EDWARD V. APPLETON

The ionosphere

Nobel Lecture, December 12, 1947

In this lecture I wish to draw your attention to certain features of the electrical state of the higher reaches of the earth's atmosphere. This is a region which human beings have not yet visited, and so the information which we have accumulated about it is of an indirect character.

Now the most striking feature of the atmospheric air at high levels is that it is ionized, and for that reason the spherical shell surrounding the earth at the levels with which we are concerned is called the ionosphere. It has, of course, been suspected for many years that when there is an aurora1 display, yielding intense luminosity, the upper atmosphere must be strongly ionized like the gas in a Geissler tube. But I am not so much concerned today with irregular events of that kind, which occur mainly in high latitudes, as with the permanent shell of ionization which exists at all times and at all latitudes, even above the equator.

There were originally two lines of evidence which suggested that the upper atmosphere might be electrically conducting. In the first place Balfour Stewart, in 1882, put forward the hypothesis that the small daily rhythmic changes of the earth's magnetic field were due to the magnetic influence of electric currents flowing at high levels. Balfour Stewart pictured such currents as arising from electromotive forces generated by periodic movements of the electrically conducting layer across the earth's permanent magnetic field. The movements, he suggested, were largely tidal in character and therefore due to the gravitational influence of the sun and the moon.

The second indication of the possible existence of a conducting layer in the upper atmosphere came from the study of the long-distance propagation of radio waves. The successful communication established by Marconi between England and Newfoundland in 1901 prompted many theoretical studies of the bending of electric waves round a spherical earth. These mathematical investigations of radio-wave propagation round a conducting sphere showed conclusively that Marconi's results could not be explained in terms of wave diffraction alone. Some factor favouring radio transmission over long distances had evidently not been taken into account.

Suggestions as to the nature of this factor were fortunately to hand, for in 1902 Kennelly and Heaviside had independently pointed out that, if the upper atmosphere were an electrical conductor, its influence would be such as to guide the radio waves round the earth's curvature, energy being conserved between the two concentric conducting shells and so not lost in outer space.

The Kennelly-Heaviside theory did not, however, gain universal acceptance, for direct evidence of the existence of the conducting layer was lacking. Opponents of the theory, for example, sought to explain Marconi's results in terms of the refractive bending of the waves due to the stratification of the air and water vapour in the lower atmosphere near ground level.

During the 1914-1918 War, when I served as a radio officer in the British Corps of Royal Engineers, I became interested in the problems of radio propagation and the fading of radio signals. As a result, after the war, when I returned to Cambridge, I began to work on the subject, starting first to develop more accurate methods of radio-signal measurement. The initiation of broadcasting in Britain in 1922 greatly assisted these experiments, for powerful continuous wave senders became generally available for the first time. Measurements of received signal intensity, made at Cambridge on waves emitted by the London B.B.C. sender, showed that, whereas the signal strength was sensibly constant during the daytime, slight fading was experienced at night. A possible explanation was that such fading was due to interference effects between waves which had travelled straight along the ground, from sender to receiver, and waves which had travelled by an overhead route by way of reflection in the upper atmosphere.

We may picture such a state of affairs as shown in Fig. 1. Here we see that radio waves can travel from the sender to the receiver by two paths - one direct and one indirect. Now if there is a whole number of wavelengths in

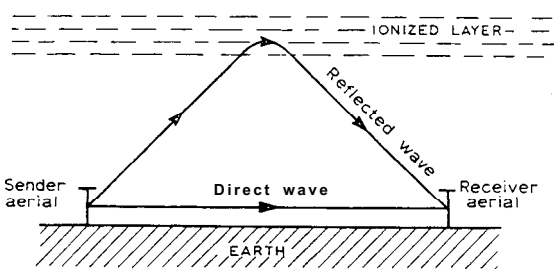

Fig. I.

the path-difference between the ground and the atmospheric ray paths, there will be a maximum of radio-signal intensity at the receiver; while if the path-difference is equal to an odd number of half-wavelengths a minimum of signal will be experienced. Let us suppose now that the wavelength of the radiation emitted by the sender is slowly and continuously altered. This will produce a succession of maxima and minima of signal intensity at the receiver, and, if the number, n, of either is counted, and the initial and final wavelengths (λ_1 and λ_2 respectively) of the change are known, the difference in length, D, between the ground and atmospheric paths, may be found. We have, in fact, the relation:

$$n = \frac{D}{\lambda_1} - \frac{D}{\lambda_2} \qquad (1)$$

if D does not alter sensibly with wavelength. When the path-difference, D, is known, the equivalent height of the reflecting layer can be found by simple triangulation.

