2 Radio Waves

Around 1864, Maxwell was able to show that his equations implied that electromagnetic fields satisfied a wave equation. Remarkably, the speed of these waves turned out to be exactly that of light, suggesting that light is an electromagnetic wave phenomenon. It turns out that light is a very high-frequency form of electromagnetic energy and it was not long before physicists started looking for electromagnetic energy at lower frequencies, those we now associate with radio waves. This led to the classic work of Heinrich Hertz around 1886 in which he demonstrated the generation and detection of radio waves. In this chapter we will show that Maxwell's equations predict fields with wavelike behaviour and show that accelerating charge is the source of such waves. The chapter then concludes with a description of the experiments that led to the discovery of radio waves and the early technological developments that turned radio from a scientific curio into a technology.

2.1 Waves

We will start our discussion of waves by considering those waves that appear as ripples on the surface of water, something with which most of us are familiar. If we have initially calm water, we can cause a disturbance in the water through the forced motion of a platform at the surface of the water (see Figure 2.1) and this disturbance will then travel outwards from the object at a finite speed (c say). The disturbances will be in sympathy with the motion of the object, but a point at distance x from the object will not receive the effect of the motion until a time x/c later. If the motion of the body starts at time 0, at time T later the water will still be calm at distances greater than cT from the object. If the wave is confined to a channel, the wave will maintain its shape as it travels out, i.e. we have a wave height of form $\psi_+(x+ct)$ leftwards and $\psi_-(x-ct)$ rightwards. A plane with $x-ct=$ constant, or $x+ct=$ constant, is known as a *wavefront* (the boundary between still and disturbed water being an important example of a wavefront). A further property of water waves is that they transfer energy from their source (the platform) to a detector at some distance from the source. This can be seen from the fact that when a water wave passes a previously motionless platform it will be set into motion in sympathy with the wave. Such a process is utilised these days in the form of a wave power generator.

Water waves exhibit all of the important features of waves. These include the ability to transport energy over large distances, the energy being transported at a finite speed c. Further, that the source can create an arbitrarily shaped wave that maintains its shape as

https://doi.org/10.1017/9781108684514.003 Published online by Cambridge University Press

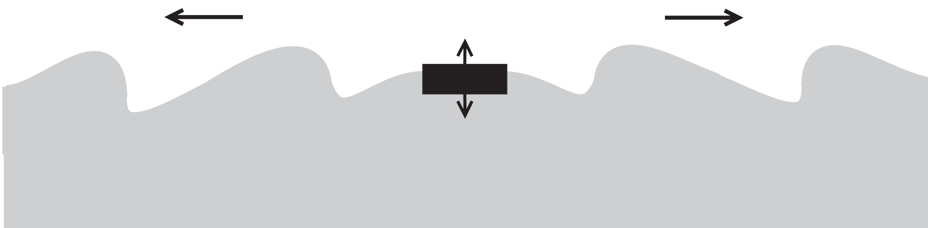

Fig. 2.1 Water waves generated by an oscillating platform.

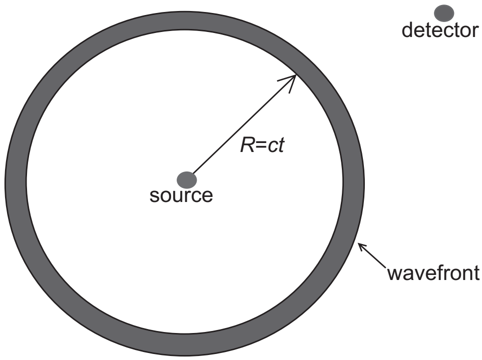

Fig. 2.2 Pulse travelling out from a bounded source.

it travels outwards and, as a consequence, can transfer arbitrary information over large distances. Water waves will satisfy a *wave equation* of the form

$$\frac{\partial^2 \psi}{\partial x^2} - \frac{1}{c^2}\frac{\partial \psi^2}{\partial t^2} = 0. \tag{2.1}$$

In the case of the rightwards wave ψ_-, this can be seen from the following argument. It will be noted that ψ_- is a function of a single variable $\phi = x - ct$. If f is a function of ϕ, an important result of calculus tells us that $\partial f/\partial x = (\partial\phi/\partial x)\mathrm{d}f/\mathrm{d}\phi = \mathrm{d}f/\mathrm{d}\phi$ and that $\partial f/\partial t = (\partial\phi/\partial t)\mathrm{d}f/\mathrm{d}\phi = -c\mathrm{d}f/\mathrm{d}\phi$. As a consequence, $\partial^2\psi_-/\partial x^2 = \mathrm{d}\psi_-^2/\mathrm{d}\phi^2$ and $\partial^2\psi_-/\partial t^2 = c^2\mathrm{d}\psi_-^2/\mathrm{d}\phi^2$, from which it can be seen that (2.1) will be satisfied for arbitrary $\psi_-(x-ct)$. In a similar fashion, (2.1) will also be satisfied by arbitrary $\psi_+(x+ct)$.

A more realistic wave is one that is generated by a bounded source (see Figure 2.2). The waves will travel outwards from such a source and so the wave field ψ will be a function of $R - ct$ where R is the distance from the source (the wavefront is spherical in this case). For most physical fields, the energy in the field is proportional to the field squared (the field is the wave height in the case of a water wave). Consequently, as the wave propagates it will spread out and the rate of spread will need to conserve energy. In the case of water waves, these are two-dimensional (they only exist on the surface of the water) and so the field will need to fall off as $1/\sqrt{R}$. Many fields (including sound and electromagnetic fields) will exist in three-dimensional space and so, in order to conserve energy, the field will need to fall off as $1/R$.

2.2 Electromagnetic Waves

We want to show that Maxwell's equations allow solutions that behave as waves. However, we first need to introduce some further ideas concerning vectors. The space we live is said to be three-dimensional and this means that we can specify any point through three numbers, its coordinates. There are many different sorts of coordinate system, but we will concentrate on what are known as Cartesian coordinates. These are arguably the simplest coordinates and were invented in the seventeenth century by the French mathematician René Descartes. In Cartesian coordinates we define an origin O for which the coordinates are $(0,0,0)$ and then the position of any point $\mathbf{r}$ is given by the vector $x\hat{\mathbf{x}}+y\hat{\mathbf{y}}+z\hat{\mathbf{z}}$, where $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ are unit length vectors ($|\hat{\mathbf{x}}|=|\hat{\mathbf{y}}|=|\hat{\mathbf{z}}|=1$) that are mutually orthogonal ($\hat{\mathbf{x}}\cdot\hat{\mathbf{y}}=\hat{\mathbf{y}}\cdot\hat{\mathbf{z}}=\hat{\mathbf{x}}\cdot\hat{\mathbf{z}}=0$). The vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ are known as *basis vectors*. Essentially, the coordinates are the distances that we must travel in the direction of the associated basis vectors in order to travel from the origin O to position $\mathbf{r}$. Any vector field $\mathbf{F}(\mathbf{r})$ can also be represented in terms of the above basis vectors as $F_x(\mathbf{r})\hat{\mathbf{x}}+F_y(\mathbf{r})\hat{\mathbf{y}}+F_z(\mathbf{r})\hat{\mathbf{z}}$ where $F_x(\mathbf{r})$, $F_y(\mathbf{r})$ and $F_z(\mathbf{r})$ are known as the components of the vector field (often abbreviated to just F_x, F_y and F_z). The important thing about the Cartesian coordinate system is that basis vectors are the same vectors at all points in space (not so for more exotic coordinate systems such as those of the polar variety). Concerning vectors expressed in component form, there are two important results that the reader should note

$$\mathbf{a}\cdot\mathbf{b}=a_xb_x+a_yb_y+a_zb_z \tag{2.2}$$

and

$$\mathbf{a}\times\mathbf{b}=(a_yb_z-a_zb_y)\hat{\mathbf{x}}+(a_zb_x-a_xb_z)\hat{\mathbf{y}}+(a_xb_y-a_yb_x)\hat{\mathbf{z}}. \tag{2.3}$$

