The Bell System Technical Journal

Vol. XXVI *April, 1947* *No. 2*

Radar Antennas

By H. T. FRIIS and W. D. LEWIS

Table of Contents

Introduction

RADAR proved to be one of the most important technical achievements of World War II. It has many sources, some as far back as the nineteenth century, yet its rapid wartime growth was the result of military necessity. This development will continue, for radar has increasing applications in a peacetime world.

In this paper we will discuss an indispensable part of radar—the antenna. In a radar system the antenna function is two-fold. It both projects into space each transmitted radar pulse, and collects from space each received reflected signal. Usually but not always a single antenna performs both functions.

The effectiveness of a radar is influenced decisively by the nature and quality of its antenna. The greatest range at which the radar can detect a target, the accuracy with which the direction to the target can be determined and the degree with which the target can be discriminated from its background or other targets all depend to a large extent on electrical properties of the antenna. The angular sector which the antenna can mechanically or electrically scan is the sector from which the radar can provide information. The scanning rate determines the frequency with which a tactical or navigational situation can be examined.

Radar antennas are as numerous in kind as radars. The unique character and particular functions of a radar are often most clearly evident in the design of its antenna. Antennas must be designed for viewing planes from the ground, the ground from planes and planes from other planes. They must see ships from the shore, from the air, from other ships, and from submarines. In modern warfare any tactical situation may require one or several radars and each radar must have one or more antennas.

Radar waves are almost exclusively in the centimeter or microwave region, yet even the basic microwave techniques are relatively new to the radio art. Radar demanded antenna gains and directivities far greater than those previously employed. Special military situations required antennas with beam shapes and scanning characteristics never imagined by communication engineers.

It is natural that war should have turned our efforts so strongly in the direction of radar. But that these efforts were so richly and quickly rewarded was due in large part to the firm technical foundations that had been laid in the period immediately preceeding the war. When, for the common good, all privately held technical information was poured into one pool, all ingredients of radar, and of radar antennas in particular, were found to be present.

A significant contribution of the Bell System to this fund of technical knowledge was its familiarity with microwave techniques. Though Hertz himself had performed radio experiments in the present microwave region, continuous wave techniques remained for decades at longer wavelengths. However, because of its interest in new communication channels and broader bands the Bell System has throughout the past thirty years vigorously pushed continuous wave techniques toward the direction of shorter waves. By the middle nineteen-thirties members of the Radio Research Department of the Bell Laboratories were working within the centimeter region.

Several aspects of this research and development appear now as particularly important. In the first place it is obvious that knowledge of how to generate and transmit microwaves is an essential factor in radar. Many lower frequency oscillator and transmission line techniques are inapplicable in the microwave region. The Bell Laboratories has been constantly concerned with the development of generators which would work at higher and higher frequencies. Its broad familiarity with coaxial cable problems and in particular its pioneering work with waveguides provided the answers to many radar antenna problems.

Another telling factor was the emphasis placed upon measurement. Only through measurements can the planners and designers of equipment hope to evaluate performance, to chose between alternatives or to see the directions of improvement. Measuring techniques employing double detection receivers and intermediate frequency amplifiers had long been in use at the Holmdel Radio Laboratory. By employing these techniques radar engineers were able to make more sensitive and accurate measurements than would have been possible with single detection.

Antennas are as old as radio. Radar antennas though different in form are identical in principle with those used by Hertz and Marconi. Consequently experience with communication antennas provided a valuable background for radar antenna design. As an example of the importance of this background it can be recalled that a series of experiments with short wave antennas for Transatlantic radio telephone service had culminated in 1936 in a scanning array of rhombic antennas. The essential principles of this array were later applied to shipborne fire control antenna which was remarkable and valuable because of the early date at which it incorporated modern rapid scanning features.

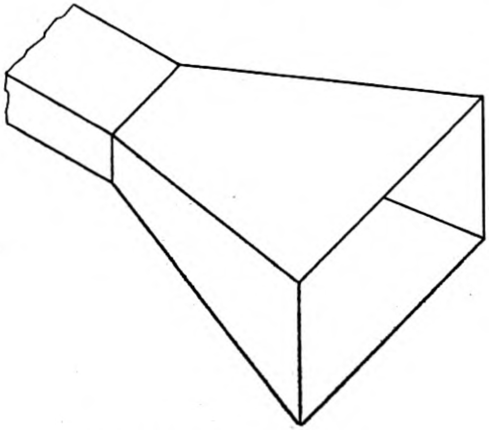


Fig. 1—An Electromagnetic Horn.

In addition to the antenna arts which arose directly out of communication problems at lower frequencies some research specifically on microwave antennas was under way before the war. Early workers in waveguides noticed that an open ended waveguide will radiate directly into space. It is not surprising therefore that these workers developed the electromagnetic horn, which is essentially a waveguide tapered out to an aperture (Fig. 1).

One of the first used and simplest radio antennas is the dipole (Fig.

2). Current oscillating in the dipole generates electromagnetic waves which travel out with the velocity of light. A single dipole is fairly non-directive and consequently produces a relatively weak field at a distance. When the wave-length is short the field of a dipole in a chosen direction can be increased many times by introducing a reflector which directs or 'focusses' the energy.

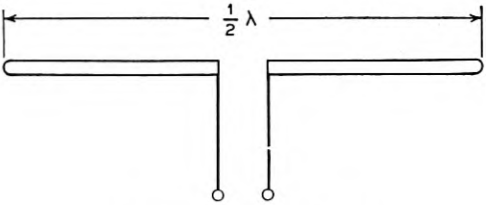


Fig. 2—A Microwave Dipole.

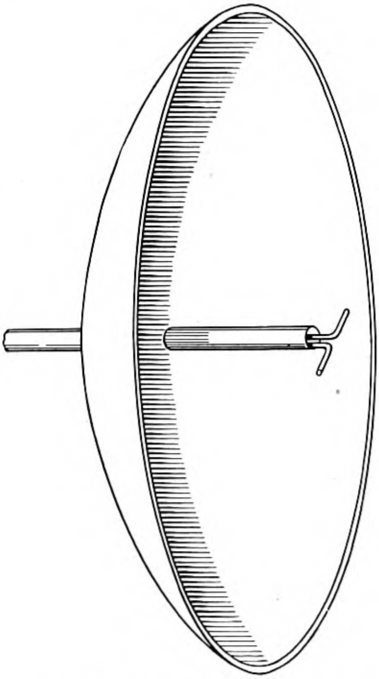


Fig. 3—A Dipole Fed Paraboloid.

In communication antennas the focussing reflector is most commonly a reflecting wire array. Even at an early date in radar the wavelength was so short that 'optical' reflectors could be used. These were sometimes paraboloids similar to those used in searchlights (Fig. 3). Sometimes they were parabolic cylinders as in the Mark III, an early shipborne fire control radar developed at the Whippany Radio Laboratory.

From these relatively simple roots, the communication antenna, the electromagnetic horn and the optical reflector, radar antennas were developed tremendously during the war. That this development in the Bell Laboratories was so well able to meet demands placed on it was due in large part to the solid foundation of experience possessed by the Research and Development groups of the Laboratories. Free interchange of individuals and information between the Laboratories and other groups, both in the United States and Great Britain, also contributed greatly to the success of radar antenna development.

Because of its accelerated wartime expansion the present radar antenna field is immense. It is still growing. It would be impossible for any single individual or group to master all details of this field, yet its broad outline can be grasped without difficulty.

The purpose of this paper is two-fold, both to provide a general discussion of radar antennas and to summarize the results of radar antenna research and development at the Bell Laboratories. Part I is a discussion of the basic electrical principles which concern radar antennas. In Part II we will outline the most common methods of radar antenna construction. Practical military antennas developed by the Bell Laboratories will be described in Part III.

The reader who is interested in general familiarity with the over all result rather than with technical features of design may proceed directly from this part to Part III.

PART I

ELECTRICAL PRINCIPLES

1. General

Radar antenna design depends basically on the same broad principles which underlie any other engineering design. The radar antenna designer can afford to neglect no aspect of his problem which has a bearing on the final product. Mechanical, chemical, and manufacturing considerations are among those which must be taken into account.

It is the electrical character of the antenna, however, which is connected most directly with the radar performance. In addition it is through attention to the electrical design problems that the greatest number of novel antennas have been introduced and it is from the electrical viewpoint that the new techniques can best be understood.

An antenna is an electromagnetic device and as such can be understood through the application of electromagnetic theory. Maxwell's equations provide a general and accurate foundation for antenna theory. They are the governing authority to which the antenna designer may refer directly when problems of a fundamental or baffling nature must be solved.

It is usually impracticable to obtain theoretically exact and simple solutions to useful antenna problems by applying Maxwell's Equations directly. We can, however, use them to derive simpler useful theories. These theories provide us with powerful analytical tools.

Lumped circuit theory is a tool of this sort which is of immense practical importance to electrical and radio engineers. As the frequency becomes higher the approximations on which lumped circuit theory is based become inaccurate and engineers find that they must consider distributed inductances and capacitances. The realm of transmission line theory has been invaded.

Transmission line theory is of the utmost importance in radar antenna design. In the first place the microwave energy must be brought to the antenna terminals over a transmission line. This feed line is usually a coaxial or a wave-guide. It must not break down under the voltage which accompanies a transmitted pulse. It must be as nearly lossless and reflectionless as possible and it must be matched properly to the antenna terminals.

The importance of a good understanding of transmission line theory does not end at the antenna terminals. In any antenna the energy to be transmitted must be distributed in the antenna structure in such a way that the desired radiation characteristics will be obtained. This may be done with transmission lines, in which case the importance of transmission line theory is obvious. It may be done by 'optical' methods. If so, certain transmission line concepts and methods will still be useful.

While it is true that transmission line theory is important it is not necessary to give a treatment of it in this paper. Adequate theoretical discussions can be found elsewhere in several sources.[1] It is enough at this point to indicate the need for a practical understanding of transmission line principles, a need which will be particularly evident in Part II, Methods of Antenna Construction.

We may, if we like, think of the whole radar transmission problem in terms of transmission line theory. The antenna then appears as a transformer between the feed line and transmission modes in free space. We cannot, however, apply this picture to details with much effectiveness unless we have some understanding of radiation.

In the sections to follow we shall deal with some theoretical aspects of radiation. We shall begin with a discussion of fundamental *transmission*

[1] See, for example, S. A. Schelkunoff, *Electromagnetic Waves*, D. Van Nostrand Co., Inc., 1943, in particular, Chapters VII and VIII, or F. E. Terman, *Radio Engineer's Handbook*, McGraw-Hill Book Co., Inc., 1943, Section 3.

principles. This discussion is applicable to all antennas regardless of how they are made or used. When applied to radar antennas it deals chiefly with those properties of the antenna which affect the radar range.

Almost all microwave radar antennas are large when measured in wavelengths. When used as transmitting antennas they produce desired radiation characteristics by distributing the transmitted energy over an area or 'wave front'. The relationships between the phase and amplitude of electrical intensity in this wave front and the radiation characteristics of the antenna are predicted by *wave front analysis*. Wave front analysis is essentially the optical theory of diffraction. Although approximate it applies excellently to the majority of radar antenna radiation problems. We shall discuss wave front analysis in Section 3.

2. Transmission Principles

2.1 *Gain and Effective Area of an Antenna*

An extremely important property of any radar antenna is its ability to project a signal to a distant target. The *gain* of the antenna is a number which provides a quantitative measure of this ability. Another important property of a radar antenna is its ability to collect reflected power which is returning from a distant target. The *effective area* of the antenna is a quantitative measure of this ability. In this section these two quantities will be defined, and a simple relation between them will be derived. Their importance to radar range will be established.

Definition of Gain. When power is fed into the terminals of an antenna some of it will be lost in heat and some will be radiated. The *gain G* of the antenna can be defined as the ratio

$$G = P/P_O \tag{1}$$

where P is the power flow per unit area in the plane linearly polarized electromagnetic wave which the antenna causes in a distant region usually in the direction of maximum radiation and P_O is the power flow per unit area which would have been produced if all the power fed into the terminals had been radiated equally in all directions in space.

Definition of Effective Area. When a plane linearly polarized electromagnetic wave is incident on the receiving antenna, received power P_R will be available at the terminals of the antenna. The *effective area* of the antenna is defined, by the equation

$$A = P_R/P' \tag{2}$$

where P' is the power per unit area in the incident wave. In other words the received power is equal to the power flow through an area that is equal to the effective area of the antenna.

2.2 *Relationship between Gain and Effective Area*

Figure 4 shows a radio circuit in free space made up of a transmitting antenna T and a receiving antenna R. If the transmitted power P_T had been radiated equally in all directions, the power flow per unit area at the receiving antenna would be

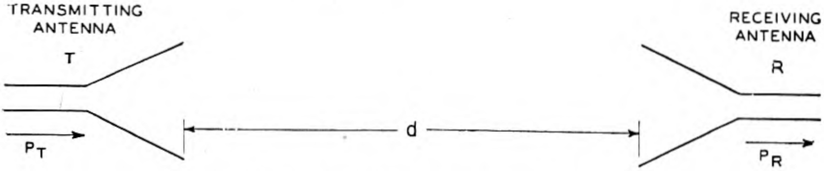


Fig. 4—Radio Circuit in Free Space.

$$P_0 = P_T \frac{1}{4\pi d^2} \tag{3}$$

Definition (1) gives, therefore, for the power flow per unit area at the receiving antenna

$$P = P_0 G_T = \frac{P_T G_T}{4\pi d^2} \tag{4}$$

and definition (2) gives for the received power

$$P_R = P A_R = \frac{P_T G_T A_R}{4\pi d^2} \tag{5}$$

From the law of reciprocity it follows that the same power is transferred if the transmitting and receiving roles are reversed. By (5) it is thus evident that

$$G_T A_R = G_R A_T$$

or

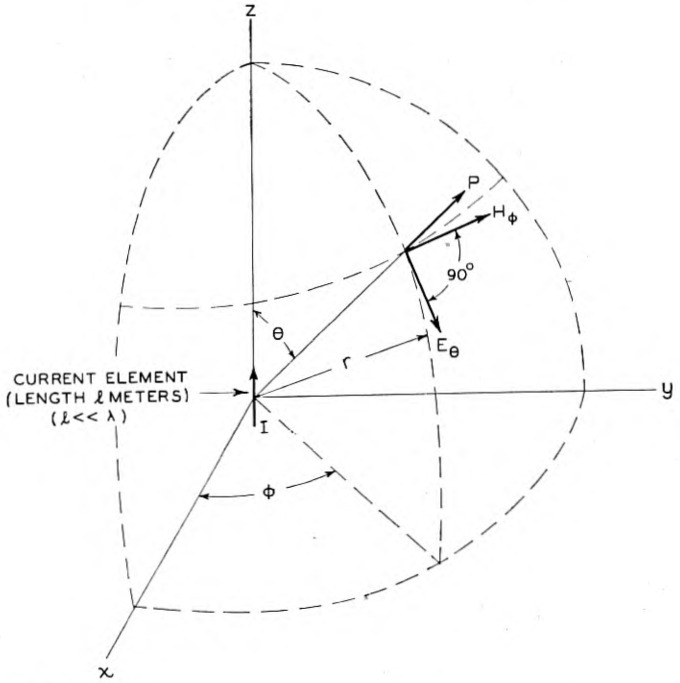
$$G_T/A_T = G_R/A_R \tag{6}$$

Equation (6) shows that the ratio of the gain and effective area has the same constant value for all antennas at a given frequency. It is necessary, therefore, to calculate this ratio only for a simple and well known antenna such as a small dipole or uniform current element.

2.3 *The Ratio G/A for a Small Current Element*

In Fig. 5 are given formulas[2] in M.K.S. units for the free space radiation from a small current element with no heat loss. We have assumed that

[2] See S. A. Schelkunoff, *Electromagnetic Waves*, D. Van Nostrand Co., Inc., 1943, p. 133



FOR $r \gg \lambda$:

$$\text{MAGNETIC INTENSITY} = H_\phi = i\frac{I\ell}{2\lambda r}e^{-i\frac{2\pi}{\lambda}r}\,\text{SIN}\,\theta \quad \frac{\text{AMPERES}}{\text{METER}} \qquad (1)$$

$$\text{ELECTRIC INTENSITY} = E_\theta = 120\pi H_\phi \quad \frac{\text{VOLTS}}{\text{METER}} \qquad (2)$$

$$\text{POWER FLOW} = P = \left|H_\phi E_\theta\right| = 30\pi\left[\frac{I\ell}{\lambda r}\right]^2 \text{SIN}^2\theta \quad \frac{\text{WATTS}}{\text{METER}^2} \qquad (3)$$

$$\text{P IS MAXIMUM FOR } \theta = 90^\circ \; \textit{i.e.,}\; P_{\text{MAX}} = 30\pi\left[\frac{I\ell}{\lambda r}\right]^2 \frac{\text{WATTS}}{\text{METER}^2} \qquad (4)$$

POWER FLOW ACROSS SPHERE OF RADIUS r OR

$$\text{TOTAL RADIATION} = W = \int_0^\pi P\, 2\pi r\,\text{SIN}\,\theta\, r\,d\theta = 80\pi^2\left[\frac{I\ell}{\lambda}\right]^2 \text{WATTS} \qquad (5)$$

$$\text{RADIATION RESISTANCE} = R_{\text{RAD}} = \frac{W}{I^2} = 80\pi^2\left(\frac{\ell}{\lambda}\right)^2 \text{OHMS} \qquad (6)$$

$$\text{BY } (4) \text{ AND } (5):\; P_{\text{MAX}} = \frac{3}{8\pi r^2}\,W \quad \frac{\text{WATTS}}{\text{METER}^2} \qquad (7)$$

Fig. 5—Free Space Radiation from a Small Current Element with Uniform Current I Amperes over its Entire Length.

this element is centered at the origin of a rectangular coordinate system and that it lies along the Z axis. At a large distance r from the element the maximum power flow per unit area occurs in a direction normal to it and is given by

$$P_{\max} = \frac{3W}{8\pi r^2}\,\frac{\text{watts}}{\text{meter}^2} \tag{7}$$

where W is the total radiated power. If W had been radiated equally in all directions the power flow per unit area would be

$$P_0 = \frac{W}{4\pi r^2}\,\frac{\text{watts}}{\text{meters}^2} \tag{8}$$

It follows that the gain of the small current element is

$$G_{\text{diople}} = \frac{P_{\max}}{P_0} = 1.5 \tag{9}$$

The effective area of the dipole will now be calculated. When it is used to receive a plane linearly polarized electromagnetic wave, the available output power is equal to the induced voltage squared divided by four times the radiation resistance. Thus

$$P_R = \frac{E^2\ell^2}{4R_{\text{rad}}}\text{ Watts} \tag{10}$$

where E is the effective value of the electric field of the wave, ℓ is the length of the current element and R_{rad} is the radiation resistance of the current element. From Fig. 5 we see that $R_{\text{rad}} = \dfrac{80\pi^2\ell^2}{\lambda^2}$ ohms. Since the power flow per unit area is equal to the electric field squared divided by the impedance of free space, in other words $P_0 = \dfrac{E^2}{120\pi}$ we have

$$A_{\text{dipole}} = \frac{P_R}{P_O} = \frac{3\lambda^2}{8\pi}\text{ meter}^2 \tag{11}$$

We combine formulas (9) and (11) to find that

$$\frac{G_{\text{dipole}}}{A_{\text{dipole}}} = \frac{4\pi}{\lambda^2}$$

Since, as proved in 2.2 this ratio is the same for all antennas, it follows that for any antenna

$$\frac{G}{A} = \frac{4\pi}{\lambda^2} \tag{12}$$

2.4 *The General Transmission Formula*

Transmission loss between transmitter and receiver through the radio circuit shown in Fig. 4 was given by equation (5). By substituting the relation (12) into (5) we can obtain the simple free space transmission formula:

$$P_R = P_T \frac{A_T A_R}{\lambda^2 d^2} \text{ watts} \tag{13}$$

Although this formula applies to free space only it is believed to be as useful in radio engineering as Ohm's law is in circuit engineering.

2.5 *The Reradiation Formula*

One further relation, the radar reflection formula is of particular interest. Consider the situation illustrated in Fig. 6. Let P_T be the power radiated from an antenna with effective area A_T, A_S the area of a reflecting object at distance d from the antenna and P_R the power received by an antenna of effective area A_R. By equation (13) the power striking A_S is $\frac{P_T A_T A_S}{\lambda^2 d^2}$. If this power were reradiated equally in all directions the reflected power flow at the receiving antenna would be $\frac{P_T A_T A_S}{4\pi d^4 \lambda^2}$ but since the average reradiation is larger toward the receiving antenna, the power flow per unit area there is usually $K \frac{P_T A_T A_S}{4\pi d^4 \lambda^2}$ where $K > 1$. It follows from (2) that

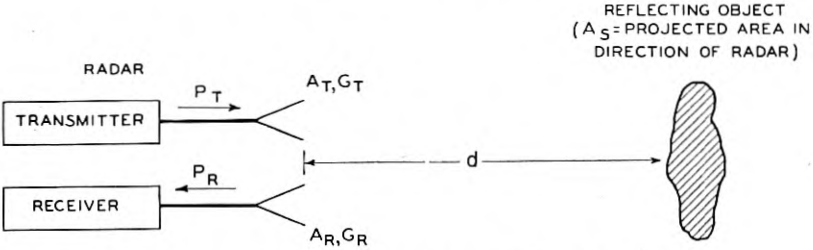


Fig. 6—Radar with Separate Receiving and Transmitting Antennas.

$$P_R = K \frac{P_T A_T A_R A_S}{4\pi \lambda^2 d^4} \tag{14}$$

Formula (14) shows clearly why the use of large and efficient antennas will greatly increase the radar range.

Formula (14) applies to free space only. Application to other conditions may require corrections for the effect of the "ground", and for the effect of the transmission medium, which are beyond the scope of this paper.

2.6 *The Plane, Linearly Polarized Electromagnetic Wave*

In the foregoing sections we have referred several times to 'plane, linearly polarized electromagnetic waves'. These waves occur so commonly in antenna theory and practice that it is worth while to discuss them further here.

Some properties of linearly polarized, plane electromagnetic waves are illustrated in Fig. 7. At any point in the wave there is an electric field and a magnetic field. These fields are vectorial in nature and are at right angles to each other and to the direction of propagation. It is customary to give the magnitude of the electric field only.

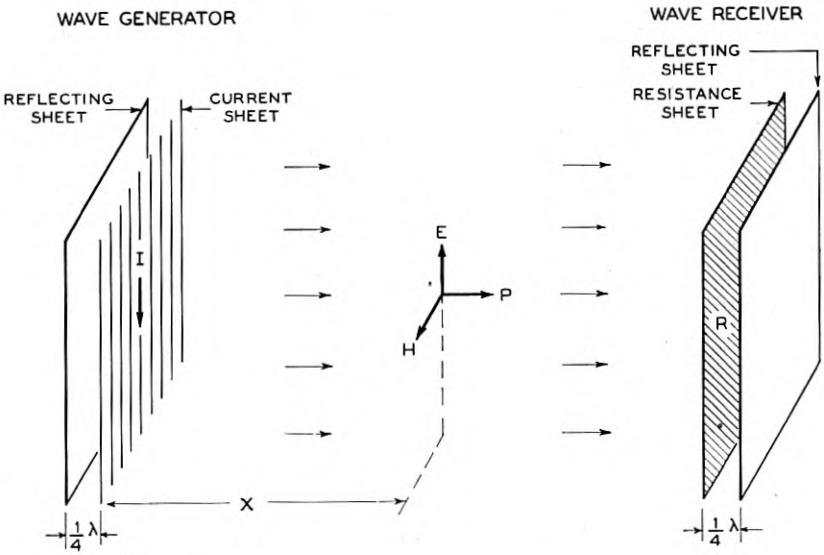
If we use the M.K.S. system of units the magnitudes of the fields are expressed in familiar units. Electric intensity appears as volts per meter and magnetic intensity as amperes per meter. The ratio of electric to magnetic intensity has a value of 120π or about 377 ohms. This is the 'impedance' of free space. The power flow per unit area is expressed in watts per square meter. We see, therefore, that the electromagnetic wave is a means for carrying energy not entirely unlike a familiar two wire line or a coaxial cable.

Electromagnetic waves are generated when oscillating currents flow in conductors. We could generate a plane linearly polarized electromagnetic wave with a uniphase current sheet consisting of a network of fine wires backed up with a conducting reflector as shown in Fig. 7. This wave could be absorbed by a plane resistance sheet with a resistivity of 377 ohms, also backed up by a conducting sheet. The perfectly conducting reflecting sheets put infinite impedances in parallel with the current sheet and the resistance sheet, since each of these reflecting sheets has a zero impedance at a spacing of a quarter wavelength.

A perfectly plane electromagnetic wave can exist only under certain ideal conditions. It must be either infinite in extent or bounded appropriately by perfect electric and magnetic conductors. Nevertheless thinking in terms of plane electromagnetic waves is common and extremely useful. In the first place the waves produced over a small region at a great distance from any radiator are essentially plane. Arguments concerning receiving antennas therefore generally assume that the incident waves are plane. In the second place an antenna which has dimensions of many wavelengths can be analyzed with considerable profit on the basis of the assumption that it transmits by producing a nearly plane electromagnetic wave across its aperture. This method of analysis can be applied to the majority of microwave radar antennas, and will be discussed in the following sections.

3. Wave Front Analysis

The fundamental design question is "How to get what we want?" In a radar antenna we want specified radiation characteristics; gain, pattern and polarization. Electromagnetic theory tells us that if all electric and magnetic currents in an antenna are known its radiation characteristics may be derived with the help of Maxwell's Equations. However, the essence of electromagnetic theory insofar as it is of use to the radar antenna designer can usually be expressed in a simpler, more easily visualized and thus more useful form. This simpler method we call *wave front analysis.*



MAGNETIC INTENSITY = $H = Ie^{-i\frac{2\pi}{\lambda}x}$ $\frac{\text{AMPERES}}{\text{METER}}$

ELECTRIC INTENSITY = $E = 120\pi H$ $\frac{\text{VOLTS}}{\text{METER}}$

POWER FLOW = $P = EH$ $\frac{\text{WATTS}}{\text{METER}^2}$

CURRENT DENSITY = I $\frac{\text{AMPERES}}{\text{METER}}$

RESISTIVITY = $R = 120\pi$ OHMS

Fig. 7—Linearly Polarized Plane Electromagnetic Waves.

In a transmitting microwave antenna the power to be radiated is used to produce currents in antenna elements which are distributed in space. This distribution is usually over an area, it may be discrete as with a dipole array or it may be continuous as in an electromagnetic horn or paraboloid. These currents generate an advancing electromagnetic wave over the aperture of the antenna. The amplitude, phase and polarization of the electric intensity in portions of the wave are determined by the currents in the antenna and thus by the details of the antenna structure. This advancing wave can be called the 'wave front' of the antenna.

When the wave front of an antenna is known its radiation characteristics may be calculated. Each portion of the wave front can be regarded as a secondary or 'Huygens' source of known electric intensity, phase and polarization. At any other point in space the electric intensity, phase and polarization due to a Huygens source can be obtained through a simple expression given in the next section. The radiation characteristics of the antenna can be found by adding or integrating the effects due to all Huygens sources of the wave front.

This procedure is based on the assumption that the antenna is transmitting. A basic law of reciprocity assures us that the receiving gain and radiation characteristics of the antenna will be identical with the transmitting ones when only linear elements are involved.

This resolution of an antenna wave front into an array of secondary sources can be justified within certain limitations on the basis of the induction theorem of electromagnetic theory.[3] These limitations are discussed in a qualitative way in section 3.13.

3.1 *The Huygens Source*

Consider an elementary Huygens source of electric intensity E_0polarized parallel to the X axis with area dS in the XY plane (Fig. 8). This can be thought of as an element of area dS of a wave front of a linearly polarized plane electromagnetic wave which is advancing in the positive z direction.[3] From Maxwell's Equations we can determine the field at any point of space due to this Huygens Source. The components of electric field, are found to be

$$\begin{aligned} E_r &= 0 \\ E_\theta &= i\,\frac{E_0\,dS}{2\lambda r}\,e^{-i(2\pi/\lambda)r}\,(1+\cos\theta)\cos\phi \\ E_\phi &= -i\,\frac{E_0\,dS}{2\lambda r}\,e^{-i(2\pi/\lambda)r}\,(1+\cos\theta)\sin\phi \end{aligned} \tag{15}$$

where λ is the wavelength.

We see at once that this represents a vector whose absolute magnitude at all points of space is given by

$$|E| = \frac{E_0\,dS}{2\lambda r}\,(1+\cos\theta). \tag{16}$$

[3] S. A. Schelkunoff, Loc. Cit., Chap. 9.

Here $\frac{E_0\,dS}{\lambda}$ is an amplitude factor which depends on the wavelength, intensity and area of the elementary source and $1/r$ is an amplitude factor which specifies the variation of field with distance. $(1 + \cos\theta)$ is an amplitude factor which shows that the directional pattern of the elementary source is a cardioid with maximum radiation in the direction of propagation and no radiation in the reverse direction.

When we use the properties of the Huygens source in analyzing a microwave antenna we are usually concerned principally with radiation in or near the direction of propagation. For such radiation Equation 16 takes a particularly simple form in Cartesian Coordinates

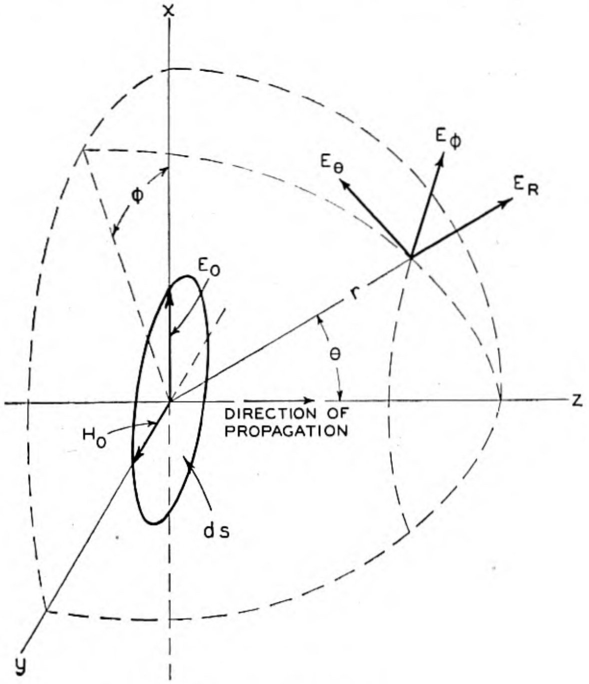


Fig. 8—The Huygens Source.

$$E_x \cong i\,\frac{E_0\,dS}{\lambda r}\,e^{-i(2\pi/\lambda)r};\ E_y \cong 0;\ E_z \cong 0. \tag{17}$$

This represents an electric vector nearly parallel to the electric vector of the source. The amplitude is given by the factor $\frac{E_0\,dS}{\lambda r}$ and the phase by the factor $i\ e^{-i(2\pi/\lambda)r}$. With this equation as a basis we will now proceed to study some relevant matters concerning radar antennas.

3.2 *Gain and Effective Area of an Ideal Antenna*

On the basis of (17) we can now determine the gain of an ideal antenna of area S ($S \gg \lambda^2$). This antenna is assumed to be free of heat loss and to transmit by generating an advancing wave which is uniform in phase and amplitude in the XY plane. Let the electric intensity in the wave front of the ideal antenna be E_0 polarized parallel to the X axis (Fig. 9). The transmitted power P_T is equal to the power flow through S and is given by

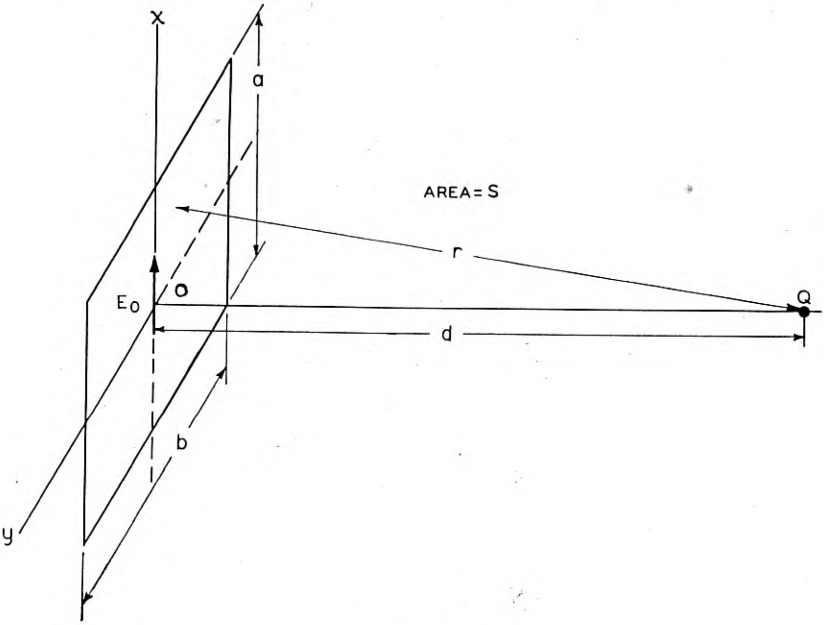


Fig. 9—An Ideal Antenna.

$$P_T = \frac{E_0^2}{120\pi} S. \tag{18}$$

At a point Q on the Z axis the electric intensity is obtained by adding the effects of all the Huygens sources in S. If the distance of Q from O is sc great that

$$r = d + \Delta$$

where Δ is a negligibly small fraction of a wavelength for every point on S then we see from (17) that the electric vector at Q is given by

$$E_x = \int_S i\,\frac{E_0\,dS}{\lambda r}\,e^{-i(2\pi/\lambda)r} = i\,e^{-i(2\pi/\lambda)d}\,\frac{E_0\,S}{\lambda d}\,;\ E_y = 0;\ E_z = 0. \tag{19}$$

The power flow per unit area at Q is therefore

$$P = \frac{1}{120\pi}\,\frac{E_0^2 S^2}{\lambda^2 d^2} = \frac{P_T S}{\lambda^2 d^2}$$

P_0 the power flow per unit area at Q when power is radiated isotropically from O is found by assuming that P_T is spread evenly over the surface of a sphere of radius d.

$$P_0 = \frac{P_T}{4\pi d^2} \tag{20}$$

The gain of a lossless, uniphase, uniamplitude, linearly polarized antenna is, by the definition of equation 1, the ratio of 19 and 20.

$$G_s = \frac{P_T S}{\lambda^2 d^2}\bigg/\frac{P_T}{4\pi d^2} = \frac{4\pi S}{\lambda^2}\,. \tag{21}$$

It follows from 12 that the effective area of the ideal antenna is

$$A = S \tag{22}$$

In other words in this ideal antenna the effective area is equal to the actual area. This is a result which might have been obtained by more direct arguments.

3.3 *Gain and Effective Area of an Antenna with Aperture in a Plane and with Arbitrary Phase and Amplitude*

Let us consider an antenna with a wave front in the XY plane which has a known phase and amplitude variation. Let the electric intensity in the wave front be

$$E(x, y) = E_0 a(x, y) e^{i\phi(x,y)} \tag{23}$$

polarized parallel to the x axis. The radiated power is equal to the power flow through S and is given by

$$P_{\text{rad}} = \frac{E_0^2 \int a^2(x, y)\,dS}{120\pi}\,. \tag{24}$$

The input power to the antenna is

$$P_T = P_{\text{rad}}/L \tag{25}$$

where L is a loss factor (<1). At a point Q on the Z axis the electric intensity is obtained by adding the effects of all the Huygens sources in S. If OQ is as great as in the above derivation for the gain of an ideal antenna then we see from 17 that the electric intensity at Q is

$$E_x = i\,\frac{e^{i(2\pi/\lambda)d}}{\lambda d}\,E_0 \int a(x, y)e^{i\phi(x,y)}\,dS;\; E_y = 0;\; E_z = 0. \tag{26}$$

The power flow per unit area at Q is given by

$$P = \frac{1}{120\pi}\,|\,E_x\,|^2 \tag{27}$$

and P_0 the power flow per unit area at Q when P_T is radiated isotropically is given by equation (3).

The power gain of the antenna, by definition 1 is therefore

$$G = \frac{P}{P_0} = \frac{|\,E_x\,|^2}{120\pi} \Bigg/ \frac{P_T}{4\pi d^2} = \frac{4\pi L}{\lambda^2}\,\frac{\left|\int_S a(x, y)e^{i\phi(x,y)}\,dS\right|^2}{\int_S a^2(x, y)\,dS}. \tag{28}$$

The gain expressed in *db* is given by

$$G_{db} = 10\log_{10} G \tag{29}$$

We combine 12 and 28 to obtain

$$A = L\,\frac{\left|\int_S a(x, y)e^{i\phi(x,y)}\,dS\right|^2}{\int_S a^2(x, y)\,dS} \tag{30}$$

a formula for the effective area of the antenna.

3.4 *The Significance of the Pattern of a Radar Antenna*

The accuracy with which a radar can determine the directions to a target depends upon the *beam widths* of the radar antenna. The ability of the radar to separate a target from its background or distinguish it from other targets depends upon the beam widths and the *minor lobes* of the radar antenna. The efficiency with which the radar uses the available power to view a given region of space depends on the *beam shape* of the antenna. These quantities characterize the antenna pattern. In the following sections means for the calculation of antenna patterns in terms of wave front theory will be developed, and some illustrations will be given.

3.5 *Pattern in Terms of Antenna Wave Front*

If the relative phase and amplitude in a wave front are given by

$$E(x, y) = a(x, y)e^{i\phi(x,y)} \tag{31}$$

the relative phase and amplitude at a distant point Q not necessarily on the Z axis (Fig. 10) in the important case where the angle QOZ between the direction of propagation and the direction to the point is small, is given from (17) by adding the contributions at Q due to all parts of the wave front. This gives

$$E_Q = \frac{i}{\lambda d} \int_S e^{-i(2\pi/\lambda)r} e^{i\phi(x,y)} a(x, y) \, dS. \tag{32}$$

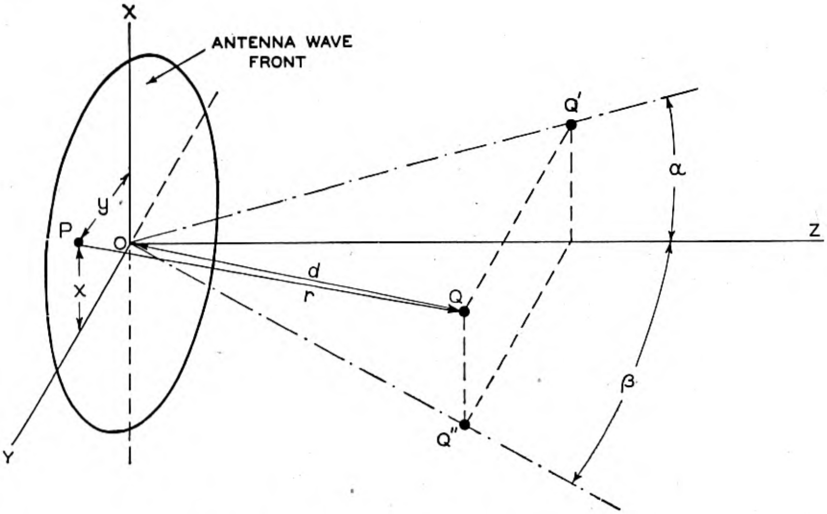


Fig. 10—Geometry of Pattern Analysis.

The quantity r in (32) is the distance from any point P with coordinates x, y, Q, in the XY, plane to the point Q (Fig. 10). Simple trigonometry shows that when OQ is very large

$$r = d - x \sin \alpha - y \sin \beta \tag{33}$$

where d is the distance OQ, α is the angle ZOQ' between OZ and OQ' the projection of OQ on the XZ plane and β is similarly the angle ZOQ''. The substitution of 33 into 32 gives

$$E_Q = \frac{i\,e^{-i(2\pi/\lambda)d}}{\lambda d} \int_S e^{i(2\pi/\lambda)(x\sin\alpha + y\sin\beta) + i\phi(x,y)} a(x, y) \, dS. \tag{34}$$

In most practical cases this equation can be simplified by the assumptions

$$\phi(x, y) = \phi'(x) + \phi''(y)$$

$$a(x, y) = a'(x)a''(y)$$

from which it follows that

$$| E_Q | = F(d)F(\alpha)F(\beta) \tag{35}$$

where $F(d)$ is an amplitude factor which does not depend on angle,

$$F(\alpha) = \int e^{i(2\pi/\lambda)x\sin\alpha + i\phi'(x)} a'(x)\, dx \tag{36}$$

is a directional factor which depends only on the angle α and not on the angle β or d, and $F(\beta)$ similarly depends on β but not on α or d. The pattern of an antenna can be calculated with the help of the simple integrals as in 36, and illustrations of such calculations will be given in the following sections.

3.6 *Pattern of an Ideal Rectangular Antenna*

Let the wave front be that of an ideal rectangular antenna of dimensions a, b; with linear polarization and uniform phase and amplitude. The dimensions a and b can be placed parallel to the X and Y axes respectively as sketched in Fig. 9. Equation 36 then gives

$$F(\alpha) = \int_{-a/2}^{a/2} e^{i(2\pi/\lambda)x\sin\alpha}\, dx = a\,\frac{\sin\psi}{\psi} \tag{37}$$

where $\psi = \dfrac{\pi a \sin\alpha}{\lambda}$.

Similarly

$$F(\beta) = b\,\frac{\sin\psi'}{\psi'} \tag{38}$$

where $\psi' = \dfrac{\pi b \sin\beta}{\lambda}$.

The pattern of the ideal rectangular aperture, in other words the distribution of electrical field in angle is thus given approximately by

$$F(\alpha)F(\beta) = ab\,\frac{\sin\psi}{\psi}\,\frac{\sin\psi'}{\psi'}. \tag{39}$$

The function $\dfrac{\sin\psi}{\psi}$ is plotted in Fig. 11. It is perhaps the most useful function of antenna theory, not because ideal antennas as defined above are particularly desirable in practice but because they provide a simple standard with which more useful but more complex antennas can profitably be compared.

3.7 *Effect on Pattern of Amplitude Taper*

The $\frac{\sin\psi}{\psi}$ pattern which results from an ideal wave front has undesirably high minor lobes for most radar applications. These minor lobes will be reduced if the wave front of constant amplitude is replaced by one which retains a constant phase but has a rounded or 'tapered' amplitude distribution.

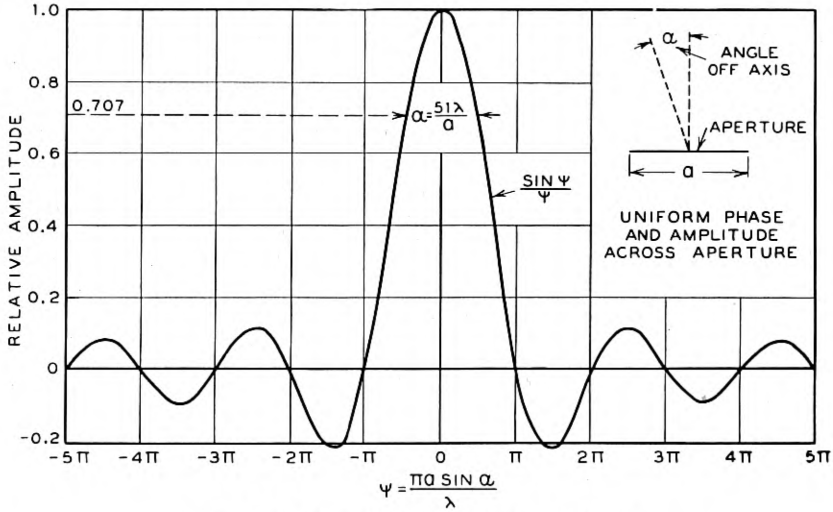


Fig. 11—Pattern of Ideal Rectangular Antenna.

If such an amplitude taper is represented analytically by the function

$$a'(x) = C_1 + C_2 \cos\frac{\pi x}{a} \tag{40}$$

then equation (36) is readily integrable. To integrate it we utilize the identity

$$\cos\frac{\pi x}{a} = \frac{e^{+i\pi x/a} + e^{-i\pi x/a}}{2}$$

upon which the integral becomes the sum of three simple integrals of the form

$$\int_{-a/2}^{a/2} e^{ikx}dx = a\left[\frac{\sin\frac{ka}{2}}{\frac{ka}{2}}\right] \tag{41}$$

We therefore obtain

$$F(\alpha) = aC_1 \frac{\sin \psi}{\psi} + a \frac{C_2}{2} \left[\frac{\sin\left(\psi + \frac{\pi}{2}\right)}{\left(\psi + \frac{\pi}{2}\right)} + \frac{\sin\left(\psi - \frac{\pi}{2}\right)}{\left(\psi - \frac{\pi}{2}\right)} \right] \tag{42}$$

The patterns resulting from two possible tapers are given by substituting $C_1 = 0, C_2 = 1$ and $C_1 = 1/3, C_2 = 2/3$ in (42). These patterns are evidently calculable in terms of the known function $\frac{\sin \alpha}{\alpha}$. They are plotted in Figs. 12 and 13.

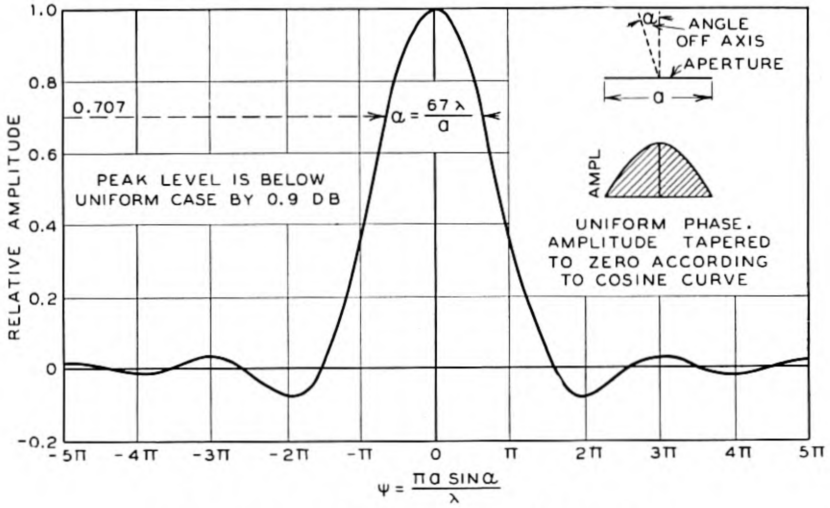


Fig. 12—Pattern of Tapered Rectangular Antenna.

