Millimetre-wave Optics, Devices and Systems

J. Lesurf

2018 free PDF version (first 7 chapters so far!)
New Introduction (2018)

This book was initially written to serve as a general ‘research level introduction’ to the field specified in the title. I wrote it to save my having to repeatedly pass on the same basic knowledge to each new research student, etc, who came to work on the topics. The original version was published in 1990 by the UK Institute of Physics (IoP) via their Adam Hilger imprint. Some years later the IoP decided to cease book publication. Most of their titles were sold on to another publisher, but I took back the rights to the book at that time, and since then the book has been out of print.

This is a newly regenerated version of the book which is freely downloadable. The fundamentals have not changed, although the details of the current technology and its performance will be much improved and extended compared with what was possible in 1990! Despite that, it may prove useful as a basic reference wrt the concepts and physics involved for anyone new to the topics covered.

This version is not identical to the printed book. That was written in a past era before modern desktop publishing software was available that could easily cope with equations, etc. Nor in those days could I email the results as a PDF or PostScript file to be edited and printed directly. The original manuscript was printed out on paper, then posted to the publishers. They then went thought it making corrections in red pen, and sent the result back to me for me to make further corrections. The result was then sent back to the printers who duly typeset the book and printed it. One result of this process is that I have not seen the final corrected manuscript since, and I don’t have any computer files which contain the precise printed text/equations/etc. All I have is some files of a ‘draft’ version of the text and equations made with a text editor I no longer use. Plus a copy of the printed book and the – hand drawn in pen on drafting film – illustrations.

As a result, for this version I have had to reprocess the original text files, re-type the equations, and scan the old drawings to reconstruct what you see. Since the text was an early draft and I’m retyping the equations, this version will have some errors, etc, which were removed for the printed version before it was printed! I have tried to find and correct these. And have in a few places altered the text for the sake of clarity. But I am sure some errors will remain.

The bottom line is that I then offer this version, free, ‘as is’, in the hope some will find it of interest. Apart from the main intent of providing an introduction for research students the book did have one other objective. This was encapsulated in Chapter 10 on the design of ‘optical circuits’. The intent was to make clear the concept that arrangements of optical elements can be treated in a way similar to electronics circuits. From a suitable arrangement and selection of optical ‘components’ an arbitrarily complex information/signal processing ‘circuit’ can be constructed. As in conventional electronics, only a few basic types of ‘component’ are needed. The function then depends on the way these are selected and connected together – in this case by mm-wave beams of EM radiation rather than via wires. Instead of resistors, capacitors, etc, think of polarisers, roof-mirrors, etc.

Jim Lesurf 2018
Signal transmission, modes, and Gaussian Beams.

1.1 Waveguides and Mode Properties

Conventional electronic instruments use wires to send signals from place to place. As the signal frequency rises it becomes increasingly difficult to make wired systems which work well. This is because - as was mentioned in the introduction - the actual signal is transmitted as an electromagnetic field which moves along just outside the wire. Some of this field can, therefore, be radiated away or coupled onto any other nearby wires. Changes in the potential of a wire can be detected as a varying force on nearby charges. Similarly, changes in current alter the surrounding magnetic field and may induce currents on other wires. These effects are normally dealt with in electronics as 'stray' capacitances and inductances. They mean that some of the power we wish to transmit via a wire may fail to arrive at its intended destination because it has been diverted elsewhere. It may also mean that some of the variations in potential and current arriving at the signal destination are the result of unwanted signals coupled onto the wire.

In many cases the signal wavelengths are far larger than the lengths of the transmitting wires and we can think of variations in current and charge as being uniform along the wire. If the frequency is increased sufficiently (or the wire extended) this assumption ceases to be reliable. Then the potential and current may be seen to vary along the length of the wire - i.e. there is a noticeable electric field and variable magnetic field along the wire. Both of these fields vary periodically in time at the signal frequency. Now the wire will act as an antenna, radiating some signal power into the surrounding space. As a result of these effects the efficiency of signal coupling along a wire tends to fall as the frequency rises. Various measures can be adopted to try and counteract these problems. One of the most useful is to replace the wires with metallic waveguide. Although mainly used at microwave frequencies metal waveguide is worth discussing in some detail here as it is used in many mm-wave systems and components. Many of the properties of waves propagating in guides of this type also turn out to be applicable in general to other forms of waveguide and to beams in space.

If we regard a normal wire as a length of metal surrounded by space we can think of metal waveguide as a length of space surrounded by metal. As with the wire, signal power is transmitted as an electromagnetic wave which moves along in the space outside the metal - i.e. the wave moves down the hole in the center of the guide. The most common form of waveguide is rectangular in cross-section. Consider Fig 1.1 which represents a rectangular guide.

![Figure 1.1 A Short length of rectangular waveguide.](image)
In order to determine the field pattern inside the waveguide we have to find the appropriate solution for Maxwell's equations. We need to know at least some of the field components at some point within the guide. How can we do this for a rectangular metal waveguide?

A rectangular waveguide can be regarded simply as a set of four mirrors placed so as to form a long metal box with open ends. With this in mind we can understand its behaviour by considering what happens when an electromagnetic wave is incident upon a metal surface. The incident wave sets up a current in a thin layer at the surface. If the wave is a plane parallel one moving perpendicular to the surface then the current distribution is uniform over the surface. Hence there will be no net accumulation of charge at any point on the surface. No potential differences are produced and the electric field along the surface must remain zero everywhere.

The velocity of light inside a metal is generally far lower than in free space - i.e. the refractive index is very high. Hence very little of the incident wave penetrates into the metal. In order for the incident energy not to vanish mysteriously it must be reflected. Since the total electric field along the surface must be zero it follows that the incident and reflected waves have equal and opposite electric fields at the surface. A wave incident upon a metal surface at an angle can be regarded as a combination of two waves arriving simultaneously - one perpendicular to the surface, the other parallel to it. The effect of the perpendicular wave is as described. The behaviour of the wave moving parallel to the surface depends upon the orientation of its electric field with respect to the metal surface. However, any current flow set up in the metal will be such that there can be no electric field at the surface perpendicular to the direction of propagation of the wave.

For the case of a set of mirrors arranged to form a rectangular guide we find that we cannot produce an electric field at a guide wall which is both parallel to the wall and perpendicular to the axis of propagation along the guide. We can consider the signal power propagating along the waveguide as being guided by repeated reflection between opposite walls. Although this is a simple model of what takes place it serves to introduce some of the basic properties of signal transmission using a metallic waveguide.

The first property we may expect is that the electric field component parallel to a metal wall should be zero at the wall surface. We can also expect the rate of progress of a signal along the guide to be less than for a wave in free space. This is because the signal does not simply move parallel to the guide axis but takes a longer path, 'tacking' along the guide like a yacht travelling upwind. From knowing that particular electric field components must be zero at the walls we can find the appropriate solutions for Maxwell's equations and determine the actual field pattern within the guide. For a rectangular waveguide the solutions fall into two general classes, referred to as Transverse Electric (TE) and Transverse Magnetic (TM) waves. The electric field of a TE wave is everywhere perpendicular to the guide axis of propagation. Similarly, the magnetic field of a TM wave is everywhere perpendicular to the guide axis.

It is useful to examine the properties of these waves in a little detail because this will enable us to see some of the basic properties which are common to all wave guiding systems - even those consisting of a beam in space. Most metallic waveguide systems use TE waves, hence we can use them for the purpose of example. The solutions of Maxwell's equations for TE waves in a rectangular metal waveguide are of the general form.
\[ E_x = \frac{j\omega H_0 n\pi}{k_c^2 b} \cos \left( \frac{m\pi x}{a} \right) \sin \left( \frac{n\pi y}{b} \right) \exp \{ j(\omega t - \beta z) \} \]  
\[ E_y = -\frac{j\omega H_0 m\pi}{k_c^2 a} \sin \left( \frac{m\pi x}{a} \right) \cos \left( \frac{n\pi y}{b} \right) \exp \{ j(\omega t - \beta z) \} \]  
\[ E_z = 0 \]

Where
\[ k_c^2 = \left( \frac{m\pi}{a} \right)^2 + \left( \frac{n\pi}{b} \right)^2 \]
and
\[ \beta^2 = \omega^2 \mu_0 \varepsilon_0 - k_c^2 \]

The Cartesian co-ordinates are defined as shown in Figure 1 with the \( z \) axis parallel to the guide propagation axis. The guide width and height are \( a \) & \( b \); \( \varepsilon_0 \) and \( \mu_0 \) are the permittivity and permeability of free space; \( t \) is the time; \( \omega \) is \( 2\pi \) times the signal frequency; and \( H_0 \) is a measure of the maximum magnetic field in the guide. Hence the amount of power transported varies with \( |H_0|^2 \). In general \( H_0 \) is a complex number which may also be used to specify the phase of the wave when \( t = 0 \).

\( \beta \) is called the propagation constant and defines how the phase of the wave varies along the length of the guide. \( m \) and \( n \) are non-negative integers which are called the mode numbers. In fact, in metal waveguides of this type a mode with \( m = n = 0 \) cannot exist because this would require a non-zero electric field along a guide wall. From the expressions given above we can deduce some of the general properties of guided waves. It can be seen that there are a range of possible TE waves corresponding to the choice of \( m \) & \( n \). These are conventionally distinguished by calling each solution for particular values the TE\( mn \) mode. It can also be seen that each mode has a unique field distribution in the \( x,y \) plane and that the form of this distribution does not change as the mode propagates along the guide. It is also important to note that the propagation constant depends upon the mode numbers. As a consequence the effective wavelength along the guide varies from mode to mode.

It is a general property of the solutions of a differential equation that, if \( A_1, A_2, A_3, \) etc, are solutions then any linear combination of them is also a solution. Hence the wave in a guide may consist of various amounts of a number of modes, each moving along with its own effective wavelength and keeping its own characteristic mode field pattern.

For most practical purposes it is undesirable to have a multi-mode signal. Although each mode propagates keeping its own field distribution the total field at any point in the guide would be produced by adding together the contributions of the various modes present. As the mode propagation constants differ this means that the actual field distribution at a given plane depends upon where we are placed along the guide. A signal sensor placed in the guide will have been designed to respond to a particular field pattern. If the transmitted wave does not match the desired pattern the sensor will not work efficiently and some signal power will be lost. Furthermore, the propagation constants of the modes are all frequency dependent. Hence the signal losses will display a frequency sensitivity which makes the sensor difficult to use over a wide frequency range.

In principle it is possible to design a sensor which matches the multi-mode field in some specific circumstances, but in general it is preferable to produce a system where only one mode can be transmitted. This makes design much simpler and usually means that better performance can
obtained with much less effort. For rectangular metallic waveguide we find that the propagation constant is zero when the signal frequency is such that

\[ \omega^2 \mu_0 \varepsilon_0 = k_0^2 \]  

... (1.6)

At frequencies below this value the propagation constant is imaginary. This means that the field in the guide decays exponentially with z rather than being a sinusoid - i.e. the wave cannot propagate along the guide. The mode is said to be cut off at this frequency. The frequency at which \( \beta \) becomes zero depends upon the guide width & height, and also upon the mode numbers. Microwave engineers have exploited this behaviour to develop what is referred to as standard rectangular waveguide. This has a ratio of 2:1 between its width and height.

We can characterise the cut-off point of a mode in terms of the wavelength in free space which corresponds to the frequency at which the propagation constant becomes zero. If we choose a guide where \( a = 2b \) the cut-off wavelengths, \( \lambda_c \), of the first few modes will be

\[ m = 1 \quad n = 0 \quad \Rightarrow \quad \lambda_c = 2a \]
\[ m = 0 \quad n = 1 \quad \Rightarrow \quad \lambda_c = a \]
\[ m = 1 \quad n = 1 \quad \Rightarrow \quad \lambda_c = (0.89...) \lambda_c \]

As the mode numbers increase so the cut-off wavelength reduces. In the wavelength range, \( 2a > \lambda > a \), only one mode is possible. Over the corresponding range of frequencies the guide is referred to as being single mode. In this range the field shape and propagation rate can be uniquely defined within the guide. This greatly simplifies the design of waveguide-based systems.

Standard rectangular waveguide as a single-mode transmission system is widely used at microwave frequencies. However, it requires the guide width and height to be around one-half to one-quarter of the free space wavelength of the radiation. This becomes increasingly difficult to achieve at mm-wave frequencies because it requires us to manufacture rectangular pipes with dimensions of a millimetre or less. Even when we are able to make such small guides we discover that their performance is poor compared to the larger guides used at lower frequencies. In part this is due to problems in accurately making the guide, but we also encounter a more fundamental problem.

As the wave moves down the guide it sets up currents in the surface of the walls. These currents must exist if the walls are to reflect the wave and confine it to the guide. As was mentioned in the introduction, the velocity of an electromagnetic wave is far higher in air than in a normal metal. Only electrons near the surface have time to respond to the field before it is reflected. As the signal frequency rises the depth of the layer in which the electrons respond becomes less. In effect we find that the electrons are confined to a thinner conductor layer. In a normal resistive material this means that the resistance through which the wall current must flow increases with frequency. Similarly, as we try to reduce the guide size and maintain single-mode operation we are reducing the effective width of the conductor available for the wall current, increasing the resistance still further.

The losses in normal microwave guides can be almost unmeasurably small, but at frequencies around 100GHz a standard waveguide may have a loss of around 1db per cm or more. These losses rise rapidly at higher frequencies, making standard waveguide unusable save over very short distances. Optics and laser engineers also need to be able to transmit electromagnetic waves from place to place as efficiently as possible. Instead of using metal waveguides they employ either beams in free space or various forms of dielectric structures - e.g. fibre guides.
Dielectric fibres (and similar structures) can work very well for visible and near-visible frequencies. Unfortunately there are, as yet, no mm-wave dielectrics which have come anywhere near matching the low losses of the best materials for the near-visible region. Free space, by its very nature, is a low-loss transmission medium - and it does not present us with any difficulties of supply or manufacture! For this reason it has proved possible to design mm-wave systems which work very well over a wide frequency range using signal coupling by beams in free space.

Visible and near visible radiation is typified by a wavelength of less than a micron. It is usually not very difficult to make systems of optics for visible light where the sizes of the beams, lenses, etc, are many thousands of wavelengths or more. These systems can, for most purposes, be analysed or designed using the traditional methods of ray-optics. Light can be treated as being transmitted from place to place via plane wave beams.

### 1.2 Gaussian Beams in Free Space

In order to treat mm-wave optical systems in the same way we would have to use beams and optical elements which were perhaps a metre or more across. This would make mm-wave optical systems impractical for all but a few purposes. Clearly we need to be able to design compact mm-wave optical systems - ones where the beam and optical element sizes are no more than a few tens or hundreds of wavelengths.

During the 1960's, laser engineers were coming to grips with designing optical systems which were not many wavelengths across. At about the same time other workers were looking into the possibility of using chains of lenses to couple radiation from place to place. Out of this work grew the subject of Gaussian Beam Mode Optics which has proved to be an outstanding method for analysing compact optical systems. (Optical engineers also use a technique called Gaussian Ray Optics. It is important to note that is not the same as the Beam Mode technique which is being discussed here.)

Gaussian Beam Mode (GBM) theory generally makes the following assumptions:-

- That the radiation is moving in a paraxial beam whose cross-sectional size is not sufficiently large that it may be treated as plane parallel.
- That the radiation can be represented as a scalar field distribution.

By ‘paraxial’ we mean that the signal is moving essentially in a beam along a given axis but that some diffraction is taking place. Having represented the field as a scalar we can then associate the scalar magnitude and phase with, say, a specific electric vector component of the actual electromagnetic wave. These assumptions allow us to simplify Maxwell's equations and obtain a solution appropriate to a beam moving through space. It should be noted, however, that this means the results are only useful in real situations which mimic the assumptions.

We can write the scalar version of the wave equation as

\[ \nabla^2 \psi + k^2 \psi = 0 \]

where \( k = \frac{2\pi f}{c} \).

Here, \( \psi \) represents the scalar field distribution appropriate for the beam, \( c \) is the velocity of light and \( f \) is the signal frequency.
For a beam moving essentially paraxially in the $z$-direction (Cartesian co-ordinates) we can define a function, $\psi$, such that

$$\psi = u(x, y, z) \exp\{-jkz\} \exp\{j2\pi ft\} \quad \ldots (1.8)$$

Since the beam is paraxial along the $z$-direction we can assume that virtually all the $z$-dependance of $\psi$ is contained in the exponential terms which multiply $u()$ above. Hence we can assume that

$$\frac{\partial^2 u}{\partial z^2} \ll \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \quad \ldots (1.9)$$

From the scalar wave equation we can, therefore, write

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - 2jk\frac{\partial u}{\partial z} = 0 \quad \ldots (1.10)$$

The solutions of this differential equation are known. From them we can say

$$\psi_{mn} = \frac{1}{\omega} H_m\left(\frac{\sqrt{2}}{\omega}\right) H_n\left(\frac{\sqrt{2}}{\omega}\right) \exp\{-j(kz - \Phi_{mn} - 2\pi ft) - r^2\left(\frac{1}{\omega^2} + \frac{jk}{2R}\right)\} \quad \ldots (1.11)$$

Where $H_m$ and $H_n$ are Hermite polynomials of degree $m$ and $n$ and

$$\Phi_{mn} = (m + n + 1) \arctan\left(\frac{\lambda z}{\pi\omega_0^2}\right) \quad \ldots (1.12)$$

$$\omega^2 = \omega_0^2\left[1 + \left(\frac{\lambda z}{\pi\omega_0^2}\right)^2\right] \quad \ldots (1.13)$$

$$R = z\left[1 + \left(\frac{\lambda z}{\pi\omega_0^2}\right)^2\right] \quad \ldots (1.14)$$

$$r^2 = x^2 + y^2 \quad \ldots (1.15)$$

$E$ is a complex number which defines the overall amplitude and phase of the beam at the time $t = 0$. The beam power will be proportional to $E^2$. The parameter $\omega_0$ is called the Beam Waist Radius (or size). From the above expressions we can see that this parameter determines how the beam’s width, etc, vary as it propagates in the $z$-direction.

As we found for the case of a metallic waveguide, the solutions represent a set of modes. Each mode propagates keeping its own specific form and the rate of change of phase along the beam depends upon the mode numbers $m$ and $n$. As before, we may expect that a single-mode beam would be preferred to a multi-mode one because this ensures that the field profile at any place along the beam is easy to define and is not strongly frequency dependent.

Most practical systems involve beams with circular or square symmetry - i.e. the beam size is identical in both the $x$- and $y$-directions at any place along the beam. Expressions 1.11 to 1.14 are in the form appropriate for a beam of this type. In some cases the beam may have elliptic or rectangular symmetry and has a beam size, phasefront curvature, etc, which differ in the $x$- and $y$-directions. Under these more general circumstances we may define the beam as being the product

$$\Psi(x, y) = \Psi(x)\Psi(y) \exp\{-j(kz - 2\pi ft)\} \quad \ldots (1.16)$$

Where $\Psi(x)$ and $\Psi(y)$ are linear combinations of one-dimensional mode expressions of the form
J. C. G. Lesurf — 9 — Millimetre-wave Optics, etc

\[ \Psi_m(a) = \sqrt{\frac{1}{\omega_a}} E_a H_m \left( \sqrt{\frac{2}{\omega_a}} \right) \exp \left\{ -j \Phi_m - a^2 \left[ \left( \frac{1}{\omega_a} \right)^2 + \frac{j k}{2 R_a} \right] \right\} \] 

... (1.17)

Where the subscript \( a \) denotes the value appropriate for either the \( x \)- or \( y \)-direction. The beam size and phasefront curvature may be obtained from equations 1.13 and 1.14 as relevant for each direction. For the one-dimensional case

\[ \Phi_m(a) = \left( m + \frac{1}{2} \right) \arctan \left( \frac{\lambda z}{\pi \omega_a^2} \right) \] 

... (1.18)

Although free space modes share many properties with those propagating along a metallic waveguide there are a few important differences.

The \( m = n = 0 \) mode is allowed in free space. Indeed, this mode - usually called the **Fundamental Mode** - is the one which proves to be the best for many purposes. It is fairly easy to produce and detect. It also proved to be the best choice when the effects of truncation by apertures are taken into account.

There is no cut-off wavelength for a free space mode. The propagation constant - which determines the rate of phase change along the beam axis - cannot be zero unless the signal frequency or beam cross-sectional size go to zero. Also unlike the metal waveguide modes, the propagation constant of a free space mode is not constant along the beam. From eqn. 1.11 we may see that changes in phase along the beam axis of a Gaussian mode may be obtained from

\[ \theta = k z - \Phi_m \] 

... (1.19)

where \( \theta \) represents the relative phase at a distance \( z \) along the axis from the location of the beam waist. The term, \( \Phi_m \), is sometimes called the **anomalous phase term** because it indicates a phase difference between the behaviour of a Gaussian Mode and a plane wave.

The zeroth degree Hermite polynomial is equal to unity. Hence the fundamental mode is of the form

\[ \psi_0 = \frac{1}{\omega} E_0 \sqrt{\frac{2}{\pi}} \exp \left\{ -j(k z - \Phi_0 - 2 \pi f t) - r^2 \left[ \frac{1}{\omega^2} \right]^2 + \frac{j k}{2 R} \right\} \] 

... (1.20)

(Note that I have adopted the convention of using a single subscript zero to represent the fundamental mode beam.)

The expressions used above have been simplified by assuming that the beam axis is along the line \( x = y = 0 \). Hence \( \psi \) is the field at a place a distance \( r \) from the beam axis. For the fundamental mode we can see that the field amplitude varies as

\[ \exp \left( \frac{-r^2}{\omega^2} \right) \] 

... (1.21)

i.e. the amplitude variation across the beam has a Gaussian shape. \( \omega \) is a measure of the width of the distribution at any place along the beam. For this reason \( \omega \) is often referred to as the beam size or the beam radius. In the special case of the fundamental mode this the beam radius is also the distance from the beam axis at which the field amplitude has fallen to 1/e of its value on the beam axis.

Figure 1.2 shows how the field of a fundamental Gaussian mode beam varies across the beam. It also illustrates how, \( \omega \), the beam radius varies along the beam. This variation is hyperbolic, hence the radius has a minimum value at some plane along the beam. This plane is called the **beam waist plane** and the value of the beam size here is defined as, \( \omega_0 \), the beam waist radius.
Fig. 1.2 Cross section of Gaussian field pattern and variations of size and curvature along the beam

The beam is clearly not a plane parallel one. Its amplitude distribution is not uniform across the beam and its size varies along the beam. Hence its phasefront cannot, in general, be plane. The phase distribution in a given plane depends upon the parameter, $R$, in the above expressions. The form of this phase variation is that of a spherical phasefront of radius, $R$. Although the width of the field distribution varies along the beam the shape remains Gaussian. Similarly, although the phasefront curvature alters along the beam it remains spherical. For higher order (i.e. when $m \neq 0$ and/or $n \neq 0$) the amplitude distribution for a single mode beam may not be Gaussian but the same rules apply - i.e. the beam amplitude pattern does not alter along the beam. Although the scaling width will vary hyperbolically just as it does for the fundamental mode.

Throughout most of this book it will be assumed that the beams being considered are composed of just a single (usually the fundamental) mode with circular or square symmetry. Most of the arguments and methods discussed can, however, be applied to a multi-mode beam. The advantages of a single mode beam are that the computation is generally easier and the overall performance of a practical system is optimised. The details of multi-mode analysis are considered in Chapter 3.
Chapter 2

Beam coupling, lenses, and mirrors.

2.1 Coupling between Gaussian Beams

In the previous chapter I described the basic expressions which define the way in which a beam of electromagnetic power propagates in free space. In this chapter we see how we can use beams to efficiently couple this power from one component to another of a mm-wave system.

Many mm-wave devices are mounted in a small length of standard rectangular waveguide. A typical system will consist of sources, detectors, etc, each in a short piece of waveguide. The advantage of this method is that it allows us to define the field pattern surrounding the device, making design and analysis easier. In order to avoid the problems mentioned in Chapter 1, however, the signals being coupled between components should be carried by free space beams over as much of the intervening distance as possible.

The device's waveguide must be terminated by an arrangement designed to couple the field propagating along the guide to a free space beam. We require a suitable antenna for the devices. We also must be able to couple the beam produced by a source antenna into that required by another antenna attached to a detector in order to communicated signal power between them.

The subject of antennas will be discussed in Chapter 4. Here we will look at how we can couple power between beams in free space.

Figure 2.1 Field patterns of two beams with the same axis but different waist locations.

Figure 2.1 illustrates the coupling of two beams in space. Clearly, we may expect that any misalignment of the two beams will limit the efficiency of signal power coupling. Here we will therefore make the simplest possible assumption, namely that the beams share a common axis of propagation. The two beams can then differ in the following ways:

a) their beam waist sizes may differ
b) their beam waist planes may not coincide
c) the beam profiles may not be the same (i.e. they may be different modes, or be composed of different linear combinations of modes.)
In order to obtain perfect coupling the two beams should be identical. Then power in one beam is indistinguishable from power in the other and power is transferred with ideal efficiency. In practice, however, the beams may differ in one or more of the respects mentioned above.

Consider the case where, at an arbitrary plane normal to the beam axes, the two field distributions are $\Psi$ and $\Phi$ respectively. Each of these is a Gaussian mode (or combination of modes). For the sake of this example we may take $\Phi$ to represent the beam field pattern produced by a signal power source and $\Psi$ to be the pattern of sensitivity to field of a detector.

At a point $(x,y)$ on our plane, $\Phi(x, y)$ represents the field produced by the source. $\Psi(x, y)$ represents the sensitivity of the detector to field at that point. Both $\Psi$ and $\Phi$ are complex values which define the magnitude and phase distribution of the fields. What fraction of the power contained by the beam, $\Phi$, will be coupled into the detector beam, $\Psi$, and hence into the detector? Since the fields are complex we determine the power coupling from calculating the total field coupling between the beams. This can be done by integrating the incremental amount of coupling for each small area, $(dx,dy)$, over the plane where the fields are known.

The total field coupling will be

$$\int^{+\infty} \int^{-\infty} \Phi \Psi^* \, dx \, dy \quad \text{... (2.1)}$$

It should be noted that the integral is in terms of the product of one field and the complex conjugate of the other.

The total power coupled into the detector beam will therefore be

$$P = \left| \int^{+\infty} \int^{-\infty} \Phi \Psi^* \, dx \, dy \right|^2 \quad \text{... (2.2)}$$

The power contained by the input beam can be defined to be the integral of the modulus squared of its field distribution. i.e. the source power may be defined as

$$S = \int^{+\infty} \int^{-\infty} \Phi \Phi^* \, dx \, dy \quad \text{... (2.3)}$$

The power coupling efficiency, $N$, can then be defined to be the ratio

$$N = \frac{P}{S} \quad \text{... (2.4)}$$

We would associate a power coupling efficiency, $N$, of unity with perfect, loss free, coupling. This means it makes sense to normalise the output beam to arrange that

$$\int^{+\infty} \int^{-\infty} \Psi \Psi^* \, dx \, dy \equiv 1 \quad \text{... (2.5)}$$

which also implies that a ‘perfect’ detector with such a pattern will collect all the power from an input beam, $\Psi$, with an efficiency of unity.

The actual coupling efficiency between source and detector beams in a specific practical system cannot depend upon which $xy$ plane we chose for calculating the coupling integrals. In reality a definite amount of power is coupled and our calculation must yield this result if it is to prove correct. Hence we may expect that the choice of coupling plane should not alter the value of $N$ we obtain. Similarly, our choice of co-ordinate system should not alter the result.

