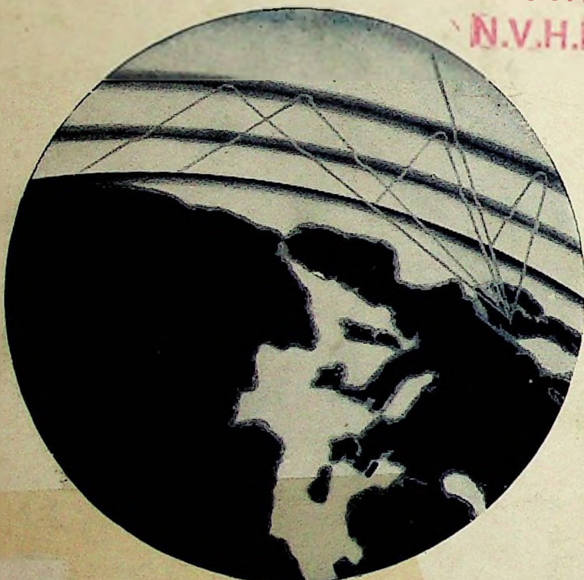
RADIO WAVES AND THE IONOSPHERE



...tific basis of short-wave radio explained in simple ...or all interested in long distance communication.

T. W. BENNINGTON

RADIO WAVES
AND
THE IONOSPHERE

By

T. W. BENNINGTON

Engineering Division, British Broadcasting Corporation

with foreword by

Sir Edward V. Appleton, K.C.B., F.R.S.

Wireless World

ILIFFE & SONS LTD

DORSET HOUSE, STAMFORD STREET, LONDON, S.E.1

BOOK
PRODUCTION
WAR ECONOMY
STANDARD

THIS BOOK IS PRODUCED IN COMPLETE CONFORMITY
WITH THE AUTHORIZED ECONOMY STANDARDS

MADE AND PRINTED IN GREAT BRITAIN
AT THE CHAPEL RIVER PRESS
ANDOVER, HANTS

2/44

FOREWORD

by

Sir Edward Appleton, K.C.B., F.R.S.

SHORT radio waves are great travellers, as every wireless man knows. Not many years ago there seemed something of a mystery about the route they take in spanning continents and oceans in long-distance communication. But recent research has enabled us to unravel much that was previously obscure. In particular, it has emphasized the essential rôle played by the ionosphere in radio reflection, and has shown us why short waves travel such amazing distances without serious loss of strength.

In this volume Mr. Bennington seeks to make us " ionosphere-minded ". Although it is primarily written for the professional radio technician who wishes to understand more about his own subject, I recommend it as a friendly and well-informed guide to anyone interested in long-distance radio communication.

EDWARD V. APPLETON.

PREFACE

THE rôle of the ionosphere in long-distance short wave communication is one of paramount importance; indeed, without it such communication would be impossible. The story of its discovery, of the fitting together of the new knowledge acquired so as to build up a complete picture of its structure, and of the gradual utilisation of this knowledge in practical communication, is one of absorbing interest to the radio engineer and operator, as well as to the physicist and astrophysicist. And it is a story which is as yet by no means complete.

At the present time, with the vast war-time development of radio communication which has taken place, there must be many people who, whatever may have been their occupations and hobbies of a year or two ago, have now a new interest—that of understanding something of short-wave communication and of the part played by the ionosphere in its maintenance. Such individuals are not likely to be highly trained mathematicians and are equally unlikely to have more than a smattering of scientific knowledge at their disposal. It is for these that this book has been written. It is not a "text" book. It contains no mathematics. It merely aims to explain the phenomena in as simple language as is possible—language that should, at least, be understandable to those with an elementary knowledge of radio, or alternatively, to those who have taken "physics" at school. Anyone, therefore, who is capable of reading a simple technical article should be able to read and understand this book, and if he understands it there is occasion for hope that his knowledge of short-wave problems will be considerably improved.

In writing the book the author has, sometimes consciously and often unconsciously, made use of information which has been published before, and he wishes adequately to acknowledge this debt.

In particular, his thanks and acknowledgments are due to the British Broadcasting Corporation for permission to publish information which has been collected and used in the development of the Corporation's Overseas Services, to the National Bureau of Standards of the U.S.A. for various ionosphere data and graphs abstracted from its publications, to Newbern Smith, Theodore R. Gilliland and S. S. Kirby for similar data, to the " Admiralty Handbook of Wireless Telegraphy " for diagrams, to Sir Edward Appleton and his colleagues for information obtained from various of their published Papers, and to T. L. Eckersley for information from his Paper " On the Existence of a Bi-Annual Component in the F Layer Ionisation ".

October, 1943. T. W. B.

CONTENTS

RADIO WAVES AND THE IONOSPHERE

CHAPTER I

GROUND WAVES AND SKY WAVES

RADIO communication over very long distances—whether by broadcasting, point-to-point telegraphy or telephony, or by any other means—is usually carried out on the short waves, that is to say on waves between about 100 metres and 10 metres in wavelength. It is not possible to carry out such long-distance communication on the medium waves (those between about 100 metres and 3,000 metres in wavelength) and, though it is both possible and practicable to do so on the long waves (those over 3,000 metres in wavelength), it is such a costly business that present-day long-distance communication is, for economic reasons, confined almost exclusively to the short waves.

1. Long Distance Communication

These waves are therefore of immense importance in modern communication systems, and it is safe to say that in the future—when international inter-communication will, it is hoped, be even more highly developed than it was before the war—the short waves will be still further exploited as the means of linking closely together every nation and region of the World.

From the foregoing the reader will have gathered that there is something about the short radio waves which makes them peculiarly suitable for long-distance communication, and that they have some characteristic which is not shared by the other waves. This is indeed true, though it is not because of any fundamental difference in the *nature* of the short and of the other radio waves. The difference in their behaviour lies rather in the fact that phenomena which are of little importance when the wavelength is long become of great significance when it is

2. Different Behaviour of Different Waves

reduced, and conversely, phenomena which are of prime importance in the case of the long waves are of little consequence in the short-wave part of the spectrum. Hence, although the nature of the waves is the same throughout the wavelengths usable for radio communication—the only difference being that of wavelength or frequency—the actual behaviour of the short waves during transmission is indeed quite different from those of longer wavelength. And it is this difference in behaviour which makes them so suitable for long-distance communication.

It must not be imagined, however, that there is any sharp dividing line between waves which travel in one way and those which travel in another. That is not so—in fact the very division of wavelengths into long, medium and short is to a large extent an arbitrary one, though for the sake of clarity it is necessary to make some such division. As would be expected from what has already been said, there is in fact a gradual change in the transmission performance of the waves as the wavelength is altered. All that can really be said for the wavebands as defined is that the waves within those bands usually and primarily exhibit the characteristics we associate with long, medium or short waves, as the case may be.

In order to understand why the behaviour of the waves should vary with wavelength we must examine, in a very brief way, something of the nature of a radio wave.

In a book such as this we cannot go into the highly complicated details concerning the exact way in which a radio wave is produced. It would be well, however, to grasp certain fundamentals, as we shall find these necessary in order to understand the reasons for the special behaviour of the short waves.

3. Radiation of Radio Waves

It can be assumed in the first place that a radio transmitting aerial along which an electric charge is oscillating produces lines of electric strain in the space surrounding it. Since it would do this equally as well if it were in a vacuum it is evident that the air surrounding the aerial has nothing to do with the matter. The electric strain lines do not, as a matter of fact, require any material medium to support them—they can exist in " free space ", that is to say in space which contains no material medium whatever.*

Owing to the rapidity with which the charge oscillates, the electric strain lines are continually being " broken off " and losing their contact with the aerial wire. They are accompanied by lines of magnetic strain, which act in a direction at right angles to the electric strain lines. Such a disturbance in space—consisting of electric and magnetic

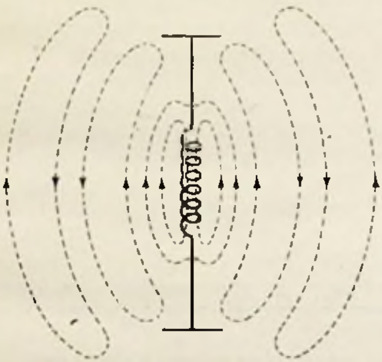


Fig. 1 : Radiation of an electric field from a dipole aerial

strains—constitutes an electromagnetic wave, and this travels through space at its own natural velocity of 300,000,000 metres (186,000 miles) per second. Its velocity through ordinary air is the same as through " free space ".

The outer edge of the advancing disturbance is called the wave front, and in any portion of the wave front near the aerial the direction of the electric strain, the direction of the magnetic strain and the direction of advance of the

* The modern theory discards the concept of an " ether " altogether, but the reader of a book like this will perhaps find that the idea of a medium through which the waves can travel will assist him in understanding their behaviour.

wave are all at right angles to each other. Thus in Fig. 2, where we picture a wave which has just left an aerial, A is the wave front which is advancing toward the right, the direction of the electric strain is vertical and that of the magnetic strain horizontal, i.e., up and down through the paper.

It is important to remember, however, that the wave is not radiated in just one direction only, but—we can assume for the moment—in all directions in the horizontal plane and also at all angles to the horizontal.

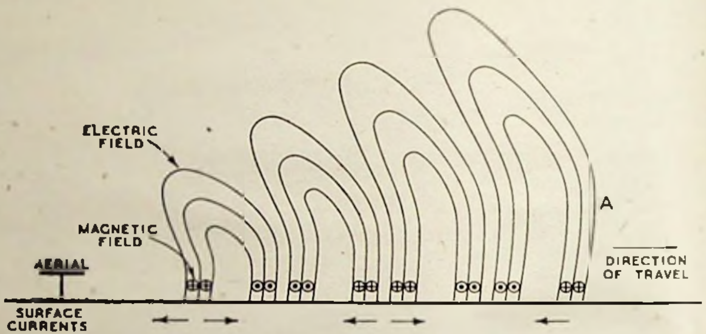


Fig. 2 : Illustrating the elements of the "ground" wave

As the charge continues to oscillate in the aerial electromagnetic waves continue to be radiated from it. The rate at which the waves are emitted will depend on the rate or "frequency" of the electric oscillation which is being fed to the aerial from the radio transmitter. For each complete oscillation of the charge in the aerial one complete wave is emitted.

4. Velocity, Wavelength and Frequency

The velocity at which the wave travels always being the same, i.e., that of light, the wavelength and frequency are connected by the relation :—Wavelength (in metres) =

$$\frac{300{,}000{,}000}{\text{frequency. (in cycles per sec.)}}$$

It is rather important to understand the relation between these three quantities, so let us try and form a visual picture which will help in the matter. Suppose that the radiated waves are visible, and that we are able to watch them coming from the aerial and to follow them as they travel outward from it. We know the rate at which the waves will travel, namely, 300 million metres per second. Suppose the electric charge to be oscillating up and down the aerial 300 million times per second, in which case we say that it has a frequency of 300 million cycles per second. For each complete oscillation—or cycle—of the charge one complete wave is radiated. At the end of a second the wave front will be 300 million metres away, whilst the last wave will just be leaving the aerial, and the distance between the wave front and the aerial will be occupied by the 300 million waves which have been emitted. The distance occupied by each wave will therefore be 1 metre, i.e., the wavelength is 1 metre.

Now let us decrease the frequency of the oscillating charge and cause it to oscillate 100 times more slowly than before, namely at a frequency of 3 million cycles per second. At the end of a second there will now be only 3 million waves occupying the 300 million metres between wave front and aerial. The wavelength will therefore be

$$\frac{300{,}000{,}000}{3{,}000{,}000} = 100 \text{ metres.}$$

We thus see that a wave of low frequency has a long wavelength, whilst a high frequency wave has a short wavelength. We can refer to a wave either in terms of its frequency or of its wavelength, though frequency is perhaps the better designation.

In order to avoid getting confused by the large figures involved when dealing with frequency we can express this quantity in more convenient terms than that of cycles, viz.:

1 Kilocycle per second (kc/s) = 1,000 cycles per second.

1 Megacycle per second (Mc/s) = 1,000,000 cycles per second.

If we now understand the meaning of the terms "frequency" and "wavelength" it will be well, at this stage, to make a classification of the wave-lengths used for radio communication, according to their main characteristics and uses. The classification given in Table I is admittedly a rough one, but a too-rigid classification would only serve to create confusion. Nor, as has already been said, is this strictly possible, for the characteristics of different waves vary considerably from time to time.

5. Classification of Radio Waves

The term " attenuation " used in the Table means the weakening of the wave which takes place as it travels onward, because of the fact that it is continually losing energy for one reason or another as it goes along.

TABLE I

Class	Wavelength Range Metres	Frequency Range Kc/s	Main Characteristics	Principal Uses
Long Waves	Above 3,000	Below 100	Wave travels over earth's surface, i.e., between ground and lower edge of ionosphere. Low attenuation at all times.	Medium- and long-distance point-to-point communication.
Medium Waves	3,000 to 100	100 to 3,000	Wave travels over earth's surface during the day, some energy coming from ionosphere at night. High attenuation during day. Low attenuation at night.	Broadcasting, marine and aircraft communication, direction finding.
Short Waves	100 to 10	3,000 to 30,000	Wave travels up to ionosphere, whence it is reflected back to earth. Conditions for reflection vary greatly with time of day and season. Extremely small attenuation if conditions favourable.	Long distance broadcasting, point-to-point communication, etc.
Ultra-Short Waves	Below 10	Above 30,000	Wave travels over earth's surface only, and for relatively short distances.	Short distance communication, television, aircraft guidance systems.

We have already made the assumption that our transmitting aerial will radiate waves, not only in all directions parallel to the earth's surface, but also in all upward directions as well.

6. The Ground Wave

We should, perhaps, require a rather special form of aerial to do this in practice, but for the moment, let us continue to assume that equal amounts of energy are radiated in all directions, both horizontally and vertically.

If we could stop the radiation an instant after it had started and hold the waves still in space whilst we examined the situation, we should get the picture of a huge hemisphere of radiated energy surrounding the aerial and supported on the ground, with the aerial at the centre of the circle described on the ground by the bottom of the hemisphere of energy.

It is rather difficult to show this in a diagram drawn on

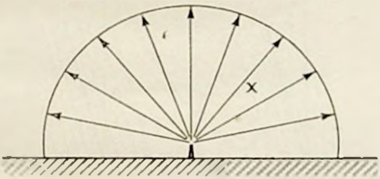


Fig. 3 : Showing direction of travel of some of the " rays " of radio energy

a plane surface, but it will suffice for our purpose to show a section cut through the hemisphere, as in Fig. 3.

The arrows indicate the direction of travel of some of the radiated waves, and it is seen at once that some of them are going out in a very inappropriate direction if they are to reach and actuate a receiver located on the earth's surface.

However, we shall return to these upward-going waves later. For the moment we consider only the waves which are travelling outwards in directions parallel to the surface of the ground. This part of the radiated disturbance is called the " ground " wave, since it remains in contact with the ground throughout its journey. As it travels along it sets up electric currents in the earth itself. These currents cause weakening or " attenuation " of the wave, for energy is taken from it in order to maintain them, and this loss of energy is said to be due to " ground absorption ". The part of the wave which touches the ground is thus constantly losing its energy, though this is being partly replaced by energy present in that part of the wave which is immediately overhead. The upper part of the wave must, therefore, " bend " slightly forwards and downwards in order to replace the energy lost at the foot, and also to follow the curvature of the earth. This slight bending of the wave away from the straight line in which it commences its travel is known as " diffraction ", and by this means the waves can, to a limited extent, follow the earth's curved surface.

But the energy at the foot of the wave is only partly

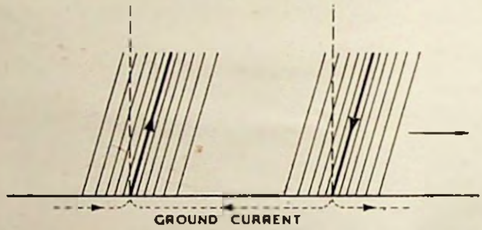


Fig. 4 : Showing forward tilt in the ground wave front

replaced in this way, and, since it goes on inducing currents in the earth and losing more and more of its energy the further it goes, it eventually becomes so greatly attenuated that—to all intents and purposes—it dies away altogether.

There is one interesting point which it may be as well to mention here. As the wave travels onward the electric strain lines remain roughly vertical, as they were when they left the aerial, and a wave such as this is said to be vertically " polarised ". But due to the losses occurring in the earth, the wave acquires a slightly forward tilt, the foot of the wave, as it were, lagging behind the upper part, as in Fig. 4. Thus, at a distance from the aerial the direction of the electric and magnetic strains in the wave front are still at right angles, but the whole disturbance slopes forward from the earth's surface.