It will be observed that minor lobe suppression through tapering is obtained at the expense of beam broadening. In addition to this the gain is reduced by tapering, as could have been calculated from 28. These undesirable effects must be contended with in any practical antenna design. The choice of taper must be made on the basis of the most desirable compromise between the conflicting factors.

3.8 *Effect on Pattern of Linear Phase Variation*

If we assume a constant amplitude and a linear phase variation

$$\phi'(x) = -k_1 x$$

over an aperture $-a/2 < x < a/2$ then 36 becomes a simple integral of the form (41) and we obtain

$$F(\alpha) = a\,\frac{\sin\psi''}{\psi''} \quad \text{where} \quad \psi'' = \frac{\pi a}{\lambda}\sin\alpha - \frac{k_1 a}{2} \tag{43}$$

The physical interpretation of (43) is simply that the pattern is identical to the pattern of an antenna with constant amplitude and uniform phase but rotated through an angle θ where

$$\sin\theta = \frac{k_1\lambda}{2\pi}$$

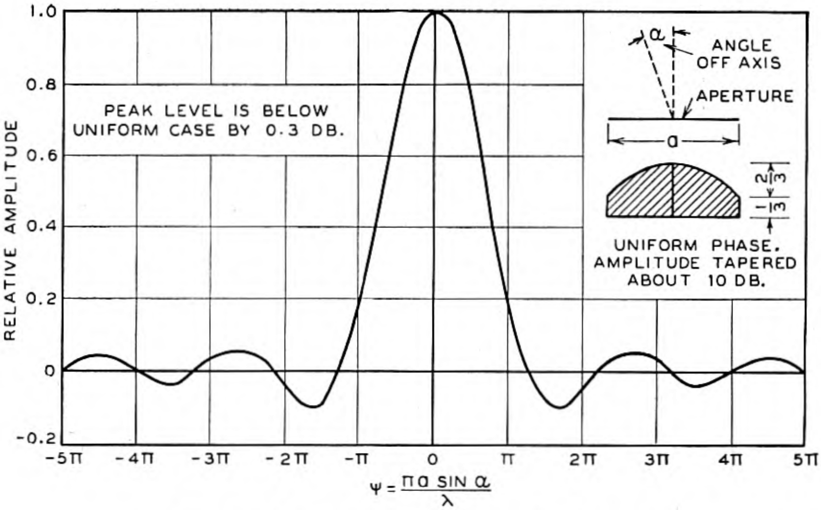


Fig. 13—Pattern of Tapered Rectangular Antenna.

Simple examination shows that the new direction of the radiation maximum is at right angles to a uniphase surface, as we would intuitively expect. This phenomenon has particular relevance to the design of scanning antennas.

3.9 *Effect on Pattern of Square Law Phase Variation*

If we assume a constant amplitude and a square law phase variation

$$\phi'(x) = -k_2^2x^2$$

over the aperture $a/2 < x < a/2$ then the substitution

$$x = \frac{1}{k_2}\left[X + \frac{\frac{2\pi}{\lambda}\sin\alpha}{2k_2}\right] \tag{44}$$

reduces (36) to the form

$$F(\alpha) = \frac{1}{k_2} e^{i\left(\frac{2\pi}{\lambda}\right)^2 \sin^2 \alpha / 4k_2^2} \int e^{-iX^2}\, dX \tag{45}$$

Equation (45) can be evaluated with the help of Fresnel's Integrals

$$\int \cos X^2\, dX, \qquad \int \sin X^2\, dX$$

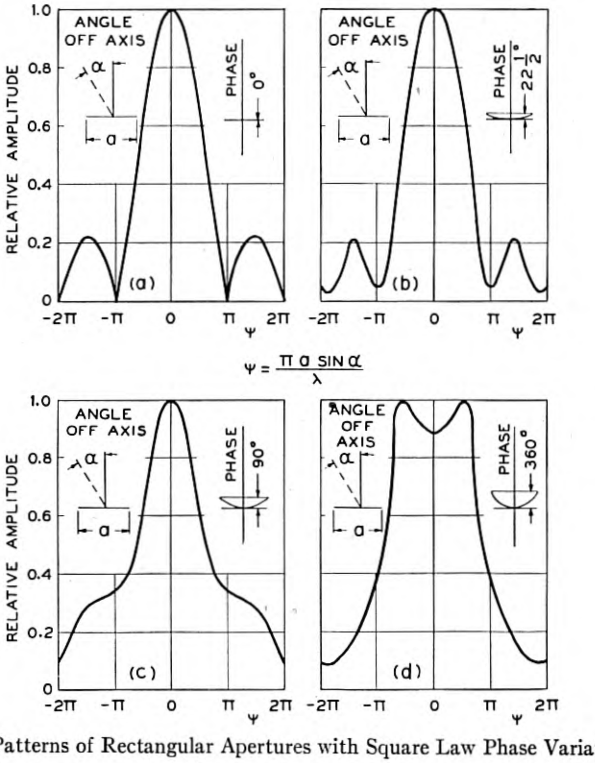


Fig. 14—Patterns of Rectangular Apertures with Square Law Phase Variation.

which are tabulated[4], or from Cornu's Spiral which is a convenient graphical representation of the Fresnel Integrals.

Typical computed patterns for apertures with square law phase variations are plotted in Fig. 14. These theoretical curves can be applied to the following important practical problems.

(1) The pattern of an electromagnetic horn.

[4] For numerical values of Fresnel's Integrals and a plot of Cornu's Spiral see Jahnke and Emde, *Tables of Functions* B. G. Teubner, Leipzig, 1933, or Dover Publications, New York City, 1943.

(2) The defocussing of a reflector or lens due to improper placing of the primary feed.

(3) The defocussing of a zoned reflector or lens due to operation at a frequency off mid-band.

In addition to providing distant patterns of apertures with curved wave fronts (44) provides theoretical 'close in' patterns of antennas with plane wave fronts. This arises from the simple fact that a plane aperture appears as a curved aperture to close in points. The degree of curvature depends on the distance and can be evaluated by extremely simple geometrical considerations. When this has been done we find that Fig. 14 represents the so-called Fresnel diffraction field.

With this interpretation of square law variation of the aperture we can examine several additional useful problems. We can for instance justify the commonly used relation

$$D = \frac{2a^2}{\lambda}$$

for the minimum permissible distance of the field source from an experimental antenna test site. This distance produces an effective phase curvature of $\lambda/16$. We can examine optical antenna systems employing large primary feeds, in particular those employing parabolic cylinders illuminated by line sources.

3.10 *Effect on Pattern of Cubic Phase Variation*

If we assume a constant amplitude and a cubic phase variation $\phi'(x) = -k_3^3x^3$ over the aperture from $-a/2 < x < a/2$ then equation (36) becomes

$$F(\alpha) = \int_{-a/2}^{a/2} e^{-ik_3^3x^3} \cdot e^{i(2\pi/\lambda)x \sin \alpha} \cdot dx \tag{46}$$

If $k_3^3x^3 \leq \frac{\pi}{2}$ then it is a fairly good approximation to write

$$c^{-ik_3^3x^3} = 1 - ik_3^3x^3 - \frac{k_3^6x^6}{2} + \cdots \tag{47}$$

from which it follows that (46) can be integrated since it reduces to a sum of three terms each of which can be integrated.

Typical computed patterns for apertures with cubic phase variation are plotted in Figs. 15 and 16. Cubic phase distortions are found in practice when reflectors or lenses are illuminated by primary feeds which are off axis either because of inaccurate alignment or because beam lobing or scanning through feed motion is desired. The beam distortion due to cubic phase variation is known in optics as 'coma' and the increased unsymmetrical lobe which is particularly evident in Fig. 16 is commonly called a 'coma lobe'.

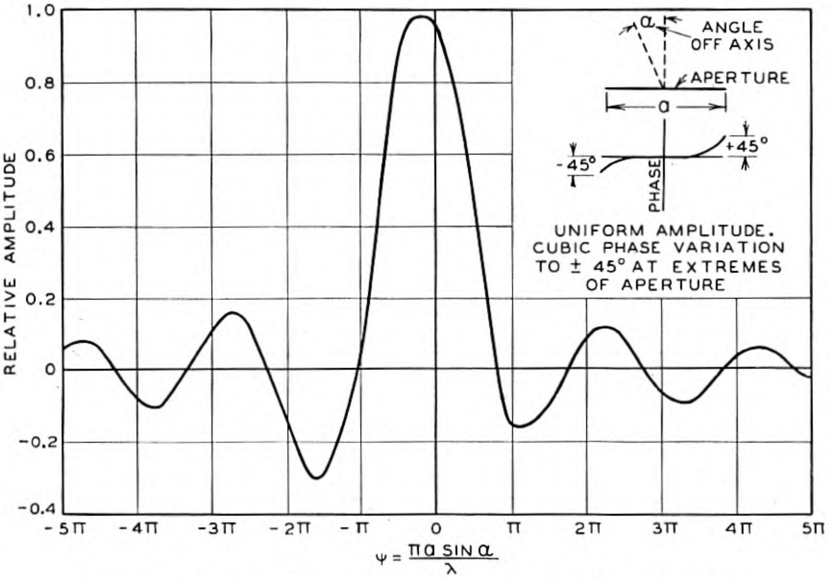


Fig. 15—Pattern of Rectangular Antenna with Cubic Phase Variation.

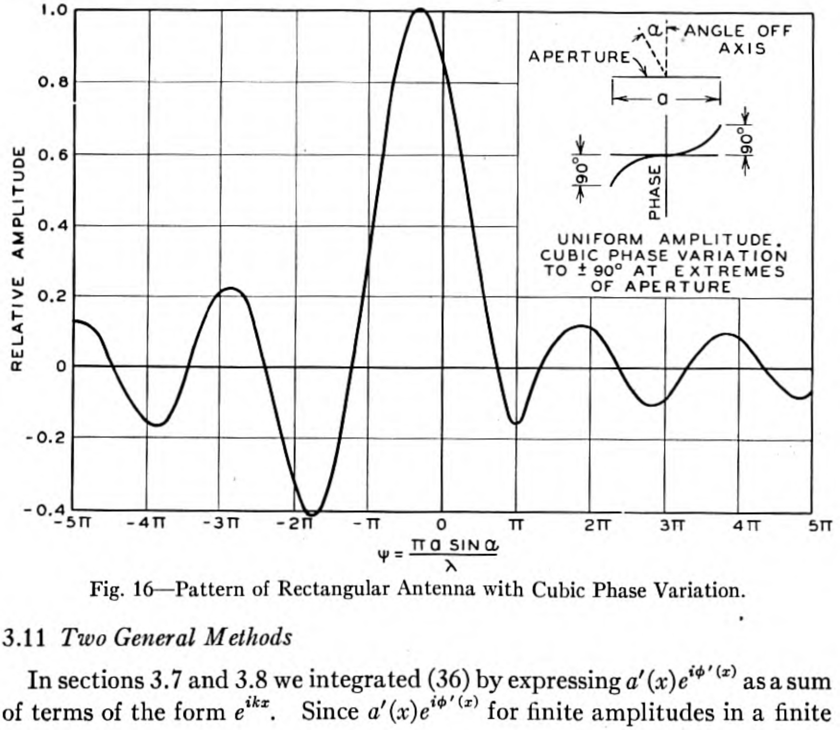


Fig. 16—Pattern of Rectangular Antenna with Cubic Phase Variation.

3.11 *Two General Methods*

In sections 3.7 and 3.8 we integrated (36) by expressing $a'(x)e^{i\phi'(x)}$ as a sum of terms of the form e^{ikx}. Since $a'(x)e^{i\phi'(x)}$ for finite amplitudes in a finite aperture can always be expressed as a Fourier sum of this form this solution can in principle always be found.

Alternatively in section 3.10 the integral was evaluated as a sum of integrals of the general type $\int x^n e^{ikx}\,dx$. Since $a'(x)e^{i\phi'(x)}$ for finite amplitudes in a finite aperture can always be expressed in terms of a power series, this solution can also in principle always be found.

3.12 *Arrays*

When the aperture consists of an array of component or unit apertures the evaluation of (36) must be made in part through a summation. When all of the elementary apertures are alike this summation can be reduced to the determination of an 'Array Factor'. The pattern of the array is given by multiplying the array factor by the pattern of a single unit.

The pattern of an array of identical units spaced equally at distances somewhat less than a wavelength can be proved to be usually almost equivalent to the pattern of a continuous wave front with the same average energy density and phase in each region.

3.13 *Limitations to Antenna Wave Front Analysis*

Through the analysis of antenna characteristics by means of wave front theory as based on equation (17) we have been able to demonstrate some of the fundamental theoretical principles of antenna design. The use of this simple approach is justified fully by its relative simplicity and by its applicability to the majority of radar antennas. Nevertheless it cannot always be used. It will certainly be inaccurate or inapplicable in the following cases:

(1) When any dimension of the aperture is of the order of a wavelength or smaller (as in many primary feeds).

(2) Where large variations in the amplitude or phase in the aperture occur in distances which are of the order of a wavelength or smaller (as in dipole arrays).

(3) Where the antenna to be considered does not act essentially through the generation of a plane wave front (as in an end fire antenna or a cosecant antenna).

When the wave front analysis breaks down alternative satisfactory approaches based on Maxwell's equation are sometimes but not always fruitful. Literature on more classical antenna theory is available in a variety of sources. For much fundamental and relevant theoretical work the reader is referred to Schelkunoff.[5]

[5] S. A. Schelkunoff, Loc. Cit.

4. Application of General Principles

In the foregoing sections we have provided some discussion of what happens to a radar signal from the time that the pulse enters the antenna on transmission until the time that the reflected signal leaves the antenna on reception. We have for convenience divided the principles which chiefly concern us into three groups, transmission line theory, transmission principles and wave front theory.

With the aid of transmission line theory we can examine problems concerning locally guided or controlled energy. The details of the problems of antenna construction, such as those to be discussed in Part II frequently demand a grasp of transmission line theory. With it we can study local losses, due to resistance or leakage, which affect the gain of the antenna. We can examine reflection problems and their effect on the match of the antenna. Special antennas, such as those employing phase shifters or transmission between parallel conducting plates, introduce many special problems which lie wholly or partly in the transmission line field.

An understanding of the principles which govern transmission through free space aids us in comprehending the radar antenna field as a whole. Through a general understanding of antenna gains and effective areas we are better equipped to judge their significance in particular cases, and to evaluate and control the effects of particular methods of construction on them.

Wave front theory provides us with a powerful method of analysis through which we can connect the radiation characteristics produced by a given antenna with the radiating currents in the antenna. Through it we can examine theoretical questions concerning beam widths and shape, unwanted radiation and gain.

An understanding of theory is necessary to the radar antenna designer, but it is by no means sufficient. It is easy to attach too much importance to theoretical examination and speculation while neglecting physical facts which can 'make or break' an antenna design. Theory alone provides no substitute for the practical 'know how' of antenna construction. It cannot do away with the necessity for careful experiment and measurement. Least of all can it replace the inventiveness and aggressive originality through which new problems are solved and new techniques are developed.

PART II

METHODS OF ANTENNA CONSTRUCTION

5. General

Techniques are essential to technical accomplishment. An understanding of general principles alone is not enough. The designing engineer must have at his disposal or develop practical methods which can produce the results he requires. The effectiveness and simplicity of these methods are fair measures of the degree of technical development.

The study of methods of radar antenna construction is the study of the means by which radar antenna requirements are met. In a broader sense this includes an examination of mechanical structures, of the metals and plastics from which antennas are made, of the processes by which they are assembled, and of the finishes by which they are protected from their environment. It might include a study of practical installation and maintenance procedures. But these matters, which like the rest of Radar have unfolded widely during the war, are beyond the scope of this paper. An adequate discussion of them would have to be based on hundreds of technical reports and instruction manuals and on thousands of manufacturing drawings. The account of methods which is to follow will therefore be restricted to a discussion, usually from the electrical point of view, of the more useful and common radar antenna configurations.

6. Classification of Methods

During the history of radar, short as it is, many methods of antenna construction have been devised. To understand the details of all of these methods and the diverse applications of each is a task that lies beyond the ability of any single individual. Nevertheless most of the methods fall into one or another of a limited collection of groups or classifications. We can grasp most of what is generally important through a study of these groups.

In order to provide a basis for classification we will review briefly, from a transmitting standpoint, the action of an antenna. Any antenna is in a sense a transformer between a transmission line and free space. More explicitly, it is a device which accepts energy incident at its terminals, and converts it into an advancing electromagnetic wave with prescribed amplitude, phase and polarization over an area. In order to do this the antenna must have some kind of energy distributing system, some means of amplitude control and some means of phase control. The distributed energy must be suitably controlled in phase, amplitude and polarization.

All antennas perform these functions, but different antennas perform them by different means. Through an examination of the means by which they are performed and the differences between them we are enabled to classify methods of antenna construction.

To distribute energy over its aperture an antenna can use a branching system of transmission lines. When this is done the antenna is an array. Arrays are particularly common in the short wave communication bands, but somewhat less common in the microwave radar bands. In a somewhat simpler method the antenna distributes energy over an area by radiating it from an initial source or 'primary feed'. This distribution can occur in both dimensions at once, as from a point source. Alternatively the energy can be radiated from a primary source but be constrained to lie between parallel conducting plates so that it is at first distributed only over a long narrow aperture or 'line source'. Distribution over the other dimension occurs only after radiation from the line source.

In order to control the amplitude across the aperture of an array antenna we must design the branching junctions so that the desired power division occurs in each one. When the energy is distributed by radiation from a primary source we must control the amplitude by selecting the proper primary feed directivity.

We can control the phase of an array antenna by choosing properly the lengths of the branching lines. Alternatively we can insert appropriate phase changers in the lines.

When the energy is distributed by primary feeds, methods resembling those of optics can be used to control phase. The radiation from a point source is spherical in character. It can be 'focussed' into a plane wave by means of a paraboloidal reflector or by a spherical lens. The radiation from a point source between parallel plates or from a uniphase line source is cylindrical in character. It can be focussed by a parabolic cylinder or a cylindrical lens.

In Table A we have indicated a possible classification of methods of radar antenna construction. This classification is based on the differences discussed in the foregoing paragraphs.

TABLE A

CLASSIFICATION OF METHODS OF RADAR ANTENNA CONSTRUCTION

Methods of Radar Antenna Construction	Arrays of	Dipoles		
		Polyrods		
		Optical Elements		
	Optical Methods	Spherical Optics	Point sources	Dipole Arrays
				Wave Guide Apertures
			and	
			Spherical Elements	Reflectors
				Lenses
		Cylindrical Optics	Line sources	Arrays
				Reflectors
				Lenses
			and	
			Cylindrical Elements	Reflectors
				Lenses

7. Basic Design Formulation

Certain design factors are common to almost all radar antennas. Because of their importance it would be well to consider these factors in a general way before proceeding with a study of particular antenna techniques.

Almost every radar antenna, regardless of how it is made, has a well defined aperture or wave front. Through wave front analysis we can often examine the connections between the Huygens sources in the antenna aperture and the radiation characteristics of the antenna. We can, in other words, use wave front analysis to study the fundamental antenna design factors, provided the analysis does not violate one of the conditions of section 3.13.

7.1 *Dimensions of the Aperture*

The dimensions of the aperture of a properly designed antenna are related to its gain by simple and general approximate relations. If the aperture is uniphase and has an amplitude distribution that is not too far from constant the relation

$$G = \frac{4\pi A}{\lambda^2}$$

is useful in connecting the gain of an antenna with the area of its aperture. The effective area is related to the area of the aperture by the equation

$$A = \eta S$$

where η is an efficiency factor. In principle η could have any value but in practice for microwave antennas η has always been less than one. Its value for most uniphase and tapered amplitude antennas is between 0.4 and 0.7. In special cases, e.g. for cosecant antennas or for some scanners its value may be less than 0.4.

The necessary dimensions for the aperture may be determined from the required beam widths in two perpendicular directions. Beam widths are usually specified as half power widths, that is by the number of degrees between directions for which the one way response is 3 *db* below the maximum response. Figure 11 shows that for an ideal rectangular antenna with uniform phase, polarization and amplitude $\alpha_{P/2} = 51 \frac{\lambda}{a}$ degrees where $\alpha_{P/2} =$ half power width in degrees, $a =$ aperture dimension and $\lambda =$ wavelength. The relation $\alpha_{P/2} = 65 \frac{\lambda}{a}$ degrees is more nearly correct for the majority of practical antennas with round or elliptical apertures and with uniform phase and reasonably tapered amplitudes.

7.2 *Amplitude Distribution*

Except where special, in particular cosecant, patterns are desired the principle factors affecting amplitude distribution are efficiency and required minor lobe level. The amplitude distribution or taper of an ideal uniphase rectangular wave front affects the minor lobe level as indicated by Figures 11, 12 and 13. Practical antennas tend to fall somewhat below this ideal picture because of non-uniform phase and because of variations from the ideal amplitude distribution due to discontinuities in the aperture and undesired leakage or spillover of energy. Nevertheless a commonly used rule of thumb is that minor lobes 20 *db* or more below the peak radiation level are tolerable and will not be exceeded with a rounded amplitude taper of 10 or 12 *db*.

7.3 *Phase Control*

Uniphase wave fronts are used whenever a simple pattern with prescribed gain, beam widths and minor lobes is to be obtained with minimum aperture dimensions. When special results are desired such as cosecant patterns or scanning beams the phase must be varied in special ways.

Mechanical tolerances in the antenna structure make it impossible to hold phases precisely to the desired values. The accuracy with which the phases can be held constant in practice varies with the technique, the antenna size and the wave length. Undesired phase variations increase the minor lobes and reduce the gain of an antenna. The extent to which phase variations can be expected to reduce the gain is indicated in Fig. 17.

8. Parabolic Antennas

The headlights of a car or the searchlights of an antiaircraft battery use reflectors to produce beams of light. Similarly the majority of radar antennas employ reflectors to focus beams of microwave energy. These reflectors may be exactly or approximately parabolic or they may have special shapes to produce special patterns. If they are parabolic they may be paraboloids which are illuminated by point sources and focus in both directions, or they may be parabolic cylinders which focus in only one direction. If they are parabolic cylinders they may be illuminated by line sources or they may be confined between parallel conducting plates and illuminated by point sources to produce line sources.

8.1 *Control of Phase*

A simple and natural way to distribute energy smoothly in space is to radiate it from a relatively nondirectional 'primary' source such as a dipole array or an open ended wave guide. This energy will be formed into a directive beam if a reflector is introduced to bring it to a plane area or wave front with constant phase. If the primary source is effectively a point as far as phase is concerned, that is if the radiated energy has the same phase for all points which are the same distance from a given point, then the reflector should be parabolic. This can be proved by simple geometrical means.

In Fig. 18 let the point source S coincide with the point $x = f$, $y = O$ of a coordinate system and let the uniphase wave front coincide with the line $x = f$. Let us assume that one point O of the reflector is at the origin. Then it can be shown that any other point of the reflector must lie on the curve

$$y^2 = 4fx$$

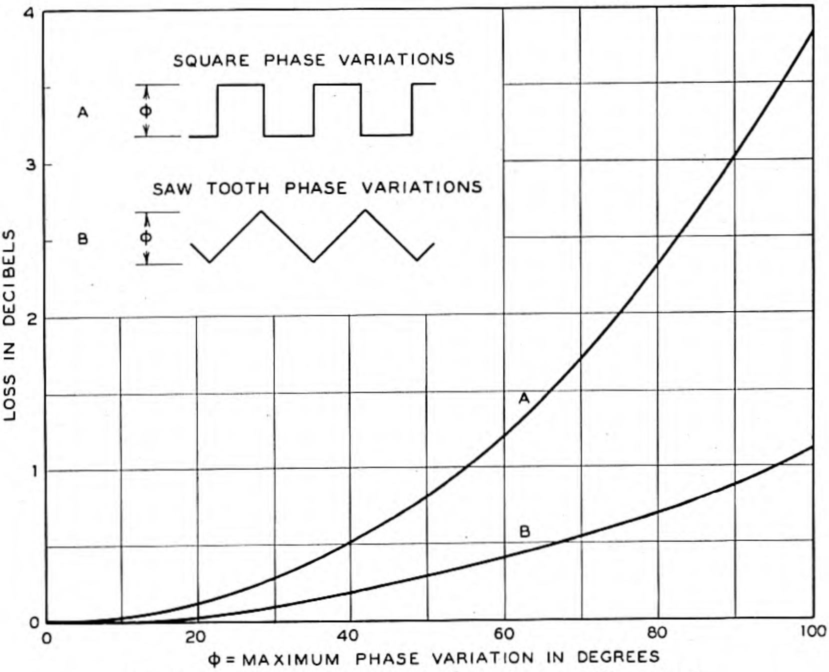


Fig. 17—Loss due to Phase Variation in Antenna Wave Front.

This is a parabola with focus at f, o and focal length f.

The derivation based on Fig. 18 is two dimensional and therefore in principle applies as it stands only to line source antennas employing parabolic cylinders bounded by parallel conducting planes (Fig. 24 and 25). If Fig. 18 is rotated about the X axis the parabola generates a paraboloid of revolution (Fig. 3). This paraboloid focusses energy spreading spherically from the point source at S in such a way that a uniphase wave front over a plane area is produced. Alternatively Fig. 18 can be translated in the Z direction perpendicular to the XY plane. The parabola then generates a parabolic cylinder and the point source S generates a line source at the focal line of the parabolic cylinder (Fig. 19). The energy spreading cylindrically from the line source is focussed by the parabolic cylinder in such a way that a uniphase wave front over a plane area is again produced. Parabolic cylinders and paraboloids are both used commonly in radar antenna practice.

In the discussion so far it has been assumed that the primary source is effectively a point source and that the reflector is exactly parabolic. If the primary source is not effectively a point source, in other words if it produces waves which are not purely spherical, then the reflector must be distorted from the parabolic shape if it is to produce perfect phase correction. When this occurs the correct reflector shape is sometimes specified on the basis of an experimental determination of phase.

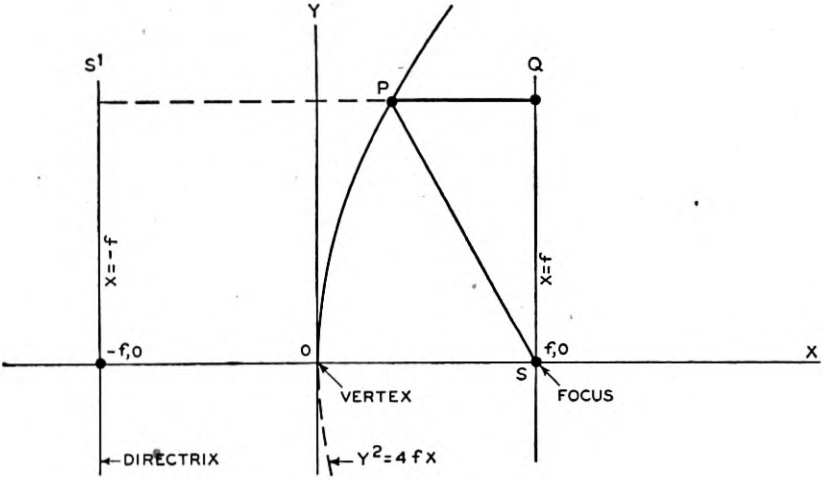


Fig. 18—Parabola.

8.2 *Control of Amplitude*

When a primary source is used to illuminate a parabolic reflector there are two factors which affect the amplitude of the resulting wave front. One of these is of course the amplitude pattern of the primary source. The other is the geometrical or space attenuation factor which is different for different parts of the wave front. In most practical antennas each of these factors tends to taper the amplitude so that it is less at the edges of the antenna than it is in the central region. The effective area of the antenna is reduced by this taper.

In any finite parabolic antenna some of the energy radiated by the primary source will fail to strike the reflector. The effective area must also be reduced by the loss of this 'spill-over' energy.

The maximum effective area for a parabolic antenna is obtained by designing the primary feed to obtain the best compromise between loss due to taper and loss due to spill-over. It has been shown theoretically that this best compromise generally occurs when the amplitude taper across the aperture is about 10 or 12 *db* and that in the neighborhood of the optimum the efficiency is not too critically dependent on the taper.[6]

This theoretical result is well justified by experience and has been applied to the majority of practical parabolic antennas. It applies both when the reflector is paraboloidal so that taper in both directions must be considered and when the reflector is a parabolic cylinder with only a single direction of taper. It is a fortunate by-product of a 10 or 12 *db* taper that it is generally sufficient to produce satisfactory minor lobe suppression.

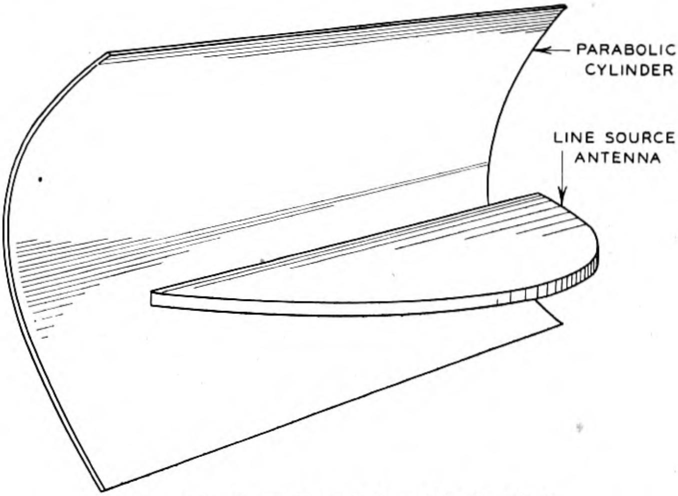


Fig. 19—A Parabolic Cylinder with Line Feed.

8.3 *Choice of Configuration*

We have shown how a simple beam can be obtained through the use of a paraboloidal reflector with a point source or alternatively through the use of a reflecting parabolic cylinder and a line source. The line source itself can be produced with the help of a parabolic cylinder bounded by parallel conducting plates. We will now outline certain practical considerations. These considerations may determine which of the two reflector types will be

[6] C. C. Cutler, Parabolic Antenna Design for Microwaves, paper to be published in Proc. of the I. R. E.

used for a particular job. They may help in choosing a focal length and in determining which finite portion of a theoretically infinite parabolic curve should be used. Finally they may assist in determining whether reflector technique is really the best for the purpose at hand or whether we could do better with a lens or an array.

In designing a parabolic antenna it must obviously be decided at an early stage whether a paraboloid or one or more parabolic cylinders are to be employed. This choice must be based on a number of mechanical and electrical considerations. Paraboloids are more common in the radar art than parabolic cylinders and are probably to be preferred, yet a categorical a priori judgment is dangerous. It will perhaps be helpful to compare the two alternatives by the simple procedure of enumerating some features in which each is usually preferable to the other.

Paraboloidal antennas

(a) are simpler electrically, since point sources are simpler than line sources.

(b) are usually lighter.

(c) are more efficient.

(d) have better patterns in the desired polarization.

(e) are more appropriate for conical lobing or spiral scanning.

Antennas employing parabolic cylinders

(a) are simpler mechanically since only singly curved surfaces are required.

(b) have separate electrical control in two perpendicular directions.

This last advantage of parabolic cylinders is important in special antennas, many of which will be described in later sections. It is useful where antennas with very large aspect ratios (ratio of dimensions of the aperture in two perpendicular directions) are desired. It is highly desirable where control in one direction is to be achieved through some special means, as in cosecant antennas, or in antennas which scan in one direction only.

Let us suppose that we have selected the aperture dimensions and have decided whether the reflector is to be paraboloidal or cylindrical. The reflector is not yet completely determined for we are still free in principle to use any portion of a parabolic surface of any focal length. In order to obtain economy in physical size the focal length is generally made between 0.6a and 0.25a where 'a' is the aperture. For the same reason a section of the reflecting surface which is located symmetrically about the vertex is often chosen (Figures 3 and 19).

When a symmetrically located section of the reflector is used certain difficulties are introduced. These difficulties, if serious enough so that their removal justifies some increase in size can be bypassed through the use of an offset section as shown in Fig. 20. We can comment on these difficulties as follows:

1. The presence of the feed in the path of the reflected energy causes a region of low intensity or 'shadow' in the wave front. The effect of this shadow on the antenna pattern depends on the size and shape of the feed and on the characteristics of the portion of the wave front where it is located. Its effect is to subtract from the undisturbed pattern a 'shadow pattern' component which is broad in angle. This decreases the gain and increases the minor lobes as indicated in Fig. 21.[7]

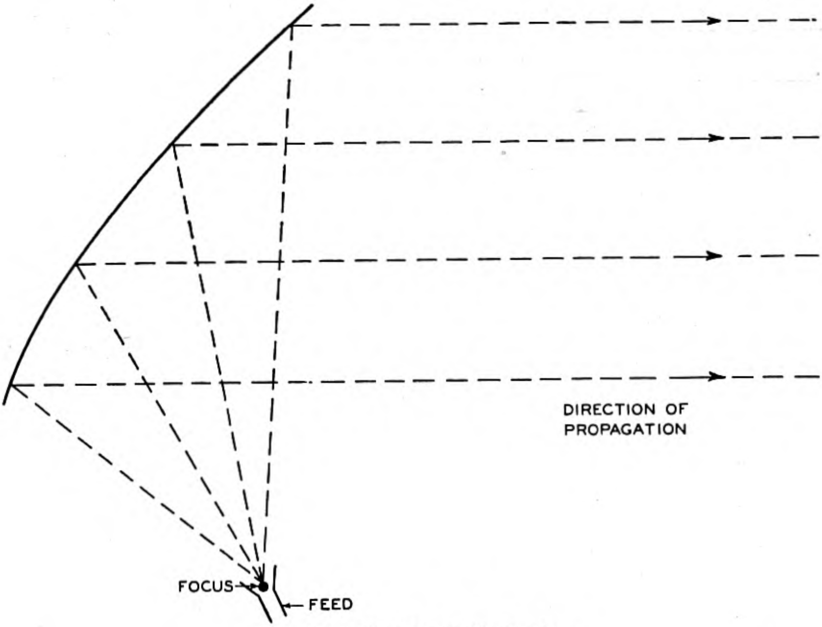


Fig. 20—Offset Parabolic Section.

2. Return of reflected energy into the feed introduces a standing wave of impedance mismatch in the feed line which is constant in amplitude but varies rapidly in phase as the frequency is varied. A mismatch at the feed which cancels the standing wave at one frequency will add to it at another frequency. A mismatch which will compensate over a band can be introduced by placing a raised plate of proper dimensions at the vertex of the reflector as indicated in Fig. 22, but such a plate produces a harmful effect on the pattern. In an antenna which must operate over a broad band it is consequently usually better to match

[7] Figures 21, 22, and 23 are taken from C. C. Cutler, loc. cit.

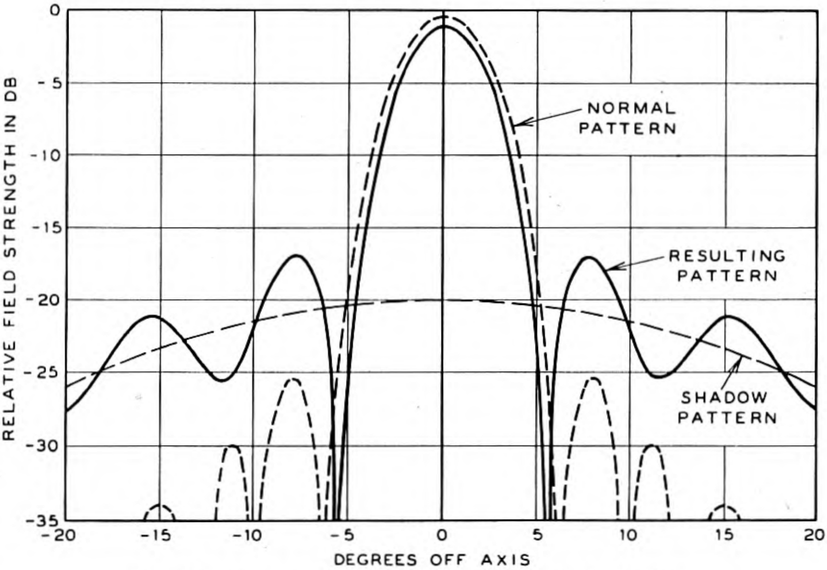


Fig. 21—Effect of Shadow on Paraboloid Radiation Pattern.

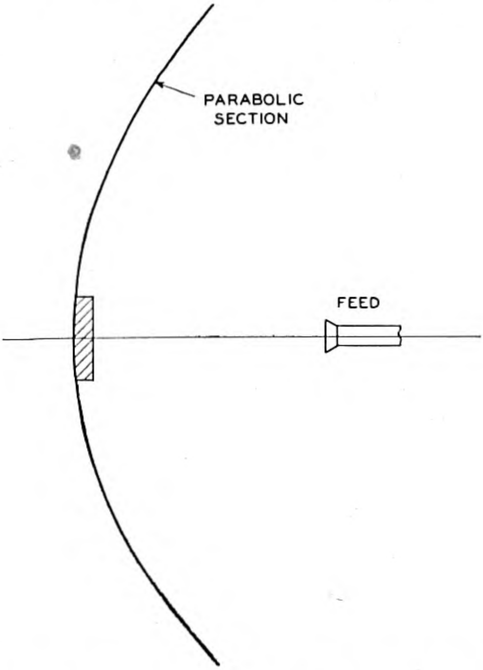


Fig. 22—Apex Matching Plate for Improving the Impedance Properties of a Parabola.

the feed to space and accept the residual standing wave, or if this is too great to use an offset section of the parabolic surface.

8.4 *Feeds for Paraboloids*

We have seen that an antenna with good wave front characteristics and consequently with a good beam and pattern can be constructed by illuminating a reflecting paraboloid with a properly designed feed placed at its focus. In this section we will examine the characteristics which the feed should have and some of the ways in which feeds are made in practice.

A feed for a paraboloid should

a. be appropriate to the transmission line with which it is fed. This is sometimes a coaxial line but more commonly a waveguide.

b. Provide an impedance match to this feed line. This match should usually be obtained in the absence of the reflector but sometimes, for narrow band antennas, with the reflector present.

c. have a satisfactory phase characteristic. For a paraboloid the feed should be, as far as phase is concerned, a true point source radiating spherical waves. As discussed at the end of 8.1, if the wave front is not accurately spherical, a compensating correction in the reflector can be made.

d. have a satisfactory amplitude characteristic. According to 8.2 this means that the feed should have a major radiation lobe with its maximum striking the center of the reflector, its intensity decreasing smoothly to a value about 8 to 10 *db* below the maximum in the direction of the reflector boundaries and remaining small for all directions which do not strike the reflector.

e. have a polarization characteristic which is such that the electric vectors in the reflected wave front will all be polarized in the same direction.

f. not disturb seriously the radiation characteristics of the antenna as a whole. The shadow effect of the feed, the feed line and the necessary mechanical supports must be small or absent. Primary radiation from the feed which does not strike the reflector or reflected energy which strikes the feed or associated structure and is then reradiated must be far enough down or so controlled that the antenna pattern is as required.

In addition to the electrical requirements for a paraboloid feed it must of course be so designed that all other engineering requirements are met, it must be firmly supported in the required position, must be connected to the antenna feed line in a satisfactory manner, must sometimes be furnished with an air tight or water tight seal, and so forth.

From the foregoing it is evident that a feed for a paraboloid is in itself a small relatively non-directive antenna. Its directivity is somewhat less than that obtained with an ordinary short wave array. It is therefore not surprising that dipole arrays are sometimes used in practice to feed paraboloids.

A simple dipole or half wave doublet can in itself be used to feed a paraboloid, but it is inefficient because of its inadequate directivity. It is preferable and more common to use an array in which only one doublet is excited directly and which contains a reflector system consisting of another doublet or a reflecting surface which is excited parasitically.

Dipole feeds although useful in practice have poor polarization characteristics and although natural when a coaxial antenna feed line is used are less convenient when the feed line is a waveguide. Since waveguides are more common in the microwave radar bands it is to be expected that waveguide feeds would be preferred in the majority of paraboloidal antennas.

The most easily constructed waveguide feed is simply an open ended waveguide. It is easy to permit a standard round or rectangular waveguide transmitting the dominant mode to radiate out into space toward the paraboloid. It will do this naturally with desirable phase, polarization and amplitude characteristics. It is purely coincidental, however, when this results in optimum amplitude characteristics. It is usually necessary to obtain these by tapering the feed line to form a waveguide aperture of the required size and shape. The aperture required may be smaller than a standard waveguide cross section so that its directivity will be less. In this case it may be necessary to 'load' it with dielectric material so that the power can be transmitted. It may be greater, in which case it is sometimes called an 'electromagnetic horn'. It may be greater in one dimension and less in the other, as when a paraboloidal section of large aspect ratio is to be illuminated.

If a single open ended waveguide or electromagnetic horn is used to feed a section of the paraboloid which includes the vertex, the waveguide feed line must partially block the reflected wave in order to be connected to the feed. To avoid this difficulty several rear waveguide feeds have been used. In this type of feed the waveguide passes through the vertex of the paraboloid and serves to support the feed at the focus. The energy can be caused to radiate back towards the reflector in any one of several ways, some of which involve reflecting rings or plates or parasitically excited doublets. The 'Cutler' feed[8] is perhaps the most successful and common rear feed. It operates by radiating the energy back towards the paraboloid through two apertures located and excited as shown in Fig. 23.

[8] C. C. Cutler, Loc. Cit.

8.5 *Parabolic Cylinders between Parallel Plates*

In 8.0 we saw that parabolic cylinders may be illuminated by line sources or that they may be confined between parallel plates and illuminated by point sources to produce line sources. In either of these two cases the characteristics which the feed should have are specified accurately by the conditions stated at the beginning of 8.4 for paraboloidal feeds with the exceptions that condition *c* must be reworded so that it applies to cylindrical rather than to spherical optics.

We will first consider parabolic cylinders bounded by parallel plates because in doing so we describe in passing one form of feed for unbounded parabolic cylinders. Two forms of transmission between parallel plates are used in practice.

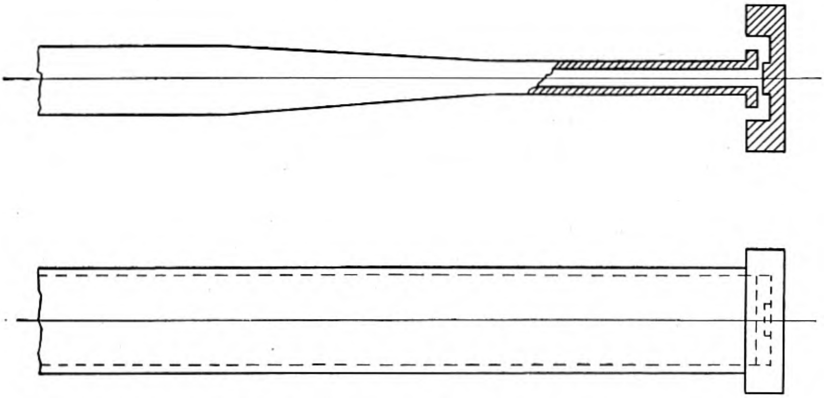


Fig. 23—Dual Aperture Rear Feed Horn.

a. The transverse electromagnetic (TEM) mode in which the electric vector is perpendicular to the plates. This is simply a slice of the familiar free space wave and can be propagated regardless of the spacing between the plates. It is the only mode that can travel between the plates if they are separated less than half a wavelength. Its velocity of propagation is independent of plate spacing.
b. The TE_{01} mode in which the electric vector is parallel to the plates. This mode is similar to the dominant mode in a rectangular waveguide and differs from it only in that it is not bounded by planes perpendicular to the electric vector. It can be transmitted only if the plate spacing is greater than half a wavelength, is the *only* parallel mode that can exist if the spacing is under a wavelength and is the only symmetrical parallel mode that can exist if the plate spacing is under three

halves of a wavelength. Its phase velocity is determined by the plate spacing in a manner given by the familiar waveguide formula

$$V_g = \frac{c}{\sqrt{\epsilon - \left(\frac{\lambda}{2a}\right)^2}}$$

where '*c*' is the velocity of light, ϵ is the dielectric constant relative to free space of the medium between the plates, λ is the wavelength in air and '*a*' is the plate spacing.

The TEM mode between parallel plates can be generated by extending the central conductor of a coaxial perpendicularly into or through the wave space and backing it up with a reflecting cylinder as indicated in Fig. 24.

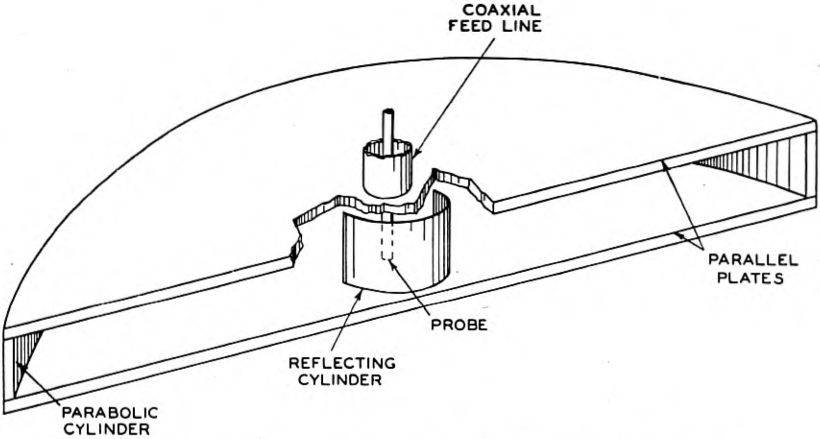


Fig. 24—Parabolic Cylinder Bounded by Parallel Plates. Probe Feed.