We will assume that there is an EM wave travelling in the $\hat{\mathbf{x}}$ direction. Consequently, the fields will depend on x and t alone (i.e. $\boldsymbol{\mathcal{E}}(\mathbf{r})=\boldsymbol{\mathcal{E}}(x,t)$ and $\boldsymbol{\mathcal{B}}(\mathbf{r})=\boldsymbol{\mathcal{B}}(x,t)$) and then only through the combination $\phi(x,t)=x-ct$ where c is some, as yet to be determined, propagation speed (we assume a right-moving wave for the present). We will now study the implication of these assumptions for the Maxwell equations. We first consider the equations

$$\int_S \boldsymbol{\mathcal{B}}\cdot\mathbf{n}dS=0 \text{ and } \int_S \epsilon\boldsymbol{\mathcal{E}}\cdot\mathbf{n}dS=Q \tag{2.4}$$

and apply these equations over a small cube that does not contain any charge ($Q=0$), the cube having centre at the general point $\mathbf{r}$ and sides of length Δ.

There will be contributions to the integral from the six faces of the cube, the integral on each being approximated as the area of the face multiplied by the normal component of the field at the centre of each face. Consequently,

$$\begin{aligned}\int_S \boldsymbol{\mathcal{B}}\cdot\mathbf{n}dS \approx&\mathcal{B}_x(x+\Delta/2,t)\Delta^2-\mathcal{B}_x(x-\Delta/2,t)\Delta^2\\&+\mathcal{B}_y(x,t)\Delta^2-\mathcal{B}_y(x,t)\Delta^2\\&+\mathcal{B}_z(x,t)\Delta^2-\mathcal{B}_z(x,t)\Delta^2,\end{aligned} \tag{2.5}$$

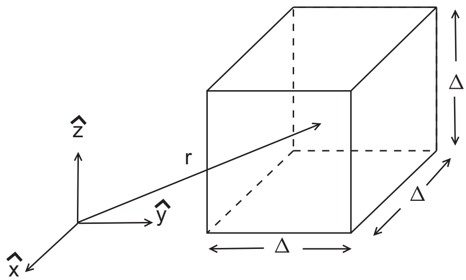

Fig. 2.3 Cube for evaluating Maxwell's equations.

where the contributions on the right-hand side are from, in order, the sides with normals $\hat{\mathbf{x}}$, $-\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, $-\hat{\mathbf{y}}$, $\hat{\mathbf{z}}$ and $-\hat{\mathbf{z}}$. The equation $\int_S \boldsymbol{\mathcal{B}} \cdot \mathbf{n} \mathrm{d}S = 0$ will then imply that

$$\mathcal{B}_x(x + \Delta/2, t)\Delta^2 - \mathcal{B}_x(x - \Delta/2, t)\Delta^2 = 0. \tag{2.6}$$

From (2.6) we will have that $(\mathcal{B}_x(x+\Delta/2,t) - \mathcal{B}_x(x-\Delta/2,t))/\Delta = 0$ and so, in the limit $\Delta \to 0$, this will imply that $\partial\mathcal{B}_x/\partial x = 0$. The only way that this can hold is if $\mathcal{B}_x$ is itself constant and so we will take $\mathcal{B}_x = 0$ since a constant non-zero field throughout space is not the wavelike solution we are looking for. In a similar fashion, we will have $\mathcal{E}_x = 0$.

We now consider Eqs. (1.35) and (1.36) when applied over small rectangular circuits that are not threaded by any current ($\mathcal{I} = 0$). The circuits are static and so we can bring the time derivative under the integral to obtain

$$\int_C \frac{\boldsymbol{\mathcal{B}}}{\mu} \cdot \mathrm{d}\mathbf{r} = \int_S \epsilon \frac{\partial \boldsymbol{\mathcal{E}}}{\partial t} \cdot \mathbf{n} \mathrm{d}S \tag{2.7}$$

and

$$\int_C \boldsymbol{\mathcal{E}} \cdot \mathrm{d}\mathbf{r} = -\int_S \frac{\partial \boldsymbol{\mathcal{B}}}{\partial t} \cdot \mathbf{n} \mathrm{d}S. \tag{2.8}$$

Since the rectangles are small, the surface integral can be approximated by $\mathbf{n} \cdot \partial\boldsymbol{\mathcal{B}}/\partial t$ when evaluated at the centre of the rectangle and multiplied by the area of the rectangle. On each side of the rectangle, the line integral can be approximated by $\mathbf{t} \cdot \boldsymbol{\mathcal{B}}/\mu$ evaluated at the mid point of the side ($\mathbf{t}$ is a unit vector along the side) multiplied by the length of the side. Consider a rectangle in the xy plane (see Figure 2.4a), then (2.7) will become

$$-\frac{\mathcal{B}_x(x,t)}{\mu}\Delta - \frac{\mathcal{B}_y(x-\Delta/2,t)}{\mu}\Delta + \frac{\mathcal{B}_x(x,t)}{\mu}\Delta + \frac{\mathcal{B}_y(x+\Delta/2,t)}{\mu}\Delta = \epsilon\frac{\partial\mathcal{E}_z}{\partial t}\Delta^2, \tag{2.9}$$

where the terms on the left hand side are, in order, the contributions from the sides with directions $-\hat{\mathbf{x}}$, $-\hat{\mathbf{y}}$, $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. As a consequence,

$$\frac{1}{\Delta}\left(\frac{\mathcal{B}_y(x+\Delta/2,t)}{\mu} - \frac{\mathcal{B}_y(x-\Delta/2,t)}{\mu}\right) = \epsilon\frac{\partial\mathcal{E}_z}{\partial t} \tag{2.10}$$

and then, in the limit $\Delta \to 0$,

$$\frac{1}{\mu}\frac{\partial\mathcal{B}_y}{\partial x} = \epsilon\frac{\partial\mathcal{E}_z}{\partial t}. \tag{2.11}$$

https://doi.org/10.1017/9781108684514.003 Published online by Cambridge University Press

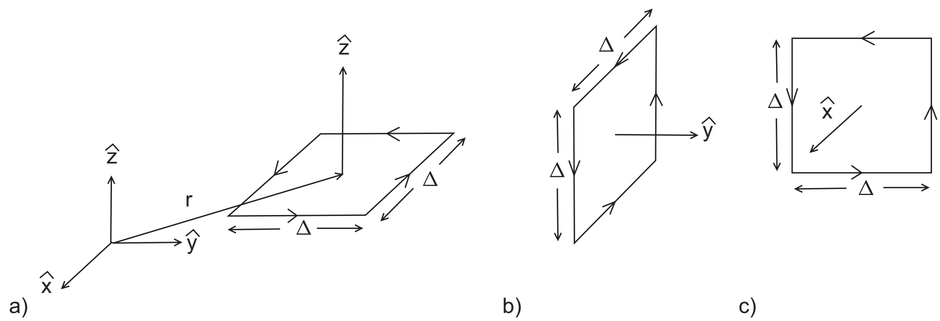

Fig. 2.4 Rectangular circuit for evaluating Maxwell's equations.