It is convenient when considering the coupling integrals which must be evaluated to obtain a power coupling efficiency to use an expression which avoids being specific about the co-ordinate system. This can be used to make the actual expressions more compact and the
argument clearer. It is also a useful continuing reminder that the choice of co-ordinate system is one which can be made on the basis of computational convenience in a particular case. Here we will, therefore, adopt the notation that, in Cartesian co-ordinates

\[ \langle \Phi | \Psi \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Phi^* \Psi \, dx \, dy \quad \cdots (2.6) \]

The notation, \( \langle \Phi | \Psi \rangle \), can now be used to represent the integral product over the plane irrespective of the co-ordinate system being used to define this coupling integral or calculate its value in a specific case. We may now re-define the power coupling efficiency, \( N \), via the expression

\[ N \equiv \frac{|\langle \Phi | \Psi \rangle|^2}{\langle \Phi \Phi \rangle \langle \Psi | \Psi \rangle} \quad \cdots (2.7) \]

Note that expression 2.7 also takes into account the possibility that the field distribution, \( \Psi \), has not already been normalised in accord with expression 2.5.

Having obtained a suitable definition for power coupling efficiency we may use it to determine the fraction of power coupled between beams in specific cases.

As an example, let us take the simplest case: that of the coupling between two fundamental Gaussian modes. In Cartesian co-ordinates we may write

\[ \Phi = \frac{1}{\omega} \exp \left\{ -r^2 \left[ \left( \frac{1}{\omega} \right)^2 + \frac{jk}{2R} \right] \right\} \quad \cdots (2.8) \]

\[ \Psi = \frac{1}{\omega'} \exp \left\{ -r^2 \left[ \left( \frac{1}{\omega'} \right)^2 + \frac{jk}{2R'} \right] \right\} \quad \cdots (2.9) \]

and because we are concerned here only with the power coupling efficiency we can omit those terms which only affect the phase of the coupling integral. In these expressions, for our chosen coupling plane, we have used \( \omega \), and \( R \) to represent the beam beam size, and phasefront curvature of one beam, and \( \omega' \), and \( R' \) of the other. We may now say that

\[ \langle \Phi | \Psi \rangle = AB \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp \left\{ -r^2 \left[ \left( \frac{1}{\omega} \right)^2 + \left( \frac{1}{\omega'} \right)^2 + \frac{jk}{2} \left( \frac{1}{R} - \frac{1}{R'} \right) \right] \right\} \, dx \, dy \quad \cdots (2.10) \]

\[ \langle \Phi | \Phi \rangle = A \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp \left\{ -r^2 \left( \frac{1}{\omega} \right)^2 \right\} \, dx \, dy \quad \cdots (2.11) \]

\[ \langle \Psi | \Psi \rangle = B \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp \left\{ -r^2 \left( \frac{1}{\omega'} \right)^2 \right\} \, dx \, dy \quad \cdots (2.12) \]

where

\[ A \equiv \frac{1}{\omega} \quad \cdots (2.13) \]

\[ B \equiv \frac{1}{\omega'} \quad \cdots (2.14) \]

By consulting a suitable table of standard integrals we find that

\[ \int_{-\infty}^{\infty} \exp (-at^2) \, dt = \sqrt{\pi \over a} \quad \cdots (2.15) \]

from which we can obtain the result

\[ N = \frac{4}{\left( \omega / \omega' + \omega' / \omega \right)^2 + \left( \frac{k\omega'}{2} \right)^2 \left( 1 / R' - 1 / R \right)^2} \quad \cdots (2.16) \]

As we would expect, expression 2.16 confirms that \( N \) will only be unity when \( \omega = \omega' \) and \( R = R' \) at the same plane - i.e. when the two beams are identical. The method employed to
obtain this result may be applied for beams composed of other single- or multi-mode beams (see, for example, Chapter 3).

In most practical cases we are presented with two beams which are essentially fundamental mode gaussians but which may differ in beam waist size and location. Provided the beams have a common axis there will always be at least one plane where their sizes are equal. It is at this plane we may choose to place a lens or mirror in order to optimise the beam coupling.

### 2.2 Lenses and Beam Coupling

A lens may be regarded as a device which alters the radius of phasefront curvature of a beam passing through it. A suitable lens for our purposes is one which changes the radius of curvature of an incident beam so that it matched exactly that of another beam into which we wish to couple power. The modified input beam and the output beam now have the same size and curvatures and - barring other problems - the coupling efficiency would be unity.

For a ‘thin’ lens the focal length, $f$, may be defined simply in terms of the required change in phasefront curvature

$$\frac{1}{f} = \frac{1}{R} - \frac{1}{R'}$$

... (2.17)

(It is assumed that the sign of $R$ determines which side of the lens it is centered.)

In practice, all lenses possess various defects which may affect the performance of the system in which they are placed. These may be summarised as:-

i) distortions and aberrations

ii) absorption losses

iii) surface reflections.

The expression 2.16 gives the coupling efficiency between two fundamental mode beams which may have differing sizes and curvatures. If the lens which seeks to equalise the curvatures alters the field pattern so that it is no longer simply Gaussian then 2.16 ceases to be the appropriate expression and the coupling efficiency will be reduced.

In conventional ray optics we could draw a fan of rays from a focal point to the surface of a lens. The directions of the locally refracted rays may then be obtained from Snell's Law and the focussing action - and distortions - of the lens obtained. By applying this method appropriate lens profiles can then be calculated.

For a Gaussian mode we may assume that power propagates locally in the direction normal to the beam phasefront. As the phasefront is everywhere spherical this means that at any plane normal to the beam axis the power appears to diverge/converge as if from a point source. However the apparent location of this point varies as we move along the beam. This would not be the case for a ‘ray optics’ beam and means that a lens designed using simple ray optics may not work as expected when the beam is compact.

From expression 1.14 we may see that, when $z \gg \pi \omega_0^2 / \lambda$, then $R \approx z$ and the phasefront appears essentially fixed at the center of the beam waist plane. This condition is equivalent to saying that the beam size is much larger than the wavelength and the normal assumptions of ray optics can be made. In general, however, we must take the Gaussian nature of the beam into consideration when designing lenses if we wish to avoid unexpected distortions and coupling losses.
Fig. 2.2  Focussing effect of a convex lens surface

Figure 2.2 illustrates a Gaussian mode incident upon a lens surface. On the beam axis the change in phase between the beam waist plane and \( z_p \) is

\[
\alpha = -k z_0 + \text{Arctan} \left( \frac{\lambda z_0}{\pi \omega_0^2} \right) - k \mu (z_p - z_0) \quad \ldots (2.18)
\]

and at the point where \( z_p \) intersects the lens surface the phase change compared with the waist plane is

\[
\beta = -k z_p + \text{Arctan} \left( \frac{\lambda z_p}{\pi \omega_0^2} \right) - \frac{k r^2}{2 R} \quad \ldots (2.19)
\]

It is assumed that the lens is in air and that \( \mu \) is the refractive index of the lens material. It has also been assumed that the beam is a fundamental mode one. One of the useful properties of Gaussian modes is that \( R \) and \( \omega \) do not depend upon the mode numbers. However, \( \Phi \), the anomalous phase term does depend upon the mode numbers and this may need to be taken into account in some cases.

If we arrange that these phase changes shall be equal for every value of the distance, \( r \), from the beam axis then we have essentially produced a plane wave within the lens. From the above expressions it can be seen that the requirement for us to obtain this result is that

\[
r^2 = \omega_0^2 \left( \hat{z}_p + \frac{1}{\hat{z}_p} \right) \left( \Gamma + \text{Arctan} \left( \hat{z}_p \right) - \text{Arctan} \left( \hat{z}_0 \right) \right) \quad \ldots (2.20)
\]

where

\[
\hat{z}_p = \frac{2 z_p}{k \omega_0^2} \quad ; \quad \hat{z}_0 = \frac{2 z_0}{k \omega_0^2} \quad \ldots (2.21 ; 2.22)
\]

\[
\Gamma = \frac{1}{2} \left( \mu - 1 \right) k^2 \omega_0^2 \left( \hat{z}_p - \hat{z}_0 \right) \quad \ldots (2.23)
\]

These expressions link \( r \) with \( z_p - z_0 \) and allow us to calculate a lens surface which produces a nominally plane beam inside the lens. A real lens, curved on both faces, can be treated as a 'back-to-back' combination of a pair of such lenses. Each produces the desired free space beam and the two lens surface are coupled via an essentially parallel beam within the lens.

A lens of this type will work well provided it is 'thin' - i.e. provided that its axial thickness is
small compared with its effective focal length. When this is not the case, problems will arise which distort the beam and reduce the reliability of the calculation.

i) Some of the power incident upon a dielectric interface will be reflected. The amount reflected depends upon the angle of incidence. On-axis the power arrives at normal incidence. As we move away from the axis the angle of incidence steadily changes, increasing the reflective loss.

ii) It is inevitable that the lens material will absorb some of the power passing through it. For example, in a convex lens this means a higher absorption loss near the axis than at the edges.

iii) Power falling on the surface at an angle appears 'foreshortened'. The curved geometry of the refractive surface will alter the field distribution of the beam passing through it.

iv) We may no longer regard the beam within the lens as being essentially parallel. Instead we must treat it as a Gaussian beam and take diffraction effects into account. This problem is, however, rarely significant unless the lens is exceptionally thick and has a small diameter compared with the radiation wavelength.

2.3 Dielectric materials and surface reflections

Provided that the lens thickness/focal length and diameter/focal length ratios are less than about 0.2 the geometric losses are less than those due to absorption or reflection for most mm-wave lenses. Reflection and absorption losses depend upon the lens material and the signal frequency. Table 2.1 summarises the dielectric properties of some of the better mm-wave dielectric materials.

<table>
<thead>
<tr>
<th>Material</th>
<th>Refractive Index</th>
<th>Loss Tangent (x 10^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Density Polyethylene (HDPE)</td>
<td>1.53</td>
<td>0.3</td>
</tr>
<tr>
<td>Low Density Polyethylene (LDPE)</td>
<td>1.51</td>
<td>0.3</td>
</tr>
<tr>
<td>Polytetrafluorethylene (PTFE)</td>
<td>1.43</td>
<td>0.8</td>
</tr>
<tr>
<td>Quartz</td>
<td>2.11 / 2.15</td>
<td>0.3 / 0.8</td>
</tr>
<tr>
<td>Poly 4 Methyl Pentane-1 (TPX)</td>
<td>1.44</td>
<td>4.0</td>
</tr>
<tr>
<td>Sapphire</td>
<td>3.06 / 3/42</td>
<td>1.7 / 2.9</td>
</tr>
</tbody>
</table>

It is convenient to quote the loss tangent of the materials since, in general they have a loss per centimetre which is proportional to the signal frequency over the mm-wave region. At a frequency, \( f \), in GHz the loss of a material, \( \alpha \), in dB/cm may be obtained from the loss tangent, \( \delta \), via the expression

\[ \alpha = 0.91 \mu \delta f \]

where \( \mu \) is the refractive index of the material.

The values quoted for Quartz and Sapphire in Table 2.1 were obtained from crystalline samples. Two values are given because these crystals are birefringent. This makes them unsuitable for lenses in some cases as the lens properties will be polarisation sensitive. The materials may be useful, however, in cases where a particularly high refractive index is required.

HDPE (High-Density Polyethylene) and PTFE have about the lowest losses known for mm-wave dielectrics. Although they cannot be polished or ground like Quartz or Sapphire they can be turned quite easily on a lathe. For mass production purposes they could also be pre-cast to the required shape. Both materials are reasonably strong and inert.
TPX is another polymer material which has been used on a number of occasions to manufacture mm-wave lenses. Although significantly more lossy than either HDPE or PTFE it does possess one useful property. TPX is reasonably transparent to visible light and has a refractive index in the visible similar to its mm-wave index. For this reason systems including TPX lenses may be aligned by eye or using visible lasers as 'ray' sources. HDPE and PTFE are opaque white materials and their use prevents alignment in this way.

LDPE (Low-Density Polyethylene) is virtually identical in its mm-wave properties to HDPE. It is, however, a relatively soft material, difficult to machine.

There is evidence that the mm-wave dielectric properties of most commercial polymers are somewhat variable. The manufacturers are usually more concerned to produce a material suitable for manufacturing milk crates than mm-wave lenses! Most commercial polymers will include small amounts of other materials intended to improve, for example, the way in which the material flows when being pressure moulded. The polymerization process followed at different times by various manufacturers are also varied. This can produce 5% to 10% variations in refractive index - and may cause the absorbivity to rise in some cases by an order of magnitude. In most cases the dielectric constants are close to the values quoted in Table 2.1, but it is a good idea to select lens materials with caution for demanding applications.

The reflectivity of a dielectric interface depends upon the refractive indices and the angle of incidence. A wave moving through a medium of refractive index \( \mu_1 \) and striking another medium of refractive index \( \mu_2 \) at normal incidence will produce a reflected field whose magnitude is

\[
\frac{\mu_2 - \mu_1}{\mu_2 + \mu_1}
\]

... (2.24)

times the incident field’s magnitude.

Taking the example of an air-HDPE interface this means that we may expect an 0.23dB loss in the power transmitted due to reflection. This compares with an absorption loss in the material of around 0.1dB per cm at 250GHz. For a dielectric lens in air a reflection loss will occur at both of the lens surfaces. The reflected waves will interfere and the total reflection loss then depends upon the lens size and shape compared with the radiation wavelength.

The effects of lens reflections will be considered in more detail in a later chapter. Here we may simply look at the general behaviour by regarding a thin lens as being essentially a plane parallel sided slab of dielectric that has negligible internal losses. The power reflection coefficient, \( |\Gamma|^2 \), of such a slab will be

\[
|\Gamma|^2 = \frac{(r_1 + r_2)^2 - 4r_1r_2 \sin \{x\}}{(1 + r_1r_2)^2 - 4r_1r_2 \sin^2 \{x\}}
\]

... (2.25)

where \( r_1 \) and \( r_2 \) are the field reflectivities of the two surfaces, and

\[
x = \frac{2\pi d}{\lambda} \sqrt{\mu^2 - \sin^2 \{\theta\}}
\]

... (2.26)

where \( d \) is the slab thickness, \( \mu \) its refractive index, \( \lambda \) the radiation wavelength in free space, and \( \theta \) the angle of incidence of the beam moving inside the slab. For simplicity the beam has been assumed to behave as if plane parallel and the slab absorption loss is neglected. For such a slab

\[
r_1 = -r_2
\]

... (2.27)

At normal incidence the reflectivity varies with \( d / \lambda \) between a maximum value
\[ |\Gamma|^2 \text{ (max)} = \frac{4 r^2}{(1 + r^2)^2} \] ... (2.28)

where \( r \) can be taken as being equal to either \( r_1 \) or \( r_2 \), and a minimum value of zero.

For HDPE, \( r = 0.21 \), and \( |\Gamma|^2 \text{ (max)} = 0.16 \), leading to a maximum power loss by reflection of around 0.75db. From this example it may be seen that for the best lens materials reflection losses may be significantly higher than those produced by absorption. They are also strongly frequency dependent because of the interference effects between the waves reflected from different surfaces.

Various methods have been evolved to reduce the reflection losses of mm-wave lenses. The simplest being to carefully choose a lens thickness which causes the two surface reflections to cancel as closely as possible. Whilst this method works fairly well, it suffers from two drawbacks. Firstly, the surfaces of a lens are not plane parallel. The reflected waves cannot be made to exactly overlay and cancel perfectly. This means that it may not be possible to obtain a true zero total reflection, although the reflections may be significantly reduced.

Secondly, a typical lens will have a thickness with is somewhat longer than a wavelength. Adjacent maximum and minimum values of \( |\Gamma|^2 \) are separated by a change of just 0.25 in the ratio \( d\mu/\lambda \). Consider the example of an HDPE lens which is approximately 10 mm thick. This thickness corresponds to around 15 wavelengths at 300 GHz. If the lens is designed to have a minimum reflectivity at 300 GHz then the its power reflectivity will have maxima at about 295 and 305 GHz. Hence trying to minimise reflection losses by choosing an optimum lens thickness isn’t likely to be satisfactory for wideband use.

![Fig. 2.3 Modifications of a dielectric surface to reduce reflection losses.](image)

Alternative methods are based upon blazing or blooming the lens surfaces. This involves either creating a suitable layer on each lens surface or patterning the surface so as to reduce the reflection losses. If we place an intermediate dielectric layer, of refractive index \( \mu' \), in between the air and the lens dielectric we create two reflective interfaces at each side of the lens. The amplitude reflectivities at these two interfaces will now be

\[ r_1 = \frac{\mu' - 1}{\mu' + 1} \quad \text{and} \quad r_2 = \frac{\mu - \mu'}{\mu + \mu'} \] ... (2.29 and 2.30)

From expression 2.25 it may be seen that, for normal incidence, the total power reflection may
be zero if
\[ \mu' = \sqrt{\mu} \] ... (2.31)
and the layer thickness, \( t \), is such that
\[ t = \frac{(2n + 1)\lambda}{4\mu'} \] ... (2.32)
Where \( n \) is a non-negative integer.

In this way we may cancel the reflections of each lens surface, independently. If the layers are of uniform thickness the two reflected fields at each lens surface will have essentially identical curvatures, permitting the fields to cancel almost exactly. Furthermore, by choosing \( n = 0 \) we have selected a layer thickness of just one quarter wavelength. This means that the reflection reduction extends over a much wider frequency range than would be the case if we try to employ cancellation effects from surfaces many wavelengths apart. Hence this method is preferable when a lens is to be used over a wide frequency range.

In order to produce successful anti-reflection coatings for HDPE and PTFE lenses we require low-loss materials with a refractive index around \( \sqrt{1.5} \) which can be attached to these materials in uniform layers. In practice, suitable materials for this method haven’t been easily obtained! Fortunately, an alternative approach is to cut a series of grooves into each lens surface. Figure 2.3 illustrates a series of narrow rectangular grooves cut into a dielectric surface. The action of this surface contour can be viewed as synthesising a layer of intermediate refractive index by removing a fraction of the dielectric material. In order to prevent interference from generating scattered waves being reflected or transmitted at an angle to the normal beams the distance between successive grooves should be small compared with the wavelength. Concentric grooves (or even a spiral groove) of this type can be cut on HDPE or PTFE using standard lathe machining techniques.

A problem which arises with grooved material is that the effective refractive index does, in fact, depend upon the orientation of the incident wave’s electric vector with respect to the groove direction. This may mean that a pattern of groove will alter the polarisation pattern of a beam. Typical blazed layers on lenses are quite thin and hence this effect is not normally significant, but it should be considered where a system is particularly sensitive to unwanted polarisation effects. Modified arrangements have been employed in a number of systems. For example, the polarisation sensitivity of grooves may be avoided if the surface contour is produced by drilling an array of holes into the dielectric surface.

Grooves of triangular cross-section have also been used. These are particularly useful if they can be cut as a close pattern of deep grooves. The arrangement then behaves as a smoothly graded change in refractive index as the wave moves into the lens material. By cutting grooves (or triangular holes) a few wavelengths deep we avoid any abrupt changes in refractive index and the surface reflectivity may be reduced almost to zero over a very wide frequency range. The distances between adjacent grooves or holes must, however, remain small compared with the wavelength. Unfortunately, it is difficult to manufacture such long narrow grooves or holes in most mm-wave dielectric lens materials.

Two beams may be coupled by placing a lens at the plane where the beam sizes are equal. The action of the lens being to equalise the beam phasefront curvatures, hence maximising the power coupling. In many cases we are concerned with using a series of lenses (or mirrors) to guide a beam. We must therefore be able to specify how power may be coupled efficiently between focussing elements.
Consider the example of a pair of lenses a distance, \( d \), apart. The size of the beam emerging from one lens is \( \omega \). By altering the focal length of the lens we may alter, \( R \), the phasefront radius of the emerging beam, but we cannot alter the beam size at the lens in this way. The beam formed by the lens will have a waist size given by

\[
\omega_0^2 = \frac{\omega^2}{1 + \frac{\omega^2}{2\lambda}}
\]  
... (2.33)

located at a distance

\[
z = \frac{R}{1 + (1/\hat{z})^2}
\]  
... (2.34)

from the lens, where we have defined

\[
\hat{z} \equiv \frac{\pi\omega^2}{\lambda R}
\]  
... (2.35)

At a following lens the beam will then have a size, \( \omega' \), and radius of curvature, \( R' \), given by the expressions

\[
\omega'^2 = \omega_0^2 \left[ 1 + \left( \frac{\lambda(d - z)}{\pi\omega_0^2} \right)^2 \right] 
\]  
... (2.36)

\[
R' = (d - z) \left[ 1 + \left( \frac{\pi\omega_0^2}{\lambda(d - z)} \right)^2 \right] 
\]  
... (2.37)

These expressions may now be used to select a value for \( R \), the radius of the phasefront leaving the first lens, which produces the required beamsize, \( \omega' \), when it arrives at the second lens.

From expressions 2.34 and 2.35 we can see how the beam waist location varies with \( R \). When \( R \gg \pi\omega^2/\lambda \) then \( z \approx R/R^2 \) and as \( R \) increases \( z \) falls to zero. When \( R \ll \pi\omega^2/\lambda \) then \( z \approx R \) and as \( R \) decreases \( z \) falls to zero. Re-arranging 2.34 and 2.35 we obtain

\[
R^2a z - R + z = 0 
\]  
... (2.38)

where

\[
a = \left( \frac{\lambda}{\pi\omega^2} \right)^2 
\]  
... (2.39)

2.38 is a quadratic equation whose roots are

\[
R = \frac{1 \pm \sqrt{1 - 4az^2}}{2az}
\]  
... (2.40)

Two conclusions may be drawn from this result. Firstly that the distance to the beam waist, \( z \), must always be such that

\[
z \ll \frac{\pi\omega^2}{2\lambda}
\]  
... (2.41)

in order to ensure that \( R \) is a real number. This means that we cannot produce a beam waist at an arbitrary distance from a lens (or mirror). The distance obtained by setting 2.41 to an equality is often called the maximum throw of a lens.

The second conclusion is that, for any distance to the waist which is less than the maximum throw we have a choice of two solutions for \( R \). These produce different beam waist sizes at the chosen beam waist plane. These are sometimes distinguished by calling the smaller a focused beam waist and the larger a parallel beam waist.
2.4 Mirrors and beam coupling

Metallic mirrors have a number of advantages over lenses. The reflectivity of most metals is essentially unity for mm-wave radiation. Mirrors are not subject to the problems of frequency dependent performance created in lenses by unwanted surface reflections. It is also possible to manufacture very large mirror systems to act as telescope or antenna systems.

The chief disadvantage of the use of mirrors as beam coupling elements stems from their the reflection property itself. If used near normal incidence the input and output beams largely occupy the same volume of space. This means that there is a tendency for the source or detector - whichever is the closer - to block a part of the beam coupling the other to the mirror.

Some mirror arrangements, for example the Cassegrain system, accept the limitations this may impose in order to keep the convenience of an axially symmetric optical system. When the mirrors are relatively compact, however, the diffraction problems created by any significant amount of centre-blocking are usually better avoided by employing mirrors 'off-axis'.

Mirrors of this type are often considered as if the system was employing just a small part of a much larger reflector well away from its axis of symmetry - hence the term 'off-axis'. The performance of axially symmetric systems such as Cassegrain will be considered in the chapter on antennas. Here we will consider the behaviour of off-axis mirrors used as beam coupling elements.

![Image](image-url)

Fig 2.4 Focussing behaviour of an off-axis concave mirror.

Figure 2.4 illustrates two beams being coupled via a curved reflecting surface. For simplicity we will examine the case where the input and output beams are at 90° to each other. The mirror surface intersects the beam axies at a distance, z, from one beam waist, and z’, from the other.

As with the case of a lens we can define the required mirror surface shape by considering the phase variations we require along the beam. Measured along the beam axis, the phase difference between the beam waist planes will be

$$\delta(\text{axis}) = k \left( z + z' \right) + \Phi'(z') + \Phi(z) \quad \ldots (2.42)$$

and along the path which touches the mirror surface above the axial plane of reflection at the point \((\alpha, \beta)\) the phase shift in moving between waist planes will be...
\[ \delta (\alpha, \beta) = k (z + \beta + z' - \alpha) + \Phi'(z' - \alpha) + \Phi (z + \beta) + \frac{jka^2}{2R} + \frac{jk\beta^2}{2R'} \quad \ldots (2.43) \]

The correct mirror surface for coupling the beams whose waists are as shown in Figure 2.4 may be obtained by requiring that \( \delta(\text{axis}) = \delta (\alpha, \beta) \) for all points on the surface. We may then use the above expressions to determine the mirror profile. In many cases \( \alpha \) and \( \beta \) are sufficiently small for us to be able to make use of the approximations

\[ \Phi (z + \beta) - \Phi (z) \approx \frac{\partial \Phi}{\partial z} \beta = \frac{\lambda \beta}{\pi \omega_0^2} (1 + \hat{z}^2)^{-1} \quad \ldots (2.44) \]

\[ \Phi'(z' - \alpha) - \Phi'(z') \approx \frac{\partial \Phi'}{\partial z'} (-\alpha) \approx -\frac{\lambda \alpha}{\pi \omega_0'^2} (1 + \hat{z}^2)^{-1} \quad \ldots (2.45) \]

In practice, however, mirrors of this general type tend to produce distortions in the amplitude profile of reflected beams due to the off-axis geometry. The shapes required as a solution produced by the method described can also be difficult to manufacture. Hence in practice it tends to be advisable to ensure the mirror is far enough from the beam waists for the \( \partial \Phi / \partial z \) values to approach zero whenever this is possible.

The above enables us to simplify the above expressions which then tend towards the form recognisable from conventional ray optics. The mirror surface then becomes part of a standard parabolic/elliptic/hyperbolic curve family. However, whilst this helps with design and performance it does so at the expense of accepting a less compact mirror because we have essentially retreated back into the province of ray optics. Failing this we must evaluate the distortions produced in a specific case and decide if they are acceptable for a given purpose.

The polarisation of a beam is also affected by being reflected from a curved off-axis surface. As with the amplitude distortion, this polarisation distortion arises from the geometry of the arrangement. Unfortunately, this effect does not necessarily reduce if we attempt to retreat to the ray optics domain. The amount of cross-polar radiation generated by such a reflector can be a serious problem in a number of applications. It is, for example, one of the reasons why axially symmetric reflecting antennas are preferred for many communications or radar antennas.

![Fig. 2.5 Four alternative arrangements for focussing mirror pairs coupled by either a parallel or focussed beam waist.](image)
In some optical systems it is possible to employ off-axis mirrors in pairs, arranged such that each mirror counteracts the distortions generated by the other. Four possible arrangements are illustrated in Figure 2.5. They can be distinguished using two criteria. Firstly, whether or not the emergent beam is moving in the same direction as the input. Secondly, the choice of a focussed or parallel beam waist between the mirrors. Fig 2.5a and 2.5c show a parallel beam waist between the mirrors, 2.5b and 2.5d show a focussed beam waist. The beam emerges in the same direction it entered in 2.5c and 2.5d but not in 2.5a and 2.5b.

When we wish the output beam to move in the same direction as the input the beam distortions can be minimised by producing a focussed beam waist (as shown in 2.5d) between the mirrors. When the output is moving in the opposite direction to the input a parallel beam waist (as 2.5a) is to be preferred. These two results may be combined when repeated mirror pairs are used to carry a beam over a long distance. In such a case the waist between each mirror of a pair should be focussed and the waist of the beam projected between pairs should be parallel.

Under some conditions it is possible to employ an arrangement of mirrors which are used on-axis, without any centre blockage. This allows us to avoid the distortions and losses which usually arise with off-axis reflectors. Figure 2.6 illustrates one type of focussing system which use axially symmetric mirrors.

![Fig 2.6 Axially symmetric mirror focussing without blockage.](image)

This arrangement uses two wire-grid polarisers and two curved mirrors and is a very simple optical circuit of the sort discussed in more detail in later chapters. In order for this system to operate the input beam must be plane polarised. (Similar systems may, however, be employed for arbitrary beam polarisation states.)

For the system illustrated, the input beam is plane polarised with its electric field parallel to the horizontal plane. Hence it is initially transmitted through the 'horizontal' polariser. The beam is then incident upon another polariser which is positioned so as to act as a 50:50 beamsplitter. The input beam is hence divided into two orthogonally polarised components of equal amplitude. These components are reflected by two axially symmetric focussing mirrors and returned to the polariser which recombines them.