Alternatively this mode can be generated as indicated by Fig. 25 by a waveguide aperture with the proper polarization.

The TE_{01} mode, when used, is usually generated by a rectangular waveguide aperture set between the plates with proper polarization as indicated in Fig. 25. Care must be taken that only the desired mode is produced. The TEM mode will be unexcited if only the desired polarization is present in the feed. The next parallel mode is unsymmetrical and therefore even if it can be transmitted will be unexcited if the feed is placed symmetrically with respect to the two plates.

Parallel plate antennas as shown in figures 24 and 25 are useful where particularly large aspect ratios are required. The aperture dimension perpendicular to the plates is equal to the plate spacing and therefore small.

It can be increased somewhat by the addition of flares. The other dimension can easily be made large.

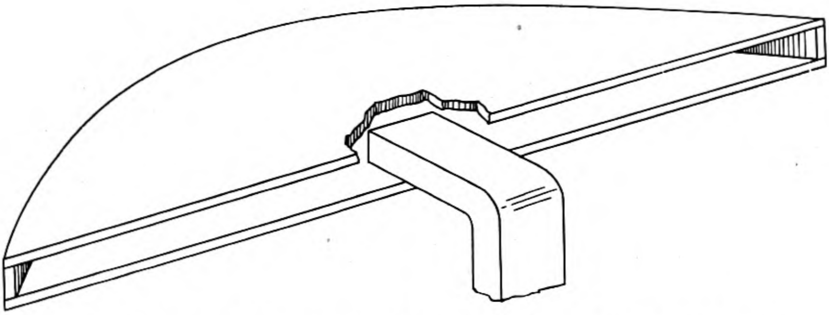


Fig. 25—Parabolic Cylinder Bounded by Parallel Plates. Wave Guide Feed.

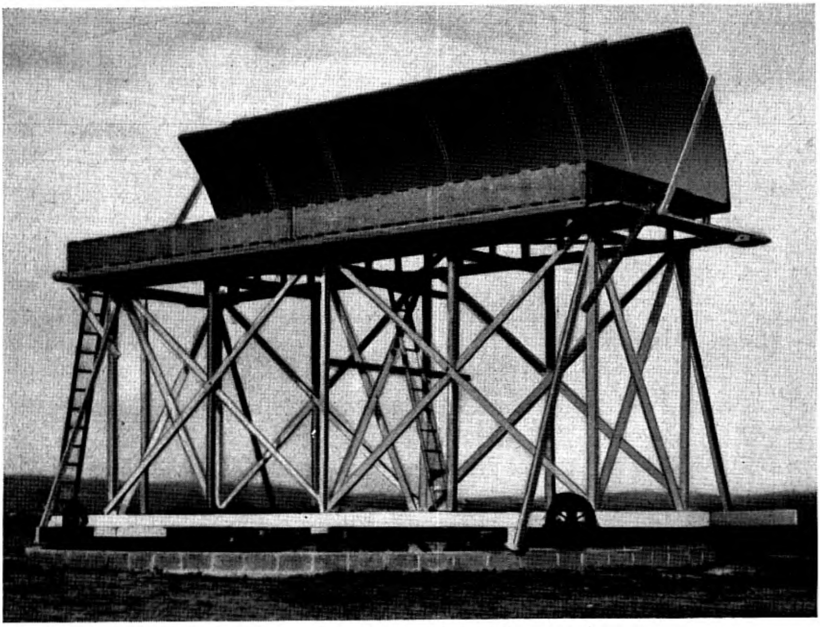


Fig. 26.—Experimental 7′ x 32′ Antenna.

8.6 *Line Sources for Parabolic Cylinders*

A line source for a parabolic cylinder is physically an antenna with a long narrow aperture. Any means for obtaining such an aperture can be used in producing a line sourçe. Parallel plate systems as described in 8.5 have been used as line sources in several radar antennas. The large (7′ x 32′)

experimental antenna shown in Fig. 26 was one of the first to illustrate the practicality of this design.

The horizontal pattern of the 7′ x 32′ antenna is plotted in Fig. 27. The horizontal beam width is seen to be of the order of 0.7 degrees.

The antenna illustrated in Fig. 26 is interesting in another way for it is a good example of a type of experimental construction which was extremely useful in wartime antenna development. Research and development engineers found that they could often save months by constructing initial models of wood. Upon completion of a wooden model electrically important surfaces were covered with metal foil or were sprayed or painted with metal. Thus, where tolerances permitted, the carpenter shop could replace the relatively slow machine shop.

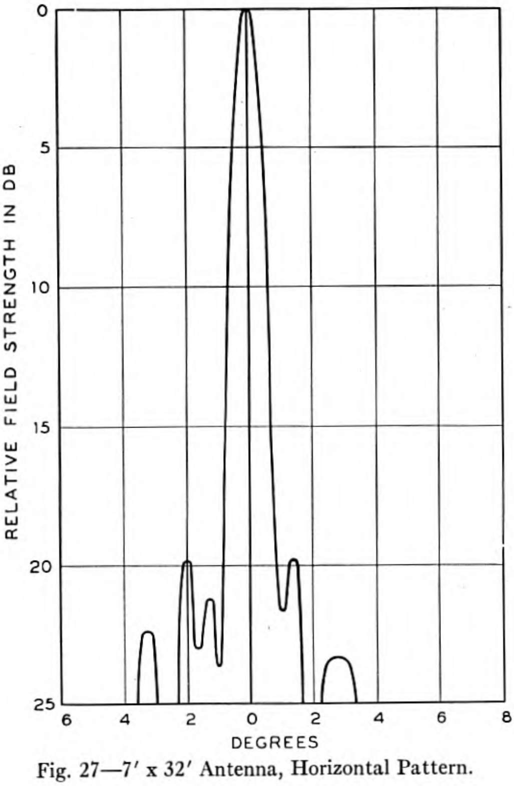


Fig. 27—7′ x 32′ Antenna, Horizontal Pattern.

Another form of parallel plate line feed results when a plastic lens is placed between parallel plates and used as the focussing element. A linear array of elements excited with the proper phase and amplitude can also be used. Some discussion of alternative approaches will appear in the section on scanning techniques.

8.7 *Tolerances in Parabolic Antennas*

The question of tolerances will always arise in practice. Ideal dimensions are only approximated, never reached. The ease of obtaining the required accuracy is an important engineering factor.

The tolerances in paraboloidal antennas or in parabolic cylinders illuminated by line sources can be divided into three general classes:

(a) Tolerances on reflecting surfaces.

(b) Tolerances on spacial relationships of feed and reflector.

(c) Tolerances on the feed.

When the actual reflector departs from the ideal parabolic curve deviations in the phase will result. These will tend to reduce the gain and increase the minor lobes. The effects of such deviations on the gain can be estimated with the help of Fig. 17. We should recall that an error of σ in the reflector surface will produce an error of about 2σ in the phase front. Based on this kind of argument and on experience reflector tolerances are generally set in practice to about $\pm \frac{\lambda}{16}$ or $\pm \frac{\lambda}{32}$ depending on the amount of beam deterioration that can be permitted.

In Fig. 28 are compared some electrical characteristics of two paraboloidal antennas, one employing a precisely constructed paraboloidal searchlight mirror and the other a carefully constructed wooden paraboloidal reflector with the same nominal contour. This comparison is revealing for it shows the harm that can be done even by small defects in the reflector surface. Although the two patterns are almost identical in the vicinity of the main beam, the general minor lobe level of the wooden reflector remains higher at large angles and its gain is less.

It must not, however, be assumed that a solid reflecting surface is necessary to insure excellent results. Any reflecting surface which reflects all or most of the power is satisfactory provided that it is properly located. Perforated reflectors, reflectors of woven material and reflectors consisting of gratings with less than half wavelength spacing are commonly used in radar antenna practice. These reflectors tend to reduce weight, wind or water resistance and visibility. Many of them will be described in Part III of this paper.

The feed of a parabolic reflector should be located so that its phase center coincides with the focus of the reflector. If it is located at an incorrect distance from the vertex a circular curvature of phase results and the system is said to be 'defocussed' (Sec. 3.9). As the feed is moved off the axis of the reflector the first effect is a shifting of the beam due to a linear variation of the phase (Sec. 3.8). For greater distances off axis a cubic component of phase error becomes effective (Sec. 3.10). Phase error, whether circular, cubic or more complex, results in a reduction in gain and usually in an increase of minor lobes. Although the effects of given amounts of phase curvature on the radiation characteristics of an antenna can be estimated by theoretical means, it is usually easier and quicker to find them experimentally.

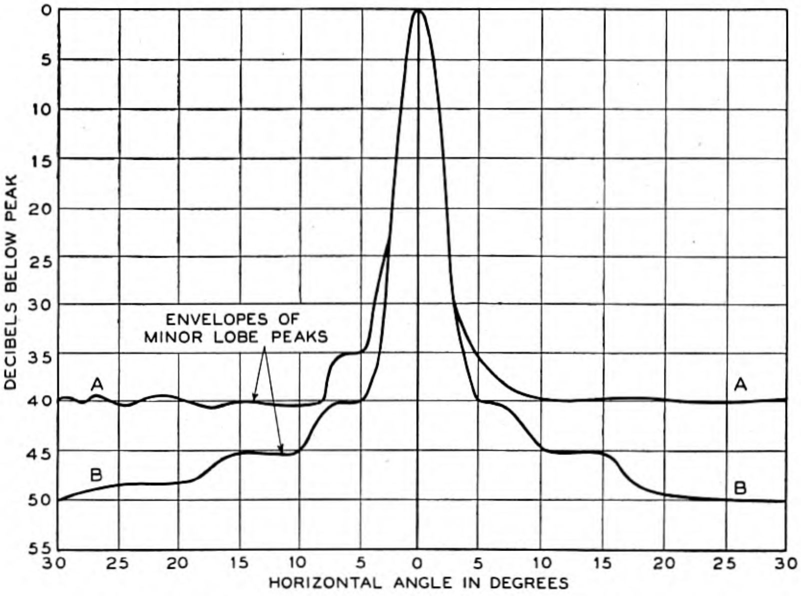


Fig. 28—Effect of Small Inaccuracies in Reflector.

The tolerances on the feed itself appear in various forms, many of which can be examined with the aid of transmission line theory and most of which are too detailed for discussion in this paper. It is generally true here also that experiment is a more effective guide than theory.

Experience has shown that when parallel plate systems are used, either as complete antennas or as line feeds for other elements, tolerances on the parallel conducting plates must be considered carefully. It is obvious that when the TE_{01} mode is used the plate spacing must be held closely, since the phase velocity is related to the spacing. This spacing can be controlled through the use of metallic spacers perpendicular to the plates. These spacers, if small enough in cross section, do not disturb things unduly. The velocity of the TEM mode is, on the other hand, almost independent of the plate spacing. This mode is, however, more likely to cause trouble by leaks through joints and cracks in the plates.

9. Metal Plate Lenses

At visible wavelengths lenses have, in the past, been far more common than in the microwave region, due chiefly to the absence of satisfactory lens materials. A solid lens of glass or plastic with a diameter of several feet is a massive and unwieldy object. By zoning, which will be discussed below, these difficulties can be reduced but they still remain.

A new lens technique, particularly effective in the microwave region was developed by the Bell Laboratories during the war.[9] It is evident that any material in which the phase velocity is different from that of free space can be used to make a phase correcting lens. The material which is used in this new technique is essentially a stack of equally spaced metal plates parallel to the electric vector of the wave front and to the direction of propagation. Lenses made from this material are called 'Metal Plate Lenses'.

When the spacing between neighboring plates is between $\lambda/2$ and λ only one mode with electric vector parallel to the plates can be transmitted. This is the TE_{01} mode for which the phase velocity is given in Sec. 8.5. When the medium between the plates is air this equation can be converted into the expression

$$N = \sqrt{1 - \left(\frac{\lambda}{2a}\right)^2}$$

for the effective index of refraction. Here λ is the wavelength in air and a is the plate spacing.

As a varies between $\lambda/2$ and λ, N varies as indicated in Fig. 29. In the neighborhood of $a = \lambda$, N is not far from 1 and as a approaches $\lambda/2$, N approaches 0. Since N is always less than 1 we see that there is an essential difference between metal plate lenses and glass or plastic lenses for which N is always greater than 1. This difference is seen in the fact that a glass lens corrects phases by slowing down a travelling wave front, while a metal lens operates in the reverse direction by speeding it up. This means that a convergent lens with a real focus must be thinner in the center than the edge, the opposite of a convergent optical lens (Fig. 30).

Unless the value of N is considerably different from 1 it is evident that very thick lens sections must be used to produce useful phase corrections. For this reason values of 'a' not far from $\lambda/2$ should be chosen. On the other hand values of 'a' too close to $\lambda/2$ would cause undesirably large reflections

[9] W. E. Kock, "Metal Lens Antennas", Proc. I. R. E., Nov., 1946.

from the lens surfaces and impose severe restrictions on the accuracy of plate spacings. The compromises that have been used in practice are $N = 0.5$ for which $a = 0.577\lambda$ and $N = 0.6$ for which $a = 0.625\lambda$.

Even with $N = 0.5$ or 0.6 lenses become thick unless inconveniently long focal distances are used. Thick lenses are undesirable not only because they occupy more space and are heavier but also because the plate spacing must be held to a higher degree of accuracy if the phase correction is to be as required. To get around these difficulties the technique of zoning is used.

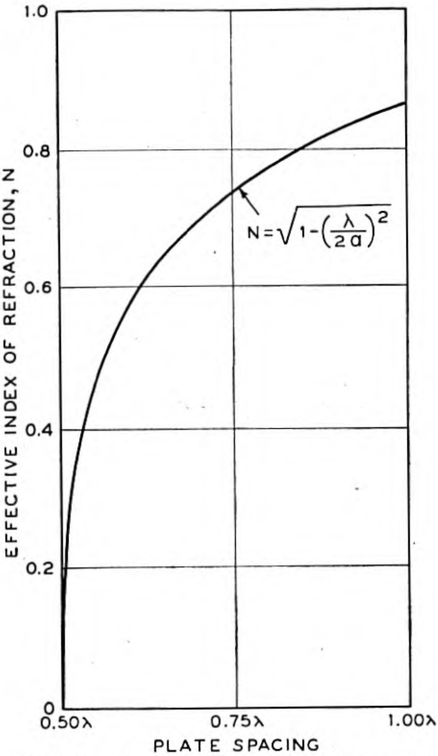


Fig. 29—Variation of Effective Index of Refraction with Plate Spacing in a Metal Plate Lens.

Zoning makes use of the fact that if the phase of an electromagnetic vector is increased or decreased by any number of complete cycles the effect of the vector is unchanged. When applied to a metal plate lens antenna this means simply that wherever the phase correction due to a portion of the lens is greater than a wavelength this correction can be reduced by some integral number of wavelengths such that the residual phase correction is under one wavelength. If this is done it is evident that no portion of the lens needs to correct the phase by more than one wavelength. It follows that no portion of the lens need to be thicker than $\lambda/(1 - N)$.

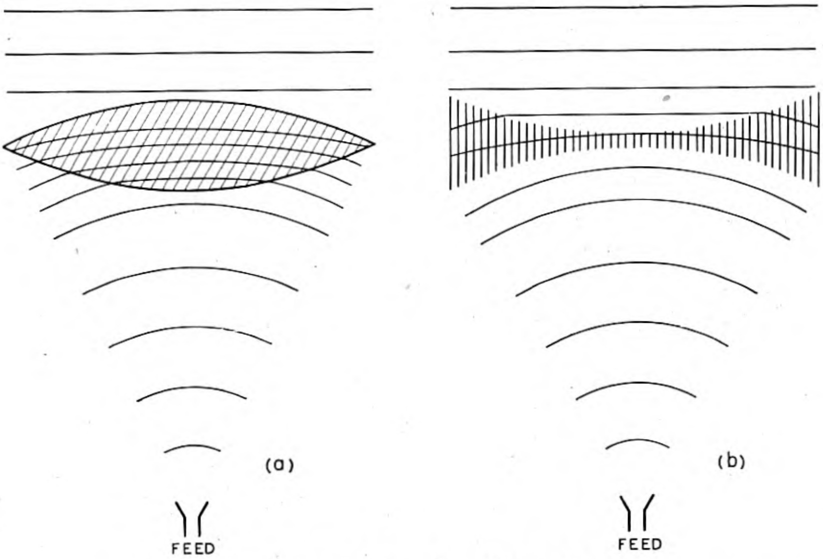


Fig. 30—Comparison of Dielectric and Metal Plate Lenses.

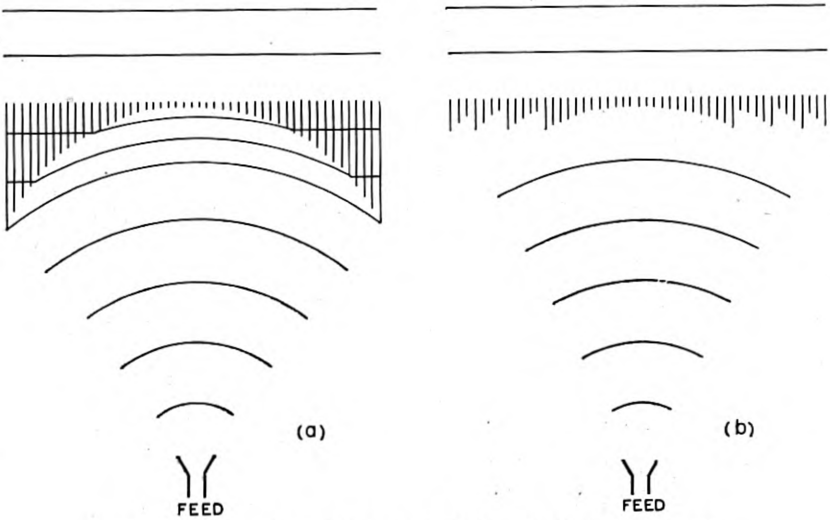


Fig. 31—Comparison of Unzoned and Zoned Metal Plate Lenses.

A cross section of a typical metal plate lens before and after zoning is illustrated in Fig. 31. This figure shows clearly why zoning reduces considerably the size and mass of a lens.

Zoning is not without disadvantages. One disadvantage is obviously that a zoned lens which is designed for one frequency will not necessarily work well at other frequencies. It is in principle possible to design a broad band zoned metal plate lens corresponding to the color compensated lenses used in good cameras. So far, however, this has not been necessary since band characteristics of simple lenses have been adequate.

Another difficulty that zoning introduces is due to the boundary regions between the zones. The wave front in this region is influenced partly by one zone and partly by the other and may as a result have undesirable phase and amplitude characteristics. This becomes serious only if especially short focal distances are used.

9.1 *Lens Antenna Configurations*

Any of the configurations which are possible with parabolic reflectors have their analogues when metal plate lenses are used. Circular lenses illuminated by point sources and cylindrical lenses illuminated by line sources are not only theoretically possible but have been built and used. Since a lens has two surfaces there is actually somewhat more freedom in lens design than in reflector design. Metal Plate Lenses have usually been designed with one surface flat, but the possibility of controlling both surfaces is emerging as a useful design factor where special requirements must be met.

Feeds for lenses should fulfill most of the same requirements as feeds for reflectors. We find a difference in size in lens feeds in that they must generally be more directive because of greater ratios of focal length to aperture. A difference in kind occurs because the feed is located behind the lens where none of the focussed energy can enter the feed or be disturbed by it. As a result some matching and pattern problems which arise in parabolic antennas are automatically absent when lenses are used.

In choosing a design for a lens antenna system with a given aperture one must compromise between the large size which is necessary when a long focal length is used and the more zones which result if the focal length is made short. Most metal plate lenses so far constructed have had focal lengths somewhere between 0.5 and 1.0 times the greatest aperture dimension.

9.2 *Tolerances in Metal Plate Lenses*

It is not difficult to see that phase errors resulting from small displacements or distortions of a metal plate lens are much less serious than those due to comparable distortions of a reflector surface. This follows from the fact that the lens operates on a wave which passes through it. If a portion of the lens is displaced slightly in the direction of propagation it is still operating on roughly the same portion of the wave front and gives it the same phase correction. If a portion of a reflector were displaced in the same way the error in the wave front would be of the order of twice the displacement. Quantitative arguments show that less severe tolerances apply to all major structural dimensions of a metal lens antenna.

It is true of course that the dimensions of individual portions of the metal lens must be held with some accuracy. The metal plate spacing determines the effective index of refraction of the lens material. Where $N = 0.5$ it is customary to require that this be held to $\pm\lambda/75$, and where $N = 0.6$ to $\pm\lambda/50$. The thickness of the lens in a given region is less critical, and must be held to $\pm \dfrac{\lambda}{16\,(1 - N)}$ where it is desired to hold the phase front to $\pm\lambda/16$.

Fig. 32 illustrates clearly the drastic way in which the location of a lens can be altered without seriously affecting the pattern. It shows, incidentally, how a lens may behave well when used as the focussing element in a moving feed scanning antenna.

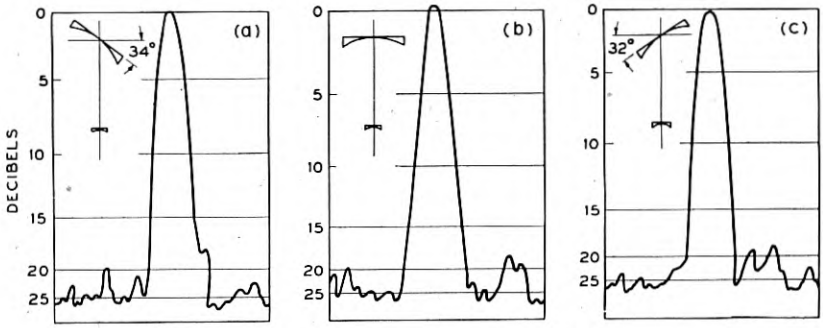


Fig. 32—Effect on Pattern of Lens Tilting.

9.3 *Advantages of Metal Plate Lenses*

On the basis of the above discussion we can see that metal plate lenses have certain considerable advantages. The most important of these is perhaps found in the practical matter of tolerances. It is a comparatively simple matter to hold dimensions of small objects to close tolerances but quite another thing to hold dimensions of large objects closely under the conditions of modern warfare. This advantage emerges with increasing importance as the wavelength is reduced.

Metal plate lenses have contributed a great degree of flexibility to radar antenna art. When they are used two surfaces rather than one may be controlled, and the dielectric constant can be varied within wide limits. Independent control in the two polarizations may be applied. We can confidently expect that they will become increasingly popular in the radar field.

10. Cosecant Antennas

One of the earliest uses of radar was for early warning against aircraft.

The skies were searched for possible attackers with antennas which rotated continuously in azimuth. An equally important but later use appeared with the advent of great bombing attacks. Bombing radars 'painted' maps of the ground which permitted navigation and bombing during night and under even the worst weather conditions. In these radars also the antennas were rotated in azimuth, either continuously through 360° or back and forth through sectors.

The majority of radars designed to perform these functions provided vertical coverage by means of a special vertical pattern rather than a vertical scan. It can easily be seen that such a pattern would have to be 'special.' If we assume, for example, that a bombing plane is flying at an altitude of 10,000 feet, then the radar range must be 10,000 *csc* $60^\circ = 11{,}500$ feet if a target on the ground at a bomb release angle of 60° from the horizontal is to be seen. Such a range would by no means be enough to pick up the target at say 10° in time to prepare for bombing, for then a range of 10,000 csc $10^\circ = 57{,}600$ feet would be required. This range is far more than is necessary for the 60° angle. It appears then that in the most efficient design the radar range and therefore the radar antenna gain, must be different in different directions.

The required variation of gain with vertical direction could be specified in any one of several ways. It seems natural to specify that a given ground target should produce a constant signal as the plane flies towards it at a constant altitude. Neglecting the directivity of the target this will occur if the amplitude response of the antenna is given by $E = E_0 csc\theta$. This same condition will apply by reciprocity to an early warning radar antenna on the ground which is required to obtain the same response at all ranges from a plane which is flying in at a constant altitude.

This condition is not alone sufficient to specify completely the vertical pattern of an antenna. For one thing it can obviously not be followed when $\theta = 0$, for this would require infinite gain in this direction. Therefore a lower limit to the value of θ for which the condition is valid must be set. In addition an upper limit less than 90° is specified whenever requirements permit, since control at high angles is especially difficult. When the limits have been set it still remains to specify the magnitude of the constant E_0. This can be done by specifying the range in one particular direction. This specification must of course be consistent with all the factors that determine gain, including the reduction due to the required vertical spread of the pattern.

10.1 *Cosecant Antennas based on the Paraboloid*

It is evident that the standard paraboloidal antennas so far discussed will not produce cosecant patterns. These patterns being unsymmetrical will result only if the wave front phase and amplitude are especially controlled.

On the other hand, because paraboloidal antennas are simple and common it is natural that many cosecant designs should be based on them. These designs can be classified into two groups, those in which the reflector is modified and those in which the feed is modified.

Some early cosecant antennas were made by introducing discontinuities in paraboloidal reflectors as illustrated in Fig. 33. These controlled the radiation more or less as desired over the desired cosecant pattern but produced rather serious minor lobes elsewhere. This difficulty can be overcome through the use of a continuously distorted surface as illustrated in Fig. 34. This reflector, first used at the Radiation Laboratories, is a normal paraboloid in the lower part whereas the upper part is the surface that would be obtained by rotating the parabola through the vertex of the upper part about its focal point.

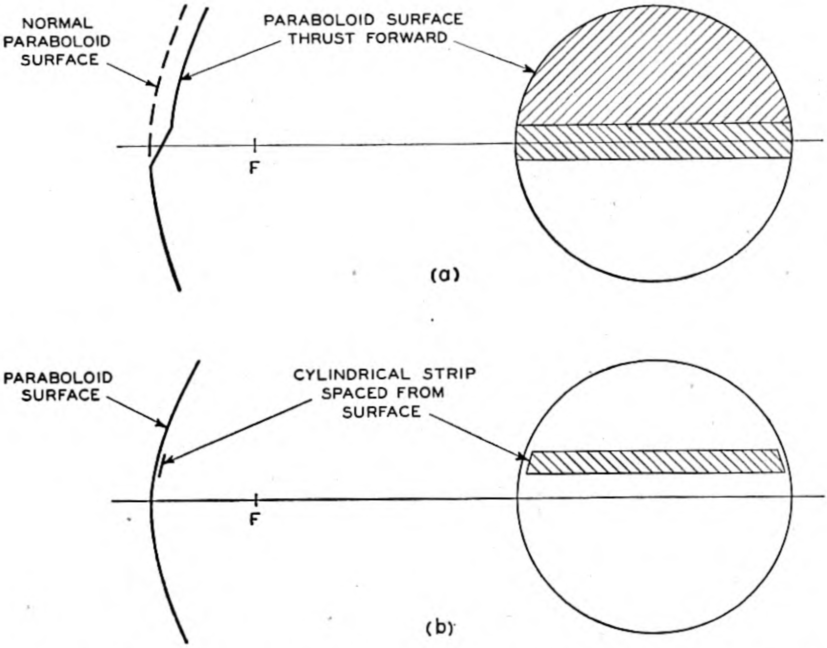

NORMAL PARABOLOID SURFACE
PARABOLOID SURFACE THRUST FORWARD
F
(a)
PARABOLOID SURFACE
CYLINDRICAL STRIP SPACED FROM SURFACE
F
(b)

Fig. 33—Some Cosecant Antennas Based on the Paraboloid (Cosecant Energy Downward).

Several types of feed have been used in combination with paraboloids to produce cosecant patterns. These are usually arrays which operate on the principle that each element is a feed which contributes principally to one region of the vertical pattern. The elements may be dipoles or waveguide apertures fed directly through the antenna feed line or they may be reflectors which reradiate reflected energy originating from a single primary source. No matter how excited they must be properly controlled in phase, amplitude and directivity.

Cosecant antennas based on the paraboloid are common and can sometimes fulfill all requirements with complete satisfaction. Nevertheless they suffer from certain disadvantages. The most serious of these is that they lack resolution at high vertical angles, that is the beam is wider horizontally at high angles. This is to be expected for reasons of phase alone, for a paraboloidal reflector is, after all, designed to focus in only one direction. If phase difficulties were completely absent however, azimuthal resolution at high angles would still be destroyed because of cross polarized components of radiation. These components arise naturally from doubly curved reflectors, even simple paraboloids. They are sometimes overlooked when antennas are measured in a one way circuit with a linearly polarized test field, but must obviously be considered in radar antennas.

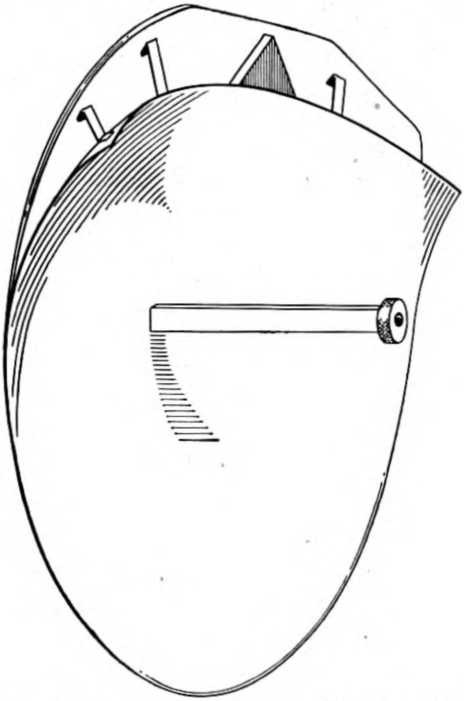


Fig. 34—Barrel Cosecant Antenna (Cosecant Energy Downward).

10.2 *Cylindrical Cosecant Antennas*

Harmful cross polarized radiation is produced by doubly curved reflectors. This radiation is difficult to control and therefore undesirable where a closely controlled cosecant characteristic at high angles is required. Although not at first evident, it seems natural now to bypass polarization difficulties through the use of singly curved cylindrical reflectors. These reflectors if illuminated with a line source of closely controlled linear polarization provide a beam which is linearly polarized. This beam has also in azimuth approximately the directivity of the line source at all vertical angles. It is thus superior in two significant respects to cosecant beams produced by doubly curved reflecting surfaces.

A cylindrical cosecant antenna consists of a cylindrical reflector illuminated by a line source. Part of the cylinder is almost parabolic and contributes chiefly to the strong part of the beam which lies closest to the horizontal. This part is merged continuously into a region which departs considerably from the parabolic and contributes chiefly to the radiation at higher angles.

Although wave front principles can be used and certainly must not be violated, the principles of geometrical optics have been particularly effective in the determination of cosecant reflector shapes. The detailed application of these principles will not be discussed here. While applying the geometrical principles the designer must be sure that the over-all size and configuration of the antenna can produce the results he wants. He must design a line source with the desired polarization and horizontal pattern and a vertical pattern which fits in with the cosecant design. In addition he must take particular care to reduce sources of pattern distortion to a level at which they cannot interfere significantly with the lowest level of the cosecant 'tail'.

11. Lobing

In many of the tactical situations of modern war radar can be used to provide fire control information. Radar by its nature determines range and microwave radar with its narrow well defined beams is a natural instrument for finding directions to a target, whether the missile to be sent to that target is a shell, a torpedo or a bomb. In fire control radar, as opposed to search or navigational radar, two properties of the antenna deserve particular attention. These are the *accuracy* and the *rate* with which direction to a target can be measured.

Lobing is a means which utilizes to the fullest extent the accuracy available from a given antenna aperture and which increases, usually as far as is desired, the rate at which this information is provided and corrected.

A lobing antenna which is to provide information concerning one angle only, azimuth for example, is capable of producing two beams, one at a time, and of switching rapidly from one to the other. This process is called *Lobe Switching.* The two beams are nearly coincident, differing in direction by about one beam width. When the signals from the two beams are compared, they will be equal only if the target lies on the bisector between the beams (Fig. 35). The two signals can be compared visually on an indicator screen of the radar or they can be compared electrically and fed directly into circuits which control the direction of fire.

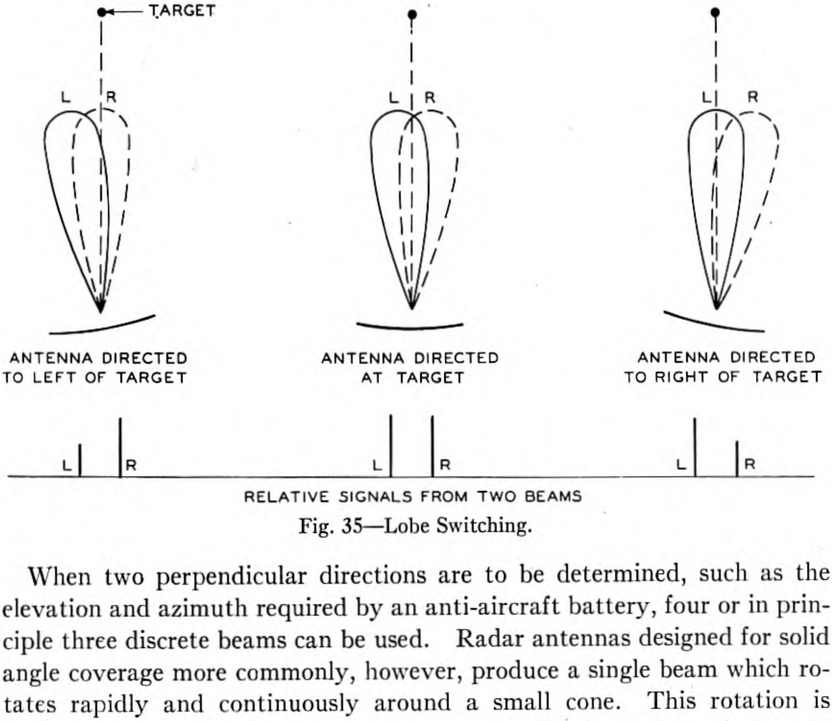

TARGET
L
R
ANTENNA DIRECTED TO LEFT OF TARGET
ANTENNA DIRECTED AT TARGET
ANTENNA DIRECTED TO RIGHT OF TARGET
RELATIVE SIGNALS FROM TWO BEAMS

Fig. 35—Lobe Switching.

When two perpendicular directions are to be determined, such as the elevation and azimuth required by an anti-aircraft battery, four or in principle three discrete beams can be used. Radar antennas designed for solid angle coverage more commonly, however, produce a single beam which rotates rapidly and continuously around a small cone. This rotation is known as *conical lobing.* A comparison of amplitudes in a vertical plane can then be used to give the elevation of the target and a similar comparison in a horizontal plane to give its azimuth. Here too the electrical signals can be compared visually on an indicator screen, but an electrical comparison will provide continuous data which can be used to aim the guns and at the same time to cause the radar antenna to follow the target automatically.

11.1 *Lobe Switching*

Two methods of lobe switching are common. In one of these the lobing antenna is an array of two equally excited elements. Each of these elements occupies one half of the final antenna aperture, and provides a uniphase front across this half. If the two elements were excited with the same phase the radiation maximum of the resulting antenna beam would occur in a direction at right angles to the combined phase front. If the phase of one element is made to lag behind that of the other by a small amount, 60° say, the phase of the combined aperture will of course be discontinuous with a step in the middle. This discontinuous phase front will approximate with a small error, a uniphase wave front which is tilted somewhat with respect to the wave fronts of the elements. The phase shift will therefore result in a slight shift of the beam away from the normal direction. When the phase shift is reversed the beam shift will be reversed. Two properly designed elementary antennas in combination with a means for rapidly changing the phase will therefore constitute a lobe switching antenna. Such an antenna is described more in detail in Sec. 14.6.

Another method of lobe switching is more natural for antennas based on optical principles. In this method two identical feeds are placed side by side in the focal region of the reflector. When one of these feeds illuminates the reflector a beam is produced which is slightly off the normal axial direction. Illumination by the other feed produces a second beam which is equally displaced in the opposite direction. The lobe of the antenna switches rapidly when the two feeds are activated alternately in rapid succession. The antenna must use some form of rapid switching appropriate to the antenna feed line. In several applications switches are used which depend on the rapid tuning and detuning of resonant cavities or irises.

11.2 *Conical Lobing*

A conically lobing antenna produces a beam which nutates rapidly about a fixed axial direction. This is usually accomplished by rotating or nutating an antenna feed in a small circle about the focus in the focal plane of a paraboloid or lens. This antenna feed can be a spinning asymmetrical dipole or a rotating or nutating waveguide aperture. It can result in a beam with linear polarization which rotates as the feed rotates, or preferably in a beam for which the polarization remains parallel to a fixed direction. The beam itself must be nearly circularly symmetric so that the radar response from a target in the axial direction will not vary with the lobing. The reflector or lens aperture is consequently usually circular.

When the antenna is small it is sometimes easier to leave the feed fixed and to produce the lobing by moving the reflector.

12. RAPID SCANNING

A lobing radar can provide range and angular information concerning a single target rapidly and accurately but these things are not always enough.

It is sometimes necessary to obtain accurate and rapid information from all regions within an agular sector. It may be necessary to watch a certain region of space almost continuously in order to be sure of picking up fast moving targets such as planes. To accomplish any of these ends we must use a *rapid scanning* radar. A rapid scanning radar antenna produces a beam which scans continuously through an angular sector. The beam may sweep in azimuth or elevation alone or it may sweep in both directions to cover a solid angle. An azimuth or elevation scan may be sinusoidal or it may occur linearly and repeat in a sawtooth fashion. Solid angle scanning may follow a spiral or flower leaf pattern or it might be a combination of two one way scans. A combination of scanning in one direction and lobing in the other is sometimes used.

Scanning antennas must, unfortunately, be constructed in obedience to the same principles which regulate ordinary antennas. The same attention to phase, amplitude, polarization and losses is necessary if comparable results are to be obtained. When scanning requirements are added to these ordinary ones new problems are created and old ones made more difficult.

An antenna in order to scan in any specified manner must act to produce a wave front which has a constant phase in a plane which is always normal to the required beam direction. This can be done in several different ways. The simplest of these, electrically, is to rotate a fixed beam antenna as a whole in the required fashion. This can be called *mechanical scanning.* Alternatively an antenna array can be scanned if it is made up of suitable elements and the relative phases of these elements can be varied appropriately. This can be called *array scanning.* Thirdly, *optical scanning* can be produced by moving either the feed or the focussing element of a suitably designed optical antenna.

12.1 *Mechanical Scanning*

Electrical complexities of other types of rapid scanners are such that it is probably not going too far to say that the required scan should be accomplished by mechanical means wherever it is at all practical. This applies to radar antenna scans which occur at a slow or medium rate. Search antennas, whether they rotate continuously through 360° or back and forth over a sector are scanners in a sense but the scan is usually slow enough to be performed by rotating the antenna structure as a whole. As the scan becomes more rapid, mechanical problems become more severe and electrically scanning antennas appear more attractive.

Mechanical ingenuity has during the war extended the range in which mechanical scanners are used. One important and eminently practical mechanical rapid scanner, the 'rocking horse' is now in common use (Fig. 36). This antenna is electrically a paraboloid of elliptical aperture illuminated by a horn feed, a combination which produces excellent electrical characteristics. The paraboloid and feed combination is made structurally strong and is pivoted to permit rotational oscillation in a horizontal plane. It is forced to oscillate by a rigid crank rod which is in turn driven by an eccentric crank on a shaft. The shaft is belt driven by an electric motor and its rotational rate is held nearly constant by a flywheel. The mechanical arrangement described so far would oscillate rotationally in an approximately sinusoidal fashion. Since every action has an equal and opposite reaction it would, however, react by producing an oscillatory torque on its mounting. Since the antenna is large and the oscillation rapid this would produce a severe and undesirable vibration. To get around this difficulty an opposite and balancing rotating moment is introduced into the mechanical system. This appears in the form of a pivoted and weighted rod which is driven from the same eccentric crank by another and almost parallel crank arm.

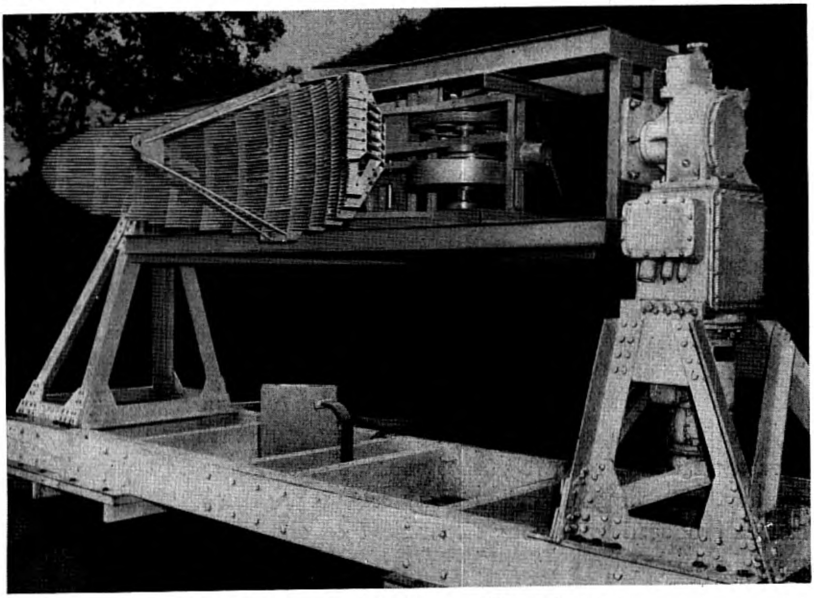


Fig. 36—Experimental Rocking Horse Antenna.

Although not theoretically perfect the rotational 'dynamic' balancing described permits the antenna to scan without serious vibration. One form of this antenna will be described in a later section.

12.2 *Array Scanning*

During our discussion of general principles in Part II, we saw that an antenna wave front can be synthesized by assembling an array of radiating elements and distributing power to it through an appropriate transmission line network. If the radiation characteristics of the array are to be as desired the electrical drive of each element must have a specified phase and amplitude. In addition each element must in itself have a satisfactory characteristic and the elements must have a proper spacial relationship to each other.

Such array antennas have been extremely useful in the 'short wave' bands where wavelengths and antenna sizes are many times larger than at most radar wavelengths but for fixed beam radar antennas they have been largely superceded by the simpler optical antennas. Where a rapidly scanning beam is desired, however, they possess certain advantages which were put to excellent use in the war. These advantages spring from the possibility of scanning the beam of an array through the introduction of rapidly varying phase changes in its transmission line distributing system.

Let us first examine certain basic conditions that must be fulfilled if an array antenna is to provide a satisfactory scan. The pattern of any array is merely the sum of the patterns of its elements taking due account of phase, amplitude and spacial relationships. If all elements are alike and are spaced equally along a straight line it is not difficult to show that a mathematical expression for the pattern can be obtained in the form of a product of a factor which gives the pattern of a single element and an array factor. The array factor is an expression for the pattern of an array of elements each of which radiates equally in all directions. Since each of the elements is fixed in direction it is only through control of the array factor that the scan can be obtained.

If we excite all points of a continuous aperture with equal phase and a smoothly tapered amplitude the aperture produces a beam with desirable characteristics at right angles to itself and no comparable radiation elsewhere. Similarly if we excite all elements of an array of identical equally spaced circularly radiating elements with equal phase and a smoothly tapered amplitude the array will produce a beam with desirable characteristics at right angles to itself. It will also produce a beam in any *other* direction for which waves from the elements can add up to produce a wave front. Such other directions will exist whenever the array spacing is greater than one wavelength.

In order to see this more clearly let us examine Fig. 37, where line XX' represents an array of elements. From each element to the line AA' is a constant distance, so AA' is obviously parallel to a wave front when the elements are excited with equal phase. If we can find a line BB' to which the distance from each element is exactly one wavelength more or less than from its immediate neighbors then it too is parallel to a wavefront, for energy reaching it from any element of the array will have the same phase except for an integral number of cycles. The same will apply to a line CC', to which the distance from each element is exactly two wave lengths more or less than from its immediate neighbors, or to any other line where this difference is any integral number of wavelengths.

Now in no radar antenna do we desire two or more beams for they will result in loss of gain and probably in target confusion. The array must therefore be designed so that for all positions of scan all beams except one will be suppressed. This will automatically occur if the array spacing is somewhat less than one wavelength. If the array spacing is greater than one wavelength these extra beams will appear in the array factor; they must therefore be suppressed by the pattern of a single element. The pattern of an element must in other words, have no significant components in any direction where an extra beam can occur.

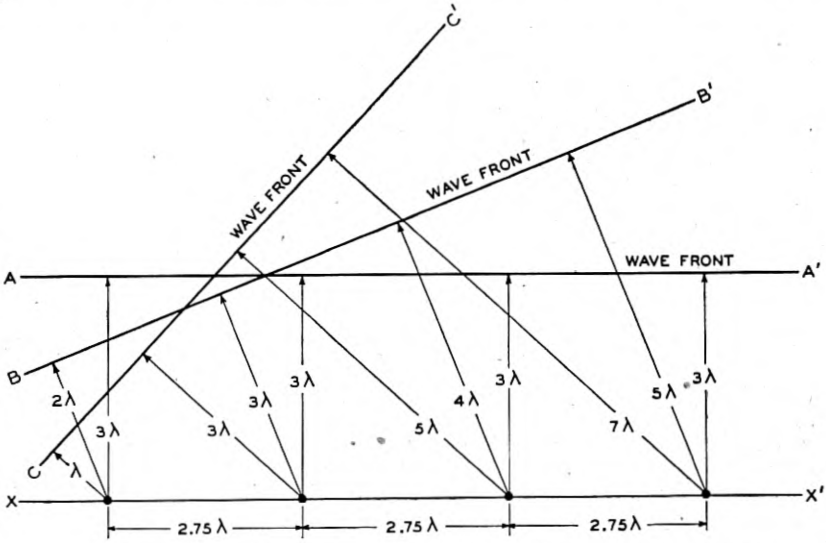


Fig. 37—Some Possible Wave Fronts of an Array of Elements Spaced 2.75 λ.

