PROPAGATION

There are various ways by which a radio wave may travel from a transmitting aerial to a receiving aerial. The more important of these are: the ground wave, the sky wave and the space wave.

The ground wave can exist when the transmitting and receiving aerials are close to the surface of the earth and are vertically polarised. This wave supported at its lower edge by the presence of the ground, is of practical importance at broadcast and lower frequencies (i.e. below about 2 Mc/s).

The sky wave represents energy that reaches the receiving aerial as a result of the bending of the wave path introduced by the ionisation in the upper atmosphere. This ionised region is called the ionosphere and accounts for nearly all long distance radio communication.

The space wave represents energy that travels from the transmitting to receiving aerials in the earth's troposphere, i.e. the portion of the earth's atmosphere in the first ten miles adjacent to the earth. The space wave commonly consists of two components, one of these is a ray that travels directly from transmitter to receiver, while the other is a wave that reaches the receiver as a result of reflection from the surface of the earth.

Radiation from an aerial

Radio waves represent electrical energy that has escaped into free space. Radio waves are produced to some extent whenever a wire in open space carries a high frequency current. The laws governing such radiation are obtained by using Maxwell's equations to express the fields associated with the wire.

The calculation of the field radiated from even such a simple structure as a dipole, is a complicated procedure. It is possible however to discuss the form of the result.

Let the dipole be a thin conductor of length l, carrying a uniform alternating current which in this case is most conveniently expressed in its complex form

$$I_o \, e^{j\omega t}$$

It can be shown that the magnetic field in free space

$$H_{\phi} = \frac{I_o \, l \sin \theta \, e^{j\omega t}}{4\pi} \left(\frac{j\omega}{cr} + \frac{1}{r^2} \right)$$

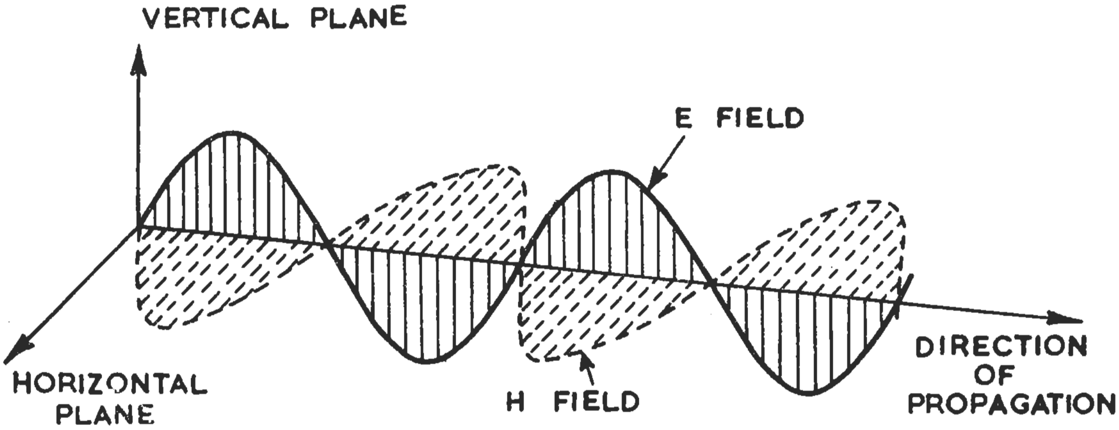
where c is the speed of light,
and r is the distance from the dipole.

It will be seen that the magnetic field is composed of two terms, one varying inversely as r, the other inversely as r^2. When r is small the second term will predominate and it represents what is called the induction field or near field. When r is large the second term becomes negligible compared with the first, which represents the radiation or far field.

The results obtained here clarify two very important points.

Firstly, it is useless to try to measure the radiated field from an aerial at a distance that is less than 5 times the wavelength from the aerial, because of interference from the induction field. Secondly it explains how the radiated magnetic field is in phase in time with the electric field and 90° out of phase with the induction field. The phase change occurs because of the operator j in the function $\frac{j\omega}{cr}$.

The radiated field is the field that is used in communications.



Diag. 8.1

The electric and magnetic fields are in phase in time, but in quadrature in space.

The plane of the electric field is the Polarisation of the e.m. wave.

A vertically radiated signal from a vertical aerial can only be received by a vertical receiving aerial.

The Ground Wave

Ground wave propagation is propagation of r.f. energy along the curved surface of the earth, without using the earth's ionosphere. The ground wave is vertically polarised, because any horizontal component of electric field in contact with the earth is short circuited by the earth. The ground wave induces charges in the earth, which travel with the wave and so constitute a current. In carrying this induced current, the earth behaves like a leaky capacitor and so can be represented by a resistance or conductance shunted by a capacitive reactance. The characteristics of the earth as a conductor can therefore be described in terms of conductivity and dielectric constant. As the ground wave passes over the surface of the earth, it is weakened as a result of energy absorbed by the earth in order to supply the power loss resulting from the induced current flowing in the earth's resistance.