Now we come to an important point which will help to explain the reasons for the different performance of waves of different wavelength or frequency.

7. Variation in Performance of Ground Wave with Wavelength

The losses to which the ground wave is subject, because of the earth currents at its foot, besides varying with the nature of the soil or water over which it is travelling, vary also with the wavelength. The faster the wave is going through its complete changes, i.e., the greater its frequency, the

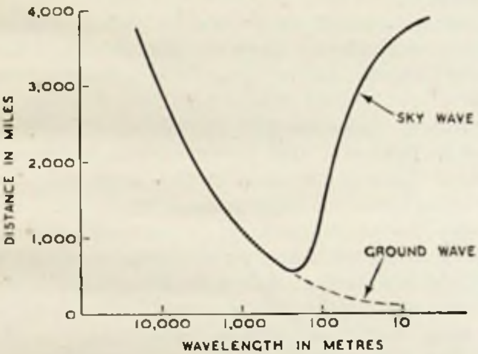


Fig. 5 : Variation of range with wavelength

greater is the amount of energy lost in the earth and the sooner does the wave become completely attenuated and die way. In other words, the longer the wavelength (lower the frequency) the less are the earth losses, and the longer does the wave persist, so that with a given amount of energy radiated the greater is the *range* of the transmitting station.

Hence, when the wavelength is long the ground wave is suitable for really long-distance communication. But as we reduce the wavelength—by increasing the frequency at the transmitter—the ground losses increase, and the range of the station gets less and less. Fig. 5 shows how the range of a station would vary as the wavelength was reduced. On the longest waves the ground absorption is slight and so the range is great. Coming down to the medium waves (below about 3,000 metres), the losses are increasing rapidly and the ground wave range is so severely restricted, that, as was stated in Table I, these waves are only of use for short-distance communication. From Fig. 5, it will be seen that this reduction in range continues—so far as the ground wave is concerned—right down to the shortest wavelengths. Below about 100 metres therefore, the ground wave is of little consequence and is not—except for certain special services—relied upon to provide communication.

But the *actual* range, as shown by the full line curve of Fig. 5, is seen to increase very rapidly at about 100 metres, so that on waves below this the greatest ranges are obtainable.

Furthermore, these ranges are obtainable at a fraction of the cost in material and power of similar ranges on the long waves. This is because of the large dimensions and high operating costs of long-wave installations. For, considering only the aerial arrangements and bearing in mind that to be a good radiator its dimensions must approach in magnitude the order of a wavelength, it is easily seen how difficult and costly this is to achieve where the wavelength is long, and how simple and cheap in the case of the short waves.

But we have not yet explained why the range of short waves should suddenly increase as shown in Fig. 5. This is the subject with which the rest of this book will primarily deal, so we may as well examine the fundamentals of the matter straight away.

In considering Fig. 3, we agreed to deal for the time being, only with those waves which were going out at small angles to the earth's surface.

8. The Sky Wave

We will now consider that other portion of the radiated energy which is contained in the top portion of our hemisphere of radiated energy, and which represents the upward going waves, shown by the sloping arrows in Fig. 3. These waves, on leaving the aerial, commence to travel up towards the sky, and hence this portion of the radiated energy is known as the " sky " wave. In short-wave work we rely entirely on this sky wave to provide the energy which will actuate the receiver at the far-distant point.

But observing the direction of the arrows and remembering that the waves will normally advance in straight lines, it is at once seen that after they have been travelling for a second or so, they will be many thousands of miles away from the earth and will never be in a position to operate a radio receiver located upon it. That would indeed be so if they were travelling all the time in normal air—such as exists at the earth's surface. This acts as an electric insulator. But, fortunately for us—in more respects than one—the air surrounding the earth is not all in this normal state. For high in the atmosphere, and surrounding the earth like a shell, is a region where the air has been turned into an electrical " conductor ", and air in such a state acts upon radio waves very differently from that at the earth's surface. This " shell " of conducting air is said to be " ionised ", and hence the whole region has been given the name " Ionosphere ". It extends from about 30 miles to about 300 miles above the earth's surface.

Referring back to Fig. 3, let us consider that part of the radiated energy going out in the direction shown by the arrow marked *X*. When we consider the energy going upwards in one single direction only, such as that shown by arrow *X*—the energy contained within an extremely thin sector of the radiated hemisphere—then it is convenient to look upon it as a " ray " of radio energy, having very similar characteristics to those of a ray of light. We shall, therefore, refer to that part of the wave whose direction is shown by the arrow as a " ray ", but we must remember that in actual fact there will always be, not one, but a great many such rays travelling upwards side by side.

9. What Happens to the Wave in the Ionosphere

The ray we are considering travels onward and upwards in a straight line and with the velocity of light. When it enters the ionosphere, with its layers of conducting air, its behaviour alters. It is deflected from its straight course and commences to curve round so that it is travelling at a smaller angle to the earth's surface than when it left the ground. The curving process continues as it penetrates deeper into the ionised region until it ultimately describes a wide sweep, and eventually emerges again from the under side of the ionosphere, whence it continues in a straight line toward the earth. This curving process is known as refraction, and is illustrated in Fig. 6. Having lost very little of its energy during its journey—it *will* lose some in the ionosphere, particularly under certain conditions—the ray is able to actuate a radio receiver at the point on earth where it returns, which will be many miles distant from the transmitting station. Furthermore, on

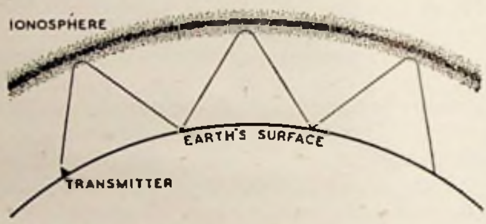


Fig. 6 : Short-wave transmission round the curvature of the earth

reaching the earth's surface it is reflected like a light ray from a mirror, and is sent off upwards again at the same angle at which it started. Reaching the ionosphere again it is again refracted and returned to earth at a point twice as far from the transmitter as that at which it first came down. These processes are repeated again and again so that the wave travels to the greatest distances on earth in a series of hops, as is pictured in Fig. 6. This is very convenient and fortunate for us, for when the spherical shape of the earth is considered it is impossible to see how we should ever get a radio wave to Australia, for example, if it persisted in travelling in a straight line.

We have, in our examination, considered only one ray of radio energy, but if we bear in mind that all the adjacent rays are being affected by the ionosphere in a similar way we shall see that a considerable portion of the earth's surface at the distant points will be covered by the down-coming rays, so that many thousands of radio receivers will be able to pick up the signals sent out.

We shall examine the structure of the ionosphere and its effect on the radio waves in greater detail later on. Let us now return for a moment to a consideration of Fig. 5.

Earlier on we discussed the different performance of the ground wave at different wavelengths and saw that the attenuation due to the earth was least on the long and greatest on the short wavelengths. What of the sky wave? What happens to it on the long wavelengths?

10. Different Ionosphere Effects at Different Wavelengths

The truth is that sky wave attenuation varies with wavelength in the opposite way to ground wave attenuation, being greatest at the long, and least at the short, wavelengths.

Now we see the full reason for the slope of the curve in Fig. 5. On the longest wavelengths no sky wave is returning at all, that is to say, no part of the received signal is due to waves which have been propagated by refraction in the ionosphere. As we reduce the wavelength (increase the frequency) the range gets less and less because the ground wave losses increase while still the upward-going waves do not return. They are completely attenuated in the lower part of the ionosphere. Then, as we continue to reduce the wavelength from about 100 metres, whilst the attenuation of the ground wave still increases, the sky wave attenation is reduced to the point where we begin to get the waves returned from the ionosphere. Thus our range goes up, and from then downwards in wavelength it further increases, because the sky wave gets stronger and stronger until the greatest distances are reached. We depend, therefore, on the sky wave alone for long-distance communication on the short waves.

It should be added that there are many factors affecting the matter which will introduce modifications to Fig. 5. However, we shall come to these in good time. In the next chapter it will be well to examine briefly the way in which the ionosphere is produced, and to learn something of its structure.

CHAPTER 2

THE SUN AND THE IONOSPHERE

IT may be wondered how the existence of the ionosphere ever came to be suspected, or anything of its nature learnt, since no man has ever been to that region, and it is far too high in the atmosphere to be reached by an unmanned balloon carrying recording instruments. Well, in the first place there was merely a postulation that a conducting region *must* exist somewhere in the atmosphere ; otherwise, how could a radio wave manage to travel round the earth ? If the high atmosphere was in the same electrical state as that at ground level the wave would simply continue straight onwards and be lost in space. Kennelly and Heaviside independently, but almost simultaneously, suggested its existence in 1902, and one part of the region carries their names.

11. Discovery of the Ionosphere

It was then soon seen that the air, which had always been regarded as a good insulator, might, in the high atmosphere, be converted into a conductor, by the process of ionisation. For in that region there would exist the atoms and molecules of gas capable of being made into ions, and also the ultra-violet light from the sun, capable of doing the work. Since then has come the invention of apparatus and the development of a technique by which the ionosphere can be, and constantly is being, explored and " sounded " by means of a special type of radio signal. This is sent straight upwards from the ground and soon comes echoing back from the sky, bringing messages which tell us a great deal about its journey, and about the ionosphere conditions encountered on the way. The first experiments of this kind, which directly proved the existence of the Kennelly-Heaviside layer of the ionosphere, were made in 1924 by Sir Edward Appleton and Dr. M. A. F. Barnett, using the Bournemouth B.B.C. transmitter.

By this means, not only have the theories of Kennelly and Heaviside been amply confirmed, but a great store of new knowledge has been acquired, enabling us to understand the reasons for the existence of such a region and to know something of the details of its structure and of the continual variations which occur within it. There are, however, a number of points yet to be explained—matters on which the scientists can only speculate—but about which they are constantly increasing their knowledge. So in what follows, we will attempt to give an idea of the main details as to the causation and structure of the ionosphere, with the reservation that, on some points, future knowledge may lead to some modifications of this idea.

The ionosphere is brought into existence by energy which is radiated from the sun. This fact very quickly becomes evident when we study some of the variations in the conductivity of the air which occur within it. Perhaps the most striking evidence of its dependence on the sun's radiations is its behaviour during a total eclipse of the sun. During such an event, when the sun's rays begin to be cut off by the moon, the conductivity or degree of " ionisation " of the air in the ionosphere begins to decrease, and this decrease continues as the eclipse proceeds towards totality. A minimum in the

12. The Sun's Radiations

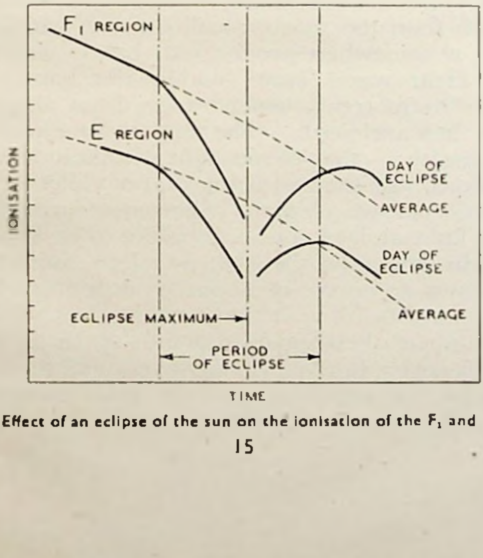


Fig. 7 : Effect of an eclipse of the sun on the ionisation of the F_1 and E layer

ionisation occurs at about the centre of the eclipse, and when it is over the ionisation increases again, and soon returns to a normal state. This shows that the agent responsible for the ionisation of the air has been prevented from reaching it during the eclipse, because the ionising rays have been blocked, as it were, by the presence of the moon in between the sun and earth, in the same way that the sun's light and heat have been cut off from the earth's surface during the eclipse. Furthermore, the fall in the ionisation is observed to start at the same time that the amount of light and heat reaching the earth starts to decrease, thus proving that all three radiations—light, heat and ionising radiations—are travelling towards the earth at the same speed.

Again, there are very marked variations in the degree of ionisation of the air as between night and day, and also as between summer and winter, as would be expected if the sun were responsible for it. All the evidence goes to show, therefore, that the agent causing the ionisation is indeed a part of the energy emitted by the sun.

Radiant heat and light are electromagnetic waves, and, as such, have much in common with radio waves. They are propagated according to the same general laws, and at the same velocity. They are, however, of much shorter wavelength than the shortest radio waves, which we may take to be of somewhere around 0·01 cm. in wavelength at present. Heat waves come within the band of waves known as "infra red", which ranges from about 0·04 to 0·002 cm. in wavelength. The longest wave to which our eyes are sensitive—that of red light—is about 0·00008 cm. in wavelength, the shortest being that of violet light, which is of 0·00004 cm. wavelength. Shorter in wavelength than the violet light and, of course, invisible to us, is a range of wavelengths, known as the " ultra-violet " radiation which extends down to waves of about 0·000006 cm. in wavelength.

From the sun there pours out into space an enormous amount of energy, in the form of waves of all these different wavelengths. A small part of the total energy emitted travels in the direction of this planet, and again a small part of this reaches the earth's surface. This we are very well aware of, because our eyes and other organs are sensitive to certain ranges of wavelengths, which we know as light and heat. But the waves of the visible light are accompanied by the other waves, both longer and shorter in wavelength. We are here most interested in the shorter variety, namely those known as ultra-violet rays. A large amount of energy comes from the sun towards the earth in this form but most of it never reaches the ground, because it is " absorbed " by the gases of the atmosphere. The energy present in these rays is expended in performing the work of ionising these gases.

The numerous gases constituting the earth's atmosphere —the principal of which are oxygen and nitrogen—are not uniformly distributed throughout the whole thickness of the atmosphere. Nor do they exist in the same state at all levels. The details of the actual distribution of the gases in the high atmosphere are not yet definitely known, though new information is fast being acquired. It is known, for example, that the distribution of the gases themselves is to some extent affected by the rays which pass into the atmosphere from the sun, and this fact accounts, in part, for the different character of the ionosphere at different times of day, and at different seasons of the year. For if the height above ground at which a particular gas predominates is first of all determined by the amount of a ray of particular wavelength coming down, then the further work of ionisation of the gas which is later performed by the rays will manifest itself at this height, as predetermined by the original rays. So the heights at which the ionisation appears will vary with time of day and season.

13. The Atmospheric Gases

We cannot, however, pursue this matter too far in a book such as this. We do need to know, however, something of the process of ionisation of the gas. Let us, therefore, follow the down-coming rays from the sun and observe what happens when they reach the atmosphere, imagining for the moment that the heights at which the different gases lie are constant.

Matter may be subdivided into small particles or molecules, and the molecules of all matter—whether gaseous, liquid or solid—are made up of atoms.

14. Ionisation of the Gases

A constituent part of the atom is the electron, which may be regarded as a particle of electricity, and should be visualised as being loosely held in its place by the electric charge in the rest of the atom, rather than as being embedded in the solid mass. Thus, if sufficient energy be communicated to it, the electron may leave its parent atom altogether, and take on a separate existence. When the ultra-violet rays come in from the sun they set the electrons of the gas molecules into a state of electrical oscillation, and when in this state the molecules dissipate energy which is taken from the wave. The wave thus loses its energy ; it is " absorbed " by the gas, and so eventually dies away. Molecules of different gases respond most readily to radiations of different wavelengths, according to the " frequency " of the down-coming waves and the nature of the gas. Thus a radiation of one wavelength may set the electrons of the molecules of nitrogen oscillating, whilst it would require radiation of a different wavelength to affect the oxygen molecules in the same way.

If the radiation is of the right wavelength the oscillations within the gas molecules become so violent that their structure is destroyed. The molecules lose some of their electrons and these float about independently of the parent molecules, thus constituting what are called " free " electrons. True, they may not be free for very long, for if they come near enough to another molecule which has an electron missing they rapidly become attached to it. But the chance of finding such a molecule in the upper atmosphere is much less than at ground level, because of the comparative rarity of the air at those heights and thus of the sparsity of the gas molecules. Also, as fast as—sometimes faster than—the free electrons becomes re-attached to such molecules, others are set free by the ultra-violet rays which continue to arrive from the sun, so that we have a more or less steady supply of them.

This is the process of ionisation about which we have been speaking. The parent molecules, when they have lost an electron, are known as " ions ", and in this condition they possess an electrical " positive charge ", whilst the free electrons have a " negative charge " of their own. A gas which is in this ionised state possesses the properties of an electrical conductor—because its free electrons are capable of independent movement—and acts upon radio waves in the way we have already briefly mentioned. It is therefore, the splitting up of the neutral gas molecules into ions and free electrons which changes the electrical nature of the air and so renders it impervious to radio waves coming up from the earth.

