi (100 in linear terms), we mean that an isotropic antenna would have to radiate 100 times more power to give the same intensity at the same distance as that of the directional antenna.

Directivity, or directive gain, is the power that would be radiated by an antenna if there were no losses. A tuned antenna such as a half-wave dipole has a resistance of $\sim 73 \Omega$. The RF cable feeding the antenna usually has a characteristic impedance (see Section 1.6.2.1) of 50Ω . Thus there is a mismatch which results in some power being reflected, that is, not all the power is transmitted. Note that domestic RF cables usually have a characteristic impedance of 75Ω and therefore match the impedance of dipole aerials more closely.

The efficiency η of an antenna is the ratio of the power delivered at the terminals of the antenna to the radiated power, expressed as a percentage:

$$\eta = \frac{G}{D} \times 100, \quad (1.12)$$

where

D is the directivity in linear terms

G is the gain in linear terms.

The efficiencies of antennas vary between about 50% and 100%.

If the gain and directivity are in dB then the efficiency would also be in dB and is not usually expressed as a percentage in this case:

$$\eta_{\text{dB}} = G_{\text{dB}} - D_{\text{dB}}. \quad (1.13)$$

Efficiencies of 50% and 100% are 0.5 and 1.0 and, since $10 \log(0.5)$ is -3 and $10 \log(1)$ is zero, the values in dBs would be -3 and 0 dB, respectively.

1.5.3 Electrical and Mechanical Boresight

For directional antennas that have a main lobe the position of the peak radiation is the electrical boresight of the antenna. In the case of an aperture antenna such as a horn or reflector, ideally the electrical boresight would be expected to be along the physical main axis through the centre of the aperture and coincident with the mechanical boresight. However, for the real antenna, as shown in Figure 1.37, this is not the case since the main lobe tends to be slightly squinted and at a small angle to the mechanical boresight. The radiation pattern is often positioned so that its electrical boresight is coincident with the zero angular position of the graph.

1.5.4 Beamwidth and Gain of the Main Lobe

The beamwidth is inversely proportional to the gain of a directional antenna. Thus the narrower the beamwidth the higher the gain on boresight.

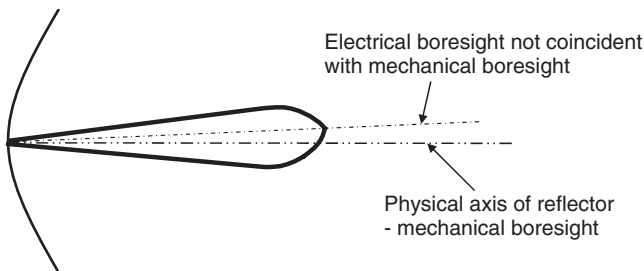


Figure 1.37 Difference between electrical and mechanical boresight.

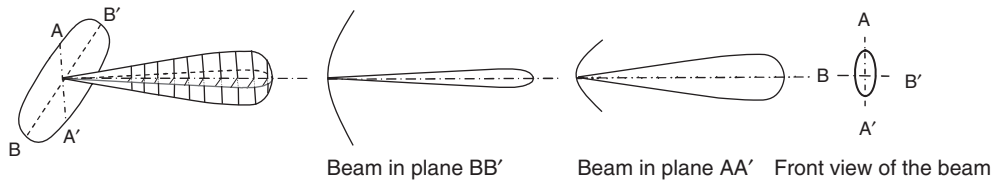


Figure 1.38 Beamwidths of an elliptical reflector antenna.

1.5.4.1 Beamwidth

When we refer to the beamwidth of an antenna we are referring only to the width of the main lobe/beam of the antenna and not the sidelobes. In general, the larger the antenna the smaller is its beamwidth for the same frequency, so that the beamwidth of an antenna is inversely proportional to its physical size. If the antenna does not have the same dimensions in all planes, the plane containing the largest dimension will have the narrowest beamwidth. In the case of a elliptical reflector, as shown in Figure 1.38, the beamwidth in the horizontal or azimuth plane BB' is narrower than that in the vertical or elevation plane, AA', since the major axis of the ellipse is horizontal.

The beamwidth of an antenna can be defined in three ways:

1. half power beamwidth
2. 10 dB beamwidth
3. first null beamwidth.

When the beamwidth is quoted, it is usually assumed that it is the half power or 3 dB beamwidth, that is, the width in degrees (or sometimes in radians) of the main beam across which the gain drops to 3 dB below that of its level at boresight.

In the case of highly directional antennas that have a narrow main lobe, the 10 dB beamwidth is often used, that is, the width at the points on either side of main beam where the radiated power is one-tenth of the maximum value. Large reflector antennas may have gains as high as 60 dBi, that is, linear gains 1 million times greater than that of an isotropic antenna.