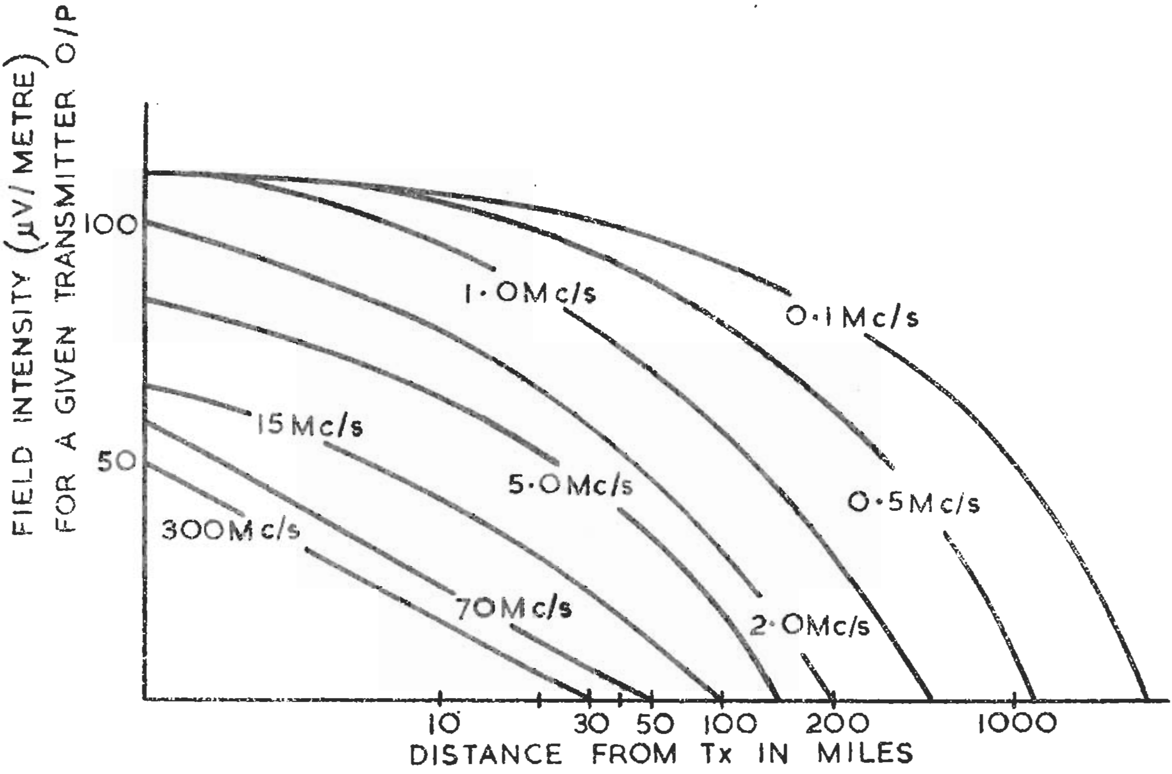
The factors governing the ground wave range are:

(a) The power of the transmitter.

(b) The frequency - the lower the frequency the greater the range.

(c) The nature of the ground. The better the conductor the greater the range. This means that greater ground wave range (for a given power and frequency) can be obtained over the sea, than over the land.

Ground wave communication is used for most broadcasts in the frequency range of under 2 Mc/s, using relatively high powers. Long range communications utilise the VLF (15 - 30 kc/s) and the LF band (30 - 300 kc/s). These bands are relatively free from the effects of disturbances in the ionosphere and can be generally relied upon at all periods of day and night and the seasons. However, at these low frequencies, the power required is enormous and the size of the aerials very large indeed. This limits the use of these frequencies to large static stations.



