RADIO WAVES

What makes them go

by "Cathode Ray"

Someone who read what I had to say in the September and October 1974 issues, on magnetism being a side effect of electricity (and it's nice to know that at least one person did so) asked me if I would care to go on and deal with electromagnetic waves, and in particular to derive from first principles their velocity and "the impedance of space."

There are three approaches to electromagnetic (or radio) waves. Those of you who are well versed in vector analysis and three-dimensional differential equations will no doubt follow in the footsteps of Clerk Maxwell, with the advantage over him of being able to see and hear the multifarious practical results now obtained with the waves that Maxwell predicted mathematically. Others will be content to enjoy those results, without any overwhelming urge to inquire into their theory. Members of these two classes may now disperse and employ their time more profitably elsewhere, as I am about to address myself exclusively to any who do wish to know what makes electromagnetic waves go, but lack the mathematical expertise needed for taking Maxwell's way.

Electromagnetic (e-m) waves consist entirely of electric and magnetic fields. Most of us are more at home with circuits, amps and volts than with fields. Transmission lines (or high-frequency cables) offer themselves as a bridge from one to the other. So let us adopt that way of approach to free-travelling e-m waves.

In ordinary circuits, resistance, inductance and capacitance are regarded as if they were confined to the places indicated by their symbols in the circuit diagram – the components. The rest of the circuit – the wiring – is there just for connecting up the components, and not for contributing any R, L and C of its own. In so far as these qualities are inevitably present to some extent in the wiring, they are just unwanted complications which we hopefully neglect.

If transmission lines are regarded in this way, as they well might be by an electrician, they look like just wiring, needed to connect units that unfortunately have to be installed at a distance from one another; the radio-frequency counterpart of the flex needed to connect the TV set to the mains socket. It is true that in both of these types of electrical link we would like the resistance to be small enough to neglect. If the resistance of the flex is enough to cause a noticeable loss of volts at the appliance, a heavier gauge of wire is indicated. Transmission lines, being in general much longer than flex leads, their resistance usually does cause appreciable loss en route. But at least they do not (as would too-resistive a flex) constitute a fire risk!

At the mains frequency the inductance and capacitance of a few yards of flex are truly negligible. But at the multi-million times greater frequency of the incoming signals, and with the greater length of the cable, one would quite rightly estimate that the capacitance between its two conductors would be very far from negligible, and perhaps expect this capacitance to be almost a short-circuit for the signals, allowing very little to reach the receiving end. But in spite of the two conductors being so close together – in the common coaxial type, one actually surrounding the other – so that the magnetic effects of currents in them tend to cancel out, there is enough inductance to have a profound effect. Just as the total capitance is distributed in parallel all along the line (assumed to be uniform), so the total inductance is distributed in series. Electrically, the line can be represented as in Fig. 1,

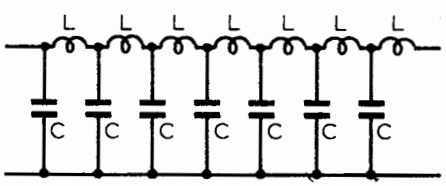

Fig.1 An ideal transmission line can be considered as a circuit in which the distributed capacitance and inductance are represented as a very large number of very small capacitors and inductors.

where L and C are respectively the inductance and capacitance per (very small) unit length of the cable.

Next, let us consider the effective resistance of the TV set or whatever the cable is feeding into, R in Fig. 2(a). It probably won't be a pure resistance, but it can always be made so by tuning; and that gets rid of one complication. Now we connect to it one of our very short unit lengths of cable, Fig. 2(b). It is so short that the series inductive reactance X_L, which is $2\pi fL$, is very small compared with R; and the parallel capacitive reactance X_C, which is $1/2\pi fC$, is very large compared with R. That being so, the series capacitance C' in Fig. 2 (c) is (near enough) electrically equivalent to C in (b) if its reactance, X'_C is equal to R^2/X_C.* X_L and X_C', being respectively positive and negative reactances, cancel out if $X_L' = X_L$. Fig 2(c), and therefore very nearly (b) also, is electrically the same as (a). For this to be true, R^2/X_C must be equal to X_L, so

$$R^2 = X_L X_C = \frac{2\pi fL}{2\pi fC} = \frac{L}{C}$$

$$\text{So } R = \sqrt{\frac{L}{C}} \qquad (1)$$

Notice that frequency doesn't come into this at all, except that if it is very high