If the path lengths to the two mirrors are identical the system acts just as if the polarisers were absent and just one mirror were used - i.e., the focussed beam is directed back towards the input. If we arrange for the path lengths to differ, however, then the polarisation state of the reflected beam will depend upon the path difference and the signal wavelength. The system may be considered as a simple form of polarising two-beam interferometer, similar to the instruments
considered in later chapters.

By choosing a path difference which is \( m + \frac{1}{2} \) wavelengths, where \( m \) is an integer, we can arrange that the recombined reflected beam will be plane polarised orthogonally to the input. In these circumstances the output beam is reflected by the horizontal polariser and is not returned to the input. Hence for appropriate signal frequencies the system will focus a beam passing between the input and output.

A system of this form can provide a very low level of beam distortion and loss and this makes it valuable in a number of applications. It has one significant disadvantage: the system is ‘tuned’ by the choice of a specific path difference. For signal wavelengths other than those which satisfy the above requirement some power will be directed back towards the input. This particular system is hence unsuitable for signals which cover a wide frequency range.
Chapter 3
Multi-mode beams and apertures.

A common problem in the propagation of electromagnetic waves is illustrated in Fig 3.1. Here we know the field pattern, $\Psi$, on a given plane and wish to calculate the field, $\Psi'$, this produces at some other plane. In conventional ray optics a problem of this type will typically be dealt with using methods such as Fraunhofer or Fresnel diffraction theory, or by applying a technique known as the Geometric Theory of Diffraction. These techniques are valuable, but they do suffer from some practical difficulties.

![Fig. 3.1 Effect of diffraction upon a field pattern moving from one plane to another where $z$ is the nominal direction of propagation for a beam](image)

Consider the situation shown in Fig 3.1. We may define the field at a position, $x$, $y$, on the first plane to be $\Psi(x, y)$. An incremental area, $dA$, centred on $(x, y)$ will couple a given amount of field onto an incremental area, $dA'$, centred at $x'$, $y'$ on the second plane such that

$$\delta\Psi'(x', y') \, dA' = f[\Psi(x, y)] \ldots (3.1)$$

where the form of the function, $f[]$, may be derived from the properties of the diffraction theory chosen. The total field produced at $x'$, $y'$ can then be obtained by integrating over all the contributions of the various incremental areas in the first plane. Hence we may say that the total field, $\Psi'(x, y)$, may be given by

$$\Psi'(x', y') = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f[\Psi(x, y)] \, dx \, dy \ldots (3.2)$$

A double integral is therefore necessary in order to evaluate the field at each location on the second plane.

In many cases of practical interest it is found that the required integral does not have a known analytic solution. Then we must resort to a numerical summation over the a set of finite incremental areas to calculate $\Psi'$ at each particular place. This means that we have to carry out a four-dimensional summation to calculate the field pattern on the second plane. With the advent of powerful computers, numerical computations of this type have become increasing common,
being particularly widespread in the areas of Antenna Analysis and Design.

When using a method of replacing an integral with summation over a set of finite elements it is necessary to pay some attention to the need to establish the accuracy of the result. Clearly, if our elemental areas are too large and too few the answers obtained may be unacceptably inaccurate. On the other hand, choosing to employ too many incremental areas produces an unnecessary increase in computational difficulties. Some time must, therefore, be devoted to assessing the optimum choices for obtaining results which are sufficiently accurate. Additional effort may also have to be devoted to checking that the level of accuracy obtained is that which was expected and required. These peripheral requirements can often represent a significant portion of the total effort required to obtain a result.

The numerical approach may also prove unhelpful when the result calculated is not quite what we were looking for. In some cases we wish to obtain an output field, $\Psi'$, which differs in some ways from the result obtained. The question which then arises is, how should we alter $\Psi$ to obtain the required result?

When dealing with the propagation of a beam we may wish to know the field patterns produced at a number of planes along the beam direction. We may also wish to find a plane where the field has a particular pattern. This would require us to repeatedly evaluate a four-dimensional summation, searching for the plane we require. The method may therefore involve us with a considerable amount of computation in order to discover the answer to a fairly simple question.

Where we are interested in power propagating as a paraxial beam we may regard the field as being produced by a specific combination of Gaussian beam modes. The corresponding fields at any other points along the beam may then be calculated from the properties of Gaussian modes propagating in free space. This technique possesses two strong advantages over the conventional numerical techniques. Firstly, it allows us to obtain an algebraic expression which defines the field’s magnitude and phase at any point on any plane along the beam. Once the initial calculation has been performed we do not need to carry out any further integrals or summations to evaluate the field at some new plane. Secondly, the propagation behaviour of the beam can be described in a way simple enough to allow some physical insight. For example, the manner in which $\omega$ and $R$ vary along the beam allow us to rapidly identify the optimum position for a coupling lens or mirror. It also becomes possible in many cases to examine a result and swiftly deduce the effects of various changes.

Consider again the field pattern, $\Psi$, shown in Fig 3.1. We may regard this as being the field produced in a particular plane by a multi-mode Gaussian beam propagating along an axis perpendicular to the plane. The field distribution at this plane is then simply the result of superimposing various amount of different Gaussian modes. i.e. we may write that

$$\Psi = \sum_m \sum_n A_{mn} \psi_{mn} \quad \ldots (3.3)$$

Where $\psi_{mn}$ is the $mn$'th Gaussian mode, normalised such that

$$\langle \psi_{mn} | \psi_{mn} \rangle = 1 \quad \ldots (3.4)$$

and the $A_{mn}$ values are a set of a complex numbers which determines the amplitude and phase of each mode’s contribution. The modes are all chosen such that, by definition, they all have the same value of beam size, $\omega$, and phasefront radius of curvature, $R$, at the plane where we have defined the field $\Psi$. This being the case it can be shown to follow that the functions, $\psi$, are mutually orthogonal. i.e. we may say that
Now

\[ \langle \Psi | \psi_{mn} \rangle = \sum_p \sum_q \langle A \psi_{pq} | \psi_{mn} \rangle \]  \hspace{1cm} \text{(3.6)}

so if we combine 3.4 - 3.6 we can obtain

\[ A_{mn} = \langle \Psi | \psi_{mn} \rangle \]  \hspace{1cm} \text{(3.7)}

which provides us with an expression for calculating the $A_{mn}$ values once $\omega$ and $R$ are chosen. From $\omega$, $R$, and the signal frequency we may then also calculate the distance to the beam waist, $z_0$, and the beam waist radius, $\omega_0$.

Each mode propagates maintaining its own characteristic field pattern. The variation of $\omega$ and $R$ as we move along the beam do not depend upon the mode numbers and all the modes have been chosen to share common values of $\omega$ and $R$ at the plane where $\Psi$ is known. Hence all the modes share identical $\omega$ and $R$ values at any plane along the beam.

For the $mn$'th mode, the phase change in moving along the beam axis from a plane, $(\omega, R)$, to the beam waist plane $z_0$ away will be

\[ \delta(m, n) = k z_0 - \Phi_{mn}(z_0) \]  \hspace{1cm} \text{(3.8)}

where $\Phi_{mn}(z_0)$ is the anomalous phase term for the $mn$'th mode evaluated at the plane a distance $z_0$ from the beam waist. We can now write an expression of the form

\[ \Psi(x, y, z) = \sum_m \sum_n A_{mn} \psi_{mn} \text{Exp}\{-j\delta\} \]  \hspace{1cm} \text{(3.9)}

where $\psi_{mn}$ is a Gaussian mode of the form defined by expressions 1.11 to 1.15 in chapter 1. The result of the above process is that we can now write an algebraic form for the field at all locations along the beam.

At first glance it may seem as if the above result implies that the field pattern will change only in size and phasefront curvature as we move along the beam - after all, each mode propagates maintaining its own pattern. $R$ and $\omega$ vary along the beam, but these variations are the same for all modes. Such a result would be contrary to our expectation that, save for a few special cases, diffraction will alter the field profile of a beam as it propagates.

Although the field profile of a single-mode beam remains constant along the beam this is not generally the case for a multi-mode beam. This is because of the presence of the anomalous phase terms, $\Phi_{mn}$, which do depend upon the mode numbers. The total field at any point is obtained by adding together the various mode contributions taking their relative phases into account. The phase relationship will alter along the beam, altering the resultant total field pattern.
Fig 3.2 Column (a) shows the cross-sectional field patterns of the first four Gaussian-Hermite modes. Column (b) shows the field of a multimode beam composed of equal amounts of these four modes, illustrated at different planes along the beam. The modes are all in phase at the waist plane, but differ in phase at other planes along the beam.

Figure 3.2 illustrates the alteration in field profile along the beam for a beam composed of a few modes of comparable power levels. For these graphs $\hat{z}$ is given its standard definition

$$\hat{z} \equiv \frac{\lambda z}{\pi \omega_0^2}$$

... (3.10)

The graphs plotted in the left hand column of Figure 3.2 show the field patterns of four Gaussian-Hermite modes. The graphs in the right-hand column show the field pattern of a multi-mode beam at four planes. The continuous line shows the variation across the beam of the modulus of the field. The dotted lines in 3.2 show how the phase distribution in a plane may depart from following a spherical phasefront. It can be seen that the phase distribution need not be simply spherical. This is also a consequence of the changing mode phase relationship as we move along the beam.

It may also be seen that the ‘far field’ ($\hat{z} = \infty$) pattern of the multimode beam has a uniform, spherical, phase distribution given a set of modes that share the same phase at their common beam waist. The anomalous phase change between $\hat{z} = 0$ and $z = \pm \infty$ for any mode may be given by the expression

$$\delta (m, n) = \frac{\pi}{2} (m + n + 1)$$

... (3.11)

In the example $m = n$ for each of the modes, and this leads to the above result. Expression 3.10, combined with the uniform phase distribution at the beam waist plane, determines that the far field must also have a uniform phase distribution.

This special result is useful as it applies in many cases of practical interest. For example, beams
with circular/square symmetry will be composed only of modes of this type. Because of this we may expect as a general result that a beam of circular/square symmetry with a uniform phase distribution in the beam waist plane must also have a uniform phase distribution in the far field. A uniform phase distribution at the waist plane also leads to a uniform far field phase distribution when a beam is composed only of modes for which \((m+n+1)\) is even. However, this situation does not arise very often in practice.

### 3.2 Apertures and beam truncation

One of the most valuable applications for multi-mode analysis is in understanding the effects of apertures upon a beam. This becomes an important subject when we design compact optical systems as we then need to determine how small the optical elements (lenses, mirrors, etc) can be without the effects of beam truncation by the finite apertures becoming too severe.

Most optical systems employ elements which are circular and the beam axis is generally arranged to pass through their centres. We can therefore regard such an element as being placed within a circular aperture through which the beam must pass. Given the circular symmetry which arises in most cases it is convenient to make use of a cylindrical co-ordinate system.

The derivation of the mode expressions appropriate for Cartesian co-ordinates is given in Chapter 1. In cylindrical co-ordinates a similar method may be used to obtain suitable mode expressions. The solutions obtained may usefully be divided into two types. Firstly, the modes, \(\psi_p\), which have circular symmetry

\[
\psi_p = \frac{1}{\omega} L_p \left( \frac{2r^2}{\omega^2} \right) \exp \left\{ -j(kz - 2\pi ft) \right\} \exp \left\{ j\Phi_p + r^2 \left[ \left( \frac{1}{\omega} \right)^2 + \frac{jk}{2R} \right] \right\} \quad \ldots (3.12)
\]

where \(r\) is the radial co-ordinate, \(p\) is the radial mode number, \(L_p\) is a Laguerre polynomial of degree \(p\) as defined in appendix 1, and

\[
\Phi_p = (2p + 1) \arctan \left( \frac{\lambda z}{\pi \omega \theta_0} \right) \quad \ldots (3.13)
\]

Secondly, modes where the field pattern varies with the angular rotation co-ordinate, \(\theta\), are of the form

\[
\psi_{pi} = \frac{1}{\omega} \sqrt{\frac{p!}{(p - i)!}} \left( \frac{2r^2}{\omega^2} \right)^i L_p \left( \frac{2r^2}{\omega^2} \right) \exp \left\{ -j(kz - 2\pi ft) + j\Phi_{pi} + r^2 \left[ \left( \frac{1}{\omega} \right)^2 + \frac{jk}{2R} \right] + ji\theta \right\} \quad \ldots (3.14)
\]

where \(L_p^i\) is a generalised Laguerre polynomial and \(i\) the rotational mode number. Note that when \(i = 0\) the modes revert to the form shown in expression 3.12 and have rotational symmetry.

In many cases of practical interest the beams and apertures in compact systems may be assumed to have circular symmetry and the \(\phi p\) modes may be neglected. In order to illustrate the way in which we may employ Gaussian mode techniques to analyse the effects of aperture truncation we can use the simple example of a beam with circular symmetry being coupled through a hole of radius, \(a\). We can represent the beam, \(\Psi\), directed onto the aperture plane as a linear combination of \(\psi_p\) modes:

\[
\Psi = \sum_p E_p \psi_p \quad \ldots (3.15)
\]

Where the complex values, \(E_p\), represent the magnitude and relative phase of each contribution.
made by the relevant \( p \)'th mode. The truncated field distribution, \( \Psi' \), in a plane immediately following the aperture will be of the form

\[
\Psi' = \sum_p E_p \psi_p \text{ when } r < a
\]

and

\[
\Psi' = 0 \text{ when } r \geq a
\]...

(3.16)

The total beam power, \( P \), coupled through the aperture may be obtained from the integral

\[
P = \int_0^{+\infty} 2\pi r \Psi' \Psi'^* \, dr
\]...

(3.17)

which, from expression 3.15, is equivalent to

\[
P = \int_0^a 2\pi r \sum_p \sum_q E_p \psi_p E_q^* \psi_q^* \, dr
\]...

(3.18)

In chapter 1 we introduced the notation \( \langle \rangle \) to represent the integral over the whole plane normal to the beam axis. Here it is convenient to introduce a similar notation which indicates integration over a finite portion of a plane. We hence may define the notation \( \{ \} \) to represent an integration over a limited plane area, such that

\[
\{ \Psi|\Psi \} = \int_0^a 2\pi r \Psi \Psi^* \, dr
\]...

(3.19)

and

\[
\{ \psi_p|\psi_q \} = \int_0^a 2\pi r \psi_p \psi_q^* \, dr
\]...

(3.20)

It is also useful to note that, because \( \Psi' \) is zero outside the aperture,

\[
\langle \Psi'|\Psi' \rangle = \{ \Psi'|\Psi' \} = \{ \Psi|\Psi \}
\]...

(3.21)

and we can substitute any of these integrals for each other whenever it is convenient to do so - e.g. we may replace \( \Psi' \) with \( \Psi \) in most of the integrals over the aperture opening or where the field is zero outside this region.

We may write that

\[
\{ \Psi|\Psi \} = \sum_p \sum_q E_p E_q^* \{ \psi_p|\psi_q \}
\]...

(3.22)

and the total power, \( I \), incident upon the aperture will be

\[
I = \langle \Psi|\Psi \rangle
\]...

(3.23)

The fraction of power transmitted through the aperture may then be defined as the power transmission efficiency, \( T \), which may be obtain from

\[
T = \frac{I}{P} = \frac{\{ \Psi|\Psi \}}{\langle \Psi|\Psi \rangle}
\]...

(3.24)

which allows us to determine how effectively power is transmitted through the aperture.

Power passing through the aperture may be coupled into a beam, \( \Phi \), which represents the pattern of sensitivity of a given detector. In the absence of an aperture ideal coupling would occur if \( \Psi = \Phi \). However the aperture alters the field pattern at the plane just beyond the aperture. The
amount of power coupled, \( P_c \), between the truncated source beam and the detector beam can be obtained from
\[
P_c = |\langle \Phi | \Psi \rangle|^2 \tag{3.25}
\]
provided that the detector beam has been normalised correctly. From this we can say that the efficiency of power coupling, \( N \), between the beams will be
\[
N = \frac{|\langle \Phi | \Psi \rangle|^2}{\langle \Phi | \Phi \rangle \langle \Psi | \Psi \rangle} \tag{3.26}
\]
where we have allowed for the possibility that the detector beam may not be pre-normalised correctly.

In many practical cases we are concerned simply with the use of fundamental mode Gaussian beams. For an input beam of this type we may say that for all values of \( p \) greater than zero, and we can write that
\[
\Psi' = E \psi_0 \text{ for } r < a \tag{3.27}
\]
The fractional power transmission through an aperture, radius \( a \), centred on the beam axis, will hence be
\[
T_0 = \frac{\langle \psi_0 | \psi_0 \rangle}{\langle \psi_0 | \psi_0 \rangle} \tag{3.28}
\]
Having defined \( \langle \psi_0 | \psi_0 \rangle \) to be unity it follows that
\[
T_0 = \int_0^{a} 2\pi \left( \frac{1}{\omega} \right)^2 \frac{2}{\pi} \text{Exp} \left( -\frac{2r^2}{\omega^2} \right) r \, dr \tag{3.29}
\]
By reference to a suitable table of standard integrals we may find the solution of this integral and obtain the result
\[
T_0 = 1 - \text{Exp} \left( -\frac{2a^2}{\omega^2} \right) \tag{3.30}
\]
Expression 3.28 allows us to specify the amount of power coupled through the aperture. This in itself doesn’t tell us how much power emerges in some specific pattern or mode. Hence we will usually then wish to discover either how efficiently power is coupled into a chosen detector beam or determine the resulting beam pattern which radiates along from the aperture.

The truncated pattern at the aperture may be regarded as generating a new multi-mode beam, whose field distribution at the aperture plane is \( \Psi' \). We can therefore define an appropriate set of modes to be such that
\[
\Psi' = \sum_{q} A_q \psi'_q \tag{3.31}
\]
where the \( \psi'_q \) are a series of mutually orthogonal Gauss-Laguerre modes which have a beam size, \( \omega' \), and phasefront radius, \( R' \), at the aperture plane. The coefficients, \( A_q \), determine the magnitude and relative phase of the contribution the \( q'\text{th} \) mode makes to the truncated beam. These coefficients may be calculated using the methods outlined earlier. This produces the result
\[
A_q = \{ \Psi | \psi_q' \} \tag{3.32}
\]
i.e. for the example under consideration where the input beam is simply a fundamental Gaussian we may write that
\[
A_q = \int_{0}^{a} \frac{1}{\omega \omega'} L_q \left( \frac{2r^2}{\omega'^2} \right) \text{Exp} \left[ j \delta_q - r^2 (\alpha + j\beta) \right] 2\pi r \, dr \tag{3.33}
\]
where

\[ \delta_q = (2q + 1) \arctan \left( \frac{\lambda z}{\pi \omega_0} \right) - \arctan \left( \frac{\lambda z}{\pi \omega_0} \right) \quad \ldots (3.34) \]

\[ \alpha = \frac{1}{\omega^2} + 1 \quad \ldots (3.35) \]

\[ \beta = \frac{k}{2} \left( \frac{1}{R} - \frac{1}{R'} \right) \quad \ldots (3.36) \]

We are often only concerned to establish the amount of power coupled into the fundamental mode, \( q = 0 \), as this enables us to determine how efficiently power passing through the aperture may be coupled to a detector beam of this simple form. By setting \( q = 0 \) only, we obtain

\[ A_0 = \frac{2 \exp(j\delta_0)}{\omega \omega'} \int_0^a \exp \left[ -r^2(\alpha + j\beta) \right] 2\pi r \, dr \quad \ldots (3.37) \]

Now

\[ \langle \psi_0 | \psi_0 \rangle = \langle \psi'_0 | \psi'_0 \rangle = 1 \quad \ldots (3.38) \]

hence the power coupling efficiency, \( N_0 \), through the aperture between the input and output fundamental modes will be

\[ N_0 = |A_0|^2 \quad \ldots (3.39) \]

which, by referring to a standard table of integrals, leads to the result

\[ N_0 = \frac{4 \left( 1 - 2 \exp(-a^2\alpha) \cos(a^2\beta) + \exp(-2a^2\alpha) \right)}{(\omega \omega')^2 (\alpha + \beta)^2} \quad \ldots (3.40) \]

The power coupling efficiency, \( N_0 \), obtained here includes both the losses caused by the failure of some field to pass the aperture and those which arise because the field pattern following the aperture is no longer a simple Gaussian and cannot be ideally coupled into a fundamental mode output beam.

It is useful to note that, in general, \( N_0 \) is not invariably maximised by setting \( \omega' \) equal to \( \omega \). This may be seen as an effect of the aperture tending to 'narrow' the beam width by removing the field beyond \( r = a \). When the aperture size, \( a \), is far larger than the input beam size, \( \omega \), then the effects of truncation are negligible and optimum coupling will arise when \( \omega' \approx \omega \) and \( R' \approx R \). However, when \( a \) is comparable with (or less than) \( \omega \) then the beam coupling efficiency may be improved by altering \( \omega' \).
Figure 3.3 illustrates how the fractional power transmission, $T_0$, and fundamental mode coupling efficiency, $N_0$, vary with the ratio $a/\omega$. For comparison, $N_0$ is shown both for $\omega' = \omega$ and for $\omega' = 0.8\omega$. It can be seen that, once $a/\omega$ is small enough, the power coupling efficiency can be improved by reducing $\omega'$. For the graphs shown it is assumed that $\beta = 0$ (i.e. that the input and output beams have identical phasefront curvatures at the aperture plane).

By their very nature, when designing compact optical systems we usually want to make the optical elements as small as possible. Clearly the choice of beam size will have consequences for the system performance, hence there will be a minimum beam waist size acceptable for a given application. Similarly, the choice of $a/\omega$, will determine the degree to which aperture truncation effects may degrade system performance. The use of single-mode beams is generally advisable as it eases the design process and usually enables the optical system to work well over a wide frequency range. In practice the fundamental mode is used most often. This is partly because it is relatively easy to produce, but it also proves to be the mode which generally offers the lowest truncation losses. This is because, as the mode number is increased, there is a tendency for the field pattern to have more of the beam power further from the axis.

Whilst there is no obvious ‘ideal’ choice for $a/\omega$, the value of $a/\omega = 1.5$ has come to be commonly adopted in practical designs. A single aperture of this size will produce a coupling efficiency, $N_0 = 0.989$, i.e. a beam power loss of -0.05dB when transmitting a fundamental mode beam. This is low enough to be insignificant in most circumstances. If $a/\omega$ is reduced to 1.3 the coupling loss rises to around -0.15dB, which is comparable with the reflection and absorption losses of a typical lens. The losses when $a/\omega = 1.5$ are so low that there is rarely any point in increasing the value significantly. The choice is also convenient in that it gives us a simple ‘rule’ that lenses, etc, should have a diameter of three times the beam size.
3.3 Beam offsets and measuring beam patterns

A knowledge of how power coupling varies when the beam axes are not coincident is useful when assessing the magnitude of any problems which may arise due to system misalignments. It can also provide us with a way of measuring the power or field profile of an unknown beam. For the sake of example we can concentrate on the simplest case and evaluate the coupling between two fundamental Gaussian modes whose beam axes are parallel but separated by a distance, $d$.

The beams, $\Psi$, and $\Psi'$ can be written as the product of an $x$- and $y$-dependent pair of one-dimensional mode expressions. i.e.

$$\Psi = \Psi(x)\Psi(y) \quad ; \quad \Psi' = \Psi'(x)\Psi'(y)$$  \hspace{1cm} \text{... (3.41)}

where $\Psi(x)$, etc, are of the form defined previously. The total field coupling, $C$, between the beams will be

$$C = \langle\Psi|\Psi'\rangle = \langle\Psi(x)|\Psi'(x)\rangle\langle\Psi'(y)|\Psi'(y)\rangle$$  \hspace{1cm} \text{... (3.42)}

For simplicity we can assume that the axes have been arranged so that the offset, $d$, is in the $x$-direction and that the beams are symmetric. We can then say that, ignoring terms which only affect the absolute phase,

$$\langle\Psi(y)|\Psi'(y)\rangle = S \int_{-\infty}^{\infty} \exp\left[-y^2\left((\alpha + \alpha') + j(\beta + \beta')\right)\right] \, dy$$  \hspace{1cm} \text{... (3.43)}

$$\langle\Psi(x)|\Psi'(x)\rangle = S \int_{-\infty}^{\infty} \exp\left[-x^2(\alpha + j\beta - (x - d)^2(\alpha' - j\beta')\right]\, dx$$  \hspace{1cm} \text{... (3.44)}

where

$$S = \frac{2}{\pi^{1/2}}$$  \hspace{1cm} \text{... (3.45)}

$$\alpha = \frac{1}{\omega^2} \quad ; \quad \alpha' = \frac{1}{\omega'^2}$$  \hspace{1cm} \text{... (3.46 ; 3.47)}

$$\beta = \frac{k}{2R} \quad ; \quad \beta' = \frac{k}{2R'}$$  \hspace{1cm} \text{... (3.48 ; 3.49)}

referring to a suitable table of integrals we may find the standard result

$$\int_{-\infty}^{\infty} \exp\left[-(ax^2 + 2bx + c)\right] \, dx = \frac{\pi}{a} \exp\left\{\frac{b^2}{a} - c\right\}$$  \hspace{1cm} \text{... (3.50)}

(provided that a $a \neq 0$)

Using this to solve expressions 3.43 and 3.44 we obtain

$$\langle\Psi(y)|\Psi'(y)\rangle = \sqrt{\frac{2}{A_{\omega\omega'}}}$$  \hspace{1cm} \text{... (3.51)}

where

$$A_y = (\alpha + \alpha') + j(\beta - \beta')$$  \hspace{1cm} \text{... (3.52)}

and

$$\langle\Psi(x)|\Psi'(x)\rangle = \exp(-d^2B_x)\sqrt{\frac{2}{A_{\omega\omega'}}}$$  \hspace{1cm} \text{... (3.53)}

where
J. C. G. Lesurf — 35 — Millimetre-wave Optics, etc

\[ A_x = (\alpha + \alpha') + j(\beta - \beta') \quad ; \quad B_x = (\alpha' - j\beta')(1 - \frac{\alpha' - j\beta'}{A_x}) \quad \text{... (3.54 ; 3.55)} \]

By substituting expressions 3.50 to 3.54 into 3.41 we can now define how the field coupling, \( C \), between the two beams will vary as a function of the offset distance, \( d \).

In practice we often wish to check that the beam waist location and size of a beam are correct. This can be done by employing a detector with an antenna which responds to a known beam pattern and observing how the coupled field magnitude and phase vary as the detector offset position (i.e., \( d \)) is altered. The expressions given above then enable us to use the measurements to establish beam waist parameters. Unfortunately, we are frequently only able to employ a power detector which cannot provide us with information on the relative phase variations as we move around the beam we wish to measure.

The power coupling efficiency, \( N(d) \), between the offset beams may be calculated from

\[ N(d) = |C|^2 \quad \text{... (3.56)} \]

In principle we can then demonstrate that, if we measure how \( N(d) \) varies at three separate planes along a beam we can deduce the beam’s waist size and location. Hence we can choose to make our measurements at any three suitable planes. If we then place our power detector, whose beam waist size is \( \omega_0' \), such that its waist plane is coincident with that of the beam we wish to measure, at the common waist plane we can calculate that \( N(d) \) will be of the form

\[ N(d) = \frac{4 \text{ Exp}\left(-2d^2/\xi^2\right)}{(\omega_0'/\omega_0 + \omega_0/\omega_0')^2} \quad \text{... (3.57)} \]

where \( \omega_0' \) is the beam waist size of the detector beam or antenna, and

\[ \xi^2 = \omega_0^2 + \omega_0'^2 \quad \text{... (3.58)} \]

From this result we can see that the measured pattern of \( N(d) \) has a Gaussian profile. Knowing \( \omega_0' \) we can use the measured profile to determine the beam’s waist size.