Where elements with only side fire directivity are spaced more than a wavelength apart in a scanning array it is almost impossible to obtain adequate extra lobe suppression. If these elements are spaced by the minimum amount, that is by exactly the dimensions of their apertures and all radiate in phase they may indeed just manage to produce a desirable beam. A little analysis shows however that an appreciable phase variation from element to element, even though linear, will introduce a serious extra lobe. To get around this difficulty elements with some end fire directivity must be used.

A simple end fire element, and one that has been used in practice, is the 'polyrod' (Fig. 38). A polyrod, is as its name implies, a rod of polystyrene. This rod, if inserted into the open end of a waveguide, and if properly proportioned and tapered, will radiate energy entering from the waveguide from points which are distributed continuously along its length. If the wave in the polyrod travels approximately with free space velocity it will produce a radiation maximum in the direction of its axis. The radiation pattern of the polyrod will have a shape which is characteristic of end fire arrays, narrower and flatter topped than the pattern of a side fire array which occupies the same *lateral* dimension. This elementary pattern can be fitted in well with the array factor of a scanning array.

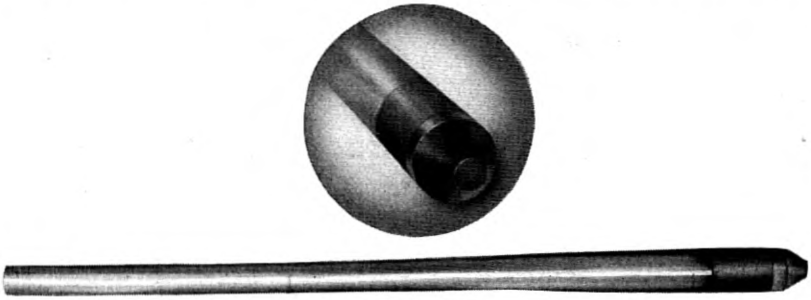


Fig. 38—A Polyrod.

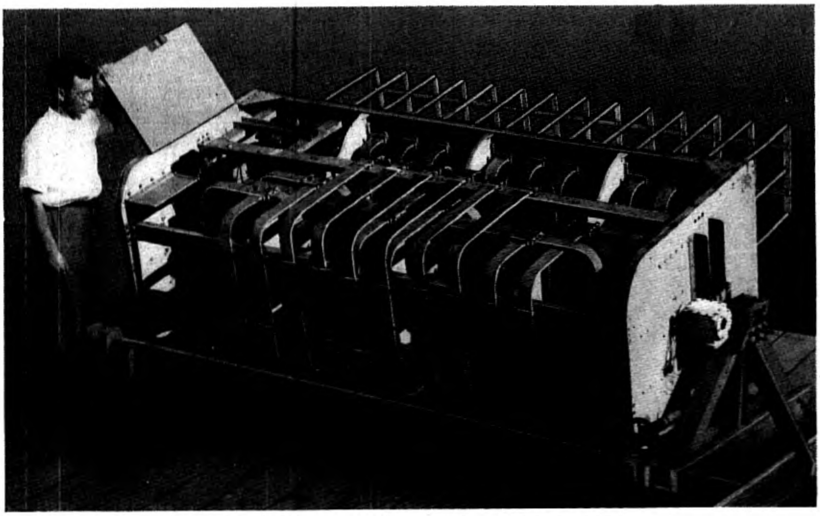


Fig. 39.—Experimental Polyrod Array.

Such a scanning array is shown in Fig. 39 and will be described in greater detail in section 14.8. Each element of this array consists of a fixed vertical array of three polyrods. This elementary array provides the required vertical pattern and has appropriate horizontal characteristics. Fourteen of these elements are arranged in a horizontal array with a spacing between neighbors of about two wavelengths. Energy is distributed among the elements with a system of branching waveguides. Thirteen rotary phase changers are inserted strategically in the distributing system. Each phase change is rotating continuously and shifts the phase linearly from 0° to 360° twice for each revolution. As the phase changers rotate the array produces a beam which sweeps repeatedly linearly and continuously across the scanning sector.

When elements of a scanning array are spaced considerably less than one wavelength it is a very simple matter to obtain a suitable elementary pattern, for the array factor itself has only a single beam. This advantage is offset by the greater number of elements and the consequent greater complexity of distributing and phase shifting equipment. In one useful type of scanning antenna however distributing and phase shifting is accomplished in a particularly simple manner. Here the distributing system is merely a waveguide which can transmit only the dominant mode. The wide dimension of the guide is varied to produce the phase shifts required for scanning. The elements are dipoles. The center conductor of each dipole protrudes just enough into the guide to pick up the required amount of energy.

It is evident from the above discussion that such a waveguide fed dipole array will produce a single beam in the normal direction only if the dipoles are all fed in phase and are spaced less than a wavelength. It is therefore not satisfactory to obtain constant phase excitation by tapping the dipoles into the guide at successive guide wavelengths for these are greater than free space wavelengths. Consequently the dipoles are tapped in at successive half wavelengths in the guide and reversed successively in polarity to compensate for the successive phase reversals due to their spacing.

This type of array provides a line source which can be scanned by moving the guide walls. In order to leave these mechanically free suitable wave trapping slots are provided along the length of the array.

A practical antenna of this type will be described in Sec. 16.3.

12.3 *Optical Scanning*

With a camera or telescope all parts of an angular sector or field are viewed simultaneously. We would like to do the same thing by radar means, but since this so far appears impossible we do the next best thing by looking at the parts of the field in rapid succession. Nevertheless certain pcints of similarity appear. These points are emphasized by a survey of the fixed beam antenna field for there we find optical instruments in abundance, parabolic reflectors and even lenses.

It is not a very big step to proceed from an examination of optical systems to the suggestion that a scanning antenna can be provided by moving a feed over the focal plane of a reflector. Nevertheless experience shows that this will not be especially profitable unless done with due caution. The first effect of moving the feed away from the focus in the focal plane of a paraboloid is indeed a beam shift but before this process has gone far a third order curvature of the phase front is produced and is accompanied by a serious deterioration in the pattern and reduction in gain. This difficulty or aberration is well known in classical optical theory and is called coma. Coma is typified by patterns such as the one shown in Fig. 16. It is the first obstacle in the path of the engineer who wishes to design a good moving feed scanning antenna.

Coma is not an insuperable obstacle however. Its removal can be accomplished by the application of a very simple geometrical principle. This principle can be stated as follows: "The condition for the absence of coma is that each part of the focussing reflector or lens should be located on a circle with center at the focus."

This condition can be regarded as a statement of the spacial relationship required between the feed and all parts of the focussing element. It is a condition which insures that the phase front will remain nearly linear when the feed is moved in the focal plane. It can be applied approximately whether the focussing element is a reflector or a lens and to optical systems which scan in both directions as well as those which scan in one direction.

Coma is usually the most serious aberration to be reckoned with in a scanning optical system, but it is by no means the only one. Any defect in the phase and amplitude characteristic which arises when the feed is moved can cause trouble and must be eliminated or reduced until it is tolerable. Another defect in phase which arises is 'defocussing'. Defocussing is a square law curvature of phase and arises when the feed is placed at an improper distance from the reflector or lens. Its effect may be as shown in Fig. 14. It can in principle always be corrected by moving the feed in a correctly chosen arc, but this is not always consistent with other requirements on the system. In addition to troubles in phase an improper amplitude across the aperture of the antenna will arise when the feed is translated unless proper rotation accompanies this motion.

To combat the imperfections in an optical scanning system we can choose over-all dimensions in such a way that they will be lessened. Thus it is generally true that an increase in focal length or a decrease in aperture will increase the scanning capabilities of an optical system. This alone is usually not enough, however, we must also employ the degrees of freedom available to us in the designing of the focussing element and the feed motion to improve the performance. If the degrees of freedom are not enough we must, if we insist on an optical solution introduce more. This could in principle result in microwave lenses similar to the four and five element glass lenses found in good cameras, but such complication has not as yet been necessary in the radar antenna art.

Since military release has not been obtained as this article goes to press we must omit any detailed discussion of optically scanning radar antenna techniques.

PART III

MILITARY RADAR ANTENNAS DEVELOPED BY THE BELL LABORATORIES

13. General

In the final part of this paper we will describe in a brief fashion the end products of radar antenna technology, manufactured radar antennas. Without these final practical exhibits the foregoing discussion of principles and methods might appear academic. By including them we hope to illustrate in a concrete fashion the rather general discussion of Parts I and II.

The list of manufactured antennas will be limited in several ways. Severe but obviously essential are the limitations of military security. In addition we will restrict the list to antennas developed by the Bell Laboratories. In cases where invention or fundamental research was accomplished elsewhere due credit will be given. Finally the list will include only antennas manufactured by contract. This last limitation excludes many experimental antennas, some initiated by the Laboratories and some by the armed forces.

It is worthwhile to begin with an account of the processes by which these antennas were brought into production. The initiating force was of course military necessity. The initial human steps were taken sometimes by members of the armed forces who had definite needs in mind and sometimes by members of the Laboratories who had solutions to what they believed to be military needs.

With a definite job in mind conferences between military and Laboratories personnel were necessary. Some of these dealt with legal or financial matters, others were principally technical. In the technical conferences it was necessary at an early date to bring military requirements and technical possibilities in line.

As a result of the conferences a program of research and development was often undertaken by the Laboratories. An initial contract was signed which called for the delivery of technical information, and sometimes for manufacturing drawings and one or more completed models. Usually the antenna was designed and manufactured as part of a complete radar system, sometimes the contract called for an antenna alone.

After preliminary work had been undertaken the status of the job was reviewed from time to time. If preliminary results and current military requirements warranted a manufacturing contract was eventually drawn up and signed by Western Electric and the contracting government agency. This contract called for delivery of manufactured radars or antennas according to a predetermined schedule.

Research and development groups of the Laboratories cooperated in war as in peace to solve technical problems and accomplish technical tasks. Under the pressure of war the two functions often overlapped and seemed to merge, yet the basic differences usually remained.

Members of the Research Department, working in New York and at the Deal and Holmdel Radio Laboratories in New Jersey were concerned chiefly with electrical design. It was their duty to understand fully electrical principles and to invent and develop improved methods of meeting military requirements. During the war it was usually their responsibility to prescribe on the basis of theory and experiment the electrical dimensions of each new radar antenna.

A new and difficult requirement presented to the Research Department was sometimes the cause of an almost personal competition between alternative schemes for meeting it. Some of these schemes were soon eliminated by their own weight, others were carried side by side far along the road to production. Even those that lost one race might reappear in another as a natural winner.

In the Development Groups working in New York and in the greatly expanded Whippany Radio Laboratory activity was directed towards coordination of all radar components, towards the establishment of a sound, well integrated mechanical and electrical design for each component and towards the tremendous task of preparing all information necessary for manufacture. It was the job of these groups also to help the manufacturer past the unavoidable snarls and bottlenecks which appeared in the first stages of production. In addition development personnel frequently tested early production models, sometimes in cooperation with the armed forces.

As we have intimated, research and development were indistinguishable at times during the war. Members of the research department often found themselves in factories and sometimes in aircraft and warships. Development personnel faced and solved research problems, and worked closely with research groups.

For several years when pressure was high the effort was intense; at times feverish. Judging by military results it was highly effective. Some of the material results of this effort are described in the following pages.

14. Naval Shipborne Radar Antennas

14.1 *The SE Antenna*[10]

Very early in the war, the Navy requested the design of a simple search radar system for small vessels, to be manufactured as quickly as possible in order to fill the gap between design and production of the more complex search systems then in process of development. The proposed system was to be small and simple, to permit its use on vessels which otherwise would be unable to carry radar equipment because of size or power supply capability. This class of vessel included PT boats and landing craft.

The antenna designed for the SE system is housed as shown in Fig. 40. It was adapted for mounting on the top or side of a small ship's mast, and is rotated in azimuth by a mechanical drive, hand operated. The paraboloid reflector is 42 inches wide, 20 inches high, and is illuminated by a circular aperture 2.9 inches in diameter. In the interests of simplicity, the polarization of the radiated beam was permitted to vary with rotation of the antenna.

The SE antenna was operated at 9.8 cm, and fed by $1\frac{1}{2}$ x 3 rectangular waveguide. At the antenna base, a taper section converted from the rectangular waveguide to 3″ round guide, through a rotating joint directly to the feed opening.

Characteristics of the SE antenna are given below:

Wavelength	9.7 to 10.3 cm	
Reflector	42″ W x 20″ H	
Gain	25 db	
Horizontal Beam Width	6°	
Vertical Beam Width	12°, varying somewhat with polarization	
Standing Wave	9.7to10.0 cm	4.0 db
	10.0 to 10.3 cm	6.0 db

14.2 *The SL Radar Antenna*[11]

The SL radar is a simple marine search radar developed by Bell Telephone Laboratories for the Bureau of Ships. During the war, over 1000 of these radars were produced by the Western Electric Company and installed on Navy vessels of various categories. The principal field for installation was destroyer escort craft ("DE"s). Figure 41 shows an SL antenna installation aboard a DE. The antenna is covered, for wind and

[10] Written by R. J. Phillips.

[11] Written by H. T. Budenbom.

weather protection, in a housing which can transmit 10 cm radiation. Visible also is the waveguide run down the mast to the r.f. unit.

The SL radar provides a simple non-stabilized PPI (Plan Position Indicator) display. The antenna is driven by a synchronous motor at 18 rpm. Horizontal polarization is used to minimize sea clutter. The radiating structure, shown in Figure 42, consists of a 20″ sector of a 42″ paraboloid. The resulting larger beam width in the vertical plane is provided in order to improve the stability of the pattern under conditions of ship roll. Figure 43 illustrates the path of the transmitted wave from the SL r.f. unit to the antenna. It also illustrates the manner in which horizontally polarized radiation is obtained. The diagram shows the position of

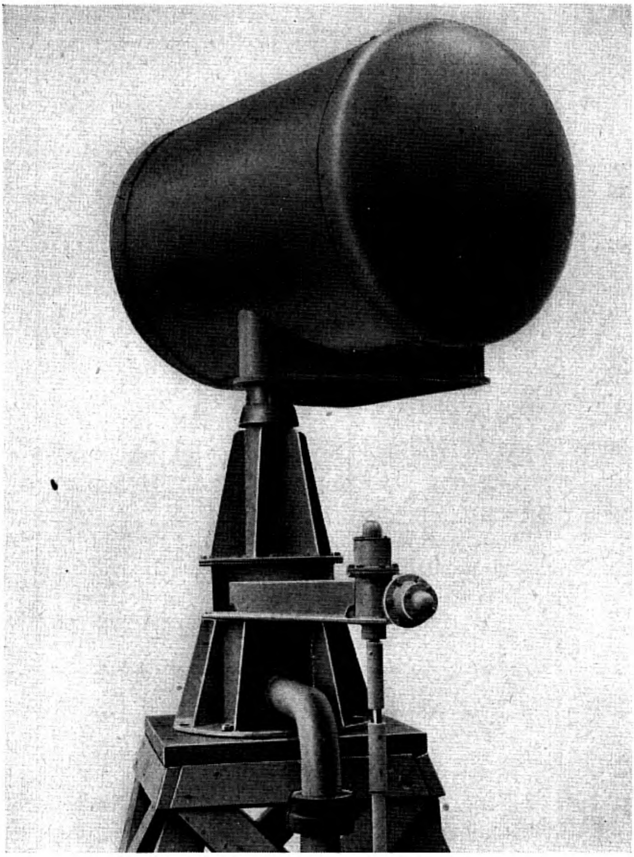


Fig. 40—SE Antenna.

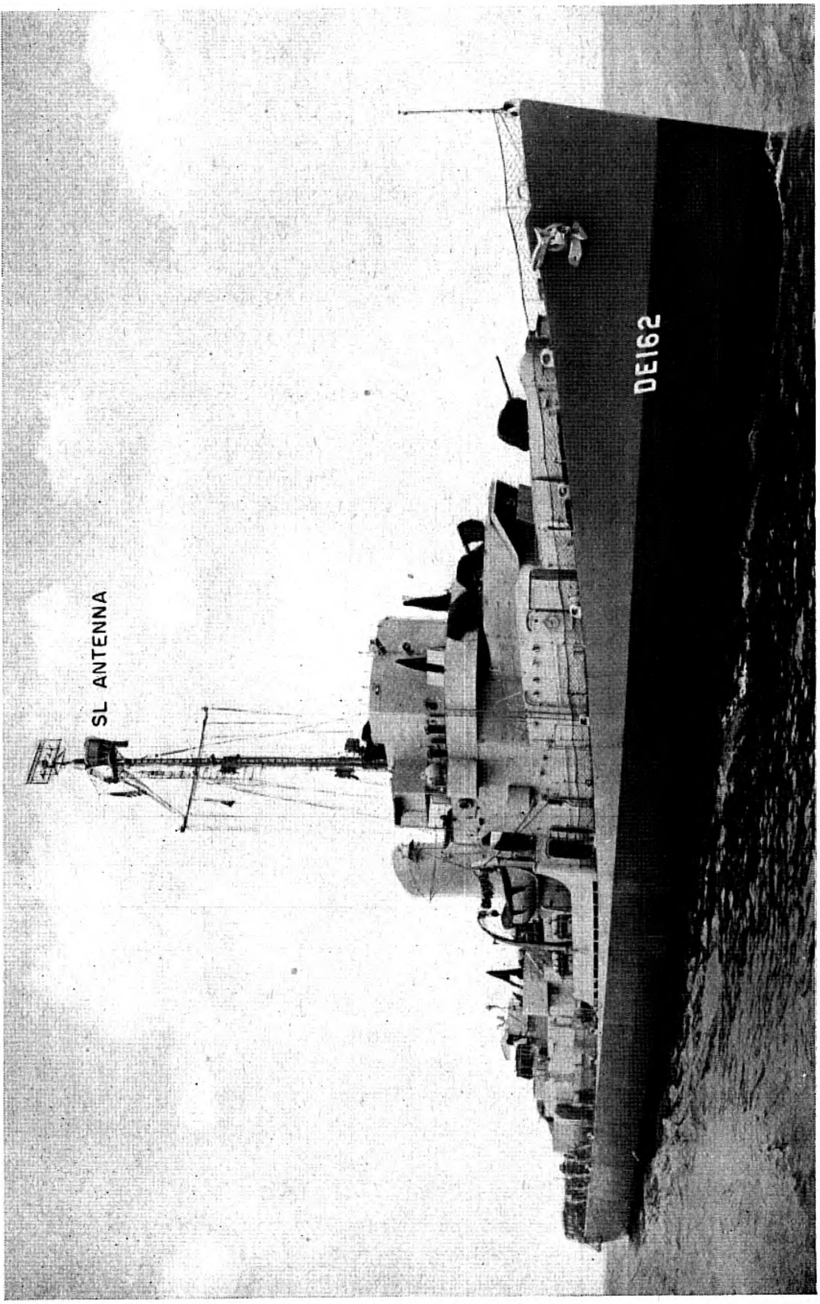


Fig. 41—SL Antenna Aboard DE.

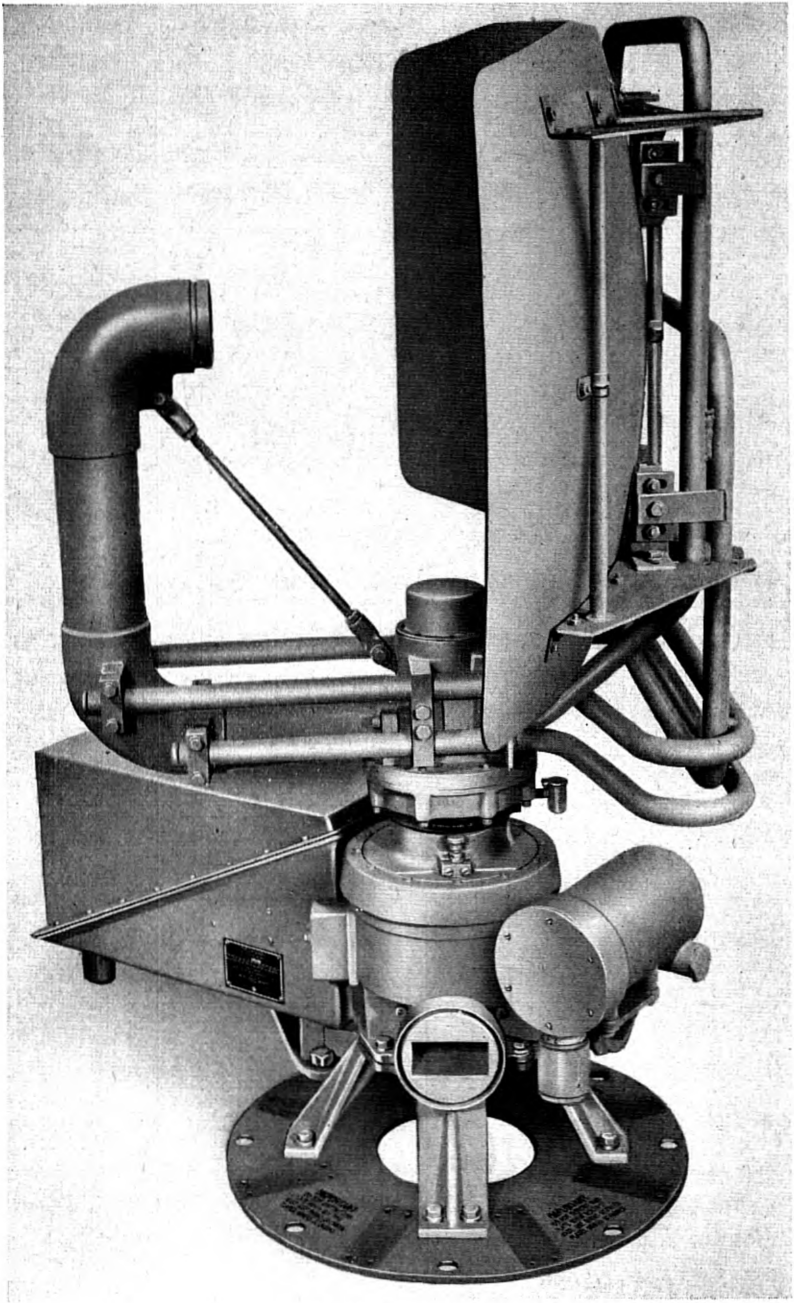


Fig. 42—SL Antenna.

the electric force vector in traversing the waveguide run. The path from the r.f. unit is in rectangular guide ($TE_{1,0}$ mode) through the right angle bend, to the base of the rotary joint. A transducer which forms the base portion of the joint converts to the TM_{01} mode in circular pipe. For this mode, the electric field has radial symmetry, much as though the waveguide were a coaxial line of vanishingly small inner conductor diameter.

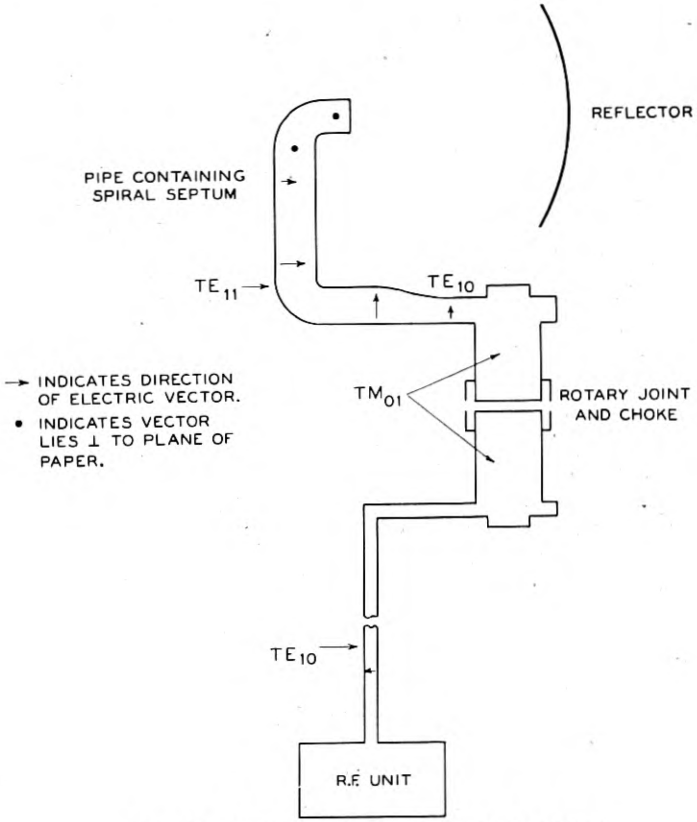

REFLECTOR
PIPE CONTAINING SPIRAL SEPTUM
TE_{11}
TE_{10}
→ INDICATES DIRECTION OF ELECTRIC VECTOR.
• INDICATES VECTOR LIES ⊥ TO PLANE OF PAPER.
TM_{01}
ROTARY JOINT AND CHOKE
TE_{10}
R.F. UNIT

Fig. 43—SL Radar Antenna—Wave Guide Path.

The energy passes the rotary joint in this mode; choke labyrinths are provided at the joint to minimize radio frequency leakage. The energy then flows through another transducer, from TM_{01} mode back to TE_{10} mode. The lower horizontal portion of the feed pipe immediately tapers to round guide, the mode being now TE_{11}. Next the energy transverses a 90° elbow, which is a standard 90° pipe casting, and enters the vertical section immediately below the feed aperture. The E vector is in the plane of the paper at this point. However, the ensuing vertical section is fitted with a spiral septum. This gradually rotates the plane of polarization until at the top of this pipe the E vector is perpendicular to the plane of the paper. Thus, after transversing another 90° pipe bend, the energy emerges horizontally polarized, to feed the main reflector.

Specific electrical characteristics of the SL antenna are:

Polarization—Horizontal
Horizontal Half Power Beamwidth—6°
Vertical Half Power Beamwidth—12°
Gain—about 22 db.

14.3 *The SJ Submarine Radar Antenna*

It had long been expected that one of the early offensive weapons of the war would be the submarine. It was therefore natural that early in the history of radar the need for practical submarine radars was felt. The principal components of this need were twofold, to provide warning of approaching enemies and to obtain torpedo fire control data. The SJ Submarine Radar was the first to be designed principally for the torpedo fire control function.

Work on the SJ system was under way considerably before Pearl Harbor. When this work was initiated the advantages of lobing fire control systems were clearly recognized, but no lobing antennas appropriate for submarine use had been developed. Requirements on such an antenna were obviously severe, for in addition to fulfilling fairly stringent electrical conditions, it would have to withstand very large forces due to water resistance and pressure.

The difficulties evident at the outset of the work were overcome by an ingenious adaptation of the simple waveguide feed. It was recognized that a shift of the feed in the focal plane of a reflector would cause a beam shift. Why not, then, use two waveguide feeds side by side to produce the two nearly coincident beams required in a lobing antenna? When this was tried it was found to work as expected.

It remained to devise a means of switching from one waveguide feed to the other with the desired rapidity. This in itself was no simple problem, but was solved by applying principles learned through work on waveguide filters. The switch at first employed was essentially a branching filter at the junction of the single antenna feed line and the line to each feed aperture. Both branches of this filter were carefully tuned to the same frequency, that of the radar. The switching was performed by the insertion of small rapidly rotating pins successively into the resonant cavities of the two filters (Fig. 44). Presence of the pins in one of the filters detuned it and therefore prevented power from flowing through it. Rotation of the pins accordingly produced switching as desired.

In a later modification of this switch the same general principles were used but resonant irises rather than resonant cavities were employed.

The SJ Submarine Radar was in use at a comparatively early date in the war and saw much service with the Pacific submarine fleet. Despite some early doubts, submarine commanders were soon convinced of its powers. It is believed that in the majority of cases it replaced the periscope as the principle fire control instrument. In addition it served as a valuable and unprecedented aid to navigation.

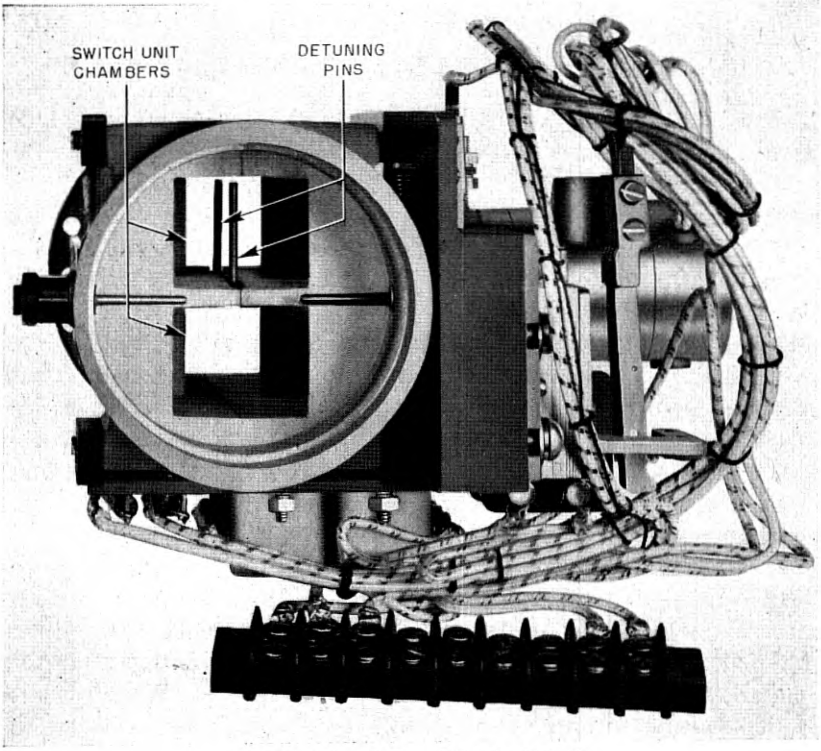


Fig. 44—The SJ Tuned Cavity Switch.

It is interesting and relevant to quote from two letters to Laboratories engineers concerning the SJ. One dated October 3, 1943, from the radar officer of a submarine stated that there were twenty "setting sun" flags painted on the conning tower and asked the engineer to "let your mind dwell on the fact that you helped to put more than 50% of those flags there".

The commander of another submarine wrote in a similar vein, "You can rest assured that we don't regard your gear as a bushy-brain space taker, but a very essential part of our armament".

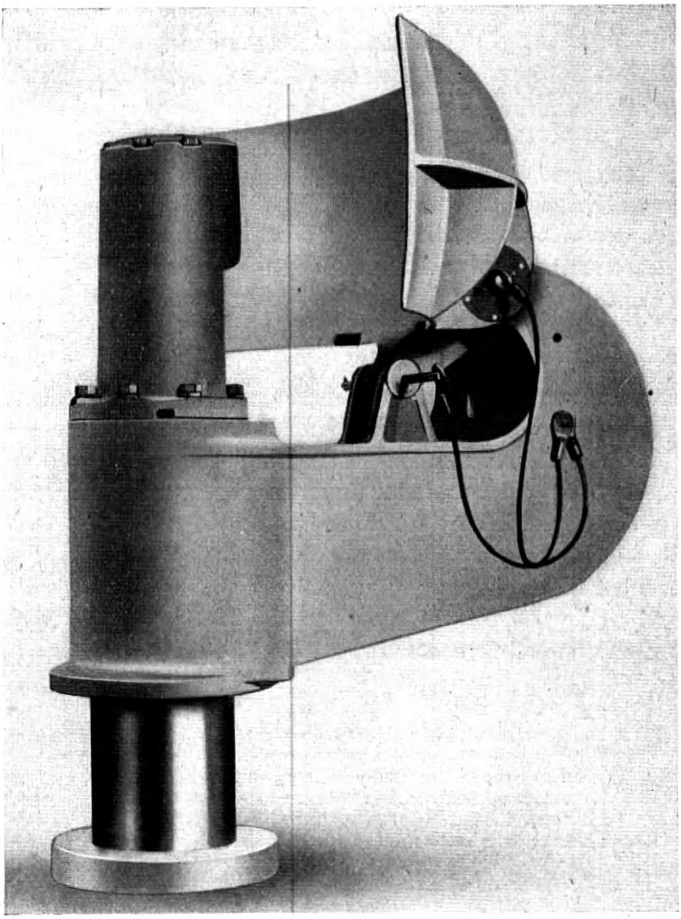


Fig. 45—The SJ Submarine Radar Antenna.

Figure 45 is a photograph of an SJ antenna. Its principal electrical characteristics are as follows:

Gain > 19 db
Horizontal Half Power Beamwidth 8°
Vertical Half Power Beamwidth 18°
Vertical Beam Character—Some upward radiation
Lobe Switching Beam Separation—approximately 5°
Gain reduction at beam cross-over < 1 db
Polarization—Horizontal

14.4 *The Modified SJ/Mark 27 Radar Antenna*

The SJ antenna described above performed a remarkable and timely fire control job as a lobing antenna but was found to be unsatisfactory when rotated continuously to produce a Plan Position Indicator (PPI) presentation. In the PPI method of presentation range and angle are presented as radius and angle on the oscilloscope screen. Consequently a realistic map of the strategic situation is produced. This map is easily spoiled by false signals due to large minor lobes of the antenna.

Since it had been established that the PPI picture was valuable for navigation and warning as well as for target selection it was decided to modify the antenna in a way that would reduce these undesirably high minor lobes. These were evidently due principally to the shadowing effect of the massively built double primary feed. Accordingly a new reflector was designed which in combination with a slightly modified feed provided a much improved pattern.

The new reflector was different in configuration principally in that it was a partially offset section of a paraboloid. The reflector surface was also markedly different in character since it was built as a grating rather than a solid surface. This reduced water drag on the antenna. In addition the grating was less visible at a distance, an advantage that is obviously appreciable when the antenna is the only object above the water.

This modified antenna was used not only on submarines as part of the SJ-1 radar but also on surface vessels as the Mark 27 Radar Antenna. Figure 46 shows one of these antennas. Its electrical characteristics are as follows:

Gain > 20 db
Horizontal Half Power Beamwidth = 8°
Vertical Half Power Beamwidth = 17°
Vertical Beam Character—Some upward radiation
Lobe Switching Beam Separation—approximately 5°
Gain reduction at beam cross-over < .1 db
Polarization—Horizontal

14.5 *The SH and Mark 16 Antenna*[12]

The antennas designed for the SH and Mark 16 Radar Equipments are practically identical. The SH system was a shipborne combined fire control and search system, and the Mark 16 its land based counterpart was used by the Marine Corps for directing shore batteries.

These systems operated at 9.8 cm. The requirement that the system, operate as a fire control as well as a search system imposed some rather stringent mechanical requirements on the antenna. For search purposes, the antenna was rotated at 180 rpm, and indications were presented on a plan position indicator. For fire control data, slow, accurately controlled motion was required. Bearing accuracy is attained by lobe switching in

[12] Written by R. J. Philipps.

much the same manner as in the SJ and SJ-1 antennas previously described.

The antenna is illustrated in Fig. 47. With the SH system, the unit is mast mounted; for the Mark 16, the unit is mounted atop a 50 foot steel tower which can be erected in a few hours with a minimum of personnel. The electrical characteristics are as follows:

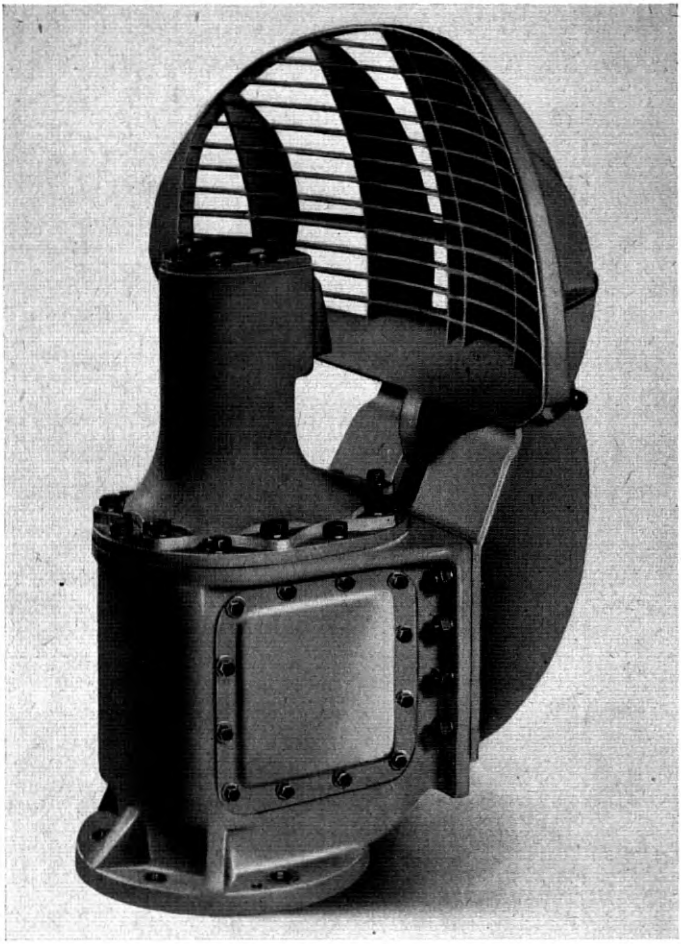


Fig. 46—The SJ-1/Mark 27 Submarine Radar Antenna.

Gain—21. db
Reflector Dimensions 30″ W x 20″ H
Horizontal beam width—7.5°
Vertical beam width—12°
Lobe separation—5° approximately
Loss in gain at lobe crossover—1 db approximately
Scan—(1) 360°, at 180 rpm for PPI operation
(2) 360°, at approximately 1 rpm for accurate azimuth readings, with lobe switching

SH systems were most successfully used in invasion operations in the Aleutians. They were installed on landing craft, and the use of the high speed scan enabled the craft to check constantly their relative positions in the dense fogs encountered during the landing operations.

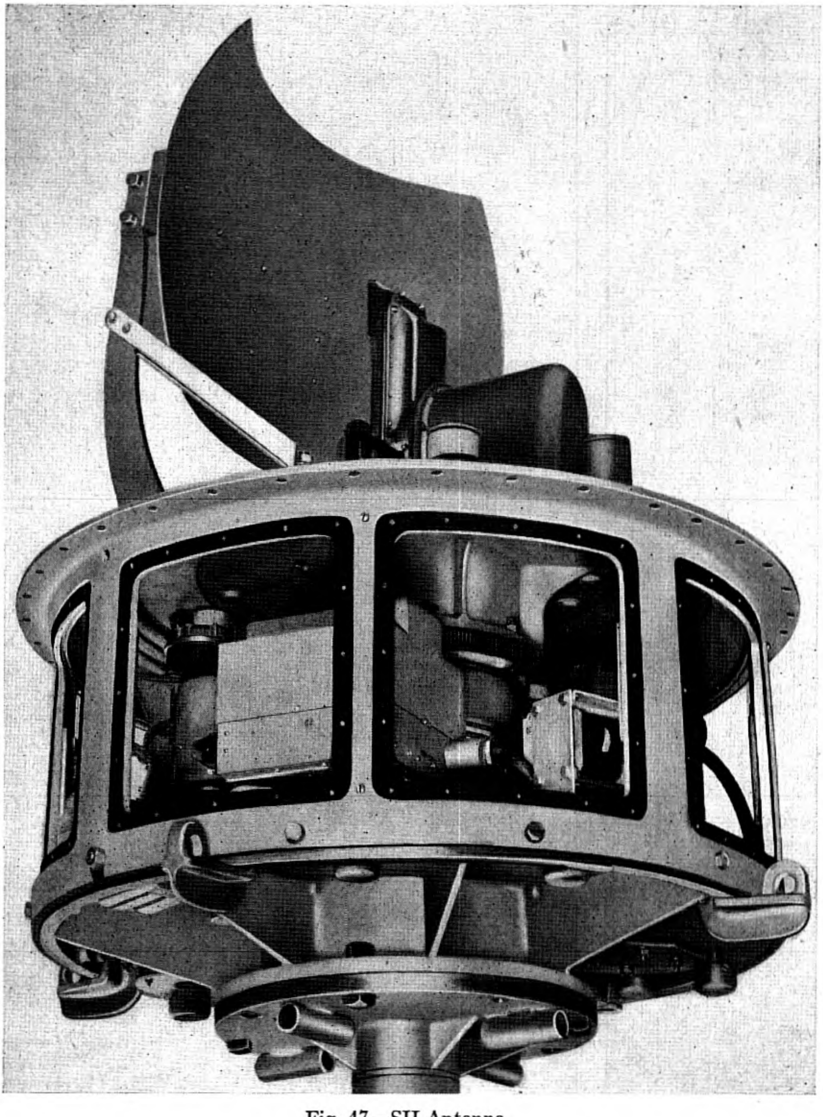


Fig. 47—SH Antenna.

14.6 *Antennas for Early Fire Control Radars*[13]

The first radars to be produced in quantity for fire control on naval vessels were the Mark 1, Mark 3 and Mark 4 (originally designated FA, FC and FD). These radars were used to obtain the position of the target with sufficient accuracy to permit computation of the firing data required by the guns. The first two (Mark 1 and Mark 3) were used against enemy surface targets while the Mark 4 Radar was a dual purpose system for use against both surface and aircraft targets. These radars were described in detail in an earlier issue.[14] However, photographs of the antennas and pertinent information on the antenna characteristics are repeated herein for the sake of completeness. (See Table B and Figures 48, 49 and 50.)

TABLE B

	Radar			
	Mark 1	Mark 3		Mark 4
Dimensions	6′ x 6′	3′ x 12′	6′ x 6′	6′ x 7′
Operating Frequency	500 or 700 MC	680–720 MC		680–720 MC
Beam Width in Degrees (Between half power points one way.)				
Azimuth	12°	6°	12°	12°
Elevation	14°	30°	14°	12°
Antenna Gain	22 db	22 db	22 db	22.5 db.
Beam Shift in Degrees				
Azimuth	0°	±1.5°	±3°	±3°
Elevation	0°	0°	0°	±3°

An antenna quite similar to the Mark 3, 6 ft. x 6 ft. antenna, was also used on Radio Set SCR-296 for the Army. This equipment was similar to the Mark 3 in operating characteristics but was designed mechanically for fixed installations at shore points for the direction of coast artillery gun fire. For these installations the antenna was mounted on an amplidyne controlled turntable located on a high steel tower. The entire antenna and turntable was housed within a cylindrical wooden structure resembling a water tower. Equipments of this type were used as a part of the coastal defense system of the United States, Hawaiian Islands, Aleutian Islands and Panama.

[13] Written by W. H. C. Higgins.

[14] "Early Fire Control Radars for Naval Vessels," W. C. Tinus and W. H. C. Higgins, B. S. T. J.

14.7 *A Shipborne Anti-Aircraft Fire Control Antenna*[15]

A Shipborne Anti-Aircraft Fire Control Antenna is shown in Fig. 51. This antenna consists of two main horizontal cylindrical parabolas in each of which two groups of four half-wave dipoles are mounted with their axes in a horizontal line at the focus of the parabolic reflectors. The four groups of dipoles are connected by coaxial lines on the back of the antenna to a lobe

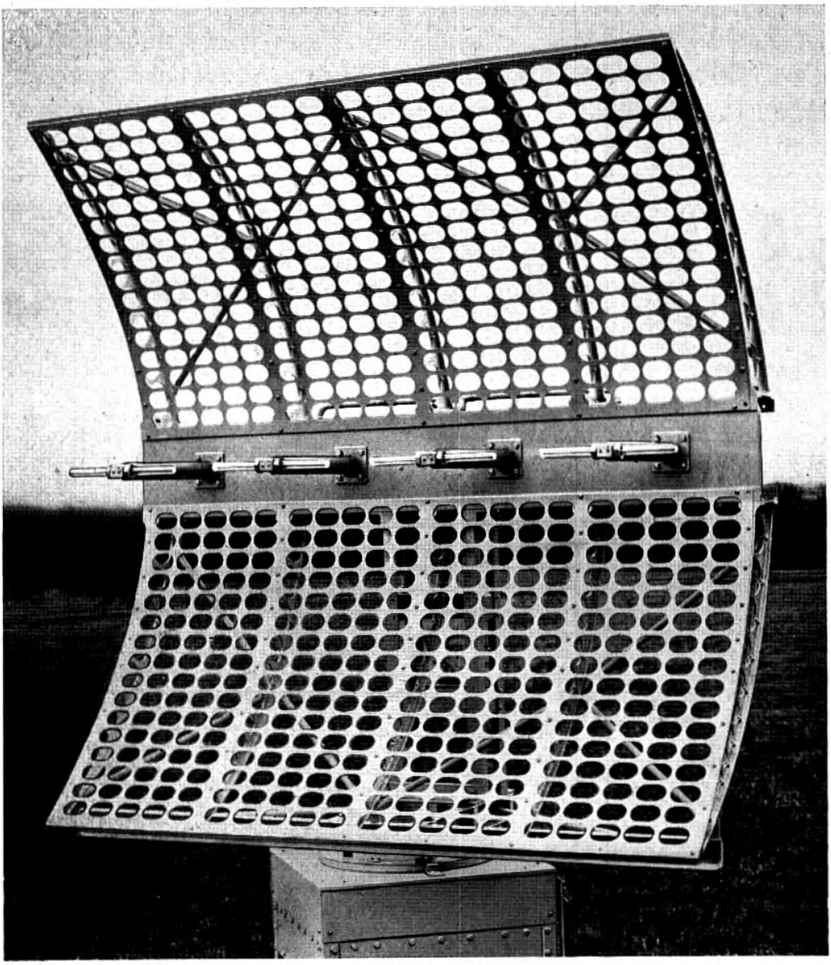


Fig. 48—Mark 1 Antenna.

[15] Written by C. A. Warren.

switcher, which is a motor driven capacitor that has a single rotor plate and four stator plates, one for each group of dipoles. The phase shift introduced into the four feed lines by the lobe switching mechanism causes the antenna beam to be "lobed" or successively shifted to the right, up, left and down as the rotor of the capacitor turns through 360 degrees.