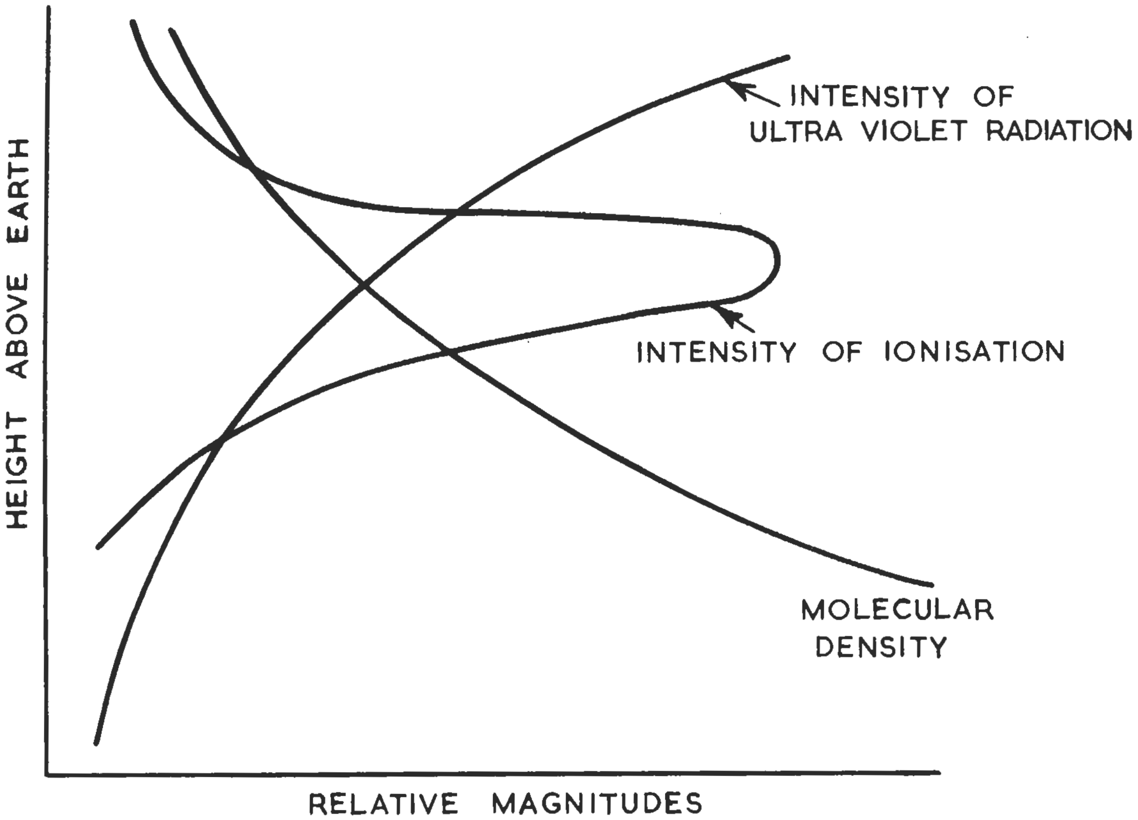
Diag. 8·2

The diagram gives an indication of the range that can be expected over good ground for a given power output of transmitter.

The nature of the ionosphere

The upper parts of the earth's atmosphere absorb large quantities of radiant energy from the sun, which not only heats the atmosphere but also produces some ionisation in the form of free electrons and positive ions.

When an electromagnetic wave strikes an atom, it is capable of moving an electron from an inner to an outer orbit. When this occurs, the electron has absorbed energy from the wave. If the frequency of this incident wave is sufficiently high, such as in ultra-violet waves, an electron may be knocked completely out of an atom. When this occurs, a positively charged atom or ion remains in space together with a free electron. The rate of ion and free electron formation depends upon the density of the atmosphere and the intensity of the ultra-violet wave. As the ultra-violet wave produces positive ions and free electrons, its intensity diminishes. Therefore, the ionised region will tend to form in a layer, forming few positive ions and free electrons due to the less dense atmosphere when the ultra-violet is most intense, forming more positive ions and free electrons due to the more dense atmosphere when the ultra-violet wave is of moderate strength and again forming few positive ions and free electrons due to the low intensity of the ultra-violet wave when in the most dense atmosphere. This relationship between ultra-violet intensity, rate of ionisation and atmospheric density is shown in the Diagram 8.3.



Diag. 8·3

The formation of ions and free electrons is not, in itself, sufficient reason to account for the existence of an ionic layer, because the positive ions and free electrons tend to recombine due to the inherent attraction of unlike charges. The recombination rate is directly related to the molecular density of the atmosphere, because the more dense the atmosphere, the smaller is the mean free path of the electrons. The recombination rate is also directly proportional to the density of positive ions and free electrons. Therefore, as the ultra-violet waves continue to produce ions and free electrons, a free electron density will be reached where the recombination rate just equals the rate of formation. In this state of equilibrium, a free electron density exists for every set of given conditions.