To check that we have correctly identified the beam waist plane we need only measure the beam widths at two other planes, spaced equal distances along the beam on either side of the waist. From the symmetric nature of a Gaussian beam these other planes should produce beam sizes which are; i) bigger than \( \xi \), and ii) equal to each other. If this is not the case, then we did not initially have the waist planes collocated.
Chapter 4
Antennas and feed systems

In the first three chapters of this book we have established how we can deal with transmitting signals from place to place using beams in free space. Many of the actual signal sources, detectors, etc, are mounted in a length of metal waveguide. We therefore require some way of coupling signal power back and forth between free space beams and waveguides. We need some form of antenna or feed system to transform one sort of mode into another.

![Diagram of antenna, transmission line, and back-short arrangement](image)

**Fig. 4.1** Combination of antenna, transmission line, and back-short arranged to couple a device to an electromagnetic signal propagating in free space.

Figure 4.1 illustrates one of the most common arrangements used for mm-wave devices. The arrangement represents how power in a free space beam may be coupled into a small device such as a detector diode mounted in a short length of metal waveguide. The manner in which signal power is coupled from a waveguide into a device will be discussed in a later chapter. Here we will concentrate on how power is transferred from free space into (or out of) a waveguide.

### 4.1 Basic Coherent Antenna Theory

Before analysing the properties of mm-wave systems in terms of Gaussian modes it is useful to establish some of the general properties of antennas. An antenna may be regarded as a sort of 'transformer', taking one form of field pattern and converting it into another. Looked at in this way a number of the basic properties become fairly clear.

Firstly, an antenna cannot of itself produce any 'extra' signal power. Even a perfect antenna can only output the same amount of signal power we initially put in. For any real antenna, of course, some signal power will be 'lost' - for example, currents set up in metal parts will cause signal power to be wasted warming up surface resistances.

Secondly, an antenna is a reciprocal system. i.e. it works in much the same way irrespective of which way we pass signal power through it. To see just what this means, consider the arrangement shown in Fig 4.1 but with the detector replaced by a signal source. The source will produce a mode pattern, $\Psi_s$, which propagates along the guide to the antenna. The antenna then transforms this into a pattern, $\Psi_a$, which it radiates out into space.

Imagine taking a video of the field pattern moving through the system. If we replay the video backwards, the field, $\Psi_s$, appears to enter the antenna from free space and is transformed into
the pattern, $\Psi_g$, moving along the guide away from the antenna. For an antenna dealing with a coherent signal we cannot tell by looking at the video which way we are running the playback! This is a very important feature of antennas. It means that the only real difference between a ‘transmitting’ antenna (one used to radiate a signal out into space) and a ‘receiving’ one (used for collecting power from free space) is the direction we have chosen to pass signal power through the system.

The reciprocal nature of antennas means that we often obtain good results by using essentially identical antennas to radiate and receive beams. It also turns out to be very helpful when we analyse or design antennas. This is because we can often choose to carry out our analysis by thinking of the antenna as either transmitting or receiving a signal irrespective of its actual role.

Figure 4.2 illustrates one of the simplest theoretical possible forms of antenna, the Hertzian Dipole. These are defined to have a length, $L \ll \lambda$, where $\lambda$ is the free space wavelength of the radiation produced. Antennas of this form are not normally employed for mm-wave systems, but it is a useful example which serves to establish some basic antenna properties. More complicated and practical antenna arrangements may often be analysed or designed in terms of considering them as an array of elemental Hertzian Dipoles. So they form a good starting point for considering antenna properties.

We can inject a signal at some chosen frequency into the antenna via the pair of parallel wires connected to the break near its center. The actual antenna consists of two short (compared with the signal wavelength) pieces of wire and the signal sets up a uniform alternating current on the antenna. In order for the charge moving at the wire ends to have ‘somewhere to go’ we need to arrange for some capacitance between the wire ends. This can be done by attaching the ends to some large discs, spheres, or whatever. The electric field between their end areas then acts as a capacitor into which end current may flow.

For simplicity, here we will ignore any effects of the end shapes upon the antenna and regard the arrangement as simply a length, $L$, of wire upon which we have set up a uniform alternating current

$$I(t) = I_0 \sin(\omega t)$$

... (4.1)

By referring to a suitable book on electromagnetism it may be found that this current will generate an electric field, $E$, in the ‘far field’ given by the expression
\[ E = \frac{60\pi LI_0 \sin(\theta) \cos(\omega t - kr)}{\lambda r} \quad \text{(4.2)} \]

where \( r \) is the distance from the dipole centre to the point where we wish to determine the field, \( E \), and \( \theta \) is the angle between the dipole wire and the line connecting the point and the dipole center. (Here the term ‘far field’ means we are considering locations sufficiently distant from the antenna that \( r \gg \lambda \).

The rms field at the point \((r, \theta)\) will therefore be

\[ E(r, \theta) = \frac{60\pi LI_0 \sin(\theta)}{\sqrt{2} \lambda r} \quad \text{(4.3)} \]

If we imagine the dipole as being placed at the center of a sphere of radius, \( r \), then the power per unit area, \( P(\bar{E}) \), passing through the surface of the sphere around the point \((r, \theta)\) will be

\[ P(\theta) = \frac{E^2(\theta)}{Z} \quad \text{(4.4)} \]

where \( Z \) is the impedance of free space. i.e. we may say that,

\[ P(\theta) = \frac{(60\pi)^2 l^2 r_0^2 \sin^2(\theta)}{2Z\lambda^2 r^2} \quad \text{(4.5)} \]

The way in which \( P \) varies with direction is generally referred to as the Power Pattern or Antenna Pattern of the antenna.

More generally for various kinds of antenna, in a spherical co-ordinate \((r, \theta, \phi)\) system the power pattern often depends upon both angles and may be given as an appropriate function, \( P(\theta, \phi) \). In this particular case the pattern has rotational symmetry in the \( \phi \) direction - i.e. the power does not vary when we move around the antenna provided we keep \( \theta \) constant.

Some authors use the term ‘antenna pattern’ to mean \( E(\theta, \psi) \) whilst others use it to mean \( P(\theta, \psi) \). To avoid this ambiguity, here we will use the term ‘power pattern’ for \( P(\theta, \psi) \) and ‘antenna pattern’ for \( E(\theta, \psi) \).

From expression 4.5 we can see that the power pattern is directional, more power being transmitted in some directions than others. In particular no power is transmitted along the two directions \( \theta = 0 \) and \( \theta = \pi \). Antenna engineers often invoke the concept of an ideal omnidirectional antenna, the isotropic radiator which is assumed to radiate power equally in all directions. The behaviour of a particular arrangement may then be compared with an isotropic antenna.

Consider comparing a pair of antennas, one a Hertzian dipole, the other omnidirectional, being used in turn to radiate the same total power out into space. The Hertzian dipole will radiate less power in some directions (those around \( \theta = 0 \) or \( \theta = \pi \)) than the omnidirectional antenna. Since the total radiated powers are equal it follows that more power is transmitted by the Hertzian dipole in some other directions (in fact, in the directions around \( \theta = \pi/2 \)).

As a consequence, if we were to set up a detector some distance from an omnidirectional antenna and measure what happens when it is replaced by a Hertzian dipole we find that - in the \( \theta = \pi/2 \) plane - the received signal increases by around 50 per cent. So far as the detector is concerned the result is no different than if we had amplified the transmitted power. From this has grown the concept of antenna gain.

We can define the gain of an antenna as being the factor by which we have increased the
received signal by changing the antenna from some standard type (e.g. isotropic). When using the term 'gain' however it is important to note that the total power radiated has not been increased. Instead, we have arranged not to 'waste' power by transmitting in directions where it will not be received. Because of this, the gain and the directionality of an antenna are closely linked.

In fact, if we are dealing with a single coherent mode, it can be shown that it is impossible to produce a truly isotropic radiator. This may be seen by appeal to a rule from topology - "You can't comb a hairy billiard ball smooth!". The coherent wave radiated from an antenna will have its electric field perpendicular to its direction of propagation. If we consider the field moving out through a large sphere centered on the antenna we find that the power is moving radially. Hence the field electric vector is everywhere parallel to the local surface of the sphere.

In order to satisfy Maxwell's equations in free space, we cannot let the field distribution have any discontinuities. i.e. the field vector direction and magnitude must vary smoothly from place to place on the sphere. We can then think of the field as like the hairs on a hairy billiard ball. No matter how they are combed, there will always be a tuft somewhere and a gap somewhere else. This would represent a discontinuity as the field direction (e.g. the direction the hairs point) changes abruptly as we move through the tuft or gap. To avoid this we can arrange for the magnitude of the field (e.g. length of the hairs) to fall smoothly to zero as we approach either the tuft or gap.

When dealing with a vector field, where magnitude, direction (or polarisation), and phase are uniquely defined we cannot, therefore, make a real isotropic radiator. However we can - in principle at least - make one if we are only concerned with power.

For example: a single dipole will transmit no field in the directions along its axis. We can, however, use a pair of crossed dipoles and supply them with a signal at the same frequency, but which are 90° out of phase. Now we are essentially radiating two fields which are orthogonal in phase. The 'gaps' in each pattern are filled by radiation from the other dipole. The power pattern is now much more even than for a single dipole. The radiated field of such a pair is elliptically polarised. The precise polarisation state of the wave we observe depends upon the angle between the local wave propagation direction and the plane of the two dipoles.

Conventionally, despite the inability to make a real isotropic antenna, the gain of an antenna is often quoted as being so many dB compared with an (assumed) isotropic radiator. This gain is, in practice, normally inferred from measuring the directional pattern of the antenna, or by comparing it with a 'standard' antenna whose gain or pattern are taken as 'known'.
Figure 4.3 shows a possible power pattern for a 'high-gain' antenna. A real pattern will be fairly complex. Most of the transmitted power will be beamed into a fairly narrow solid angle which is conventionally called the main lobe. The remaining power being sent in other directions into what are called the sidelobes. The detailed analysis of such a complex pattern is quite difficult. Fortunately, for most purposes we can replace it with a much simpler 'idealised' antenna pattern. Figure 4.3 also illustrates an 'idealised' pattern. Note that both patterns only show how the radiated power varies with one of the co-ordinate angles. Strictly speaking we should picture the variations in power with both \( \theta \) and \( \phi \). For many antennas the actual pattern can be assumed to have circular symmetry about the axis of the main lobe, hence only one 'slice' through the pattern need be shown.

The main lobe of the real pattern is taken to be the solid angle within the power minima which surround the direction of maximum gain. We can calculate the amount of power radiated in the main lobe by integrating \( P(\theta, \phi) \) over this solid angle. i.e. we can say that

\[
\int_{\text{Main}} P(\theta, \phi) \, d\Omega = P_M \Omega_M \quad \ldots \quad (4.6)
\]

where \( P_M \) is the maximum value of \( P(\theta, \phi) \) and the integral is over the main lobe only. By defining the normalised power pattern

\[
P_n(\theta, \phi) \equiv \frac{P(\theta, \phi)}{P_M} \quad \ldots \quad (4.7)
\]

this can also be written in the form

\[
\int_{\text{Main}} P_n(\theta, \phi) \, d\Omega \equiv \Omega_M \quad \ldots \quad (4.8)
\]

This lets us define the solid angle, \( \Omega_M \), of an idealised main lobe which would radiate the same total power as the real main lobe but has a uniform value of \( P(\theta, \phi) \) equal to the maximum value, \( P_M \), of the real antenna. The value, \( \Omega_M \), is conventionally referred to at the main beam solid angle. Any remaining power, radiated in the sidelobes of the real antenna, is assumed for the idealised pattern to be radiated uniformly in all directions outside \( \Omega_M \). It is also useful to define another conventional quantity, \( \Omega_A \), the antenna solid angle, via the expression
\[ \int_{4\pi} P_n(\theta, \phi) \, d\Omega = \Omega_A \]  
\[ \text{(4.9)} \]

Hence \( \Omega_A \) is the solid angle into which all the output power would be directed if the pattern were uniform, and no power was radiated into the sidelobes. Power radiated into the sidelobes is generally wasted. Many directional antennas are designed to minimise their sidelobes and thus, in practice, we can often assume these two solid angle values are effectively the same.

The antenna gain only arises because power is largely directed into a limited solid angle, \( \Omega_A \), instead of being radiated into \( 4\pi \) sterads. From the solid geometry of the radiation pattern we can observe that the gain, \( G \), of the idealised antenna pattern compared with an isotropic radiator will be such that

\[ G = \frac{4\pi}{\Omega_A} \]  
\[ \text{(4.10)} \]

Hence we need only be given either the gain or the directionality (i.e. the antenna solid angle) of an antenna to be able to calculate the other value.

Consider now a receiving antenna, placed so that it intercepts a field, \( E \), over a collecting area, \( A \). The received power will be

\[ W = \frac{|E|^2 A}{Z} \]  
\[ \text{(4.11)} \]

where \( Z \) is the impedance of free space.

Earlier in this chapter it was pointed out that the distinction between a receiving and transmitting antenna is solely one of use - i.e. of which direction the field passes through the antenna. We can, therefore, expect that there should be some relationship between the ‘collecting area’ of an antenna when used to receive radiation and it’s gain or directionality when used to transmit. By referring to a suitable book on antennas it will be found that the effective area of an antenna, \( A \), and its antenna angle, \( \Omega_A \), are related via the expression

\[ \lambda^2 = A\Omega_A \]  
\[ \text{(4.12)} \]

where \( \lambda \) is the radiation wavelength. Once the wavelength is known the main properties of the antenna can be determined by knowing the value of just one of the three parameters, \( A \), \( \Omega_A \), or \( G \). The other two may be then be calculated if required.

When a system of the type illustrated in 4.1 is used with a waveguide mounted source a field, \( \Psi_g \), in the waveguide will be transformed into a field, \( \Psi_s \), radiated into space. If only one mode pattern, \( \Psi_g \), may propagate along the guide at this frequency the pattern radiated into space is uniquely defined. The antenna is a reciprocal system. Hence when used as a receiver it will be sensitive to the specific field pattern, \( \Psi_s \), in free space. Any power that arrives which is in this particular pattern will be efficiently coupled into the waveguide.

Any field pattern, \( \Psi \), in the region surrounding the antenna may be thought of as being a sum of two orthogonal components, \( \Psi_s \) and \( \Psi_o \). Power in the field, \( \Psi_s \) is transformed into the guide pattern, \( \Psi_g \). Power in the orthogonal field, \( \Psi_o \) may attempt to produce some other pattern in the guide. However, we have arranged that only one guide mode is possible. Hence any power in the field \( \Psi_o \) will be rejected or ignored by the antenna.

Communications engineers, radioastronomers, etc, often use antennas to detect signals which are arriving in the form of plane waves radiated by a distant source. Under these circumstances the
main concern is to arrange for an antenna pattern which responds as strongly as possibly to a plane wave field coming from the direction of the source of interest. i.e. the antenna field distribution, $\Psi_s$, should be as close as possible to a plane wave.

Similarly, in compact optical systems where we are dealing with beams, we would wish $\Psi_s$ to be a similar as possible to the signal field of the beam we wish to receive or transmit. Radioastronomer, etc, may also wish to receive power from an ‘extended’ source - i.e. one which has a measurable angular extent on the sky. Here the ideal antenna field may be designed to respond over a specific, limited solid angle.

In each case the efficiency of power coupling through the antenna will depend upon what fraction of an incoming field, $\Psi$, couples into the pattern, $\Psi_s$, to which the antenna will respond.

When dealing with plane waves, or fields from a source of limited angular size it is very convenient to treat the antenna in terms of its effective area, or gain, or directionality. When dealing with beams, however, we often find that we need to employ the arguments of field coupling outlined in the first three chapters.

The effective antenna area we have defined may not always be clearly identified with a specific physical area. Consider, for example, a dipole. The actual surface area of the dipole wires is very small and varying the wire diameters to alter their area has little or no effect upon the antenna’s behaviour. For a dipole there is no clear physical area which can be seen to relate to the antenna’s ‘effective’ area. This having been said, for many of the most common forms of mm-wave or microwave antennas the physical collecting areas of the lenses or mirrors involved are usually reasonably large compared with $\lambda^2$ and in these cases the antenna’s effective area tends to be at least similar to an identifiable physical area.

### 4.2 Feed horns

In contrast to the dipole, most mm-wave antennas can best be analysed as an aperture which frames a defined field pattern. This means that we can usually identify a physical area which is closely related to the effective area, $A$.

Perhaps the simplest form of antenna we can use with metal waveguide is to leave the end of the guide open to free space. The signal travelling along the guide produces a field pattern at the open end which we can define by knowing the details of the signal modes in metal waveguide. We can then treat the open end as being a rectangular aperture.

For the case of a standard rectangular waveguide carrying just the fundamental TE mode and the electric field will be

$$E = E_0 \sin \left( \frac{\pi x}{a} \right) \exp(\imath \omega t) \quad \text{within the aperture}$$

and

$$E = 0 \quad \text{outside the aperture} \quad \ldots (4.13)$$

The electric field being plane polarised with its vector parallel to the short wall of the guide. The $x$-direction is assumed to be parallel to the longer wall which is of size, $a$.

For standard guide the width and height will be around $\lambda/2$ and $\lambda/4$, hence the effective area
will be approximately $\lambda^2/8$. Using expression 4.12 we can see that for an open waveguide end this implies an antenna angle, $\Omega_A \approx 8$ sterads i.e. about three-quarters of a complete sphere!

Clearly, in order to produce a reasonably directional beam the effective area of the antenna needs to be somewhat larger than $\lambda^2/8$. Hence an open waveguide end is an unsatisfactory antenna for most purposes.

The simplest way to obtain a larger area is to flare out the part of the waveguide near the open end into a pyramidal shape. A wave travelling along the guide sees this as a length of guide whose cross-sectional area increases gradually as it moves towards the open end. This has the effect of ‘stretching out’ the field pattern without drastically altering its shape. If required, the angles may be different for the two pairs of walls and we can, for example, have this lead to a square opening at the ‘base’ of the pyramid shape.

In this way we can produce an open end which is, say, $5\lambda \times 5\lambda$ across. If we input the fundamental TE mode the field at the open end will be

$$E = E_0' \sin \left( \frac{\pi x}{a'} \right) \exp (j\omega t)$$

at an aperture of width, $a' = 5\lambda$. Note that the overall field magnitude must have been reduced by this ‘stretching’ process because the total power leaving the open end cannot exceed what we put in via the waveguide.

For an aperture of area, $A$, and a waveguide of area $\lambda^2/8$ conservation of power allows us to write that

$$E_0' = \frac{E_0 \lambda}{\sqrt{8A}}$$

Taking the example of $A = 25\lambda^2$ we obtain an antenna angle, $\Omega_A = 1/25$ Sterads. i.e. if the main lobe was symmetric it would be about 12 degrees across in any direction. This is now directional enough to be coupled reasonably effectively into a beam. (Or, of course, to collect power from a suitably focussed beam.) Antennas of this general type are called feed horns partly as a reference to their shape, partly because the are used to ‘feed’ radiation from one place to another.

Some care must be taken when designing or building feed horns which have a very steep flare angle or which have a large open end. As a field moves closer to the open end the effective waveguide size increases and the cut-of frequencies for higher TE (or TM) modes will fall below the signal frequency. Any small imperfections in the horn may then alter the field pattern, changing the behaviour of the antenna. Also, the field moving within the horn does not have a plane phasefront (if it did, it would not spread out!). The mode inside the horn will tend to have a spherical phasefront centered on the horn apex. Hence there may not be a perfect transfer of signal between the guide and horn modes. This problem becomes more severe if the horn angle is increased.

In practice, the description of horn behaviour given above is fairly reliable provided:

a) the aperture is no more than a few wavelengths across

b) the flare angle is not more than around 10 - 20 deg. Beyond these limits the analysis and design of feed horns should be undertaken using a more involved method. Fortunately, we can normally make feed horns offering acceptable beam coupling performance within these limitations.
4.3 The Scalar Feed

Having obtained a suitable antenna area, gain, and directivity, we ideally require an antenna which produces (or is sensitive to) a single free space mode - preferably the fundamental Gaussian one with circular symmetry. For the fundamental TE mode, the field at the aperture of a pyramidal horn varies sinusoidally with $x$ but does not alter as we move in the $y$-direction (unless we move outside the aperture!). Hence the pattern radiated by such a feed cannot have the same form in the $x$- and $y$-planes - even if the aperture is square.

If we require a free space pattern which is simply a fundamental Gaussian we need to generate good approximation to a Gaussian field pattern at the horn’s aperture. Various methods have been employed to achieve this. One of the most successful arrangements is called the scalar or conical corrugated feed. This differs from the pyramidal horn in two ways:

i) As the guide size is gradually increased it is changed from rectangular to circular cross-section;

ii) A series of radial corrugations are introduced into the metal walls.

The detailed analysis of the scalar feed is quite complex. For our purposes, however, it is sufficient just to give an outline of the general behaviour.

For a TE mode in a rectangular guide we can expect that, just above the surface of a metal wall, the $E$-field component parallel to the surface and perpendicular to the axis must be zero. However, the $E$-field component perpendicular to the surface can be just as large as it is further towards the centre of the guide. At points a half-wavelength apart along the guide the perpendicular $E$-fields will be out of phase. Hence there will be a potential difference along the guide wall which sets up a current flow along the guide, parallel to the axis of propagation.

Consider what would happen when we cut a series of narrow grooves, each a quarter-wavelength deep, into the guide walls. The grooves are placed so that current wanting to flow along the guide is forced to travel down the groove, across the bottom, and back up the other side. Because of the time this takes the currents at the two edges of a groove are 180 deg out of phase. Although the currents are forced to go 'the long way round' the groove edges are very close to each other. As a result, any fields set up at a groove edge are effectively cancelled by an equal and opposite adjacent field. It becomes impossible to set up a potential difference or current along the wall.

The effect of the grooves is to 'kill' the $E$-field near the wall surface which cannot now have a non-zero component perpendicular to the surface. The field pattern in the guide must now fade gracefully to zero as we approach any of the walls.

If we wish, we can make a square pyramidal horn with corrugated walls. This can be used to produce an aperture pattern which varies symmetrically. The resulting pattern does not, however, radiate a simple fundamental Gaussian. Instead it produces a multi-mode beam with quite a lot of power spread over the first few modes. By altering the waveguide profile to circular we can change the field profile into a shape which is a good approximation of Gaussian.
Figure 4.4 illustrates a typical conical corrugated feed. For a real feed horn of this type it is useful to introduce two extra features not mentioned above. Firstly, the conversion from rectangular to circular guide takes place before the guide is flared out into a cone. Secondly, the first few groove at the ‘throat’ of the horn (i.e. the end nearest the waveguide) are different in length from the rest. Both of these features make the transition from the normal rectangular TE mode to the field pattern in the corrugated horn more gradual. If the change from rectangular guide to corrugated cone is too abrupt the field patterns do not couple efficiently into one another and a significant amount of signal power may be reflected.

A detailed analysis of this type of horn reveals that the field at the aperture can be well described as a spherical ‘cap’. The radius of curvature, $R$, of the phasefront which passes through the aperture rim is at the cone’s apex. It may be shown that the field distribution at the aperture is plane polarised and of the form

$$E_y = AJ_0(a r) \exp\left(\frac{j k r^2}{2R}\right) \exp(j 2\pi ft)$$

... (4.16)

where

$$a = \frac{2.405}{\lambda}$$

... (4.17)

and

$$J_0^T(\alpha r) = J_0(\alpha r)$$

when $r < a$ ; $J_0^T = 0$ when $r > a$

... (4.19)

where, $J_0$, is a Bessel function of the first kind, zeroth order and, $a$, the aperture radius. Note that $J_0(2.405) = 0$, hence the distribution is arranged so that the field falls smoothly to a zero at the aperture rim.

By using the methods outlined in Chapter 3 it is possible to define a series of Gauss-Laguerre free space modes which would produce the aperture distribution of expression 4.19. From this analysis we find that the aperture field may be represented as

$$J_0^T(\alpha r) = \sum_p \frac{1}{\omega^p} L_p \left(\frac{2r^2}{\omega^2}\right) \exp\left(-\frac{r^2}{\omega^2}\right)$$

... (4.20)
where $\omega$ is the beam size which, at the aperture plane may be taken from

$$\omega = 0.6435 a$$

... (4.21)

The power radiated in the $p$'th mode will be $|A_p|^2$. If we arrange that the total radiated power, $P_T$, is unity, i.e. if we arrange that

$$P_T = \sum_p |A_p|^2 = 1$$

... (4.22)

then it can be shown that

$$|A_0|^2 = 0.9792, \quad |A_1|^2 = 4.9 \times 10^{-9}, \quad |A_2|^2 = 1.45 \times 10^{-2}$$

$$|A_4|^2 = 1.86 \times 10^{-3} \text{ etc}$$

From this result it can be seen that around 98 per cent of the power radiated by the conical corrugated feed will be in the fundamental Gaussian mode. For many purposes we may regard such a feed horn as an ideal antenna for coupling between a waveguide device and a free space Gaussian beam. From the beam size, $\omega$, and the phasefront radius of curvature, $R$, at the aperture we can also work out an effective beam waist radius and location for the antenna beam produced by the horn. It is then normally convenient to regard the horn performance as depending upon this waist size and position.

The main problems of conical corrugated horns are consequences of the grooved structure which produces their highly useful antenna pattern. The grooves must be around one-quarter of a wavelength deep. They must also be fairly narrow and closely packed. Millimetre-wave horns are usually produced by electroforming. A steel or aluminium 'mandrel' is turned on a lathe. The actual horn is then formed by copper electroforming onto the outside of the mandril.

When making a smooth horn the mandril may be pulled out and re-used. This is impossible when the mandril is radially grooved because this locks it into the grown copper. Some manufacturers have used a technique of combining a smooth metal horn with a plastic (usually injection moulded) pattern for the grooved outer part. The horn is pulled out and the plastic then dissolved by a suitable solvent. An alternative approach is to use an aluminium mandril which is then eaten out using a reactive liquid which attacks aluminium but not copper. This produces excellent results, but means that a new mandril must be made for every horn.

Above around 300 GHz (1 mm wavelength) it becomes difficult to machine and form the required grooves as they are becoming too small to cope with accurately. Hence corrugated feeds are rare above a couple of hundred GHz.

The use of quarter-wavelength grooves also 'tunes' the feed horn. i.e. we can only obtain the desired behaviour for signals whose wavelength is appropriate for a given horn. In practice this tends to mean that corrugated feeds work well over about a 20 per cent frequency range. Outside this range the antenna pattern deteriorates. The grooves also tend to make the horn couple less well to the waveguide at frequencies well away from the 'design' frequency. Hence the signal losses also increase outside a limited range.

Perhaps ironically, the imperfections of a smooth metal horn can sometime mean that they work better than a corrugated one at high frequencies! At frequencies around 300 GHz and above the losses due to surface resistance in a copper horn become significant. This tends to absorb signal power from the field near the walls of the horn. The result may be to produce an aperture field pattern which is closer to Gaussian than would otherwise occur. Hence the antenna pattern of a
simple smooth horn may be quite useful although the result is obtained by a process of selective signal loss.

### 4.4 Vivaldi and other E-plane antennas.

A number of alternatives to the conical corrugated feed have been proposed. Each offering a particular combination of good and bad points. Rather than spending time listing them and discussing their relative merits it is more profitable to compare the general idea of feed horn 'aperture' antennas with another class of antennas based upon a different sort of waveguide. The behaviour of conical and rectangular feed horns can be described by considering them as tapered lengths of standard metallic waveguide. In a similar way, the properties of another class of antenna can be obtained from their similarity to E-plane guide.

![Fig. 4.5 Linear (a) and ‘Vivaldi’ (exponential) (b) planar antennas](image)

**Fig. 4.5** Linear (a) and ‘Vivaldi’ (exponential) (b) planar antennas

Figure 4.5 illustrates a couple of pieces of E-plane waveguide, linked to two slightly different antenna arrangements. The guiding structure consists of a narrow gap between the parallel edges of two metal plates. A signal may be transmitted along the gap, power being carried by the EM field which surrounds it. The properties of this type of guide are almost the same as those of twin-feeder which will be discussed in the next chapter.