In a similar fashion, 2.8 will imply that

$$\frac{\partial \mathcal{E}_y}{\partial x} = -\frac{\partial \mathcal{B}_z}{\partial t}. \tag{2.12}$$

If we now consider a rectangle in the xz plane (see Figure 2.4b), similar arguments imply that

$$\frac{\partial \mathcal{E}_z}{\partial x} = \frac{\partial \mathcal{B}_y}{\partial t} \tag{2.13}$$

and

$$-\frac{1}{\mu}\frac{\partial \mathcal{B}_z}{\partial x} = \epsilon \frac{\partial \mathcal{E}_y}{\partial t}. \tag{2.14}$$

For a rectangle in the zy plane (see Figure 2.4c), the equations are identically satisfied. Bringing all of this together, we have $\mathcal{B}_x = \mathcal{E}_x = 0$ with the four remaining field components ($\mathcal{E}_y$, $\mathcal{E}_z$, $\mathcal{B}_y$ and $\mathcal{B}_z$) satisfying Eqs. (2.11) to (2.14).

If we take the t derivative of (2.11) and the x derivative of (2.13) we obtain

$$\frac{1}{\mu}\frac{\partial^2 \mathcal{B}_y}{\partial x \partial t} = \epsilon \frac{\partial \mathcal{E}_z^2}{\partial t^2} \tag{2.15}$$

and

$$\frac{\partial^2 \mathcal{E}_z}{\partial x^2} = \frac{\partial^2 \mathcal{B}_y}{\partial t \partial x}. \tag{2.16}$$

Then, eliminating $\partial^2 \mathcal{B}_y / \partial t \partial x$ between these two equations, we obtain the wave equation

$$\frac{\partial^2 \mathcal{E}_z}{\partial x^2} - \mu\epsilon \frac{\partial^2 \mathcal{E}_z}{\partial t^2} = 0, \tag{2.17}$$

where $c = 1/\sqrt{\epsilon\mu}$ is the propagation speed of the wave ($c = 3 \times 10^8$ m/s for an electromagnetic wave in a vacuum and this is often denoted by c_0). In a similar fashion

$$\frac{\partial^2 \mathcal{E}_y}{\partial x^2} - \mu\epsilon \frac{\partial^2 \mathcal{E}_y}{\partial t^2} = 0, \quad \frac{\partial^2 \mathcal{B}_z}{\partial x^2} - \mu\epsilon \frac{\partial^2 \mathcal{B}_z}{\partial t^2} = 0 \quad \text{and} \quad \frac{\partial^2 \mathcal{B}_y}{\partial x^2} - \mu\epsilon \frac{\partial^2 \mathcal{B}_y}{\partial t^2} = 0, \tag{2.18}$$

i.e. all field components satisfy the same wave equation.

Although the above equations imply that the field components $\mathcal{E}_y$, $\mathcal{E}_z$, $\mathcal{B}_y$ and $\mathcal{B}_z$ all satisfy wave equations, they are not independent of each other. For simplicity, we will assume a right-travelling wave in the $\hat{\mathbf{x}}$ direction. All of the field components will then be functions of $\phi = x - ct$ and, by similar arguments to the previous section, we find from (2.11) that

$$\frac{\mathrm{d}\mathcal{B}_y}{\mathrm{d}\phi} = -c\epsilon\mu\frac{\mathrm{d}\mathcal{E}_z}{\mathrm{d}\phi}. \tag{2.19}$$

Integrating with respect to ϕ, and noting that $\epsilon\mu = 1/c^2$, we obtain

$$\mathcal{B}_y = -\frac{1}{c}\mathcal{E}_z. \tag{2.20}$$

In a similar fashion, from (2.12) to (2.14),

$$\mathcal{E}_y = c\mathcal{B}_z\ ,\ \mathcal{E}_z = -c\mathcal{B}_y \text{ and } \mathcal{B}_z = \frac{1}{c}\mathcal{E}_y. \tag{2.21}$$

Since $\mathcal{E}_x = \mathcal{B}_x = 0$, it is clear that the propagation direction (the $\hat{\mathbf{x}}$ direction) will be orthogonal to both $\boldsymbol{\mathcal{E}}$ and $\boldsymbol{\mathcal{B}}$. Further, noting the relations in Eqs. (2.20) and (2.21), we obtain that $\boldsymbol{\mathcal{E}} \cdot \boldsymbol{\mathcal{B}} = \mathcal{E}_y\mathcal{B}_y + \mathcal{E}_z\mathcal{B}_z = c\mathcal{B}_z\mathcal{B}_y - c\mathcal{B}_y\mathcal{B}_z = 0$. As a consequence, we have that the propagation direction, the magnetic field $\boldsymbol{\mathcal{B}}$ and the electric field $\boldsymbol{\mathcal{E}}$ are all mutually orthogonal, a basic property of EM waves. Noting the relations (2.20) and (2.21), we also have the relation

$$\boldsymbol{\mathcal{B}} = -\frac{1}{c}\hat{\mathbf{x}} \times \boldsymbol{\mathcal{E}}, \tag{2.22}$$

where $\hat{\mathbf{x}}$ is a unit vector in the propagation direction. It is clear that an electric wave will always have a corresponding magnetic field. However, providing it is orthogonal to the propagation direction $\hat{\mathbf{x}}$, the components of the electric field $\boldsymbol{\mathcal{E}}$ can be arbitrary functions of $x - ct$, i.e. $\boldsymbol{\mathcal{E}} = \boldsymbol{\mathcal{E}}(x - ct)$. Such waves are known as *plane waves* due to the fact that the fields are constant across planes that are orthogonal to the propagation direction. Although plane waves might seem like a mathematical idealisation, they are found to give a good representation of real waves at large distances from their sources. Further, the orthogonality properties, and the relation (2.22), are retained by real waves. Figure 2.5 shows an example of a plane wave pulse that propagates in the $\hat{\mathbf{x}}$ direction and for which $\boldsymbol{\mathcal{E}}$ has the form $\mathcal{E}_y\hat{\mathbf{y}}$. Then, from (2.21), we then find that $\boldsymbol{\mathcal{B}} = (\mathcal{E}_y/c)\hat{\mathbf{z}}$. In general, a plane wave has a propagation direction $\mathbf{p}$ (a unit vector) with the electric field orthogonal to this direction. The electric field $\boldsymbol{\mathcal{E}}$ can be an arbitrary function of $\mathbf{p} \cdot \mathbf{r} - ct$ and the magnetic flux density is given by

$$\boldsymbol{\mathcal{B}} = -\frac{1}{c}\hat{\mathbf{p}} \times \boldsymbol{\mathcal{E}}. \tag{2.23}$$

Before we proceed any further, we need to address the issue of the energy content of an EM field. This is important since one of the major properties of waves is that they transport energy. In the case of the electric field, energy is most simply understood by considering the example of a capacitor. We consider the parallel plate capacitor of Figure 2.6a and calculate the work required to move a charge Q between the capacitor plates starting from a condition of no charge. This will be the energy of the field inside

https://doi.org/10.1017/9781108684514.003 Published online by Cambridge University Press

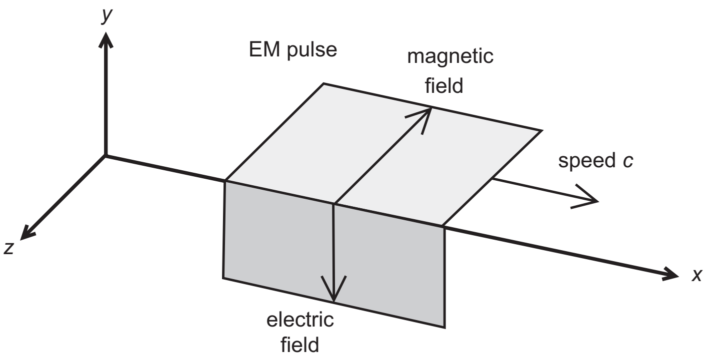

Fig. 2.5 Plane electromagnetic wave consisting of a pulse.