Mounted centrally on the front of the antenna at the junction of the two parabolic antennas is a smaller auxiliary antenna consisting of two dipole elements and a parabolic reflector, the purpose of which is to reduce the minor lobes that are present in the main antenna beam. The auxiliary antenna beam is not lobe switched and is sufficiently broad in both the horizontal and vertical planes to overlap both the main antenna beam and the first minor lobes. The auxiliary antenna feed is so designed that its field is in phase with the field of the main beam of the main antenna. This causes the feed of the auxiliary antenna to "add" to the field of the main antenna in the region of its main beam, but to subtract from the field in the region of its first minor lobes. This occurs because the phase of the first minor lobes differs by 180 degrees from that of the main beam. As a result, the field of the main beam is increased and the first minor lobes are greatly

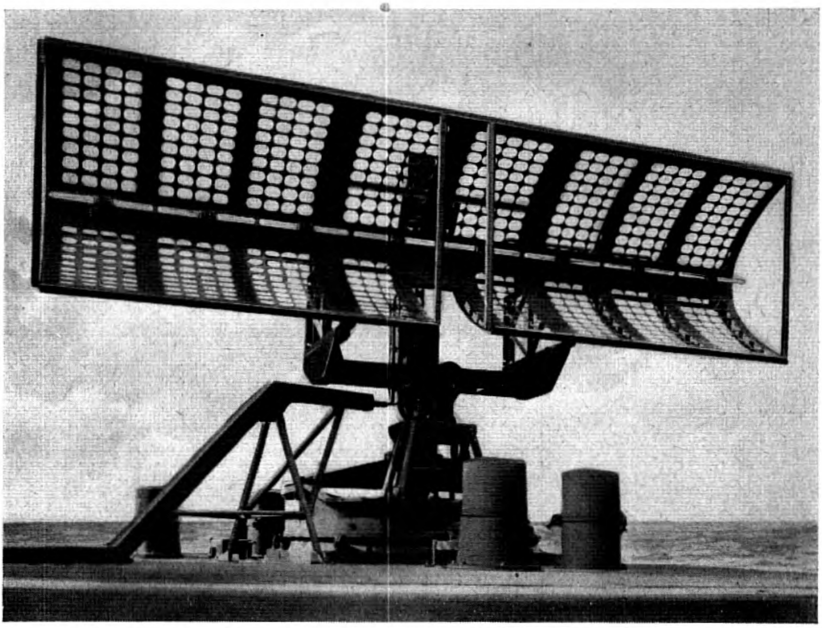


Fig. 49—Mark 3 Radar Antenna on Battleship New Jersey.

reduced. By reducing these minor lobes to a low value, the region around the main beam is free of lobes, thus greatly reducing the possibility of false tracking due to "cross overs" between the main beam and the minor lobes.

14.8 *The Polyrod Fire Control Antenna*

The Polyrod Fire Control antenna is an array scanner employing essentially the same principles as those used in the multiple unit steerable antenna system (MUSA) developed before the war for short-wave transatlantic telephony. Some of these principles have been discussed in Sec. 12.2. That they could be applied with such success in the microwave region was due to a firm grounding in waveguide techniques, to the invention of the polyrod antenna and the rotary phase changer, and especially to excellent technical work on the part of research, development and production personnel. It is perhaps one of the most remarkable achievements of wartime

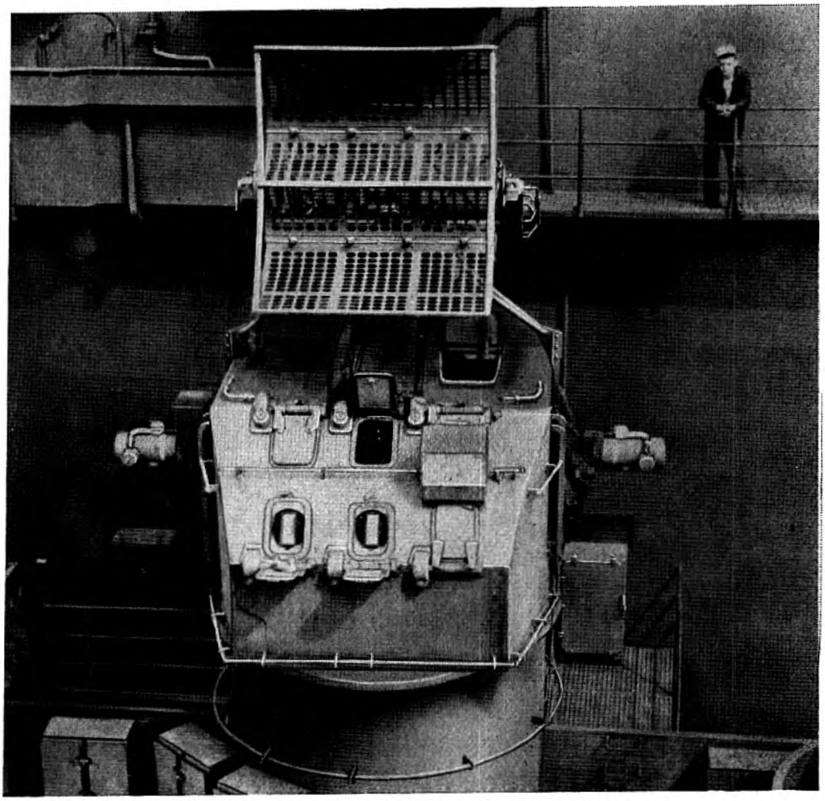


Fig. 50—Mark 4 Radar Antenna on Battleship Tennessee.

radar that the polyrod antenna emerged to fill the rapid scanning need a early and as well developed as it did.

The Polyrod Fire Control antenna is a horizontal array of fourteen identical fixed elements, each element being a vertical array of three polyrods. Energy is distributed to the elements through a waveguide manifold. The phase of each element is controlled and changed to produce the desired scan by means of thirteen rotary phase changers. These phase shifters are geared together and driven synchronously. Figure 52 is a schematic diagram of the waveguide and phase changer circuits.

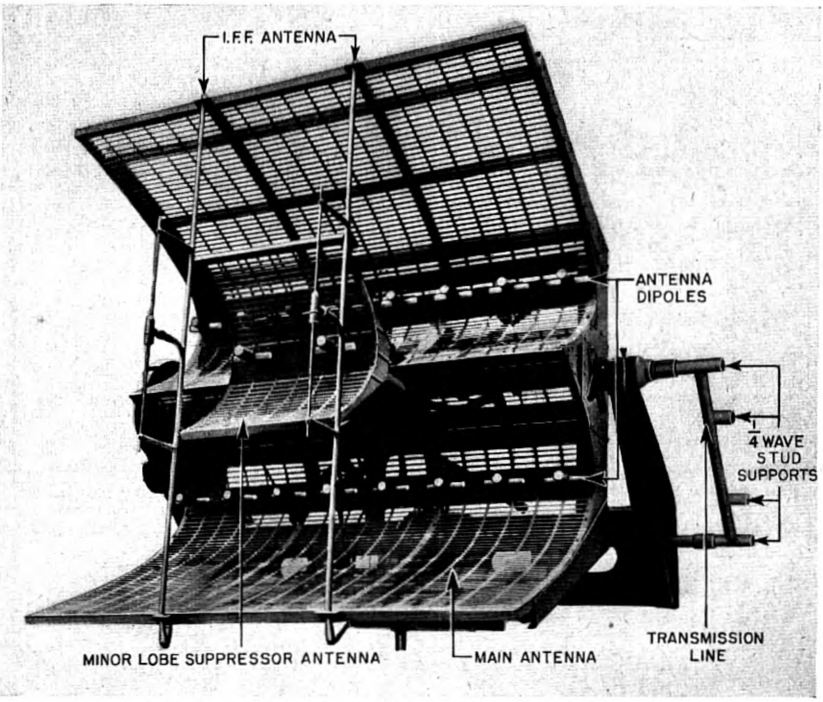


Fig. 51.—Shipborne Anti-Aircraft Fire Control Antenna

Figure 39 shows an experimental polyrod antenna under test at Holmdel. Figure 53 is another view of the Polyrod antenna.

14.9 *The Rocking Horse Fire Control Antenna*

It was long recognized that an important direction of Radar development lay towards shorter waves. This is particularly true for fire control antennas where narrow, easily controlled beams rather than great ranges are needed. The Polyrod antenna had pretty thoroughly demonstrated the value of rapid scanning, yet the problem of producing a rapid scanning higher frequency antenna of nearly equal dimensions was a new and different one.

Several possible solutions to this problem were known. The array technique applied so effectively to the polyrod antenna could have been applied here also, but only at the expense of many more elements and greater complexity.

After much preliminary work it was finally concluded that a mechanically scanning antenna, the "rocking horse," provided the best solution to the higher frequency scanning problem. This solution is practical and relatively simple.

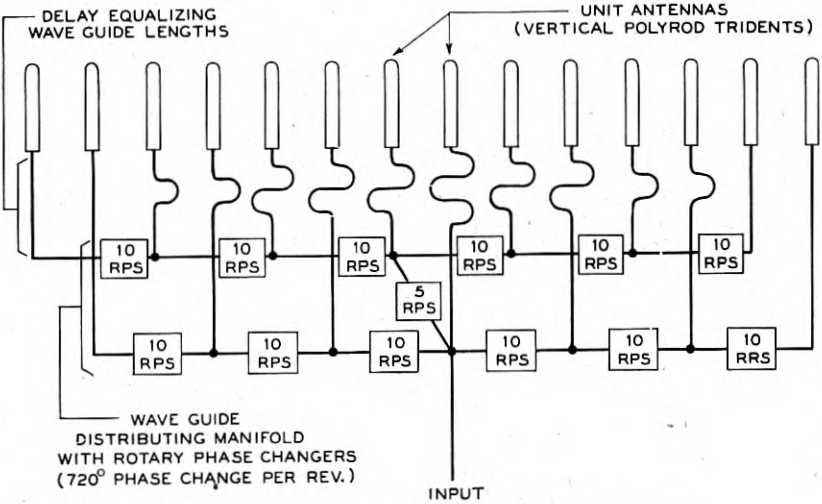


Fig. 52.—Schematic Diagram of Polyrod Fire Control Antenna.

The operation of the rocking horse is described in Sec. 12.1. It is essentially a carefully designed and firmly built paraboloidal antenna which oscillates rapidly through the scanning sector. Its oscillation is dynamically balanced to eliminate undesirable vibration.

Figure 54 is a photograph of a production model of the rocking horse antenna.

14.10 *The Mark 19 Radar Antenna*[16]

In Anti-aircraft Fire Control Radar Systems for Heavy Machine Guns it is necessary to employ a highly directive antenna and to obtain continuous rapid comparison of the received signals on a number of beam positions

[16] Sections 14.10, 14.11 and 14.12 were written by F. E. Nimmcke.

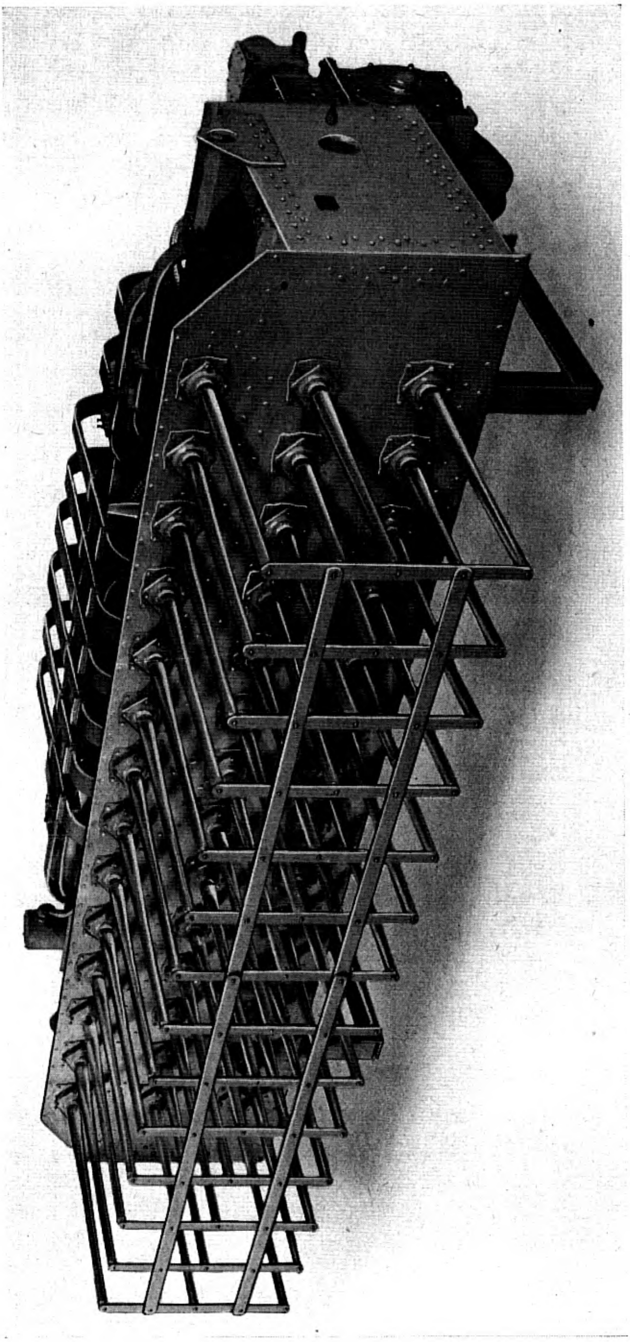


Fig. 53.—Polyrod Fire Control Antenna.

as discussed in Section 11.2. Such an antenna is also required to obtain the high angular precision for anti-aircraft fire control. These requirements are achieved by the use of a conical scanning system. The beam from the antenna describes a narrow cone and the deviation of the axis of the cone from the line of sight to the target can be determined and measured by the phase difference between the amplitude modulated received signal and the frequency of the reference generator associated with the antenna. This information is presented to the pointer-trainer at the director in the form of a wandering dot on an oscilloscope.

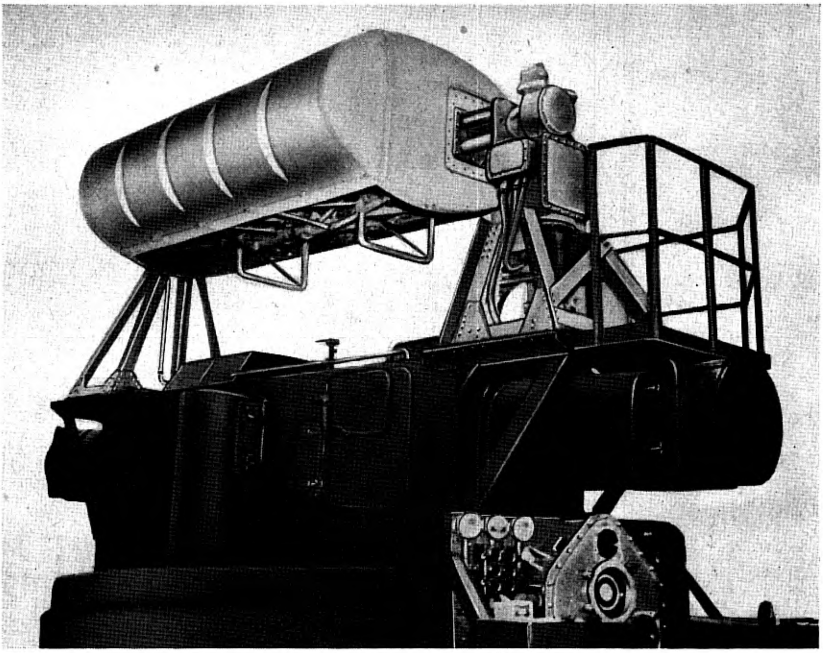


Fig. 54.—Rocking Horse Fire Control Antenna.

The antennas described in sections 14.10, 14.11 and 14.12 were all designed by the Bell Laboratories as anti-aircraft fire control radar systems, particularly for directing heavy machine guns. They were designed for use on all types of Naval surface warships.

In Radar Equipment Mark 19, the first system to be associated with the control of 1.1 inch and 40 mm anti-aircraft machine guns, the antenna was designed for operation in the 10 cm region. This antenna consisted of a spinning half dipole with a coaxial transmission line feed. The antenna was driven by 115-volt, 60 cycle, single phase motor to which was coupled a two-phase reference voltage generator. The motor rotated at approximately 1800 rpm which resulted in a scanning rate of 30 cycles per second. This antenna was used with a 24-inch spun steel parabolic reflector which provided, at the 3 db point, a beam width of approximately 11° and a beam shift of 8.5° making a total beam width of approximately 20° when scanning. The minor lobes were down more than 17 db (one way) from the maximum; and the gain of this antenna was 21 db. This antenna assembly was integral with a transmitter-receiver (Fig. 55) which was mounted on the associated gun director. Consequently, the size of the reflector was limited by requirements for unobstructed vision for the operators in the director. As a matter of fact, for this type of radar system serious consideration must be given to the size and weight of the antenna and associated components.

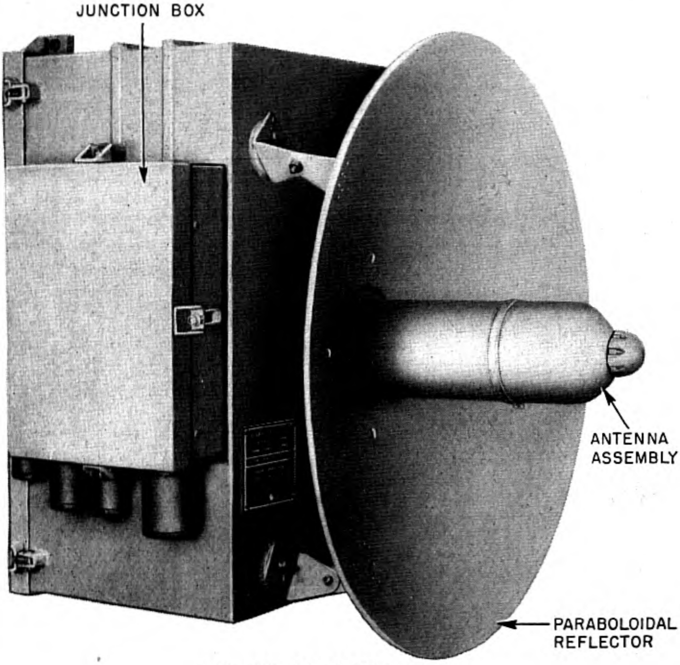


Fig. 55—Mark 19 Antenna.

14.11 *The Mark 28 Radar Antenna*

The beam from the antenna used in Radar Equipment Mark 19 was relatively broad and to improve target resolution, the diameter oi the reflector for the antenna in Mark 28 was approximately doubled. The Mark 28 is a 10 cm system and employs a conical scanning antenna similar to that described for Mark 19. The essential difference is that the spun steel parabolic reflector is 45 inches in diameter which provides a beam width of approximately 6.5° and a beam shift of 4.5° making a total of 11°.

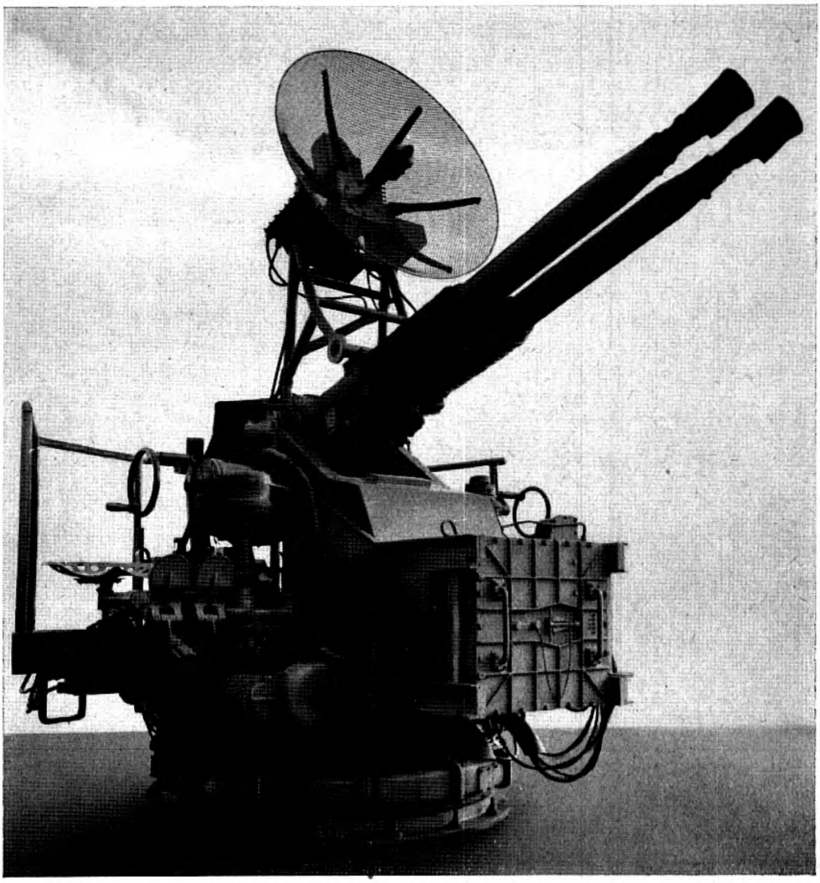


Fig. 56—Mark 28 Antenna Mounted on 40 MM Gun.

The minor lobes are down more than 17 db (one way) from the maximum; and the gain of this antenna is 26 db. It was found necessary to perforate the reflector of this dimension in order to reduce deflection caused by gun blast and by wind drag on the antenna assembly. The antenna assembly for Radar Equipment Mark 28 is shown in Fig. 56. This assembly is shown mounted on a 40 mm Gun.

14.12 *A 3 CM Anti-Aircraft Radar Antenna.*

To obtain greater discrimination between a given target and other targets, or between a target and its surroundings, the wavelength was reduced to the 3 cm region. An antenna for this wavelength was designed to employ the conical scan principle. In this case the parabolic reflector was 30 inches in diameter and transmitted a beam approximately 3° wide at the 3db point with a beam shift of 1.5° making a total of 4.5° with the antenna scanning. The minor lobes are down more than 22 db (one way) from the maximum; and the gain of this antenna is 35 db.

In the 3 cm system in which a Cutler feed was used, the axis of the beam was rotated in an orbit by "nutation" about the mechanical axis of the antenna. This was accomplished by passing circular waveguide through the hollow shaft of the driving motor. The rear end of the feed (choke coupling end) was fixed in a ball pivot while the center (near the reflector) was off set the proper amount to develop the required beam shift. This off set was produced by a rotating eccentric driven by the motor. The latter was a 440 volt, 60 cycle, 3 phase motor rotating at approximately 1800 rpm which resulted in a scanning rate of 30 cycles per second. The two-phase reference voltage generator was integral with the driving motor.

It was found necessary at these radio frequencies to use a cast aluminum reflector and to machine the reflecting surface to close tolerances in order to attain the consistency in beam width and beam direction required for accurate pointing. An antenna assembly for the 3 cm anti-aircraft radar is shown in Fig. 57.

15. Land Based Radar Antennas

15.1 *The SCR-545 Radar "Search" and "Track" Antennas*[17]

The SCR-545 Radar Set was developed at the Army's request to meet the urgent need for a radar set to detect aircraft and provide accurate target tracking data for the direction of anti-aircraft guns.

This use required that a narrow beam tracking antenna be employed to achieve the necessary tracking accuracy, furthermore, a narrow beam antenna suitable for accurate tracking has a very limited field of view and requires additional facilities for target acquisition. This was provided by the search antenna which has a relatively large field of view and is provided with facilities for centering the target in its field of view. These two antennas are integrated into a single mechanical structure and both radar axes coincide.

The "Search" antenna operates in the 200 mc band and is com-

[17] Section 15.1 was written by A. L. Robinson.

posed of an array of 16 quarter wave dipoles spaced 0.1 wave-length in front of a flat metal reflector. All feed system lines and impedance matching devices are made up of coaxial transmission line sections. The array is divided into four quarters, each being fed from the lobe switching mechanism. This division is required to permit lobe switching in both horizontal and vertical planes. The function of the lobe switching mechanism is to introduce a particular phase shift in the excitation of the elements of one half of the antenna with respect to the other half. The theory of this type of lobe switching is discussed in section 11.1. The antenna beam spends approximately one quarter of a lobing cycle in each one of the four lobe positions. Each of the four lobe positions has the same radiated field intensity along the antenna axis and therefore when a target is on axis equal signals will be received from all four lobe positions.

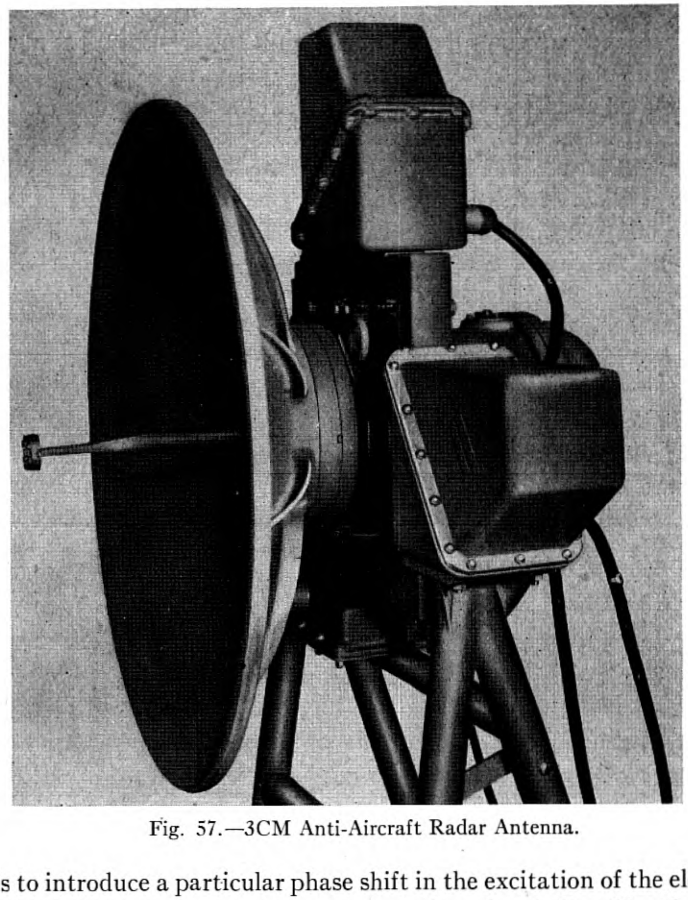


Fig. 57.—3CM Anti-Aircraft Radar Antenna.

The "Track" antenna operates in the 10 cm. region and consists of a reflector which is a parabola or revolution, 57 inches in diameter, illuminated by a source of energy emerging from a round waveguide in the lobing mechanism. Conical lobing is achieved by rotating the source of energy around the parabola axis in the focal plane of the parabola. Conical lobing is discussed in section 11.2. The round waveguide forming the source is filled with a specially shaped polystyrene core to control the illumination of the parabola and to seal the feed system against the weather. The radio frequency power is fed through coaxial transmission line to a coaxial-waveguide transition which is attached to the lobing mechanism.

The "Search" and "Track" antenna lobing mechanisms are synchronized and driven by a common motor.

The radio frequency power for both antennas is transmitted through a single specially constructed coaxial transmission line to the common antenna structure, where a coaxial transmission line filter separates the power for each antenna.

Figure 58 is a photograph of a production model of the SCR-545 Radar Set. The principal electrical characteristics of the antennas are tabulated below:

	Antennas	
	Search	Track
Gain	14.5 db	30 db
Horizontal Beamwidth	23.5°	5°
Vertical Beamwidth	25.5°	5°
Polarization	Horizontal	Vertical
Type of Lobing	Lobe switching	Conical lobing
Angle between lobe positions	10°	3°
Lobing rate	60 cycles/sec.	60 cycles/sec.

The SCR-545 played an important part in the Italian campaign, particularly in helping to secure the Anzio Beach Head area, as well as combating the "V" bombs in Belgium. However the majority of SCR-545 equipments were sent to the Pacific Theater of Operations and played an important part in operations on Leyte, Saipan, Iwo Jima, and Okinawa.

15.2 *The AN/TPS-1A Portable Search Antenna*[18]

In order to provide early warning information for advanced units, a light weight, readily transportable radar was designed under Signal Corps contract.

[18] Written by R. E. Crane.

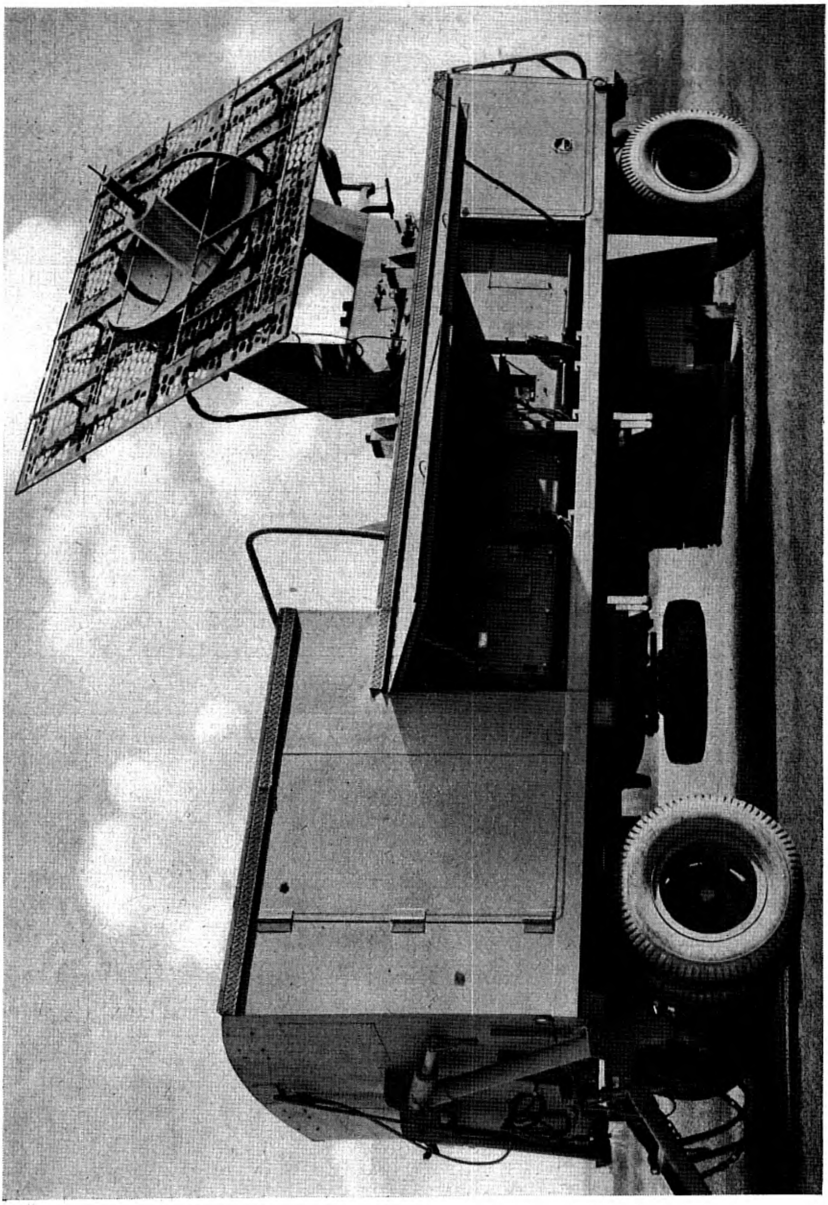


Fig. 58—SCR/545 Antenna.

The objective was to obtain as long range early warning as possible with moderate accurracy of location. Emphasis was placed on detection of low flying planes.

The objectives for the set indicated that the antenna should be built as large as reasonable and placed as high as reasonable for a portable set. Some latitude in choice of frequency was permitted at first. For ruggedness and reliability reasons which seemed controlling at the time, the frequency was pushed as high as possible with vacuum tube detectors and R.F. amplifiers. This was finally set at 1080 mc.

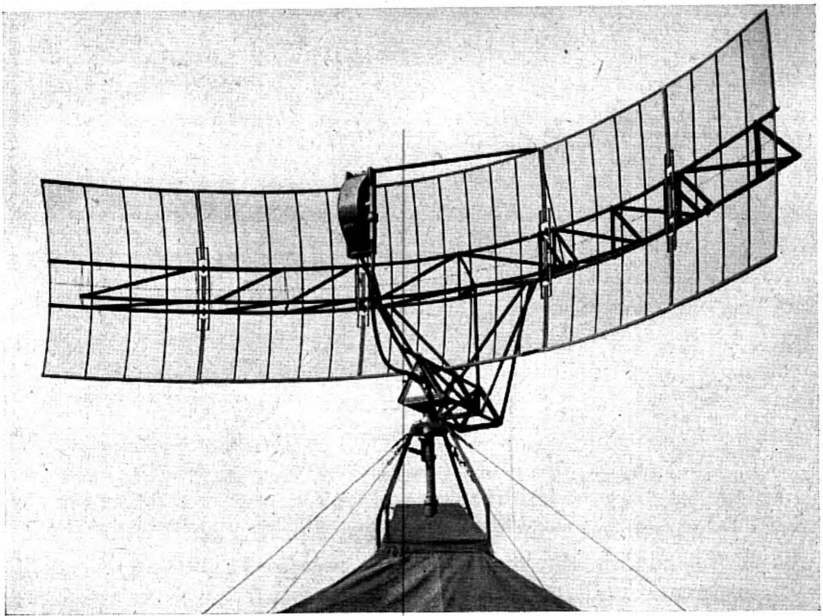


Fig. 59—AN/TPS-1A Antenna.

The antenna as finally produced was 15 ft. in width and 4 ft. in height˙ The reflecting surface was paraboloidal. The mouth of the feed horn was approximately at the focus of the generating parabola. The feedhorn was excited by a probe consisting of the inner conductor of the coaxial transmission line extended through the side of the horn and suitably shaped. To reduce side lobeṡ and back radiation the feedhorn was dimensioned to taper the illumination so that it was reduced about 10 db in the horizontal and vertical planes at the edges of the reflector. Dimensions of probe and exact location of feed, etc. were determined empirically to secure acceptable impedance over the frequency band needed. This band, covered by spot frequency magnetrons, was approximately ±2.5% from mid frequency.

Figure 59 shows the antenna in place on top of the set.

The characteristics of this antenna are summarized below:

Gain	27.3 db.
Horizontal Half Power Beamwidth	4.4°
Vertical Half Power Beamwidth	12.6°
Vertical Beam Characteristic	Symmetrical
Polarization	Horizontal
Impedance (SWR over ±2.5% band)	<4.0 db

16. Airborne Radar Antennas

16.1 *The AN/APS-4 Antenna*[19]

AN/APS-4 was designed to provide the Navy's carrier-based planes with a high performance high resolution radar for search against surface and airborne targets, navigation and interception of enemy planes under conditions of fog and darkness. For this service, weight was an all important consideration and throughout a production schedule that by V-J day was approaching 15,000 units, changes to reduce weight were constantly being introduced. In late production the antenna was responsible for 19 lbs. out of a total equipment weight of 164 lbs. The military requirements called for a scan covering 150° in azimuth ahead of the plane and 30° above and below the horizontal plane in elevation. To meet this requirement a Cutler feed and a parabolic reflector of 6.3″ focal length and $14\frac{1}{2}$″ diameter was selected. Scanning in azimuth was performed by oscillating reflector and feed through the required 150° while elevation scan was performed by tilting the reflector. Beam pattern was good for all tilt angles. In early flight tests the altitude line on the B scope due to reflection from the sea beneath was found to be a serious detriment to the performance of the set. To reduce this, a feed with elongated slots designed for an elliptical reflector was tried and found to give an improvement even when used with the approximately round reflector. The elliptical reflector was also tried, but did not improve the performance sufficiently to justify the increased size.

As will be noted in Fig. 60, the course of the mechanical development brought the horizontal pivot of the reflector to the form of small ears projecting through the parabola. No appreciable deterioration of the beam pattern due to this unorthodox expedient was noted.

The equipment as a whole was built into a bomb-shaped container hung in the bomb rack on the underside of the wing. Various accidents resulted in this container being torn off the wing in a crash landing in water or dropped on the deck of the carrier. After these mishaps, the equipment was frequently found to be in good working order with little or no repair required.

[19] Written by F. C. Willis.

Gain	28 db
Beamwidth	6° approx. circular
Polarization	Horizontal
Scan	Mechanical
Scanning Sector—Azimuth	150°
Scanning Sector—Elevation	60°
Scanning Rate	one per sec.
Total weight	19 lbs.

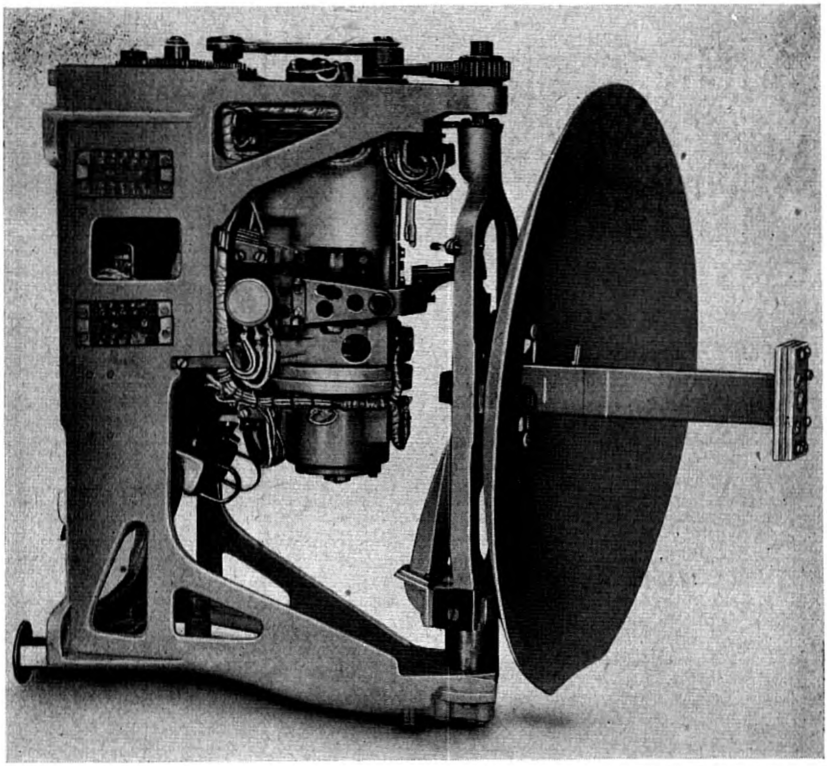


Fig. 60—AN/APS-4 Antenna.

16.2 *The SCR-520, SCR-717 and SCR-720 Antennas*[20]

The antenna shown in Fig. 61 is typical of the type used with the SCR-520 and SCR-720 aircraft interception (night fighter) airborne radar equipment, as well as the SCR-717 sea search and anti-submarine airborne radar equipment. The parabolic reflector is 29 inches in diameter and produces a radiation beam about 10° wide. The absolute gain is approximately 25 db. RF energy is supplied to a pressurized emitter through a pressurized transmission line system which includes a rotary joint located on the ver-

[20] Written by J. F. Morrison.

tical axis and a tilt joint on the horizontal axis. Either vertical or horizontal polarization can be used by rotating the mounting position of the emitter. Vertical polarization is preferred for aircraft interception work and horizontal polarization is preferred for sea search work.

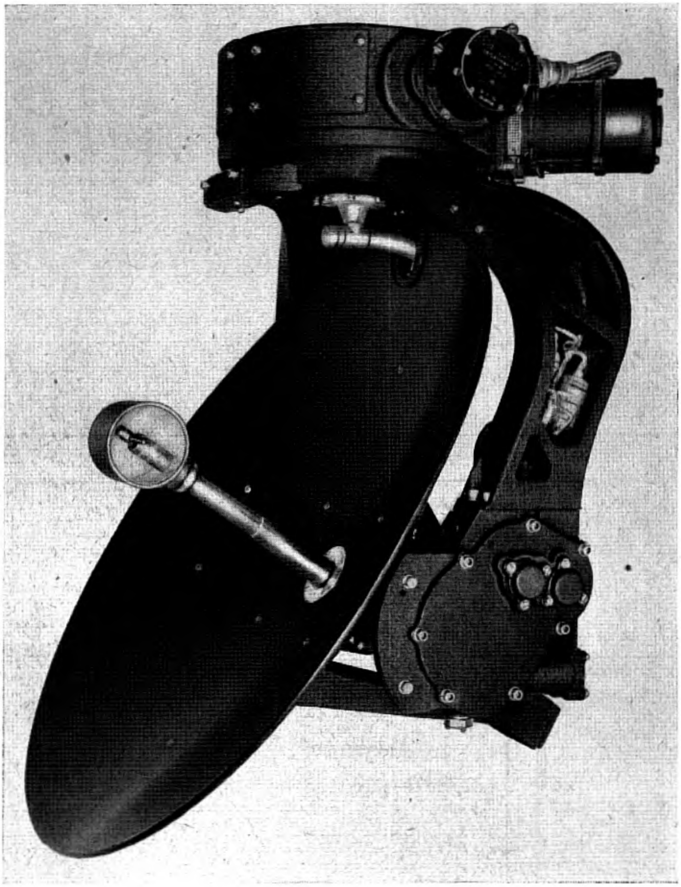


Fig. 61—SCR-520 Antenna.

For aircraft interception the military services desired to scan rapidly a large solid angle forward of the pursuing airplane, i.e. 90° right and left, 15° below and 50° above the line of flight. The data is presented to the operator in the form of both "B" and "C" presentations and for this purpose potentiometer data take-offs are provided on the antenna. The reflector is spun on a vertical axis at a rate of 360 rpm and at the same time it is made to nod up and down about its horizontal axis by controllable amounts up to a total of 65° and at a rate of 30° per second.

In the sea search SCR-717 equipment, selsyn azimuth position data take-offs are provided which drive a PPI type of indicator presentation. The rotational speed about the vertical axis in this case is either 8 or 20 rpm as selected by the operator. The reflector can also be tilted about its horizontal axis above or below the line of flight as desired by the operator.

It will be noted that the emitter moves with the reflector and accordingly it is always located at the focal point throughout all orientations of the antenna.

16.3 *The AN/APQ-7 Radar Bombsight Antenna*[21]

Early experience in the use of bombing-through-overcast radar equipment indicated that a severe limitation in performance was to be expected as the result of the inadequate resolution offered by the then available airborne radar equipments. This lack of resolution accounted for gross errors in bombing where the target area was not ideal from a radar standpoint.

To meet this increased resolution requirement in range, the transmitted pulse width was shortened considerably. In attempting to increase the azimuthal resolution, higher frequencies of transmission were employed. This enabled an improvement in azimuthal resolution without resorting to larger radiating structures, a most important consideration on modern high speed military aircraft.

To extend the size of the radiating structure without penalizing the aircraft performance, the use of a linear scanning array which would exhibit high azimuthal resolution was considered. This array was originally conceived in a form suitable to mount within the existing aircraft wing and transmit through the leading edge. As development proceeded, the restrictions imposed on the antenna structure as well as the aircraft wing design resulted in the linear array scanner being housed in an appropriate separate air foil and attached to the aircraft fuselage (Fig. 62).

The above study resulted in the development of the AN/APQ-7 radar equipment, operating at the X-band of frequencies.[22] This equipment provided facilities for radar navigation and bombing.

The AN/APQ-7 antenna consisted of an array of 250 dipole structures spaced at $\frac{1}{2}$ wavelength intervals and energized by means of coupling probes extending into a variable width waveguide. The vertical pattern was arranged to exhibit a modified $\csc^2$ distribution by means of accurately shaped "flaps" attached to the assembly.

[21] Written by L. W. Morrison.

[22] A large part of the antenna development was carried out at the M. I. T. Radiation Laboratory.

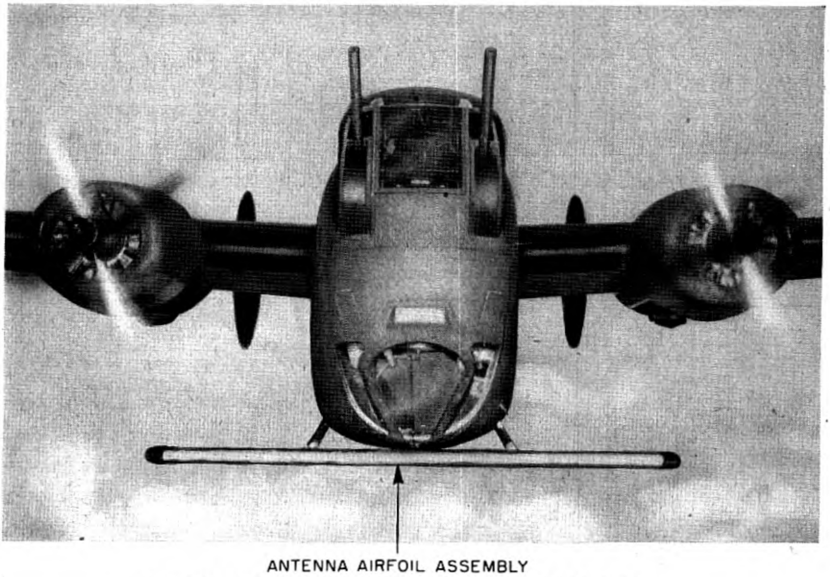


Fig. 62—AN/APQ-7 Antenna Mounted on B24 Bomber.

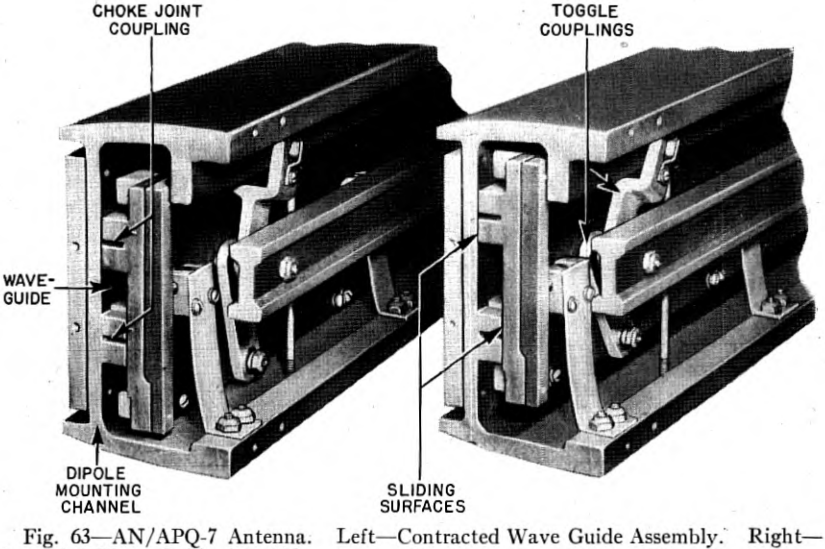


Fig. 63—AN/APQ-7 Antenna. Left—Contracted Wave Guide Assembly. Right—Expanded Wave Guide Assembly.