The first experiment of this type, using the variation of wavelength (or frequency) was carried out on December 11th, 1924, with the assistance of M. A. F. Barnett, my first research student. The B.B.C. sender at Bournemouth was used and the receiving station was established at Oxford. This experiment immediately yielded evidence of a sequence of artificially produced maxima and minima of received signal intensity. The estimated height of reflection was found to be about 90 km above the ground.

In another series of experiments, the angle of incidence of the reflected waves on the ground was measured. This was done by comparing the simultaneous signal variations on two receivers, one with a loop aerial and the other with a vertical antenna. These results also indicated the reception of downcoming waves from about the same level in the upper atmosphere. The two sets of observations therefore directly established the existence of the Kennelly-Heaviside Layer.

In the winter of 1926–1927, using experimental methods of the same type, I found that, before dawn, the ionization in the Kennelly-Heaviside Layer («E Layer») had been sufficiently reduced by recombination to permit of its penetration. Reflection, however, was found to take place at an upper layer which was richer in ionization and which I termed the «F Layer», the lower boundary of which was found to be situated at a level 230 km above the earth.

Since the positions of the two main reflecting layers were first established, experiments of the kind I have described have been continued and extended. It was soon found that the technique could be so improved as to permit the study of radio reflection of radio waves incident normally on the reflecting layers. This greatly simplified the interpretation of the results. In addition to using the frequency-modulation method of measuring the distance of the reflecting stratum, which has been described above, the elegant pulse-modulation method of making the same type of measurement, which Breit and Tuve had invented in 1925, was also adopted and developed. This method has proved the most powerful tool in ionospheric research. In its application in England, cathode-ray oscillograph delineation of the ground pulse and the subsequent echo pulses was also employed. It was by using a technique of this kind that the phenomenon of magneto-ionic splitting of echoes was discovered by G. Builder and myself. This confirmed, in a direct manner, what J. A. Ratcliffe and I had previously suspected from our experiments on the circular polarization of downcoming waves, that the ionosphere was a doubly-refracting medium due to the influence of the earth's magnetic field. The result indicated that free electrons, and not atomic or molecular ions, were the effective electrical particles in the ionosphere and paved the way for the development of the basic theory of a method of measuring electron densities in the ionosphere.

In the very early experiments on the ionosphere it was customary to use a constant frequency for the exploring waves and study the variation of equivalent height of reflection (hi), as a function of time *(t).* In 1930, however, I suggested that perhaps more information could be obtained by studying the relation between height of reflection *(h')* and frequency *(f),* since, in this way, it would be possible to find the « critical » frequency of penetration for any layer. I had found, for example, that the « critical » penetration frequency for the E Layer, just before dawn, was about 0.75 Mc/s but that its value in summer daytime was about 3.0 Mc/s. It was from considerations such as these that the critical frequency method of measuring upper-atmospheric ionization was evolved. From a general theory of propagation of radio waves in a magneto-ionic medium I had found, in 1927, that the refractive index of such a medium became reduced to zero, permitting the reflection of radio waves at vertical incidence, when

$$N = \frac{\pi e^2}{m} f_o^2 \quad \text{(ordinary wave)} \qquad (2)$$

and

$$N = \frac{\pi e^2}{m}\left(f_x^2 - f_x f_H\right) \text{ (extraordinary wave)} \qquad (3)$$

for the conditions which usually obtain in practice in temperate latitudes. Here N is the electron density (electrons per cc) at the atmospheric level at which the refractive index becomes zero for an ordinary wave of frequency f_o and an extraordinary wave of frequency f_x. The quantities e and m are, respectively, the charge and the mass of the electron, while f_H is the gyro-angular-frequency (eH/mc), with which the electrons spiral round the lines of the earth's total magnetic force, H, at the level of reflection, c being the velocity of light.

Now, if, in Eqs. (2) and (3) the values of f_o or f_x refer to critical penetration, and therefore maximum, values of frequency, the corresponding values of N refer to maximum electron densities for the ionized layer which is just penetrated.

It will therefore be seen that, by finding the critical penetration frequency for either the ordinary ray or the extra-ordinary ray (which may be distinguished by their characteristic polarizations) it is possible to find the maximum electron density for any layer at any given time. For experimental convenience it is usual to employ the value of the ordinary-ray critical frequency f_o (see Eq. (2)), since that quantity is the easier one to determine experimentally.

(It should, however, be pointed out that, if both f_o and f_x are determined for the same conditions, it is possible to calculate f_H and so determine the value of H at the level of reflection. Using this method I have determined the value of the earth's total magnetic force H *at* the level of about 300 km, and shown that its value is approximately 10% less than its value at the ground.)