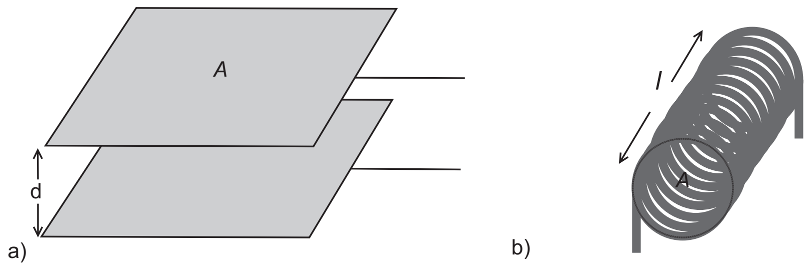

Fig. 2.6 Parallel plate capacitor and long solenoid.

the capacitor. If there are charges q and $-q$ on the plates of the capacitor, there will be a field of strength $\mathcal{E} = q/\epsilon A$ between the plates (A is the area of the capacitor plates). If we move a small amount of charge $\mathrm{d}q$ from the bottom plate to the top plate, this will require an amount of work $\mathrm{d}W = \text{force} \times \text{distance} = \mathrm{d}q\mathcal{E}d = \mathrm{d}q(qd/\epsilon_0 A)$ to be performed (d is the distance between the capacitor plates). If we build up the charge from 0 to Q in $\mathrm{d}q$ stages, the total work done will be sum of these stages, i.e. the integral

$$W_e = \int_0^Q \frac{qd}{\epsilon_0 A}\mathrm{d}q = Q^2 \frac{d}{2\epsilon_0 A} \tag{2.24}$$

and, since $\mathcal{E} = Q/\epsilon_0 A$,

$$W_e = \frac{1}{2}\epsilon_0 \mathcal{E}^2 V, \tag{2.25}$$

where V is the volume over which the capacitor field exists. For a general electric field $\boldsymbol{\mathcal{E}}$, the energy of the field contained inside a volume V will be

$$W_e = \frac{1}{2}\int_V \epsilon \boldsymbol{\mathcal{E}} \cdot \boldsymbol{\mathcal{E}} \mathrm{d}V. \tag{2.26}$$

The above volume integral can be understood by dividing V into N smaller volumes δV_i which are small enough for the electric field to be represented by a constant value $\boldsymbol{\mathcal{E}}_i$. Over ith volume, the energy field can be approximated by $\epsilon \boldsymbol{\mathcal{E}}_i \cdot \boldsymbol{\mathcal{E}}_i \delta V_i/2$ and then the

https://doi.org/10.1017/9781108684514.003 Published online by Cambridge University Press

total energy will be the sum of these contributions. The integral is then the limit as the size of the volumes tends to zero.

As we have seen above, the electric field cannot exist without the magnetic field in the case of an EM wave. Consequently, we also need to discuss the energy content of the magnetic field. This is most simply understood by considering the example of an inductor. We consider the long solenoid shown in Figure 2.6b and need to calculate the work required to establish a current flow of magnitude $\mathcal{I}$ at time T, starting from zero current at time 0. This will then be the energy W_h of the field inside the inductor. If a current i is to be driven through the inductor, the source will need to have an EMF equal to $-\mathcal{E}$ where $\mathcal{E}$ is the EMF that is generated in the inductor by the current flow. If we consider a time interval dt, a charge idt will flow into the inductor and hence the source will supply the inductor with energy $-i\mathcal{E}\mathrm{d}t$. Since $\mathcal{E} = -L\mathrm{d}i/\mathrm{d}t$ for an inductor with inductance L, the source will do the work $\mathrm{d}W_h = iL(\mathrm{d}i/\mathrm{d}t)\mathrm{d}t$. Summing the work done from time 0 to time T (i.e. integrating with respect to time)

$$W_h = \int_0^t Li\frac{\mathrm{d}i}{\mathrm{d}t}\mathrm{d}t = \int_0^I Li\mathrm{d}i = \frac{1}{2}L\mathcal{I}^2. \tag{2.27}$$

For a solenoid of length l, and cross-sectional area A, $L = \mu_0 NA/l$ and the field inside the solenoid is given by $\mathcal{B} = \mu_0 n\mathcal{I}$. As a consequence,

$$W_h = \frac{1}{2}\frac{B^2}{\mu_0}V, \tag{2.28}$$

where $V(= lA)$ is the volume occupied by the solenoid. For a general magnetic field $\boldsymbol{\mathcal{B}}$, the energy of the field contained inside a volume V will be

$$W_h = \frac{1}{2}\int_V \frac{\boldsymbol{\mathcal{B}}\cdot\boldsymbol{\mathcal{B}}}{\mu}\mathrm{d}V. \tag{2.29}$$

The total EM field energy can now be stated as

$$W = W_e + W_h = \frac{1}{2}\int_V \left(\epsilon\boldsymbol{\mathcal{E}}\cdot\boldsymbol{\mathcal{E}} + \frac{\boldsymbol{\mathcal{B}}\cdot\boldsymbol{\mathcal{B}}}{\mu}\right)\mathrm{d}V. \tag{2.30}$$

We now consider the energy flow of an EM wave. On noting the vector identity $(\mathbf{a}\times\mathbf{b})\cdot(\mathbf{a}\times\mathbf{b}) = (\mathbf{a}\cdot\mathbf{a})(\mathbf{b}\cdot\mathbf{b}) - (\mathbf{a}\cdot\mathbf{b})^2$, we obtain from (2.23) that $\boldsymbol{\mathcal{B}}\cdot\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{E}}\cdot\boldsymbol{\mathcal{E}}/c^2$. As a consequence, for an EM wave,

$$W = \int_V \epsilon\boldsymbol{\mathcal{E}}\cdot\boldsymbol{\mathcal{E}}\mathrm{d}V. \tag{2.31}$$

Consider the rate at which energy is transported by an EM wave, i.e. the power. Let A be an area that is orthogonal to the propagation direction, then an amount of energy

$$dW = \epsilon\boldsymbol{\mathcal{E}}\cdot\boldsymbol{\mathcal{E}}Acdt \tag{2.32}$$

will be transported across the area A in the time interval dt. Consequently the power density, the rate of transport of energy across a unit area, will be given by

$$\mathcal{P} = \frac{1}{\eta}\boldsymbol{\mathcal{E}}\cdot\boldsymbol{\mathcal{E}}, \tag{2.33}$$

https://doi.org/10.1017/9781108684514.003 Published online by Cambridge University Press

where $\eta = \sqrt{\frac{\mu}{\epsilon}}$ is known as the *impedance* of the medium (in a space that is free of matter $\eta = \eta_0 = \sqrt{\frac{\mu_0}{\epsilon_0}} \approx 120\pi$ ohms). The use of the terminology 'impedance' can be justified on dimensional grounds. Firstly, the electric field $\mathcal{E}$ has dimensions V/L where V is the dimension of voltage and L is the dimension of length. From the Maxwell equation (1.35) we have that $\mathcal{B}/\mu$ has the dimensions of current I/L where I is the dimension of current. Consequently, noting relation (2.22), $\eta = c\mu$ must have the dimensions of V/I, i.e. the dimensions of impedance.