When the sun's rays are cut off from the atmosphere—as at night or during a total eclipse—then the molecules and free electrons *do* recombine so as to cause the density of the free electrons to diminish. The *rate* of recombination will depend upon the density of the gas, being low at the outer part of the atmosphere where the gas is very rare, and greater at lower levels where more gas molecules per unit of volume are to be found.

So we have the ultra-violet rays travelling from the sun towards the earth and meeting the gases of the

15. Distribution of the Ionisation

atmosphere. As they meet a concentration of any particular gas the rays of a certain wavelength are absorbed, and their energy is expended in ionising that gas. As they travel on the density of the free electrons produced is at first increased, because the gas is getting denser, and the more and

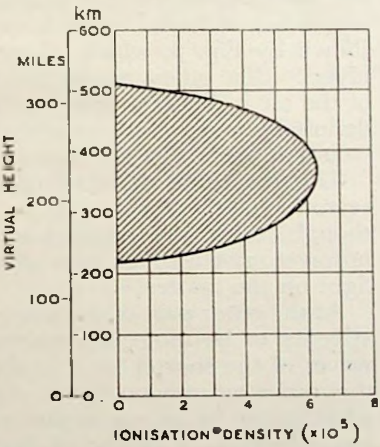


Fig. 8 : Variation of Ionisation density with height—F_2 region

more atoms do the rays find on which to expend their energy. Then the rays themselves become weaker because of this energy expenditure and the density of the free electrons produced decreases, until finally the rays are completely used up.

The density of the free electrons is therefore not constant throughout the ionised gas, but is a minimum at the top and bottom edges of the region where a particular gas is ionised and increases to a maximum at the centre. We may show this for one of the ionised gases in the graph of Fig. 8, where electronic density is plotted against height above the ground.

This sort of thing happens at the different levels in the atmosphere where the different gases predominate, rays of different frequency being expended at the different levels. The result is that throughout the ionosphere there are formed a number of *layers* of ionised gas, the height at which the layers lie being determined by the levels at which the different gases predominate.

16. "Layers" of Ionised Gas

It is important to appreciate this fact that the ionisation is not uniformly distributed with altitude, but is stratified into a number of layers, each having the main characteristics shown by Fig. 8, which, however, only shows one of the layers. The others are formed beneath it. The condition of the air in between the well defined ionised layers is not definitely known, but certainly the electronic density there would appear to be less than in the layers themselves.

The scientists have acquired a lot of information as to the reasons for the heights taken up by the various layers, though much still remains to be discovered. We may, however, give a rough idea which will serve to throw some light on the matter.

At the outer part of the atmosphere the main constituent appears to be molecular nitrogen, and this is ionised by waves of the shorter kind in the ultra-violet range, i.e., of the order of 0·0000085 c. wavelength. This produces what is, so far as we know, the uppermost layer of the ionosphere, at a height of from about 300 to 200 miles above the earth's surface. This region—named after its discoverer, Sir Edward Appleton, the " Appleton Region " —is also called the F layer of the ionosphere, and is the principal refracting layer for the short waves. It may here be remarked that the allocation of letters to designate the different layers was also initiated by Sir Edward, who, when he discovered the upper layer in 1925 called it the F, and the other layer then known the E, whilst a still lower layer which he located was named the D. This, as he says, left several letters at the disposal of future workers for allocation to other layers above and below these three.

Longer ultra-violet waves pass through the F region without losing an appreciable amount of energy, and continue on until, at lower levels, other gases are encountered. At a height of about 90 miles there begins to be encountered molecular oxygen, which quickly rises to high densities. This appears to be ionised by waves of about 0·000011 cm. wavelength, so that between about 90 and 50 miles there appears the Kennelly-Heaviside region, or E layer of the ionosphere. This also has important effects upon the short radio waves. At a height of about 38 miles a range of waves relatively near to those of visible light are absorbed. This gives rise to the D layer of the ionosphere. There does not appear to be any concentration of gas such as would give rise to further ionised regions lower than this ; at least, one does not *regularly* exist in the lower atmosphere.

The position of the *main* ionised layers thus seems to be accounted for. The wavelengths given for the rays absorbed at different levels are only intended to be approximate. It is more likely that a considerable band of waves is absorbed by each gas, and there is also the possibility that some of the ionisation may be due to particles of matter which have been shot out of the sun.

Finally we are left with those solar rays which have not been absorbed and which reach the ground. These comprise a band of waves including the heat and visible light waves, and cutting off fairly sharply at the short-wave end at about 0·00003 cm. wavelength, owing to absorption by the atmospheric ozone.

At the other end the band cuts off at about 0·00025 cm., the longer solar rays being absorbed by water vapour in the atmosphere.

Let us now look at a diagram showing the various layers of ionised gas which we have enumerated, so that we may get a good idea of the main structure of the ionosphere. As has been mentioned, owing to the solar rays affecting to some extent the distribution of the gases, the layers will not always be at the same height, nor will the electron density always be the same

17. Structure of the Ionosphere

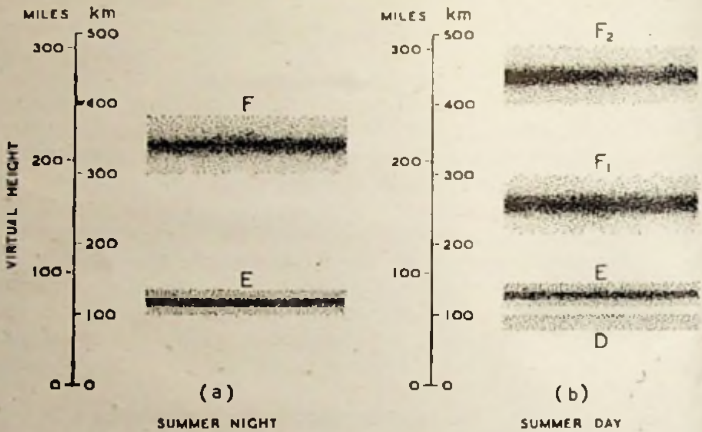


Fig. 9 : How the structure of the ionosphere changes from night to day

within them. Both height and density will depend upon the intensity of the solar rays and so will vary diurnally, seasonally and otherwise. So we shall need different diagrams to show the conditions at different times.

Fig. 9 will help us to picture the ionosphere structure by night and by day during the summer.

The F layer exists at night as a single layer at the " top " of the ionosphere, while below it is the E, of much lower electronic density than during the day. The D layer does not exist at night, or at least its ionisation becomes very weak. During the day the F layer divides into two layers—

probably owing to some redistribution of the gases—the upper part of which is called the F_2 and the lower the F_1 layer. The F_1 comes into being lower down than where the night-time F existed, while the F_2 is higher than the night-time F in summer and somewhat lower in winter.

Below this is the E—at about the same height as during the night—and below this again the D layer has come into being.

We shall talk about the variations in height and density

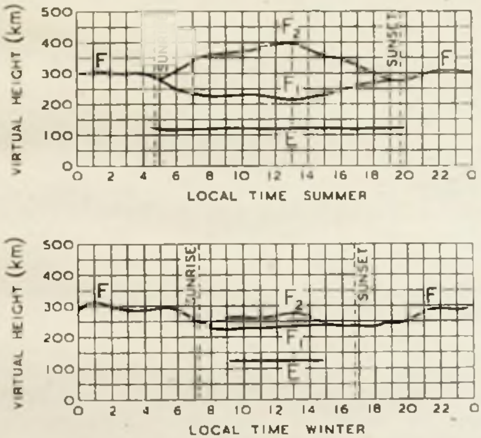


Fig. 10 : *(Courtesy U.S. National Bureau of Standards).*
Monthly average virtual heights of the principal ionosphere layers for each hour of the day. From observations made at Washington

later on, but it may help if, at this stage, we give some typical measured values for the heights (of the lower edge) of the F, F_1, F_2 and E layers for typical days during summer and winter. These are shown in Fig. 10 and it is seen how the variations in height took place at different times of day. The D layer is not shown in this diagram and the E only during the day, when it is most highly ionised.

These variations need not, however, be studied very deeply at this stage. It will be sufficient for our purpose to picture the ionosphere as a region comprising 3 main layers, the F, E and D, and we can imagine, for the time being, that they remain at the same height all the time, lying one above the other in the atmosphere.

This will simplify matters somewhat, and we can now go on to learn something of the way in which the heights and other characteristics of the layers are actually measured and studied.

CHAPTER 3

HOW THE IONOSPHERE IS SOUNDED

IN order to understand how the ionosphere is sounded, and its characteristics measured and recorded, we shall have to go a little deeper into the *reason* for the different behaviour of a radio wave in ionised and in ordinary air. This is a rather complex subject, and one somewhat difficult to explain in a few simple words. But if we neglect some of the details, and confine ourselves to a few essential facts, we ought to be able to get a fairly clear idea of what happens to cause an upward-going radio signal to return to earth as an echo from the ionosphere.*

18. Production of the Echo

It will be remembered that a radio wave has been said to consist—in part—of a series of electric strain lines in space, the electric field as a whole being in a state of continual variation—its *rate* of variation depending on the frequency of the wave. Now the velocity of a wave depends upon the amount of current set up by the oscillating electric field. In ordinary air—which is an insulator—the electrons are held fast in the molecules, and the wave is unable to set them in motion so as to produce any current. It, therefore, travels with the velocity of light ; i.e., at a speed of 300,000,000 metres per second. It does not travel at an infinite speed, because there is in fact what is equivalent to a current, namely, the constant change in magnitude and direction of the electric field and this limits its velocity to the figure just given. We may regard this as the natural velocity of all electromagnetic waves in " free space ". The equivalent current due to the rate of change of the oscillating electric field is called the " displacement current ".

* In particular it is considered inadvisable here to go into the difference between the " phase velocity " and the " group velocity " of a wave in the ionosphere. Readers desiring fuller information are referred to " Short Wave Wireless Communication ", by A. W. Ladner and C. K. Stoner or to articles on " Electro-magnetic Fields in Radio," by Dr. Martin Johnson, *Wireless World*, June and July 1943.

When the wave travels in *ionised* air, however, the situation is different, because the electrons are not all held fast in the molecules—some of them are free. The electric field sets into motion all these free electrons and these, oscillating in time to the frequency of the wave, do comprise an actual current, which tends to cancel out the displacement current. We may refer to this actual current, due to the moving electrons, as the " conduction current ". Hence the velocity of the wave is altered, because the total current set up by the electric field is now different from that which existed in ordinary air. The velocity at which it now travels will depend on the number of electrons per unit of volume in the ionised region, or in other words, upon the strength of the conduction current which it sets up.

Now the type of signal used for ionosphere measurement work is a very short, sharp burst of energy, like the dot in the morse code, though much shorter even than this. It is called a " pulse ", and usually lasts only a few thousandths of a second. Nevertheless, during that time several complete waves are emitted, and we may say that the pulse is made up of a small group or " train " of such waves. We will consider the group of waves comprising a pulse as one entity, and endeavour to follow its course during its journey to the ionosphere and back.

19. Pulses

In order to obtain the data on ionosphere conditions which we desire, we arrange the aerial system of our transmitter so that the pulses are sent straight upwards ; i.e., vertically up towards the sky. We arrange to have our receiver very near to the transmitter, so that it will be actuated by some of the energy radiated when the pulse is sent off, and also by the energy in the sky wave when it comes echoing back from the ionosphere.

The wave group radiated from the aerial thus ascends vertically upwards, travelling at the speed of light, until it reaches the layer of ionised air and sets the free electrons into oscillation. The current represented by these immediately reacts upon the wave group, cancelling some of the displacement current and causing the group to travel *slower* than it did in ordinary air. As it penetrates further into the ionised region, the density of the free electrons gets greater and greater, and the conduction current set up by the wave thus gets stronger and stronger. The result is that the wave group continues its ascent at an ever decreasing speed. Eventually, when the density of the electrons reaches a certain critical value, the velocity of the wave group is reduced to zero. It stops still in the ionosphere. But only for an infinitely short time, for it is then turned completely round or " reflected ", and immediately starts to travel away from the critical density region and towards a region where the electron density is lower ; i.e., downwards again towards the lower boundary of the ionised layer. Its velocity on reaching this has again increased, and it emerges from the layer and travels toward the ground with the velocity of light.

So we can imagine our pulse signal ascending vertically towards the ionosphere, slowing down when it reaches that region, and eventually being reflected and reversing its direction so as to return again to earth. We may look upon it as if the group were reflected from a certain level in the ionosphere, as it were from a metallic surface, or like a ray of light reflected from a mirror. We might also regard the density of the electrons necessary to reflect the wave as having something in common with the mirror, and call it the " mirror " density of electrons.

Suppose, therefore, that we have arranged our transmitter so that it sends the short-wave group directly upwards, and also that we have available a radio receiver whose output is connected to an oscillograph.

20. Observation of the Echo

What we desire to do is to measure the *time* taken for the pulse to go up and down again, and then, knowing the speed at which it has travelled (over the greater part of its journey) we shall know at what height it was reflected. The time taken will only be a matter of a few thousandths of a second, so we cannot hope to measure this by merely listening to the echo. That is why we connect the output of our receiver to the oscillograph so that we can actually have a visual record of it. Then our receiver will pick up the pulse immediately it is sent off—by getting it direct from the transmitter—and a moment later it will receive the echo, or signal which has been reflected back from the ionosphere.

Our oscillograph record is arranged to move from right to left at a known speed, and what we shall get on it is shown on Fig. 11 (*a*) and (*b*). In (*a*) G is the pulse picked up by the receiver immediately it was sent off and F is the echo received a moment later. As we know the speed at which the record is moving we can measure the time to which the distance between G and F corresponds. This is the time taken for the pulses to go up to the ionosphere and return to earth again. If we multiply this by the velocity of light we have the distance it has travelled, and half of this is the height above ground at which the signal was reflected. Fig. 10 (*b*) shows a case where we have a certain amount of energy returned from the E layer, though the majority was reflected by the F, as shown by the longer time delay.

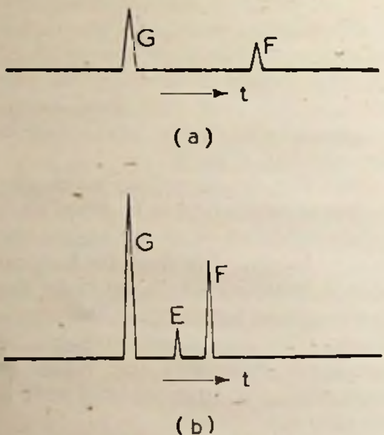


Fig. 11 : Echoes returning (*a*) from the F layer, (*b*) from the E and F layers. G is the original pulse signal, received at the moment it is sent

In practice the oscillograph equipment can be calibrated directly in terms of height rather than in terms of time delays, so we can read off the height of reflection of the signal without any calculation. What we actually measure is a quantity known as the " virtual " height and not the true height to which the wave has reached. We shall, however, explain the meaning of this later on and for the time being it will do no harm if we regard the height recorded as the true one.

Now we come to a very important point. When the wave sets the free electrons into motion the amplitude and average velocity of their vibrations is the greater the lower the frequency of the wave. The impetus given to the moving electrons, and the magnitude of their oscillation will be the greater the longer the electric force in the wave continues to act in one direction. Hence the oftener it changes its direction (the higher the frequency) the less will be the motion imparted to the electrons. Consequently, the magnitude of their effect upon the wave group—in slowing it down—will be the greater the lower the frequency. It varies, in point of fact, inversely as the square of the frequency or directly as the square of the wavelength.

21. Variation of Height with Frequency

Let us pause here to see that we fully appreciate the implication of this. It means that a pulse sent up on a long wavelength (low frequency) will slow down more quickly and be returned from a point lower in the ionosphere than one on a short wavelength (high frequency). For, if the magnitude of the electronic effect upon the wave is greater for the low frequency than for the high frequency, then it will not require such a high electron density to " reflect " the low frequency as is necessary for the high frequency. Here we have one of the fundamental points, not only of ionosphere measurement work, but also of practical short-wave communication. It may be summarised like this. As the frequency is increased the wave penetrates further and further into the ionosphere layer before it is reflected, in order to reach the point where the electron density is great enough to do this. And there is a definite limit to the highest frequency which will be reflected by an ionised layer, corresponding to the maximum electron density existing within it. This is called the " critical " frequency of that layer, a term introduced by Appleton, who first carried out experiments of this kind.

22. Plotting the Ionosphere Characteristics

Suppose now that we are about to start sending a whole series of pulses up to the ionosphere with the object of finding out what the prevailing conditions are. We arrange to start sending our pulses upon a relatively low frequency (long wavelength) and gradually to increase the frequency as we proceed. We will read off the height at which each pulse is reflected and plot this in a graph. Suppose we start on a frequency of 1 Mc/s (300 metres) and then increase the frequency of the transmitter (reduce its wavelength) in steps of, say 2 kc/s. We shall obtain a graph or curve somewhat like that in Fig. 12.