The scanning of the beam is accomplished by varying the width of the feed waveguide. This is accomplished by means of a motor driven actuated cam which drives a push rod extending along the waveguide assembly back and forth. Toggle arms are attached to this push rod at frequently spaced intervals which provides the motion for varying the width of the waveguide while assuring precise parallelism of the side walls throughout its length (Fig. 63).

The normal range of horizontal scanning exhibited by this linear array, extends from a line perpendicular to the array to 30° in the direction of the feed. By alternately feeding each end, a total scanning range of ±30° from the perpendicular is achieved. Appropriate circuits to synchronize the indicator for this range are included.

The use of alternate end feed on the AN/APQ-7 antenna requires that the amount of energy fed to the individual dipoles is somewhat less than if a single end feed is employed.

The AN/APQ-7 antenna is 16½ feet in length and weighs 180 pounds exclusive of air foil housing.

The following data applies:

Gain = 32.5 db
Horizontal beamwidth = 0.4°
Vertical beam characteristic = modified $\csc^2$
Scan—Array scanning
Scanning Sector—± 30° Horizontal
Scanning Rate = 45°/second

Acknowledgments

Contributors to the research and development of the radar antennas described in this paper included not only the great number of people directly concerned with these antennas but also the many people engaged in general research and development of microwave components and measuring techniques. A complete list of credits, therefore, will not be attempted.

In addition to the few individuals mentioned in footnotes throughout the paper, the authors would like to pay special tribute to the following co-workers in the Radio Research Department: C. B. H. Feldman who with the assistance of D. H. Ring made an outstanding contribution in the development of the polyrod array antenna; W. A. Tyrrell for his work on lobe switches; A. G. Fox, waveguide phase changers; A. P. King, paraboloids and horn antennas; A. C. Beck, submarine antennas; G. E. Mueller, polyrods.

Probability Functions for the Modulus and Angle of the Normal Complex Variate

By RAY S. HOYT

This paper deals mainly with various 'distribution functions' and 'cumulative distribution functions' pertaining to the modulus and to the angle of the 'normal' complex variate, for the case where the mean value of this variate is zero. Also, for auxiliary uses chiefly, the distribution function pertaining to the reciprocal of the modulus is included. For all of these various probability functions the paper derives convenient general formulas, and for four of the functions it supplies comprehensive sets of curves; furthur, it gives a table of computed values of the cumulative distribution function for the modulus, serving to verify the values computed by a different method in an earlier paper by the same author.[1]

Introduction

IN THE solution of problems relating to alternating current networks and transmission systems by means of the usual complex quantity method, any deviation of any quantity from its reference value is naturally a complex quantity, in general. If, further, the deviation is of a random nature and hence is variable in a random sense, then it constitutes a 'complex random variable,' or a 'complex variate,' the word 'variate' here meaning the same as 'random variable' (or 'chance variable'—though, on the whole, 'random variable' seems preferable to 'chance variable' and is more widely used).

Although a complex variate may be regarded formally as a single analytical entity, denotable by a single letter (as Z), nevertheless it has two analytical constituents, or components: for instance, its real and imaginary constituents (X and Y); also, its modulus and amplitude ($|Z|$ and θ). Correspondingly, a complex variate can be represented geometrically by a single geometrical entity, namely a plane vector, but this, in turn, has two geometrical components, or constituents: for instance, its two rectangular components (X and Y); also, its two polar components, radius vector and vectorial angle ($R \equiv |Z|$ and θ).

This paper deals mainly with the modulus and the angle of the complex variate,[2] which are often of greater theoretical interest and practical im-

[1] "Probability Theory and Telephone Transmission Engineering," *Bell System Technical Journal,* January 1933, which will hereafter be referred to merely as the "1933 paper".

[2] Throughout the paper, I have used the term 'complex variate' for any 2-dimensional variate, because of the nature of the contemplated applications indicated in the first

portance than the real and imaginary constituents. The modulus variate and the angle variate, individually and jointly, are of considerable theoretical interest; while the modulus variate is also of very considerable practical importance, and the angle variate may conceivably become of some practical importance.

The paper is concerned chiefly with the 'distribution functions'[3] and the 'cumulative distribution functions' pertaining to the modulus (Sections 3 and 5) and to the angle (Sections 6 and 7) of the 'normal' complex variate, for the case where the mean value of this variate is zero. The distribution function for the reciprocal of the modulus is also included (Section 4).

The term 'probability function' is used in this paper generically to include 'distribution function' and 'cumulative distribution function.'

To avoid all except short digressions, some of the derivation work has been placed in appendices, of which there are four. These may be found of some intrinsic interest, besides facilitating the understanding of the paper.

1. Distribution Function and Cumulative Distribution Function in General: Definitions, Terminology, Notation, Relations, and Formulas

The present section constitutes a generic basis for the rest of the paper.

Let τ denote any complex variate, and let ρ and σ denote any pair of real quantities determining τ and determined by τ. (For instance, ρ and σ might be the real and imaginary components of τ, or they might be the modulus and angle of τ.) Geometrically, ρ and σ may be pictured as general curvilinear coordinates in a plane, as indicated by Fig. 1.1.

Let τ' denote the unknown value of a random sample consisting of a single τ-variate, and ρ' and σ' the corresponding unknown values of the constituents of τ'.

Further, let $G(\rho, \sigma)$ denote the 'areal probability density' at any point ρ,σ in the ρ,σ-plane, so that $G(\rho,\sigma)dA$ gives the probability that τ' falls in a differential area dA containing the point τ; and so that the integral of

paragraph of the Introduction, and also because the present paper is a sort of sequel to my 1933 paper, where the term 'complex variate' (or rather, 'complex chance-variable') was used throughout since there it seemed clearly to be the best term, on account of the field of applications contemplated and the specific applications given as illustrations. However, for wider usage the term 'bivariate' might be preferred because of its prevalence in the field of Mathematical Statistics; and therefore the paper should be read with this alternative in view.

[3] The term 'distribution function' is used with the same meaning in this paper as in my 1933 paper, although there the term 'probability law' was used much more frequently than 'distribution function,' but with the same meaning.

$G(\rho,\sigma)dA$ over the entire ρ,σ-plane is equal to unity, corresponding to certainty.

For the sake of subsequent needs of a formal nature, it will now be assumed that $G(\rho,\sigma) = 0$ at all points ρ,σ outside of the ρ_1, ρ_2, σ_1, σ_2 quadrilateral region in the ρ,σ-plane, Fig. 1.1, bounded by arcs of the four heavy curves, for which ρ has the values ρ_1 and ρ_2 and σ the values σ_1 and σ_2, with ρ_1 and σ_1 regarded, for convenience, as being less than ρ_2 and σ_2 respectively. Further, $G(\rho,\sigma)$ will be assumed to be continuous inside of this quadrilateral region, and to be non-infinite on its boundary. Hence, for probability purposes, it will suffice to deal with the open inequalities

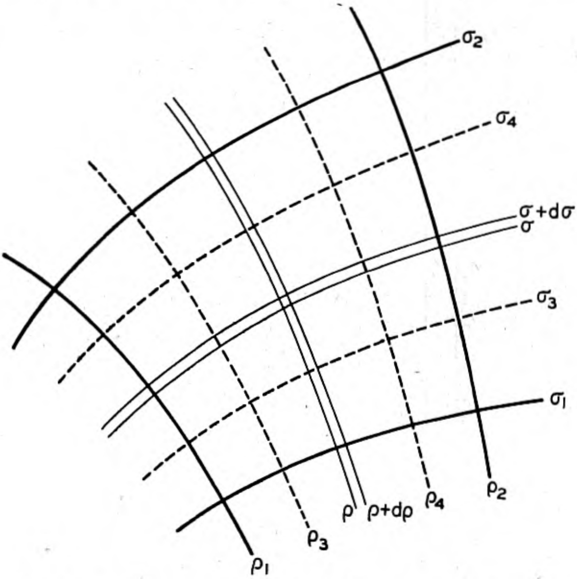


Fig. 1.1—Diagram of general curvilinear coordinates.

$$\rho_1 < \rho < \rho_2, \qquad (1.1) \qquad\qquad \sigma_1 < \sigma < \sigma_2, \qquad (1.2)$$

which pertain to this quadrilateral region excluding its boundary; and thus it will not be necessary to deal with the closed inequalities $\rho_1 \leqq \rho \leqq \rho_2$ and $\sigma_1 \leqq \sigma \leqq \sigma_2$, which include the boundary.[4]

[4] The matters dealt with generically in this paragraph may be illustrated by the following two important particular cases, which occur further on, namely:

POLAR COORDINATES: $\rho = |\tau| = R$, $\sigma = \theta =$ angle of τ. Then $\rho_1 = R_1 = 0$, $\rho_2 = R_2 = \infty$, $\sigma_1 = \theta_1 = 0$, $\sigma_2 = \theta_2 = 2\pi$, whence (1.1) and (1.2) become $0 < R < \infty$ and $0 < \theta < 2\pi$, respectively.

RECTANGULAR COORDINATES: $\rho = \text{Re}\,\tau = x$, $\sigma = \text{Im}\,\tau = y$. Then $\rho_1 = x_1 = -\infty$, $\rho_2 = x_2 = \infty$, $\sigma_1 = y_1 = -\infty$, $\sigma_2 = y_2 = \infty$, whence (1.1) and (1.2) become $-\infty < x < \infty$ and $-\infty < y < \infty$, respectively.

A generic quadrilateral region contained within the quadrilateral region ρ_1, ρ_2, σ_1, σ_2 in Fig. 1.1 is the one bounded by arcs of the dashed curves ρ_3, ρ_4, σ_3, σ_4, where $\rho_3 < \rho_4$ and $\sigma_3 < \sigma_4$. Here, as in the preceding paragraph, it will evidently suffice to deal with open inequalities.

Referring to Fig. 1.1, the probability functions with which this paper will chiefly deal are certain particular cases of the probability functions $P(\rho, \sigma)$, $P(\rho \mid \sigma_{34})$ and $Q(\rho_{34}, \sigma_{34})$ occurring on the right sides of the following three equations respectively:

$$p(\rho < \rho' < \rho + d\rho, \sigma < \sigma' < \sigma + d\sigma) = P(\rho,\sigma)d\rho d\sigma, \tag{1.3}$$

$$p(\rho < \rho' < \rho + d\rho, \sigma_3 < \sigma' < \sigma_4) = P(\rho \mid \sigma_{34})d\rho, \tag{1.4}$$

$$p(\rho_3 < \rho' < \rho_4, \sigma_3 < \sigma' < \sigma_4) = Q(\rho_{34}, \sigma_{34}). \tag{1.5}$$

These equations serve to define the above-mentioned probability functions occurring on the right sides in terms of the probabilities denoted by the left sides, each expression $p(\quad)$ on the left side denoting the probability of the pair of inequalities within the parentheses.[5] Inspection of these equations shows that: $P(\rho,\sigma)$ is the 'distribution function' for ρ and σ jointly; $P(\rho \mid \sigma_{34})$ is a 'distribution function' for ρ individually, with the understanding that σ' is restricted to the range σ_3-to-σ_4; $Q(\rho_{34}, \sigma_{34})$ is a 'cumulative distribution function' for ρ and σ jointly.

Since the left sides of (1.3), (1.4) and (1.5) are necessarily positive, the right sides must be also. Hence, as all of the probability functions occurring in the right sides are of course desired to be positive, the differentials $d\rho$ and $d\sigma$ must be taken as positive, if we are to avoid writing $|d\rho|$ and $|d\sigma|$ in place of $d\rho$ and $d\sigma$ respectively.

Returning to (1.3), it is seen that, stated in words, $P(\rho,\sigma)$ is such that $P(\rho,\sigma)d\rho d\sigma$ gives the probability that the unknown values ρ' and σ' of the constituents of the unknown value τ' of a random sample consisting of a single τ-variate lie respectively in the differential intervals $d\rho$ and $d\sigma$ containing the constituent values ρ and σ respectively. Thus, unless $d\rho d\sigma$ is the differential element of area, $P(\rho,\sigma)$ is not equal to the 'areal probability density,' $G(\rho,\sigma)$, defined in the fourth paragraph of this section. In general, if E is such that $Ed\rho d\sigma$ is the differential element of area, then $P(\rho, \sigma) = EG(\rho, \sigma)$. (An illustration is afforded incidentally by Appendix A.)

$P(\rho,\sigma)$, defined by (1.3), is the basic 'probability function,' in the sense that the others can be expressed in terms of it, by integration. Thus

[5] Thus p in $p(\quad)$ may be read 'probability that' or 'probability of.'

$P(\rho \mid \sigma_{34})$ and $P(\sigma \mid \rho_{34})$, defined respectively by (1.4) and by the correlative of (1.4), can be expressed as 'single integrals,' as follows[6]:

$$P(\rho \mid \sigma_{34}) = \int_{\sigma_3}^{\sigma_4} P(\rho,\sigma)\, d\sigma, \quad (1.6) \qquad P(\sigma \mid \rho_{34}) = \int_{\rho_3}^{\rho_4} P(\rho,\sigma)\, d\rho. \quad (1.7)$$

$Q(\rho_{34}, \sigma_{34})$, defined by (1.5), can be expressed as a 'double integral,' fundamentally; but, for purposes of analysis and of evaluation, this will be replaced by its two equivalent 'repeated integrals':

$$Q(\rho_{34}, \sigma_{34}) = \int_{\rho_3}^{\rho_4} \left[\int_{\sigma_3}^{\sigma_4} P(\rho,\sigma)\, d\sigma \right] d\rho = \int_{\sigma_3}^{\sigma_4} \left[\int_{3}^{\rho_4} P(\rho,\sigma)\, d\rho \right] d\sigma, \quad (1.8)$$

the set of integration limits being the same in both repeated integrals because these limits are constants, as indicated by Fig. 1.1. On account of (1.6) and (1.7) respectively, (1.8) can evidently be written formally as two single integrals:

$$Q(\rho_{34}, \sigma_{34}) = \int_{\rho_3}^{\rho_4} P(\rho \mid \sigma_{34})\, d\rho = \int_{\sigma_3}^{\sigma_4} P(\sigma \mid \rho_{34})\, d\sigma, \quad (1.9)$$

but implicitly these are repeated integrals unless the single integrations in (1.6) and (1.7) can be executed, in which case the integrals in (1.9) will actually be single integrals, and these will be quite unlike each other in form, being integrals with respect to ρ and σ respectively—though of course yielding a common expression in case the indicated integrations can be executed.

The particular cases of (1.4) and (1.5) with which this paper will chiefly deal are the following three:

$$p(\rho < \rho' < \rho + d\rho,\ \sigma_1 < \sigma' < \sigma_2) = P(\rho \mid \sigma_{12})\, d\rho \equiv P\,(\rho)\, d\rho, \quad (1.10)$$

$$p(\rho_1 < \rho' < \rho,\ \sigma_1 < \sigma' < \sigma_2) = Q(<\rho,\sigma_{12}) \equiv Q(\rho), \quad (1.11)$$

$$p(\rho < \rho' < \rho_2,\ \sigma_1 < \sigma' < \sigma_2) = Q(>\rho,\sigma_{12}) \equiv Q^*(\rho). \quad (1.12)$$

[6] The single-integral formulation in (1.6) can be written down directly by mere inspection of the left side of (1.4). Alternatively, (1.6) can be obtained by representing the left side of (1.4) by a repeated integral, as follows:

$$P(\rho \mid \sigma_{34}) d\rho = \int_{\rho}^{\rho+d\rho} \left[\int_{\sigma_3}^{\sigma_4} P(\rho,\ \sigma) d\sigma \right] d\rho = \left[\int_{\sigma_3}^{\sigma_4} P(\rho,\ \sigma) d\sigma \right] d\rho,$$

whence (1.6); the last equality in the above chain equation in this footnote evidently results from the fact that, in general, $\int_{x}^{x+dx} f(x)dx = f(x)dx$, since each side of this equation represents dA, the differential element of area under the graph of $f(x)$ from x to $x + dx$.

In each of these three equations the very abbreviated notation at the extreme right will be used wherever the function is being dealt with extensively, as in the various succeeding sections. Such notation will not seem unduly abbreviated nor arbitrary if the following considerations are noted: In (1.10), σ_{12} corresponds to the entire effective range of σ, so that $P(\rho \mid \sigma_{12})$ is the 'principal' distribution function for ρ. Similarly, in (1.11), $Q(< \rho, \sigma_{12})$ is the 'principal' cumultive distribution function for ρ. In (1.12), the star indicates that $Q^*(\rho)$ is the 'complementary' cumulative distribution function, since $Q(\rho) + Q^*(\rho) = Q(\rho_{12}, \sigma_{12}) = 1$, unity being taken as the measure of certainty, of course.

For occasional use in succeeding sections, the defining equations for the probability functions pertaining to four other particular cases will be set down here:

$$p(\rho < \rho' < \rho + d\rho, \sigma_1 < \sigma' < \sigma) = P(\rho \mid < \sigma)\, d\rho, \tag{1.13}$$

$$p(\rho < \rho' < \rho + d\rho, \sigma < \sigma' < \sigma_2) = P(\rho \mid > \sigma)\, d\rho, \tag{1.14}$$

$$p(\rho_1 < \rho' < \rho, \sigma_1 < \sigma' < \sigma) = Q(< \rho, < \sigma), \tag{1.15}$$

$$p(\rho < \rho' < \rho_2, \sigma_1 < \sigma' < \sigma) = Q(> \rho, < \sigma). \tag{1.16}$$

It may be noted that (1.13) and (1.14) are mutually supplementary, in the sense that their sum is (1.10). Similarly, (1.15) and (1.16) are mutually supplementary, in the sense that their sum is $Q(\rho_{12}, < \sigma) = Q(< \sigma, \rho_{12})$, which is the correlative of (1.11).

This section will be concluded with the following three simple transformation relations (1.17), (1.18) and (1.19), which will be needed further on. They pertain to the probability functions on the right sides of equations (1.3), (1.4) and (1.5) respectively. h and k denote any positive real constants, the restriction to positive values serving to simplify matters without being too restrictive for the needs of this paper.

$$P(h\rho, k\sigma) = \frac{1}{hk} P(\rho, \sigma), \tag{1.17}$$

$$P(h\rho \mid k\sigma_{34}) = \frac{1}{h} P(\rho \mid \sigma_{34}), \tag{1.18}$$

$$Q(h\rho_{34}, k\sigma_{34}) = Q(\rho_{34}, \sigma_{34}). \tag{1.19}$$

Each of the three formulas (1.17), (1.18), (1.19) can be rather easily derived in at least two ways that are very different from each other. One way depends on probability inequality relations of the sort

$$p(t < t' < t + dt) = p(gt < gt' < gt + d[gt]), \tag{1.20}$$

$$p(t_3 < t' < t_4) = p(gt_3 < gt' < gt_4), \tag{1.21}$$

where t stands generically for ρ and for σ, and g is any positive real constant, standing generically for h and for k; (1.20) and (1.21) are easily seen to be true by imagining every variate in the universe of the t-variates to be multiplied by g, thereby obtaining a universe of (gt)-variates. A second way of deriving each of the three formulas (1.17), (1.18), (1.19) depends on general integral relations of the sort

$$\int_a^b f(t)\,dt = \frac{1}{g}\int_{ga}^{gb} f(t)\,d(gt) = \frac{1}{g}\int_{ga}^{gb} f\left(\frac{\lambda}{g}\right)d\lambda. \tag{1.22}$$

A third way, which is distantly related to the second way, depends on the use of the Jacobian for changing the variables in any double integral; thus,

$$\frac{P(\rho,\sigma)}{P(\lambda,\mu)} = \left|\frac{d\lambda d\mu}{d\rho d\sigma}\right| = \left|\frac{\partial(\lambda,\mu)}{\partial(\rho,\sigma)}\right| = 1 \div \left|\frac{\partial(\rho,\sigma)}{\partial(\lambda,\mu)}\right|, \tag{1.23}$$

the first equality in (1.23) depending on the fact that the two sets of variables and of differentials have corresponding values and hence are so related that

$$p(\rho<\rho'<\rho+d\rho,\ \sigma<\sigma'<\sigma+d\sigma) = p(\lambda<\lambda'<\lambda+d\lambda,\ \mu<\mu'<\mu+d\mu), \tag{1.24}$$

whence

$$P(\rho,\sigma)\,|\,d\rho d\sigma\,| = P(\lambda,\mu)\,|\,d\lambda d\mu\,|.$$

2. The Normal Complex Variate and Its Chief Probability Functions

The 'normal' complex variate may be defined in various equivalent ways. Here, a given complex variate $z = x + iy$ will be defined as being 'normal' if it is possible to choose in the plane of the scatter diagram of z a pair of rectangular axes, u and v, such that the distribution function[7] $P(u,v)$ for the given complex variate with respect to these axes can be written in the form[8]

$$P(u,v) = \frac{1}{2\pi S_u S_v}\exp\left[-\frac{u^2}{2S_u^2} - \frac{v^2}{2S_v^2}\right] = P(u)P(v). \tag{2.1}$$

We shall call $w = u + iv$ the 'modified' complex variate, as it represents the value of the given complex variate $z = x + iy$ when the latter is referred to the u,v-axes; $P(u)$ and $P(v)$ are respectively the individual distribution functions for the u and v components of the modified complex variate; and

[7] Defined by equation (1.3) on setting $\rho = u$ and $\sigma = v$.

[8] This equation is (12) of my 1933 paper. It can be easily verified that the (double) integral of (2.1) taken over the entire u, v-plane is equal to unity.

S_u and S_v are distribution parameters called the 'standard deviations' of u and v respectively. If t stands for u and for v generically, then

$$P(t) = \frac{1}{\sqrt{2\pi}S_t}\exp\left[-\frac{t^2}{2S_t^2}\right], \qquad (2.2) \qquad S_t^2 = \int_{-\infty}^{\infty} t^2 P(t)\,dt. \qquad (2.3)$$

From the viewpoint of the scatter diagram, the distribution function $P(u,v)$ is, in general, equal to the 'areal probability density' at the point u,v in the plane of the scatter diagram, so that the probability of falling in a differential element of area dA containing the point u,v is equal to $P(u,v)dA$; similarly, $P(u)$ and $P(v)$ are equal to the component probability densities. In particular, the probability density is 'normal' when $P(u,v)$ is given by (2.1).

Geometrically, equation (2.1) evidently represents a surface, the normal 'probability surface,' situated above the u, v-plane; and $P(u,v)$ is the ordinate from any point u,v in the u,v-plane to the probability surface.

The u,v-axes described above will be recognized as being the 'principal central axes,' namely that pair of rectangular axs which have their origin at the 'center' of the scatter diagram of $z = x + iy$ and hence at the center of the scatter diagram of $w = u + iv$, so that $\overline{w} = 0$, and are so oriented in the scatter diagram that $\overline{uv} = 0$ (whereas $\bar{z} \neq 0$ and $\overline{xy} \neq 0$, in general).

In equation (2.1), which has been adopted above as the analytical basis for defining the 'normal' complex variate, the distribution parameters are S_u and S_v; and they occur symmetrically there, which is evidently natural and is desirable for purposes of definition. Henceforth, however, it will be preferable to adopt as the distribution parameters the quantities S and b defined by the pair of equations[9]

$$S^2 = S_u^2 + S_v^2, \qquad (2.4) \qquad bS^2 = S_u^2 - S_v^2, \qquad (2.5)$$

whence

$$b = \frac{S_u^2 - S_v^2}{S_u^2 + S_v^2} = \frac{1 - (S_v/S_u)^2}{1 + (S_v/S_u)^2}. \qquad (2.6)$$

From (2.4), S is seen to be a sort of 'resultant standard deviation.' The last form of (2.6) shows clearly that the total possible range of b is

$$-1 \leqq b \leqq 1, \qquad \text{corresponding to} \qquad \infty \geqq S_v/S_u \geqq 0.$$

The pair of simultaneous equations (2.4) and (2.5) give

$$2S_u^2 = (1+b)S^2, \qquad (2.7) \qquad 2S_v^2 = (1-b)S^2, \qquad (2.8)$$

which will be used below in deriving (2.11).

[9] Equations (2.4) and (2.6) are respectively (14) and (13) of my 1933 paper.

With the purpose of reducing the number of parameters by 1 and of dealing with variables that are dimensionless, we shall henceforth deal with the 'reduced' modified variate $W = U + iV$ defined by the equation

$$W = w/S = u/S + iv/S = U + iV. \tag{2.9}$$

Thus we shall be directly concerned with the scatter diagram of $W = U + iV$ instead of with that of $w = u + iv$.

The distribution function $P(U,V)$ for the rectangular components U and V of any complex variate $W = U + iV$ is defined by (1.3) on setting $\rho = U$ and $\sigma = V$; thus,

$$P(U,V)dUdV = p(U<U'<U+dU, V<V'<V+dV). \tag{2.10}$$

When the given variate $z = x + iy$ is normal, so that the modified variate $w = u + iv$ is normal, as represented by (2.1), then, since S is a mere constant, the reduced modified variate $W = U + iV$ defined by (2.9) will evidently be normal also, though of course with a different distribution parameter. Its distribution function $P(U,V)$ is found to have the formula[10]

$$P(U,V) = \frac{1}{\pi\sqrt{1-b^2}} \exp\left[-\frac{U^2}{1+b} - \frac{V^2}{1-b}\right] = P(U)P(V), \tag{2.11}$$

where $P(U)$ and $P(V)$ are the component distribution functions:

$$P(U) = \frac{1}{\sqrt{\pi(1+b)}} \exp\left[-\frac{U^2}{1+b}\right], \tag{2.12}$$

$$P(V) = \frac{1}{\sqrt{\pi(1-b)}} \exp\left[-\frac{V^2}{1-b}\right]. \tag{2.13}$$

These three distribution functions each contain only one distribution parameter, namely b; moreover, the variables $U = u/S$ and $V = v/S$ are dimensionless.

The distribution function $P(R,\theta)$ for the polar components R and θ of any complex variate $W \equiv R(\cos\theta + i \sin\theta)$ is defined by (1.3) on setting $\rho = R$ and $\sigma = \theta$; thus

$$P(R,\theta)dRd\theta = p(R<R'<R+dR, \; \theta<\theta'<\theta+d\theta). \tag{2.14}$$

For the case where W is 'normal,' it is shown in Appendix A that

$$P(R,\theta) = \frac{R}{\pi\sqrt{1-b^2}} \exp\left[\frac{-R^2}{1-b^2}(1 - b\cos 2\theta)\right] \tag{2.15}$$

$$= \frac{\sqrt{L}}{\pi} \exp[-L(1 - b\cos 2\theta)], \tag{2.16}$$

[10] This formula can be obtained from (2.1) by means of (2.7), (2.8), (2.9) and (1.17) after specializing (1.17) by the substitutions $\rho = u$, $\sigma = v$ and $h = k = 1/S$. It is (16) of my 1933 paper, but was given there without proof.

where

$$L = R^2/(1-b^2). \tag{2.17}$$

In $P(R,\theta)$ it will evidently suffice to deal with values of θ in the first quadrant, because of symmetry of the scatter diagram.

The fact that $P(R,\theta)$ depends on b as a parameter when W is ‘normal’ may be indicated explicitly by employing the fuller symbol $P(R,\theta;b)$ when desired; thus the former symbol is here an abbreviation for the latter.

In $P(R,\theta) \equiv P(R,\theta;b)$ it will suffice to deal with only positive values of b, that is, with $0 \leqq b \leqq 1$ (whereas the total possible range of b is $-1 \leqq b \leqq 1$). For (2.15) shows that changing b to $-b$ has the same effect as changing 2θ to $\pi \pm 2\theta$, or θ to $\pi/2 \pm \theta$; that is, $P(R,\theta;-b) = P(R,\pi/2 \pm \theta;b)$.

Seven formulas which will find considerable use subsequently are obtainable from the integrals corresponding to equations (1.13) to (1.16), by setting $\rho = R$ and $\sigma = \theta$ or else $\rho = \theta$ and $\sigma = R$, whichever is appropriate, and thereafter substituting for $P(R,\theta)$ the expression given by (2.16), and lastly executing the indicated integrations wherever they appear possible.[11] The resulting formulas are as follows:

$$P(R \mid <\theta) = \frac{\sqrt{L}}{\pi} \exp(-L) \int_0^\theta \exp(bL \cos 2\theta)\, d\theta, \tag{2.18}$$

$$P(\theta \mid <R) = \frac{\sqrt{1-b^2}}{2\pi} \, \frac{1-\exp[-L(1-b\cos 2\theta)]}{1-b\cos 2\theta}, \tag{2.19}$$

$$P(\theta \mid >R) = \frac{\sqrt{1-b^2}}{2\pi} \, \frac{\exp[-L(1-b\cos 2\theta)]}{1-b\cos 2\theta}. \tag{2.20}$$

$$Q(<R, <\theta) = \frac{1}{\pi} \int_0^R \left[\sqrt{L} \exp(-L) \int_0^\theta \exp(bL\cos 2\theta)\, d\theta \right] dR \tag{2.21}$$

$$= \frac{\sqrt{1-b^2}}{2\pi} \int_0^\theta \frac{1-\exp[-L(1-b\cos 2\theta)]}{1-b\cos 2\theta}\, d\theta, \tag{2.22}$$

$$Q(>R, <\theta) = \frac{1}{\pi} \int_R^\infty \left[\sqrt{L} \exp(-L) \int_0^\theta \exp(bL\cos 2\theta)\, d\theta \right] dR \tag{2.23}$$

$$= \frac{\sqrt{1-b^2}}{2\pi} \exp(-L) \int_0^\theta \frac{\exp(bL\cos 2\theta)}{1-b\cos 2\theta}\, d\theta. \tag{2.24}$$

Formulas (2.21) to (2.24) are obtainable also by substituting (2.18) to (2.20) into the appropriate particular forms of (1.9).

When a θ-range of integration is 0-to-$q(\pi/2)$, where $q = 1, 2, 3$ or 4, this

[11] Except that in (2.22) the part $1/(1 - b\cos 2\theta)$ is integrable, as found in Sec. 7, equations (7.6) and (7.7).

range can be reduced to 0-to-$\pi/2$ provided the resulting integral is multiplied by q; that is,

$$\int_0^{q(\pi/2)} F(\theta)d\theta = q\int_0^{\pi/2} F(\theta)d\theta, \tag{2.25}$$

because of symmetry of the scatter diagram.

3. The Distribution Function for the Modulus

The distribution function $P(R \mid \theta_{12}) \equiv P(R)$ for the modulus R of any complex variate $W \equiv R(\cos\theta + i\sin\theta)$ is defined by equation (1.10) on setting $\rho = R$, $\sigma = \theta$, $\sigma_1 = \theta_1 = 0$ and $\sigma_2 = \theta_2 = 2\pi$; thus

$$P(R)dR = p(R<R'<R+dR,\ \ 0<\theta'<2\pi). \tag{3.1}$$

An integral formula for $P(R)$ is immediately obtainable from (1.6) by setting $\rho = R$, $\sigma = \theta$, $\sigma_3 = \sigma_1 = \theta_1 = 0$ and $\sigma_4 = \sigma_2 = \theta_2 = 2\pi$; thus

$$P(R) = \int_0^{2\pi} P(R,\theta)\,d\theta. \tag{3.2}$$

The rest of this section deals with the case where $W \equiv R(\cos\theta + i\sin\theta)$ is 'normal.' Since this case depends on b as a parameter, $P(R)$ is here an abbreviation for $P(R;b)$. A formula for $P(R;b)$ can be obtained by substituting $P(R,\theta)$ from (2.15) into (3.2) and executing the indicated integration by means of the known Bessel function formula

$$\int_0^{\pi} \exp(\eta\cos\psi)\,d\psi = \pi I_0(\eta), \tag{3.3}$$

$I_0(\ )$ being the so-called 'modified Bessel function of the first kind,' of order zero.[12] The resulting formula is found to be[13]

$$P(R;b) = \frac{2R}{\sqrt{1-b^2}}\exp\left[\frac{-R^2}{1-b^2}\right] I_0\left[\frac{bR^2}{1-b^2}\right]. \tag{3.4}$$

This can also be obtained as a particular case of the more general formula (2.18) by setting $\theta = 2\pi$ in the upper limit of integration and then applying (3.3).

In $P(R;b)$ it will suffice to deal with positive values of b, that is, with $0 \leqq b \leqq 1$, as (3.4) shows that $P(R;-b) = P(R;b)$.

[12] It may be recalled that $I_0(z) = J_0(iz)$, and in general that $I_n(z) = i^{-n}J_n(iz)$.

In the list of references on Bessel functions, on the last page of this paper, the 'modified Bessel function' is treated in Ref. 2, p. 20; Ref. 3, p. 102; Ref. 4, p. 41; Ref. 1, p. 77.

Regarding formula (3.3), see Ref. 1, p. 181, Eq. (4), $\nu = 0$; Ref. 1, p. 19, Eq. (9), fourth expression, $\nu = 0$; Ref. 2, p. 46, Eq. (10), $n = 0$; Ref. 3, p. 164, Eq. 103, $n = 0$.

[13] This formula was given in its cumulative forms, $\int P(R;\,b)dR$, as formulas (51-A) and (53-A) of the unpublished Appendix A to my 1933 paper.

It will often be advantageous to express $P_{R;\,b}$ in terms of b and one or the other of the auxiliary variables L and T defined by the equations

$$L = \frac{R^2}{1 - b^2}, \quad (3.5) \qquad\qquad T = bL = \frac{bR^2}{1 - b^2}. \quad (3.6)$$

Formula (3.4) thereby becomes, respectively,

$$P(R;b) = 2\sqrt{L}\,\exp(-L)I_0(bL), \tag{3.7}$$

$$P(R;b) = 2\sqrt{\frac{T}{b}}\,\exp\left[\frac{-T}{b}\right] I_0(T). \tag{3.8}$$

Formula (3.8) will often be preferable to (3.7) because the argument of the Bessel function in (3.8) is a single quantity, T.

Because tables of $I_0(X)$ are much less easily interpolated than tables of $M_0(X)$ defined by the equation

$$M_0(X) = \exp(-X)I_0(X), \tag{3.9}$$

extensive tables of which have been published,[14] it is natural, at least for computational purposes, to write (3.4) in the form

$$P(R;\,b) = \frac{2R}{\sqrt{1 - b^2}}\,\exp\left[\frac{-R^2}{1 + b}\right] M_0\left[\frac{bR^2}{1 - b^2}\right]. \tag{3.10}$$

For use in equation (3.16), it is convenient to define here a function $M_1(X)$ by the equation

$$M_1(X) = \exp(-X)I_1(X), \tag{3.11}$$

corresponding to (3.9) defining $M_0(X)$. $M_1(X)$ has the similar property that it is much more easily interpolated than is $I_1(X)$; and extensive tables of $M_1(X)$ are constituent parts of the tables in Ref. 1 and Ref. 6.

The quantity $bR^2/(1-b^2) \equiv T$, which occurs in (3.4) and (3.8) as the argument of $I_0(\ \ )$, and in (3.10) as the argument of $M_0(\ \ )$, evidently ranges from 0 to ∞ when R ranges from 0 to ∞ and also when b ranges from 0 to 1. Formula (3.10) is suitable for computational purposes for all values of the above-mentioned argument $bR^2/(1-b^2) \equiv T$ not exceeding the largest values of X in the above-cited tables in Ref. 1 and Ref. 6. For larger values of the argument, and partiularly for dealing with the limiting

[14] Ref. 1, Table II (p. 698–713), for $X = 0$ to 16 by .02. Ref. 6, Table VIII (p. 272–283), for $X = 5$ to 10 by .01, and 10 to 20 by 0.1. Each of these references conveniently includes a table of $\exp(X)$ whereby values of $I_0(X)$ can be readily and accurately evaluated if desired. Values of $I_0(X)$ so obtained would enable formulas (3.4), (3.7) and (3.8) of the present paper to be used with high accuracy without any difficult interpolations, since the table of $\exp(X)$ is easily interpolated by utilizing the identity $\exp(X_1 + X_2) = \exp(X_1)\exp(X_2)$.

case where the argument becomes infinite, formula (310)—and hence (3.4)—may be advantageously written in the form

$$P(R;\,b) = \frac{2}{\sqrt{2\pi b}}\exp\left[\frac{-R^2}{1+b}\right] N_0\left[\frac{bR^2}{1-b^2}\right], \tag{3.12}$$

where

$$N_0(X) = \sqrt{2\pi X}\exp(-X)I_0(X) = \sqrt{2\pi X}\,M_0(X), \tag{3.13}$$

an extensive table of which has been published.[15] The natural suitability of the function $N_0(X)$ for dealing with large values of X is evident from the structure of the asymptotic series for $N_0(X)$, for sufficiently large values of X, which runs as follows:[16]

$$N_0(X) \sim 1 + \frac{1^2}{1!8X} + \frac{1^2 3^2}{2!(8X)^2} + \frac{1^2 3^2 5^2}{3!(8X)^3} + \cdots, \tag{3.14}$$

whence it is evident that

$$N_0(\infty) = 1. \tag{3.15}$$

For use in Appendix C, it is convenient to define here a function $N_1(X)$ by the equation[17]

$$N_1(X) = \sqrt{2\pi X}\exp(-X)I_1(X) = \sqrt{2\pi X}M_1(X), \tag{3.16}$$

corresponding to (3.13) defining $N_0(X)$, with $M_1(X)$ defined by (3.11). The asymptotic series for $N_1(X)$, which will be needed in Appendix C, is[18]

$$N_1(X) \sim 1 - 3\left[\frac{1}{1!8X} + \frac{(1\cdot 5)}{2!(8X)^2} + \frac{(1\cdot 5)(3\cdot 7)}{3!(8X)^3} + \cdots\right], \tag{3.17}$$

whence it is evident that

$$N_1(\infty) = 1. \tag{3.18}$$

When b is very nearly but not exactly equal to unity, so that

$$\frac{bR^2}{1-b^2} \approx \frac{R^2}{1-b^2} \approx \frac{R^2}{2(1-b)}, \tag{3.19}$$

it is seen from (3.4) that $P(R;b)$ is, to a very close approximation, a function

[15] Ref. 7, pp. 45–72, for X = 10 to 50 by 0.1, 50 to 200 by 1, 200 to 1000 by 10, and for various larger values of X.

[16] Ref. 1, p. 203, with (ν, m) defined on p. 198; Ref. 5, p. 366; Ref. 2, p. 58; Ref. 3, p. 163, Eq. 84; Ref. 4, pp. 48, 84.

[17] $N_1(X)$ is tabulated along with $N_0(X)$ in Ref. 7 already cited in connection with equation (3.13).

[18] Ref. 1, p. 203, with (ν, m) defined on p. 198; Ref. 5, p. 366; Ref. 2, p. 58; Ref. 3, p. 163, Eq. 84.

of only a single quantity, which may be any one of the three very nearly equal expressions in (3.19)—but the last of them is evidently the simplest.

Fig. 3.1 gives curves of $P(R;b)$, with the variable R ranging continuously from 0 to 2.8 and the parameter b ranging by steps from 0 to 1 inclusive, which is the complete range of positive b. Fig. 3.2 gives an enlargement (along the R-axis) of the portion of Fig. 3.1 between $R = 0$ and $R = 0.4$,

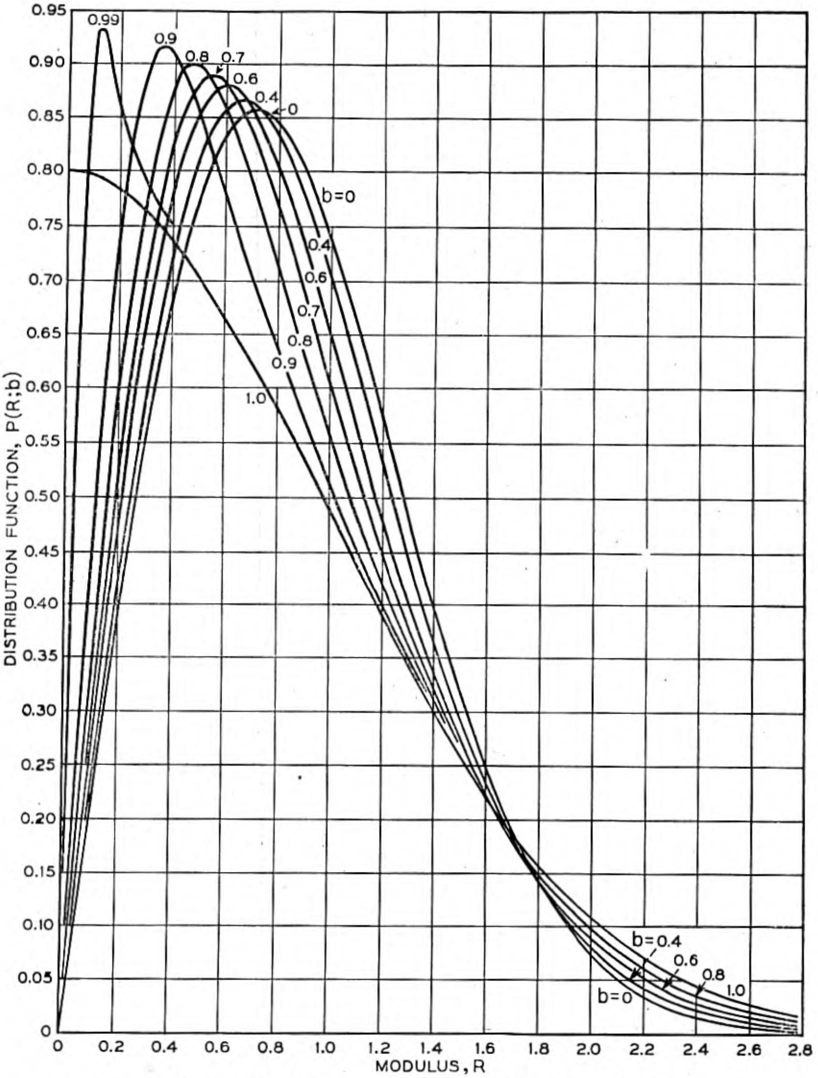


Fig. 3.1—Distribution function for the modulus (R = 0 to 2.8).

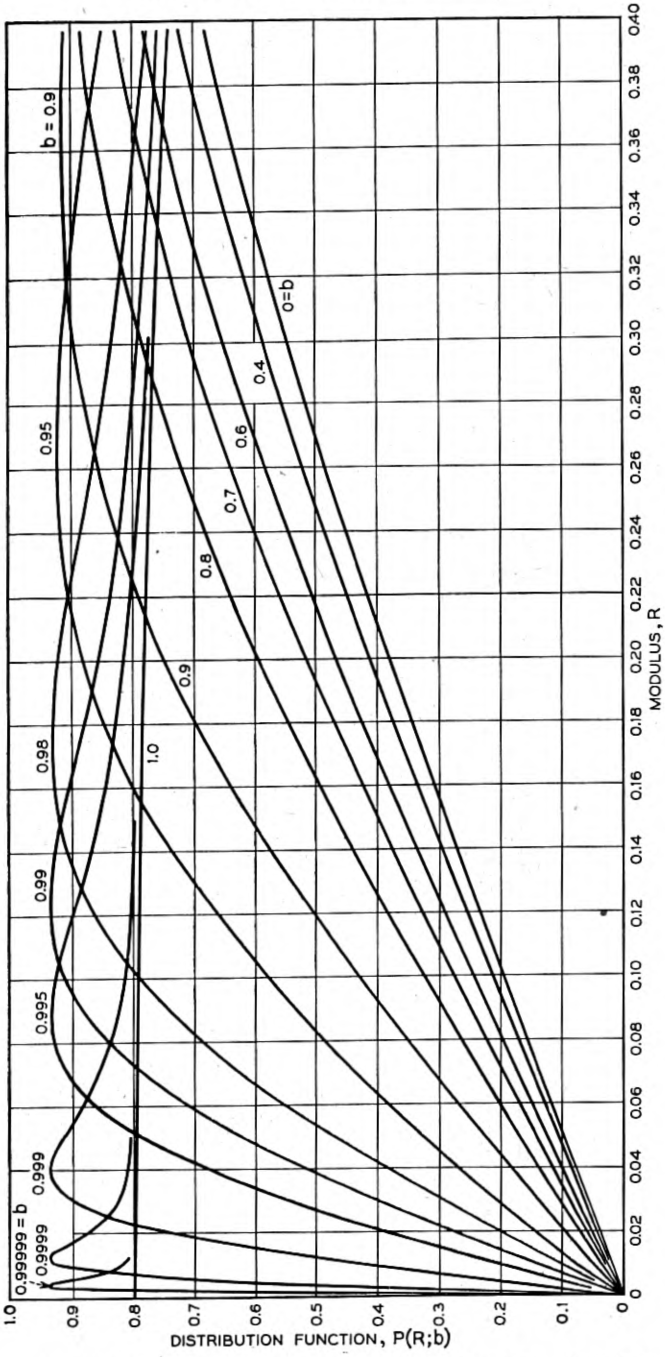


Fig. 3.2—Distribution function for the modulus ($R = 0$ to 0.4).

and includes therein curves for a considerable number of additional values of b between 0.9 and 1 so chosen as to show clearly how, with b increasing toward 1, the curves approach the curve for $b = 1$ as a limiting particular curve; or, conversely, how the curve for $b = 1$ constitutes a limiting particular curve—which, incidentally, will be found to be a natural and convenient reference curve. This curve, for $b = 1$, will be considered more fully a little further on, because it is a limiting particular curve and because of its resulting peculiarity at $R = 0$, the curve for $b = 1$ having at $R = 0$ a projection, or spur, situated in the $P(R;b)$ axis and extending from 0.7979 to 0.9376 therein (as shown a little further on).