2.3 The Field of an Accelerating Charge

Consider a charge q that is initially resting at some origin O, in which case it will have an electric field $\mathcal{E} = (q/4\pi\epsilon_0 r^3)\mathbf{r}$. At time 0 we accelerate the charge to a velocity $\Delta\mathbf{v}$ over a time period Δt, after which the charge continues at a constant speed $\Delta\mathbf{v}$. At time t after the acceleration period, the field lines will be those shown in Figure 2.7 (Purcell, 2013). Outside a sphere of radius $c(t+\Delta t)$, the field will still be that of a charge fixed at the origin O since the information about the changes to the source has not yet had time to arrive. Inside a sphere of radius ct about the charge, however, the field will have settled down to that of a charge moving with constant velocity Δv and hence will have the form $\mathcal{E} = (q/4\pi\epsilon_0\tilde{r}^3)\tilde{\mathbf{r}}$ where $\tilde{\mathbf{r}} = \mathbf{r} - t\Delta\mathbf{v}$. A fundamental requirement of EM fields is that field lines be continuous, except at sources (i.e. current and charge). As a result, we obtain the field lines for the accelerating period by joining up the corresponding field lines on these two spheres. It will now be noted that the field lines in the period of acceleration are not completely radial and have a component $\mathcal{E}_t$ that is transverse to the radial direction. We will now assume that t is very much greater than Δt, i.e. we will consider the effect of the acceleration a long time after it has happened. Referring

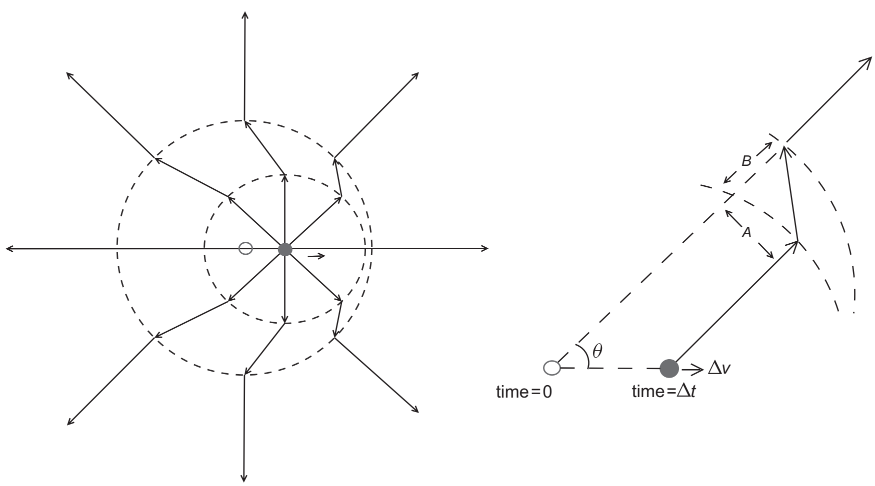

Fig. 2.7 Radiation by an accelerating charge.

https://doi.org/10.1017/9781108684514.003 Published online by Cambridge University Press

to Figure 2.7, we have that $A \approx t\Delta v \sin\theta$ and $B \approx c\Delta t$. It therefore follows that

$$\frac{\mathcal{E}_t}{\mathcal{E}_r} = \frac{A}{B} \approx \frac{t\Delta v \sin\theta}{c\Delta t}, \tag{2.34}$$

where $\mathcal{E}_r$ is the radial component of the electric field. We further assume that Δv is very much smaller than the speed of light c and therefore $\tilde{r} \approx ct$. As a consequence, $\mathcal{E}_r \approx q/4\pi\epsilon_0 c^2 t^2$ for points between the spheres. From (2.34) we will now have that

$$\mathcal{E}_t = f\frac{q\sin\theta}{4\pi\epsilon_0 c^2 \tilde{r}}, \tag{2.35}$$

where $f = \Delta v/\Delta t$ is the acceleration of the charge. This expression also holds for a particle that is continuously accelerating, but we must now remember that a point at a distance $|\tilde{\mathbf{r}}|$ is responding to an acceleration that happened at time $t - |\tilde{\mathbf{r}}|/c$. As a consequence

$$\mathcal{E}_t = f(t - |\tilde{\mathbf{r}}|/c)\frac{q\mu_0 \sin\theta}{4\pi\tilde{r}}, \tag{2.36}$$

where the acceleration is now a function $f(t)$ of time. The important thing to note is that the field of an accelerating charge falls away as $1/|\tilde{\mathbf{r}}|$ as we move far away from the charge, whereas that of a charge in uniform motion falls away as $1/|\tilde{\mathbf{r}}|^2$. By accelerating a charge, we have extended its influence way beyond that of a static charge.

At large distances from the accelerating charge, the $\mathcal{E}_t$ term of the electric field will dominate and this will behave like a plane wave travelling radially outwards from the charge. Importantly, the wave can be given an arbitrary shape by controlling the behaviour of the charge acceleration. We now show that this wave carries energy away from the charge. In order to find the power P travelling outwards from the charge, we will need to integrate the power density $\mathcal{P}$ over a large sphere of radius R that surrounds the charge, i.e.

$$P = \frac{1}{\eta}\int_S \boldsymbol{\mathcal{E}} \cdot \boldsymbol{\mathcal{E}} \mathrm{d}S. \tag{2.37}$$

The sphere can be considered as a series of rings of area $2\pi R^2 \sin\theta d\theta$ that are parameterised by angle θ (see Figure 2.8). The power travelling through this ring will be $(\mathcal{E}_t^2/\eta)R^2 \sin\theta \mathrm{d}\theta$ and adding up the contributions from all rings, we obtain

$$P = \frac{1}{\eta}\int_0^{\pi} \mathcal{E}_t^2 2\pi R^2 \sin\theta \mathrm{d}\theta, \tag{2.38}$$

where $\mathcal{E}_t = f(t - R/c)q\mu_0 \sin\theta/4\pi R$. Then, noting that $\int_0^{\pi} \sin^3\theta \mathrm{d}\theta = 4/3$, we obtain the Larmor formula

$$P = \frac{f_{\mathrm{ret}}^2 q^2 \mu_0^2}{6\eta\pi}, \tag{2.39}$$

where $f_{\mathrm{ret}} = f(t - R/c)$ denotes the acceleration at a retarded time $t - R/c$. It will be noted that the power carried away by waves is proportional to the square of the acceleration.

https://doi.org/10.1017/9781108684514.003 Published online by Cambridge University Press

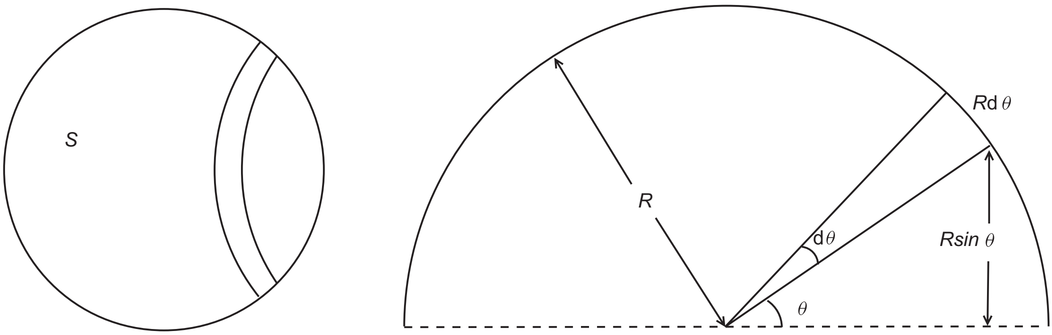

Fig. 2.8 Geometry for calculating power.