For wider beamwidth antennas, the beamwidth is quoted to the first nulls, that is, the width across which the main beam drops to the first nulls. These are depicted in Figure 1.39, where the angle between the 3 dB points K and L, 10 dB points M and N, and first nulls are shown as 25°, 44° and 60°, respectively.

A tuned half wave dipole has a beamwidth of about 78°, whereas a Hertzian dipole has a beamwidth of about 90°.

1.5.4.2 Gain of Main Lobe

The radiation pattern of an antenna shows the power on boresight as 0 dB and the power in other directions as negative values. The gain in all directions is plotted relative to the gain on boresight. Thus we cannot find out the absolute power radiated in a particular direction or at a particular distance, unless we know the absolute gain of the antenna, as

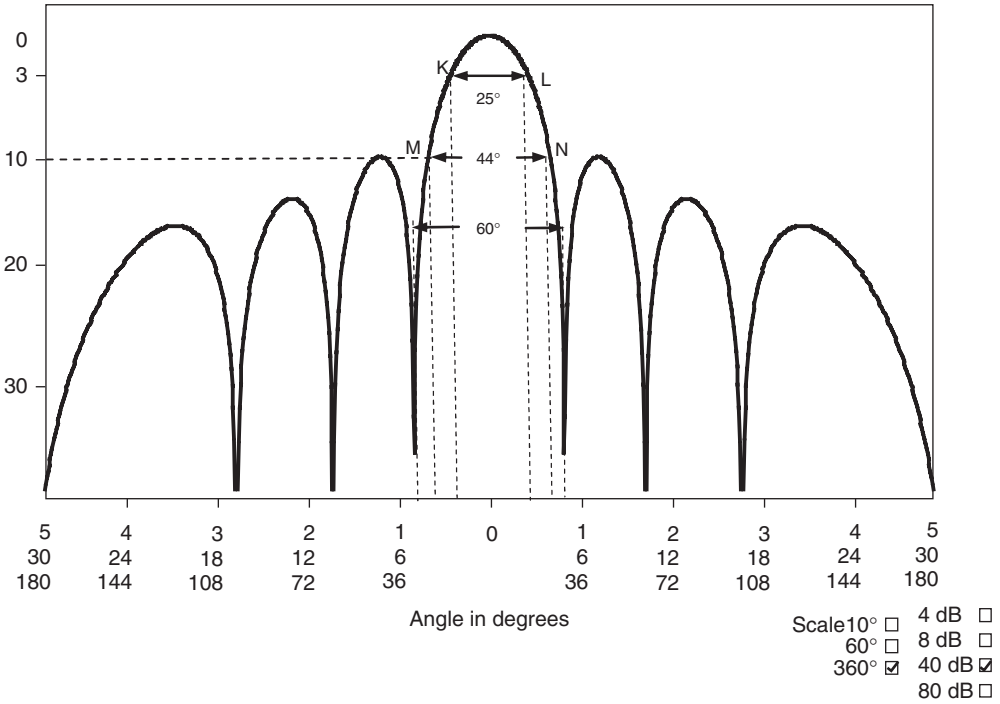


Figure 1.39 Definition of 3 dB and 10 dB beamwidths and the width to the first nulls of an antenna.

well as the power actually radiated by the antenna. In order to find the absolute gain in any particular direction the absolute gain on boresight must be known. If this gain is in decibels then this value can just be added to the gain at any point to give the absolute gain. The absolute gain on boresight is measured using a standard gain horn or other standard antenna. Often the plot is normalized. When a plot is normalized the peak value is set to the maximum (usually zero) shown on the radiation pattern and all other values are adjusted accordingly. Thus, for instance, if the absolute peak gain is 2.3 dBi this value is shown as 0 dB and the level of the near-in sidelobe (see Section 1.5.5) which is actually -14 dBi, would be shown as -16.3 dB ($-14-2.3$) on the plot.

1.5.5 Position and Magnitude of the Lobes

The lobes are the sidelobes as well as the backlobe that is 180° away from the boresight. When sidelobes are referred to, it is often just the near-in sidelobes, that is, the two lobes on either side of the main lobe marked A in Figure 1.40. It is very important to know where sidelobes are, as well as their magnitudes. In the case of a radar antenna, for instance, if a return is received on a sidelobe instead of on the main lobe, the return would be a lower level and thus it could be mistaken for a target that is further away than its true range.