At first (on 1 Mc/s) we obtain no echo at all, because the wavelength is too long, and the upgoing energy is all absorbed in the D layer, as explained for medium waves in Chapter I. At about 1·7 Mc/s we commence to get reflections from a height of 110 km. and this continues until about 2·8 Mc/s, when the height from which the pulses are

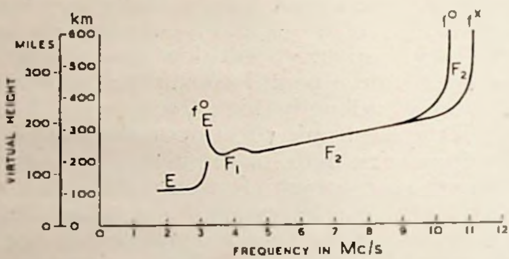


Fig. 12: Curve showing heights from which the echoes are returned for various frequencies during winter day. fo relates to the "ordinary" and fx to the extraordinary wave, see page 34

reflected starts to increase. The pulses over this band of frequencies are being reflected from the lower part of the E layer, as is shown by the height registered (see Fig. 9). The lowest layer in the ionosphere, the D, does not reflect waves—it only acts as an absorbing medium to them. The upward curl at the right-hand end of the curve is occasioned by the penetration of the wave into the E as the frequency is raised, until at 3·2 Mc/s the pulses penetrate the E altogether and go up to the F_1. 3·2 Mc/s is thus the critical frequency of the E; i.e., the highest frequency returned by it at vertical incidence.

As we continue raising the frequency—the pulses now coming down from F_1 layer—we find at first that the height apparently *decreases* with increasing frequency, as shown by the upward curl at the left-hand end of the F_1 curve. There appears to be something wrong here, as we should not expect the wave to be reflected *lower down* with each increase of frequency. Well, it must be remembered that, in whatever way our oscillograph is calibrated, we are actually measuring the *time* taken for the wave group to go up and come down again. At frequencies near the critical frequency of the E layer the wave group will be slowed up considerably as it passes through this layer, although it does not actually undergo reflection there. If the wave is thus retarded in the E we shall get a falsely great height recorded for the F_1, and so we may disregard the heights given by the curl. They are too great because of this retardation in the E. As we get away from the critical frequency of the E the retardation of the wave in that layer soon becomes negligible, and at about 3·5 Mc/s the curve for the F_1 flattens out and our height of 220 km. is really indicative of the height of reflection at the F_1. The curve now shows slightly increasing height with frequency—because the wave goes a little further into the layer with each increase of frequency—until at 4·2 Mc/s we get a decided kink in the curve. This shows the point at which the wave penetrates the F_1 and begins to be reflected at the F_2. During the winter day—as will be seen from Fig. 10—there is very little difference in height between the F_1 and F_2, so that the wave goes from one layer to the other almost imperceptibly. Nevertheless, the kink in the curve has this definite meaning, and does enable the change from F_1 to F_2 to be detected. It is due to a repetition of the retardation phenomena which we have just described for the first reflections from the F_1.

Our pulses are now up to the under-side of the F_2—the highest layer in the ionosphere. As we continue to increase frequency the wave begins to penetrate further into the layer, so that the height shown gets slowly greater. Then—following only the upper branch of the curve—the penetration (and retardation) in the F_2 is seen to increase rapidly from about 9·0 Mc/s. Finally, at about 10·4 Mc/s, the wave is penetrating to the point of maximum electron density and so the curve cuts off, for any increase in frequency merely means that the wave goes right through the layer, there being an insufficient electron density at any point to return it. 10·4 Mc/s is, therefore, the highest frequency returned from the ionosphere at *vertical* incidence—if we neglect the lower branch of the curve—and is the critical frequency of the F_2 layer.

Now we come to a rather difficult matter, and one which, although the reader need not let it worry him unduly, nevertheless merits some explanation. We refer to the splitting or forking of the curve which is seen to commence at 8·4 Mc/s.

23. Effect of the Earth's Magnetic Field

This is due to the effect of the earth's magnetic field, which does in fact split a wave travelling in the ionosphere into two components. When the wave is travelling in ordinary air the magnetic field has no effect upon it, but as soon as it sets the free electrons of the ionosphere into motion—the behaviour of the wave itself being dependent on the character of their motion—then the magnetic field does begin to affect it. For the paths of the moving electrons are altered by the magnetic field, which exerts a twisting effect upon them, and thus their effect upon the wave is different. So that the magnetic field affects the radio wave through its influence on the moving electrons—it influences the wave at second-hand so to speak.

We may perhaps explain it in this way. It was mentioned earlier that when the wave leaves the aerial the direction of its electric strain lines is quite definite, and in the ionosphere the electrons will initially vibrate in paths determined by the direction in which the strain lines are acting. But the electronic motion, when affected by the earth's magnetic field, causes the polarisation if the wave to change in a very complicated manner. We need not, however, go into this.

But suppose in the case of our exploring wave, sent vertically up, that when it enters the ionosphere the electric field is acting so that the electrons are set vibrating in a direction exactly parallel to that of the earth's magnetic field. The field, in such a case, will have no effect on them, and consequently its effect will not be apparent in the behaviour of the wave itself. The pulse signal will ascend until the magnitude of the conduction current in sufficient to cause complete reflection, and then it will commence to descend.

Suppose now that the electric field is acting so as to set the electrons vibrating in a direction *transverse* to that of the magnetic field. The field will now have the maximum effect upon them ; its twisting effect upon their paths will be at its greatest. And this twisting effect is tantamount to an increase in the strength of the conduction current itself, so that the wave is more affected than before. The pulse signal is slowed down more quickly and it is completely reflected with a lesser density of electrons than before. It therefore is reflected lower down in the ionosphere than is the wave we first considered. In practical cases—when the wave enters the ionosphere with the direction of its electric field at an angle to that of the earth's field—the wave is resolved by the ionosphere into two separate components, each behaving differently and according to the general cases stated above. They become differently polarised, travel with different velocities and require different electronic densities to ensure their reflection. That behaving according to the first case is called the " ordinary " wave, and its performance is represented by the upper or left-hand fork of our curve. That behaving according to the second case is the " extraordinary " wave, and its behaviour is recorded in the lower or right-hand fork. As will be seen, after a frequency is reached such that the ordinary wave has penetrated the layer, echoes of the extraordinary are still received, because it requires less electrons to reflect it than does the other.

The result of this is that, after reaching a frequency where the ordinary ray has penetrated the ionosphere, as at 10·4 Mc/s in Fig. 12, we can still get echoes of the extraordinary ray coming down. For on this frequency there is still a sufficient electron density to cause reflection of the extraordinary ray, though the ordinary ray has escaped. As we increase frequency still further the extraordinary ray ascends further into the layer (and is retarded more) until it too fails to find an electron density sufficient to reflect it. Se we have two critical frequencies for the one layer—one for the ordinary and one for the extraordinary ray. The frequency separation between the two depends on the strength of the earth's magnetic field—because that will determine the magnitude of the twisting effect on the moving electrons—and thus it will vary at different points on the earth's surface, being about 0·7 Mc/s in these latitudes.

In short-wave practice it is the ordinary ray critical frequency which is most important, and this is the value which is generally considered to be the safe upper limit for use in the calculation of the highest frequencies for actual use over short-wave circuits.

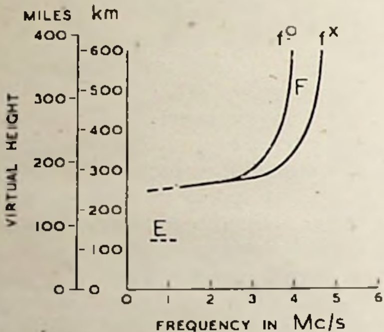


Fig. 13 : Curve showing heights from which the echoes are returned for various frequencies during winter night

Fig. 13 shows the sort of curve we might have obtained had we conducted our experiment at night. Although we can detect the presence of the E at this time most of our energy penetrates this layer and goes up to the F. There is no kink in the F curve and this shows that the F now exists as a single layer (see Fig. 9). Furthermore, the F critical frequency—for the ordinary ray—is not only 3·9 Mc/s as against 10·4 Mc/s during the day. There is thus a large difference in the highest frequency returned from the ionosphere as between day and night.

Returning to the daytime curve, we can now read off the critical frequencies of the E, F_1 and F_2 layers, and the virtual height of the layer for a given frequency. The critical frequency of a layer is denoted by a symbol such as f_E in which the E subscript indicates the layer in question. The superscripts O or X,

24. Data Obtained from the Curves

denoting the ordinary or extraordinary rays, have already been mentioned.

Since the critical frequency is a measure of the electron density or ionisation—the density of the free electrons being proportional to the square of the critical frequency—it is seen that the ionisation is least in the E layer, greater in the F_1 and much greater still in the F_2. As to the heights given, it must be stressed that what we are measuring is not actual but virtual height. The reason for the difference is this. As we arrive at the height by noting the time taken for the echo to return, and assume that the wave travels all the way with the velocity of light, we fail to take account of the slowing up of the wave group in the ionosphere. We, therefore, place the height as somewhat greater than it should be. And we are unable to allow for this slowing up because we do not know how the electron density is varying with height and so how the wave is being retarded. However, since the wave travels for the greater part of its journey in ordinary air where we do know its velocity, the heights given do yield us some information as to the atmospheric levels taken up by the various layers.

We have, therefore, in the curves plotted, quite a lot of information about the ionosphere above the transmitter at the time the pulses were sent up. This information obtained at vertical incidence is of great use in telling us the correct frequencies to use in practice over different short-wave circuits, for we can, by calculation, apply the data for vertical incidence to any case of oblique incidence, such as will occur in practical communication. But we must remember that, since the ionisation is produced by the sun, it will vary according to the time of day, season of the year, etc. This means that it will vary greatly with latitude and longitude. So we really require a series of curves obtained at different times of day and at different locations on the earth's surface. As a matter of fact there are a number of stations located in different parts of the world engaged in making these curves, and it is from their records that we obtain the information which we require in short-wave work.

At some of these stations the transmissions are sent up automatically and the curves themselves are automatically plotted by photographic means. An apparatus is arranged at the transmitter to send out the pulses several times in a second, whilst the frequency is automatically increased by about 2 kc/s between the sending of each pulse. The photographic film on which the height is recorded is moved each time the pulse is sent, so that the whole curve is traced out photographically, the time taken to cover the

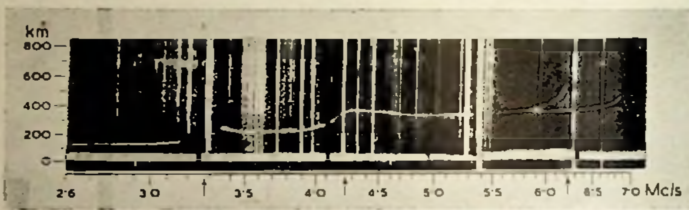


Fig. 14 : (*Courtesy National Physical Laboratory*).
Curve of virtual height against frequency taken by photographic means at the Radio Research Station, Slough

whole frequency range being about 15 minutes. The apparatus then repeats the whole process over and over again, so that continuous recording is going on. An example of some curves obtained in this way is given in Fig. 14.

In the next chapter we shall examine the nature of the continual variations which occur in the critical frequencies and virtual heights so recorded.

CHAPTER 4

IONOSPHERE VARIATIONS

AS would be expected from what was said in Chapter 2, large changes in the ionisation of the upper atmosphere take place because of the variations in the amount of solar radiation acting upon it. The ionisation varies, firstly in accordance with the relative position of the sun and earth, and secondly in accordance with the general activity of the sun itself. For the amount of ionising radiation actually emitted from the sun is not constant, but shows a considerable variation with time.

25. Causes of Variation in the Ionisation

Thus the ionosphere is subject, because of the variation in the relative position of sun and earth, to regular daily and seasonal changes of ionisation, and also, because of the variations in the sun's emissions themselves, to changes which, though of an erratic nature from day to day, nevertheless have a long-term periodic character. This is such as to show maxima in the ionisation about every eleven years, with minima *about* half-way between the maxima. As has been said, it is the amount of ionisation —or free electron density in an ionosphere layer which determines its critical frequency. The greater the density of the free electrons the higher is the critical frequency, and also—what is more important in practice—the higher is the frequency which the layer will refract when the wave strikes it at oblique incidence, as it will do when we send it out so as to communicate over a long distance. In practical short-wave communication, therefore, we are vitally interested in the changes in the ionisation of the layers, and hence in the variations of their critical frequencies, for these will determine the frequencies which we shall be able to use at any particular time for long-distance communication. If we ignore these variations, and attempt to operate our short-wave stations on frequencies chosen haphazardly, and without regard to the ionosphere variations, the chances of our being able to maintain good communication will be very remote, for part of the time our waves will fail to arrive at the receiving point because they have penetrated through the ionosphere altogether, while at other times they will fail to do so because of complete absorption in the lower ionosphere. If, on the other hand, we carefully choose our working frequency to suit the conditions of ionisation prevailing at any time, then the wave will be properly refracted, and will travel to great distances with extraordinarily little loss of energy.

We can best study these variations in terms of the critical frequency, because it is easiest to relate this to the actual working frequencies for various distances. We might remember that the actual ionisation in the layer is proportional to the square of the critical frequency.

26. Diurnal Variations

Figure 15 gives the critical frequencies and virtual heights of the three principal layers, as measured at Washington during winter and summer of years when the sun's activity was at a minimum and at a maximum. It should be mentioned that the last minimum of solar activity occurred in 1933 and the last maximum in 1937. Let us examine these curves in some detail.

Concentrating first on the diurnal variations and looking only at those of the E and F_1 layers, we see that at all seasons of the year and at all epochs of the solar cycle the critical frequency varies in direct accordance with the altitude of the sun, increasing from sunrise, reaching a diurnal maximum at noon, and then decreasing again towards sunset. The critical frequency is, as a matter of fact, proportional to the cosine of the zenithal angle of the sun. The layers thus behave according to the simple theory of ionisation by ultra-violet radiation from the sun. When the amount of ultra-violet light affecting them is greatest, i.e., at noon, then the ionisation is at a maximum, and when it is negligible, as at night, the ionisation falls to a low value.

We must here digress for a moment to remind the reader that the free electron density in a layer at any time will depend, not only upon the rate of ion production, but also upon the rate of recombination of the electrons and ions. The density will only increase when the production rate is greater than the recombination rate. The recombination rate will depend, in the main, upon the density of the air itself, i.e., upon the number of molecules of gas per unit of space. For it will be clear that the more numerous the molecules are, i.e., the less the distance between them, the greater will be the chance that a free electron will come into contact with an ionised particle and so recombine with it so as to re-establish the electrical neutrality of the structure. In the lower part of the ionosphere a high molecular density does exist, and thus when the sun's rays diminish or cease to affect the layers, the ionisation rapidly decreases.

So far as the E layer is concerned, then, there is a fall in the critical frequency towards sunset, and during the night—though not shown in Fig. 15—it is of such a low value as hardly to affect the short waves at all. It should be remarked, however, that the E ionisation does not *entirely* disappear at night—it remains at a value, for example, that has some effect in long and medium wave propagation. As to the F_1 layer, it ceases to exist as a separate layer at night, but merges into the F_2 to form a single night-time F layer.