2.4 The Field of an Oscillating Charge

As we have seen, an accelerating charge will cause radio waves. For practical radio, however, we will need to produce accelerations in a sustainable fashion. It turns out that we can do this through an oscillating charge and, furthermore, this will have properties that are beneficial to radio. Consider a charge that performs harmonic oscillations over a distance $2l$ along a line that is parallel to unit vector $\hat{\mathbf{t}}$, i.e

$$\mathbf{r}_s = l\sin(\omega t + \phi)\hat{\mathbf{t}}, \tag{2.40}$$

where ω is the *oscillation frequency* in radians per second and ϕ is an arbitrary phase factor. The acceleration of the charge will be $f(t) = -\omega^2 l\sin(\omega t + \phi)$ and so, from (2.36), its field will have the transverse component

$$\mathcal{E}_t = -\omega^2 l\sin(\omega t - 2\pi R/\lambda + \phi)\frac{q\mu_0 \sin\theta}{4\pi R}, \tag{2.41}$$

where R is the distance from the charge and $\lambda = 2\pi c/\omega$. Furthermore, at large distances (many wavelengths away), this will be the dominant component. Figure 2.9 shows the field lines for a charge oscillating at a frequency of 3×10^7 Hz. (Note that Hz, short for *hertz*, is one cycle per second. Further, frequency f in hertz is related to angular frequency ω through $\omega = 2\pi f$.) The field is shown at equally spaced times over a full period of oscillation and it will be noted that the field lines exhibit a wavelike structure that moves outwards with time. The waves have a wavelength (distance between crests) $\lambda = 2\pi c/\omega$ which is 10 m in this case (the scales on the Figures are in metres).

In practical terms we can produce the effect of an oscillating charge by driving an oscillating current into a metal structure that is known as an antenna. The current will cause electrons to move backwards and forwards on the structure in a similar fashion to the single charge above. Consequently, at large distances, this will result in a field that is similar in nature to that of an oscillating charge. On the antenna structure, it will be noted that there will be a counterbalancing positive charge for each electron. However, since these charges are static, they have no effect upon the far field. The simplest practical antenna consists of a metallic rod with a generator driving oscillating current into its centre (see Figure 2.11). The electrons will move coherently and so the current can be

https://doi.org/10.1017/9781108684514.003 Published online by Cambridge University Press

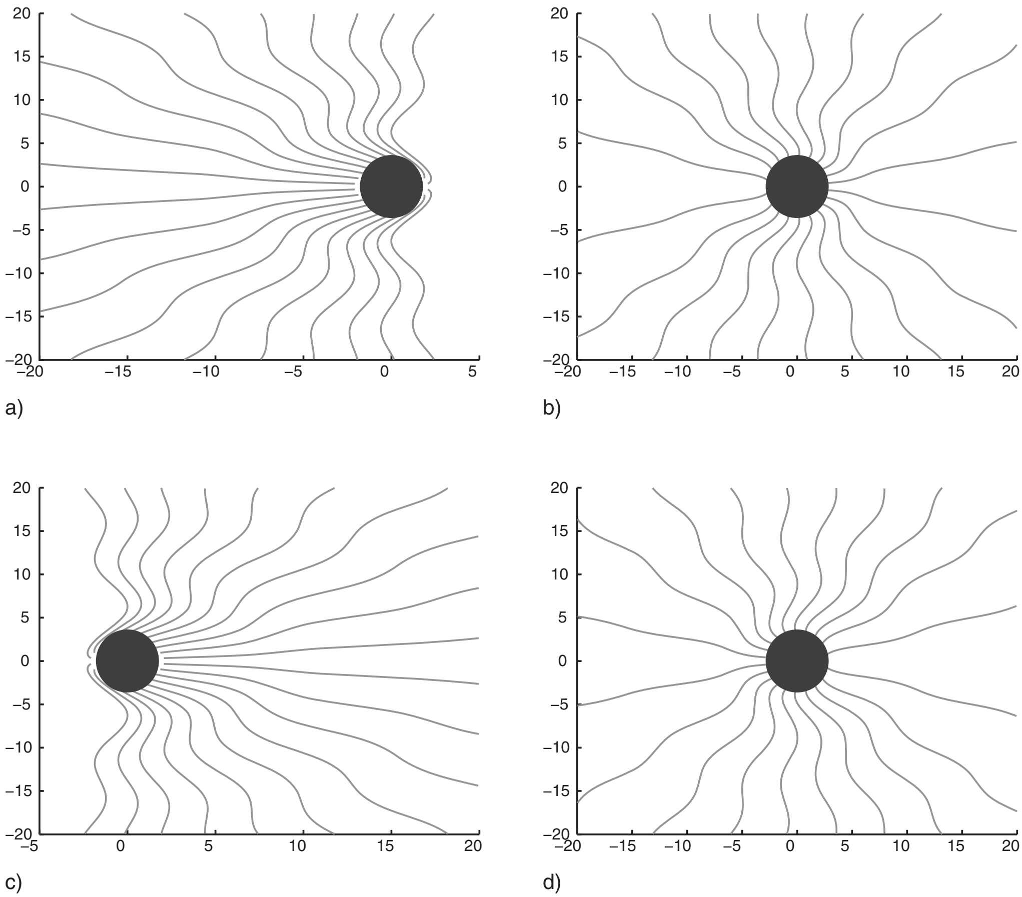

Fig. 2.9 Electric field lines of an oscillating charge over one period.

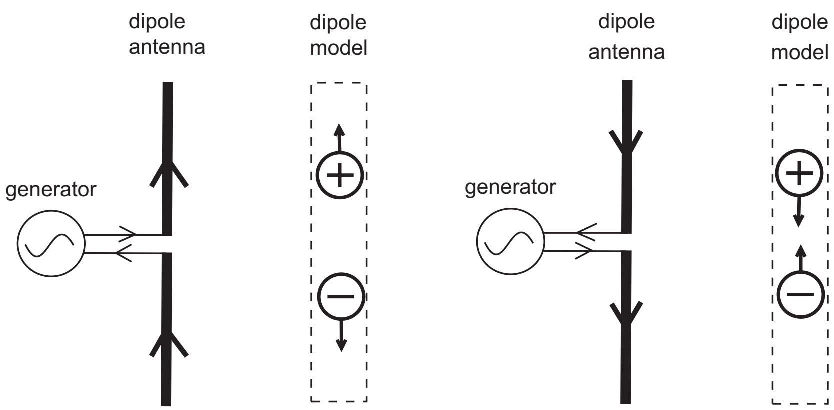

Fig. 2.10 Dipole antenna and model.

https://doi.org/10.1017/9781108684514.003 Published online by Cambridge University Press

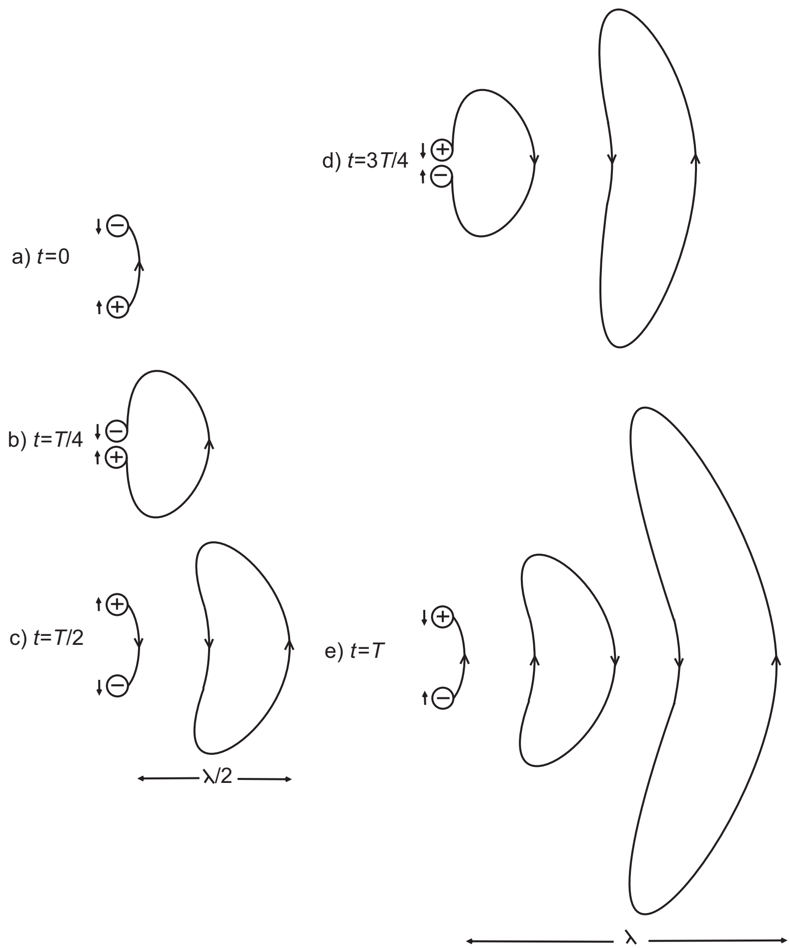

Fig. 2.11 Development of the dipole field over a period T of oscillation.