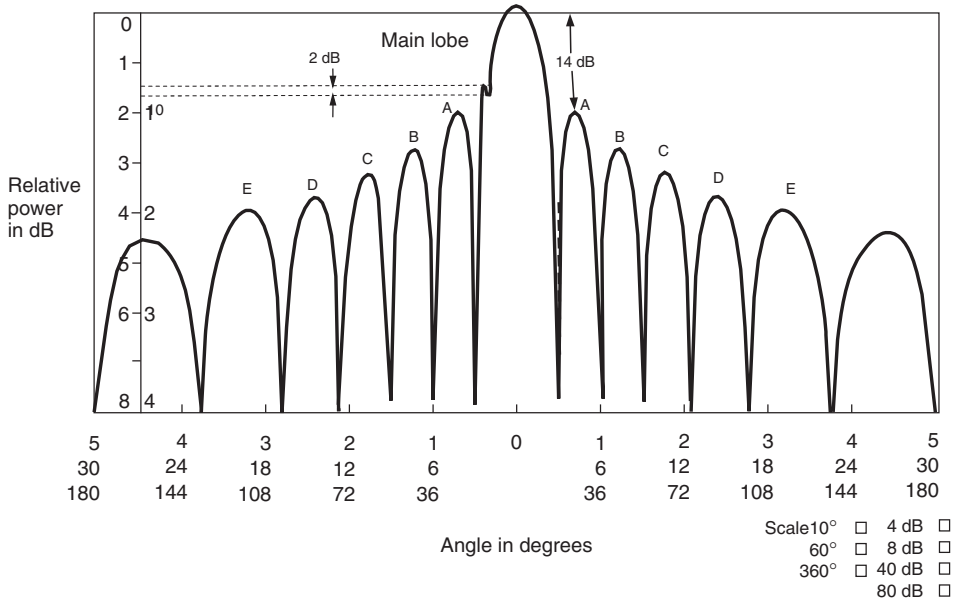


Figure 1.40 The sidelobes (near-in) and definition of sidelobe.

The sidelobe level is the level of the near-in sidelobe below the peak of the main lobe. For instance, in Figure 1.40 the near-in sidelobe marked A is 14 dB below the peak. Thus it is -14 dB compared with the peak. However, it is usually just stated that ‘the sidelobe level is 14 dB’ instead of stating that the near-in sidelobe is 14 dB below the peak.

Measured patterns of directional antennas are often not smooth or symmetrical and may have several dips and peaks. Thus it may be difficult to distinguish between a sidelobe and an irregularity of the radiation pattern. In some cases a sidelobe is defined as the case where the difference between a peak and a null is greater than 2 dB. For instance, in Figure 1.40 there is an irregularity shown where the difference is 2 dB, and therefore this would not be considered a sidelobe.

1.5.6 Bandwidth

When we refer to the bandwidth, we are usually referring to the frequency bandwidth. The IEEE defines the frequency band limits as exclusive of the lower limit and inclusive of the upper limit. Thus a frequency band of 3–30 MHz refers to frequencies above 3 and up to 30 MHz. An antenna has its maximum gain at its tuned frequency, but can still receive other frequencies, albeit with reduced gain. The response of an ideal tuned antenna over its frequency band is similar to that of the current in an inductance-capacitance-resistance (LCR) circuit, as shown in Figure 1.41. In the case of an antenna (as in LCR circuits) the resonant frequency is the point at which the impedance is purely resistive. The bandwidth is commonly defined as the frequency band over which the linear gain of the antenna is at least half of its gain at its resonant frequency. Expressed in decibels, the bandwidth would

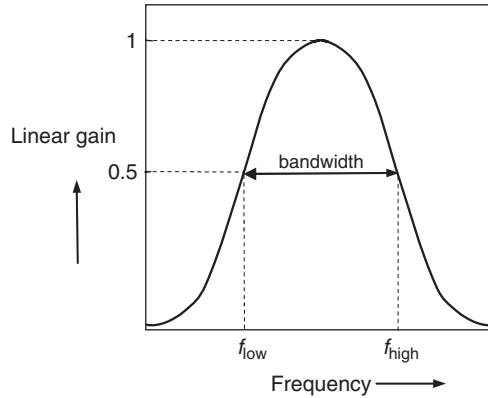


Figure 1.41 Frequency response of an ideal antenna.

be the frequency band over which the gain of the antenna is at least -3 dB compared with its resonant frequency gain. Thus this definition is similar to that of the beamwidth of an antenna, although we are dealing with frequency instead of angles. Note that since the gain has the dimensions of power, when the power is halved the electric field is $1/\sqrt{2}$ (0.707) times that of the electric field at resonance.

A wide bandwidth is usually achieved at the expense of gain. Thus if the antenna had the same size but a narrower bandwidth, its gain would be higher.

The bandwidth can be expressed:

1. by quoting the actual frequencies
2. as a percentage or as a fraction
3. as a multiple of an octave.

1.5.6.1 Bandwidth by Quoting the Absolute Frequencies

The absolute bandwidth is specified by quoting the upper and lower frequencies, or by quoting the midband frequency plus or minus (\pm) half the frequency band. Thus if the bandwidth of the antenna is from 300 to 1000 MHz, these frequencies are quoted, or the bandwidth is expressed as the centre frequency (arithmetic mean of the upper and lower frequencies) and the difference between the centre and band edge frequencies, that is, 650 ± 350 MHz.