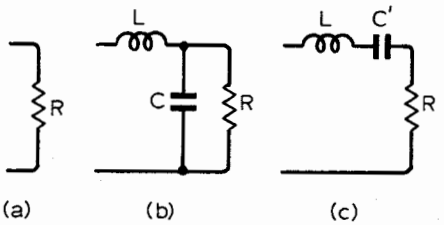

Fig.2 By selecting a suitable value of load resistance R, a pair of the small units of L and C in Fig.1 connected to it (b) can be made of no effect, because the series equivalent of C (C' at (c)) cancels out L. This process can be repeated until any length of line terminated by R is found to be equivalent, as an impedance, to R alone.

then L and C have to very small indeed to fulfil the condition that $X_C \gg R \gg X_L$

The process of finding a Fig. 2(b) equivalent to (a) can then be repeated indefinitely, so that any length of cable terminated by a resistance is electrically the same as the resistance alone, provided that eqn.(1) is true, subject to the approximation we used. The smaller X_C and X_L are, the smaller is the error in assuming $X_C' = R^2/X_C$, so by making them smaller and smaller and increasing their number correspondingly, ultimately making Fig. 1 equivalent to a real cable, we can make the error as near zero as we like.

So to make a line or cable ideal for

*Foundations of Wireless & Electronics, 8th edition, M. G. Scroggie; Sec. 8.16. See also Sec.16.2.

conveying radio signals from one place to another we have to ensure that its inductance and capacitance per unit length (not necessarily small at low frequency) are related to the load resistance R thus: $R = \sqrt{(L/C)}$. L and C depend on the cross-sectional dimensions of the cable, and there is only a limited range of practical values of these, so it is usual to fit R to them, rather than $\sqrt{(L/C)}$ to R. The resistance $\sqrt{(L/C)}$ is usually called the characteristic resistance of the cable and denoted by R_0. (It is also called characteristic impedance and denoted by Z_0; this covers the fact that the effect of *resistance* of the line conductors introduces *reactance* along with R_0.). Owing to the way in which it was derived with the help of Fig. 2, R_0 can also be regarded as the input resistance of an infinitely long line.

How does one find L and C? Well, of course, they can be measured. Calculating them for practical lines and cables (parallel-wire or coaxial) is rather complicated. But although not a very practical form, there is no theoretical reason why a transmission line should not consist of two parallel metal strips as in Fig. 3. This is much easier for calculating L and C at least

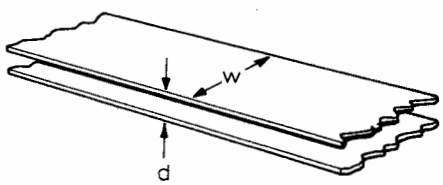

Fig. 3 A simple, if unusual, form of transmission line is a pair of parallel metal strips.

approximately, and, as we shall see, it is a very convenient form for studying the whole subject.

Although thinking a thing out from fundamental principles is usually harder work than remembering a handy formula or looking it up in a book, it should be worth it in this case. So we start with the standard definition of the capacitance between two conductors as the electric charge (positive on one conductor; negative on the other) per volt needed to put it there:

$$C = \frac{Q}{V} \qquad (2)$$

(Until further notice I'm going to use the symbols C and L in a general sense; not per unit length as in Figs. 1 and 2.) And the inductance of a circuit or part thereof is defined as the voltage induced in it per amp-per-second variation in the current flowing through it:

$$L = \frac{V}{dI/dt} \qquad (3)$$

Although defined thus, capacitance and inductance are really effects of electric and magnetic fields respectively, and we won't be able to get far without

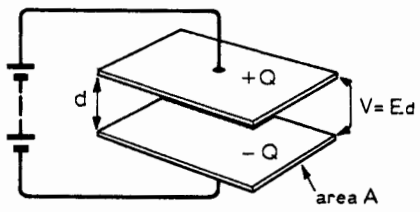

Fig.4 A short length of the Fig.3 type of line can be treated as a capacitor.