modelled as a pair of opposing charges that oscillate backwards and forwards along the antenna axis (i.e. an oscillating dipole). These charges move in opposite directions (they are 180° out of phase) and so will represent, at any time, a current in the same direction. For a dipole, the charges will oscillate backwards and forwards between the ends of the dipole, a distance $2l$, and the effective current will be $I = I_0 \cos(\omega t + \phi)$, where $I_0 = q\omega$. Consequently, at large distances, the transverse component of electric field will be

$$\mathcal{E}_t = -\sin(\omega t - 2\pi R/\lambda + \phi)\frac{\omega\mu_0 I_0 l \sin\theta}{2\pi R} \tag{2.42}$$

(this is the sum of the fields for positive and negative charge). Figure 2.11 shows the development of the dipole field through a full period of the charge oscillations. Field lines cannot cross and so, as the charges themselves cross, the field lines break off from the charges and the ends join together to form closed field lines. These closed field lines will then move away at the speed of light to make room for the new field lines that join

https://doi.org/10.1017/9781108684514.003 Published online by Cambridge University Press

the charges. Because of its similarity with a simple dipole of two charges, a driven metal rod is known as a *dipole antenna*.

The use of a sinusoidal generator will produce a radio wave at a single frequency and this can be advantageous when many radio users must share the radio spectrum. In this case, a receiver of radio signals can be designed to *filter* out all but the frequency of the desired signal. A sinusoidal signal, however, carries no information and must be modulated in some fashion if it is to convey information. We can achieve modulation by varying a suitable combination of I_0, ω and ϕ. If we vary I_0 alone we have *amplitude modulation*, ω alone we have *frequency modulation* and ϕ alone we have *phase modulation*. Unfortunately, the modulation will cause the radio signal to spread out around the *carrier frequency* ω and the width of the spread is known as the *bandwidth* of the signal.

2.5 The First Radio Systems

As we have seen in the previous two sections, we need to have accelerating charge in order to generate radio waves and charge oscillations are a convenient means of producing the necessary acceleration in a sustainable fashion. Around the year 1886, German physicist Heinrich Hertz demonstrated a practical device for producing radio waves. Using this device he was able to verify the existence of radio waves. The generator that Hertz invented is what we now call a *spark transmitter* and this was the mainstay of radio for the next thirty years. The essence of a spark transmitter and Hertz's receiving system is shown in Figure 2.12. In this Figure, when the key closes, the capacitor charges up until it reaches a point where the voltage across the spark gap is great enough for the insulation afforded by the air to break down and current to flow (the spark). The circuit, consisting of a series capacitor and inductor, is now closed and, as we have seen in Chapter 1, there is a damped oscillation that drives oscillatory current into a dipole antenna through a secondary winding on the inductor. When the oscillations in the circuit have dropped below a certain level, the sparking stops and the circuit breaks. The charge now builds up again on the capacitor until the break down point in the spark gap is once again reached and the oscillation starts again. As a consequence, the transmitted radio signal will consist of a sequence of damped oscillations. The radio wave generates an EMF in a receiving loop that is suitably orientated and this can cause a spark across the gap in

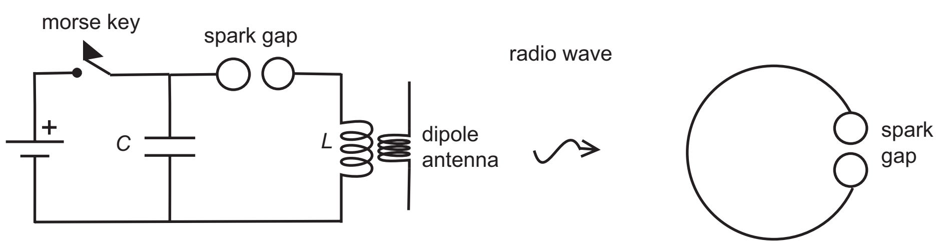

Fig. 2.12 A basic spark transmitter and a detector circuit.

https://doi.org/10.1017/9781108684514.003 Published online by Cambridge University Press

the loop (Hertz needed to use a microscope to see this very small spark). According to Faraday's law, the EMF in the loop will be given by $\mathcal{E} = -A\partial\mathcal{B}_n/\partial t$ where $\mathcal{B}_n$ is the component of magnetic field that is normal to the plane of the loop and A is the area of the loop.

Whilst the transmitter shown in Figure 2.12 could, in theory, work for a small enough spark gap, the strength of the radio waves would have been far too small for the experiments of Hertz. To overcome this problem, Hertz used an induction coil to produce a high voltage. (This is the way that your car produces a spark to ignite the petrol in its engine.) Figure 2.13 illustrates the way in which this is achieved. Initially a current flows through a primary winding with a low number of turns. This causes a magnetic field in an iron core that attracts the lever on the right, which breaks the circuit. This sudden change in current produces a voltage pulse in the secondary winding. The secondary will have a far greater number of turns than the primary and so the pulse at the output will have a much larger voltage than that of the battery. While the key is closed, this process is repeated continuously resulting in a sequence of high-voltage pulses. In the Hertz transmitter, the battery and key of Figure 2.12 are replaced by the induction system of Figure 2.13. Spark transmitters were the predominant form of radio transmitter until well into the twentieth century and for this reason a radio operator was often known as a 'sparks', the terminology remaining into the current era.

Although Hertz verified the existence of radio waves, his apparatus was not practical as a piece of technology, mainly due to the inadequacy of the detector. Indeed, Hertz expressed the view that his discoveries would never amount to anything practical. Unfortunately, he did not live long enough to see how wrong he was on this score (he died in 1894). The first demonstration of the practical possibilities of radio were performed by the Indian physicist Sir Jagadish Chandra Bose. In November of 1894, Bose gave a public demonstration of radio-wave transmission over a distance of 23 metres, using his transmissions to set off a bell and some gunpowder. Also in that year, the British physicist Sir James Lodge gave a practical demonstration of the transmission of telegraph signals using radio. It was, however, left to the Italian inventor Guglielmo Marconi, to

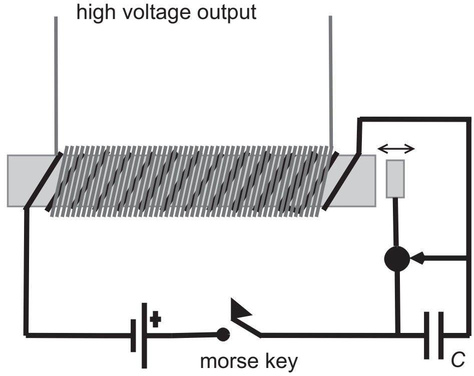

Fig. 2.13 Induction coil for providing high voltage pulses.