Turning to the diurnal variation of critical frequency in the F and F_2 layers, we see that the critical frequency increases from sunrise but that it is higher during the hours after noon than during those before noon. In fact, during the summer it actually goes on *increasing* during the afternoon, and the diurnal maximum does not occur till about sunset. There is a rapid decrease after sunset and, generally speaking, a slow decrease during the night, though in winter there is a slight pre-sunrise peak in the critical frequency. It appears that these characteristics may be accounted for by saying that in the first place the layer has a diurnal variation lagging behind the altitude of the sun. This is probably because of the rarity of the gas molecules at the height at which

27. The F_2 Layer

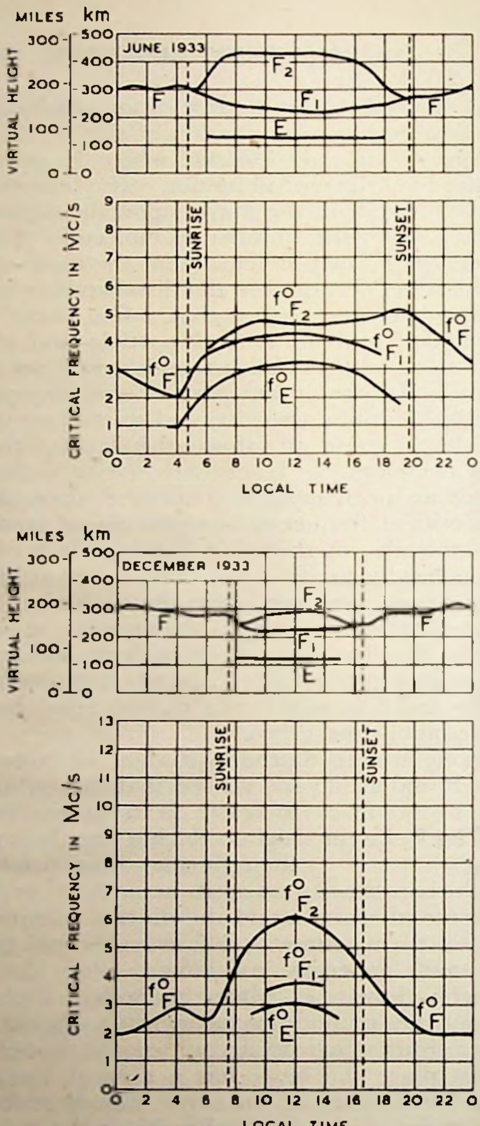
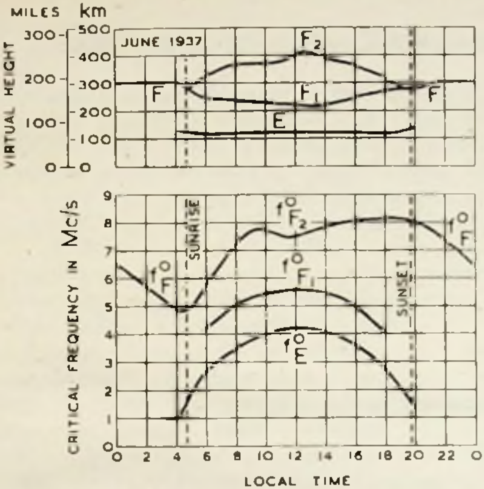
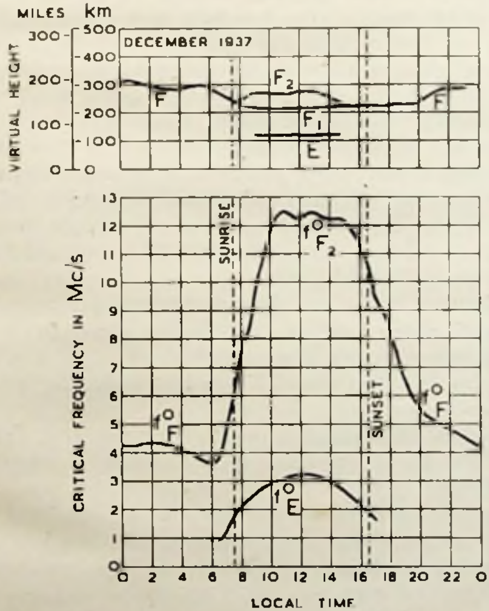


Fig. 15 : (*Courtesy U.S. Nat*

Monthly graphs of critical frequencies and vi
and Winter of years of mini





eau *of Standards*).

ghts observed at Washington during Summer
maximum solar activity

the F layer lies, the consequent low recombination rate resulting in a sort of " hangover " of electron density after the free-electron-producing agency has diminished. In the second place there appears to be a temperature effect which is responsible for the pre-sunrise peak and for other effects, which however, we can more conveniently discuss when we come to the seasonal variations.

28. Seasonal Variation in Critical Frequency

The seasonal variations in the critical frequency of E and F_1 layers are again simple and straightforward, the critical frequencies being higher in the summer than in the winter of any one year, and the diurnal maxima being greater in summer than in winter. This again is consistent with the simple theory of ionisation by ultra-violet radiation, the critical frequency being proportional to the sun's zenithal angle and thus reaching a peak in summer.

In the F_2 layer exactly the opposite sort of variation is seen to occur, so far as the day-time critical frequency is concerned, this being greatest in winter and least in summer. So far as is known this effect has never been quite satisfactorily explained, though theories have been advanced to account for it. The most satisfactory one is that due to Sir Edward Appleton, who ascribes it as due either to pronounced heating of the uppermost regions of the atmosphere, or to the presence of a very light gaseous constituent in summer. In either case there would be an increased extension of the atmosphere to great heights during the summer months. In an expanded atmosphere the ionisation is more widely distributed than in a contracted atmosphere, and the number of free electrons per unit of space is decreased, consequently the critical frequency decreases.

So in summer the ionisation and hence the critical frequency of the F_2 is less than it is in winter, when the gas is more dense.

The pre-sunrise increase in winter is accounted for on a similar basis. During the long winter night the gas cools very considerably and contracts during the process, so that the number of ions per unit of space is increased.

Thus, before dawn there is a definite rise in the critical frequency, due, not to the beginning of free-electron production, but to the crowding of the free electrons already there into a smaller space. Apart from this pre-sunrise increase in the critical frequency the night-time critical frequencies are lower in winter than in summer. This is due to the shorter period of time during which the ionising radiation is operative in winter, the earlier onset of darkness—during which only recombination takes place—resulting in the critical frequency falling to very low values during the winter night.

Combining the diurnal and seasonal variations in F layer critical frequency we see that it reaches its highest values during the winter day and its lowest during the winter night. The variation between day and night critical frequencies is therefore, great in winter and relatively small in summer. The greatest rate of variation between day and night frequencies thus occurs in winter just after dawn and just after sunset.

It should be noted that in the ionosphere there is no lag in the seasonal effects such as occurs in the seasons of weather, which follow the sun's seasonal position a month or two later. The ionosphere effects coincide with the sun's position and in the northern hemisphere winter conditions may be said to prevail from October to March and summer conditions from April to September. At the equinoxes conditions alternate between the summer and winter types, and are usually erratic at this time.

It will be seen from Fig. 15 that the critical frequencies of all the layers, both by day and night, were considerably greater in 1937 than in 1933. This is because of the variation in the sun's activity itself during that period.

29. Long-period Variations

Such a variation in solar activity has long been known to occur—indeed, records of it go back to the time of Galileo, who in 1610 was keeping records of the sunspots. These sunspots are only one evidence of solar activity and of the variation taking place in it, though they are the most convenient of solar phenomena to observe. It is found that there is a mean period of

11.1 years in the degree of solar activity, as evidenced by the number and size of the sunspots observed. That is to say, a mean period of 11.1 years is occupied by a complete cycle of activity from minimum, through maximum to minimum activity again. By this it is not meant that the sun's activity increases in a smooth and regular manner from minimum to maximum—there are, in fact, large and erratic variations from day to day. Nor are the cycles

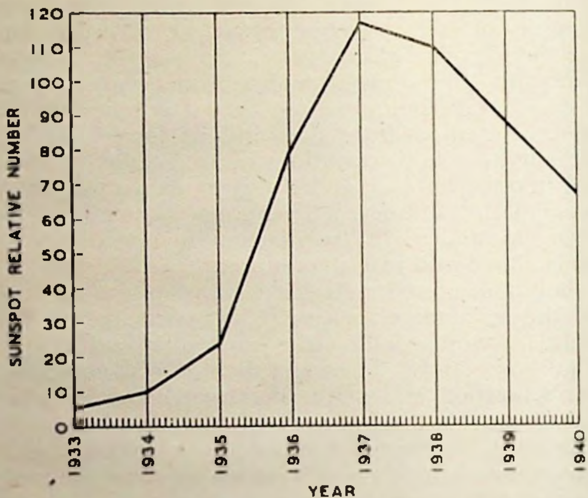


Fig. 16 : Yearly means of sunspot relative numbers

themselves perfectly regular, either in amplitude or frequency. Nevertheless they are regular enough to show the cyclic change in *average* activity quite markedly.

If we take the sunspots as evidence of the solar activity and plot the yearly mean value of the sunspot " relative number " we get a graph such as is shown in Fig. 16. We see from this that 1933 was a year of minimum solar activity whilst during 1937 the activity of the sun was at its peak.

Now if the solar radiation of ultra-violet light—which gives rise to the layers of the ionosphere—increases and decreases according to the variations in the sun's activity, we should expect a greater degree of ionisation—and hence higher critical frequencies—to prevail at the maximum than at the minimum of the solar cycle. Fig. 15 clearly shows that this is true, and the difference in the critical frequencies in those two years is thus correctly accounted for. The variations in the critical frequency over part of a solar cycle are better shown in a graph such as Fig. 17,

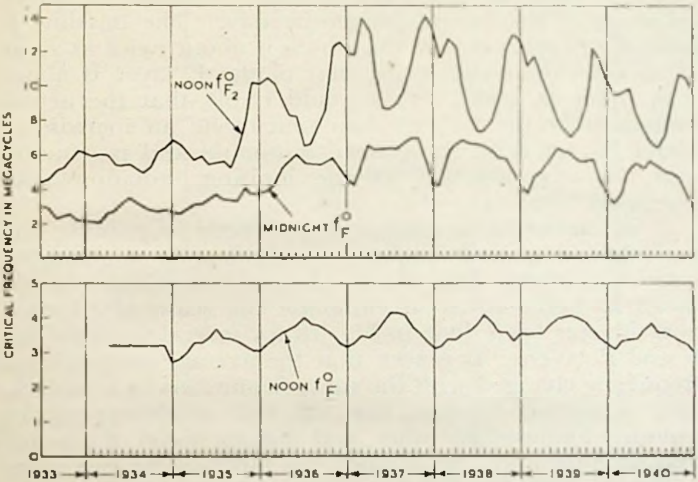


Fig. 17 : Monthly average of noon (f^o_E and $f^o_{F_2}$) and midnight (f^o_F) critical frequencies at Washington

which gives the monthly mean of the noon critical frequency of the F_2 and E layers as well as the midnight values for the F layer for the years 1933 to 1940. It is seen that, as the solar activity increases there is a large increase in the average value of the critical frequency, particularly in that of the F_2 layer.

Of course, in Fig. 17 we have the large seasonal variations super-imposed on that due to the solar cycle, and it will be seen that at noon, as has already been pointed out, these are in the opposite sense in the two layers, i.e., the E critical frequency is highest in summer whereas that of the F_2 layer is lowest.

As we shall see later, this long-period variation will have its effect on the frequencies we are able to use in short-wave communication, leading to important changes from year to year. As to the magnitude of the change in critical frequency over the solar cycle it will be seen that it is greater in the F than in the E layer, whilst it is greater in the F layer by day than by night, because the direct influence of the sun is greater by day. The daytime F critical frequency at the maximum is about twice as great as at the minimum, whilst that of the E layer is about 1.25 times as great. This would imply that the actual ionisation in the E layer had undergone an increase of about 60 per cent. as between minimum and maximum, and that the intensity of the ionising radiations had increased by about 150 per cent.

It is interesting to compare, as directly as possible, the variation in critical frequency with that in the sunspot relative number. This can best be done by taking a yearly mean of both—so as to eliminate the seasonal effects—and this has been done in Fig. 18 for several years for the F and E layers. It is seen that the average noon critical frequency changed with the sunspot numbers in a remarkably consistent manner, that the rate of change of one quantity followed the other, and that the times of maxima and minima were *approximately* the same in the two cases.

Fig. 15 also shows the minimum virtual heights of the different layers in 1933 and 1937. That of the E layer remains practically constant throughout the day and at all seasons of the year and epochs of the cycle. The F_1 also maintains an *approximately* constant virtual height, and this also applies to the night-time F layer.

30. Changes in Virtual Height

The daytime F_2, on the other hand, lies at a much greater height during the summer day than during the winter day. This variation in height is probably due to some redistribution of the gases by the action of the solar rays themselves, or possibly it is connected with the expansion of the gas during the summer day because of heating. As will be seen from Fig. 15, the night virtual height is usually in the vicinity of 300 km. both in winter and in summer. Just before sunrise there is a considerable and rapid decrease, and then, in winter, a small increase—the layer remaining during the day at a lower level than

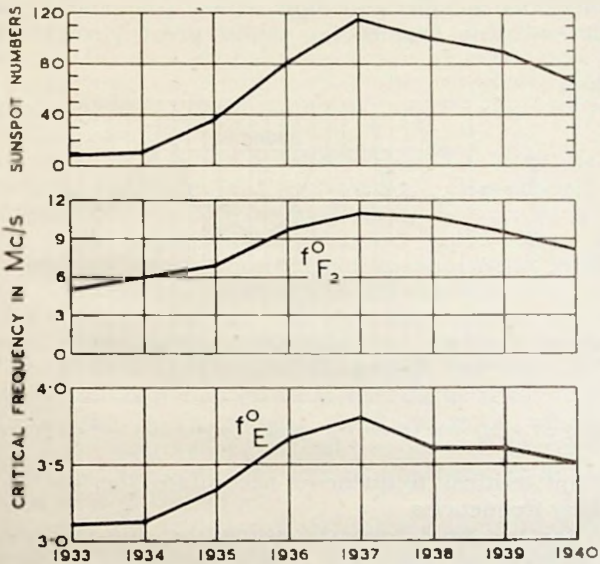


Fig. 18 : Comparison of yearly means of sunspot numbers and of critical frequencies of the E and F_2 layers at noon

at night. In summer, however, there is a large increase in virtual height after the sunrise drop, and the layer during the day is at a far greater height than at night. Both in summer and winter there is a further decrease around sunset and then a gradual increase to the night-time values.

Not enough is known about the virtual height variation of F_2 to say whether there is any considerable change in connection with the sunspot cycle, though there is some indication that such a change may occur. Fig. 19 shows the variations in the F_2 minimum virtual height during the years 1936 and 1937.

As we shall later see, changes in the virtual height affect the frequencies to be used for long-distance short-wave working, as well as do the variations of critical frequency. It may here be mentioned that in general small virtual heights and high critical frequencies lead to higher working frequencies, whilst great virtual heights

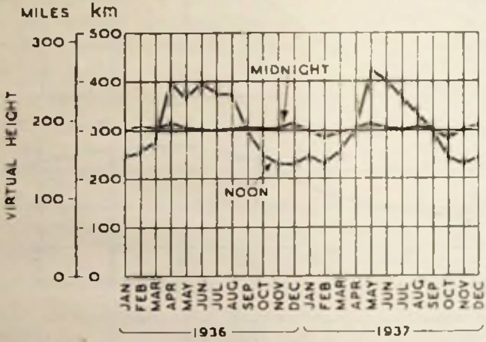


Fig. 19 : Monthly mean of noon and midnight virtual height for F and F_2 layers

and low critical frequencies necessitate the use of low working frequencies.

So far we have been discussing the diurnal, seasonal and long-period changes in critical frequency on the basis of measurements made at Washington. The variations will, however, be similar for all middle latitudes of the Northern Hemisphere. But in practical communication we shall interested in ionosphere conditions, not at the location of the transmitting station, but over most of a transmission path which may well extend to the other side of the world. And the condition of the ionosphere varies with latitude and longitude. The latter we have already taken account of, for we have

31. Critical Frequency Variations in the Southern Hemisphere

discussed the variations with time of day, which is another way of saying the same thing. But, at any one time of day we should expect considerable variations in the ionisation, and hence in the critical frequency, with latitude, for with changing latitude the zenithal angle of the sun alters. The influence of the sun at the equator should, for example, be greater than in higher latitudes, for there it is more directly overhead and hence its radiations are stronger. We cannot say a great deal about this matter at present but our expectations are in general borne out. For as the equator is approached the critical frequencies of all the layers do become higher, both for day and night and at all seasons of the year.

South of the equator we should expect similar seasonal variations, to those of the Northern Hemisphere, but displaced from these by about six months, since when it is winter in one hemisphere it is summer in the other. That is to say, we should expect that in equivalent latitudes in the Southern Hemisphere the critical frequency would be of a similar value but with variations opposite in phase to those of the Northern Hemisphere. In June we should look for low day values of F_2 critical frequency in the Northern and high day values in the Southern Hemisphere, and *vice versa* in December. In June again we should expect night values of F_2 critical frequency to be relatively high in the Northern and low in Southern Hemispheres and *vice versa* in December.

But again the F_2 behaves in an anomalous manner, for its variations in the Southern Hemisphere are not by any means in opposite sense to those in the Northern. We saw that its anomalies in the Northern Hemisphere could be explained by the heating effect and that the theory of its production by ultra-violet sunlight is not invalidated by reason of these anomalies. But the heating effect itself would vary in opposite sense in the two hemispheres, so the critical frequency variation should still be in anti-phase, the maximum occurring in one hemisphere six months after that in the other.

In point of fact the daytime critical frequency in the Southern Hemisphere varies in an *altogether* different manner to that in the Northern, high values occurring at the equinoxes and relatively low values in summer and winter. So that we have low values of day critical frequency in June in *both hemispheres.*

We cannot, in a book like this, speculate upon the reasons for these anomalous variations, nor go into details of the theories advanced to account for them. In practice, we shall simply take account of the measured—or predicted—critical frequency over the whole transmission path and relate our working frequency to this.