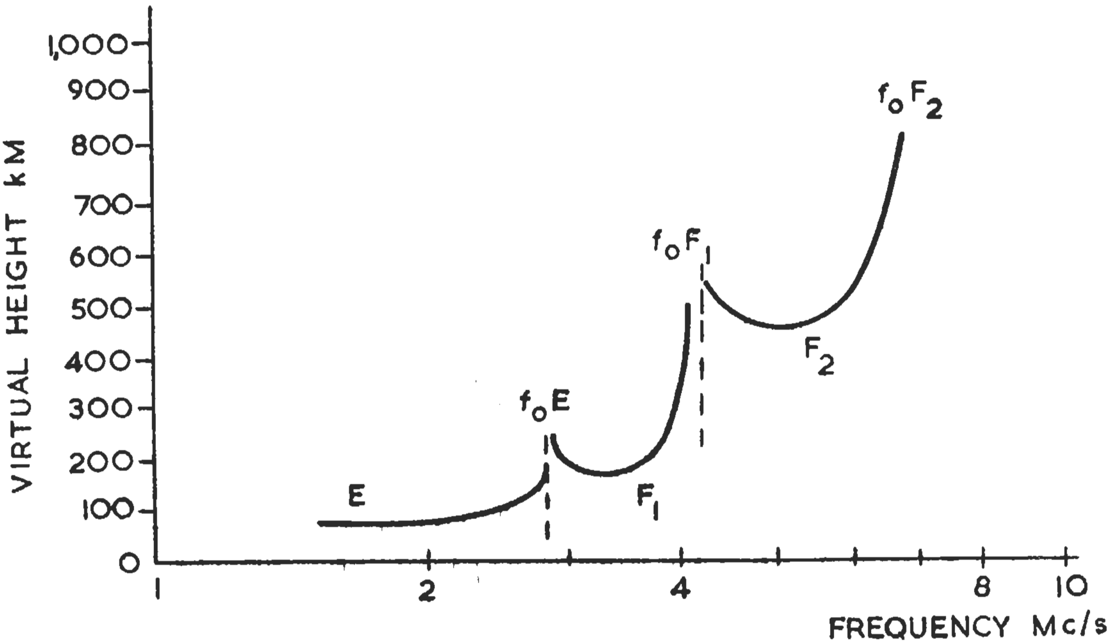
That more than one layer exists is explained by the existence of different ultra-violet wave frequencies. The lower frequency ultra-violet waves tend to produce a higher altitude layer, expending all their energy at the high altitude. The higher frequency ultra-violet waves tend to penetrate deeper into the atmosphere before producing appreciable ionisation. In addition to the ultra-violet waves from the sun, particle radiation, cosmic radiation and meteors all produce ionisation of the earth's atmosphere, particularly in a higher altitude layer.

The three principal layers formed in the daytime are called the E, F_1 and F_2 layers. In addition to these regular layers, there is a region below the E layer which is responsible for much of the daytime attenuation of high frequency radio waves. Called the D region, or layer, it lies between heights of 50 and 90 km. The heights of the maximum density of the regular E and F_1 layer are relatively constant, with only a small diurnal and seasonal change. The F_2 layer is more variable with typical heights lying within the range 200 - 400 km.

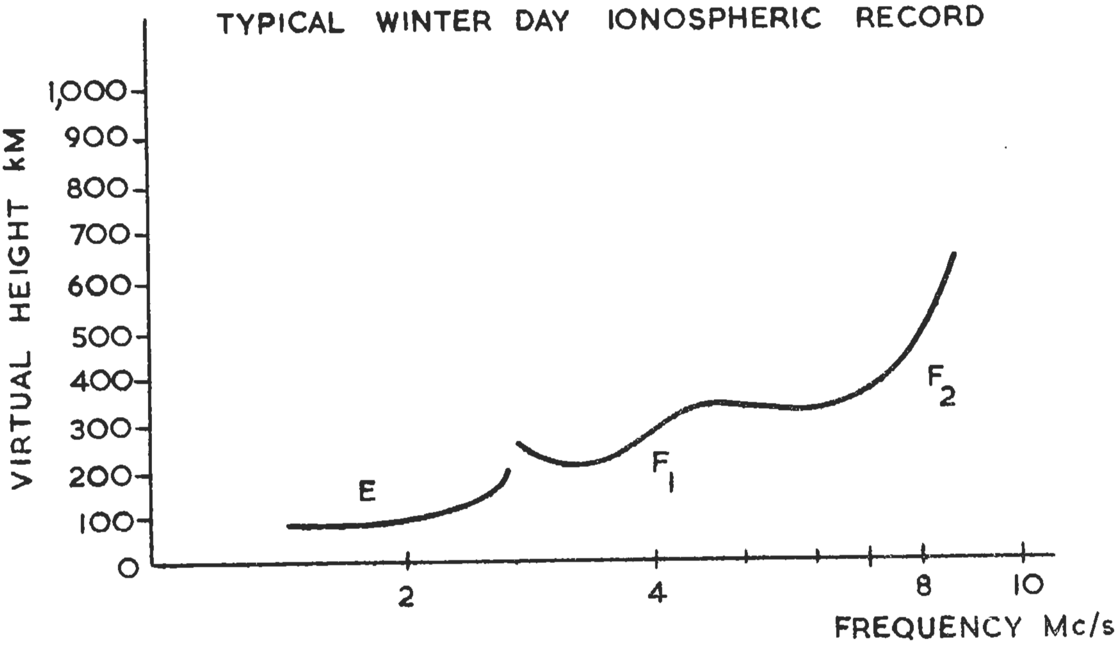
At night the F_1 and F_2 layers join to form a single night time F_2 layer. The regular E layer is governed closely by the amount of ultra-violet radiation from the sun and at night tends to decay uniformly with time. The D layer almost entirely disappears at night due to recombination.