accepting that fact. Electric field is denoted by E, and is the voltage per metre between two points at different potential (Fig. 4). So

$$E = \frac{V}{d} \qquad (4)$$

That connects with the V in (2). Associated with E is what is called electric flux density or displacement, D, which is equal to the charge on the plate per unit area:

$$D = \frac{Q}{A} \qquad (5)$$

(This equation is a form of Gauss's theorem.) D *and* E are related to one another by a property of whatever material or non-material fills the space between the plates – its permittivity, ϵ:

$$D = \epsilon E \qquad (6)$$

So, substituting for Q and V in (2), from (4)-(6), we get

$$C = \frac{AD}{dE} = \frac{A\epsilon}{d} \text{ farads} \qquad (7)$$

which I hope you will recognise as the well known formula for the capacitance of a parallel-plate capacitor in SI units. It would be very inaccurate for a capacitor like the one in Fig. 4 because there would be a lot of stray field besides that directly between the plates; this is usually referred to as edge effect, and is much less if (as in practice) the plates are very close together. Obviously we can apply (7) to calculating the capacitance between the strips in Fig. 3, per small length, or per metre, or for the whole

But now let us go back to inductance, eqn. (3). A coil has an inductance of 1 henry if 1 volt is induced in it when the current is changing at a steady rate of 1 amp per second. But what induces the e.m.f. is not the varying current itself but the varying magnetic flux due to the current and linked with the coil. In SI units the voltage induced is equal to the rate at which the flux is changing, so if 1 amp in the coil causes Φ units of flux the inductance is equal to Φ. In other words, the inductance is equal to the flux per amp:

$$L = \frac{\Phi}{I} \qquad (8)$$

We can make a sort of coil of Fig. 3 if we short-circuit a length W at both ends and circulate a current I around this "coil". The flux passes through the "core" of this coil and doubles back to complete a loop around the current, as indicated in Fig. 5 by only two dotted lines, which represent the continuous flux filling the whole core. Its conventional direction, for a clockwise current, is inwards as shown (corkscrew rule). The flux density, denoted by B, is equal to Φ divided by the cross-sectional area A of the "window" inside the coil:

$$B = \frac{\Phi}{A} \qquad (9)$$

This A is not the same as in (5) of course, but in this case is equal to Wd. The current itself is also involved because the magnetic field strength H inside the coil is equal to the encircling current per unit length of the Φ path (Ampere's law):

$$H = \frac{I}{l} \qquad (10)$$

But B and H are related to one another by a property of the material or non-material in which they occur – its permeability, μ:

$$B = \mu H \qquad (11)$$

So, substituting for Φ and I in (8) we get

$$L = \frac{A\mu}{l} \qquad (12)$$

Strictly, except where H is constant all the way l around its loop (which is very rarely) Hl in (10) must be $\int H.dl$;

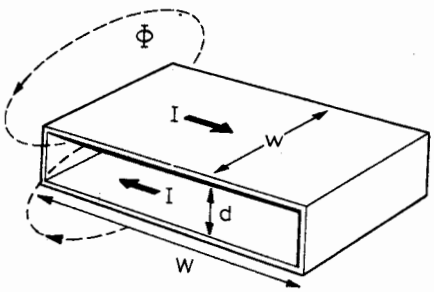

Fig.5 By adding short-circuiting pieces at the ends, the bit of line in Fig.4 can be made into an elementary coil.

but in Fig. 5 the external path of H (and Φ) is so much fatter than the internal one that, provided d/w is small, the only part of l that need be counted is the internal part, w. So approximately in this case.

$$L = \frac{A\mu}{w} \text{ henries} \qquad (13)$$

This is analogous to (7), but we must remember that A means different things in (7) and (13). Incidentally, the approximation error ("end effect") in (13) due to w not being the whole path of B is analogous to the "edge effect" in (7); but whereas (7) gives too low a value for C, it should be fairly obvious that (13) gives too high a value for L.