But it is interesting to note that the implication is that the ionisation in the F_2 layer is only partly—though perhaps it is largely—due to ultra-violet sunlight. For the most acceptable explanation of the F_2 anomalies is that they are due to components in the ionisation which are not produced by ultra-violet sunlight, and therefore, do not vary in accordance with the sun's position. Summarising *all* the variations in F_2 ionisation it appears that they may be due to :—

(1) A long-period effect due to the variation in ultra-violet radiation over the solar cycle.

(2) An anomalous seasonal effect, described by Appleton and attributed by him to different atmospheric conditions in summer and winter (e.g., heating or the presence of a light gas in summer). This seasonal effect causes high values of ionisation in local winter and low values in local summer, which is just the opposite of what we might have expected.

(3) An annual effect, suggested by Berkner and Wells, and thought to be due to ionising radiation from outside the solar system. It is in the same phase in both hemispheres and tends to produce maxima in December and minima in June.

(4) A bi-annual effect suggested by Eckersley, and thought to be due to increased radiation from the sun affecting the earth at the equinoxes. It tends to produce maxima in both hemispheres at the equinoxes.

If all these effects are added algebraicly then the F_2 variations in both hemispheres are reasonably accounted for. Incidentally they help to explain the Northern

Hemisphere drop in daytime critical frequency which, from Fig. 17, is seen to occur in mid-winter.

As far as variations in the E and F_1 critical frequencies in the Southern Hemisphere are concerned, there are no anomalies, the variations being similar in degree but opposite in phase to those of the Northern Hemisphere. This incidentally is further evidence of their causative agent being, in its entirety, the ultra-violet light from the sun.

In the next chapter we shall go on to see how the critical frequency values, and their variations, are likely to affect the working frequencies of use for short-wave communication.

CHAPTER 5

LONG DISTANCE TRANSMISSION

SO far we have dealt mainly with the behaviour of a radio wave when it is sent vertically upwards, so as to enter the ionosphere at right angles to its lower boundary. This direction—at right angles to the boundary of a refracting medium—is called the " normal " to the refracting surface or boundary. A ray of radio energy entering a layer in this manner will, if it is of such a frequency as to be reflected, return to earth at a point immediately beneath that where it entered the layer. Rays which are going to return to earth at points distant from the transmitter must go upwards so as to strike the refracting layer at an angle to the normal, and the greater the distance to the point on earth where they are to return, the greater must this angle be. So for any layer height which may be encountered there is, for every distance out from the transmitter, a definite angle at which the wave will enter the ionosphere. Fig. 20 should make this clear, and also help to explain some of the terms which may come into this chapter. From the Figure it will be seen that (*a*) for a given layer the greater the distance, or (*b*) for a given distance the lower the layer, the greater will be the angle of incidence.

32. Obliquely Incident Rays

Earlier we were considering a transmitting aerial from which rays were sent out at all angles, both vertical and horizontal, and in such a case there would be rays striking the ionosphere at a large number of different angles of incidence. In a practical short wave station we should arrange things somewhat differently to this, and should see that our radiated energy was more in the nature of a " beam " or " pencil " of rays, going up at a suitable angle according to the distance over which we wished to transmit, as pictured in Fig. 21. Even so there would still be a large number of rays going up side by side and making a number of different angles of incidence at the ionosphere boundary.

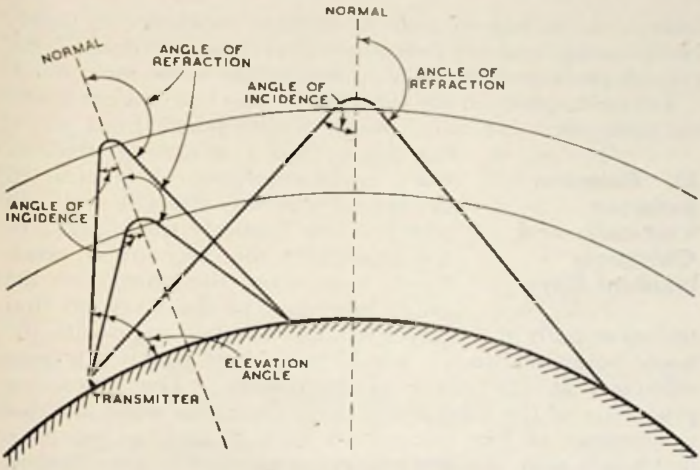


Fig. 20 : Showing how the angle of incidence varies with transmission distance and with layer height

Now such obliquely incident rays can be returned from the ionosphere with a much lower ionisation in the layer than is required to return a vertically incident ray. Or—another way of looking at it—with a given ionisation density in the layer the greater the angle of incidence the higher the frequency which will be returned. Thus the critical frequency for vertical incidence is related to the highest frequency returned at oblique incidence—the latter being the " maximum usable frequency " (MUF) for the distance considered.

The theory concerning the behaviour of obliquely incident rays and relating the penetration frequency at

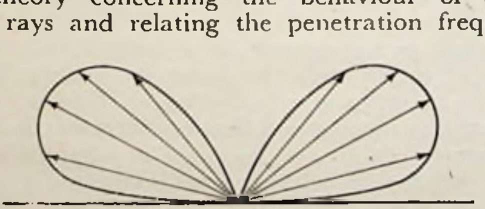


Fig. 21 : Section of radiated energy when using aerial directive in the vertical plane

oblique incidence to that at vertical incidence is highly complicated, and we cannot go into it in any detail here, though perhaps a few of its points ought to be mentioned.

Let us imagine, for the time being, that both the earth and the ionosphere are flat. When a wave passes from air or free space into a refracting medium (such as the ionosphere) it is refracted or bent from the straight path in which it was formerly travelling. In the ionosphere the direction of bending is away from the normal to the lower boundary of the layer, so that the wave path in the layer makes a smaller angle with the lower boundary than it would have done had it not been refracted at the point of incidence. The refracting properties of the ionosphere upon the radio wave are due to presence of free electrons within it, and, as we have previously seen, the electron concentration is not constant throughout the medium, but increases with height up to the point of maximum concentration. Suppose, however, that the free electron density within the layer was itself in a series of " layers ", as in Fig. 22, each thin " layer " having a constant electron density which was greater than that of the " layer " next below it. Then we should get the effect shown in the Figure, the wave being refracted each time it came to the under-edge of a " layer " of greater electron density.

33. Relation Between Vertically and Obliquely Incident Rays

It is easily seen that its direction would be so altered by the successive refractions that it would strike each succeeding " layer " more and more obliquely. Eventually it would strike one of the " layers " so obliquely that the angle of refraction would be greater than 90°, i.e., no energy would penetrate into the " layer ", but the wave would be " totally reflected " at its surface and sent downwards again, as at x in Fig. 22.

In actual fact the electron density increases continuously, in the manner which was described in Chapter 2, and in Fig. 8, so that the refractive process is constantly increasing as the wave penetrates further into the layer. Perhaps a good way to picture the effect is to imagine the top part of the wave as being further into the ionosphere than the bottom part and, therefore, because it is in a region where the free electrons are denser, travelling faster* than the bottom part. If the top part travels faster than the bottom the wave direction cannot be in a straight line, but will bend away from the regions of high electron density towards those where it is low, and the wave is, therefore, bent away from the normal. As the process is

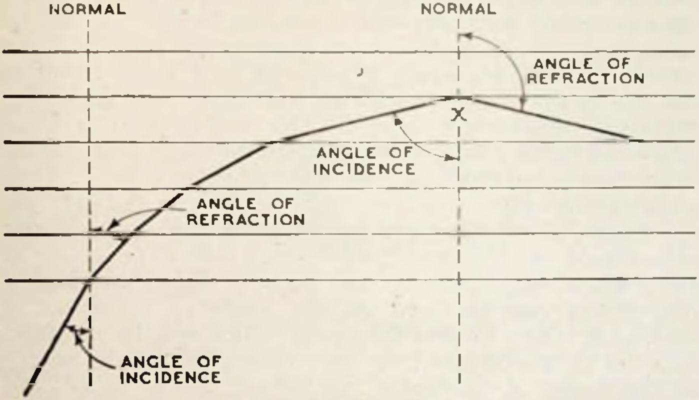


Fig. 22 : Showing successive refractions at imaginary thin layers of constant electronic density, each layer having greater electronic density than that next below it

continuous the wave will curve round in a gradual sweep and eventually arrive at the lower boundary again.

In order to understand the relation between the critical frequency at vertical incidence and the MUF at oblique incidence it may be best to look upon the ionosphere as having a refractive index, which is smaller than that of ordinary air. The amount of " bending " to which a ray of radio energy is subject when it passes from one medium to another depends upon the refractive index of the new medium, and also upon the angle of incidence, the relation

* This is not a mistake. We are now speaking of wave velocity, *i.e.*, the velocity of individual waves, as distinct from the velocity of a whole group. This latter is the group velocity and is less in the ionosphere than in ordinary air, whilst the wave velocity is greater.

being such that the sine of the angle of incidence is equal to the sine of the angle of refraction multiplied by the refractive index. Therefore at the top of the path, where the wave is travelling parallel to the ionosphere boundary and the angle of refraction is thus 90°, the refractive index will equal the sine of the angle of incidence, because the sine of 90° is 1·0. The refractive index at any level in the ionosphere depends upon the free electron density: the greater this is the smaller is the refractive index, which thus gradually decreases with increasing height. It will be seen, therefore, that the wave will penetrate into the layer until the refractive index is reduced to a value equal to the sine of the angle of incidence and that it will then start travelling downwards again. This means that it must penetrate furthest into the layer when the angle of incidence is 0° (vertical incidence) and least when it is 90°. Thus with a given refractive index (given electron density) the wave will penetrate less and less far as the angle of incidence is increased, or, if it is allowed in each case to penetrate to the point of maximum electron density, higher and higher frequencies can be used as the angle of incidence is increased. The highest frequency which will be returned is equal to the critical frequency multiplied by the secant of the angle of incidence (the secant is 1·0 at 0° and infinity at 90°. It is not possible, of course, to strike the ionosphere with an angle of incidence of 90° in practice).

The curvature of the ionosphere is responsible for some modifications to this law. Because of the curvature the angle between the wave path and the thin "layers" of electron concentration which we used in our illustration would alter with increasing height. The result of this is that the wave path is completely turned round in a region where the refractive index is higher (the electron density is lower) than would be the case if the ionosphere were flat. Also because of the curvature there is a limit to the angle of incidence which can be made at the lower boundary of the ionosphere, for there comes a time when this angle no longer alters in spite of alteration of the "elevation" angle of the ray. Summing up the effect of the curvature it may be said that (*a*) it results in higher frequencies being returned for any angle of incidence than would be the case if the ionosphere were flat, and

(*b*) the modification to the flat earth case which it introduces is least when there is a thin sharp reflecting surface of ionisation and greatest when the ionosphere gradient is low, i.e., when the electron density increases only slowly with height.

There are other modifications due to the presence of the earth's magnetic field, but we shall not find it necessary to discuss them here.

It will be clear, then, that in a curved ionosphere the highest frequency returned (MUF) for a given maximum density of electrons in the layer will increase as the angle of incidence is increased, that the increase will be greater the lower the layer, and that, for a given set of ionosphere conditions of height and critical frequency the greater the distance of transmission—up to the limit it is possible to achieve in one hop—the greater will be the MUF. If, then, we know the critical frequency for any time—or have available predictions of what its value should be—we can calculate the MUF for any distance by multiplying it by a factor which has been worked out on the basis of the theory we have just briefly discussed. We may call this the " MUF factor " and it will obviously vary for the different layers because of the different heights at which they lie. Also for the F layer—the principal refracting layer—it will vary considerably for different seasons and also, to some extent, for different times of day, because the height and thickness of the layer varies in this way.

Table II gives some typical MUF factors for transmission by way of the E or F layer over various distances at different times of day and seasons.

34. The Maximum Usable Frequency

It is obviously not possible to take account of all the variations in layer height and thickness in a Table such as this, and, as has been stated, the accurate derivation of the MUF factors is complicated. But the factors given will yield MUF values which are near enough for most practical purposes, though a certain amount of interpolation for times of day, seasons and distances not given, will be found to be necessary. The F_1 so seldom acts as the refracting layer that we need not bother to consider it.

TABLE II

MUF FACTORS FOR DIFFERENT DISTANCES

	Transmission Distance, in Miles						
	310	620	930	1240	1550	1860	2200
E LAYER (all seasons).	2·0	3·4	4·4	5·0	—	—	—
F OR F2 LAYER							
Winter—							
Midnight	1·2	1·4	1·8	2·1	2·6	2·8	2·9
Sunrise and Sunset	1·3	1·9	2·4	2·9	3·2	3·5	3·6
Noon	1·2	1·8	2·3	2·8	3·1	3·4	3·5
Summer—							
Midnight	1·2	1·4	1·7	2·1	2·4	2·6	2·8
Sunrise and Sunset	1·3	1·6	1·9	2·4	2·6	2·8	3·0
Noon	1·2	1·5	1·8	2·3	2·5	2·7	2·9

If, therefore, having available curves like those of Fig. 15 showing the critical frequency at all times of day, we wish to ascertain the MUF for transmission over any distance, then we read off the critical frequency for the appropriate time *at the centre of the transmission path*, and multiply it by the appropriate factor from the Table. The maximum distance over which it is possible to transmit with one hop is determined by the height at which the refracting layer lies, and is, on the average, about 2,200 miles by the F or F_2 layers and 1,250 miles by the E. Beyond 2,200 miles the transmission will be by multiple hops, and this case will be dealt with later.

It seems appropriate at this stage to mention that there is a low limit to the range of elevation angles suitable for transmission. Those waves which are sent out at very small elevation angles—almost tangentially to the earth's surface—are not of much use for long-distance transmission because of absorption in the ground near the transmitter. For this reason, waves sent out at less than about 4° to the ground may be regarded as useless. The useful angles are, therefore, from 4° upwards, according to the distance over which we desire to transmit, and naturally we should arrange to radiate as much of the energy as possible at angles near to that appropriate to this distance. For multiple hop transmission over long distances the high-angle rays also would not be of much use, because waves sent out at such high angles would have to make too many hops. The angles between about 20° and about 4° to the horizontal may be regarded as being those useful in such long-distance transmission.

It would be useful to have available a set of curves from which one could read off the MUF directly, and these can be prepared by calculating the MUF for each hour of day and for all the distances given in Table II. Fig. 23 shows such a set of MUF curves, obtained from the critical frequency curves of Fig. 15 and by the use of the factors given in Table II. From these we can see at a glance what is the highest frequency usable for transmission over any distance.