An anomalous ionisation, termed "sporadic E", is often present in the E region in addition to the regular ionisation. "Sporadic E" ionisation usually exhibits the characteristics of patches of intense ionisation, which may appear anywhere in the height range 90 - 130 km. These patches may be from 1 km to several hundred km across. The occurrence of sporadic E is quite unpredictable and although very high frequencies are regularly returned, there is no possibility of predicting the conditions so that they may be used in communications.

TYPICAL SUMMER DAY IONOSPHERIC RECORD



TYPICAL WINTER DAY IONOSPHERIC RECORD



The changes that can occur in the ionosphere can be loosely tabulated.

1. Diurnal changes. Day and night changes in the height and density of the layers.

2. Seasonal changes. These are obviously very much tied in with geographical position.

3. Sporadic 'E'. Patches of intense ionisation in the E layer.

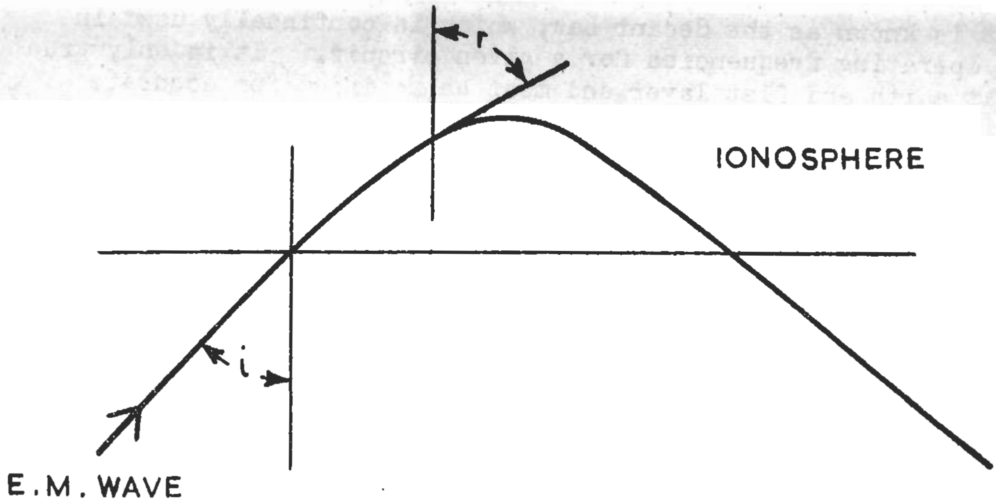
4. Sudden ionospheric disturbances, S.I.D.s or Dellinger fade outs. These are caused by intense ultra-violet radiation given out by the sun during a solar flare. The result of a S.I.D. is a sudden great increase in the ion density of the highly absorptive D region and an increase in the ion density of the moderately absorptive E layer.

5. Magnetic storms should not be confused with S.I.Ds. although they both have the same effect of reducing the probability of communication by sky wave. Magnetic storms are apparently caused by the emission by the sun of particles, these particles are emitted at the same time as the flare, but being of finite weight they take up to about 36 hours to arrive after a S.I.D. A magnetic storm may last for several days with its appearance being very sudden and recovery of the layers to normal being very slow. As the emitted particles are mostly magnetic, the effects of a magnetic storm are most severe in the geomagnetic pole regions. Magnetic storms are more likely to occur during periods of maximum sunspot activity and this occurs in a regular 11 year pattern. At the height of the 11 year sun spot cycle, emission of ultra-violet waves from the sun is considerably greater than during years of "quiet sun". During periods of intense solar activity, ionisation is much greater and much higher frequencies are regularly returned in the HF spectrum. However, communications are more likely to be interrupted by S.I.Ds. and magnetic storms.

Refraction of a radio wave by the ionosphere

When sky wave propagation is used for communication the electromagnetic wave from the aerial is transmitted towards the ionic layer at an oblique angle. The incident wave is then apparently reflected back towards the receiving aerial. Actually the wave is not reflected, though this term is commonly used for convenience, the wave is actually bent back towards the earth by refraction, just as a prism refracts light. As far as propagation is concerned, the all-important effect of ionisation is to reduce the refractive index (μ) of the ionosphere, thus causing the wave (travelling from a medium of high μ to a medium of low μ) to be refracted in accordance with Snell's optical law.