E-plane guide is often mounted on a dielectric substrate in order to give is stability and strength. This will, however, affect the field pattern near the gap. A detailed analysis of the arrangement indicates that, provided the thickness, \( t \), of any dielectric substrate satisfies the condition

\[
t \leq \frac{\lambda}{4(\varepsilon - 1)}
\]

... (4.23)

and the gap width is less than a half-wavelength, then power may be transmitted in a single TEM mode.

When discussing rectangular waveguide it was pointed out that their modes are of two types: TE modes where the electric field is wholly transverse; TM modes where the magnetic field is wholly transverse. For an electromagnetic wave in free space both the electric and magnetic fields are wholly transverse. This property is shared by the E-plane mode of interest here. TEM stands for Transverse Electric and Magnetic fields.

Just as was the case with rectangular waveguide we may produce a form of antenna by flaring open the gap before terminating the guide at an open end. Provided the flare angle is gentle the field pattern 'stretches' smoothly until it covers a significant area by the time it arrives at the end of the guide. Although the system differs from rectangular guide in that there is no obvious physical aperture, we can still define the field pattern at the final plane and use this to determine
the pattern radiated out into free space.

Two examples of E-plane antennas are shown in 4.5. The LTSA or Linear Taper Slot Antenna is shown on the left. This is a direct analog of a rectangular waveguide horn, tapered or flared at a fixed angle. The Vivaldi antenna shown on the right differs in that the flare angle varies gradually as we move along the antenna. In a LTSA arrangement, the gap increases linearly as we move away from the guide. In the Vivaldi, the gap increases exponentially.

E-plane guides and antennas are planar and can be manufactured by photolithography and printing processes very similar to those used to make ordinary electronic printed circuit boards. This opens up the possibility of, relatively cheaply, making large numbers of identical systems. It is also possible to produce much finer detail by photolithograph than by normal machining. Hence these systems may prove to be more useful at very high frequencies where rectangular waveguide and corrugated feeds are almost impossible to make.

Another advantage of E-plane antennas is related to the nature of TEM waves. It is a general property of the fundamental TEM mode that it does not cut off at any signal frequency above zero. We can exploit this to help produce antenna systems which work well over a very wide frequency range.

Consider what would happen if we made an antenna which had a very large open end and tapered smoothly down into an E-plane guide which had a very narrow gap. Over a fairly large frequency range the signal wavelength would remain large compared to the gap and small compared to the opening. Hence the behaviour of the antenna will depend more upon the taper angle or shape than the sizes of the ends. As the input guide gap can be made quite small without having to worry about low frequencies being cut off it becomes possible to make antennas which work well over wide frequency ranges.

## 4.5 Far-Field Properties of a Cassegrain system.

From the field, $\Psi$, produced in a plane nearby an antenna we can obtain appropriate values of beam waist size, location, and the mode composition of the beam. Using the methods outlined in this, and previous, chapters we can then calculate the field pattern at any plane from the antenna out to the far field (i.e. where $z$ approaches $\infty$).

As we have already discovered, the behaviour of an antenna is to a large extent determined by a simple measurement of the antenna or beam solid angle. Given the beam waist size, $\omega_0$, we can immediately calculate the angular extent of the antenna beam in the far field. From expression 1.13 the beam size, $\omega$, at a distance, $z$, from the beam waist will be such that

$$
\omega^2 = \omega_0^2 \left[ 1 + \left( \frac{\lambda z}{\pi \omega_0^2} \right)^2 \right] 
$$

Hence, when $z$ is very large,

$$
\omega \approx \frac{\lambda z}{\pi \omega_0} 
$$

and we can therefore define a beam divergence angle, $\theta_0$, in terms of the differential

$$
\theta_0 \equiv \frac{\delta \omega}{\delta z} = \frac{\lambda}{\pi \omega_0} 
$$
This angle may then be taken as a measure of the effective antenna beam’s angular size in the far field.

The antenna beam’s *solid* angle for a beam of circular symmetry, given by \( \Omega_0 = \pi \theta_0^2 \), will then be equal to

\[
\Omega_0 = \frac{\lambda^2}{\pi \omega_0^2}
\]  

which has a similar form to expression 4.12 linking an antennas beam angle with its effective aperture area. A circular disc of radius, \( \omega_0 \), would, in fact have an area, \( \pi \omega_0^2 \), hence it is convenient to identify \( \omega_0 \) as the radius of the antenna effective area.

For a pure fundamental mode beam the field amplitude falls to 1/\( e \) of its axial value when we are offset by \( \omega \) from the beam axis. i.e. the beam solid angle, \( \Omega_0 \), represents the beam size to the 1/\( e^2 \) power positions. Choosing a different measure for the beam size (e.g. the half-power size) would have produced a different measure of the far field beam solid angle.

When discussing feeds intended to produce Gaussian beams the beam size is often given as an angle measured to the angles where the *power* falls to 1/\( e \) of its level on-axis. Similarly, feeds are sometimes described in terms of an ‘F-number’. This usage sometimes differs from the more common one of the ‘F-number’ of a lens or mirror.

The 1/\( e \) power width of a fundamental Gaussian is \( \omega / \sqrt{2} \). Hence the far field *power* half-angle, \( \theta_p \), will be

\[
\theta_p = \frac{\lambda}{\sqrt{2} \pi \omega_0}
\]  

For such a horn, the *F-number*, \( F \), is often defined as the ratio of the total width, 2\( \omega / \sqrt{2} \), of the beam between the 1/\( e \) power points in the far field to the distance from the beam waist. i.e.

\[
F = \frac{z}{\sqrt{2} \omega}
\]  

or

\[
F = \frac{2}{\theta_p}
\]

The properties of a given feed may be defined in terms of any of the above quantities. Provided we keep clearly in mind which method is being used, any of the measures mentioned above may be employed. In practice the usual approach is to base descriptions or discussions of antenna beamwidths, etc, on the 1/\( e \) power. This is despite the bulk of the Gaussian Beam Mode expressions being described in terms of the field distribution.
Figure 4.6 represents a Cassegrain antenna system. The beam produced by such a system can be described in terms of the field pattern produced in an appropriate aperture plane.

For the sake of simplicity we may choose either the primary or secondary mirror locations as the nominal aperture plane. The secondary mirror can then be treated as a ‘centre-stop’ which blocks the middle portion of the aperture field. The primary size can then be treated as an aperture which truncates the outer boundary of the field pattern. The aperture field, $\Psi$, can then be described in terms of a appropriate Gaussian mode set.

In reality the primary and secondary mirrors will be curved and hence have finite depth, they will also be in distinct planes. However, in general, the effective beam waist size, $\omega_0$, at the primary will be sufficiently large that $\lambda Z / \pi \omega_0^2 \ll 1$ for any of these depths, $Z$. Hence the assumption of a single aperture plane is a good approximation for most purposes.

For the sake of example we can consider the case where the antenna is coupled to a feed which produces a fundamental mode Gaussian field pattern at the aperture plane which has a uniform phase distribution - i.e. the aperture is at the waist plane of the projected beam and thus $R = \infty$ at the aperture. For a circularly symmetric antenna system the resulting aperture field will be of the form

$$\Psi = \frac{1}{\omega_0} L_0 \left( \frac{2 r^2}{\omega_0^2} \right) \exp \left( - \frac{r^2}{\omega_0^2} \right) \quad \text{for} \quad s < r < a$$

and zero elsewhere on the plane, where $L_0$ is the zero'th Laguerre polynomial as defined in appendix 1, $a$ is the radius of the primary aperture, and $s$ is the radius of the secondary (centre-stop).

The zero'th polynomial has the simple form, $L_0 = \sqrt{2/\pi}$, hence the field within the unstopped part of the aperture is

$$\Psi = \frac{1}{\omega_0} \sqrt{\frac{2}{\pi}} \exp \left( - \frac{r^2}{\omega_0^2} \right)$$

The power coupling efficiency with which the field may be coupled will therefore be

$$N(a, s) = |\langle \Psi | \Psi \rangle|^2$$
where \( \{ \Psi | \Psi \} \) represents the coupling integral over the area bounded by the conditions \( s < r < a \).

For a fundamental Gaussian field the result of this integral is

\[
N(a, s) = \exp \left( -\frac{2s^2}{\omega_0^2} \right) - \exp \left( -\frac{2a^2}{\omega_0^2} \right)
\] ... (4.34)

The value represents the efficiency with which the aperture is exploited by the feed and is therefore often referred to as an *Aperture Efficiency*.

The beam pattern, \( \Psi' \) which corresponds to the truncated aperture field will be of the form

\[
\Psi' = \sum_p C_p \psi_p
\] ... (4.35)

where \( \psi_p \) is the \( p \)'th Gauss-Laguerre mode

\[
\psi_p = \frac{1}{\omega} L_p \left( \frac{2r^2}{\omega^2} \right) \exp \left[ -j(kz - 2\pi ft) - j\Phi_p - r^2 \left( \frac{1}{\omega^2} + \frac{jk}{2R} \right) \right]
\] ... (4.36)

The coefficients, \( C_p \), may be derived from the coupling integral

\[
C_p = \left\{ \Psi | \psi_p \right\}
\] ... (4.37)

Since the aperture field is assumed to have a uniform phase distribution it is sensible to assume that the \( \psi_p \) mode set also has its beam waist located at the aperture plane. For the sake of simplicity we may also choose a beam waist size equal to \( \omega_0 \). Hence we can write that

\[
C_p = \int_s^a \frac{1}{\omega_0^2} \sqrt{\frac{2}{\pi}} L_p \left( \frac{2r^2}{\omega_0^2} \right) \exp \left( -\frac{2r^2}{\omega_0^2} \right) 2\pi r \, dr
\] ... (4.38)

By consulting a suitable book of standard integrals and the properties of polynomials this can be shown to be equivalent to

\[
C_p = \sum_{m=0}^p \frac{(-1)^m p!}{(m!)^2 (p - m)!} \int_s^a X^m \exp(-X) \, dX
\] ... (4.39)

where

\[
A \equiv \frac{a^2}{2\omega_0^2}; \quad S \equiv \frac{s^2}{2\omega_0^2}
\] ... (4.40)

The solution of this integral leads to the result

\[
C_p = \sum_{m=0}^p \frac{(-1)^m p!}{(m!)^2 (p - m)!} \zeta
\] ... (4.41)

where

\[
\zeta = m! \left[ \left( 1 + S + \frac{S^2}{2!} + \ldots + \frac{S^m}{m!} \right) \exp(-S) - \left( 1 + A + \frac{A^2}{2!} + \ldots + \frac{A^m}{m!} \right) \exp(-A) \right]
\] ... (4.42)

Using the above expressions we can now evaluate the coefficients, \( C_p \), and hence define the field amplitude and phase at any point along the beam produced by the feed - Cassegrain system.

In many cases we are mainly interested in the antenna’s far-field pattern - i.e. the field when \( z \) approaches infinity. This field pattern will be of the form consisting of a series of modes
\[ \phi(\theta) = C_p L_p \left( \frac{2\theta^2}{\theta_0^2} \right) (-1)^p \exp \left( \frac{-\theta^2}{\theta_0^2} \right) \]...

where \( \theta_0 \) is the antenna beam’s divergence angle as defined by expression 4.26. The term, \((-1)^p\), appears in this expression because the \( p \)'th mode undergoes an anomalous phase shift of \( p\pi \) between the beam waist (antenna aperture) and far field.

The above expression defines the field at an angle, \( \theta \), to the beam axis and has been normalised such that an untruncated and unstopped fundamental Gaussian mode whose waist size is \( \omega_0 \) would provide an axial field of unity.

![Graph](image)

Fig. 4.7 Calculated far-field pattern of a typical Cassegrain system. The power and angle are normalised as described in the text. For the example shown, \( a / \omega_0 = 1.5 \); \( s / \omega_0 = 0.2 \) and the aperture efficiency is 0.91. The axial gain is -1.03 dB. A fundamental Gaussian beam profile is shown (broken line) for comparison.

Although we are generally interested in a far field pattern, it is worth noticing that this form of Gaussian beam mode analysis provides us with a simple analytic technique for calculating the field at any plane in the near, mid, or far field regions. Indeed the same approach can also be used to calculate field at any plane along the beam once the field is defined at just one plane. This can sometimes proves useful for, say, evaluating the aperture field from a measurement of the far field (or mid/near field) pattern of an antenna.

The secondary mirror of a Cassegrain antenna reflects signal power back towards the antenna feed. If we consider the situation where the antenna is being employed to transmit a beam, some of the power radiated by the feed will pass back through the center opening in the primary and re-enters the feed. As a consequence, the Cassegrain system will appear as an unmatched load to the transmission line which supplies power to the feed.

This process can be analysed by thinking of the feed as being able to observe an image of itself in the secondary mirror. The feed beam pattern at the secondary mirror can be represented by a beam, \( \psi \). The reflected field - which appears to come from the feed’s image - can be represented by a beam, \( \varphi \). For a circularly symmetric secondary mirror of diameter, \( s \), the field coupling between these two beams will be...
Taking the simplest case, where we approximate these to fundamental Gaussian modes, we can therefore say that for a secondary mirror of radius, $s$,
\[
C_s = \int_0^s \left( \frac{1}{\omega} \right)^2 \frac{2}{\pi} \exp \left[ -\frac{r^2}{\omega^2} \left( \frac{2R}{2R} - \frac{jk}{2R'} \right) \right] \quad ... (4.45)
\]
where $\omega$ is the beam size at the secondary mirror plane, $l = 2\pi/\lambda$, $R$ and $R'$ are the phasefront radii of the input and reflected beams at the secondary.

This may be rearranged into the form of a standard complex integral, leading to the result
\[
C_s = 2 \frac{1 - \exp[-s^2(\alpha + j\beta)]}{\omega^2(\alpha + j\beta)} \quad ... (4.46)
\]

Knowing the radiation wavelength, $\lambda$, secondary radius, $s$, and the beam parameters, $\omega$, $R$, and $R'$, we can therefore calculate the effective retro-reflectivity of a Cassegrain system produced by reflection at the secondary mirror.

Although the above example considers the antenna/feed as a transmitting system, this reflectivity is also important in receiving systems. The feed itself (or some other items in the receiver system) may also be mismatched, giving rise to a second partial signal reflection.

For the sake of illustration, consider the situation where the feed employed has a signal reflectivity, $\rho$, and is at a distance, $Z$, from the secondary mirror. This system now behaves as if it were a resonator, formed from two semi-reflectors whose reflectivities are $C_s$ and $\rho$, spaced a distance, $Z$, apart. The total signal power lost by reflection will now vary periodically as a function of $\lambda/Z$. In most practical cases $Z \gg \lambda$, hence quite small changes in signal wavelength may produce significant changes in signal coupling. For this reason, Cassegrain reflectivity problems may give rise to an unwanted frequency-dependent fluctuations in sensitivity in receiving or transmitting systems.
Chapter 5
Transmission lines, impedance matching, and signal reflections.

5.1 Transmission lines and terminating loads

Rectangular waveguide, E-plane guide, and even Gaussian Beams in free space, may all be considered as particular types of Transmission Line. A Transmission line will have a uniform, or periodic, structure which guides electromagnetic waves in a given direction. The details of the structure vary from type to type - metal pipes, dielectric fibres, chains of lenses or mirrors, etc - but they all share a set of basic properties. For example, the field patterns in each case may be treated as modes propagating in a particular direction.

When dealing with very low signal frequencies, electronic systems can often be analysed by assuming that voltages and currents instantly move along connecting wires. At high frequencies, however, we often need to communicate signals between points which are many wavelengths apart - i.e. the time taken for signals to be transferred is long compared with the period of a cycle of the signal frequency. Under these conditions we can understand what happens by regarding the signals as being communicated via a transmission line.

![Diagram of Twin Feeder transmission line]

**Fig. 5.1 ‘Twin feeder’ transmitting an electromagnetic signal**

Figure 5.1 illustrates a typical sort of transmission line called Twin Feeder (or Lecher Line). It consists of two parallel metal wires. The basic properties of the arrangement can be seen by considering what happens if we attach an oscillator to one end of a very long piece of twin feeder. The oscillator produces a voltage,

\[ V(t) = V_0 \exp(j 2\pi f t) \]  

between the wires at the place (i.e. at the end of the wires) where it is attached.

The applied voltage produces a potential difference between the wires, hence there is an electric field between them. As the wires have a finite surface area this means that they must become charged - i.e. the generator must produce a current flow, moving charge from one wire into the other - to produce the electric field. Hence the generator also produces a magnetic field around the wires, caused by these charge movements. As a result, an electromagnetic wave is created.
and flows along the parallel wires, away from the generator.

This wave is guided by the electrons which move around in the surfaces of the two wires. If the twin feeder is very long (i.e. essentially infinite) the wave flows away and is never seen again by the generator. Signal power flows away from the generator as if the generator were connected to a resistive load. The capacitance and inductance per unit length of the wires - and hence the effective resistance of a very long run of twin feeder - will depend upon the diameters of the wires and their spacing.

The voltage between the wires at some place a distance, $x$, from the generator will be

$$V(t, x) = V_0 \exp\left[ j2\pi \left( ft - \frac{x}{\lambda} \right) \right] \quad \text{... (5.2)}$$

where $\lambda$ is the wavelength of the signal whose frequency is $f$.

The current, $I(t, x)$, on the wires will be proportional to this voltage. We can therefore say that

$$I(t, x) = \frac{V(t, x)}{Z_C} \quad \text{... (5.3)}$$

where the value, $Z_C$, is characterised by the diameters and separation of the wires. $Z_C$ is therefore often called the characteristic impedance or waveguide impedance of the transmission line.

If we assume that the metal wires have a negligible resistivity the basic properties of the transmission line can be described in terms of just two values:-

i) The value of the characteristic impedance, $Z_C$

ii) The velocity of the signal, $c = \frac{\lambda}{f}$

For many purposes we do not need to know the details of the field patterns which propagate along the line. So far as the generator is concerned, an infinite length of line looks the same as a resistor, whose resistance, $R = Z_C$ because the power injected will never return. Twin feeder and a number of other forms of transmission line, including free space beams and coaxial cable, share the ability to transmit TEM waves which are not cut-off at a specific low frequency. Other forms of transmission line, rectangular waveguide for example, do not have this property. Despite this, all of these systems can be treated adequately as transmission lines for most purposes.

From the assumption that the wires themselves do not dissipate any signal power it follows that any signal put in by the generator should then move along an ideal transmission line without altering in any way. The signal which arrives at an ‘output’ end of a long line is identical to that which was put into the ‘input’ end by the generator some time ago. As a result, if we attach a resistor of value, $R = Z_C$, to the far end of the line it sees a ‘delayed’ version of the signal produced by the generator. (For this reason, transmission lines are sometimes used as ‘delay lines’ to produce time-delayed versions of a signal.)

Unless the resistor is provided with some sort of clock it has no way of knowing how far away it is from the generator. The voltage between the wire ends is $V$ and, obeying Ohm’s Law, the resistor passes a current, $I = \frac{V}{Z_C}$. This means that the generator cannot tell the difference between an infinite line of impedance, $Z_C$, and a shorter line terminated in a resistor, $R = Z_C$.

A resistor of value $R = Z_C$ is said to be matched to the transmission line. All of the signal
power transmitted along the line from the generator is coupled perfectly into such a resistor. If, however, the resistor is replaced by an impedance, \( Z_L \) which does not equal \( Z_C \) then the arrangement is said to be \textit{mis-matched}.

Consider the example of a termination \( Z_L = 0 \). - i.e. the far end of the line is terminated by a short-circuit. The wave propagating along the line will produce a voltage, \( V(t, x) \) and current, \( I(t, x) \). However, we cannot produce any voltage across a short circuit. Hence the voltage at the termination must always be zero, no matter how large the current. The wave will, however, carry a signal power of \( P = IV \). None of this power can be dissipated in the short-circuit because this would require us to produce a non-zero voltage across zero impedance. The power cannot just vanish quietly, hence it takes the only way out and flows back along the line. All of the signal power coupled via a transmission line onto a short-circuit is \textit{reflected} back along the line.

The line now carries two waves, propagating in opposite directions. One flows from the generator to the termination. The other flows back from the termination to the generator. The voltage and current we measure at any place along the line is composed of the sum of these two waves.

More generally, when a signal propagating along a transmission line is incident upon an unmatched termination some fraction of the signal power is absorbed by the termination and the rest is reflected back along the line. The fractional power \textit{reflectivity} will depend upon how the termination impedance, \( Z_L \), compares with the characteristic impedance, \( Z_C \). An incident signal having a voltage,

\[
V_i = V_0 \exp\left[j 2\pi \left(f t - \frac{x}{\lambda}\right)\right] \quad \ldots (5.4)
\]

will generate a reflection

\[
V_r = \rho V_0 \exp\left[j 2\pi \left(f t - \frac{x}{\lambda}\right)\right] \quad \ldots (5.5)
\]

where \( \rho \) is a \textit{reflection coefficient} which determines the relative magnitude and phase of the reflected wave. Each wave will have an associated current

\[
I_i = \frac{V_i}{Z_C} \quad ; \quad I_r = -\frac{V_r}{Z_C} \quad \ldots (5.6; 5.7)
\]

The difference in the sign of \( x/\lambda \) between the expressions 5.4 and 5.5 indicates that the waves are moving in opposite directions. The opposite signs for the currents in expressions 5.6 and 5.7 may be thought of as a consequence of the requirement that one wave carries power in one direction, from the generator to the termination and the other carries it away from the termination back towards the generator. (The above also means that the power values obtained by multiplying the voltage and current have different signs, indicating different directions of power flow.)

The voltage, \( V_L \), across, and the current, \( I_L \), through a terminating impedance, \( Z_L \), placed a distance, \( x = d \), away from the generator will hence be

\[
V_L = V_i + V_r \quad ; \quad I_L = I_i + I_r \quad \ldots (5.8; 5.9)
\]

where it must be the case that

\[
V_L = Z_L I_L \quad \ldots (5.10)
\]

combining and re-arranging the above expressions, we obtain the result

\[
\rho = \exp\left(-2j d \beta\right) \frac{Z_L - Z_C}{Z_L + Z_C} \quad \ldots (5.11)
\]

where \( \beta \) is the transmission line’s \textit{propagation constant},
\[ \beta = \frac{2\pi}{\lambda} \quad \ldots (5.12) \]

where \( \rho' \) represents the voltage reflection seen ‘at the generator’. i.e. It tells us the amplitude and phase of the reflected voltage reaching the generator relative to its own output. For an ideal loss-free transmission line the length of the line alters the phase of the returned reflection as this depends on the time taken for the signal to travel along the line and back. But the magnitude just depends on how \( Z_L \) differs from \( Z_C \).

It is useful also to define the load reflectivity, \( \Gamma \), which is equal to \( \rho \) when \( d = 0 \). i.e., we may write that

\[ \Gamma = \frac{Z_L - Z_C}{Z_L + Z_C} \quad \ldots (5.13) \]

and then

\[ \rho = \Gamma \exp(-2j\beta) \quad \ldots (5.14) \]

to represent the reflectivity experienced by the generator at the ‘input end’ of the line as a result of the load reflectivity, \( \Gamma \), at the far end of the line.

A matched load corresponds to the condition \( Z_L = Z_C \) and produces the results, \( \Gamma = 0 \) and thus \( \rho = 0 \); i.e. no power is reflected back towards the generator. We have already observed that terminating the line in a short causes all of the power to be reflected. The power reflectivity will vary as \( |\Gamma|^2 \). For a short, \( Z_L = 0 \), the load reflectivity, \( \Gamma = -1 \), i.e. the reflected and incident waves are 180 degrees out of phase at the load (short). If the termination impedance, \( Z_L \) is infinite (an ‘open’ circuit through which no current may flow) then \( \Gamma = 1 \) and, again, all the input signal power is reflected. However in this case the incident and reflected signal voltages are in-phase at the terminal load.

The effect of a given termination load and length of transmission line is often treated as being equivalent to an appropriate impedance connected directly to the generator terminals. The effective impedance, \( Z \), seen by the generator can be obtained from

\[ Z = \frac{V}{I} \quad \ldots (5.15) \]

where

\[ V = V_0 \exp(2\pi jft) + \rho V_0 \exp(2\pi jft) \quad \ldots (5.16) \]
\[ I = \frac{V_0}{Z_C} \exp(2\pi jft) - \frac{V_0}{Z_C} \rho \exp(2\pi jft) \quad \ldots (5.17) \]

hence we can show that the result will seem to the generator as if an impedance

\[ Z = Z_C \frac{1 + \rho}{1 - \rho} \quad \ldots (5.18) \]

had been connected to its output instead of the loaded transmission line.

A case which is of particular interest is the use of a short-circuit which may be moved backwards and forwards along the transmission line. The arrangement is generally referred to as a back-short. From 5.14 the reflection coefficient, \( \rho \), produced by such a movable short will be

\[ \rho = -\exp(-2j\beta) \quad \ldots (5.19) \]

and the effective load impedance presented to the generator will be

\[ Z = -jZ_C \tan(\beta d) \quad \ldots (5.20) \]
This result has two important features: - Firstly, it shows that the effective impedance produced by a back-short is purely reactive. (This can be expected because a short has zero resistance so cannot dissipate any power.) Secondly, it indicates that we may arrange to produce any effective load reactance from \(+j\infty\) to \(-j\infty\) by choosing an appropriate value for the distance, \(d\), to the location of the back short.

Thus far we have only considered a load which terminates a length of transmission line. In practice many devices act as load connected across a continuing transmission line at some place other than the end. Such a load behaves like a ‘shunt’, allowing some current to flow between the wires and producing a reflected wave. However, unlike a termination, some signal power may be transmitted past the load along the remaining length of transmission line. A similar situation arises if one length of transmission line is coupled to another which has a different characteristic impedance. In either case, some power may be reflected and some transmitted.

Figure 5.2(a) illustrates an ‘input’ length of transmission line of characteristic impedance, \(Z_c\), connected to a load, \(Z_L\), via an intermediate length, \(d\), of line whose characteristic impedance is \(Z_c'\). A signal may be injected into the system by connecting a generator to the input end of \(Z_c\).

The combination of the terminating impedance and the interconnecting line will behave as if an effective impedance, \(Z\), were connected to the the end of the input line. Combining expressions 5.11 and 5.18 the effective impedance terminating the input line will be

\[
Z = Z_c \frac{1 + \Gamma \exp(-2jd\beta)}{1 - \Gamma \exp(-2jd\beta)} \quad \ldots (5.21)
\]

where

\[
\Gamma = \frac{Z_L - Z_c'}{Z_L + Z_c'} \quad \ldots (5.22)
\]

If we can alter the length, \(d\), it is sometimes possible to arrange that \(Z = Z_c\) even though neither \(Z_c\) nor \(Z_L\) are themselves equal to \(Z_c\). In such cases the intermediate line acts as a transformer which matches the terminal load to the input transmission line. All the input signal power is then coupled into the load without any reflection loss.

5.2(b) illustrates a length of line which is terminated in an impedance, \(Z_T\), but which also has a shunt impedance, \(Z_s\), placed across it at a distance, \(d\), from the termination. In general, both \(Z_T\) and \(Z_s\) will generate reflected waves but - as in the case of a terminated length of a different line
- it is sometimes possible to arrange that these reflections combine to produce a total reflectivity of zero.

At the place where \( Z_s \) is connected, the combination of the termination, \( Z_T \), and the line of length, \( d \), will produce an effective impedance, \( Z_E \), where

\[
Z_E = Z_C \frac{1 + \Gamma \exp(-2jd\beta)}{1 - \Gamma \exp(-2jd\beta)}
\]

where

\[
\Gamma = \frac{Z_T - Z_C}{Z_T + Z_C}
\]

The total impedance, \( Z \), which appears to be placed at this position will be such that

\[
\frac{1}{Z} = \frac{1}{Z_S} + \frac{1}{Z_E}
\]

The overall reflectivity of the system will then be zero if \( Z = Z_C \) - i.e. if we can arrange that

\[
\frac{1}{Z_C} = \frac{1}{Z_S} + \frac{1}{Z_E}
\]

This example differs slightly from the previous one in that the input power may be shared between \( Z_T \) and \( Z_s \). In both cases, however, the system has been matched to the input guide and no power is lost by reflection back to the signal generator or source. In many practical systems, the arrangements discussed above are used to ensure optimised signal coupling into a device. When dealing with coherent signals, these arguments are also valid when the wave direction is reversed. Hence power coupling out of a source of a given impedance may also be optimised in an identical manner.