https://doi.org/10.1017/9781108684514.003 Published online by Cambridge University Press

turn Hertz's remarkable discovery of radio into a commercially viable technology. In 1895 Marconi successfully transmitted radio signals over distances several kilometres (Hertz had only managed a few metres). The transmitting equipment of Marconi was much the same as that used by Hertz, but the radio receiver used a detection device known as a *coherer*. The coherer was a device invented by Edouard Branly in 1890. This device consisted of two metal electrodes with the space between them filled with loose iron fillings. Initially the device exhibits a high resistance, but when a radio signal appears across its electrodes the metal fillings cling together (i.e. they 'cohere') and the device becomes highly conductive. Figure 2.14 shows a typical radio receiver that is based on such a device (note that a dipole antenna can be used to receive a well as transmit a radio signal). An incoming radio wave will generate a time-varying current in the antenna and this, in turn, will generate a time-varying voltage in the secondary of the transformer. The situation will now resemble that of Figure 1.24 of Chapter 1 with the load R_L representing the detection system (i.e. the coherer and subsequent circuitry). When a strong enough radio signal is received at the resonant frequency of L and $C1$, the coherer becomes conducting and the DC (direct current) circuit involving the battery is completed. This DC pulse will then be detected by the radio operator as a click in the speaker. Unfortunately, once a radio signal has set a coherer into its conducting state, the fillings will remain cohered. Consequently, after a detection, the coherer needs to be agitated to loosen the fillings before the next detection can take place.

The problem with a coherer-based radio receiver is its lack of sensitivity and this continued to hold back the development of radio for almost a decade. The solution to this problem eventually came from an effect that was discovered by Ferdinand Braun in 1874. Braun found that galena crystals (a mineral form of lead sulphide) could conduct electricity asymmetrically (as much as a 2:1 asymmetry between forward and backward current is possible). Figure 2.15a shows a device, known as a *crystal detector*, that is based on this property (we now call such a device a *point contact diode*). A device of this form was first patented by Jagadish Bose in 1904, but a similar device using carborundum (silicon carbide) was also patented by Henry Dunwoody in 1907 and proved to be more reliable. In the Figure 2.15a, one connection to the detector is made through a mercury bath in which the galena sits and the other through a thin wire, known as a 'cat's whisker', that must be carefully adjusted to make contact at a point on the crystal that exhibits a suitable amount of asymmetry. This device made possible the receiver architecture

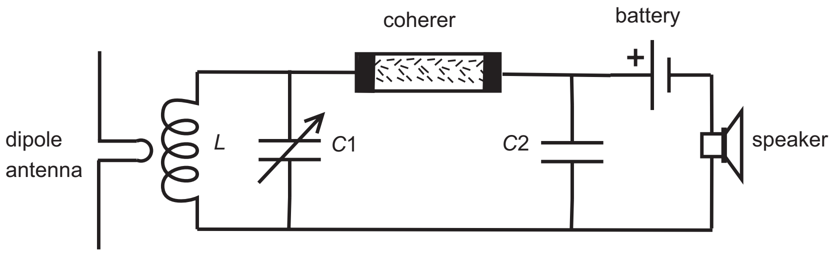

Fig. 2.14 Radio receiver based on a coherer.

https://doi.org/10.1017/9781108684514.003 Published online by Cambridge University Press

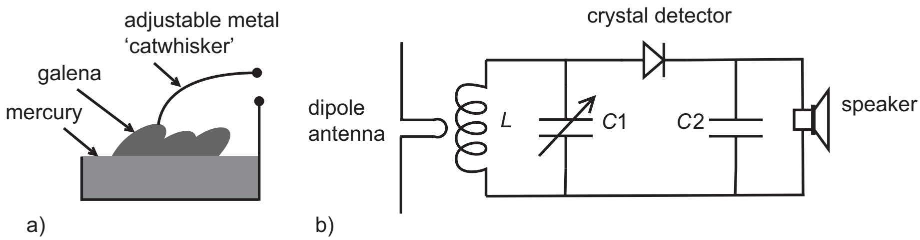

Fig. 2.15 A point contact diode and a receiver based on this device.

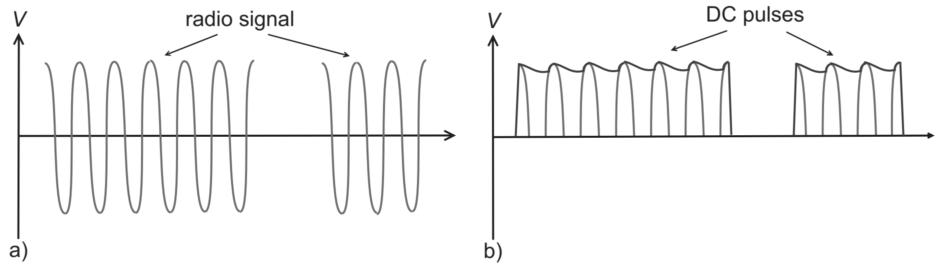

Fig. 2.16 Demodulation of radio frequency pulses by a crystal detector.

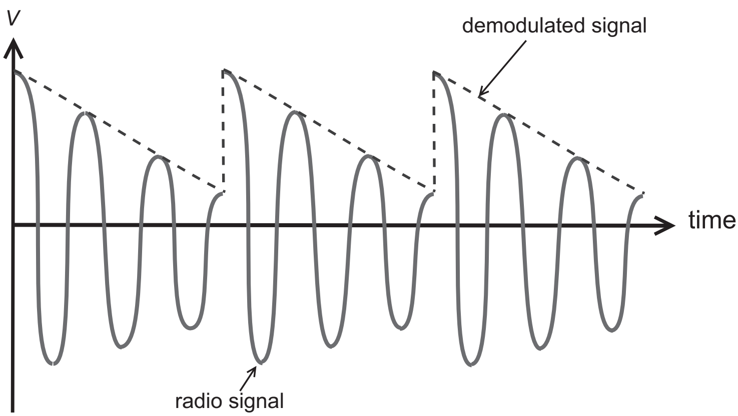

Fig. 2.17 Demodulation of an amplitude modulated signal by a crystal detector.

shown in Figure 2.15b. The field of the radio wave causes a current to flow in the antenna and hence an EMF in the circuit consisting of inductor L and capacitor $C1$ (set to resonate at the frequency of the desired incoming radio wave). The crystal detector only allows current flow in one direction and so, after smoothing by capacitor $C2$, the radio signal presents itself as a pulse at the speaker (see Figure 2.16). This DC pulse will then be detected by the radio operator as a click in the speaker.

The introduction of the crystal detector proved to be a boon for radio since it not only increased sensitivity, but also introduced a capability to demodulate signals with

https://doi.org/10.1017/9781108684514.003 Published online by Cambridge University Press

amplitude modulation. Because the output of a spark transmitter consists of a sequence of damped oscillations, it turns out to be already amplitude-modulated. The solid curve in Figure 2.17 shows typical behaviour of the amplitude of a signal that was produced by a spark transmitter and the broken curve shows the demodulated signal that presents itself to the speaker. Fortuitously, the modulation occurs at audio frequencies and this means that the output at the speaker will be a tone. Consequently, the signal from a keyed spark transmitter will consist of a sequence of tone pulses after crystal detection. This is far easier on the radio operator than a sequence of clicks.

2.6 Conclusion

In the current chapter we have developed the basic theory of radio waves, their generation, propagation and detection. Further, we have looked at some of the early developments in radio technology (see Garratt (1994) and Lee (1985) for further historical detail). Crucial to this technology is the idea of restricting the radio signal to a narrow band of frequencies in order that many users can be accommodated within the radio spectrum. In early radio, this was achieved through simple tuned circuits that were resonant at the desired frequency. However, the subsequent development of radio has required frequency selection at a much higher fidelity and so we digress in the next chapter to consider how this can be achieved.