Now at last we can combine (7) and

(13) to find approximately the characteristic resistance R_o of the Fig. 3 form of line. To simplify matters and get rid of the ambiguity of the symbol A we shall take a section of line 1 metre long, so that A in (7) is equal to w, and A in (13) is equal to d. If you object that 1 metre is not short enough to be valid in the argument based on Fig. 2, my reply is that it is good enough for relatively long waves (low frequencies) and to suit high frequencies you can reduce the scale. And if you point out that the line doesn't have short-circuits every metre along its length as in Fig. 5, the answer is that it doesn't need to, as current can flow freely along both strips, and is equal and opposite in them, just as in Fig: 5. So for Fig. 3, remembering that L and C are now per unit length again,

$$R_o = \sqrt{\frac{L}{C}} = \sqrt{\frac{\mu d}{w} / \frac{\epsilon w}{d}} = \frac{d}{w}\sqrt{\frac{\mu}{\epsilon}} \qquad (14)$$

The values of μ and ϵ for air are almost the same as for a vacuum, μ_o and ϵ_o, which are $4\pi \times 10^{-7}$ and 8.854×10^{-12} respectively. So $\sqrt{(\mu/\epsilon)} = 377$. For example, if the width of the strips in Fig. 3 was 10 times their separation, R_o for this line would be 37.7Ωapproximately (owing to edge and end effects it would be rather less).

The R_o of a line or cable terminated by a resistance equal to R_o being equivalent to that same resistance so far as any generator connected to it is concerned, R_o is also the ratio of voltage to current at the point of connection and (as will be clear from the argument illustrated by Fig.2) at every point along the line, right up to the load. This is practically so even if the line loss moderately reduces the actual values of V and I between generator and load.

The fact that a low-loss line with suitably chosen L and C is electrically equivalent to the resistance connected to the far end does not, of course, mean that the generator signal arrives at the far end instantaneously. When the generator (such as an aerial) feeds the first positive half-cycle into its end of the line, it starts to charge the capacitance of that end, say one of the C units in Fig.1, but the current that tries to go on from there to charge the next unit is delayed by the first series inductance L. And so on. So the signal waveform travels along the line at a certain speed, rather like the wave one can make by waggling the end of a long stretched rope.

What speed?

We can look again at Fig.3 and, denoting the voltage and current at the start by V and I (V/I being R_o) we calculate the charge on the upper strip (the lower one being assumed earthed) *per unit length*, from (2) and (7):

$$Q = CV = \frac{\epsilon w V}{d}$$

The current I along the line is the amount of charge passing any fixed point per second, so if we call the velocity of the charge along the line v, we have

$$I = Qv = \frac{\epsilon w V v}{d}$$

therefore $$v = \frac{Id}{\epsilon w V} = \frac{d}{\epsilon w R_o}$$

and substituting from (14)

$$= \frac{1}{\sqrt{\epsilon \mu}} \qquad (15)$$

In space, where ϵ and μ are ϵ_o and μ_o, this works out at nearly 3×10^8 metres per second, which is the speed of light, usually denoted by **c**. Together with much other convincing evidence, this discovery led to the conclusion that light is electromagnetic, differing from radio waves only in its much higher frequency.

In air, ϵ and μ are very slightly greater than ϵ_o and μ_o, so the wave speed is very slightly (negligibly for most purposes) lower. But practical lines have to rely on solid insulating spacers with values of ϵ several times greater than ϵ_o, so the wave speed therein may be much less than **c**, and the wavelength along the line, at a given frequency, much less than in air.

Note that as $V/I = R_o$ everywhere along the line, voltage and current are everywhere in phase, so they carry energy along the line, stored in the travelling electric and magnetic fields. Comparing Figs. (4) and (5) we see that these fields must be at right angles to one another and to the direction of wave motion. I don't want to get sidetracked here by the subject of polarization but just mention in passing that the direction of polarization is conventionally that of the *electric* field; vertical in Fig.3. That is why receiving dipoles for vertically polarized waves should be vertical. But e-m waves don't *have* to be like this, polarized in one direction ("linearly polarized"); they can be all mixed up.

Fig.6 shows diagrammatically the electric (E) and magnetic (H) field

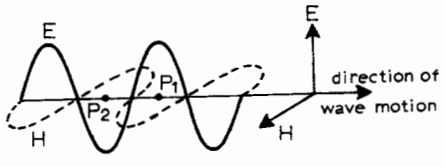

Fig.6 Plane, linearly-polarized electromagnetic waves consist of an electric field pattern, shown here in the vertical plane, accompanied by a similar magnetic field at right angles to it and to the direction in which the whole pattern is moving.

strengths in three-dimensional space between the conductors of a transmission line when vertically-polarized sinusoidal waves are going along it from left to right. If either E or H were reversed in phase, the waves would be going from right to left.