(1) $\mu = \dfrac{\sin i}{\sin r}$ where i and r are the angles (to the normal) of incidence and refraction respectively.



Diag. 8.5

The refractive index of the ionosphere is a function of the frequency, f, of the wave and of a factor D which is proportional to the density of ionisation.

$$\mu = \sqrt{1 - \frac{D}{f_2}} \quad \ldots\ldots\ldots\ldots\ldots (2)$$

where D = 81N and
N is the number of electrons per q.c.

It can be seen that for a given frequency, μ decreases as the density of ionisation increases as the wave approaches the centre of the ionised layer.

From equation (1) at a certain value of ${}_o\mu$, depending on the angle of incidence, sin r = 1 and r = 90°. This makes the path of the wave normal to the earth's radius and further refraction will cause the wave eventually to leave the ionised layer at the same angle as it entered.

In the special case of a vertically incident wave μ must reach a value of zero for the wave to be returned to earth. The highest frequency at which this occurs in a given layer is called the <u>Critical Frequency</u>, f_o, of the layer.

From equation (2)

$$\mu = \sqrt{1 - \frac{D_{max}}{f_o^{\,2}}}$$

$$\therefore f_o^{\,2} = D_{max}$$

where D_{max} is the maximum density of the layer.

From a knowledge of the critical frequency of a layer, the highest frequency at an oblique angle of incidence, i, which will be returned to earth can readily be calculated. Since r = 90°, sin r = 1, and

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

$$\therefore \sin i^2 = \sqrt{1 - \frac{f_o^{\,2}}{f^2}} \quad \text{(and since } 1 - \sin^2 = \cos^2\text{)}$$

$$\therefore \cos^2 i = \frac{f_o^{\,2}}{f^2}$$

and $f = f_o \sec i$.

This is known as the Secant Law, which is continually used in choosing operating frequencies for a given circuit. It is only true for a flat earth and flat layer and must be modified for accurate calculations.

Absorption

A radio wave entering an ionic layer interchanges energy with free electrons and ions. If the ions do not collide with gas molecules or other ions, all the energy transferred to the ionosphere is reconverted back to electromagnetic energy and the wave continues to be propagated with undiminished intensity. On the other hand, where ions and electrons engage in collisions, they dissipate the energy they have acquired from the wave, which results in attenuation of the wave. This attenuation, or absorption is proportional to the product of the number of ions N and the collision frequency f_c. Therefore the attenuation is ordinarily greatest in the region where the product of ion intensity and the collision frequency is greatest. In fact most absorption occurs in the 'D' layer, which lies closest to the earth. Considerably less absorption occurs in the E layer, and very little in the layers above. It follows that absorption is restricted mainly to the hours of daylight and falls to a low value at night. In general, absorption, during daylight hours, increases to a maximum at around 1.5 Mc/s and then decreases rapidly with increase in frequency.

Maximum usable frequency (m.u.f.)

The basis for all frequency planning is the layer critical frequency f_o and measurements are collected systematically by ionospheric sounding centres throughout the world. The practical parameter, however, is the maximum usable frequency or m.u.f. of which there are several definitions. In each case m.u.f. designates the highest frequency which, in a particular set of conditions can be used to propagate radio waves over a given route.

Predictions for f_o and m.u.f. for the lower layers can be obtained from charts for any given sun spot number and a knowledge of the sun's elevation. The m.u.f. applying to the F_2 layer is not so easily determined, because it obeys complex laws and its correlation with solar activity and geographical factors is not so close.