The first systematic experiments on the determination of the variation of the electron densities in the ionosphere were carried out in a 24-hour run on January 11-12, 1931, the E Layer being selected for study. It was then found that the E Layer maximum electron density starts to increase round about sunrise, reaches a maximum at noon and then wanes as the sun sets. Through the night, the ionization sinks to a low value, though there are often observed nocturnal sporadic increases of ionization which may possibly be due to meteoric dust. Later, the same critical frequency method was applied in the study of the F Layer. In this way there was inaugurated the long-term

study of the ionization in the various layers which has continued to the present time, when over 50 stations, using the critical-frequency method, are operating in different parts of the world.

Such continuous measurements of ionospheric densities, started in January 1931, in England, immediately showed a variation of noon ionization in sympathy with sunspot activity, which, in turn, indicated that the ultraviolet light from the sun, which is responsible for the electron production, varied substantially through the sunspot cycle. It was found, for example, that the E Layer ionization density was 50% greater in years of sunspot maximum than in years of sunspot minimum, indicating that the solar ultraviolet varied by as much as 125%, between the same two epochs. No such variation is to be noted in the heat and light we receive at ground level from the sun throughout the sunspot cycle.

Further work has confirmed the existence of a more weakly ionized region below the E Layer and which, in 1927, I had termed the « D Layer ». The D Layer acts chiefly as an absorbing stratum for high-frequency radio waves, though the reflection of extremely low-frequency waves has been detected from it. It has also been found that the F Layer, especially under summer daytime conditions, tends to bifurcate into two overlapping strata, known as the F_1 and F_2 Layers. Experimentally we can therefore determine the critical penetration frequencies for the E, F_1 and F_2 Layers and so study the variation of their maximum ionization densities from hour to hour, from season to season, and from year to year. It has not yet been found experimentally possible to study the variation of ionization density of the D Layer in a similar way, but, with the assistance of W. R. Piggott, I have been able to show that measurements of ionospheric absorption indicate that the ionization in the D Layer varies in sympathy with sunspot number, in a manner somewhat similar to that found in the E and F_1 Layers.

The F_2, Layer, on the other hand, has been found to exhibit some remarkable anomalies, the full nature of which is only now becoming clear as a result of the world-wide study of the ionosphere conducted by the network of observing stations already mentioned. For example, in 1934-1935, working with R. Naismith I found that the ionization density in the F_2 Layer was actually less at summer noon than at winter noon, a result entirely at variance with the result for the D, E, and F_1 Layers, in which the ionization, as one would expect from theory, is greater in summer than in winter. It appears, in fact, that the F_2 Layer ionization is subject to factors additional to the normal solar control by variation of solar zenith distance. In this connec-

tion I have found, from a study of the results from many stations, that the F_2 Layer noon ionization density is, to a certain extent, controlled by magnetic latitude. A complete theoretical explanation of this phenomenon is still lacking.

Although, therefore, work on the systematic study of the ionosphere has now been in progress for over a quarter of a century it will be seen that the subject still presents us with unsolved problems, largely because of the different results obtained in different latitudes and longitudes. Ionospheric phenomena in aurora1 latitudes, for example, were the subject of study by an expedition from Great Britain to Tromsö in 1932-1933. There it was found that, under weak aurora1 and magnetic storm conditions, abnormalionization was detected at night at E Layer levels, while, under strong aurora1 and magnetic storm conditions, extremely high absorption of the radio waves was experienced. No comparable studies of ionospheric conditions have yet been made in the Antarctic regions.

The work I have described was carried out with the object of exploring the electrical conditions in the higher atmosphere. It has yielded results which have shed light on many allied branches of science. In particular, the results disclosed have shown the essential correctness of the theory of Balfour Stewart concerning the origin of the rhythmic variations of terrestrial magnetism. For not only has it been shown that the upper atmosphere is electrical-conducting but it has been shown that the variation of that conductivity through the period of the sunspot cycle is of exactly the magnitude required to account for the variation of geo-magnetic changes with sunspot number. Moreover, from a careful study of layer heights the existence of large tidal movements, another essential feature of the Balfour Stewart theory, has also been demonstrated.

On the practical side of applications, ionospheric research has provided the basic ideas underlying the development of practical radiolocation of solid objects, for both the pulse-modulation and the frequency-modulation methods of measuring the distance of a reflecting surface by radio means have been used in the techniques of radar. Also, since we now have a fair understanding of the way in which the ionization in the reflecting layers varies through the day, through the season, and through the sunspot cycle, it is possible to forecast what I may call the « ionospheric weather » some time ahead. There has thus developed, on the practical side, the subject of « ionospheric forecasting» by which it is possible to forecast, say, three months ahead, the most suitable wavelengths for use at any time of the day, over

any distance of transmission, at any part of the world. In this way, scientific work conducted in the first instance with the object of exploring the wonders of the world around us, is now indicating how nation can speak unto nation with greater clarity and certainty.