All this may be all very well, you may say, but when are we going to get free from lines and cables? How can the waves exist without charges or currents? Well, we know (I hope) that although each of the imaginary flux lines between the capacitor plates in Fig.4 begins on a positive charge and ends on a negative charge, and E is inevitably present around any charge, charges are not the only cause of E. The other cause is variation of a magnetic field (Faraday's law of e-m induction)*. It happens in every power station and transformer. The basic principle of electric generators is $V = Blv$, V being the voltage generated in a straight conductor of length l cutting a magnetic field of flux density B at velocity v. But even if the conductor were not there the potential difference in space would be, and p.d. is a measure of the electric field between the ends of the length l, because over a length l it adds up to V. So a more fundamental equation is $E = Bv$. Or in terms of magnetic field strength H, $E = \mu Hv$.

But how does a magnetic field come into existence where there are no electric currents? Even the field around a permanent magnet is caused by electric currents on an atomic scale in the magnet material. We go back to eqn.(10) for the basic principle (Ampere's law). Put more correctly it says that the magnetomotive force (m.m.f.), $\int H.dl$, is equal to I, the current enclosed by an H loop of length l. Now look at Fig.7, which shows a capacitor C

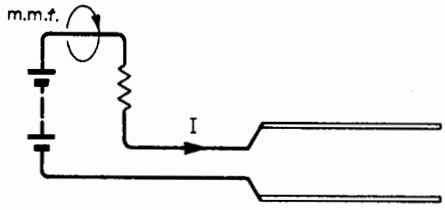

Fig.7 Although the space between capacitor plates carries no current in the ordinary sense, there is a "displacement current" which produces the same magnetomotive force around it as every other part of the circuit.

in process of being charged. The charging current is flowing in the direction conventionally shown by the arrow I, and of course is diminishing as time goes on. Exactly proportional to this current at all times, and numerically equal to it in SI units, is an m.m.f. everywhere around the current, indicated at one point in the circuit by a ring. The arrow head on this ring is conventionally related to the current arrow by the corkscrew rule. Having recapped on the familiar circuit situation, let us shift our attention to the space between the plates, which are

*Discussed in "What is e.m.f.?", August 1974 issue.

wider apart than usual in order to make this easier.

It will be generally agreed that no current, in the ordinary sense, is flowing across the space between the plates. There is, as one would say, a break or gap in the circuit. But is there a gap in the m.m.f.? I've never done the experiment, but I'm sure that if a magnetic needle were to be suspended near the plates, with precautions to prevent it from being affected by the rest of the circuit, it would respond to this non-existent current. For I trust James Clerk Maxwell, who decided theoretically (and, for all I know without being able to look it up, experimentally) that there is indeed an m.m.f. around the space between the plates, caused by what he called *displacement current*. We have come across displacement already, in eqn.(5), as the electric flux density between opposite charges. The total displacement or flux over an area A is therefore AD, and, as (5) said, this is equal to Q, the total charge on either plate. The circuit current I is equal to the rate at which charge is moving along, but in the capacitor this charge is not moving along but is accumulating on the plates. However, it makes the displacement increase. The rate of increase of total displacement being equal to the rate of increase of charge, it is also equal to the circuit current, I. So if displacement current is defined as the rate of change of total displacement, it is always equal to I. So the m.m.f. ring around C is the same amount as around I anywhere else.