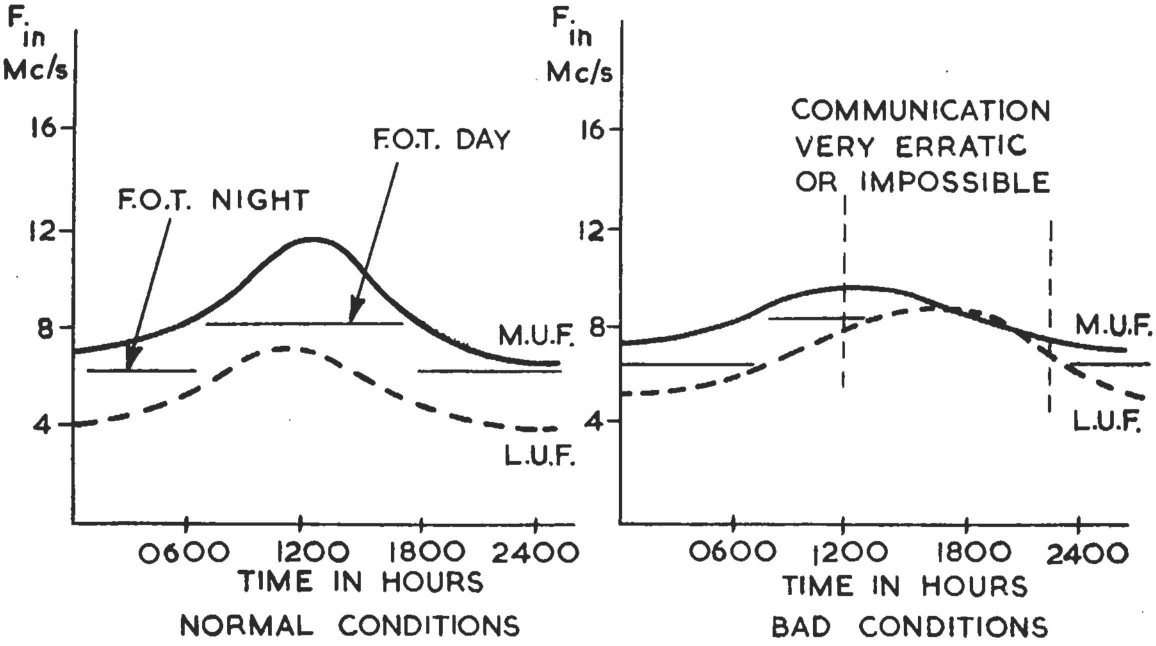
However, figures are published for a maximum usable frequency that will give a reasonable prospect of good communication. The published figures are the Optimum Working Frequencies, o.w.f. and in general this figure is about 85 - 90% of the maximum usable frequency, to allow for slight changes in the ionosphere. The R.N. Signal Orders (S) Series will provide the information that a ship requires when it is necessary to pass a message. The frequency and the shore station to be called are calculated knowing the ship's position, the time and date.

The 'S' orders also give the information required in terms of optimum frequencies, for the ship to communicate with another ship or terminal within a range of 0 - 1600 miles. One of the problems facing the communicator today is to provide gapless cover, out to several hundreds of miles. This problem was dealt with in the chapter on Aerials.

Lowest Usable Frequency (l.u.f.)

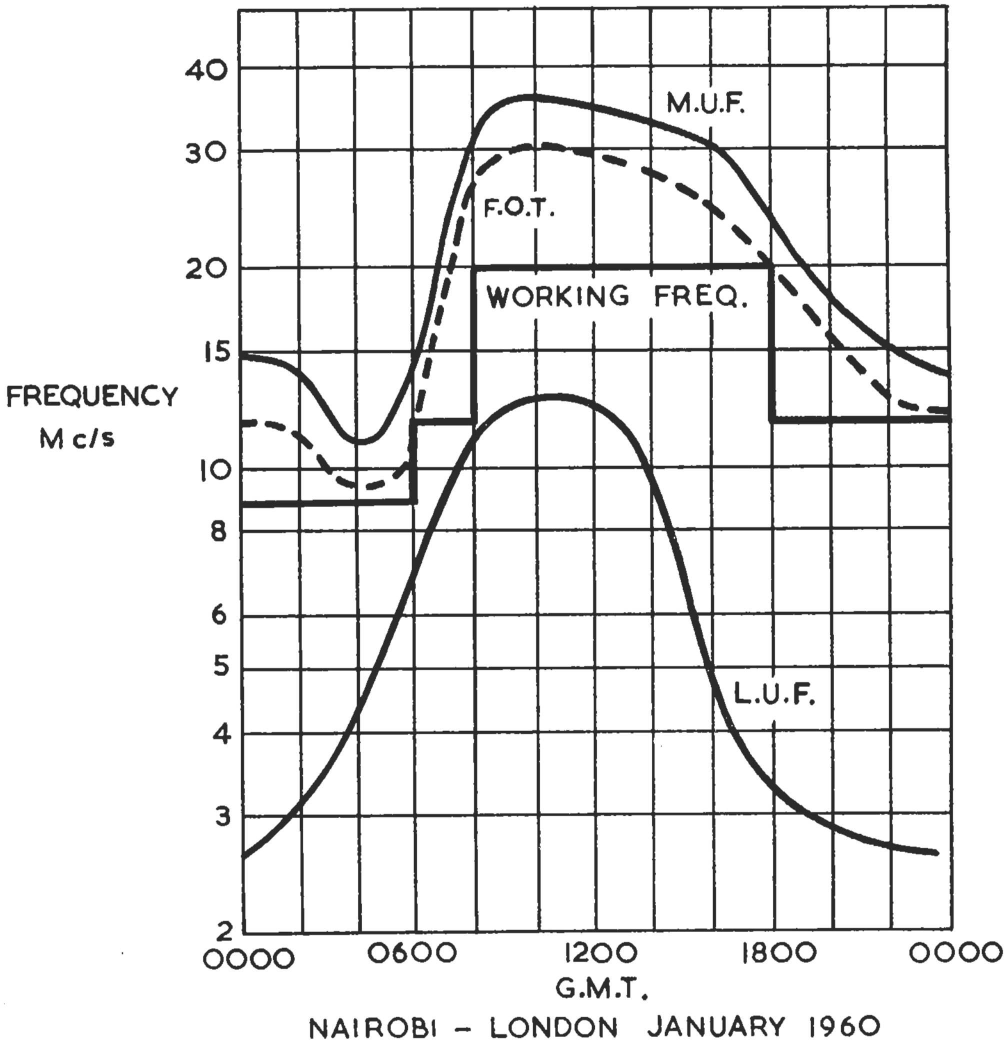
The l.u.f. is the lower limiting frequency which will provide satisfactory communication for a given link. The l.u.f. is the frequency at which the received field intensity just equals the required field intensity for reception. The received field intensity depends upon the receiver aerial system, the path length and absorption and generally increases as the frequency increases. The required field intensity for reception depends on noise limitations at the site and generally decreases with frequency