One of the simplest, and most widely used, methods for matching power into (or out of) a device is to place the device as a shunt at a given distance from a short which terminates a transmission line. For a device of impedance \( Z = X + jY \) the optimum transmission line impedance will be \( Z_C = X \left[ 1 + \left( Y^4 + X^2Y^2 \right)/\left( X^4 + X^2Y^2 \right) \right] \). The short is then best placed at a distance, \( d \), away from the device such that

\[
X^2/Y + Y = Z_C \tan(\beta d)
\]

If this can be arranged the combination of the short and the device is then matched perfectly to the transmission line and no signal power is reflected. Because the short has zero resistance this perfectly matches power into the device and gives ideal coupling.

This simple arrangement does have some drawbacks; for example, the choice of a given short distance, \( d \), ‘tunes’ the arrangement so that perfect matching only occurs at specific \( \beta \) values. If the signal frequency alters significantly it becomes necessary to move the short to re-match signal power into the device. Problems can also arise for some devices where the real part, \( X \), of the device impedance is very large or very small. This makes it difficult to produce a satisfactory transmission line with the required characteristic impedance.

### 5.2 Semi-reflectors, dielectric slabs, and mismatched lines

More complex arrangements of shunts, changes in line impedance, and backshorts may be used to obtain improvements in particular cases. By choosing an arrangement which has a high signal reflectivity at a signal frequency it becomes possible to make filters which prevent unwanted signals from being coupled into a device. Hence we can make systems which act as matching...
transformers and/or signal filters. In each case the properties of the system can be examined in terms of their signal reflection, transmission, and absorption effects.

Free space devices, acting upon Gaussian Beams, often behave as *semi-reflectors*. Various types of device exist and these will be considered in more detail in later chapters. These devices may also be treated as shunt or termination impedances coupled into a transmission line which represents the free-space beam.

When dealing with low signal frequencies we can often determine impedances by making voltage and current measurements. At millimetre-wave frequencies, and when using free-space beams, this is impractical. Instead, the behaviour of a device is determined by measuring its reflection, transmission, and absorption properties.

Consider a device whose impedance is $Z_S$ which is placed as a shunt across a line of characteristic impedance, $Z_C$. This may be taken to represent a device placed in a Gaussian Beam which will reflect, transmit, and absorb a given amount of the incident signal power.

So far as the signal incident upon the device is concerned the combination of the device and the continuing line will behave as if they were a termination whose impedance, $Z_T$, is such that

$$\frac{1}{Z_T} = \frac{1}{Z_C} + \frac{1}{Z_S} \quad \cdots (5.28)$$

In free space the impedance is defined in terms of the ratio of the electric and magnetic fields. Here we can regard the signal ‘voltage’ as related to the magnitude of the beam’s electric field distribution. The transmission line is actually propagation though free space, so we can take the value of $Z_C$ to be the impedance of free space.

The device’s field reflectivity will be

$$\Gamma = \frac{Z_L - Z_C}{Z_L + Z_C} \quad \cdots (5.29)$$

If the input signal incident upon the device is

$$V_i = V_0 \exp(2\pi j ft) \quad \cdots (5.30)$$

then the total signal magnitude at the device will be

$$V = V_i + V_r \quad \cdots (5.31)$$

where the reflected signal, $V_r$, is

$$V_r = \Gamma V_i \quad \cdots (5.32)$$

Some signal power will usually be transmitted past the device and, viewed from the ‘output’ side of the line, the device will appear as a source which transmits a signal, $V_t = V$, along the rest of the line. i.e. continuing along the beam in free space. So we can say that

$$V_t = V_i + \Gamma V_i = \left(1 + \frac{1}{\Gamma}\right) V_r \quad \cdots (5.33)$$

Hence the impedance of such a semi-reflecting device can be obtained by measuring the reflected and transmitted signals produced by a known input.

Many practical devices act as semi-reflectors without absorbing significant amounts of signal power. These devices can be regarded as purely reactive loads (i.e. having no real resistance). For such a device the impedance can be written as
J. C. G. Lesurf — 61 — Millimetre-wave Optics, etc

\[ Z_s = jX \]  \hspace{1cm} \text{... (5.34)}

this produces an apparent impedance as seen by the incident signal of

\[ Z_L = \frac{jXZ_C}{jX + Z_C} \]  \hspace{1cm} \text{... (5.35)}

Hence

\[ \Gamma = \frac{Z_C}{2jX + Z_C} \]  \hspace{1cm} \text{... (5.36)}

from which it follows that the transmitted and reflected fields are related by the expression

\[ V_t = -\left(\frac{2jX}{Z_C}\right)V_r \]  \hspace{1cm} \text{... (5.37)}

From this expression it can be seen that, for any non-absorbing semi-reflecting device, the transmitted and reflected fields are 90 degrees out of phase at the device. This result is a general one. It does not depend upon the nature of the device or the transmission line. It also has important consequences in arrangements which employ a number of semi-reflecting elements to achieve specific effects (e.g. in Fabry-Perot filters).

This result may also be seen as a consequence of the principle of energy conservation and the requirement that the field must be continuous and single-valued. i.e. The input and output fields must be such that

\[ V_i + V_r = V_t \]  \hspace{1cm} \text{... (5.38)}

and, the principle of energy conservation requires that

\[ |V_r|^2 + |V_t|^2 = |V_i|^2 \]  \hspace{1cm} \text{... (5.39)}

Unless the fields are all zero these two expressions can only simultaneously be true if we allow the voltages to be vectors and arrange that \( V_r \) and \( V_t \) are orthogonal. In the arguments presented above the vector nature of these fields shows itself as each field having real and imaginary components and \( V_t \) and \( V_r \) are at 90 degrees (i.e. orthogonal) in time.

Many of the systems and devices considered in later chapters reflect and transmit signals by exploiting polarisation effects. In these cases the reflected and transmitted signals may be orthogonally polarised in \emph{space} (e.g. plane waves polarised at right angles to each other). The fields may then be in-phase in time whilst continuing to satisfy the above requirements. However, for any loss-free semi-reflecter, it remains a basic requirement that the reflected and transmitted signals must be orthogonal - either in time or in space.

For a wave propagating in free space the ratio of the Electric to Magnetic fields has the value 377 Ohms. This resistance is the \emph{Impedance of free space} and may be used for \( Z_C \) when a free space beam is treated as a wave propagating along a transmission line. A wave moving inside a dielectric medium may have a E/H field ratio which differs from 377 Ohms. This change in E/H being described in general optics in terms of an appropriate \emph{relative permittivity}, \( \varepsilon_r \) (or a refractive index, \( \mu \)).

When considering a beam moving through a dielectric it is often convenient to regard the beam as propagating along a transmission line whose characteristic impedance differs from that of free space, i.e. just as we can regard free space as a ‘transmission line’ with a characteristic impedance, \( Z_C \), we can regard a dielectric medium as a transmission line with a different impedance, \( Z'_C \). The signal reflection which occurs at a dielectric interface can then be treated as being due to surface behaving as a junction between two transmission lines of different characteristic impedances.
The signal reflectivity, \( \Gamma \), of a junction between two lines of differing impedance, \( Z_C \) and \( Z_C' \) is

\[
\Gamma = \frac{Z_C' - Z_C}{Z_C' + Z_C}
\]

... (5.40)

for a signal arriving at the boundary along the \( Z_C \) line).

If both \( Z_C \) and \( Z_C' \) are real, \( \Gamma \) is also real - i.e. the reflected signal is either in-phase or 180 degrees out of phase with the incident wave at the junction.

If \( V_i \) represents the incident field then the total field at the junction will be

\[
V_T = V_i (1 + \Gamma)
\]

... (5.41)

The field, \( V_i \), acts as a source, transmitting a wave into and along the line whose impedance is \( Z_C' \). \( V_i \) is the transmitted signal at the junction (dielectric surface). If we define the transmission coefficient, \( \tau \), such that

\[
V_i = \tau V_i
\]

... (5.42)

then we can write that

\[
\tau = 1 + \Gamma
\]

... (5.43)

The incident signal power is \( |V_i|^2 / Z_C \) and the reflected signal power is \( |\Gamma V_i|^2 / Z_C \). Because no power is absorbed at the junction we can say that

\[
\frac{\tau^2}{Z_C} = \frac{1}{Z_C} - \frac{\Gamma^2}{Z_C}
\]

... (5.44)

Expressions 5.43 and 5.44 lead to the result

\[
\tau = \frac{2Z_C'}{Z_C + Z_C'}
\]

... (5.45)

As \( \Gamma \) and \( \tau \) are real the transmitted signal, \( V_T \), and the reflected signal, \( V_R \), must both be either in-phase or 180 degrees out of phase with the incident signal, \( V_i \), at the junction.

The reflection properties of a dielectric interface and of a parallel sided slab have already been discussed in chapter 2. There it was pointed out that the field reflectivity at an interface where an incident signal passes from a medium of refractive index \( \mu_1 \) to another of index \( \mu_2 \) will be

\[
\frac{\mu_2 - \mu_1}{\mu_1 + \mu_2}
\]

... (5.46)

hence we can identify the refractive index of a low-loss dielectric as being equivalent to the relative (to free space) characteristic impedance of the medium when considered as a transmission line. Each surface of a dielectric slab can then be treated as a junction between two transmission lines of differing impedance. The overall power transmission and reflection can then be calculated using a transmission line approach.

In general, systems involving a series of semi-reflecting and dielectric elements may also be analysed by using transmission line arguments to calculate the effective load (and hence reflectivity) presented to the input signal. These methods can be used for both conventional waveguide and free space beam systems to design a wide range of instruments and devices. Some examples of this technique will be mentioned in later chapters.
6.1 Basic properties of thermal detectors

Many different sorts of millimetre-wave detectors have been developed, each offering its own combination of advantages and drawbacks. For convenience, they may be divided into two general categories; Bolometers (heat detectors), and Mixers (heterodyne detectors). In this chapter we will consider bolometric signal detection systems.

A bolometer is a device which responds to a change in temperature produced when it absorbs incident radiation. Most bolometers are relatively unfussy regarding the frequency of incident radiation. Incident power of whatever signal frequency is absorbed, the element temperature changes, and this temperature change is sensed. Bolometers are, therefore, particularly useful when we are mainly concerned with measuring the input signal power. If we need to determine the power/frequency spectrum of the incoming signals we can employ a suitable filter which only allows given signal frequencies to fall upon the bolometer. Systems such as the Polarising Interferometer discussed in a later chapter may be regarded as a form of variable filter which permit us to explore the power/frequency spectrum of a signal without having to replace one filter with another.

The behaviour of a bolometer are basically dependent upon its thermal properties. In most cases any change in bolometer temperature is sensed electronically. For example, a bolometer may have an electrical resistance which varies with temperature. Consider the idealised bolometer illustrated in Figure 6.1.

![Fig 6.1 Thermal (left) and electronic (right) models of a bolometric detector](image)

A typical system consists of a heat sensitive detector element mounted inside a heatsink which almost surrounds the detector. Input signal power may be coupled onto the detector through an opening in the immediate surroundings. A 'window' is usually placed over the opening. This can be used to prevent power at unwanted signal frequencies from reaching the detector. The detector element is physically supported by an arrangement designed to provide a controlled heat conduction path between it and the surrounding heatsink. Wires will also be connected to the element so that electronic changes produced by input signals may be sensed.
The bolometer is assumed to have a thermal capacity, $C$, and is connected to a heatsink of temperature, $T_0$, via a thermal conductivity, $s$. For a real bolometer heat may be coupled into or out either by conduction through its physical supports or by radiation exchanging power with its surroundings. Hence the value of $s$ will depend both upon conduction and radiation. However we can assume that any difference in temperature between the bolometer element and its physical surroundings are small enough to make the assumptions behind the following analysis.

The signal power, $P$, absorbed by the detector element and its temperature, $T$, will be related by the expression

$$P = C \frac{dT}{dt} + s(T - T_0) \quad \text{... (6.1)}$$

If the input signal power is applied and held at a particular level, $P = P_0$, then the element temperature will change exponentially towards the value

$$T = \frac{P_0}{s} + T_0 \quad \text{... (6.2)}$$

In most practical detection applications we are concerned with how the signal power varies with time. For an input signal power which varies with time as

$$P = P_0 + \delta P \sin(\omega t) \quad \text{... (6.3)}$$

the detector element temperature will be

$$T(t) = T' + \delta T \sin(\omega t) \quad \text{... (6.4)}$$

at the time, $t$, where

$$T' = \frac{P_0}{s} + T_0 \quad \text{... (6.5)}$$

is the mean element temperature,

$$\delta T = \frac{\delta P}{\sqrt{C^2\omega^2 + s^2}} \quad \text{... (6.6)}$$

is the magnitude of the element's temperature variation, and

$$\phi = \arctan\left(\frac{C\omega}{s}\right) \quad \text{... (6.7)}$$

is the phase lag between the input power fluctuation and the temperature variation it produces.

Any change, $\delta T'$, in the detector temperature will produce a change in the electrical properties of the device. In Figure 6.1 it is assumed that the electrical resistance of the bolometer changes with its temperature. A bias current may be passed through the bolometer so that this change in resistance will produce a change, $\delta V$, in the voltage across the element.

If a given change in input signal power, $\delta F$, produces a voltage change, $\delta V$, then the sensitivity of the bolometer used as a detector can be defined in terms of the Responsivity, $R$,,

$$R = \frac{\delta V}{\delta P} \quad \text{... (6.8)}$$

(In some cases the responsivity may be defined in terms of a current change rather than a voltage.)

The responsivity will be proportional to $\delta T' / \delta P$, hence from 6.6 we can say that

$$R \propto \frac{1}{\sqrt{C^2\omega^2 + s^2}} \quad \text{... (6.9)}$$
For $\omega \ll s / C$, this approximates to
\[ R \propto \frac{1}{s} \quad \text{... (6.10)} \]
i.e. at low input power modulation frequencies the bolometer’s responsivity essentially only depends upon the effective thermal conductivity, $s$. However when $\omega \gg s / C$,
\[ R \propto \frac{1}{C \omega} \quad \text{... (6.11)} \]
i.e. at high modulation frequencies the responsivity depends upon the bolometer’s thermal capacity and falls with increasing modulation frequency.

We can define the bolometer’s response time or time constant, $\tau$, to be
\[ \tau = \frac{C}{s} \quad \text{... (6.12)} \]
and re-write expression 6.9 as
\[ R \propto \frac{1}{s \sqrt{1 + \omega^2 \tau^2}} \quad \text{... (6.13)} \]
The phase lag, $\phi$, between the input signal modulation and the output voltage variation can also be re-written as
\[ \phi = \text{Arctan} (\tau \omega) \quad \text{... (6.14)} \]
In effect, $\tau$ indicates a measure of the time required for the detector to respond efficiently to a change in the input signal power level. In general we would like to produce a bolometer which has a high responsivity. One possible way to increase the bolometer responsivity is to reduce the conductivity, $s$. However this has the effect of increasing $\tau$, which means that - all else being equal - the detector will take longer to efficiently respond to changes in the signal power.

Reducing the thermal capacity, $C$, will improve the response time. Although this has no effect upon the responsivity at low modulation frequency it is a useful way of preventing the response time from becoming too long when the thermal conductivity is reduced. Unfortunately, reductions in thermal capacity generally involve a detector element being physically smaller. If the detector is too small it may become difficult to efficiently couple the input signal power.

The above assumes the temperature differences, etc, involved are very small and we can assume linear behaviour. However thermal energy may usually be flow in and out of a bolometer element in two ways; by conduction and by radiation. The conductivity, $s$, should therefore really be considered as the sum of two conductivities. For a constant input signal power level, $P$, the bolometer temperature, $T$, will therefore be such that
\[ P = (T - T_0) G + A\sigma (T^4 - T_0^4) \quad \text{... (6.15)} \]
where $G$ is the conductivity for thermal conduction through the detector element’s physical support, $A$ is the detector’s surface area, and $\sigma$ is Stefan’s Constant. From 6.15 we can then say that
\[ \frac{dT}{dP} = (G + 4\sigma AT_0^3)^{-1} \quad \text{... (6.16)} \]
i.e. the total thermal conductivity, $s$, will be
\[ s = G + 4\sigma AT_0^3 \quad \text{... (6.17)} \]
provided that the input signal power is small and $T$ is almost the same as the temperature, $T_0$, of the heatsink and surroundings of the bolometer element.
Expression 6.17 leads us to two conclusions. Firstly, no matter how small we make the physical conductivity, $G$, the total thermal conductivity cannot be less than $4\sigma A T_0^3$. Secondly, that when $G$ is small the overall conductivity (and hence the detector responsivity) is strongly dependent upon the temperature of the detector and its surroundings.

Assuming $G = 0$ we can calculate that the ratio of the responsivity of a given detector at 3 K to its responsivity at 300 K will (if all else is equal) be

$$\frac{R[3 \text{ K}]}{R[300 \text{ K}]} = \frac{300^3}{3^3} = 1000000 \quad \cdots (6.18)$$

This influence on the responsivity is one of the reasons why bolometric detectors are often cooled to low temperatures. In practice $G$ will not be zero and, at a suitably low temperature, $T_0$, \( G \approx 4\sigma A T_0^3 \). Reducing the temperature of the detector and its surroundings below this value will then not produce a much further improvement in responsivity. It can be seen, however, that - unless the detector cannot work well at a reduced temperature for some other reason - the responsivity can be dramatically improved by cooling.

![Typical liquid-Helium cryostat system for cooling a low-temperature bolometric detector.](image)

Figure 6.2 shows a diagram of a typical Cryostat system which allows a bolometer to work at low temperatures. The detector is mounted inside a cavity or guide which is connected to the metal baseplate of a liquid Helium vessel. This will cool the detector and its immediate surroundings to a temperature of around 4 K. By using a vacuum pump it is possible to lower the vapour pressure above the liquid Helium. In this way the detector temperature may be reduced still further.

As far as possible, unwanted thermal energy is kept away from the detector and its mount. Unwanted heat reaching the detector from its surroundings will tend to increase its temperature and degrade its performance. It also will increase the rate at which the Helium is boiled off. Apart from a series of filters which act as 'windows' designed to pass only wanted signal energy, the detector is surrounded by heat shielding. The system illustrated in 6.2 shows a Nitrogen Shield - a surrounding screen of metal cooled to around 77 K by a Liquid Nitrogen vessel.

Because of the low internal temperatures air must be excluded to prevent freezing and
condensing inside the vessel. The detector and its surroundings are therefore evacuated. One of the main functions of the outermost window is to keep air out of the system. The inner window - shown here on the Nitrogen Shield - is designed to stop radiation at unwanted frequencies from reaching the detector. A final ‘cold’ filter is often placed in the detector mount or its input optics to reduced still further the unwanted heat radiated onto the detector. These filters also prevent the detector from producing spurious output in response to signals at frequencies outside the range of interest.

Two general types of filter can be used; absorbing or reflective. An absorption filter will absorb heat radiation at unwanted signal frequencies. The emissivity and absorptivity of a material are always the same value. Hence a filter which absorbs strongly in a given frequency range will also radiate strongly as a ‘black body’ over the same range. However, the filter is much colder than the thermal sources outside the system. Hence the net effect is to greatly reduce the unwanted thermal power seen by the detector.

A reflection filter reflected unwanted input power away from the detector. At frequencies where it is highly reflective it will not emit any significant power but, instead, reflects back any radiation it receives from the detector or the inside of its mount. Hence the detector sees a cold reflection of itself in the filter.

So far we have discussed the responsivity (i.e. the sensitivity) of a bolometer. This can be regarded as a conversion efficiency which determines how large an electrical output will be produced by a given input signal. The ability of the bolometer to detect very small signals will ultimately be limited by the amount of random noise present in the system.

The presence of random noise is a general property of physical systems and is a consequence of the underlying statistical quantum mechanical nature of the real world. Electronic charge and EM radiation are both quantised. When dealing with small signals we must consider the effects of random fluctuations in the movements of electrons and photons upon any measurements we wish to make.

Fig 6.3 (a) Thermal motion of charge carriers in a resistor. (b) Flow of quantised charges along a conductor. (c) Energy exchange by photon emission and absorption.

Here we will only consider the three main types of noise which arise in bolometer systems. Figure 6.3 illustrates the physical processes which generate these types of noise.
6.3(a) shows a piece of resistive material with a pair of electrical contacts attached to its ends. Random thermal motions of the electrons inside the material cause the internal charge distribution to fluctuate unpredictably from moment to moment. The potentials of the end contacts will depend upon the nearby charge distribution, hence there is a randomly varying potential difference between the two contacts.

If we connect a voltmeter to the two contacts we observe a randomly varying voltage. The average voltage will be zero, being just as often negative as positive. The voltage’s value squared will, however, never be negative, hence the amount of noise power present may be measured in terms of the mean squared voltage. In practice, the amount of noise is often determined in terms of the time-averaged root-mean-square (rms) noise voltage or the mean noise power.

Any real measurement system will take a finite amount of time to respond to a change in voltage. Fluctuations which take place over a period shorter than this response time will therefore have no effect upon the measured noise level. The finite response time means that any measurement only takes into account noise fluctuations over a limited frequency range or bandwidth.

The amount of thermal noise generated by a resistance, \( R \), at absolute temperature, \( T \), may be represented in terms of a ‘noise voltage generator’ placed in series with the resistance which produces a mean squared voltage

\[
e_n^2 = 4kT RB
\]

where \( k \) is Boltzmann's Constant \( (k = 1.38 \times 10^{-23} \text{ Ws/K}) \), and \( B \) is the bandwidth of the measurement system.

The noise power which this produces at the input of an electronic measurement system will depend upon the system's input resistance, \( R_I \). The noise voltage, \( e_n \), sets up a current, \( i \), which must flow through \( R \) and \( R_I \) in series. Hence, from Ohm's Law

\[
i = \frac{e_n}{R + R_I}
\]

this means that the actual observed noise voltage, \( e_0 \), generated at the system input will be

\[
e_0 = i R_I = \frac{e_n R_I}{R + R_I}
\]

the mean noise power, \( N \), observed at the system’s input will therefore be

\[
N = e_0 i = \frac{e_n^2 R_I}{R + R_I}
\]

This has a maximum value if we arrange that \( R = R_I \), when we then get

\[
N \text{ (max)} = \frac{e_0^2}{4R}
\]

this value is generally referred to as the maximum available noise power.

Thermal noise is often called Johnson noise after its discoverer. It is sometimes also referred to as Nyquist noise in recognition of the person who provided the first theoretical analysis of its origin.

Another form of electronic noise arises when current flows along a conductor. Figure 6.3(b)
represents a wire along which a current is flowing. A ‘steady’ current actually consists of a stream of discrete charge carriers passing along the wire. The average current, $I_0$, flowing through a plane cutting the wire can therefore be written as

$$I_0 = \frac{q n}{t} \quad \ldots (6.24)$$

where $q$ is the magnitude of each discrete charge (i.e. charge on an electron) and $n$ is the mean number of charges (electrons) which cross the plane during the time interval $t$.

In reality the moving carriers will have a random distribution of individual velocities imposed on their overall drift velocity along the wire. (In fact this is essentially the thermal noise velocities.) The number crossing a specific plane will hence vary randomly from one short time-period, $t$, to the next. The actual current will therefore also vary randomly from one time period to another. Averaged over a particular time interval of length, $t$, the measured current, $I$, will be

$$I = I_0 + \delta I \quad \ldots (6.25)$$

where $\delta I$ is the amount by which this random noise current alters the total during the particular time interval. If we repeatedly measure such a current flow we find that the value of $\delta I$ varies from period to period with a random spread whose mean squared size is

$$\delta I^2 = 2qI_0B \quad \ldots (6.26)$$

which then acts as another source of noise when we wish to make a measurement.

This type of noise is called shot noise. As with Thermal noise the amount observed depends upon the measurement system bandwidth, $B$. This bandwidth corresponds to the range of frequencies which will produce a response in a system which averages over the time period, $t$.

A number of other physical processes generate different types of noise - for example 1/f-noise - whose magnitude depends upon the physical construction of the electronic components. Thermal and Shot noise, however, arise from the discrete nature of fundamental charge carriers and the quantum mechanical nature of the real world. Hence these other forms of excess noise may - in principle at least - be alleviated by our skill in producing electronic devices. Thermal and Shot noise cannot be reduced in this way. Instead they can only be dealt with by lowering the temperature, $T$, of a resistive element or reducing the mean current level, $I_0$.

The third form of noise which is important in bolometers is photon noise. This is illustrated in Figure 6.3(c). This arises because electromagnetic radiation is also quantised. Power is radiated back and forth between the detector element and its environment, producing a radiative contribution to the bolometer’s thermal conductivity. The quantised energies of the exchanged photons and the random statistics of the exchange mean that the net heat flow into or out of the bolometer varies randomly from instant to instant.

A similar effect also arises in the flow of heat by conduction through the physical support of the detector element. The conducted heat is also quantised into phonons. Their behaviour within the solid supports is broadly similar to that of the photons in free space.

The random fluctuations in the bolometer’s temperature will depend upon the temperature of the bolometer and its surroundings. It will also depend upon the range of photon frequencies which can be absorbed/emitted by the bolometer and the signal power level (which is also a stream of photons). The signal power received will fluctuate from instant to instant due to the random statistics of the photons which make up the signal. This effect is similar to electrical shot noise in that the noise level generated depends upon the signal level.
At very low signal power levels the photon noise becomes dominated by the thermal radiative exchange between the detector and its surroundings. It can be shown that, under these circumstances, the random fluctuations in the temperature of the bolometer element will be equivalent to a fluctuating signal power input, $\delta W$, whose mean squared size is

$$\delta W^2 = 4k T^2 G B$$

where $T$ is the temperature of the bolometer and its surroundings and $B$ is the measurement bandwidth.

The noise performance of a bolometer is often given in terms of a *Noise Equivalent Power* (NEP). This indicates the input signal power level which is equal to the mean apparent fluctuation in the input power produced by noise (usually defined for a standard bandwidth of one Hertz).

For a one Hertz bandwidth, and taking the example of a detector where the thermal conductivity, $G$, is essentially radiative the minimum NEP, ignoring electrical noise, will be

$$\text{NEP} = 4\sqrt{A\sigma_s k T^5 B}$$

where $\sigma_s$ is Stephan's Constant and $A$ the detector's absorbing area. For a detector whose area is one square centimetre the minimum NEP will therefore be $5.5 \times 10^{-11}$ Watts/Hz if $T = 300$ K. If the detector and its surroundings are cooled to 1.5 K the minimum NEP falls to $9.8 \times 10^{-17}$ Watts/Hz.

This ideal, minimum possible NEP does not take into account the effect on any signal power being absorbed by the detector. When the signal power is significantly higher than the background level radiated by the detector's immediate surroundings the photon shot noise level will also rise. This means that the signal/noise ratio observed at a particular signal power level does not simply improve in proportion with any increase in the signal power.

Noise arises in the detector element, the biassing electronics, and in the signal amplifying system. The actual NEP of a practical system will therefore always be higher than the minimum level set by photon/phonon noise. The total noise level in a practical bolometer system may be calculated by adding together the individual noise powers generated in each part of the system.

Reducing the temperature of a bolometer and its surroundings generally both improves its responsivity and lowers the NEP. The level of thermal electronic noise produced within the the bolometer element and its biassing components may also be reduced by cooling. In some cases it is also possible to cool the amplifiers used with the bolometer, hence reducing any noise they may contribute.

### 6.2 Types of bolometer

Here we will consider three examples of bolometric detectors which have been widely used to detect millimetre wave signals; the *Golay detector*, the *Bismuth bolometer*, and the *Indium-Antimonide Hot-Electron (InSb) bolometer*.