1.5.6.2 Bandwidth as a Percentage or Fraction

When it is expressed as a percentage bandwidth B_w , its centre frequency f should be quoted, and the percentage expressed relative to this centre frequency:

$$B_w = \left(\frac{\Delta f}{f} \right) \times 100, \quad (1.14)$$

where Δf is the difference between the upper and lower frequencies.

The bandwidth can be expressed as a percentage of its total frequency span or \pm half the percentage of its frequency span, but we have to quote its centre frequency. Thus, for instance, if we have an antenna that is quoted as having a bandwidth from 100 to 300 MHz, its centre frequency f would be 200 MHz and its absolute bandwidth or frequency span is 200 MHz. Using Equation 1.14, B_w would be 100%, so we could either say that the antenna has a centre frequency of 200 MHz and its bandwidth is 100% or its bandwidth is $\pm 50\%$.

Similarly, if an antenna has an operating frequency from 20 to 40 MHz, its centre frequency is 30 MHz and its bandwidth is 66.7% (20/30) or $\pm 33.3\%$.

1.5.6.3 Bandwidth as a Fraction or Multiple of an Octave

When it is expressed in octaves, its lower and upper frequencies should be quoted. An octave is a band of frequencies between one frequency and another frequency that is double or half the first frequency. Thus, for instance, we have an octave between 100 and 200 MHz, and two octaves between 100 and 400 MHz. In this case we would not use the centre frequency since we would be unable to quote the bandwidth in terms of octaves.

The bandwidth as a fraction or multiple of an octave is defined as

$$B_w = \log_2 \left(\frac{f_{\text{high}}}{f_{\text{low}}} \right), \quad (1.15)$$

where f_{high} is the highest frequency and f_{low} is the lowest frequency. Thus, for instance, if we have an antenna that is quoted to have a bandwidth from 100 to 300 MHz, its centre frequency f would be 200 MHz and its absolute bandwidth is 200 MHz. The ratio of the highest to lowest frequencies is 3, and $\log_2 3$ is 1.58.

Many calculators do not have logarithms to the base 2. In this case we can use logarithms to the base 10 of the ratio and divide the result by logarithms to the base 10 of 2. Thus $\log_{10} 3$ is 0.477 and if this is divided by $\log_{10} 2$ ($=0.301$) we get 1.58.

1.5.7 Polarization

The antenna has to be oriented so that it can receive the maximum electric field. When this occurs for a monopole, for instance, the antenna is said to be receiving the copolar field. If the monopole is then turned through 90° (at the same position) so that it receives the minimum electric field, it is said to be receiving the cross-polar field. In the case of a monopole antenna, when the antenna is vertical it would receive the maximum electric field of a vertically polarized incident wave and therefore this would be the copolar radiation. A horizontally polarized wave would be the cross-polar radiation. However, in the case of other linearly polarized antennas such as aperture antennas, the copolar orientation cannot always be determined by visual inspection and prior knowledge may be required.

All polarizations could be considered to be variations of elliptical polarization. In the case of elliptical polarization, the magnitude as well as the direction of the electric field will vary during a cycle.

Details of the different types of polarizations are given in Section 1.4.

1.5.8 Power Handling

The power that an antenna can handle depends mainly on the structure of the antenna, its feed or feeding network, and its frequency. Printed circuit board antennas, such as patches, cannot handle powers much above 5 W and are prone to delamination on aircraft at powers approaching this magnitude. In general, the higher the frequency the less power it can handle in the case of printed circuit antennas. This is because the substrate is thinner (in order to meet the characteristic impedance requirements) in the case of the higher frequency antennas, and hence the breakdown voltage is lower.

1.6 Propagation

EM waves can be guided along transmission lines or travel through free space. These waves, composed of electric and magnetic fields, change periodically in time and have clearly defined configurations that satisfy the boundary conditions of Maxwell's equations. The boundary conditions are those that would occur at the interface when a wave is travelling from one medium to a second medium. For instance, the electric field would reduce to zero at a perfect conductor. A qualitative explanation of the boundary conditions is given in Chapter 3 of [3].

In free space the electric and magnetic fields are at right angles to each other and to the direction of propagation of the wave. This type of wave is known as a transverse electromagnetic (TEM) wave, and it can also be propagated as a guided wave in any two-conductor transmission line, such as parallel wires, coaxial lines and striplines. However, a TEM wave cannot be supported by waveguides. The TEM wave is the most common mode and hence it is the main one discussed in this book.

1.6.1 Power Flux Density

In the case of the TEM wave, the electric and magnetic fields are perpendicular to each other and the plane containing them is perpendicular to the direction of propagation of the EM energy. The vector cross-product of the electric and magnetic fields is also a vector which is equal to the power flux density (power per square area) known as Poynting's vector P_d . For a simple explanation of the vector cross-product see Chapter 2 of [3].