From the previous paragraphs it can be seen that the m.u.f. is a natural limitation while the l.u.f. is a man made limitation. It must be realised that conditions can arise where the l.u.f. is higher than the m.u.f. under these conditions, propagation is impossible.



Diag. 8.6

Diagram 8.7 shows the diurnal variation on a typical route.



Fading

Fading necessitates the transmission of a great deal more power than would be required in stable conditions. Some forms of fading cause distortion in telegraphy and telephony transmissions, thus making the design of receivers and terminal equipment more complicated and increasing their cost.

The four main causes of fading are:-

(a) Absorption.)
) Slow.
(b) Skip zone variation.)

(c) Wave interference.)
) Rapid.
(d) Changes in polarisation.)

Slow fading covers day to day variations in signal strength, and rapid fading covers those which occur within periods of the order of seconds or less. In practice, the two forms occur together.

Slow fading

Average signal intensity is chiefly dependent upon the length of the radio path and the number of ground reflections but if it may be assumed for a given circuit that these factors remain constant, there will still be daily variations due to absorption in the D layer and, to a lesser extent, in the E layer. Superimposed on these are random daily variations and others related to season and the degree of solar activity.

Fading should not be considered in isolation from noise. Being the result of multipath transmission from regions of high thunderstorm activity, noise is subject to the same variations as other types of signal. In practice, an increase in noise is equivalent to a signal fade, and the two must therefore be considered together.

It follows therefore that in arriving at estimates of the minimum signal-to-noise ratios necessary for various forms of service a slow fading safety factor of about 14 dB is added in every case.

Rapid fading

Because of the irregular and unstable layer structure of the ionosphere, HF signals propagated by this medium tend to separate along a number of simultaneous paths and to arrive at the receiver as a group of independent signals with random phase relationships. Not only does the resulting wave cause wide fluctuations (up to 30 dB or more) in signal amplitude, but the time spread in between the first and last significant arrivals becomes a limiting factor in high speed telegraphy reception.

Polarisation fading is due primarily to a plane wave front being split into randomly polarised waves by the earth's magnetic field. The received field strength can be reduced by as much as 3 dB by polarisation fading.

It is obviously difficult to lay down hard and fast rules for the overall fading safety factor. However, if good communications are required for 95% of the time, it is necessary to increase the system output by a factor at least 32 dB above the level that will produce good communications 50% of the time.

Choice of frequency separation of tones to combat fading

It has been shown in Chapter 2 that to accommodate a 100 baud signal the filter characteristics have been designed to give 170 c/s separation. This choice of frequency separation of the tones also enables the optimum use to be made of the fact that in a two tone telegraph system the full telegraph information is contained in a single tone transmission. The frequency separation can be made such that when one frequency is at a maximum, the second is most affected by selective fading.

If the two signals arrive at an aerial with a path time delay of x milliseconds then they will be in phase at a certain carrier frequency f and also at $f + \frac{1000}{x}$ and in antiphase at $f + \frac{1}{2}\frac{1000}{x}$. Thus it can be seen that if the spacing between f_Z and f_A is made $\frac{1}{2}\frac{1000}{x}$ then 'Z' or idle will be at a maximum when the 'A' or active will be a minimum.

i.e. the time separation is $\frac{1}{2T}$ where T is the time delay in seconds.

In practice the time delays are of the order 1.0 to 1.5 milliseconds and these give separations of the order

for 1.0 mS $\quad \frac{1}{2T} = \frac{1}{2}{}^{10^3} = 500$ c/s

for 1.5 mS $\quad \frac{1}{2T} = \frac{1}{2}{}^{10^3}_{1.5} = 333$ c/s

Now 333 c/s is approximately 2 x 170 c/s and this is now used, so that channelling and tone separation is selected as in Diagram 8.8.