At a fixed point P_1 in Fig.6, E is at this moment at a positive maximum and H likewise (if positive is towards us). At P_2 they are both maximum negative. The rapid change from E at P_1 to $-E$ half a cycle later was supposed to be due to a negative charge on the upper metal strip being replaced by a positive charge, brought about by current along the strip between P_2 and P_1. But if we remove the strips and join up the opposite E lines at P_1 and P_2 into complete loops as in Fig.8, the movement of this loop rightwards for half a cycle is accompanied by a displacement current in space where the conduction current used to be. This rapid change in displacement D causes a magnetic field H just as any ordinary current would. We have already noted that $E=Bv=\mu Hv$. Without going into the full derivation we can now be pretty sure (by the principle of duality I so often cite) that the counterpart is true: $H'=Dv=\epsilon Ev$. I have called the generated magnetic field H' to distinguish it from H, as in general the two are not necessarily equal. But for the two fields to keep one another going, H' must be equal to H, which in the first equation is equal to $E/\mu v$. Substituting this in the second, we get

$$\epsilon Ev=\frac{E}{\mu v}$$

from which

$$v=\frac{1}{\sqrt{\epsilon\mu}} \qquad (16)$$

which is the same as (15) derived from currents and charges.

Since e-m waves are thus able to get along quite nicely without currents and charges, what exactly is the role of the hardware, especially as its resistance weakens the waves by a few dB per 100 metres? The quick answer is that it guides them from A to B, when that is what is wanted rather than broadcasting. But how?

An air-cushion-shaped wave like Fig.8 has parts at top and bottom that are not wholly vertical. These will therefore expand upwards and downwards as well as forwards. In three dimensions it will expand sideways as well, and in fact all around. (The same applies to the magnetic field, not even suggested in Fig.8 but there in real waves.) If (say) an electric wave front expanding upwards hits a horizontal conducting surface, the field lines will not be at right angles to it. So they will have a component parallel to it, along the surface of the metal. But it is impossible for two points in or on a perfect conductor to be at different potentials. So where an electric field ends on a conductor the direction of the field must be wholly at right angles to the conductor.

That being so, the wave front inside a transmission line must be a plane at right angles to the conductors and to the direction of wave travel. Which, not surprisingly, is why the waves are called plane waves. The conductors eliminate all field components of the waves that are not directly forward. (I have said nothing about the magnetic field, because it is obvious that if any part of the electric field is eliminated the corresponding part of the magnetic field has nothing left to keep it in existence.)

It is not essential to have two conductors for this guiding action; in certain circumstances an empty tube will do, called a waveguide. But that is too long a story to start on now. If anyone asks me kindly I might tell it some other time. But I do just have room to fulfil my promise about the impedance of space. We found that a transmission line or cable that is loss-free and infinitely long has an input resistance that is

$$R_o=\sqrt{\frac{L}{C}}=\frac{d}{w}\sqrt{\frac{\mu}{\epsilon}} \qquad \text{(14 again)}$$

The awkward bit about being infinitely long can be got round by substituting any length you like provided that the far end is connected to a resistance equal to R_o, nobody will notice the difference at the input end because there won't be any difference there. If we imagine ourselves inside an enormous line of the Fig.3 type, looking towards the far end, we can consider one metre square of the cross section of space confronting us. By thus making $d=w=1$ so far as the space is concerned, we get $\sqrt{(\mu/\epsilon)}$ as the resistive impedance of space (since it is not concerned at all with the dimensions of the line). Let us call this resistance R_S. In fact, the line can be removed and, provided the waves stay plane, which they will then not do, but will very nearly do at a great distance from a radio transmitter, the same applies. We have already noted that the value of $\sqrt{(\mu/\epsilon)}$ for empty space is 377Ω. Within dielectric materials μ is hardly affected, but ϵ will be greater, so the resistance of the material to plane waves will be less than 377Ω.

If we work back from $R_S=\sqrt{(\mu/\epsilon)}$ by using equations (14), (4) and (10) (with $l=w$) and $R_o=V/I$, we get

$$R_S=\frac{E}{H}$$

which is analogous to

$$R_O=\frac{V}{I}$$

The dimensions are right, because E is in volts per metre and H is in amps per metre, and the metres cancel out.

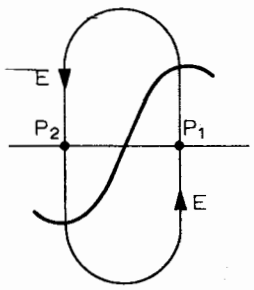

Fig.8 Although Fig.6 was based on the existence of electric charges moving along conductors (Fig.3), waves continue when the conductors are removed, because the fields join up to replace them, as far example this electric line of force at the P_1-P_2 section in Fig.6.