We can see that if we multiply V/m by A/m, we would get VA/m², which is W/m². The relationship between the three vectors shown in Figure 1.42b is defined by the equation

$$P_d = E \times H, \quad (1.16)$$

where

P_d is in W/m²

E is the electric field vector in V/m

\times represents the cross-product and

H is the magnetic field vector in A/m.

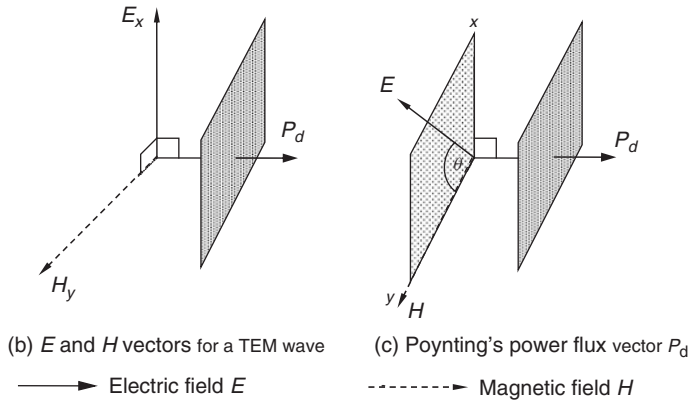


Figure 1.42 Wave propagation and Poynting's power flux density vector.

The magnitude of P_d is given by

$$|P_d| = |E||H| \sin \theta, \quad (1.17)$$

where

$|E|$ is the magnitude of E

$|H|$ is the magnitude of H and

θ is the angle between the electric and magnetic field vectors.

Note that the magnitudes of the electric and magnetic fields are those in the plane perpendicular (i.e. the transverse plane) to the direction of propagation or power flow. The maximum values of these fields occur when the E and H vectors are in this transverse plane; any component parallel to the direction of propagation will not contribute to the power flux density. We can see from Equation 1.17 that algebraically the maximum value of P_d occurs when the sine of the angle between the electric and magnetic fields is 1, that is, angle θ is 90° or the fields are perpendicular to each other in the plane transverse to the direction of the propagation of the wave. Thus the TEM wave (which has these fields perpendicular to each other) has the maximum value of Poynting's vector. The power density through a surface S , as shown in Figure 1.42, has a maximum value when Poynting's vector is perpendicular to it.

1.6.1.1 Wave Impedance

The EM wave can be considered to have an impedance, depending on its configuration. If the wave is incident on a surface with the same impedance the surface can be said to be matched, in the same way as a load is matched to a transmission line. When this occurs, no energy will be reflected and the wave is totally absorbed/transmitted. This impedance

is known as the intrinsic or characteristic impedance Z_w of the wave and in the case of a TEM wave in a medium it is given by

$$Z_w = \frac{E_x}{H_y} = \sqrt{\frac{\mu}{\varepsilon} \left(\frac{1}{1 + \sigma/j\omega\varepsilon} \right)}, \quad (1.18)$$

where

Z_w is in ohms

E_x is the electric field along the x axis in V/m

H_y is the magnetic field along the y axis in A/m

μ is the permeability of the medium in H/m

ε is the permittivity of the medium in F/m and

σ is the conductivity of the medium in S-m.

If the TEM wave is in free space, $\mu = \mu_0$ and $\varepsilon = \varepsilon_0$ and the conductivity is zero. Thus the characteristic impedance Z_0 of TEM wave in free space is given by

$$Z_0 = \sqrt{\frac{\mu_0}{\varepsilon_0}} = 120\pi, \quad (1.19)$$

since $\mu_0 = 4\pi \times 10^{-7}$ H/m and $\varepsilon_0 = (36\pi \times 10^9)^{-1}$ F/m. The characteristic wave impedance is often quoted at its approximate value of 377 Ω .

1.6.2 Guided Waves

Transmission lines that can support TEM waves require two separate conductors. The electric (solid lines) and magnetic (dotted lines) fields for circular coaxial transmission lines are shown in Figure 1.43. Note that when the magnetic lines are evenly spaced, they will also represent the lines of equipotential (with the electric field lines perpendicular to them), and the closer the spacing of the electric field lines, the greater is the magnitude of the electric field. The electric field variation with distance in the direction of propagation (longitudinal direction) is sinusoidal. We should also note that there is a sinusoidal

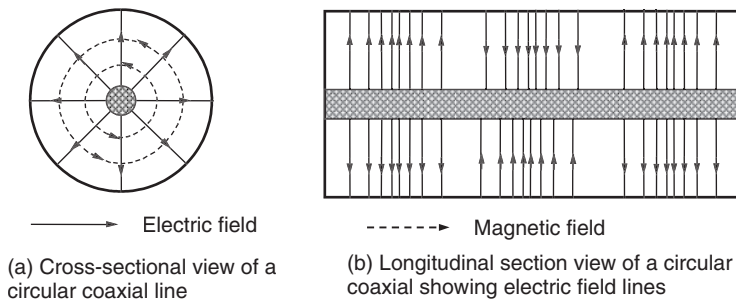


Figure 1.43 The electric and magnetic fields in a coaxial transmission line.

variation with time, which effectively moves this electric field pattern forward by half a sine variation every half period in time.