The formulas and curves for $b = 0$ and $b = 1$, being of especial interest and importance, will be considered before the remaining curves of the set.

For the case $b = 0$, formula (3.4) evidently reduces immediately to

$$P(R;0) = 2R \exp(-R^2). \tag{3.20}$$

This case, $b = 0$, is that degenerate particular case in which the equiprobability curves in the scatter diagram of the complex variate, instead of being ellipses (concentric), are merely circles, as noted in my 1933 paper, near the bottom of p. 44 thereof (p. 10 of reprint).

For the case $b = 1$, the formula for the entire curve of $P(R;\ b) = P(R;1)$, except only the part at $R = 0$, can be obtained by merely setting $b = 1$ in[19] (3.12) as this, on account of (3.15), thereby reduces immediately to

$$P'(R;\ 1) = \frac{2}{\sqrt{2\pi}} \exp\left[-\frac{R^2}{2}\right], \qquad (R \neq 0), \tag{3.21}$$

$P'(R;1)$ denoting the value of $P(R;b)$ when $b = 1$ but $R \neq 0$, the restriction $R \neq 0$ being necessary because the quantity $R^2/(1-b^2)$ in (3.12)—and in (3.4)—does not have a definite value when $b = 1$ if $R = 0$. Thus, in Figs. 3.1 and 3.2, the curve of $P'(R;1)$ is that part of the curve for $b = 1$ which does not include any point in the $P(R;\ b)$ axis (where $R = 0$) but extends rightward from that axis toward $R = +00$. The curve of $P'(R;1)$ is the 'effective' part of the curve of $P(R;1)$, in the sense that the area under the former is equal to that under the latter, since the part of the curve of $P(R;1)$ at $R = 0$ can have no area under it.

$P(0;1)$ denoting (by convention) the value, or values, of $P(R;b)$ when $R = 0$ and $b = 1$, that is, the value, or values, of $P(R;1)$ when $R = 0$, it is seen, from consideration of the curves of $P(R;b)$ in Figs. 3.1 and 3.2 when b approaches 1 and ultimately becomes equal to 1, that the curve of $P(0;1)$ consists of all points in the vertical straight line segment extending upward in the $P(R;b)$ axis, from the origin to a height 0.9376 [= Max $P(R;1)$],[20]

[19] Use of (3.12) instead of (3.4), which is transformable into (3.12), avoids the indefinite expression $\infty .0. \infty$ which would result directly from setting $b = 1$ in (3.4).

[20] As shown near the end of Appendix B, Max $P(R;1)$ is situated at $R = 0$ and is equal to 0.9376.

together with all points in the straight line segment extending downward from the point at 0.9376 to the point at 0.7979 [$= 2/\sqrt{2\pi} = P'(R;1)$ for $R = 0+$]. The curve of $P(0; 1)$, because it has no area under it, is the 'non-effective' part of the curve of $P(R;1)$.

Starting at the origin of coordinates, where $R = 0$, the complete curve of $P(R;1)$ consists of the curve of $P(0;1)$, described in the preceding paragraph, in sequence with the curve of $P'(R;1)$, given by (3.21). Thus the complete curve of $P(R;1)$ is the locus of a tracing point moving as follows: Starting at the origin of coordinates, the tracing point first ascends in the $P(R; b)$ axis to a height 0.9376 [$=$ Max $P(R;1)$]; second, descends from 0.9376 to 0.7979 [$= 2/\sqrt{2\pi} = P'(R;1)$ for $R = 0+$]; and, third, moves rightward along the graph of $P'(R;1)$ [$b = 1$] toward $R = +\infty$. The locus of all of the points thus traversed by the tracing point is the complete curve[21] of $P(R;1)$.

In addition to being the principal part ('effective' part) of the curve of $P(R;1)$, the curve of $P'(R;1)$, whose formula is (3.21), has a further important significance. For the right side of (3.21), except for the factor 2, will be recognized as being the expression for the well-known 1-dimensional 'normal' law; the presence of the factor 2 is accounted for by the fact that the variable $R = |R|$ can have only posiive values and yet the area under the curve must be equal to unity. This case, $b = 1$, is that degenerate particular case in which the equiprobability curves, instead of being ellipses, are superposed straight line segments, so that the resulting 'probability density' is not constant but varies in accordance with the 1-dimensional 'normal' law (for real variates), as noted in my 1933 paper, at the top of p. 45 thereof (p. 11 of reprint).

All of the curves of $P(R;b)$, where $0 \leqq b \leqq 1$, pass through the origin, the curve of $P(R;1)$ [$b = 1$] being no exception, since the part $P(0;1)$ passes through the origin.

Formula (3.12), supplemented by (3.15), shows that $P(R; b) = 0$ at $R = \infty$; and this is in accord with the consideration that the total area under the curve of $P(R;b)$ must be finite (equal to unity).

Since $P(R;b) = 0$ at $R = 0$ and at $R = \infty$, every curve of $P(R;b)$ must have a maximum value situated somewhere between $R = 0$ and $R = \infty$— as confirmed by Figs. 3.1 and 3.2. These figures show that when b increases from 0 to 1 the maximum value increases throughout but the value of R where it is located decreases throughout.

The maxima of the function $P(R;b)$ and of its curves (Figs. 3.1 and 3.2) are of considerable theoretical interest and of some practical importance.

[21] The presence, in the curve of $P(R; 1)$, of the vertical projection, or spur, situated in the $P(R; b)$ axis and extending from 0.7979 to 0.9376 therein, is somewhat remindful (qualitatively) of the 'Gibbs phenomenon' in the representation of discontinuous periodic functions by Fourier series.

The cases $b = 0$ and $b = 1$ will be dealt with first, and then the general case ($b = b$).

For the case $b = 0$ it is easily found by differentiating (3.20) that $P(R; b) = P(R; 0)$ is a maximum at $R = 1/\sqrt{2} = 0.7071$ and hence that its maximum value is $\sqrt{2}\exp(-1/2) = 0.8578$, agreeing with the curve for $b = 0$ in Fig. 3.1.

For the case $b = 1$, which is a limiting particular case, the maximum value of $P(R;b) = P(R;1)$ apparently cannot be found driectly and simply, as will be realized from the preceding discussion of this case. Near the end of Appendix B, it is shown that the maximum value of $P(R;1)$ occurs at $R = 0$ (as would be expected) and is equal to 0.9376. This is the maximum value of the part $P(0;1$ of $P(R;1)$. The remaining part of $P(R;1)$, namely $P'(R;1)$, whose formula is (3.21), is seen from direct inspection of that formula to have a right-hand maximum value at $R = 0+$, whence this maximum value is $2/\sqrt{2\pi} = 0.7979$.

For the general case when b has any fixed value within its possible positive range ($0 \leqq b \leqq 1$), it is apparently not possible to obtain an explicit expression (in closed form) either for the value of R at which $P(R;b)$ has its maximum value or for the maximum value of $P(R;b)$; and hence it is not possible to make explicit computations of these quantities for use in plotting curves of them, versus b, of which they will evidently be functions. However, as shown in Appendix B, these desired curves can be exactly computed, in an indirect manner, by temporarily taking b as the dependent variable and taking T, defined by (3.6), as an intermediate independent variable. For let R_C denote the critical value of R, that is, the value of R at which $P(R;b)$ has its maximum value; and let T_C denote the corresponding value of T, whence, by (3.6),

$$T_C = bR_C^2/(1-b^2). \tag{3.22}$$

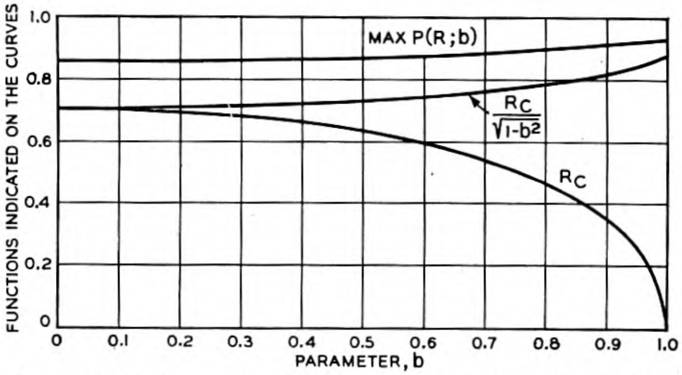


Fig. 3.3—Functions relating to the maxima of the distribution function for the modulus.

Then, computed by means of the formulas derived in Appendix B, Fig. 3.3 gives a curve of R_C and a curve of Max $P(R;b)$, each versus b. Since the curve of R_C cannot be read accurately at $b \approx 1$, there is included also a curve of $R_C/\sqrt{1-b^2}$, from which R_C can be accurately and easily computed for any value of b; incidentally, the curve of $R_C/\sqrt{1-b^2}$ is simultaneously a curve of $\sqrt{T_C/b}$, on account of (3.22). From Fig. 3.3 it is seen that R_C varies greatly with b but that Max $P_{R;b}$ varies only a little, as also is seen from inspection of Figs. 3.1 and 3.2 giving curves of $P(R;b)$ as function of R with b as parameter.

In Fig. 3.3, the curve of R_C shows that for $b = 1$ the maximum of $P(R;b)$ occurs at $R = 0$; and the curve of Max $P(R;b)$ shows that Max $P(R;1) \approx 0.94$, agreeing to two significant figures with the value 0.9376 found near the end of Appendix B.

4. The Distribution Function for the Reciprocal of the Modulus

At first, let R denote any real variate, and $P(R)$ its distribution function. Also let r denote the reciprocal of R, so that $r = 1/R$; and let $P(r)$ denote the distribution function for r. Then [22]

$$P(r) = R^2P(R) = P(R)/r^2. \tag{4.1}$$

If $P(R)$ depends on any parameters, $P(r)$ will evidently depend on the same parameters.

The rest of this section deals with the case where $W \equiv R(\cos\theta + i\sin\theta)$ is 'normal.' Since this case depends on b as a parameter, $P(R)$ and $P(r)$ are here abbreviations for $P(R;b)$ and $P(r;b)$ respectively.

As $P(R;b)$ has the distribution function given by (3.4), the distribution function for r will be

$$P(r;b) = \frac{2}{(\sqrt{1-b^2})r^3}\exp\left[\frac{-1}{(1-b^2)r^2}\right]I_0\left[\frac{b}{(1-b^2)r^2}\right], \tag{4.2}$$

obtained from the right side of (3.4) by changing R to $1/r$ and multiplying

[22] For if r and R denote any two real variates that are functionally related, say $F(r, R) = 0$, and if dr and dR are corresponding small increments, then evidently

$$P(r)\mid dr\mid = P(R)\mid dR\mid \quad \text{whence } \frac{P(r)}{P(R)} = \left|\frac{dR}{dr}\right| = \left|\frac{\partial F/\partial r}{\partial F/\partial R}\right|.$$

In particular, if $r = 1/R$, whence $F = r - 1/R$, then (4.1) results immediately.

For a somewhat different and more detailed treatment of change of the variable in distribution functions, see Thorton C. Fry, "Probability and its Engineering Uses," 1928, pp. 153–155. (Cases of more than one variate are treated on pp. 155–174 of the same reference.)

the result by $1/r^2$, in accordance with (4.1). Evidently $P(r;-b) = P(r;b)$.

By means of (4.1), formulas (3.7) and (3.8) give, respectively,

$$P(r;b) = 2(1-b^2)L^{3/2} \exp(-L)I_0(bL), \tag{4.3}$$

$$P(r;b) = 2(1-b^2)\left[\frac{T}{b}\right]^{3/2} \exp\left[\frac{-T}{b}\right] I_0(T), \tag{4.4}$$

wherein L and T are defined by (3.5) and (3.6) respectively, but will now be written in the equivalent forms

$$L = \frac{1}{(1-b^2)r^2}, \quad (4.5) \qquad\qquad T = bL = \frac{b}{(1-b^2)r^2}, \quad (4.6)$$

which are evidently more suitable for the present section.

A few particular cases that are especially important will be dealt with in the following brief paragraph, ending with equation (4.8).

For the two extreme values of r, namely 0 and ∞, $P(r;b)$ is zero for all values of b in the b-range $(0 \leqq b \leqq 1)$.

When $b = 0$,

$$P(r;b) = P(r;0) = \frac{2}{r^3}\exp\left[\frac{-1}{r^2}\right]. \tag{4.7}$$

When $b = 1$,

$$P(r;b) = P(r;1) = \frac{2}{\sqrt{2\pi}}\frac{1}{r^2}\exp\left[\frac{-1}{2r^2}\right]. \tag{4.8}$$

Fig. 4.1 gives curves of $P(r;b)$, with the variable r ranging continuously from 0 to 1.4 and the parameter b ranging by steps from 0 to 1; however, in the r-range where r is less than about 0.6, alternate curves had to be omitted to avoid undue crowding. Fig. 4.2 gives an enlargement of the section betwen $r = 0.2$ and $r = 0.5$, and includes therein the curves that had to be omitted from Fig. 4.1.

In Fig. 4.1 it will be noted that with the scale there used for $P(r;b)$ the values of $P(r;b)$ are too small to be even detectable for values of r less than about 0.25. Even in the enlargement supplied by Fig. 4.2, the values of $P(r;b)$ are not detectable for r less than about 0.2.

The curves of $P(r;b)$ in Figs. 4.1 and 4.2 would have had to be computed from the lengthy formula (4.2)—or its equivalents—except for the fact that curves of $P(R;b)$ had already been computed in the preceding section of the paper. The last circumstance enabled the $P(r;b)$ curves to be obtained from the $P(R;b)$ curves by means of the very simple relation (4.1).

It will be observed that each curve of $P(r;b)$ [Fig. 4.1] has a maximum ordinate, whose value and location depend on b. When b increases from 0 to 1, the maximum ordinate decreases throughout but the value of r where it is located remains nearly constant, at about 0.82, until b becomes about 0.7, after which the location of the maximum value moves rather rapidly to about 0.71 for $b = 1$.

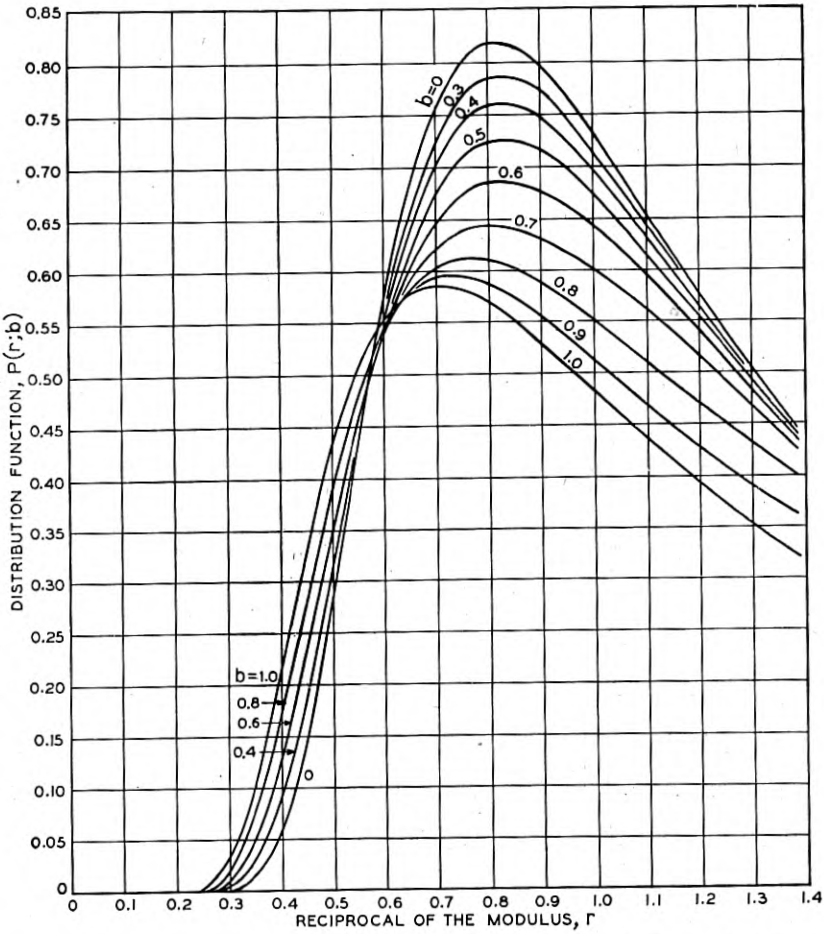


Fig. 4.1—Distribution function for the reciprocal of the modulus ($r = 0$ to 1.4).

For the cases $b = 0$ and $b = 1$, it is easily found, by differentiating (4.7) and (4.8), that the maximum ordinates are located at $r = \sqrt{2/3} = 0.8165$ and at $r = 1/\sqrt{2} = 0.7071$ respectively; and hence, by (4.7) and (4.8), that the values of these maximum ordinates are $(3\sqrt{3/2}\exp(-3/2) =$

0.8198 and $(4/\sqrt{2\pi})\exp(-1) = 0.5871$ respectively. These results for the cases $b = 0$ and $b = 1$ agree with the corresponding curves in Fig. 4.1.

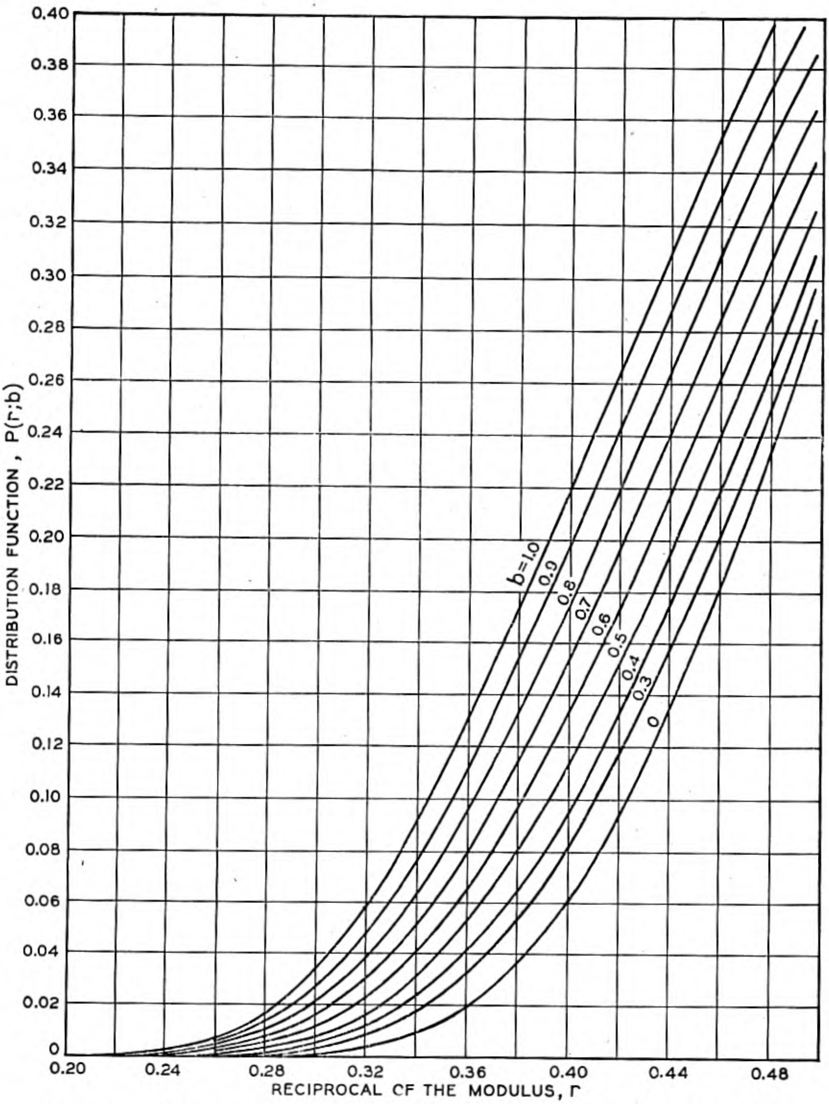


Fig. 4.2—Distribution function for the reciprocal of the modulus ($r = 0.2$ to 0.5).

For the general case where b has any fixed value in the b-range ($0 \leqq b \leqq 1$), it is apparently not possible to obtain an explicit expression (in closed form) either for the value of r at which $P(r;b)$ has its maximum value or for the maximum value of $P(r;b)$. However, as shown in Appendix C, curves of these quantities versus b can be computed, in an indirect manner, by temporarily taking b as the dependent variable and taking T, defined by (4.6), as an intermediate independent variable. For let r_c denote the critical value of r, that is, the value of r at which $P(r;b)$ has its maximum value; and let T_c denote the corresponding value of T, whence, by (4.6),

$$T_c = b/(1-b^2)r_c^2. \tag{4.9}$$

Then, computed by means of the formulas derived in Appendix C, Fig. 4.3 gives a curve of r_c and a curve of Max $P(r;b)$, each versus b. From these curves it is seen that r_c and Max $P(r;b)$ do not vary greatly with b, as also is seen from inspection of Fig. 4.1 giving curves of $P(r;b)$ as function of r with b as parameter.

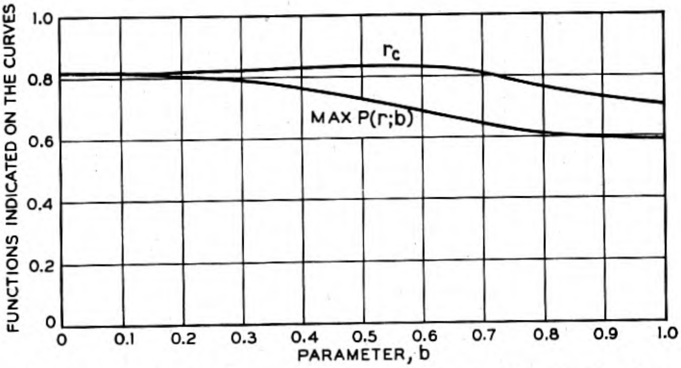


Fig. 4.3—Functions relating to the maxima of the distribution function for the reciprocal of the modulus.

5. The Cumulative Distribution Function for the Modulus

The cumulative distribution function $Q(<R,\theta_{12}) \equiv Q(R)$ for the modulus R of any complex variate $W \equiv R(\cos\theta + i\sin\theta)$ is defined by equation (1.11) on setting $\rho = R$, $\sigma = \theta$, $\rho_1 = R_1 = 0$, $\sigma_1 = \theta_1 = 0$ and $\sigma_2 = \theta_2 = 2\pi$; thus

$$Q(R) = p(0<R'<R, 0<\theta'<2\pi). \tag{5.1}$$

Similarly, from (1.12), the complementary cumulative distribution function $Q(>R,\theta_{12}) \equiv Q^*(R)$ is defined by the equation

$$Q^*(R) = p(R<R'<\infty, 0<\theta'<2\pi). \tag{5.2}$$

$Q^*(R)$ is usually more convenient than $Q(R)$ for use in engineering applications, because it is usually more convenient to deal with the relatively small probability of exceeding a preassigned rather large value of R than to deal with the corresponding rather large probability (nearly equal to unity) of being less than the preassigned value of R.

A 'double integral' for $Q(R)$, in the form of two 'repeated integrals,' can be written down directly by inspection of the $p(\ \ )$ expression in (5.1) or by specialization of (1.8); thus

$$Q(R) = \int_0^R \left[\int_0^{2\pi} P(R,\theta)\, d\theta \right] dR = \int_0^{2\pi} \left[\int_0^R P(R,\theta)\, dR \right] d\theta. \tag{5.3}$$

Evidently these can be written formally as two 'single integrals,'

$$Q(R) = \int_0^R P(R)\, dR = \int_0^{2\pi} P(\theta \mid < R)\, d\theta, \tag{5.4}$$

by means of the distribution functions $P(R) = P(R \mid \theta_{12})$ and $P(\theta \mid <R)$ given by the formulas

$$P(R) = \int_0^{2\pi} P(R,\theta)\, d\theta, \quad (5.5) \qquad P(\theta \mid < R) = \int_0^R P(R,\theta)\, dR. \quad (5.6)$$

(5.5) is the same as (3.2). (5.6) is a special case of (1.6), and the left side of (5.6) is a special case of $P(\rho \mid <\sigma)$ defined by (1.13).

Similarly, from (5.2), we arrive at the following formulas corresponding to (5.3), (5.4), (5.5), and (5.6) respectively:

$$Q^*(R) = \int_R^\infty \left[\int_0^{2\pi} P(R,\theta)\, d\theta \right] dR = \int_0^{2\pi} \left[\int_R^\infty P(R,\theta)\, dR \right] d\theta, \tag{5.7}$$

$$Q^*(R) = \int_R^\infty P(R)\, dR = \int_0^{2\pi} P(\theta \mid > R)\, d\theta, \tag{5.8}$$

$$P(R) = \int_0^{2\pi} P(R,\theta)\, d\theta, \quad (5.9) \qquad P(\theta \mid > R) = \int_R^\infty P(R,\theta)\, dR. \quad (5.10)$$

The rest of this section deals with the case where $W \equiv R(\cos\theta + i \sin\theta)$ is 'normal.'[23] Since this case depends on b as a parameter, $Q(R)$ and $Q^*(R)$ are here abbreviations for $Q(R;b)$ and $Q^*(R;b)$ respectively.

A natural and convenient way for deriving formulas for $Q(R)$ is afforded by the general formula (5.4) together with the auxiliary general formulas (5.5) and (5.6), beginning with the two latter.

For the 'normal' case, $P(R,\theta)$ is given by (2.15). When this is substituted into (5.5) and (5.6), it is found that each of the indicated integra-

[23] For the 'normal' case, the cumulative distribution function was treated in a very different manner in my 1933 paper and its unpublished Appendix A. That paper included applications to two important practical problems, and its unpublished Appendix C treated a third such problem. (The unpublished appendices, A, B and C, are mentioned in footnote 3 of the 1933 paper.)

tions can be executed, giving the two previously obtained formulas (3.4) and (2.19) for $P(R) \equiv P(R;b)$ and $P(\theta \mid <R)$ respectively. When these are substituted into (5.4), there result two types of single-integral formulas for $Q(R)$: A primary type, involving an indicated integration as to R; and a secondary type, involving an indicated integration as to θ. Formulas of these two types for $Q(R)$ will now be derived.

An integral formula of the primary type for $Q(R) \equiv Q(R;b)$ can be obtained by substituting $P(R) \equiv P(R;b)$ from (3.4) into the first integral in (5.4), giving

$$Q(R) = 2\int_0^R \frac{\lambda}{\sqrt{1-b^2}} \exp\left[\frac{-\lambda^2}{1-b^2}\right] I_0\left[\frac{b\lambda^2}{1-b^2}\right] d\lambda. \qquad (5.11)$$

This can also be obtained as a particular case of the more general formula (2.21) by setting $\theta = 2\pi$ in the upper limit of integration and then applying (3.3).

In (5.11), λ is used instead of R as the integration variable in order to avoid any possible confusion with R as an integration limit. Thus the integrand is a function of λ with b as a parameter. Evidently $Q(R;b) = Q(R;-b)$. Formula (5.11) is evidently suitable for evaluation of $Q(R)$ by numerical integration.[24]

By suitably changing the variable in (5.11), we arrive at the following various additional formulas, which, though equivalent to (5.11), are very different as regards the integrand and the limits of integration. As previously, L denotes $R^2/(1-b^2)$.

$$Q(R) = \frac{1}{\sqrt{1-b^2}} \int_0^{R^2} \exp\left[\frac{-\lambda}{1-b^2}\right] I_0\left[\frac{b\lambda}{1-b^2}\right] d\lambda, \qquad (5.12)$$

$$Q(R) = \sqrt{1-b^2}\int_0^L \exp(-\lambda)\, I_0(b\lambda)\, d\lambda, \qquad (5.13)$$

$$Q(R) = L\sqrt{1-b^2}\int_0^1 \exp(-L\lambda)\, I_0(bL\lambda)\, d\lambda, \qquad (5.14)$$

$$Q(R) = \sqrt{1-b^2}\int_{\exp(-L)}^1 I_0(b \log \lambda)\, d\lambda. \qquad (5.15)$$

These four additional formulas are of some theoretical interest, but apparently they are less suitable than (5.11) for numerical integration with respect to R. A formula differing slightly from (5.11) could evidently be obtained by taking $\lambda/\sqrt{1-b^2}$ as a new variable, and hence $R/\sqrt{1-b^2}$ as the upper limit of integration.

Corresponding formulas for $Q^*(R) \equiv Q^*(R;b)$ can of course be obtained from the preceding formulas (5.11) to (5.15) inclusive for $Q(R) \equiv Q(R;b)$

[24] In this connection, Appendix D may be of interest.

by merely changing the integration limits correspondingly—for instance, in (5.11), from 0, R to R, ∞; in (5.13), from 0, L to L, ∞; and so on. However, the first four formulas for $Q^*(R)$ so obtained would suffer the disadvantage of each having an infinite limit of integration, rendering those formulas unsatisfactory for numerical integration purposes. This difficulty can be avoided by making the substitution $R = 1/r$ in each of those formulas for $Q^*(R)$. The resulting formulas are the following five, corresponding to (5.11) to (5.15) respectively:[24]

$$Q^*(R) = \frac{2}{\sqrt{1-b^2}} \int_0^r \frac{1}{\lambda^3} \exp\left[\frac{-1/\lambda^2}{1-b^2}\right] I_0\left[\frac{b/\lambda^2}{1-b^2}\right] d\lambda, \tag{5.16}$$

$$Q^*(R) = \frac{1}{\sqrt{1-b^2}} \int_0^{r^2} \frac{1}{\lambda^2} \exp\left[\frac{-1/\lambda}{1-b^2}\right] I_0\left[\frac{b/\lambda}{1-b^2}\right] d\lambda, \tag{5.17}$$

$$Q^*(R) = \sqrt{1-b^2} \int_0^{1/L} \frac{1}{\lambda^2} \exp\left[-\frac{1}{\lambda}\right] I_0\left[\frac{b}{\lambda}\right] d\lambda, \tag{5.18}$$

$$Q^*(R) = L\sqrt{1-b^2} \int_0^{1} \frac{1}{\lambda^2} \exp\left[-\frac{L}{\lambda}\right] I_0\left[\frac{bL}{\lambda}\right] d\lambda, \tag{5.19}$$

$$Q^*(R) = \sqrt{1-b^2} \int_0^{\exp(-L)} I_0(b \log \lambda)\, d\lambda. \tag{5.20}$$

As a check on (5.16), it is obtainable from (4.2) by integrating the latter as to r.

For purposes of evaluation by numerical integration, formulas (5.11) to (5.15) inclusive may evidently differ greatly as regards the amount of labor involved and the numerical precision practically attainable. In each of these formulas except (5.14) the integrand contains only one parameter, b, while the integration range involves either R or $L \equiv R^2/(1-b^2)$. In (5.14) the integrand contains two independent parameters, b and L, while the integration range is a mere constant, 0-to-1. Similar statements apply to formulas (5.16) to (5.20) inclusive.

A partial check on any formula for $Q(R)$ can be applied by setting $R = \infty$, since $Q(\infty)$ should be equal to unity (representing certainty). If, for instance, this procedure is applied to formula (5.13), the right side is found to reduce to unity by aid of the known relation[25]

$$\int_0^\infty \exp\,(-A\lambda)\, J_0(B\lambda)\, d\lambda = \frac{1}{\sqrt{A^2+B^2}} \tag{5.21}$$

together with $I_0(B\lambda) = J_0(iB\lambda)$.

An integral formula of the secondary type for $Q^*(R) \equiv Q^*(R;b)$ can be obtained by substituting (2.20) into the last integral in (5.8), utilizing (2.25),

[25] Ref. 1, p. 384, Eq. (1); Ref. 2, p. 65, Eq. (2); Ref. 4, p. 58, Eq. (4.5).

changing the variable of integration by the substitution $6 = \phi/2$, and rearranging; thus it is found that[26]

$$Q^*(R) = \frac{\sqrt{1 - b^2}}{\pi \exp L} \int_0^\pi \frac{\exp(bL \cos \phi)}{1 - b \cos \phi} \, d\phi. \tag{5.22}$$

This formula can also be obtained as a particular case of the more general formula (2.24) by setting $\theta = 2\pi$ in the upper limit of integration, utilizing (2.25), and changing the variable of integration by the substitution $\theta = \phi/2$.

Two partial checks on any general formula for $Q(R) \equiv Q(R;b)$ or for $Q^*(R) \equiv Q^*(R;b)$ can be applied by setting $b = 0$ and $b = 1$, and comparing the resulting particular formulas with those obtained by integrating the formulas for $P(R;0)$ and $P'(R;1)$ obtained in Section 3, namely formulas (3.20) and (3.21) there. It is thus found that

$$Q^*(R; 0) = \exp(-R^2) = \int_R^\infty P(R; 0) \, dR, \tag{5.23}$$

$$Q(R; 1) = 2\left\{\frac{1}{\sqrt{2\pi}} \int_0^R \exp\left[-\frac{R^2}{2}\right] dR\right\} = \int_0^R P'(R; 1) \, dR. \tag{5.24}$$

It will be recalled that the quantity between braces in (5.24) is extensively tabulated, and that it is sometimes called the 'normal probability integral.'

Several of the above general formulas for $Q(R) \equiv p(R' < R)$ and for $Q^*(R) \equiv p(R' > R)$ are closely connected with my 1933 paper.[27] Indeed, formulas (5.11), (5.14), (5.16), (5.19) and (5.22) above are the same as (53-A), (56-A), (52-A), (55-A) and (22-A), respectively, of the unpublished Appendix A to the 1933 paper; and (5.12), (5.13), (5.15), (5.17), (5.18) and (5.20) above were derived in the same connection, although they were not included in the Appendix A.

Formula (5.22) was employed in the unpublished Appendix A of the 1933 paper, being (22-A) there, as a basis for deriving two very different kinds of series type formulas for computing the values of $p(R' > R) \equiv Q^*(R)$ underlying the values of $p_{b,0}(R' > R)$ constituting Table I (facing Fig. 8) in that paper.[28]

[26] This formula, (5.22), was derived by me in a somewhat different manner in the unpublished Appendix A to my 1933 paper. Later I found that an equivalent formula, easily transformable into (5.22), had been given by Bravais as formula (51) in his classical paper "Analyse mathématique sur les probabilités des erreurs de situation d'un point," published in Mémoires de l'Académie Royale des Sciences de l'Institut de France, 2nd series, vol. IX, 1846, pp. 255–332. (This is available in the Public Library of New York City, for instance.)

[27] There the abbreviated symbols $p(R' < R)$ and $p(R' > R)$ were used with the same meanings as the complete symbols on the right sides of equations (5.1) and (5.2), respectively, of the present paper.

[28] Each of the two kinds of series type formulas comprised a finite portion of a convergent series plus an exact remainder term consisting of a definite integral. In the

In the present paper, formulas (5.11) and (5.16) have been used for numerical evaluation of $Q(R) \equiv p(R'<R)$ and of $Q^*(R) \equiv p(R'>R)$ by numerical integration (employing 'Simpson's one-third rule'), aided by some of the considerations set forth in Appendix D. However, only a moderate number of values of these quantities have been thus evaluated—merely enough to afford a fairly comprehensive check on Table I of my 1933 paper, by means of a sample consisting of 60 values (about 26%) distributed in a somewhat representative manner over that table. These new values of $Q^*(R) \equiv p(R'>R) = 1 - Q(R)$ are presented in Table 5.1 (at the end of this section) in such a way as to facilitate comparison with the old values, namely those in the 1933 paper. Thus, for any fixed value of R in Table 5.1, there are two horizontal rows of computed values of $Q^*(R)$, the first row (top row) coming from the 1933 paper, and the second row coming from the present paper. The third row of each set of four rows gives the deviations of the second row from the first row; and the fourth row expresses these deviations as percentages of the values in the first row.

In the first row of any set of four rows, any value represents $Q^*(R) \equiv p_b(R'>R)$ obtained, in accordance with Eq. (22) of my 1933 paper, by adding $\exp(-R^2)$ to $p_{b,0}(R'>R)$ given in Table I there. In the second row of a set, any value represents $Q^*(R) = 1 - Q(R)$ as computed by formula (5.11) or (5.16) of the present paper: more specifically, the values for $R = 0.2$, 0.4, 0.6 and 0.8 were computed by (5.11); and the values for $R = 1.6$ and $R = 2$ by (5.16), taking $r = 1/1.6 = 0.625$ and $r = 1/2 = 0.5$ respectively.[29]

In the 1933 paper, the values of $p_b(R'>R) \equiv Q^*(R;b)$ for $b = 0$ and for $b = 1$ were omitted as being unnecessary there because their values could be easily obtained from the simple exact formulas to which the general formulas there reduced, for $b = 0$ and $b = 1$. Those reduced formulas were the same as (5.23) and (5.24) of the present paper, except that (5.24) gives $Q(R;1)$ instead of giving $Q^*(R;1) = 1 - Q(R;1)$. The values obtained from these two formulas, exact to the number of significant figures here retained, are given in Table 5.1 at the intersections of the first row of each set of four rows with the columns $b = 0$ and $b = 1$. Therefore in these two columns the deviations (in the third row of each set of four rows) are deviations from exact values; the values in the second row of each set are, as

use of such a formula for numerical computations, the expansion producing the convergent series was carried far enough to insure that the remainder definite integral would be relatively small, though usually not negligible; and then this remainder definite integral was evaluated sufficiently accurately by numerical integration.

[29] In the work of numerical integration, 'Simpson's one-third rule' was employed for $R = 0.2, 0.4, 0.6, 0.8$ and 2. For $R = 1.6$, so that $r = 1/1.6 = 0.625$, 'Simpson's one-third rule' was employed up to $r = 0.620$, and the 'trapezoidal rule' from $r = 0.620$ to $r = 0.625$.

already stated, those obtained by the methods of the present paper, employing numerical integration.

From detailed inspection of Table 5.1 it will presumably be considered that the agreement between the two sets of values of $Q^*(R;b) \equiv p_b(R'>R)$ is to be regarded as satisfactory, at least from the practical viewpoint, the largest deviation being less than one per cent (for $R = 0.8$, $b = 0.9$).

TABLE 5.1
VALUES OF $Q^*(R) \equiv p(R' > R)$

b....... R	0	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	1.00
0.2	.9608	.9590	.9574	.9550	.9516	.9463	.9372	.9168	.8930	.84148
"	.9623	.9605	.9590	.9567	.9528	.9473	.9387	.9206	.8925	.84124
"	.0015	.0015	.0016	.0017	.0012	.0010	.0015	.0038	−.0005	−.00024
"	.16	.16	.17	.18	.13	.11	.16	.41	−.06	−.03
0.4	.8521	.8462	.8410	.8335	.8228	.8071	.7830	.7420	.7127	.68916
"	.8537	.8477	.8427	.8351	.8240	.8081	.7841	.7459	.7125	.68897
"	.0016	.0015	.0017	.0016	.0012	.0010	.0011	.0039	−.0002	−.00019
"	.19	.18	.20	.19	.15	.12	.14	.53	−.03	−.03
0.6	.6977	.6880	.6799	.6686	.6531	.6324	.6055	.5721	.5578	.54851
"	.6992	.6892	.6814	.6698	.6540	.6334	.6065	.5764	.5572	.54831
"	.0015	.0012	.0015	.0012	.0009	.0010	.0010	.0043	−.0006	−.00020
"	.22	.17	.22	.18	.14	.16	.17	.75	−.11	−.04
0.8	.5273	.5167	.5081	.4969	.4826	.4656	.4477	.4316	.4261	.42371
"	.5290	.5183	.5099	.4982	.4840	.4672	.4488	.4357	.4266	.42355
"	.0017	.0016	.0018	.0013	.0014	.0016	.0011	.0041	.0005	−.00016
"	.32	.31	.35	.26	.29	.34	.25	.95	.12	−.04
1.6	.07730	.07986	.08207	.08522	.0891	.0938	.0990	.1042	.1070	.10960
"	.07727	.07988	.08210	.08536	.0892	.0938	.0989	.1042	.1069	.10958
"	−.00003	.00002	.00003	.00014	.0001	.0000	−.0001	.0000	−.0001	−.00002
"	−.04	.03	.04	.16	.11	.00	−.10	.00	−.09	−.02
2.0	.01832	.02153	.02394	.02681	.0301	.0337	.0375	.0414	.0435	.04550
"	.01823	.02145	.02383	.02685	.0302	.0338	.0376	.0415	.0436	.04552
"	−.00009	−.00008	−.00011	.00004	.0001	.0001	.0001	.0001	.0001	.00002
"	−.49	−.37	−.46	.15	.33	.30	.27	.24	.23	.04

6. THE DISTRIBUTION FUNCTION FOR THE ANGLE

The distribution function $P(\theta \mid R_{12}) \equiv P(\theta)$ for the angle θ of any complex variate $W \equiv R(\cos\theta + i\sin\theta)$ is defined by equation (1.10) on setting $\rho = \theta$, $\sigma = R$, $\sigma_1 = R_1 = 0$ and $\sigma_2 = R_2 = \infty$; thus

$$P(\theta)d\theta = p(\theta<\theta'<\theta+d\theta, 0<R'<\infty). \tag{6.1}$$

An integral formula for $P(\theta)$ is immediately obtainable from (1.6) by setting $\rho = \theta$, $\sigma = R$, $\sigma_3 = \sigma_1 = R_1 = 0$ and $\sigma_4 = \sigma_2 = R_2 = \infty$; thus

$$P(\theta) = \int_0^\infty P(R, \theta)\, dR. \tag{6.2}$$

The rest of this section deals with the case where $W \equiv R(\cos\theta + i\sin\theta)$ is 'normal.' Since this case depends on b as a parameter, $P(\theta)$ is here an abbreviation for $P(\theta;b)$.

A formula for $P(\theta;b) \equiv P(\theta)$ can be obtained by substituting $P(R,\theta)$ from (2.15) into (6.2) and executing the indicated integration, which can be easily accomplished. The resulting formula is found to be

$$P(\theta;\, b) = \frac{\sqrt{1 - b^2}}{2\pi(1 - b\cos 2\theta)}. \tag{6.3}$$

This formula can also be obtained as a particular case of either of the more general formulas (2.19) and (2.20) by setting $R = \infty$ in (2.19) or $R = 0$ in (2.20); also by adding (2.19) to (2.20) and then utilizing (1.10).

In $P(\theta) \equiv P(\theta;b)$ it will evidently suffice to deal with values of θ in the first quadrant, because of symmetry of the scatter diagram.

In $P(\theta;b)$ it will suffice to deal with only positive values of b, as (6.3) shows that changing b to $-b$ has the same effect as changing 2θ to $\pi\pm2\theta$, or θ to $\pi/2\pm\theta$; that is, $P(\theta;-b) = P(\pi/2\pm\theta;b)$.

Fig. 6.1 gives curves of $P(\theta;b)$, computed from (6.3), as function of θ with b as parameter, for the ranges[30] $0\leqq\theta\leqq90°$ and $0\leqq b\leqq1$.

The curves in Fig. 6.1 indicate that $P(\theta;b)$ is a maximum at $\theta = 0°$ and a minimum at $\theta = 90°$. These indications are verified by formula (6.3), as this formula shows that:

$$\text{Max } P(\theta;b) = P(0°;b) = \frac{1}{2\pi}\sqrt{\frac{1+b}{1-b}}, \tag{6.4}$$

$$\text{Min } P(\theta;b) = P(90°;b) = \frac{1}{2\pi}\sqrt{\frac{1-b}{1+b}}. \tag{6.5}$$

Thence

$$\text{Min } P(\theta;b)/\text{Max } P(\theta;b) = (1-b)/(1+b), \tag{6.6}$$

$$P(\theta;b)/\text{Max } P(\theta;b) = P(\theta;b)/P(0°;b) = (1-b)/(1-b\cos 2\theta). \tag{6.7}$$

The curves in Fig. 6.1 indicate also that $P(\theta;b)$ is independent of θ in the case $b = 0$. This is verified by formula (6.3), as this formula shows that

$$P(\theta;0) = 1/2\pi. \tag{6.8}$$

Thence (6.3) can be written

$$P(\theta;b)/P(\theta;0) = (\sqrt{1 - b^2})/(1-b\cos 2\theta). \tag{6.9}$$

[30] Beginning here, θ will usually be expressed in degrees instead of radians, for practical convenience.