Transmission on short waves is usually by way of the F or F_2 layers. The ionisation in the E layer is not usually great enough for it to act as the refracting layer, though we should always bear in mind that, during the day, if we use a frequency far below the F layer MUF appropriate to the distance over which we are transmitting the wave may be refracted at the E, and not reach the F at all. Around noon in summer, however, the E ionisation does reach a high enough value for the layer to " blot out " the F_2 layer entirely for transmission at certain angles, during which time the E is the *controlling* layer for short-wave transmission. What this means is that, at the elevation angles for which it occurs, the E-layer critical frequency and the angle of incidence the ray makes at the *E determine the MUF*, and any frequency which is high enough to penetrate the E at this elevation angle will also penetrate the F. It will not occur for the shortest distances because in this case the angle of incidence at the E is small and there is thus more tendency for the wave to penetrate the layer. Then at an angle corresponding to a certain distance it will occur and will continue out to the maximum distance possible for transmission by the E. Beyond this the F_2 layer MUF will still continue to rise because the angle of incidence at that layer still increases

35. Effect of the E Layer

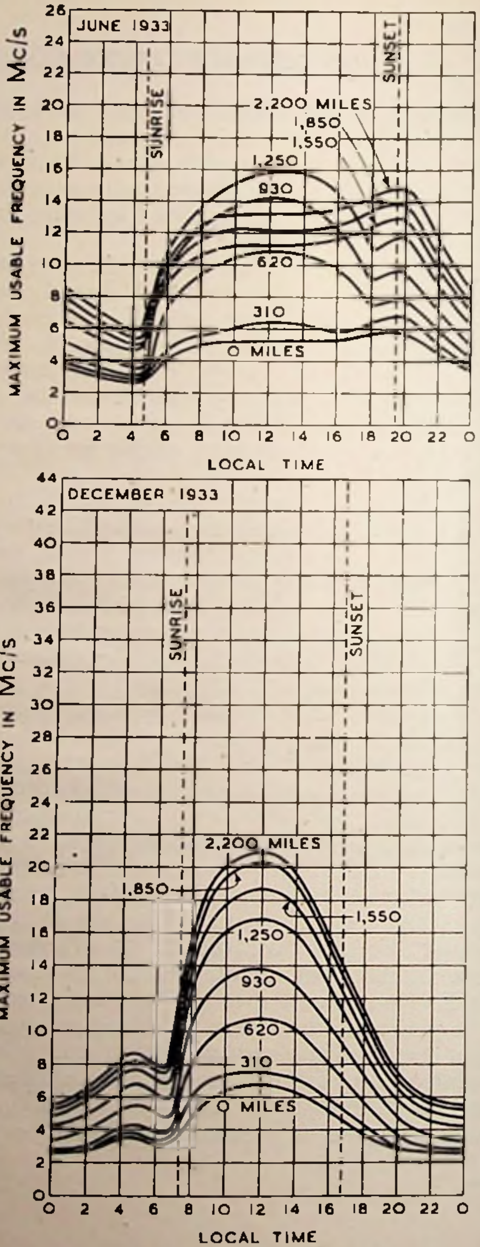
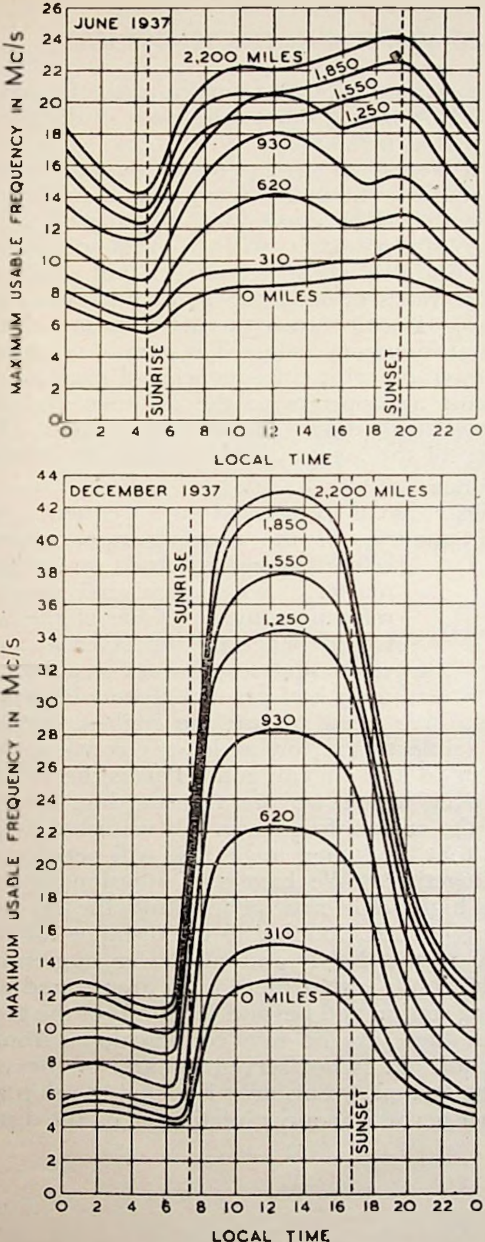


Fig. 23 : MUF's for various distances, obt
by the use of



h the critical frequency values of Fig. 15.
brs of Table II

with increasing transmission distance, and so the MUF is again controlled by the F_2 out to the limit of distance for transmission by that layer.

In Fig. 23 it is easy to see the distances and times for which the E controls the MUF by noting the " humps " which occur in the MUF curves and which peak around noon. In the afternoon, it will be remembered, the E layer ionisation falls, whilst in summer that of the F_2 is still rising, so towards evening the F_2 again takes control for all distances. In calculating the MUF for any distance from the critical frequency values it is merely necessary to multiply the daytime critical frequencies of each layer by the MUF factor appropriate to the distance and layer. Then whichever calculation yields the highest MUF indicates which of the layers will control the transmission over that distance.

It will be clear that if we worked on a frequency below the critical frequency, *all* the upgoing rays would be refracted and returned to earth, no matter at what angle they struck the refracting layer. None of the waves would penetrate the layer at such a low frequency. But, as we shall see, to work on a frequency as low as this is extremely bad practice if we wish to cover long distances. We need to use the highest frequency which it is possible to use, or, at least a convenient one which is as near to it as we can get. This is the MUF for the distance over which we are transmitting, and it is always above the critical frequency. If we use this MUF then it means that the higher angle rays will penetrate the ionosphere altogether. We have the situation shown in Fig. 24 ; the high-angle rays penetrating the layer and those at the angle corresponding to the distance for which our frequency is the MUF and all lower angles being returned. The result is that there is an area round about the transmitting station and beyond the limits of the ground wave in which there are no rays coming down from the ionosphere at all. It is not served by any of the waves radiated by the station. This area is called the skip zone, and the distance across it in any direction—i.e., the distance

36. Skip Distance

between the transmitter and the point where the first refracted ray reaches the ground—is the skip distance, the waves being pictured as "skipping" over the area.

The dimensions of the skip zone and of the skip distance will depend entirely on the ionisation of the layer and on the frequency used, and they will thus vary for a given frequency with time of day, season of year and phase of the sunspot cycle. They will not vary with the amount of power radiated, since no increase in power makes any difference as to whether the wave penetrates the ionosphere

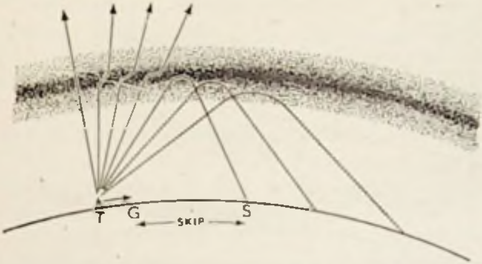


Fig. 24 : Showing the reason for the existence of a skip zone

or not—that depends simply on the frequency and the ionisation prevailing.

Now the distance at which a given frequency is the MUF is also the skip distance for that frequency, for at the angle of incidence appropriate to that distance all higher frequencies will penetrate the layer. The MUF and any *lower* frequency will be refracted so that the wave is receivable at the distance considered, though if the frequency is decreased *much* below the MUF the attenuation due to ionosphere absorption will increase, and signal strength will therefore be reduced. So from the MUF curves—such as those of Fig. 23—we can read off, for any time of day (interpolating where necessary)—the skip distance appropriate to any frequency. This will perhaps be more clearly shown if we plot the result in a curve of skip distance against frequency, as has been done for four times of day for Winter 1937 in Fig. 25. A study of this graph will yield quite a lot of information relating to skip distance. If, for example, we are interested in communicating over a fixed distance—say 1,000 miles—we can see what frequencies would skip at each time of day and thus which of our available frequencies we must use in order to avoid skipping. If on the other hand we are interested primarily in the performance of a single frequency—say 14 Mc/s—we can see from what distances it would be usable for each time of day, and over what distances it would be unusable owing to skip.

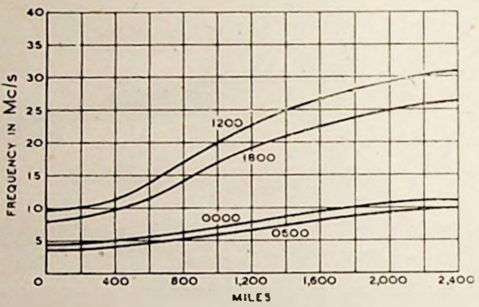


Fig. 25 : Skip distances for various frequencies at different times of day

To make sure, then, that none of the waves intended to reach a certain location penetrate the ionosphere and cause that location to fall within the skip zone, we must be certain that we always use a frequency not higher than the MUF appropriate to that distance.

It has been stated already that we should endeavour to work on a frequency *near* to the MUF appropriate to the distance over which we are communicating. The use of any higher frequency would only result in penetration of the ionosphere and loss of the radiated energy, whilst the use of too low a frequency will result in great losses due to ionospheric absorption. The curves of MUF which we shall have available will usually be those of *predicted* MUF and these will be based on the average of say a month's actual measurements. So

37. Useful Frequency Bands

what we shall actually have is a predicted *monthly average* of MUF. The variations of MUF from the monthly average, on days which are not disturbed, are not usually very great, but we should allow a small margin to account for day-to-day variations. So if we fix on a frequency for any hour of day 10% below the monthly average MUF, we can regard this as a safe " high limiting frequency " and consider ourselves permitted to use this or any lower frequency.

But as we decrease in frequency from the high limiting frequency the wave will be more and more attenuated because of absorption, which will occur mainly in the D and E layers. Thus there will also be a " low limiting frequency " below which we shall encounter such heavy absorption that the signal strength at the receiver will be too low. The *band* of frequencies between these two limits constitutes the useful frequency " band " and it is within this band that communication will be most effective.

38. Absorption

Ionosphere absorption is due to the collisions which occur between the electrons which the wave has set in motion and the neutral gas molecules. When this happens the energy which the electron has acquired from the wave is partly transferred to the gas molecule and partly dissipated as heat, so that energy is absorbed in this way from the wave. The amount of radio energy absorbed depends on the frequency with which the collisions take place, and thus upon the density of the air, or gas pressure. For, the more dense the gas molecules, the greater is the likelihood of collision with a moving electron. The absorption is, therefore, much greater in the lower layers of the ionosphere, where the gas pressure is much higher than it is in the higher ionosphere. The amount of absorption is inversely proportional to the square of the frequency, it being greatest, of course, when the motion acquired by the electrons is greatest. This, as we have seen, occurs when the frequency is low. As to the variation in the amount of absorption with time of day it is safe to say that it is at a maximum at noon and practically negligible at night.

The low limiting frequency is, however, very difficult to assess. It will vary, not only with the ionisation of the D and E layers—to which the ionospheric absorption is proportional—but also with the radiated power. For, unlike the high limiting frequency—which is independent of power—the low limit can be reduced ; i.e., the absorption overcome, by an increase in power radiated. It will also depend upon the noise conditions prevailing at the receiving location and upon the type of receiver in use, for these will determine the minimum field strength necessary to produce a workable signal. Atmospheric noise, besides increasing in an approximately square law manner with decreasing frequency, varies greatly at different locations on the earth's surface, and at any one location with time of day and season. Generally speaking, it is particularly high in tropical regions and low in temperate zones, but the precise nature of its variations cannot be stated with any degree of accuracy.

From all this it will be apparent that the low limiting frequency cannot be very precisely defined, and is perhaps best determined by experience. The guiding principle should be " Work as near to the upper limiting frequency as possible ". A rough estimation of the frequency below which one should not work is obtained by taking a frequency 50 per cent. of the MUF at any time of day, and although this is not intended to be a precise limit, it will be found to indicate pretty well the lower limit which is a safe one in practice.

We have thus enclosed between high and low limits the band of frequencies on which we should expect communication to be practicable for any distance at any time of day.

39. Multiple Hop Transmission

So far we have been discussing the way in which the useful band of frequencies may be obtained from the MUF curves for any distance up to the limit of one hop. The geographical part of the ionosphere which controls such transmission, and for which the MUF should be considered, is the part half way between transmitter and receiver. Thus the local time for which the MUF is considered is that at this point, and in this way the difference in longitude between transmitter and receiver is allowed for. For places not greatly different in latitude ionosphere conditions are not *greatly* different at the same local time, but if there are great differences in latitude then MUF data for different latitudes should be considered conjointly, if such is available.

In the case of multi-hop transmission, the matter becomes somewhat more complex, but we should always remember that we are never concerned with a single ray striking the ionosphere at certain definite points, but with a whole bunch of such rays, which do so at widely different points. It is, therefore, quite satisfactory in practice to examine the whole transmission path between transmitter and receiver, and, from the MUF data for various latitudes which is available, to define the lowest high- and the highest low-limit frequency for the most oblique transmission which occurs on the path at any time of day. The useful band is that between these two.

Nothing further can be said on this subject at the moment, but, given the MUF data from the different observing stations, there is nothing very difficult about applying it to most transmission paths.

We have now dealt with a few practical considerations affecting transmission by the ionosphere under normal conditions. In the next and final chapter we can consider some of the abnormal conditions which from time to time occur—conditions caused by disturbances in the ionosphere, such as have marked and very undesirable repercussions upon short-wave communication.

CHAPTER 6

IONOSPHERE DISTURBANCES AND OTHER ABNORMALITIES

40. Cause of Disturbances

AS we saw in Chapter 4, the ionisation of the upper atmosphere is brought about mainly by the action of the sun's ultra-violet radiation, and it therefore, follows fairly closely the activity of the sun itself. It thus varies markedly over the 11·1-year solar cycle, as is well shown by an examination of the measured critical frequencies over a number of years. Fig. 18 showed us for example, that the correlation between the annual means of the sunspot numbers and those of the critical frequencies is quite good. Thus the sunspot numbers are indicative of increased solar activity and hence of increased emission of the ionising radiations at the maximum of the cycle, which causes the ionisation and the critical frequency of the layers to rise. In this respect there is a difference between the sun's radiation of ultra-violet and of visible light, for radiation of the latter remains constant throughout the cycle.

The net result of the increased ionising radiation at the solar maximum is that short-wave propagation is improved, for, with its increased electron population the F layer is able to refract much higher frequencies than is the case at the minimum, and we are thus able to extend upwards our range of useful frequencies for short wave communication. Our high-limit frequency is much increased. At the same time the ionisation in the lower layers—whilst it is certainly greater at the maximum than at the minimum—is not increased in the same *degree* as that of the F, and so the low limiting frequency—set by the absorption occurring in the lower layers—is not unduly raised. The result is that the band of useful frequencies is broadened with increased solar activity, and beneficial results to short-wave communication ensue.

But the solar activity produces other effects in the ionosphere, for, in connection with it there occur, from time to time, great upheavals on the sun, from which *extra* radiation is emitted, such as cause abrupt changes in the ionisation in various parts of the atmosphere. This gives rise to abnormal behaviour on the part of radio waves, and to a condition often leading to their complete failure as a means of communication. These disturbances are of two distinct kinds, having, so far as their effect upon the short waves is concerned, characteristics of a very different nature. One of these is the " sudden ionosphere disturbance " and the other the " ionosphere storm ". It is important, therefore, to distinguish between the effect of the *general* increase in the sun's activity—which tends to *improve* short wave conditions—and the effect of local disturbances which occur on and within it. For these latter are now well proven to be the causes of the short-wave disturbances we have already mentioned, and, which we shall now discuss in detail. We will consider first the " sudden ionosphere disturbance " not because it constitutes the most serious form of interruption to short-wave communication but because the correlation between the terrestrial and the solar phenomena is better established than in the case of the " ionosphere storm ".

41. Sudden Ionosphere Disturbances

Fairly frequently there occur eruptive disturbances on the sun from which are emitted, not only huge quantities of gaseous matter, but also a large amount of energy in the form of radiations of various wavelengths, including those of visible light. The disturbances are thus observed from the earth as " flares " of bright light. They often take place in the vicinity of sunspots, and if they occur on that side of the sun which is facing the earth and are of sufficient intensity, they produce certain terrestrial effects which occur only on the earth's sun-lit hemisphere. The effects are produced no matter on what part of the visible disc of the sun the eruptions occur. The radiations from the disturbances would appear to be sent out in all directions and, as the terrestrial effects start at the same time the eruption is observed, they must be due to waves travelling at the speed of light. The eruptions usually last but a few minutes, though their terrestrial effects are still felt some time after they have died down. The waves emitted would appear to consist largely of wavelengths relatively near to those of the visible part of the spectrum, and so able to penetrate the outer part of the atmosphere, where the gases are transparent to such waves. They do not, as far as is known, have any appreciable effect on the upper layers of the ionosphere. They continue on until at a height of about 38 miles they reach the ozone layer, where the waves are apparently absorbed, the absorption band for this gas including longer wavelengths than those for gases which are found in the higher atmosphere. Simultaneously with the arrival of the waves in the ozone layer, there appears an ionised region of very great intensity at the height of this layer, no doubt produced by the waves. The intensity of the ionisation continues to grow so long as the waves continue to arrive, but when the eruption dies down and the wave radiation ceases, the ions and electrons start to re-combine. Though they do this at a very rapid rate owing to the great density of the gas, an hour or more may elapse before normality is restored, owing to the great intensity of the ionisation which the waves produced in the first place. Coincident with the time of observation of a bright solar eruption there is often a sudden cessation of all short-wave signals. This result of the eruption is the "sudden ionosphere disturbance", and the correlation between the two phenomena is now conclusive.

The effect on short-wave reception is very impressive, for everything appears to go "dead", even the background noise disappearing. All frequencies within the short-wave range may be affected, though the disturbance is usually more intense and lasts longer on the lower frequencies. These fade out before the high frequencies, and do not come in again until after the latter have done so. On the frequencies affected, signals usually fade away entirely within a minute or two, and may not return until anything up to two hours later, signal strength increasing from zero in a more gradual manner than that in which it was reduced. Only the sky waves are affected, and only those over transmission paths passing through the sunlit hemisphere. It has been found that on transmission paths which pass through low latitudes—where the sun's rays are more perpendicular—the disturbances are more intense than on paths which pass entirely through high latitudes.