The coaxial line is one of the most efficient ways of containing the EM energy. However, the spacing between the inner and outer conductors in a coaxial line is maintained by dielectric (in the form of beads or a continuous hollow cylinder). As the frequency increases, the losses in the dielectric also increase and the energy is increasingly attenuated. Another source of energy loss is the outer conductor, which is often of braided form. Energy can leak through holes in the braid, and the amount of the leakage is proportional to the electrical size of the hole, that is, the size of the hole as a fraction of the wavelength. Energy can escape through gaps that are as small as 0.01 times the wavelength. As the frequency increases the wavelength decreases and the gaps in the outer conductor become a larger fraction of this wavelength, causing the EM energy loss to increase.

1.6.2.1 Characteristic Impedance of Transmission Lines

When a line is greater than one-tenth of a wavelength at its highest operating frequency, we cannot ignore the properties of the line. The current distribution in it (due to the electromotive force at the input end) is not uniform. The line behaves as though it has resistive and reactive components distributed along its length. Each section of the ‘go’ and ‘return’ cables forms an unit loop, which can be represented by a shunt capacitance C and conductance G between the two cables; and a series resistance R and inductance L . If the line is a pair of parallel identical conductors, it is balanced and the resistance R and inductance L are equally divided between each cable, as shown in Figure 1.44. Normally an incident wave, travelling from left to right towards the termination, is partly reflected. The reflected and incident waves combine to form a standing wave. If the termination is ‘matched’ (made equal) to the impedance of the unit loop of the line, there is no reflected wave, and hence no standing wave.

The impedance of this unit loop is known as the characteristic impedance Z_0 and is given by

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}}, \tag{1.20}$$

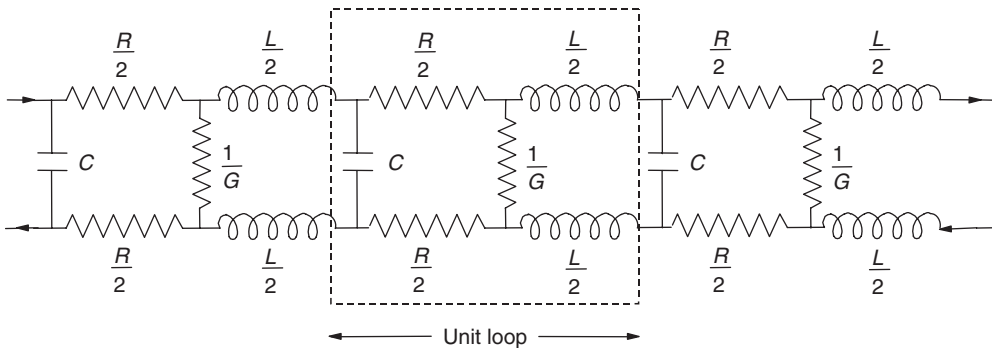


Figure 1.44 Equivalent circuit of a balanced transmission line.

where

R is the series resistance in Ω

G is the shunt resistance in S

L is the series inductance in H and

C is the shunt capacitance in F.

If the line is lossless the resistive parts are zero, so that $R = 0$ and $G = 0$.

Equation 1.20 for the characteristic impedance reduces to

$$Z_0 = \sqrt{\frac{L}{C}}. \quad (1.21)$$

This is the general formula for a lossless transmission line, and the magnitudes of L and C will depend on the shape and size of the particular transmission line and the dielectric material.

1.6.2.2 Matching Transmission Lines

When two transmission lines are connected together, they should have the same characteristic impedance. For domestic use the characteristic impedance of coaxial cable (used to connect a television to the antenna, for instance) is 75Ω , but for professional/industrial use the characteristic impedance is 50Ω . If the connected cables do not have the same characteristic impedance there is a mismatch and some of the power is reflected back towards the incident power. Thus in the case of an aircraft system, if a transmitter is connected using cables of the incorrect characteristic impedance the power reaching the antenna will be greatly reduced. Similarly, if the antenna is connected to a cable of the incorrect characteristic impedance, some of the radiated power reaching the antenna from outside the aircraft will be reflected back, thus reducing the sensitivity of the entire system.