The Golay detector and Bismuth bolometer are normally used at room temperature as convenient sensors with moderate sensitivity. Hence they do not require a cryostat system. Figure 6.4 illustrates how these two sensors work.
Fig. 6.4 Three examples of room-temperature bolometric detector arrangements. (a) Golay pressure cell. (b) Bismuth resistive film. (c) Small bismuth bolometer mounted on an E-plane line and antenna arrangement

The Golay Cell consists of an absorbing film placed inside a pressure cell. Signal power absorbed by the film is transferred by conduction to the gas inside the cell. Hence the gas pressure inside the cell will alter in response to a change in the input signal power level. In most Golay detectors, changes in pressure are detected by observing the flexing of a thin cell wall. This is done by silvering the thin part of the wall and measuring the deflection of a reflected beam of light. An alternative approach is to use a small microphone to measure the change in pressure.

The main disadvantages of the Golay are its poor response time and its sensitivity to mechanical vibration (microphony). A typical Golay requires the order of a second to respond fully to a change in the input signal power level. In order to prevent the fragile cell being destroyed by changes in atmospheric pressure or ambient temperature they generally include a small vent to allow the cell pressure to come to equilibrium with the external pressure. Hence their response is transient. Hence they work best for detecting the powers levels of signals that are modulated at frequencies of around 10 to 20 Hz.

Golay detectors have been commercially available for many years and, in some quarters, their relative age is associated with obsolescence. They are, however, still useful when very high sensitivity is not required. A typical Golay detector will have a responsivity of $10^3$ to $10^6$ Volts per Watt and an NEP of around $10^{-10}$ Watts/√Hz.

Golay detectors often respond to a very wide range of signal frequencies, from the visible to the millimetre wave region. This is useful in wide-band measurements but can cause problems such as unwanted sensitivity to room lighting. Filters can help prevent these problems. The cell wall of a sensitive Golay is very fragile. A sharp impact - or even a sudden change in the signal power level - may rupture the cell wall. These problems can be avoided by handling the detector with care, filtering the input, and avoiding any sudden change in the input level.

A typical golay cell is quite small, perhaps a few millimetres across. This is because they are generally intended for use in the near infrared where this size is convenient. In order to efficiently couple millimetre wave power onto the cell an appropriate feed horn is required.

Two general types of Bismuth bolometer are used to detect millimetre-wave radiation. Figure
6.4(b) shows a thin film of Bismuth evaporated onto a dielectric sheet. Neglecting the effect of the dielectric which is present simply as a physical support, the Bismuth film behaves as a resistive sheet onto which we may direct a millimetre wave free space beam. A pair of electrical contacts at opposite edges of the film can be used to apply a small current which lets us measure the film’s resistance.

The electrical resistivity of Bismuth is quite high (10.7 $\mu\Omega$ m at 0°C) compared to most normal metals (e.g. Copper 0.15 $\mu\Omega$ m at 0°C) and its temperature coefficient (4.2 $\mu\Omega$ m/degC) is moderately high. A thin film of Bismuth can therefore be used to absorb signal power and the change in temperature sensed by measuring the resulting change in resistivity.

The behaviour of a thin metallic film placed perpendicular to a signal beam can be analysed using the transmission line ideas outlined in Chapter 5. The incident beam behaves as a signal propagating along a line whose characteristic impedance, $Z_C$, is equal to the impedance of free space. The metal film can be regarded as a shunt load whose resistance is equal to the resistance ‘per square’ of the metal film. The units of area do not, in fact matter when the resistance per unit area of a film are calculated or measured, i.e. the amount of resistance in Ohms per cm² or per inch² (or even per acre!) of a uniform film is always the same value.

The signal reflectivity, $\Gamma$, and transmissivity, $\tau$, of a film whose resistance per square is $Z$ will be

$$\Gamma = \frac{Z' - Z}{Z' + Z}$$

where

$$Z' \equiv \frac{Z Z_C}{Z + Z_C}$$

and

$$\tau = 1 + \Gamma$$

The fraction, $A$, of the incident signal power absorbed by the film will hence be

$$A = 1 - |\Gamma|^2 - |\tau|^2$$

For a resistive film, $\Gamma$ and $\tau$ are both real, hence we may write that

$$A = -2(\Gamma + \Gamma^2)$$

At first glance expression 6.33 appears to imply a negative power absorption. However, from 6.30, it can be seen that $Z' \leq Z$. It follows from this that the reflectivity must be in the range, $-1 \leq \Gamma \leq 0$ i.e. $\Gamma$ is negative and the power absorbivity, $A$, will be positive or zero. The maximum value of $A$ occurs at a film resistance per square value $Z = Z_C / 2$, which corresponds to a resistance per square of around 180 Ohms. This produces a peak power absorbivity of 0.5, i.e. half the incident power is then absorbed. (A quarter is reflected, and another quarter transmitted.)

Bismuth films of this type are moderately simple to make by vacuum evaporation onto thin dielectric sheets. Although a practical difficulty is that they are fragile and must be protected from ageing effects caused by exposure to air, they work well over a very wide signal frequency range. They also can be used to absorb signal power almost irrespective of the beam pattern directed onto the film. Hence they are useful for measuring the signal power of a beam whose pattern is not well defined.

The amount of power absorbed may be increased by providing a suitable impedance matching or reflection/transmission cancelling arrangement around the absorbing film. For example, a mirror...
may be placed behind a film and used as an optical backshort. The film and mirror then appear to the incoming signal as a terminating load instead of a shunt across a continuing transmission line. If the film resistance per square is now chosen to be equal to $Z_c$ and the mirror is a quarter-wavelength from the film the termination is matched to free space and practically all of the incident power will be absorbed. Although the improved absorptivity is useful, the addition of a mirror backshort ‘tunes’ the detector, and the signal absorptivity now has a frequency variability which depends upon the mirror-film spacing. It also affects the effective antenna pattern of the system as the matching depends on the signal beam arriving from a given direction.

A large Bismuth film detector will be somewhat less sensitive than a Golay cell if it is designed to have a similar response time. It will not be microphonic, but may need to be protected from temperature fluctuations in the surrounding air. A typical film, designed to have a response time of just under a second, will be useful for measuring power levels around a few tens of milliwatts or more.

An alternative method for using Bismuth bolometers is illustrated in fig 6.4(c). Here the Bismuth bolometer is small compared to the signal wavelength and power is coupled into it via some form of antenna and mounting circuit. In this example an E-Plane antenna is used to coupled a free space beam into an E-Plane transmission line. A small piece of Bismuth is placed across the line. Hence some of the signal power transmitted along the line will set up a current in the Bismuth and the dissipated power may be detected. A very small Bismuth element may be treated as a shunt placed across the transmission line. A larger element may need to be considered as a length of 'lossy' transmission line. In either case, the system may be designed using normal transmission line arguments. In the illustrated example a backshort is placed behind the Bismuth element to increase the power absorbtivity at a particular signal frequency.

This arrangement differs from the simple large film in that it now possess a definite antenna pattern. The responsivity may, however, be considerably greater than that of a large film. This is because the signal power is coupled into a much smaller thermal capacity and tends to produce a greater temperature change. This in turn means that a higher thermal conductivity may be used to carry heat away from the detector, reducing the response time, without producing an unacceptable degradation of the responsivity.

The two forms of Bismuth bolometer illustrate two differing approaches to device construction. The large film acts as a ‘distributed’ or ‘bulk’ system where power is beamed directly into the active material. The small element acts as a tiny ‘two-terminal’ device into which power is coupled as a current and potential between two end-contacts. One approach is typical of optical systems, the other of electronic ones. The main distinction between these approaches is whether the device is large or small compared with the signal wavelength. In the millimetre-wave region the radiation wavelength and device size are often comparable. Either approach may then be used and the underlying Physics may be seen to be the same.

Neither the Golay detector nor the Bismuth bolometers can be cooled to very low temperatures. The gas in a Golay cell would condense if cooled to liquid Helium temperatures and the resistance properties of Bismuth become unsatisfactory. The Indium Antimonide (InSb) bolometer, however, must be cooled in order to be used as a far-infrared detector. This is because its electronic properties only become satisfactory at low temperatures.

In many types of bolometer, incoming energy is primarily absorbed by electrons which are closely bound to particular atoms. The absorbed energy and momentum is therefore quickly transformed into lattice vibrations (phonons) i.e. the crystal lattice warms up. This then affects
the resistivity of the bulk of the material. However the cooled InSb bolometer operates in a different manner. Many of the charge carriers in a cooled sample of InSb are relatively weakly coupled to the crystal lattice. Incident millimetre-wave energy tends to be absorbed by these electrons rather than by valence electrons which are closely bound to individual lattice atoms. The mobility of these weakly coupled electrons depends upon their energy. Hence any energy absorbed will tend to change the electrical resistance of the InSb crystal, allowing us to use it as a bolometer.

The weakly coupled electrons are generally referred to as an *Electron Gas* which is regarded as being able to move fairly easily about inside the crystal. The input signal power raises the mean energy of these electrons just as if the crystal lattice temperature had been increased. The electrons now act as if the crystal were hotter than its actual physical temperature. This is referred to as a *Hot Electron* effect.

In the case of an InSb bolometer it is the hot electron gas which acts as the detector and the crystal lattice acts as the heatsink. The weak coupling between electrons and lattice assumes the role of a small thermal conductance between bolometer and heatsink. The thermal capacity of the electron gas is also very small. The combination of very low thermal capacity and very low conductance produce a bolometer which has both a high responsivity and a short response time.

A typical InSb bolometer, cooled to around 3 K, will have an intrinsic response time of the order of a microsecond, a responsivity of perhaps 10 kV/W, and an NEP well below $10^{-12}$ W/Hz. This sensitivity and noise performance may also be improved by cooling below 1 K. InSb detectors have also sometimes been used at higher temperatures. This degrades the responsivity and noise performance but reduces still further the response time. At temperatures above 20 K the response time falls below 0.1 microseconds.

As a result, cooled InSb bolometers have a far higher level of performance than room temperature bolometers. The price of this performance, however, is the need to operate them at very low temperatures. Whereas detectors like the Golay can easily be used at room temperatures.

### 6.3 Phase Sensitive Detection and Background Subtraction

In many cases we need to measure a very small signal power level in the presence of noise. We may also wish to measure a signal which is dominated by the presence of an unwanted level of ‘background’ power coming from other sources in which we have no interest. As an extreme example of this we can consider the problems which arise when an astronomer wishes to use a bolometer to measure the faint millimetre wave signals radiated by a cloud of gas and dust some considerable distance from the earth.

Signal power can be collected and directed onto a bolometer using an antenna system. The input signal power level collected in this way is usually very small. An antenna system on the Earth’s surface must look through our atmosphere. This will absorb some of the signal before it can reach the detector. The atmosphere in the line of sight will also often radiate far more power onto the detector than is received from the astronomical source. To make matters worse, the amount of atmospheric absorption and emission fluctuate randomly from moment to moment.

Essentially all forms of active device, detectors, amplifiers, etc, suffer from excess noise. The forms of random noise discussed earlier all have ‘white’ frequency spectra - i.e. the amount of
noise power per Hertz bandwidth does not depend upon the frequency. More formally this is described as having a uniform power spectral density. Unfortunately, one of the most common forms of excess noise is 1/f-noise. This, as its name implies, has a power spectrum which varies inversely with the fluctuation frequency. Atmospheric fluctuations also exhibit a 1/f spectrum at low frequencies. Some detectors (e.g. the Golay) are deliberately designed in a way which makes them unsuitable for measuring fixed signal levels. Even when this is not the case it is desirable to find a way of measuring fixed or slowly-changing signals which overcomes - as far as possible - these difficulties caused by an unwanted, and randomly changing, incoming background noise level.

There are two common ways of dealing with these problems.

i) We can modulate the signal, changing a steady quantity into an alternating one.

ii) We can subtract any background effects from the signal before measurement.

*Phase Sensitive Detection* (PSD) systems employ both modulation and subtraction techniques and are often used to measure small signals in the presence of noise and unwanted background effects.

![Fig. 6.5 Example of a Phase Sensitive Detection (PSD) system.](image)

Figure 6.5 illustrates a typical PSD system, being used to measure the signal power from a weak source. In this example a beam chopper is used to periodically prevent signal power from reaching the detector. A common form of chopper is a toothed wheel which is rotated so that its blades pass through the source-detector beam. The system shown in 6.5 has a reflecting chopper. i.e., when a blade is placed in the beam, the detector sees power from a comparison source. As the chopper rotates, the signal power level reaching the detector switches rhythmically between the levels produced by the source of interest and the comparison. In the arrangement illustrated another mirror is used to ensure that the comparison source is seen against a similar background to that of the signal source.

In some cases the chopper will be made of a material which strongly absorbs (and hence emits) millimetre wave radiation. The chopper blades themselves then form the comparison source and radiate an amount of power determined by their temperature. In some other cases a reflecting chopper is used without a specific comparison source. Under these circumstances the detector will receive some power from the surroundings as seen by reflection from the chopper blades. Sometime it is possible to dispense altogether with a physical chopper and simply switch the source on and off. *(Not an option for astronomers!)*
Whatever the details of the arrangement, the result is to produce a power level at the detector
which switches in a predetermined and controlled way between the signal level we wish to
measure and a specific comparison level. If the source power level is $S$, the comparison power
level, $C$, and the common background power level is $U$, then the magnitude, $A$, of the periodic
change in detector output will be

$$A = (s + u) - (c + u) = s - c \quad \ldots (6.34)$$

By chopping or switching the signal before the detector we have produced an alternating (a.c.)
output whose size depends upon the signal $S$ but not upon any general background level, $U$. In
principle this result remains true even when $S \ll U$, provided that the background level is not
so high as to overload the detector. We may now measure the size of the A.C. signal and use this
as to measure of the signal power level.

In reality, because the background noise level varies, and we look at the source and comparison
at different times, we can expect that the above cancellation of $u$ can’t be perfect. So some of
the unwanted background noise will leak through the above chopping process. But used with
care the PSD can be used to significantly reduce the effect on the measured results. In some
cases their won’t be a convenient comparison source, $c$, and the measurement of signal power
will be made from the size of $A$ and knowing the responsivity of the detector. However in some
cases the it is possible to use a known, variable, source, $c$. This can then be adjusted to try and
null out $A$. At that point the values of $s$ and $c$ can be regarded as the same. So if we know $c$ we
have determined the value of the signal level, $s$. This method is known as a nulling technique
and helps to remove a number of potential causes of error from measurements. Hence it is often
employed in precision measurements.

The nominally steady input level has been modulated at a particular chopping frequency, $f_c$.
The detector output can now be enlarged using ac amplifiers and passed through a bandpass
filter which only lets through a limited range of frequencies centred on $f_c$. Any noise
fluctuations produced by the background, detector, or the amplifiers at frequencies significantly
different to $f_c$ will be rejected by such a filter. i.e. Only noise in the frequency range passed by
the filter will affect the output measurement. By reducing the filter’s bandpass width we can
close the amount of noise which passes through the filter. This also means that if we can choose
the chopping frequency which is above the region where $1/f$ noise is a problem we can stop any of
this extra noise from affecting the final measurement.

Unfortunately, as such a filter’s bandwidth is reduced we begin to encounter some other
problems. The most fundamental of these arises because of the fundamental properties of a
bandpass filter. A filter designed to pass only a narrow frequency range behaves like a resonator.
We can define the bandwidth, $B$, of a filter as

$$B = f(\text{max}) - f(\text{min}) \quad \ldots (6.35)$$

where $f(\text{max})$ and $f(\text{min})$ are the maximum and minimum modulation frequencies which will
pass through the filter. The time taken for a resonance to build up or decay in such a filter will
deck upon $1/B$. Hence the narrower we make the filter bandwidth in order to reduce noise, the
longer the filter output takes to respond to a change in the alternating level at its input. For
various practical reasons, filters where $B \ll f_c$ are also often difficult to make. Also we may
find that, if $B$ is too small, any small changes in the chopping frequency may cause $f_c$ to move
outside the range the filter will pass, and the signal may be lost. Phase sensitive detection
provides a method for dealing with these problems.

In the system shown in Fig 6.5 the amplified alternating. signal is passed to a parallel pair of
amplifiers whose gains are +1 and −1. A reference signal is also taken from the beam chopper
and used to operate a switch which selects which amplifier to connect to the output. The system
is then arranged so the switch connects to the amp with a gain of +1 when the chopper is
allowing the detector to see the source, but connects to the amp with a gain of −1 when the
detector is being shown the comparison, not the source. In effect, this full-wave rectifies the ac
alternation produced by the chopper. This switched signal is then passed from the switch to a
capacitor, \( C \), via a resistor, \( R \). The combination of \( R \) and \( C \) act as a time-constant which averages
the selected signal voltages over the time \( \tau = RC \). This smooths the switched signal to produce
an output dc level due to the difference between \( s \) and \( c \) which takes a time \( \tau \) to change
significantly. In effect, the overall system has taken a steady input, converted it to an alternating
signal for amplifying and processing, and then changed it back into a steady signal for final
measurement.

The reference frequency should be equal to the chopper frequency, \( f_c \), because both are
generated by the same chopping device. The reference and chopped signal should ideally also be
in phase for the same reason. It is, however, possible for there to be a phase difference between
the two. The output voltage on the capacitor will depend upon the phase between reference and
signal. If the relative phase of the reference is changed by 180 deg the switch setting will be
reversed at any particular part of the signal cycle. This produces an output which is inverted,
and the voltage on the capacitor has the same magnitude as before, but is opposite in sign. If
the reference and signal are 90 deg out of phase then the inverted and non-inverted parts of the
signal are equal and opposite, i.e. the averaged voltage is zero.

In general, the smoothed output voltage will be proportional to

\[
(s - c) \cos(\phi)
\]

where \( \phi \) is the phase difference between the reference and the chopped signal.

Any noise at a frequency, \( f_n \), which is close to the frequency, \( f_c \), may be regarded as being
proportional to

\[
\cos(2\pi f_c t + \phi_n)
\]

where the phase now varies with time according to

\[
\phi_n = 2\pi(f_n - f_c) t
\]

i.e. noise at a frequency near the chopping frequency can be regarded as being at the chopping
frequency but having a phase relative to the reference which varies linearly with time. The noise
hence produces a PSD output level which varies as

\[
\cos[2\pi(f_n - f_c) t]
\]

The time-varying output is then averaged by the output RC circuit. Since this does vary
sinusoidally with time it has no overall dc level. As a result, when \( |2\pi(f_n - f_c)| > 1/\tau \) the
noise has little or no effect upon the output dc voltage which appears on the capacitor. Thus in
effect, the PSD acts as a filter which suppresses the effects of noise outside the range

\[
f_c \pm 1/2\pi \tau
\]

The phase sensitive nature of the arrangement means that even a noise component at the
chopping frequency, \( f_c \), will be prevented from having an effect upon the output voltage if that
noise is 90 deg out of phase with the reference (and the wanted signal). There will always be
some noise at the chopping frequency. Being random, its phase will vary unpredictably. We can
treat this noise as if it were composed of two random components which, on average, have
similar power levels, one in-phase with the reference, and one in quadrature. The quadrature component cannot produce any output, hence the PSD essentially rejects half the noise power at the chopping frequency, \( f_c \).

PSD system are simple to build, flexible, and provide an excellent way of dealing with background and noise problems. They are therefore widely used. A range of slightly different practical techniques have arisen in various fields and have been given various names - an example being *sky chopping* or *beam nodding* used by astronomers to dealt with the situation mentioned at the start of this section. Here the telescope or antenna optics is moved so that the source is switched in and out of the detector beam. Provided they are not too far apart, the atmospheric emission (sky noise) in the ‘source’ and ‘off-source’ directions are much the same. A PSD technique, referenced to the antenna beam direction, can then be used to observe the source power and suppress the effects of atmospheric absorption. A similar technique is also used by radio-astronomers who refer to the method as *Dicke switching*.

PSD systems act as *comparators*, i.e. the output depends upon the difference in signal level produced by the source, \( S \), and a comparison, \( C \). This means that the signal measurement can only be as accurate as our knowledge of the level we expect from the comparison. (Note that the comparison is sometimes called a *reference source*. This usage has been avoided here as it is easily confused with the reference taken from the chopper to control the output switch.)

Most precision measurements are, in fact, comparisons rather than ‘absolute’ measurements. We measure the length of something by comparing it with a ruler, a time by counting the ticks of a clock. When buying a gallon of petrol we are less concerned with the absolute definition of a gallon than with feeling certain that this gallon is the same as that at another filling station a few days ago. Similarly, our comparison, \( C \), should be a known standard which can be reproduced or defined with whatever accuracy we require.
Chapter 7
Mixers and Heterodyne Detection.

7.1 Signal Mixing in Diodes

Bolometric signal detectors respond simply to the power level of a signal. Heterodyne detection systems produce output which contains information about the frequency distribution of incident signal power. They are, therefore, of particular value when we wish to determine the power frequency spectrum of a signal. A complete heterodyne detection system is usually called a heterodyne receiver. The detecting element in a heterodyne system is called a mixer. This is because it ‘mixes’ together two inputs to produce an output at another frequency. One of the simplest, and most commonly used, forms of mixer is a diode.

If we alter the voltage applied to a resistor we find that the resulting current varies in proportion to the voltage. The resistor obeys Ohm’s law and is described as an Ohmic device. The voltage, $V$, and current, $I$, are related by the simple expression

$$V = IR$$  \hspace{1cm} (7.1)

where $R$ is the resistance of the chosen resistor. A graph of $V$ against $I$ for such a resistor will be a straight line of slope, $R$.

An electronic diode does not obey Ohm’s Law. If we plot $V$ against $I$ for a diode we discover that the resulting line is curved. A diode exhibits a non-linear $I-V$ Curve. The behaviour of the device as a mixer depend upon this non-linear electronic property.

Most common diodes exhibit an $I-V$ curve where the device current also depends upon the sign of the applied voltage as well as its magnitude. A voltage applied in one direction generates almost no current unless the voltage is so high that the diode ‘breaks down’. A voltage applied in the other direction produces a current which increases roughly exponentially with the applied voltage. i.e. the $I-V$ behaviour of a typical diode may be represented as

$$I = 0 \quad \text{for} \quad V < 0 \quad \text{(reverse direction)}$$  \hspace{1cm} (7.2)

$$I \propto \exp(\beta V) - 1 \quad \text{for} \quad V \geq 0 \quad \text{(forward direction)}$$  \hspace{1cm} (7.3)

where we have adopted the general convention that the main current flow occurs when a ‘+ve’ voltage is applied and that is regarded as being the ‘forward’ direction of current flow.

Here we are interested in the behaviour of the diode when used in the forward direction. Expression 7.3 may be re-written as power series

$$I = a_1 V + a_2 V^2 + a_3 V^3 + \ldots$$  \hspace{1cm} (7.4)

with appropriate values of the coefficients, $a_n$. For simplicity, here we will assume that only the second coefficient is significant, i.e., we will assume that the diode’s curve is of the form

$$I = a V^2$$  \hspace{1cm} (7.5)

(Where $a$ is equivalent to the $a_2$ coefficient of expression 7.4.) This considers the diode to behave as what is called a square-law device.

This assumption of a square-law device turns out to be a fairly good one under many practical circumstances. If necessary, the same arguments as can now be illustrated using the square-law
assumption can be repeated using a more complex $I$-$V$ expression. The mathematical expressions become more involved when this is done, but the final result is similar in most practical cases.

Consider now the system illustrated in Figure 7.1. Here 7.1(a) represents a pair of bandpass filters being used to couple two input sources at the frequencies $f_s$ and $f_i$, into a diode. Filter 1 is designed to behave as a short-circuit at the frequency, $f_s$, and as an open circuit at the frequency, $f_i$. Filter 2 is designed to be a short-circuit at the frequency, $f_i$, and an open-circuit at the frequency, $f_s$. A steady bias voltage, $V_0$, is also applied to the diode via the inductor, $L$.

If the two inputs, produced by the generators attached to the filters, are

$$V_1 = V_s \sin(2\pi f_s t) ; \quad V_2 = V_i \sin(2\pi f_i t) \quad \ldots (7.6)$$

then the total voltage seen at the diode will be

$$V = V_0 + V_1 + V_2 \quad \ldots (7.7)$$

The diode current will therefore be

$$I = a \left[ V_0 + V_s \sin(2\pi f_s t) + V_i \sin(2\pi f_i t) \right]^2 \quad \ldots (7.8)$$

this expression may be re-written in the form

$$I = a V_0^2 + 2a V_0 V_s \sin(2\pi f_s t) + 2a V_0 V_i \sin(2\pi f_i t)$$

$$+ \frac{1}{2} a V_s^2 \left\{ 1 - \cos[2\pi (2f_s t)] \right\} + \frac{1}{2} a V_i^2 \left\{ 1 - \cos[2\pi (2f_i t)] \right\}$$

$$+ a V_s V_i \left\{ \cos[2\pi (f_s - f_i) t] - \cos[2\pi (2f_i + f_s) t] \right\} \quad \ldots (7.9)$$

Looking 7.9 it can be seen that the diode current has oscillatory components at the frequencies $2f_s, 2f_i, f_s - f_i, \text{ and } f_s + f_i$, as well as at the two input frequencies, $f_s, \text{ and } f_i$. The diode thus behaves as a frequency conversion device, transferring some of the power input at the frequencies $f_s, \text{ and } f_i$ to other frequencies. In general, a device which has a non-linear $I$-$V$ curve which is more complex than this simple square-law case will convert power to a range of frequencies, $m f_s \pm \pm n f_i$, where $m$ and $n$ are integers. The details will then depend on the coefficients needed for a power law like expression 7.4 above for us to define the diode curve. However here we will continue to assume the use of a square law device.

None of these newly created frequencies may pass through the input filters or the bias inductor.
We can, however, select and extract the power generated at these frequencies via and output filter. In most practical cases for a heterodyne receiver we are interested mainly in the difference frequency, \( f_s - f_l \), and the output filter will be designed to couple this component to the output so as to produce a resulting voltage

\[
V \text{ (out)} = V_s V_l \cos\left[2\pi (f_s - f_l) t\right]
\]

across the output load (which will generally be the input impedance of a following amplifier).

7.1(b) is a schematic of a typical millimetre-wave mixer system. The mixer diode is mounted as a shunt across a waveguide transmission line and a backshort is placed behind the diode to improve power matching from the waveguide into the diode. The waveguide is coupled to an antenna which collects incident millimetre-wave power.

The input filters 1 & 2 of 7.1(a) are replaced by an optical diplexer system. This takes signal power at the frequency \( f_s \) from a source of interest and couples it efficiently into the mixer's antenna. It simultaneously take the output from a local oscillator (LO) at the frequency \( f_l \) and couples this into the antenna. D.C. bias is passed into the diode via a bias choke which is designed to allow D.C. into the device whilst acting as an open circuit at the signal and LO frequencies. The bias supplied to the mixer is separated from the difference frequency output by a suitable filter - often called a bias tee.

In a mm-wave system the input signal and LO frequencies are typically around 100 GHz or more. However, by selecting an LO frequency which is reasonably similar to that of the signal, the output difference frequency may be a few GHz or less. The system thus acts as a down-converter. An input signal \( V_s \sin(2\pi f_s t) \) will generate an output \( V_s V_l \cos(2\pi f_0 t) \), where \( V_0 = f_s - f_l \ll f_c \). This relatively low frequency output can now be processed using conventional electronic techniques even though the actual input signal frequency, \( f_s \), is extremely high.

The mixer's output is usually amplified, filtered, etc, before finally being measured. It is conventional to refer to the nominal output frequency as the intermediate frequency (IF) and to the output signal as the IF signal because it represents the original signal information at a frequency intermediate between that of the initial input and the final measurement/analysis system which often converts it to a still lower frequency. Provided that the LO frequency, \( f_l \), is known and well defined the signal frequency may be deduced by measuring the frequency, \( f_0 \), of the IF output from the mixer. This is an important difference between heterodyne receivers and bolometric detectors which do not inherently supply any information regarding the signal frequency.