The cause of the short wave failure is the sudden production of greatly increased ionisation in the D layer, which is situated at about 38 miles above ground—and in which the radio wave causes collisions between the electrons and the neutral atoms. In so dense a region the number of collisions occuring is very great, and as at each collision energy taken from the wave is expended ; the result is such a heavy absorption of energy from the wave that it is often completely lost. The absorption is greater on low than on high frequencies, whilst it is greater in low than in high latitudes, because the ionisation produced is greater where the sun's radiations are most intense. For this reason transmission paths in the dark hemisphere are unaffected by the disturbance. When the ionising radiation ceases to arrive—after the eruption has died down—the electrons and ions start to recombine, and the absorption of the waves decreases. Owing to the density of the gas, the recombination rate is high ; and, as the ionisation disappears, the short-wave signals come in again, the higher frequencies first.

Such disturbances are accompanied by a sudden brief fluctuation of the earth's magnetic field, caused by the sudden presence within it of large numbers of moving ions, whose movement is in reality a vast electric current carrying an associated magnetic field. This geomagnetic disturbance, like the ionosphere disturbance, is confined to the sunlit hemisphere, and is more intense in low than in high latitudes, and in these and other respects it is quite different in character from the " magnetic storm ". A radio " fade out " of this kind, with its accompanying magnetic disturbance is illustrated by the field strength and magnetic intensity records shown in Fig. 26.

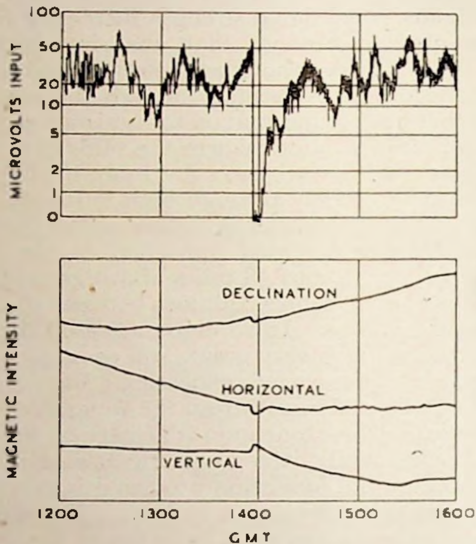


Fig. 26 : Sudden disturbance of the ionosphere on April 6, 1936, as shown by fade-out of 13·525 Mc/s transmission from GLH, Dorchester, observed at Riverhead, New York, and (below) coincidental terrestial magnetic perturbation recorded at Cheltenham, Maryland, U.S.A. *(Reproduced from "Sudden Disturbances of the Ionosphere," by J. H. Dellinger)*

The ionosphere storm constitutes the major form of disturbance to short-wave communication because, while usually its effect is not so intense as that of the sudden ionosphere disturbance, it is of much greater duration. During ionosphere storms short-wave signals on wavelengths normally well received drop to a very low level, and often disappear entirely. There is almost always a great increase in the amount of fading experienced, and a prevalence of the particular type known as flutter fading. The higher frequencies are most affected, and there is no discrimination between the sunlit and dark hemispheres. Transmission paths in low latitudes are less affected, however, than those passing through high latitudes, the paths most severely disturbed being those

42. Ionosphere Storms

which pass through certain zones centred on the geomagnetic poles. Ionosphere measurements indicate that the principal effect is in the upper part of the ionosphere, and that the F layer is usually at an abnormally great height during the disturbance, and has an abnormally low

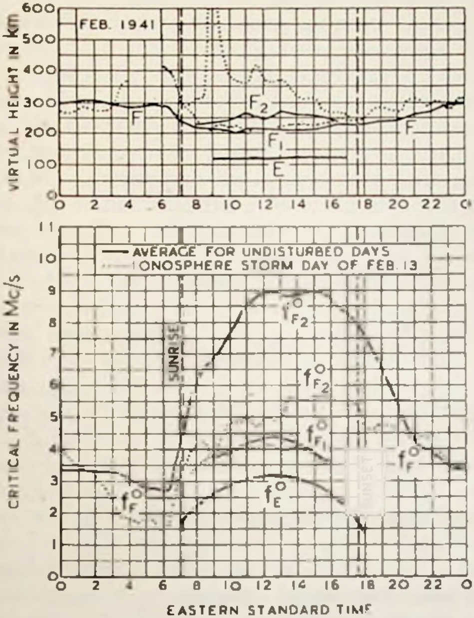


Fig. 27 : (*Courtesy, Proc. I.R.E., New York.*)
Virtual heights and critical frequencies of the ionospheric layers, observed at Washington in February, 1941, by the U.S. Bureau of Standards

ionisation density. The storms usually last for several days. The critical frequency and virtual height graphs of Fig. 27 show how the values for these quantities depart from the normal on days of ionosphere storminess.

It is thought that disturbances of this type are caused by the emission of streams of particles from the sun, and that these corpuscles—which may be particles of ionised calcium—are shot out from disturbances which occur in the vicinity of sunspots. It may be that this solar emission constitutes some of the vapour which is observed in the vicinity of the sun after a bright solar eruption. Although part of the emission is known to fall back upon the sun, some of it appears to be forced outward by radiation pressure until it escapes from the sun's atmosphere. It would thus leave the sun at the same time as the wave radiation which produces the sudden ionosphere disturbance, but the corpuscles would travel much slower than the ultra-violet electromagnetic waves, and so would not reach the earth until some time later. Furthermore, while the wave radiation is emitted in all directions, and so reaches the earth irrespective of the position of the eruption on the visible disc of the sun, the corpuscular radiation seems to be in the form of a cone-shaped jet, with the eruption at its apex. So that, unless it is emitted from a position on the sun which is " pointing " more or less towards the earth, it misses this planet altogether. It has been noticed, for instance, that when a sudden ionosphere disturbance occurs as the result of a bright solar eruption *near the central meridian* of the sun, it is often followed—about 30 hours later—by an ionosphere storm, but that disturbances due to eruptions occurring in other parts of the sun are not followed by a storm. Apart from this, it is often noticed that when an ionosphere storm starts there is a sunspot in a position about 30 hours past the central meridian in the direction of rotation of the sun. So it would appear that if there is a disturbance on the sun near its central meridian—whether visible as a bright solar eruption or indicated by a sunspot—the corpuscles which are shot out do encounter the earth and produce an ionosphere storm about 30 hours after leaving. This would indicate that they had travelled through space at a speed of about 900 miles per second.

It should be added that the correlation between the start of ionosphere storms and the solar phenomena is not by any means perfect, some storms occurring when no eruption has been seen and when there is no spot near the central meridian, though there may be spots in other parts of the sun. This may be because the solar disturbance has not yet produced a visible sunspot, or because the corpuscles are not always emitted in a direction normal to the sun's surface. On the other hand, there is a definite tendency for storms to recur at intervals of about 27 days, corresponding roughly to the average rotation period of the sun, which would indicate that the same disturbance had produced an ionosphere storm during its successive passages across the sun's central meridian.

On reaching the earth's atmosphere the corpuscles are affected by the geomagnetic field, which carries them in the direction of the magnetic poles. Consequently, their effects, both upon the ionosphere and in other ways, are most intense in zones around the poles. A state of turbulence is set up in the ionosphere, particularly in the upper layer, leading to erratic conditions for the refraction of radio waves, with consequent abnormal facing. The F layer then appears to expand and to rise, and the stratification of the ions is upset. The ionisation per unit of volume is thus reduced, so that waves which are normally refracted begin to penetrate the layer. At the same time there appears to be an increase in absorption in the lower layers.

According to one observer of an exceptionally severe storm, the F layer continued to decrease in ionisation and to rise in height until it finally diappeared altogether, when an entirely new layer appeared lower down. This, in turn, behaved in a similar manner until it, too, diappeared, to be followed again by another layer, the time between the appearance and disappearance of a layer being about three hours, and the phenomenon becoming less and less evident until eventually the turbulence subsided.

Ionosphere storms have caused the *entire disappearance* of signals on certain frequencies in a few cases for as long as two days. Though not usually as intense as this, the average time for which conditions remain *subnormal* is between one and two days, whilst they have been known to remain so for nearly a fortnight. During the storms the highest frequencies which the F layer will refract may be reduced by as much as 50 per cent. below normal.

Ionosphere storms are almost always accompanied by violent fluctuations in the geomagnetic field, which phenomenon is called a "magnetic storm". The fluctuations are rapid and irregular in character, and are quite different from the brief magnetic disturbance which occurs during a sudden ionosphere disturbance. Like the ionosphere storm, the magnetic storm occurs at the same time in all parts of the world, and is most intense in polar regions. Though the two phenomena are clearly connected, the start of the ionosphere and magnetic storms does not appear always to be exactly simultaneous, and the ionosphere storms usually persist for some time after the geomagnetic field has returned to a "quiet" state. Magnetic storms are caused by the abnormal movements of the ions in the atmosphere due to the action of the solar corpuscles, the movements of the ions constituting electric currents of great magnitude. These currents have associated magnetic fields which interfere with the normal geomagnetic field, and so produce the magnetic storm.

43. Magnetic Storms and Polar Auroræ

Another effect of the solar corpuscles is the production of the polar auroræ, which are intimately associated with variations in the geomagnetic field, and which occur frequently in the zones surrounding the magnetic poles. In these zones, incidentally, ionosphere and magnetic storms are of greater intensity and occur with greater frequency than elsewhere, so that transmission paths which pass through them are much more liable to disturbance than those traversing lower latitudes. The coloured light of the auroræ is due to the emission of visible rays by atoms of atmospheric gas when subjected to bombardment by the solar corpuscles. The height at which the auroræ occur has been measured, and it is thought that the lower limit of about 55 miles above the earth's surface represents the farthest distance that the corpuscles penetrate into the earth's atmosphere. Although the auroræ are usually confined to the zones around the magnetic poles, there are occasions when they are observed over much greater areas. The implication is that, under these conditions, the stream of solar corpuscles entering the atmosphere is of exceptional intensity, and such occurrences are almost always accompanied by ionosphere and magnetic storms of very great severity.

In the discussion above we have dealt with the two main causes of failure in short-wave communication, brought about by abnormalities in the ionisation of the layers. There is another abnormal condition which must be mentioned, though it does not result in a deterioration in reception, but rather tends to improve it at distances not too far from the transmitter. This is the phenomenon known as " sporadic E ".

44. Reception within the Skip Zone

But first we must say a little more about the effect of the E layer on reception within the normal skip zone. Anyone who has operated a short-wave receiver within the skip zone of any particular station will realise that it is quite untrue to say that *no* signals at all are *normally* obtainable within that zone. Signals of a kind *are normally* obtainable, though they are usually weak and unreliable, are much subject to distortion and fading, and are generally not of a character suitable for reliable communication. They are due to the " scattering " of some of the energy in the radio wave as it passes through the lower layers on its way up to or down from the F. This scattering is due to the presence of " clouds " of electrons in the E layer or between the E and F layers. When the up-going wave strikes these some of the energy is scattered in all directions and eventually arrives again at the earth's surface either directly from the scatter source or after further reflection at another layer. As has been said, this scattered energy provides only a weak and erratic sort of signal and one not to be compared with that due to the refracted wave, but within the skip zone, where it is the *only* energy received from the station, it does provide some sort of a signal. The amount of scattered energy received will be increased when the power radiated by the transmitting station is increased, or when it radiates this energy in the form of a very narrow beam. It will be seen that the scattered energy may arrive from any direction—not necessarily from the direction of the transmitting station—and, generally speaking, the signals due to it are weaker on the higher frequencies than on the lower. Apart from allowing one to hear a station, this type of reception is of little use in practical communication.

Sometimes, however, there are obtained, within the skip zone, signals which are strong and steady and comparable in every way to those due to the normal refracted wave. Indeed the effect is more as if the *ground* wave range of the station had been suddenly extended, and good reception is then experienced from the station out to the distance at which the sky wave would normally be received, without any skip zone being observable. Furthermore, such reception is often experienced on frequencies far above those on which it would normally be obtainable even with the most obliquely incident rays. It is due to the formation within the E layer, of a thin layer of very highly ionised air, such as will cause reflection of frequencies far higher than those which the normal E—or, indeed, the F—is capable of reflecting. This highly ionised layer is known as the " sporadic E " and usually occurs most frequently in summer and during the late afternoon and evening. It does not, however, extend over a very wide area, nor, in any one area remain in existence for very long, but appears and disappears in an erratic manner. It cannot, therefore, be put to any regular use in short-wave communication because it is so unpredictable—one can never tell whether it will be present or not. It can, therefore, be regarded as an ionosphere abnormality—but one which tends rather to improve short-wave communication than to impair it, at least at points relatively close to the transmitter.

This book would not be complete were no mention made of the phenomenon of " fading," though whether the proper place for its discussion is in this chapter is a questionable matter. For it certainly must not be regarded as an abnormality, as, unfortunately, like the poor, it is always with us. The term " fading " is here

45. Fading

used to refer to the random variation of signal strength which is almost always present, to a greater or lesser degree, in short-wave reception, and should not be confused with the " fade out " which we discussed earlier in this chapter.

Fading of the short-wave signal is due to the fact that the signal is comprised of a number of different " rays " of radio energy which have reached the receiver after travelling *via* the ionosphere over paths of different lengths. As the lengths of the different paths are constantly changing due to varying ionospheric conditions, and as the changes in path length for the various rays are not always the same, but vary for the different rays, the energy in the arriving rays " adds up " in a random manner. At one instant, for example, the path length for two of these rays may be such that they arrive at the receiver so that the energy in one reinforces that in the other, whilst at the next instant the path lengths may have changed relative one to the other so that the energy in the second ray may completely cancel out that in the first. It is the " phase " of the different rays which determines the resulting signal strength, and the interference between the different down-coming rays will cause the signal to vary in strength in a random manner. In addition to this there is the fact that during its passage through the ionosphere the wave is split into two differently polarised components, and it will readily be appreciated that because these travel by different ionosphere paths, the polarisation changes will themselves contribute something to the fading.

The sky wave signal is, therefore—unlike that due to the ground wave—almost never of constant strength. The character of the fading varies considerably with varying ionosphere conditions. It exhibits different characteristics in respect of time—" rapid " or " slow " fading—or in respect of intensity—" deep " or " shallow " fading—all depending, as has been said, upon the prevailing ionosphere conditions. Deep fading is especially prevalent near the outer edge of the skip zone, for there, at one instant there may be many rays coming down at the receiver location whilst at the next instant the paths may have so changed that all the rays pass over the receiver and reach the earth at a point more distant from the transmitter.

There is one type of fading which is particularly serious in the case of broadcast reception, since it leads to bad distortion of the received programme. This is known as " selective " fading and is due to the fact that, because the path length in the ionosphere varies with frequency, the fading may be different for frequencies which differ by only a few hundred cycles. A speech or music modulated carrier is made up of a large number of different frequencies, and the " quality " of the received programme is dependent upon these having the same relations as were present in the transmitted programme. If, therefore, the different frequencies are propagated in a different and changing fashion in the ionosphere, the result is that the received programme is distorted.

During ionosphere storms a peculiar form of fading known as " flutter " fading is often experienced. In this the variation in signal intensity takes the form of a fast, rhythmic beat, almost of the nature of a low-frequency oscillation super-imposed on the modulated carrier. It is well described by the term " flutter " fading but, apart from the fact that it is due to disturbed conditions in the F layer, not enough is known about its cause to permit of explanation of the specific reasons for its peculiar character.

Having regard to all that has been said in this chapter, the reader may begin to wonder, not that the received signal is sometimes weak and unreliable, sometimes fluctuating and distorted : but that a short-wave signal travelling *via* the ionosphere is ever receivable at a far-distant point at all. The fact remains that it is, and that the short waves have become the main—and certainly the most economical—medium for long-distance communication. True the received signal is often inferior in quality to that provided by short-range communication systems, but even that defect is being overcome. Devices and operating technique have been and are being gradually improved and perfected with a view to overcoming all the disturbing phenomena which have been discussed in this chapter. Further improvements appear to be imminent and may take the direction of the utilisation of a greater range of frequencies, of the development of technique for a greater degree of ray selection, and of improved aerial arrangements at both transmitter and receiver. It is to be hoped that these technical advances may be accompanied by an improvement—from the ethical point of view—in the uses to which the medium is put, and that the operation of the short-wave services may soon be turned away from the destruction of mankind towards its betterment—away from the furtherance of discord and war towards the promotion of international understanding and goodwill.