When an antenna is electrically very small, that is, small in terms of wavelengths, its impedance changes with frequency, and at some frequencies it has a negative reactance (capacitance) and at other frequencies it has a positive reactance (inductance). In these cases there is a matching unit, called a tuner, to prevent a large mismatch at the antenna terminals (where it is connected to the RF cable). This tuner provides a conjugate match as well as an impedance transformation between the antenna and the RF cable.

The complex conjugate of a number has the same real part as the number but its imaginary part has sign opposite to that of the number. For instance, if we have a complex number $S = a + jb$, its complex conjugate is $S^* = a - jb$. If the antenna has an impedance of, say, $R + jX$ at a particular frequency, the tuner is adjusted to have an impedance of $R - jX$, where R is the resistance and X is the reactance. A positive reactance is obtained from an inductor and a negative reactance is obtained from a capacitor.

The same problem would occur if we have a load which has a susceptance (1/reactance) of $+jB$ in parallel with its conductance of $G_l = 0.2 \text{ S}$, and we want to connect it to a line whose characteristic impedance is 50Ω (i.e. a characteristic admittance of 0.2 S). Since the load does not have the same impedance as the transmission line, some of the

transmitted wave would be reflected, and, together with the incident wave, this would result in a standing wave. This would result in a loss of power transfer, and in some cases could lead to voltage breakdown in high power systems. In order to match the load to the line, we would connect a stub, which is a short-circuited line, in parallel with the transmission line, as near to the load as possible. The conductance of the load can be varied by adjusting the position of the stub so that the conductance presented to the line equals 0.2 S . The susceptance of the stub can be varied by adjusting its length so that its susceptance is equal and opposite to that of the load, that is, $-jB$. Thus the resultant susceptance is zero, and the conductance is 0.2 S . The load now presents an impedance that is the same as that of the transmission line so that it is matched, and maximum power transfer can be effected.

1.6.2.3 Relationship between Power Density and Electric Field Strength

In the far field the electric and magnetic fields are perpendicular to each other and the plane containing them is normal to the direction of propagation. The following relationship between the power density and the power radiated by the antenna only applies to the far field. If the formula is used in the near field, we would get an infinite magnitude for the power density when the distance d is zero.

The power density P_d at a distance d in the far field is given by

$$P_d = \frac{P_t G_t}{4\pi d^2} \quad (1.22)$$

where

P_d is the power density in W m^{-2}

P_t is the power transmitted by the antenna in W

G_t is the linear or numeric gain and

d is the distance in m .

For omnidirectional antennas the gain G_t is constant in the azimuth plane, but for directive antennas the gain is a function of the angle relative to boresight. The magnitude of the power density is also equal to the real part of the cross-product of the E and H field vectors,

$$P_d = \text{Re}(E \times H), \quad (1.23)$$

where

E is the electric field intensity in V/m

H is the magnetic field intensity in A/m and

P_d is the power density in W m^{-2} .

The intrinsic wave impedance ξ of a plane wave is the ratio of the transverse E and H fields. Substituting E/ξ for H gives

$$P_d = \frac{E^2}{\xi} \quad (1.24)$$

for the power density. Combining Equations 1.22 and 1.24, we get the following expression for the electric field:

$$E = \sqrt{\frac{P_t G_t \xi}{4\pi d^2}}. \quad (1.25)$$

For a plane wave the intrinsic wave impedance is equal to 120π , and thus Equation 1.25 can be rewritten as

$$E = \sqrt{\frac{30P_t G_t}{d^2}}. \quad (1.26)$$

This expression gives us the electric field at a distance d from the transmit antenna. In the case of a directional antenna the gain will be a function of the angle relative to its boresight.

1.6.3 Free Space Waves

The propagation of the EM wave depends on the frequency, the polarization, the medium it traverses and, in the case of free space waves, whether the path is over land or sea.

The energy from a transmitting antenna reaches the receiving antenna by traversing several paths and by the mechanisms of reflection, refraction and diffraction ([5], p. 608). Waves reaching the receiving antenna can be one or all of the following types:

1. ionospheric or sky waves that are reflected or scattered by the ionosphere, which is 100–300 km above the earth's surface.
2. tropospheric waves reflected or scattered by the troposphere, which is 10 km above the earth's surface
3. ground waves.

These paths are shown in Figure 1.45.

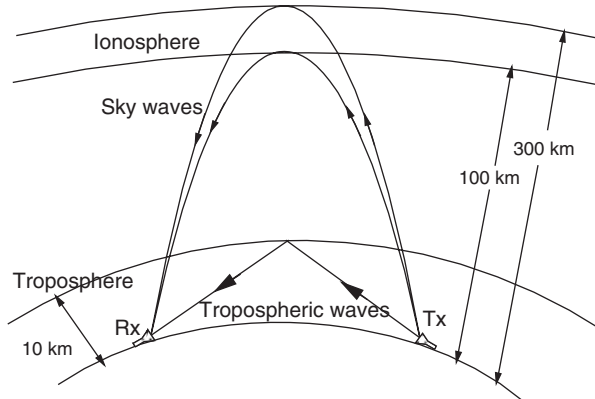


Figure 1.45 Sky and tropospheric waves propagating in free space.