Most real signals do not, in fact, consist simply of a single component of fixed amplitude and frequency. More commonly the input signal will be composed of a set of frequency components or a continuous spectral distribution. Extending the argument employed above it can be seen that an input signal of the form

\[
S(f) = \int_{f_2}^{f_1} A(f) \exp\{-j[2\pi f t + \phi(f)]\} \, df
\]

which represents a continuous spectrum between the frequencies \( f_1 \) and \( f_2 \) will - when mixed with a LO frequency - \( f_l \), produce an IF output of the form

\[
S'(f) = \int_{f_2}^{f_1} A'(f') \exp\{-j[2\pi f t + \phi'(f')]\} \, df
\]
where
\[ A'(f') \propto A(f) \quad ; \quad \phi'(f') = \phi(f) \quad \ldots (7.13 \ ; \ 7.14) \]
and
\[ f' = f - f_l \quad \ldots (7.15) \]
i.e. the output IF spectrum has an amplitude/frequency and phase/frequency spectrum which are proportional to that of the input but have been ‘shifted down’ in frequency by an amount equal to \( f_l \). It can also be seen that the relative phases of output IF components are the same as those of in the initial signal. In principle, the output IF signal contains all of the information which existed in the original input. The IF output from a mixer may therefore be amplified and passed to a suitable data analysis system which will be able to recover detailed spectral information about the input signal. This is a particularly useful capability of heterodyne detection systems. There is, however, one potential cause of confusion or loss of information which may need to be taken into account...

In practice, IF output at a frequency \( f_0 \) may be produced by a signal at either of two quite distinct input frequencies
\[ f_0 = f_1 \pm f_s \quad \ldots (7.16) \]
e.g. we may see an output at, say, 1 GHz emerging from a mixer which is driven with an LO at 100 GHz. But this might be produced by an input signal at 101 GHz or 99 GHz. This is because \( \cos(\theta) = \cos(-\theta) \). Hence the output generated according to expression 7.10 will have the same frequency for either of these two possible inputs. For a given LO frequency, a mixer which produces output over the IF range \( f_1 \) to \( f_2 \) will therefore respond to input signals in the range from \( f_1 + f_1 \) to \( f_1 + f_2 \) and to signals in the range \( f_1 - f_1 \) to \( f_1 - f_2 \). There are therefore two bands of receivable signal frequencies, placed either side of the local oscillator frequency. The range above \( f_1 \) is conventionally referred to as the upper sideband. The range below is referred to as the lower sideband. A receiver system which responds to both is called a double sideband receiver.

In some cases the existance of two distinct sidebands does not matter. The input signal may, for example, be uniform over a wide frequency range. Accepting two input bands then simply allows more signal power to be detected, making the detector more sensitive. In some other cases the source output will already be fairly well known and consists of components which fall only into one sideband. More generally, however, the chance that a given LO output may be produced by an input signal at either of two possible frequencies gives rise to ambiguities in a measurement.

The simplest way of dealing with this problem is to place a signal filter between the source and the mixer. This filter is arranged so as to pass only one sideband without loss and strongly rejects the other, unwanted, sideband. The use of such a filter produces a single sideband receiver system. The wanted frequency range in such a system is sometimes called the signal sideband. The other, unwanted, sideband is sometimes called the image.

The ability of a single sideband heterodyne receiver to accurately measure input signal powers and frequencies depends upon the frequency and amplitude stability of the LO employed. Any uncontrolled or random changes in \( f_l \) or \( V_l \) are multiplied into the output, reducing its value as a measure of the input signal. It is therefore important to provide a stable LO input. The LO generator will, itself, also sometimes generate noise at frequencies which fall into the signal and image bands. In order to obtain good performance this should be as low as possible. It is also
useful, whenever possible, to employ a diplexer which prevents unwanted LO output in these bands from reaching the mixer.

Heterodyne systems can provide frequency measurements of very high accuracy and resolution. Used at frequencies up to a few hundred GHz it is possible to measure frequencies with an accuracy and precision of a few kHz (i.e. one part in $10^8$) or better if required.

Conventional mixer diodes are *junction* devices. Their properties arise as a consequence of the behaviour of charge carriers in the region surrounding the junction between two different materials. Most of the diodes used at conventional radio frequencies are simple *pn-junctions* usually manufactured from doped Silicon. At millimetre-wave frequencies mixers are more commonly made using *Shottky barriers* which consist of a junction between a metal and a semiconductor (usually Gallium Arsenide).

Various other forms of mixer diodes have also been used. Examples included *SIS diodes* (Superconducting-Insulator-Superconducting) and *Josephson junctions*, which depend upon the properties of superconductors and very thin barriers. The active part of all of these devices are very small compared with millimetre wavelengths. A typical junction being the order of a micron or just a few microns in diameter. Hence they must be used as ‘two-terminal’ devices.

Power must be coupled into the active device via a pair of attached electronic contacts.

![Fig. 7.2](image)

Fig. 7.2 (a) Simple physical model of a junction diode. (b) Electronic model of a diode. (c) Electronic model of a diode and the surrounding circuit.

Problems arise in using junction devices because the thickness of the active junction region is usually very small - generally somewhat less than a micron. Figure 7.2 represents an ‘idealised’ form of a junction diode. The arrangement shown in 7.2(a) shows a device of cross sectional area, $A$, and overall length, $h$. The active *depletion region* is of width $d$. For simplicity it is assumed that the bulk of both the materials either side of the active junction region have a resistivity, $\rho$, and a relative permittivity, $\varepsilon$. The junction region is assumed to share the same permittivity, but may effectively have has a somewhat higher resistivity, $\rho'$. Ohmic contacts, of negligible resistivity, are placed covering the ends of the device and are used to couple signals in and out of the device via external wires.
This physical arrangement may be represented as an electronic circuit of the form shown in 7.2b, where

\[ R_s = \frac{h\rho}{A} \quad R_d = \frac{d\rho'}{A} \quad \ldots (7.17) \]

\[ C_0 = \frac{\varepsilon A}{b} \quad C_d = \frac{\varepsilon A}{d} \quad \ldots (7.18) \]

In a real device the width of the depletion region, \( d \), and its effective resistivity, \( \rho' \), will vary with the applied voltage. The resistance, \( R_d \), and capacitance, \( C_d \), do not, therefore, have unique values. Here we will assume that \( R_d \) corresponds to the gradient, \( \delta V / \delta I \) of the diode's \( I-V \) curve. The capacitance, \( C_d \), is assumed to be \( \delta q / \delta V \) where \( q \) is the charge stored in the device capacitance. Both of these quantities will depend upon the chosen bias voltage and act as the effective resistance and capacitance for small voltage changes about the chosen level. This simple model should not be taken as an accurate representation of any particular form of junction device. The physical details of a real device will differ in various respects and its behaviour is considerably more complex. However, the model is useful in demonstrating some of the basic limitations of junction devices.

The non-linear resistance of the device resides in the variable nature of \( R_d \). Hence an input signal, \( V_i \), applied between the end contacts must be coupled without significant loss to \( R_d \) if the device is to work efficiently. Between the input terminals and the 'active' diode resistance, \( R_d \), the combination of \( R_s \) and \( C_d \) form a time constant which acts as a low-pass filter. The resistance \( R_s \) is generally called the device spreading resistance as it represents the resistance through which any current must be distributed or spread to reach the active region. Provided that the device is well designed and \( R_s < R_d \) we can say that the signal voltage, \( V_d \), across \( R_d \) will be such that

\[ |V_d| \approx \frac{|V_i|}{\sqrt{1 + (2\pi f_s \tau)^2}} \quad \ldots (7.19) \]

where, \( f_s \), is the signal frequency, and

\[ \tau = R_s C_d = \frac{\varepsilon \rho b}{d} \quad \ldots (7.20) \]

where we have assumed that \( b \) is somewhat larger than \( d \). Hence, in order to be able to function efficiently we require that

\[ f < \frac{1}{2\pi \varepsilon \rho} \quad \ldots (7.21) \]

A more precise analysis of a particular device yields a result which, while differing in detail, behaves in much the same way. There is an upper frequency limitation imposed by the physical properties of the device materials which cannot be overcome by alterations in the size and shape of the device. For Gallium-Arsinide materials used in millimetre-wave mixers, this problem means that simple junction diode devices cannot function well above approximately 1000 GHz. Practical difficulties in manufacturing real devices generally mean that it is difficult to approach this upper limit. Alternate choices of device materials and techniques can improve this situation, but all junction devices must face this basic problem in some form.

A typical GaAs Shottky mixer with a diameter of the order of a micron will have a resistance of a few Ohms and a capacitance of a few femtoFarads (femto = \( 10^{-15} \)) under typical bias conditions. Whilst this is a small capacitance, it corresponds to an impedance of a few hundred Ohms at mm-wave frequencies. A typical mixer diode therefore has a signal impedance typified by a low resistance shunted by a moderate capacitance. At higher frequencies the effects of capacitance begin to dominate the impedance of the device.
A typical mixer will be mounted in a waveguide and a thin wire, or post, is placed across the guide so as to couple power into the device. This wire behaves essentially like an inductor placed across the guide in series with the mixer. 7.2(c) represents the combination of such a wire and mixer diode placed as a shunt load across a transmission line. A backshort is used to improve signal matching into the mixer. The characteristic two-terminal impedance across the center of a typical millimetre-wave waveguide is around 100 Ohms or more. A typical mixer will have a resistance of the order of a few Ohms. Hence, by itself, the mixer is poorly matched to the transmission line. The situation may, however, be improved by choosing the inductance, \( L \), of the connecting wire so as to form a resonant circuit in conjunction with the diode’s capacitance.

The properties of resonant circuits and their use to match impedance is considered in more detail in chapter 11. A similar method can be employed here to optimise the efficiency of mm-wave power coupling from waveguide to diode. The combination of the diode, inductance, and back short will produce a waveguide termination load, \( Z \), which can be described by

\[
\frac{1}{Z} = \frac{1}{\omega} + \frac{1}{jX}\quad \text{... (7.22)}
\]

where \( Z_d \) is the diode’s complex impedance and \( jX \) represents the reactance at the diode’s location which is produced by the back short. We can also say

\[
\frac{1}{Z_d} = \frac{1}{\text{Re}(Z_d) + \text{Im}(Z_d)}\quad \text{... (7.23)}
\]

where we have split the diode’s impedance into real and imaginary parts.

In order to match power ideally into the diode from a transmission line whose characteristic impedance is \( Z_c \) we would require \( Z \) to be such that

\[
\text{Re}(Z_d) = Z_c\quad \text{... (7.24)}
\]

\[
\text{Im}(Z_d) = 0\quad \text{... (7.25)}
\]

at the frequency of interest. In practice meeting both these requirements may be difficult. As a result, the best which may be possible in practice may result in some of the input signal and LO power being reflected by the diode assembly and backshort. Also, the above relies on a resonance of some kind, so the optimum power coupling may be limited to a narrow frequency range.

### 7.2 Bulk Mixers

It is very difficult to manufacture junction devices less than a micron or so in diameter. Any increase in device size will lower the resistance, \( R_d \), and increase the capacitance, \( C_d \), making it harder to obtain good signal coupling. The problems of low resistance and shunt capacitance tend to rise with increasing signal frequency.

To overcome these problems it is necessary to replace the active junction with a device which operates as a bulk mixer. To achieve this we require a material whose inherent resistivity depends upon the signal power level. To illustrate this we may consider the situation shown in Figure 7.3. Here we have a piece of material which generates (or passes in response to an applied bias voltage) a current which varies proportionally to the input signal power level.

In fact, it can be seen that most bolometric detectors which provide electrical output act in this way, i.e. they produce an output current which varies in proportion with the input signal power.
level. For this reason, most bolometric detectors can - in principle, at least - be used as heterodyne mixers. In practice their use as mixers is normally severely limited for reasons which will become clear later.

![Diagram](image)

**Fig. 7.3** (a) Mixer crystal receiving two waves, $E_s$, and $E_l$, and providing an output current, $i_0$. (b) Heterodyne receiver system.

7.3(a) shows a piece of material of surface area, $A$, illuminated with a signal field, $E_s \sin (2\pi f_s t)$, and a local oscillator field, $E_l \sin (2\pi f_l t)$. Any current produced by the material is passed through the connecting wires attached to its ends. The total power level, $P(t)$, falling upon the material at time, $t$, will therefore be

$$P(t) = \frac{A}{Z_0} [E_s \sin (2\pi f_s t) + E_l \sin (2\pi f_l t)]^2$$

... (7.26)

where $Z_0$ is the impedance of free space.

The material is assumed to behave as a photoconductor, i.e. the output current is proportional to the rate at which photons are arriving, and thus to the input power illuminating the area, $A$. Neglecting any current which flows when the input power is zero we may write that the input power generates a current

$$i(t) = \frac{e\eta P(t)}{hv}$$

... (7.27)

where $e$ is the charge on an electron. For the sake of simplicity it has been assumed that each incoming photon has an energy, $hv$, where $v$ is equivalent to a frequency mid-way between the actual signal and LO frequencies. i.e. we assume $v = (f_s + f_l)/2$. In practice these two frequencies are generally very similar and this approximation is a good one.

The incoming power level then corresponds to $P(t)/hv$ photons per unit time. The factor, $\eta$, is a quantum efficiency value which determines the probability that any individual arriving photon will produce an electron that contributes to the output current. Combining the previous two expressions we can write that

$$i(t) = i_s + i_l + 2\sqrt{i_s i_l} \cos [2\pi (f_s - f_l) t]$$

... (7.28)

where
\[
i_s = \frac{e \eta A E_s^2}{2Z_0 b v} \quad \text{and} \quad i_l = \frac{e \eta A E_l^2}{2Z_0 b v} \quad \ldots (7.29; 7.30)
\]

plus some extra terms which correspond to current oscillations at the much higher frequencies \(f_s, f_l, f_s + f_l, 2f_s, \text{ and } 2f_l\). When the input signal and LO frequencies are in the millimetre-wave region these other very high frequency components will not normally be passed along the wires to any connected measuring system. Here we will assume that they do not influence any measurement of the current generated in the wire leads. Only \(i(t)\) as defined by expression 7.28 will actually be observed to emerge along the wires.

The individual steady currents, \(i_s\) and \(i_l\), are the values which would be obtained if just the signal or LO power alone were directed onto the detector element. However, as in the case of a diode mixer, the combination of the signal and LO powers arriving simultaneously produces an oscillatory current at the difference frequency. The magnitude of this output current, \(i_0\), is

\[
i_0 = \frac{e \eta A}{Z_0 b v} E_s E_l \quad \ldots (7.31)
\]

When the input signal and LO fields are in-phase, the mean incident power will be proportional to \((E_s + E_l)^2\). When the fields are in opposition (180 degrees out of phase) then the mean incident power will be be proportional to \((E_s - E_l)^2\). The two input fields change their relative phase with the frequency \(f_s - f_l\).

A given detector will take a specific time, \(\tau\), to respond to a change in the input power level. The output from the detector therefore indicates the mean incident power level, averaged over the time, \(\tau\). The detector output will hence only oscillate significantly at the difference frequency when

\[
|f_s - f_l| < \frac{1}{2\pi \tau} \quad \ldots (7.32)
\]

For a time constant, say, of one second this requires the signal and LO frequencies to differ by less than 0.2Hz. Hence, for almost all bolometric (heat) signal detectors we find that the thermal response time means that the useable frequency bandwidth is small, making the detector unsuitable for heterodyne reception under most circumstances.

An exception to this general finding is the InSb Hot Electron bolometer. This has a response time of around a microsecond or less. As a result, InSb bulk mixers are used in some heterodyne receivers at very high frequencies. The response bandwidth of these mixers is typically around 1MHz when cooled to a few degrees above absolute zero.

### 7.3 Signal/Noise Properties of Heterodyne receivers

Figure 7.3(a) illustrates a mixer coupled to an amplifier whose input impedance is \(R_i\). If the IF current produced by the mixer is \(i_0 \cos \left[2\pi (f_s - f_l)\right]\) then the mean power supplied to the amplifier will be

\[
P_i = \langle i_0^2 \rangle R_i \quad \ldots (7.33)
\]

where \(\langle i_0^2 \rangle\) is the mean squared current which, from the expressions given above, may be written as

\[
\langle i_0^2 \rangle = \frac{1}{2} \left[ \frac{e \eta A E_s E_l}{Z_0 b v} \right]^2 \quad \ldots (7.34)
\]

The mean signal power, \(P_s\), and LO power, \(P_l\), incident upon the mixer will be
combining these expressions we obtain

\[ P_I = G P_s \]  \hspace{1cm} (7.36)

where we can define

\[ G = \frac{2e^2\eta^2 R_i}{b^2v^2} P_I \]  \hspace{1cm} (7.37)

From 7.36 it can be seen that the IF power, \( P_I \), supplied to by the mixer is simply proportional to the signal power, \( P_s \), received. The factor, \( G \), which determines the ratio of these two power levels is conventionally called the conversion gain of the mixer. The conversion gain’s value depends upon both the load resistance, \( R_i \), connected to the mixer, and the incident LO power level, \( P_I \). In theory, we could assemble a mixer system which responds to an input signal by producing an output level, \( P_I \gg P_s \) - such a system would amplify the signal as well as changing its frequency.

In practice, however, a range of problems often prevents us from producing a system with actual gain. A typical millimetre-wave mixer will have a conversion gain < 1 - i.e. the gain is, in fact, a loss, indicated by a value which is less than unity. A mixer is, by its very nature, a non-linear device. If the LO power level is increased sufficiently a real mixer will saturate. The output of the mixer then ceases to be power dependent and it no longer functions as a mixer. Hence there will be an ‘optimum’ LO power level which provides the best possible conversion gain for a particular mixer. Problems also arise if \( R_i \) is too high:- for example, any real system will have non-zero capacitance. The system bandwidth may become restricted by the time-constant this produces with the resistance, \( R_i \).

The mechanisms which produce noise in bolometer systems also occur in heterodyne systems. In addition, heterodyne detectors suffer from a particular noise problem caused by the need to illuminate the mixer with LO power. In general \( P_I \gg P_s \) and the mean current flowing in the mixer will be essentially \( i_I \). This LO induced current will produce a shot noise level

\[ \langle i_I^2 \rangle = 2e i_I B \]  \hspace{1cm} (7.38)

where \( B \) is the output measurement bandwidth (i.e. the range of frequencies we observe being produced by the mixer). Now the output IF signal power will be such that

\[ \langle i_s^2 \rangle = 2i_s i_I \]  \hspace{1cm} (7.39)

hence the output signal/noise power ratio, ignoring any other sources of noise, will be

\[ \frac{S}{N} = \frac{i_s}{eB} \]  \hspace{1cm} (7.40)

However,

\[ i_s = \frac{e\eta P_s}{bv} \]  \hspace{1cm} (7.41)

which means that

\[ \frac{S}{N} = \frac{\eta P_s}{bvB} \]  \hspace{1cm} (7.42)

The factor, \( \eta \), represents a quantum efficiency - the chance that an arriving photon will liberate an electron that then contributes to the output current. This implies that \( \eta \ll 1 \). The best possible signal/noise power ratio we can expect will therefore be

\[ \frac{S}{N} \text{ (max)} = \frac{P_s}{bvB} \]  \hspace{1cm} (7.43)
A number of important conclusions may be drawn from this result. Firstly, it can be seen that the maximum possible signal/noise ratio does not depend explicitly upon the LO power level. This may initially be a surprise as it has been established that the conversion gain, and hence the output power level, tends to increase in proportion to the LO power. However, the noise level, in, also increases in proportion with the LO power. These two effects tend to balance out, producing a fixed signal/noise ratio.

In any real device the conversion gain will tend to vary with the LO power level in a less simple way than has been assumed. There will also be a number of other processes generating noise within the receiver system. Hence the actual S/N ratio of a real system will peak at a particular LO power level. No real system, however, can be expected to provide a signal/noise performance in excess of the maximum value set by expression 7.43.

It can also be seen that the maximum possible S/N depends upon the average signal/LO frequency, $\nu$. This may be seen as a consequence of the fact that each photon has an energy, $h\nu$. This tells us that – for the same input signal power level - a high frequency signal will provide fewer photons per second than a lower frequency. Thus as a general rule the signal current for a given power will tend to fall when the signal frequency is raised. In this respect high-frequency heterodyne systems cannot work as well as those used at lower radio frequencies. Finally, we can see that the S/N depends upon the output IF bandwidth, $B$. The shot noise generated within the mixer has a uniform power spectrum. Hence if we increase the bandwidth then we will observe more noise from the mixer.

In some cases the signal power, $P_s$, is at a single frequency or is confined to a narrow frequency range. Under these circumstances it is an advantage to use a mixer (or employ suitable passband filters) which restricts $B$ to the lowest value which passes the IF output produced by the signal. The noise performance of a heterodyne system is often indicated by quoting a noise temperature value. This is a useful method as it is a figure which may be measured fairly easily and which relates directly to a common application of millimetre-wave heterodyne systems.

7.3(b) illustrates a practical mixer system being used to detect the output from a ‘black body’ thermal source. The source is large enough to completely fill the beam pattern of the detector’s antenna. The radiation output from such a source is determined simply by its physical temperature, $T$. The Wien displacement law states that the peak output power per Hz bandwidth from a thermal source occurs at the wavelength, $\lambda_m$, such that

$$\lambda_m T = 5.1 \text{ mm K} \quad \text{(7.44)}$$

This result is based upon locating the maximum of the Planck curve which describes the output spectrum of a black body thermal source.

At signal wavelengths, $\lambda \gg \lambda_m$, it can be shown from the Planck curve, that the signal power from a thermal source which fills the field of view of a detector, is simply given by

$$P_s = kT B \quad \text{(7.45)}$$

The portion of the spectrum which satisfies the above inequality is conventionally referred to as the Rayleigh-Jeans region. In almost all cases of practical interest it is found that a millimetre-wave receiver observing a thermal source is observing in this region of the spectrum. A typical example being a system being used to make thermal maps of the Earth from orbit. The temperature of the Earth would be around 300 K, implying that $\lambda_m \approx 10$ microns. A heterodyne system observing the Earth at a few millimetres wavelength would, therefore, be working well into the Rayleigh-Jeans region.
The mixer in 7.3(b) is assumed to have a conversion gain, $G$, and passes its output signal to an amplifier of power gain, $A$. A bandpass filter is used to define the signal bandwidth, $B$, and the IF signal power level is measured using an meter whose output is averaged over the time, $\tau$. Any noise generated in the mixer (or in the following amplifier) is assumed to come from a mythical thermal source of temperature, $T_n$, which injects unwanted noise power into the mixer’s input. For a single sideband receiver the mean output voltage provided by a power meter will be proportional to

$$V = kBG(A + T_n)$$

By using two thermal sources of temperatures $T_1$ and $T_2$ we can obtain two output levels, $V_1$ and $V_2$, such that

$$V_1 = kBG(A + T_n) ; \quad V_2 = kBG(A + T_n)$$

by combining and re-arranging these expressions we can show that

$$T_n = \frac{T_1V_2 - T_2V_1}{V_1 - V_2}$$

The value $T_n$ can therefore be easily be measured given two thermal sources of different known temperatures. Although it provides no information regarding the origins of the system noise, the system (or receiver) noise temperature, $T_n$, is a simple and convenient measure of the amount of noise appearing at the output and affecting measurements.

If the mixer responds equally to both input signal sidebands then the system acts as a double sideband receiver. The signal power falling upon the mixer is now $2kTB$ rather than $kTB$, and produces twice the IF output of a single sideband system. This doubles the output voltage measured by the power meter. For an unchanged level of mixer/amplifier noise this halves the value of $T_n$ obtained from expression 7.48.

To distinguish between these two cases it is necessary to specify whether the measurement includes power from both sidebands or not. This has produced the convention of specifying receiver noise temperatures as being either the double sideband or single sideband value. In practice the noise temperature of a double sideband system is sometimes measured using a thermal source and the value obtained is doubled before being quoted as a 'single sideband' result. When comparing different system care must be taken to ensure that noise temperatures are compared on a 'like with like' basis to avoid confusion.

Where we are interested in the performance of a mixer rather than that of a whole system it is conventional to quote a mixer noise temperature along with the conversion gain. Many of those working on mixers treat the mixer noise as a thermal source ‘after’ the mixer - i.e. defined so that the mixer noise power, $P_m$, presented to the following amplifier is

$$P_m = kT_mB$$

where $T_m$ is the mixer noise temperature. This definition has the advantage that it is the same for double or single sideband systems. Confusion may arise, however, because this definition omits the conversion gain factor, $G$, which appears in the definition of system noise temperatures. Hence the value of $T_m$ in a particular system will often be much lower than $T_n$.

When using a receiver the important parameter is the overall system noise. Details of the amounts of mixer noise, amplifier noise, etc, are important only when designing or modifying heterodyne systems. The actual output measurement will be of the voltage produced by the IF power detector, averaged over a time, $\tau$. One of the properties of random (‘white’) noise is that a noise power whose mean power level is $P_n$ will fluctuate from moment to moment by an amount proportional to $\sqrt{P_n}$. Hence the mean fluctuation in the measured output will vary
proportionally to $1/\sqrt{P_n}$. The spectrum of these fluctuations is also uniform (‘white’ over the frequency range $0$ to $B$ Hz.

The device which measured the IF power level is often called the detector. This usage is sometimes a little confusing, but it is reasonable because the device acts in the same way as a bolometric detector - i.e. it provides an output signal proportional to the overall input power level. The output time constant is then conventionally referred to as the post-detection filter to distinguish it from the IF filter which follows the mixer.

The time constant acts as a low pass filter, only passing fluctuations in the detected level which vary with frequencies below $1/(2\pi \tau)$. Hence the observed S/N ratio at the filter output will depend upon $1/\sqrt{\tau}$. Combining this with the effect of the IF filter we may conclude that the overall S/N behaviour of the system will allow us to detect a signal level corresponding to a change in the thermal source temperature

$$\Delta T = \frac{T_n}{\sqrt{B\tau}}$$

... (7.50)

In practice, a real receiver will differ in detail from the simple system assumed here. This will mean that various correction factors should be included in the above expressions. The results obtained here are, however, a good guide to the general properties of heterodyne signal detection systems.

From the expressions given earlier for the power from a thermal source and the maximum possible S/N ratio performance it follows that the minimum possible system noise figure for a single sideband receiver will be

$$T_n(\text{min}) = \frac{h\nu}{k}$$

... (7.51)

i.e., if $\nu$ is in GHz and $T_n$ in K,

$$T_n(\text{min}) = 0.047\nu$$

at a signal frequency of 300 GHz this implies that a single sideband receiver cannot have a system noise temperature of less than 14 K.

At present, practical millimetre-wave heterodyne receivers have system noise temperatures which are much higher than this ideal limiting value. InSb bulk mixers are generally able to provide the lowest practical noise temperatures. Over the 100 - 500 GHz range InSb receivers have been made which provide double sideband system noise temperatures of around 200 - 300 K. The InSb mixer is able to offer good performance because it is cooled to a few degrees K, reducing the background thermal noise level. Being a bulk mixer is is also often better matched to the input signal power.

In comparison, Shottky diode mixers operating at room temperature generally offer system noise temperatures around 1000 K. This can be reduced to around 500 K by cooling the mixer diode to around 20 K. At lower temperatures the performance of Shottky devices deteriorates because of the electronic properties of the materials. Although offering lower noise levels, InSb mixers are limited to an IF bandwidth of around 1 MHz if cooled to obtain the lowest possible noise. A typical Shottky device is able to offer a bandwidth which is hundreds or thousands of times greater. The choice of mixer will, therefore, be affected by the bandwidth of the incoming signal power we need to observe.

At present a considerable amount of attention is being devoted to SIS mixers. These offer wide
IF bandwidths, comparable to or better than Shottky devices, but are able to operate when cooled to much lower temperatures. Heterodyne receivers based upon SIS mixers have already produced system noise temperatures below 100 K at 100 GHz. It is probably that this level of system noise will become available at higher frequencies over the next decade or so.