1.6.3.1 Sky Waves

Sky waves are refracted by the same phenomenon of total internal reflection that occurs with light waves travelling from one medium to another with lower refractive index, as described in Section 1.2.2.1.

Sky wave propagation is used for high frequency communications, but above 30 MHz the refraction is insufficient for this type of propagation to be used. Horizontal polarization is used in the frequency range 1.5–30 MHz for ‘sky wave’ propagation. Sky waves are reflected by the ionosphere, which is 100–300 km above the earth’s surface, and the round trip back to earth is 200–600 km.

In Figure 1.45 we can see that the receiver is not in line of sight (LOS) of the transmitter, so normally the signal received would be a result of diffraction alone and hence would be very low. However, because the waves are reflected off the upper layers of the atmosphere, the signals can effectively be ‘seen’ from a greater distance than would be the case for LOS. This increased distance is commonly called the radar horizon.

The additional path travelled by the wave can be considered as a direct path to the horizon of an earth with a modified radius of ka , where a is the true radius of the earth, and k is given by

$$k = \frac{1}{1 + a \frac{dN}{dh} 10^{-6}}, \quad (1.27)$$

where

a is the true radius of the earth in km and

dN/dh is the rate of change of the refractive index with height in km^{-1} .

For a dN/dh value of -40 km^{-1} and taking the earth’s radius as 6370 km, the value of k is 1.34 or approximately $4/3$. Tropospheric propagation permits transmission beyond LOS between transmit and receive antennas. However, it entails the use of elaborate and expensive equipment, because the transmission efficiency is poor (see [6], p. 4). The extra loss is caused by the longer paths taken by the EM waves.

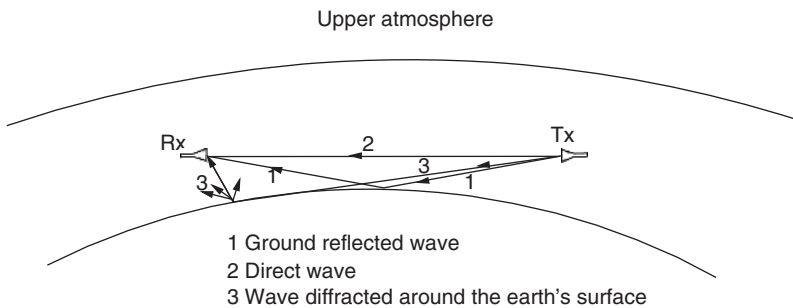


Figure 1.46 Space waves.

1.6.3.2 Ground Waves

Ground waves are made up of space and surface waves. Space waves predominate at high distances above the earth's surface, whereas surface waves are more important when an antenna is near the earth's surface.

1.6.3.3 Space Waves

A space wave is made up of three waves:

1. the ground reflected wave, by specular or diffuse reflection from the ground
2. the direct path between the transmit and receive antennas
3. the wave diffracted around the earth's surface.

Vertical polarization is used below 1.5 MHz for ground wave propagation. Ground wave propagation is also used for horizontally polarized waves, with the electric field perpendicular to the plane of incidence and parallel to the surface. The reflected wave is nearly 180° out of phase with the incident wave.

As the angle of incidence is increased the magnitude and phase of the reflection factor will change but not to any large extent. The change is greater for high frequencies and for lower ground conductivities.

When the angle of incidence is near to grazing (i.e. around 90°) the magnitude of the reflected wave is nearly equal to that of the incident wave for all frequencies and for all ground conductivities ([5], p. 612).

References

- [1] Tooth, A.R. (2004) IPAS first year activity report, Deliverable D2C, IPAS/PAR/BAES/ART 041101, 1 November. Installed Performance of Antennas on AeroStructures (IPAS), an EU Specific Targeted Research Project, November 2003 to January 2007, Contract No AST3-CT-2003-503611.
- [2] Thain, A. (2006) Computation and analysis of existing and advanced tools, Deliverable D17, IPAS/RP/EADS/AT 061219, 19 December. Installed Performance of Antennas on AeroStructures (IPAS), an EU Specific Targeted Research Project, November 2003 to January 2007, Contract No AST3-CT-2003-503611.
- [3] Macnamara, T.M. (1995) *Handbook of Antennas for EMC*, Artech House, Boston, ISBN 0-89006-549-7.
- [4] ASL (Antenna Software Ltd) P.R. Foster.
- [5] Jordan, E.C. (1953) *Electromagnetic Waves and Radiating Systems*, Constable and Company Ltd.
- [6] Powell, J. (1985) *Aircraft Radio Systems*, Pitman, London.