By setting $\cos 2\theta = 0$ in (6.3), so that $\theta = 45°$, it is found that

$$(\sqrt{1 - b^2})/2\pi = P(45°;b), \tag{6.10}$$

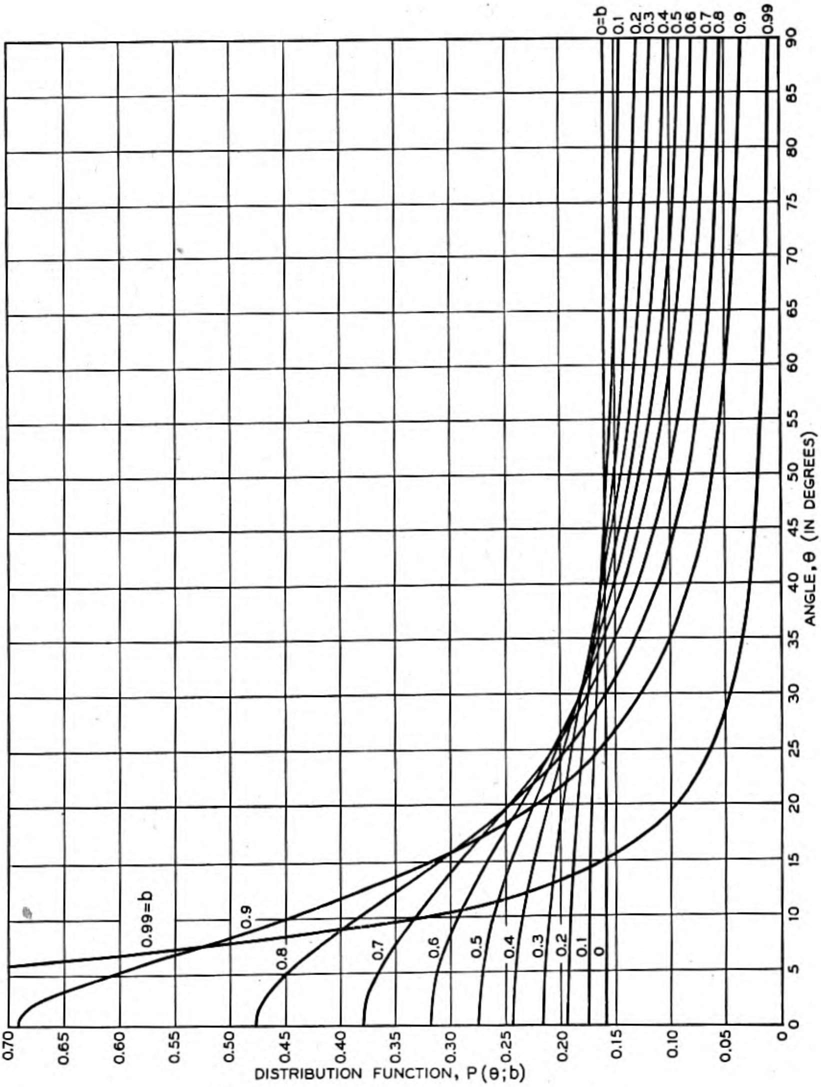


Fig. 6.1—Distribution function for the angle.

whence (6.3) can be written

$$P(\theta;b)/P(45°;b) = 1/(1-b\cos 2\theta). \tag{6.11}$$

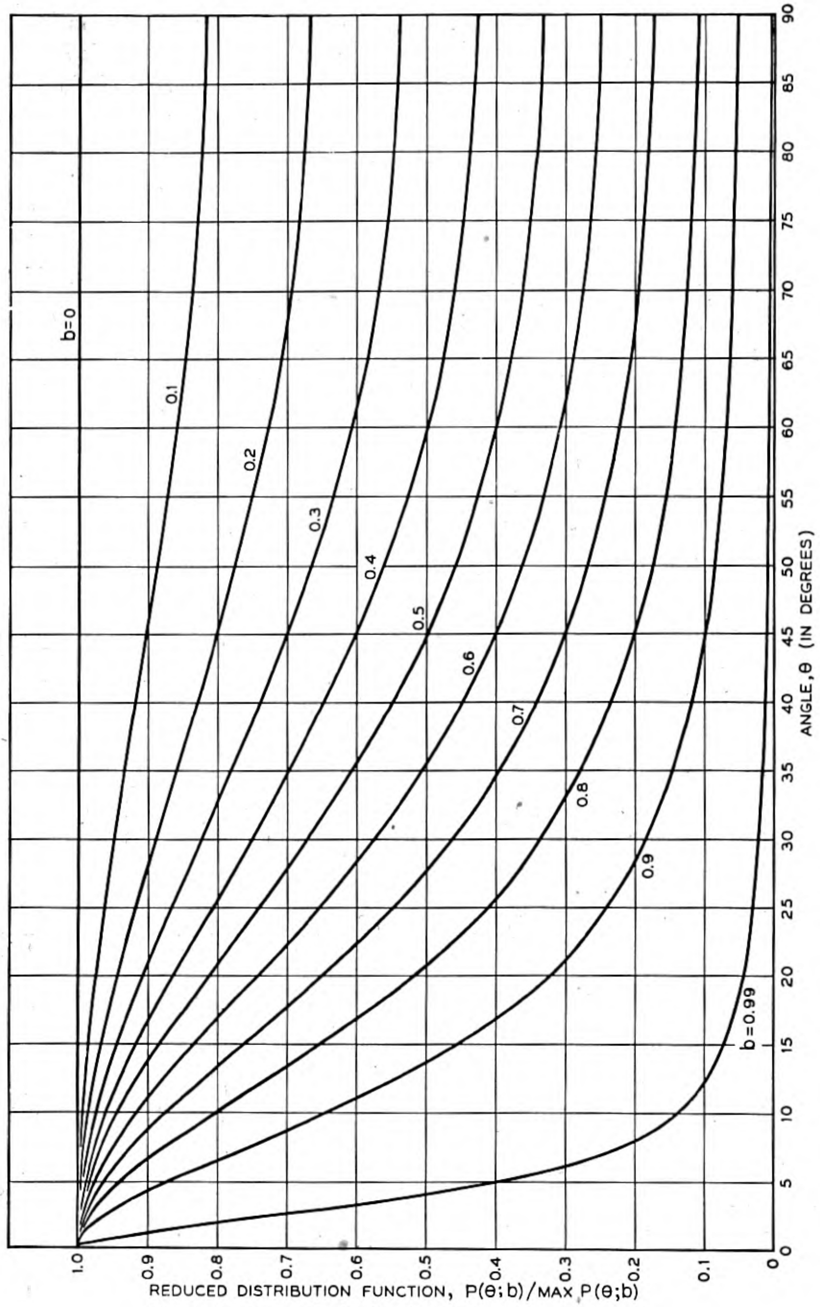


Fig. 6.2—Reduced distribution function for the angle.

In the case $b = 1$, the curves in Fig. 6.1 suggest, by limiting considerations, that $P(\theta;1)$ is zero for all θ except $\theta = 0°$, and that $P(\theta;1)$ is infinite for $\theta = 0°$. These conclusions are verified by formula (6.3), as this formula shows that:

$$P(\theta;1) = 0 \text{ for } 0°<\theta<180°;\ P(\theta;1) = \infty \text{ for } \theta = 0°,\ 180°.$$

The curves in Fig. 6.1, though having the advantage of directly representing $P(\theta;b)$ as function of θ with b as parameter, are somewhat troublesome to use because of their numerous crossings of each other. This difficulty is not present in Fig. 6.2, which gives curves of $P(\theta;b)/\text{Max } P(\theta;b)$, obtained by dividing the ordinates $P(\theta;b)$ of the curves in Fig. 6.1 by the respective maximum ordinates of those curves, as given by (6.4), so that the equation of the curves in Fig. 6.2 is formula (6.7).

7. The Cumulative Distribution Function for the Angle

The cumulative distribution function $Q(<\theta,R_{12}) \equiv Q(\theta)$ for the angle θ of any complex variate $W \equiv R(\cos\theta + i\sin\theta)$ is defined by equation (1.11) on setting $\rho = \theta$, $\sigma = R$, $\rho_1 = \theta_1 = 0$, $\sigma_1 = R_1 = 0$ and $\sigma_2 = R_2 = \infty$; thus

$$Q(\theta) = p(0<\theta'<\theta, 0<R'<\infty). \tag{7.1}$$

A 'double integral' for $Q(\theta)$, in the form of two 'repeated integrals,' can be written down directly by inspection of the $p(\ \ )$ expression in (7.1) or by specialization of (1.8); thus

$$Q(\theta) = \int_0^\theta \left[\int_0^\infty P(R, \theta)\, dR\right] d\theta = \int_0^\infty \left[\int_0^\theta P(R, \theta)\, d\theta\right] dR. \tag{7.2}$$

Evidently these can be written formally as two 'single integrals,'

$$Q(\theta) = \int_0^\theta P(\theta)\, d\theta = \int_0^\infty P(R \mid < \theta)\, dR, \tag{7.3}$$

by means of the distribution functions $P(\theta) \equiv P(\theta \mid R_{12})$ and $P(R \mid <\theta)$ given by the formulas

$$P(\theta) = \int_0^\infty P(R, \theta)\, dR, \quad (7.4) \qquad P(R \mid < \theta) = \int_0^\theta P(R, \theta)\, d\theta. \quad (7.5)$$

(7.4) is the same as (6.2). (7.5) is a special case of (1.6), and the left side of (7.5) is a special case of $P(\rho \mid <\sigma)$ defined by (1.13).

The rest of this section deals with the case where $W \equiv R(\cos\theta + i\sin\theta)$ is 'normal.' Since this case depends on b as a parameter, $Q(\theta)$ is here an abbreviation for $Q(\theta;b)$.

A natural and convenient way for deriving formulas for $Q(\theta)$ is afforded by the general formula (7.3) together with the auxiliary general formulas (7.4) and (7.5), beginning with the two latter.

It will be convenient to dispose of (7.5) before dealing with (7.4), as (7.5) turns out to be the less useful. For when $P(R,\theta)$ given by (2.16) is substituted into (7.5), the indicated integration cannot be executed in general, as (7.5) becomes (2.18), wherin the indicated integration can be executed only for certain special values of the integration limit θ—by means of the special Bessel function formula (3.3).

When $P(R,\theta)$ given by (2.15), which is equivalent to (2.16) used above, is substituted into (7.4), it is found that the indicated integration can be executed, giving the previously obtained formula (6.3) for $P(\theta) \equiv P(\theta;b)$.

A θ-integral formula for $Q(\theta) \equiv Q(\theta;b)$ can be obtained by substituting $P(\theta) \equiv P(\theta;b)$ from (6.3) into the first integral in (7.3), giving

$$Q(\theta;\,b) = \frac{\sqrt{1-b^2}}{2\pi}\int_0^\theta \frac{d\theta}{1-b\cos 2\theta} = \frac{\sqrt{1-b^2}}{4\pi}\int_0^{2\theta}\frac{d\phi}{1-b\cos\phi}. \quad (7.6)$$

This formula can also be obtained as a particular case of the more general formulas (2.22) and (2.24) by setting $R = \infty$ in (2.22) or $R = 0$ in (2.24); also by adding (2.22) to (2.24) and then utilizing (1.11).

The integral in (7.6) is of well-known form, and the indicated integration can be executed, yielding the following two equivalent formulas for $Q(\theta;b)$:

$$Q(\theta;\,b) = \frac{1}{2\pi}\left|\tan^{-1}\left[\sqrt{\frac{1+b}{1-b}}\tan\theta\right]\right| = \frac{1}{4\pi}\left|\cos^{-1}\left[\frac{\cos 2\theta - b}{1-b\cos 2\theta}\right]\right|. \quad (7.7)$$

In $Q(\theta;b)$ it will evidently suffice to deal with values of θ in the first quadrant, because of symmetry of the scatter diagram, and the resulting fact that $Q(n\,90°) = n/4$, where $n = 1, 2, 3$ or 4.

In $Q(\theta;b)$ it will suffice to deal with positive values of b, as (7.7) shows that[31]

$$Q(\theta;\,-b) = \left|\frac{1}{4} - Q\left(\frac{\pi}{2}\pm\theta;\,b\right)\right|.$$

Fig. 7.1 gives curves of $Q(\theta;b) \equiv Q(\theta)$ computed from (7.7), as function of θ with b as parameter, for the ranges $0\leqq\theta\leqq 90°$ and $0\leqq b\leqq 1$.

Consideration of the scatter diagram of W or of its equiprobability curves, which are concentric similar ellipses, affords several partial checks on the curves in Fig. 7.1 and on formula (7.7) from which they were plotted.

[31] This relation can also be derived geometrically from the fact that the scatter diagram for $-b$ is obtainable by merely rotating that for b through 90°, as shown by (2.6), or (2.7) and (2.8), or (2.11).

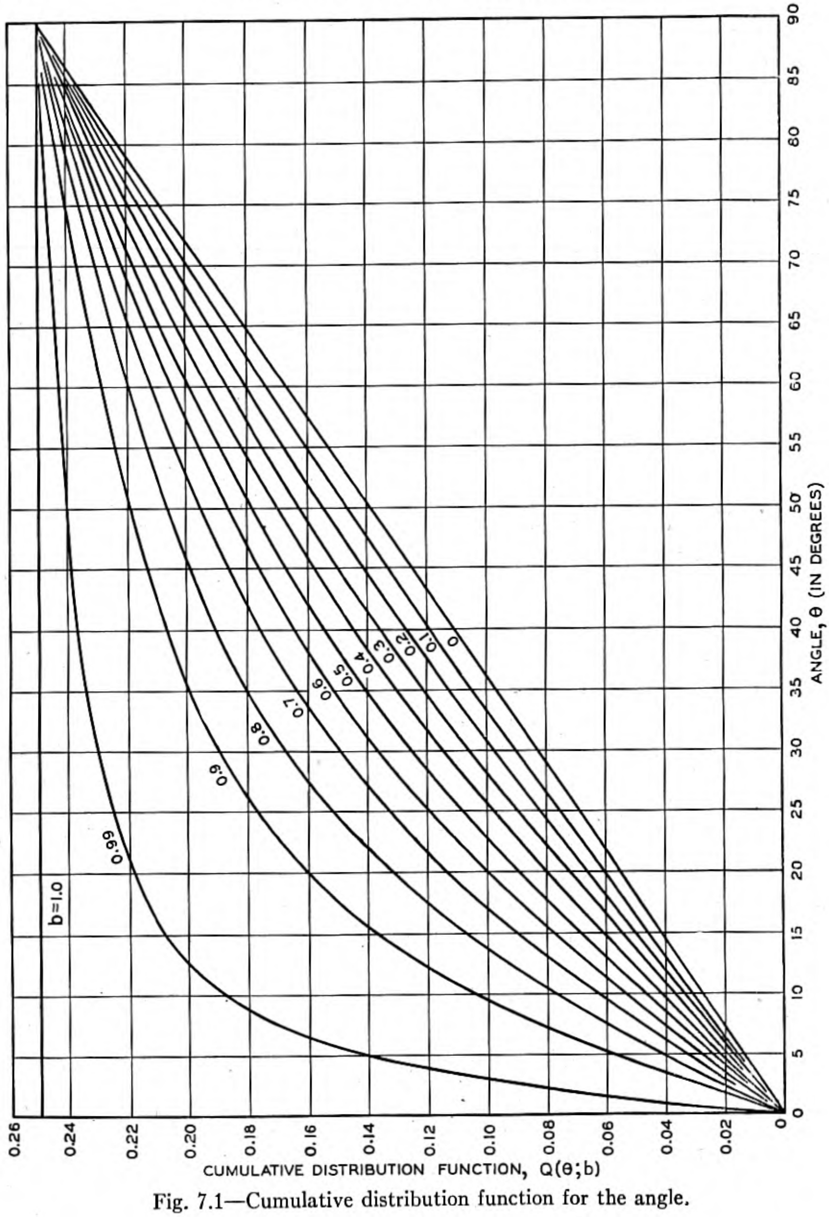


Fig. 7.1—Cumulative distribution function for the angle.

Thus, the fact that the curve for $b = 0$ is a straight line, whose equation is

$$Q(\theta;0) = \theta/2\pi = \theta^\circ/360^\circ, \qquad (b = 0),$$

corresponds to the fact that for $b = 0$ the equiprobability curves are circles.

The fact that the curve for $b = 1$ is the straight line $Q(\theta;1) = 1/4 = 0.25$ corresponds to the fact that for $b = 1$ the scatter diagram has degenerated to be merely a straight line coinciding with the real axis, so that no point outside of this line makes any contribution to $Q(\theta;1)$.

The fact that, at $\theta = 90°$, $Q(\theta;b) = Q(90°;b)$ has for all b the value $1/4 = 0.25$ corresponds to the fact that the area of a quadrant of the scatter diagram is one-fourth the area of the entire scatter diagram. Hence $Q(360°;b) = 4Q(90°;b) = 1$, which is evidently correct.

Acknowledgment

The computations and curve-plotting for this paper were done by Miss M. Darville; those for the 1933 paper, by Miss D. T. Angell.

APPENDIX A

Derivation of Formula (2.15) for $P(R,\theta)$

(2.15) will here be derived from (2.11) by utilizing the fact that the 'areal probability density', G, at any fixed point in the scatter diagram must be independent of the system of coordinates; for $G\,dA$ gives the probability of falling in any differential element of area dA, and this probability must evidently be independent of the shape of dA (assuming that all linear dimensions of dA are differential, of course). Thus, indicating the element of area by an underline, we have, in rectangular coordinates,

$$G\underline{dUdV} = P(U,V)dUdV, \qquad \text{(A1)} \qquad \text{whence} \quad G = P(U,V). \qquad \text{(A2)}$$

In polar coordinates,

$$G\underline{Rd\theta dR} = P(R,\theta)dRd\theta, \qquad \text{(A3)} \qquad \text{whence} \quad G = P(R,\theta)/R. \qquad \text{(A4)}$$

Comparing these two expressions for G shows that[32]

$$P(R,\theta) = RP(U,V). \qquad \text{(A5)}$$

Thus, a formula for $P(R,\theta)$ can be obtained from (2.11) by merelv multiplying both sides of that formula by R. However, in the resulting formula it will remain to express U and V in terms of R and θ, by means of the relations

$$U = R\cos\theta, \qquad \text{(A6)} \qquad V = R\sin\theta. \qquad \text{(A7)}$$

The final result, after a simple reduction, is (2.15), which is thus proved.

APPENDIX B

Formulas of the Curves in Fig. 3.3

As in equation (3.22), R_C will here denote the critical value of R, that is, the value of R at which $P(R) \equiv P(R;b)$ has its maximum value; and T_C

[32] Formula (A5) can be easily verified by the entirely different method which utilizes (1.23).

will denote the corresponding value of T, whence T_C is given in terms of R_C and b by (3.22).

A formula for $dP(R)/dR$ could of course be obtained directly from (3.4) but it will be found preferable to obtain it indirectly from the less cumbersome formula (3.8) containing the auxiliary variable T defined by (3.6). Evidently, since b does not depend on R,

$$\frac{dP(R)}{dR} = \frac{dP(R)}{dT}\frac{dT}{dR} = \frac{2bR}{1-b^2}\frac{dP(R)}{dT}. \tag{B1}$$

Thus, since the factor $2bR/(1-b^2)$ cannot vanish for any value of R (except $R = 0$), the only critical value of R must be that corresponding to the value of T at which $dP(R)/dT$ vanishes, namely T_C, since T_C has been defined to be the value of T corresponding to R_C. (Incidentally, equation (B1) shows that T_C is equal to the value of T at which $P(R)$ is an extremum when $P(R)$ is regarded as a function of T.) From (3.22),

$$\frac{R_C^2}{1-b^2} = \frac{T_C}{b}. \tag{B2}$$

Evidently T_C and R_C must ultimately be functions of only b. The next paragraph deals with T_C, which evidently has to be known before R_C can be evaluated.

From (3.8) it is found that, since $dI_0(T)/dT = I_1(T)$,

$$\frac{dP(R)}{dT} = P(R)\left[\frac{1}{2T} + \frac{I_1(T)}{I_0(T)} - \frac{1}{b}\right]. \tag{B3}$$

Hence, since $P(R)$ does not vanish for any value of R (except $R = 0$ and $R = \infty$), T_C will be a root of the conditional equation obtained by equating to zero the expression in brackets in (B3). This conditional equation is transcendental in T_C and apparently has no closed form of explicit solution for T_C; and its solution by successive approximation, or otherwise, would likely be rather slow and laborious. However, the bracket expression in (B3) shows that b can be immediately expressed explicitly in terms of T_C by the equation

$$b = \frac{2T_C}{1 + 2T_C I_1(T_C)/I_0(T_C)}. \tag{B4}$$

For some purposes, the following two equations, each equivalent to (B4), will be found more convenient:

$$\frac{T_C}{b} = \frac{1}{2} + T_C\frac{I_1(T_C)}{I_0(T_C)}, \tag{B5}$$

$$\frac{T_C}{b} = \frac{1/2}{1 - bI_1(T_C)/I_0(T_C)}. \tag{B6}$$

On account of (B2), the right sides of (B5) and (B6) are equal not only to T_C/b but also to $R_C^2/(1-b^2)$.

Since the utilization of formulas (B4), (B5) and (B6) for computing the curves in Fig. 3.3 will involve taking T_C as the independent variable and assigning to it a set of chosen numerical values, the natural first step is to find approximately the range of T_C corresponding to the b-range, $0 \leqq b \leqq 1$, in order to be able to choose only useful values of T_C. This step will be taken in the next paragraph.

Equation (B6) shows that $T_C/b = 1/2$ when $b = 0$, and hence that $T_C = 0$ when $b = 0$; and this last is verified by (B4). The other end-value of the T_C-range, namely the value of T_C for $b = 1$, cannot be found explicitly and exactly. However, rough values of limits between which it must lie can be found fairly easily as follows: To begin with, each of the equations (B5) and (B6) shows that $T_C \geqq b/2$, for all values of b in $0 \leqq b \leqq 1$; in particular, $T_C > 1/2$ when $b = 1$. An upper limit for T_C for any value of b can be found from (B5) by utilizing the power series expressions for $I_1(T_C)$ and $I_0(T_C)$, whereby it is found that

$$\frac{I_1(T_C)}{I_0(T_C)} = H\frac{T_C}{2}, \quad \text{(B7)} \qquad \text{where} \qquad H \approx 1 - \frac{T_C^2}{8} < 1. \quad \text{(B8)}$$

On substituting (B7) into (B5) and then solving for T_C in terms of b and H, it is found that

$$T_C = b/(1 + \sqrt{1 - Hb^2}). \tag{B9}$$

On account of (B8), (B9) shows that

$$T_C < b/(1 + \sqrt{1 - b^2}), \tag{B10}$$

whence, in particular, $T_C < 1$ when $b = 1$. By successive approximation or otherwise, it can now be rather quickly found that, when $b = 1$, $T_C = 0.79$ (to two significant figures).[33]

From the preceding paragraph, it is seen that, when b ranges from 0 to 1, T_C ranges from 0 to about 0.79; T_C/b ranges from 0.5 to about 0.79; and, on account of (B2), R_C ranges from $\sqrt{0.5} = 0.707$ down to 0.

The curves in Fig. 3.3 are constructed with the aid of the formulas and methods of this appendix as follows: First, a set of values of T_C is chosen, ranging from 0 to 0.79 and slightly larger. Second, for each such chosen T_C the right side of (B5) is computed, thereby evaluating T_C/b and also $R_C^2/(1-b^2)$, these two quantities being equal by (B2). Third, the corresponding value of b is found by dividing T_C by T_C/b; less easily, it could

[33] Because of the special importance of $b = 1$ in other connections, T_C for $b = 1$ was later evaluated to four significant figures and found to be $T_C = 0.7900$; thence, by substituting this value of T into (3.8), along with $b = 1$, it was found that Max. $P(R;1) = 0.9376$, which occurs at $R = R_C = 0$, by (B2).

be found by substituting T_C into (B4). Fourth, from T_C/b the value of $\sqrt{T_C/b}$ is found, and thereby the value of $R_C/\sqrt{1-b^2}$ and thence R_C. Finally, Max. $P(R;b)$ is computed by inserting the critical values into any of the various (equivalent) formulas for $P(R;b)$, namely (3.4), (3.7), (3.8), (3.10) or (3.12).

APPENDIX C

Fomulas of the Curves in Fig. 4.3

The first six equations of this appendix are given without derivation and almost without any comments because they correspond exactly and simply to the first six equations, respectively, of Appendix B. Beginning with the second paragraph of the present appendix, the close correspondence ceases.

$$\frac{dP(r)}{dr} = \frac{dP(r)}{dT}\frac{dT}{dr} = \frac{-2b}{(1-b^2)r^3}\frac{dP(r)}{dT}. \tag{C1}$$

$$\frac{1}{(1-b^2)r_c^2} = \frac{T_c}{b}. \tag{C2}$$

$$\frac{dP(r)}{dT} = P(r)\left[\frac{3}{2T} + \frac{I_1(T)}{I_0(T)} - \frac{1}{b}\right]. \tag{C3}$$

$$b = \frac{2T_c}{3 + 2T_c\, I_1(T_c)/I_0(T_c)}. \tag{C4}$$

$$\frac{T_c}{b} = \frac{3}{2} + T_c\,\frac{I_1(T_c)}{I_0(T_c)}. \tag{C5}$$

$$\frac{T_c}{b} = \frac{3/2}{1 - bI_1(T_c)/I_0(T_c)}. \tag{C6}$$

The bracketed expression in (C3) is seen to be obtainable from that in (B3) by merely changing T to $T/3$ wherever T does not occur as the argument of a function; hence the three equations following (C3) are obtainable from the three equations following (B3) by correspondingly changing T_C to $T_c/3$. (In this appendix, as in Section 4, small c is purposely used as a subscript to indicate a 'critical' value, whereas in Section 3 and in Appendix B, capital C is used for that purpose.)

For use below, it will here be noted that

$$I_1(T_c)/I_0(T_c) = N_1(T_c)/N_0(T_c), \tag{C7}$$

as will be seen by dividing (3.16) by (3.13). On account of (3.17) and (3.14), (C7) shows that for large values of T_c the right side of (C7) is equal to 1 as a first approximation, and to $1 - 1/2T_c$ as a second approximation; thus, for large T_c,

$$I_1(T_c)/I_0(T_c) \approx 1 - 1/2T_c \approx 1. \tag{C8}$$

The first step toward computing the curves in Fig. 4.3 is to find approximately the T_c-range corresponding to the b-range, $0 \leqq b \leqq 1$. This is done in the course of the next four paragraphs.

When $b = 0$, equation (C6) shows that $T_c/b = 3/2$ and hence that $T_c = 0$; or, what is equivalent, $b/T_c = 2/3$ and hence $1/T_c = \infty$ (since $b = 0$).

When $b = 1$, $T_c = \infty$, as can be easily verified from equation (C4), (C5) or (C6) by utilizing (C8).

Thus, from the two preceding paragraphs, it is seen that, when b ranges from 0 to 1, b/T_c ranges from 2/3 to 0; T_c/b from 3/2 to ∞; and T_c from 0 to ∞.

Since $T_c = \infty$ when $b = 1$, the choosing of a set of finite values of T_c will necessitate an approximate formula for computing T_c for values of b nearly equal to 1, which means for very large values of T. Such a formula is easily obtainable from (C5) by utilizing the approximation $1 - 1/2T_c$ in (C8), whereby it is found that, for large T_c,

$$T_c \approx b/(1-b), \tag{C9}$$

$$b/T_c \approx 1-b. \tag{C10}$$

As examples, these approximate formulas give: When $b = 0.99$, $T_c \approx 99$, $b/T_c \approx 0.01$; when $b = 0.9$, $T_c \approx 9$, $b/T_c \approx 0.1$. It will be found that even in the second example the results are pretty good approximations.

The curves in Fig. 4.3 are constructed with the aid of the formulas and methods of this appendix as follows: First, a set of values of T_c is chosen, ranging from 0 to about 100 (the latter figure corresponding approximately to $b = 0.99$). Second, for each such chosen T_c the right side of (C5) is computed, thereby evaluating T_c/b and also $1/(1-b^2)r_c^2$, these two quantities being equal by (C2). Third, the corresponding value of b is found by dividing T_c by T_c/b; less easily, it could be found by substituting T_c into (C4). Fourth, from T_c/b the value of $\sqrt{T_c/b}$ is found, and thereby the value of $1/r_c\sqrt{1 - b^2}$ and thence r_c. Finally, Max $P(r;b)$ is computed by inserting the critical values into any of the (equivalent) formulas for $P(r;b)$, namely (4.2), (4.3) or (4.4).

APPENDIX D

Some Simple General Considerations Regarding the Evaluation of Cumulative Distribution Functions by Numerical Integration

This appendix gives some simple general considerations and relations that may sometimes facilitate and render more accurate the evaluation of cumulative distribution functions by numerical integration.

Some of these considerations and relations have found application in Section 5 in the evaluation of the cumulative distribution function for the modulus $R \equiv |W|$. For this reason, the variate in the present section will be denoted by R, though without thereby restricting R to denote the modulus; rather, R will here denote any positive real variate, though it should preferably be a 'reduced' variate, so as to be dimensionless, as in equation (2.9). The restriction of R to positive values is imposed because it is strongly conducive to simplicity and brevity of treatment, without constituting an ultimate limitation. The reciprocal of R will be denoted by r, as previously.[34]

We may wish to evaluate numerically the cumulative distribution function $p(R'<R) \equiv Q(R)$ or $p(R'>R) \equiv Q^*(R)$ or both. Since these are not independent, their sum being equal to unity, the evaluation of either one determines the other, theoretically. However, when the evaluated one is nearly equal to unity, the remaining one may perhaps not be evaluable with sufficient accuracy (percentagewise) by subtracting the evaluated one from unity. Then it would presumably be advantageous to introduce for auxiliary purposes the variable $r = 1/R$, since evidently

$$p(R'>R) = p(1/R'<1/R) = p(r'<r), \tag{D1}$$

$$p(R'<R) = p(r'>r) = 1 - p(r'<r). \tag{D2}$$

Thus, if $p(R'>R)$, in (D1), is small compared to unity, it is presumably evaluable with higher accuracy percentagewise by dealing with $p(r'<r)$ than with $1 - p(R'<R)$. Incidentally, after $p(r'<r)$ has been evaluated, it might be used in (D2) to arrive at a still more accurate value of $p(R'<R)$ than had originally been obtained directly by numerical integration.

Assuming that we have a plot (or a table) of the distribution function $P(R)$, we can evidently evaluate

$$P(R'<R^0) = \int_0^{R^0} P(R)dR \tag{D3}$$

directly by numerical integration, provided the plot is sufficiently extensive to include R^0; if not, we can, by (D2), resort to

$$p(R'<R^0) = 1 - p(r'<r^0) = 1 - \int_0^{r^0} P(r)dr, \tag{D4}$$

assuming that a sufficiently extensive plot (or table) of $P(r)$ is available and applying numerical integration to it.

Even if the plot of $P(R)$ used in (D3) is sufficiently extensive to include

[34] The restriction of R, and hence of r, to positive values is seen to be absent from equations (D1), (D2), (D5) and (D6) but present in (D3), (D4), (D7) and (D8).

R^0, so that (D3) could be evaluated, it might be that (D4) would result in greater accuracy; this would presumably be the case when $p(R'<R^0)$ is nearly equal to unity.

Evidently an evaluation of

$$p(R'>R^0) = \int_{R^0}^{\infty} P(R)dR \tag{D5}$$

directly by numerical integration would be less satisfactory than the evaluation of $p(R'<R^0)$ in the preceding paragraph. For, due to the presence of the infinite limit in the integral in (D5), the plot of $P(R)$ would have to be carried to a large enough value of R so that the integral from there to ∞ would be known to be negligible. This difficulty can be avoided by starting with the relation

$$p(R'>R^0) = 1 - p(R'<R^0) \tag{D6}$$

and substituting therein the value of $p(R'<R^0)$ given by (D3) or (D4), resulting respectively in the following two formulas:

$$p(R'>R^0) = 1 - \int_0^{R^0} P(R)dR, \tag{D7}$$

$$p(R'>R^0) = p(r'<r^0) = \int_0^{r^0} P(r)dr, \tag{D8}$$

the integrals in which are evidently suitable for evaluation by numerical integration, none of the integration limits being infinite. If $p(R'>R^0)$ is small compared to unity, (D8) would presumably be more accurate (percentagewise) than (D7). If the plot of $P(R)$ is not sufficiently extensive to include R^0, (D7) evidently could not be used; but, instead, (D8) could be used if the plot of $P(r)$ were sufficiently extensive to include r^0.

References on Bessel Functions

1. Watson, "Theory of Bessel Functions," 1st. Ed., 1922; or 2nd Ed., 1944.
2. Gray, Mathews and MacRobert, "Bessel Functions," 2nd Ed., 1922.
3. McLachlan, "Bessel Functions for Engineers," 1934.
4. Bowman, "Introduction to Bessel Functions," 1938.
5. Whittaker and Watson, "Modern Analysis," 2nd Ed., 1915.
6. "British Association Mathematical Tables," Vol. VI: Bessel Functions, Part I, 1937.
7. Anding, "Sechsstellige Tafeln der Bessel'schen Funktionen imaginaren Arguments," 1911 (mentioned on p. 657 of Ref. 1).

Spectrum Analysis of Pulse Modulated Waves

By J. C. LOZIER

The problem here is to find the frequency spectrum produced by the simultaneous application of a number of frequencies to various forms of amplitude limiters or switches. The method of solution presented here is to first resolve the output wave into a series of rectangular waves or pulses and then to combine the spectrum of the individual pulses by vectorial means to find the spectrum of the output. The rectangular wave shape was chosen here as the basic unit in order to make the method easy to apply to pulse modulators.

INTRODUCTION

The rapidly expanding use of pulse modulation[1] in its various forms is bound to make the frequency spectrum of pulse modulated waves a subject of increasing practical importance. The purpose of this paper is to show how to determine the frequency spectrum of these waves by methods based as far as possible on physical rather than mathematical considerations. The physical approach is used in an attempt to maintain throughout the analysis a picture of the way in which the various factors contribute to a given result. To further this objective the fundamentals involved are reviewed from the same point of view.

The method is used here to analyze two distinct types of pulse modulation, namely, pulse position and pulse width modulation.[2] These two cases are especially important for illustrative purposes because their spectra can be tied back to more familiar methods of modulation. Thus it will be shown that, as the ratio of the pulse rate to the signal frequency becomes large, pulse position modulation becomes a phase modulation of the various carrier frequencies that form the frequency spectrum of the unmodulated pulse wave, and pulse width modulation becomes a form of amplitude modulation of its equivalent carriers. The analysis also shows certain interesting input-output relationships that may be obtained from such modulators, treating them as straight transmission elements at the signal frequency.

These relationships are of more than theoretical interest. The pulse position modulator has already been used as phase or frequency modulator to good advantage.[3] The use of a pulse width modulator as an amplifier is

[1] E. M. Deloraine and E. Labin, "Pulse Time Modulation", *Electrical Communications*, Vol. 22, No. 2, pp. 91-98, Dec. 1944; H. S. Black "AN-TRC-6 A Microwave Relay System", *Bell Labs. Record*, V. 33, pp. 445–463, Dec. 1945.

[2] By *pulse position* modulation is meant that form of pulse modulation in which the length of each pulse is kept fixed but its position in time is shifted by the modulation, and by *pulse width* modulation that form in which the length of each pulse varies with the modulation but the center of each pulse is not shifted in position.

[3] L. R. Wrathall, "Frequency Modulation by Non-linear Coils", *Bell Labs. Record*, Vol. 23, pp. 445–463, Dec. 1945.

another practical application, of which the self oscillating or hunting servomechanism is an example.

The quantitative analysis of such systems depends on the ratio of the pulse repetition rate to the signal frequency. When this ratio is low, the solution can be obtained by a method shown here for resolving the modulated waves into selected groups of effectively unmodulated components. This technique is powerful since it can be done by graphical means whenever the complexity of either the system or the signal warrants it. When the ratio of pulse rate to signal frequency becomes high enough, such methods are no longer practical. However, under these conditions other methods become available, especially in cases like those mentioned above where the spectrum of the modulation approaches one of the more familiar forms. An important example of this occurs in the case of the pulse position modulator where, as the spectrum approaches that of phase modulated waves, the solution can often be found by the conventional Bessel's function technique used in analyzing phase and frequency modulators.

The method proposed here for obtaining the spectrum analysis of pulse modulated waves is based on the use of the magnitude-time characteristic of the single pulse and its frequency spectrum as a pair of interchangeable building blocks, so that the analysis will develop this relationship. Before doing this the elementary theory of spectrum analysis will be reviewed

Review of the Elementary Theory of Spectrum Analysis

A complex wave may be represented in two ways. One way is by its magnitude at each instant of time. The other way is by its frequency spectrum, that is, by the various sinusoidal components that go to make up the wave. The two representations are interchangeable.

The transformation from a given frequency spectrum to the corresponding magnitude vs. time function is straight-forward, for it is apparent that the various components in the frequency spectrum must add up to the desired magnitude-time function. The necessary additions may be difficult to make in some cases but they are not hard to understand.

The reverse process of finding the frequency spectrum when the magnitude-time characteristic is given is more involved, though using Fourier analysis, the problem can generally be formulated readily enough. Furthermore the mathematical procedures involved can be interpreted physically in broad terms by modulation theory. However, these procedures become more difficult to perform, and the physical relationships more obscure, as the wave form under analysis becomes more complex. This is particularly true when general or informative solutions rather than specific answers are required. Pulse modulated waves are sufficiently new and complex to give such difficulties.

The process of finding the frequency spectrum of a complex wave from its magnitude-time function has a simple mathematical basis. It depends on the fact that the square of a sinusoidal wave has a positive average value over any interval of time, whereas the product of two sinusoidal waves of different frequencies will average zero over a properly chosen interval of time.[4]

In theory then, as the magnitude-time function of a complex wave is the sum of all the components of the frequency spectrum, we have only to multiply this magnitude-time function by a sinusoidal wave of the desired frequency and then average the product over the proper time interval to find the component of the spectrum at this frequency.[5]

One physical interpretation of this procedure can be given in terms of modulation theory. The product of the magnitude-time function with a sinusoidal wave will produce the beat or sum and difference frequencies between the frequency of the sinusoid and each component of the frequency spectrum. Thus, if the spectrum contains the same frequency, a zero beat or dc term is produced, and this term may be evaluated by averaging the product over an interval that is of the proper length to make all the ac components vanish.

The application of this principle for spectrum analysis is simple when the magnitude of the wave in question is a periodic function of time. The very fact that the wave is periodic is sufficient proof that the only frequencies that can be present in the wave are those corresponding to the basic repetition rate and its harmonics. Thus the frequency spectrum is confined to these specific frequencies and so it takes the form of a Fourier series. Knowing that the possible frequencies are restricted in this way, the problem of finding the frequency spectrum of a complex periodic wave is reduced to one of performing the above averaging process at each possible frequency. The period of the envelope of the Complex Wave is the proper time interval for averaging, and the integral formulation for obtaining this average is that for determining the coefficients in a Fourier series.

The principle holds equally well when the magnitude-time function is nonperiodic, but the concept is complicated by the fact that the frequency spectrum in such cases is transformed from one having a discrete number of components of harmonically related frequencies to one having a continuous-band of frequencies.[6] Such spectra contain infinite numbers of sinusoidal

[4] The proper time interval is generally some integral multiple of the period corresponding to the difference in frequency of the two sinusoid waves.

[5] In practice it is generally necessary to multiply by both sine and cosine functions because of possible phase differences.

[6] One exception to this statement is the fact that any wave made up of two or more incommensurate frequencies is nonperiodic. Yet such waves will have a discrete spectrum if the number of components is finite. This incommensurate case is neglected throughout the discussion.

components, each of infinitesimal amplitude and so close together in frequency as to cover the entire frequency range uniformly.

The continuous band type of frequency spectrum is just as characteristic of non-periodic waves as the discrete spectrum is of periodic waves. This can be shown as a logical extension of the Fourier series representation of periodic waves. The transition from a frequency spectrum consisting of a series of discrete frequencies to one consisting of a continuous band of frequencies can be made by treating the non-periodic function as a periodic function in which the period is allowed to become very large. As the period approaches infinity the fundamental recurrence rate approaches zero, so that the harmonics merge into a continuous band of frequencies.

This does not of course change the basic realtionship between the frequency spectrum of a wave and its magnitude-time function. The magnitude-time function is still the sum of the components of the frequency spectrum. Also the frequency spectrum can still be obtained frequency by frequency, by averaging the product of the magnitude-time function and a unit sinusoid at each frequency. However, the actual transformations in the case of the non-periodic functions require summations over infinite bands of frequencies and over infinite periods of time and so fall into the realm of the Fourier and similar integral transforms.

However, in any case the problem of spectrum analysis reduces to an averaging process. The process can be performed by mathematical integration in all cases where a satisfactory analytical expression for the magnitude-time function is available. Fourier analysis provides a very powerful technique for setting up the necessary integrals in such cases.

This averaging process can also be done graphically. It is apparent from the theory that if the product of the magnitude-time function and the sinusoid is sampled at a sufficient number of points, spaced uniformly over the proper time interval, then the average of the samples gives the desired value. This technique is fully treated elsewhere[7] so that it will not be considered in detail here. However, use will be made of it in a qualitative way to augment the physical picture.

Non-Linear Aspects

The use of the frequency spectrum in transmission studies is generally limited to cases where the system in question is linear; that is, where the transmission is independent of the amplitude of the signal. However, the same techniques can still be used on systems employing successive linear and non-linear components, in cases where the transmission through the non-linear elements is independent of frequency. Under these conditions, the magnitude-time representation of the wave can be used in computing

[7] **Whittaker and Robinson, Calculus of Observations.**

the transmission over each non-linear section, where the transmission is dependent only on the amplitude, and the frequency spectrum used over each linear section, where the transmission is dependent only on the frequency. This a technique can be used on most pulse modulating systems because such non-linear elements as the modulators and limiters generally encountered are substantially independent of frequency.

Frequency Spectrum of the Single Pulse

The single pulse is a non-periodic function of time and so has a continuous frequency spectrum. In this case the Fourier transforms are simple. They are derived in Appendix A. Figure 1 gives a graphical representation of the magnitude-time function and the frequency spectrum of the pulse. The expressions are general and hold for pulses of any length or amplitude.

It is instructive to note that the frequency spectrum in this case can be determined by using the graphical technique mentioned previously. For example, consider the product of the magnitude-time function of the single pulse with a sinusoidal wave of given frequency and unit amplitude, so arranged in phase that its peak coincides with the center of the pulse. Theoretically the average of this product taken over the infinite period will give the relative magnitude of the component in the frequency spectrum of the pulse having the same frequency as the sinusoidal wave. In this case however, the average need only be taken over the length of the pulse, since the product vanishes everywhere else. Thus at very low frequencies, where the period of the sinusoidal wave is very much greater than the length of the pulse, the average is proportional to $2EL$ where E is the amplitude and $2L$ the length of the pulse. Then as the frequency increases, the average of the product, and hence the relative amplitude of the component in the spectrum, will first decrease. For the particular frequency such that the length of the pulse is one half the period, the relative amplitude will have

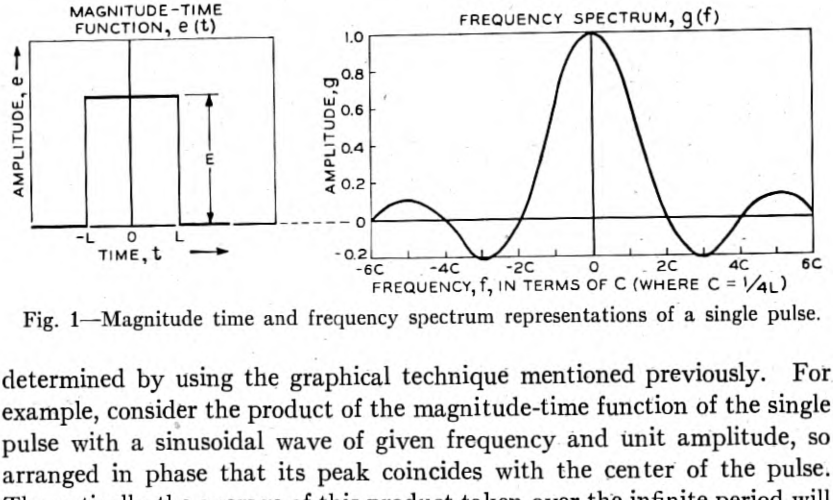


Fig. 1—Magnitude time and frequency spectrum representations of a single pulse.

fallen to $2EL \times \frac{2}{\pi}$ $\left(\frac{2}{\pi}\right.$ being the average value of a half wave of unit amplitude$\left.\right)$. Similarly when the frequency is such that the length of the pulse is a full wavelength, the average will vanish, and when the pulse length is one and a half times the wavelength, the average is negative, having two negative and one positive half waves over the length of the pulse, and the relative magnitude is $2EL \times \frac{2}{3\pi}$. These products are shown graphically on Fig. 2. Since these amplitudes correspond to those given in Fig. 1, for the spectrum components at $f = f_0 = 1/4L$, $2f_0$, and $3f_0$, it is apparent that the spectrum could be determined in this way.

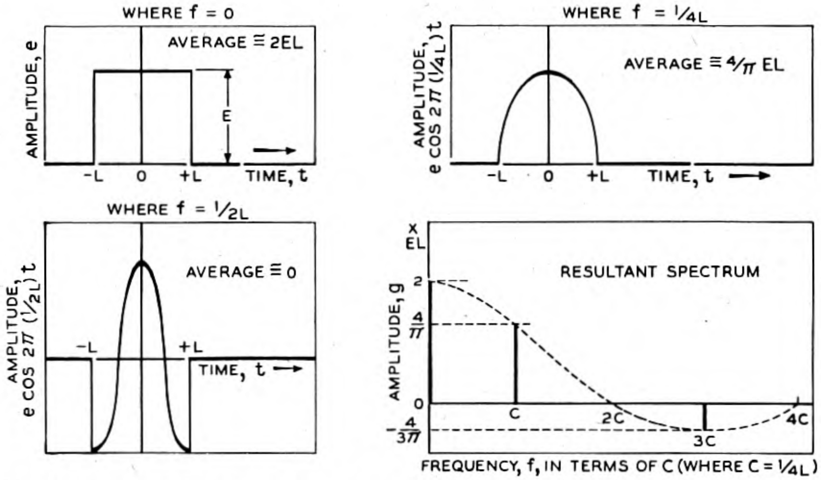


Fig. 2—Graphical derivation of spectrum of single pulse by averaging product of pulse with sinusoidal waves of various frequencies.

Basic Technique

In the analysis presented here, the single pulse and its spectrum will be used in such a way that the need for individual integral transforms for each complex wave form under study is avoided. The theory is simple.

A complex wave form may be approximated to any desired accuracy by a series of pulses, varying with respect to time in length, in amplitude, and in position. Now the spectra of these individual pulses are already known. Therefore, to find the frequency spectrum of the complex wave in question, it is necessary only to combine properly the spectra of the various pulses representing the complex wave.

Thus the process is theoretically complete. The procedure is first to break down the given complex wave into a series of single pulses. Next the spectrum of each pulse is determined separately. Then the spectrum of the complex wave is obtained by combining the spectra of the various single pulses involved. One of the things to be demonstrated here is that it is perfectly feasible in many cases to perform these summations graphically, even though basically it does involve the handling of spectra each containing an infinite number of frequency components.

There are other wave forms that could be used as the fundamental building block instead of the single pulse. The unit step function is one possibility, since it is used in transient analysis for a similar purpose. However, the single pulse has obvious advantages when the complex wave to be analyzed is itself a series of pulses, as in pulse modulation. Again it would be nice to be able to choose as the fundamental unit a wave that has a discrete rather than a continuous band frequency spectrum, but it seems that any wave flexible enough to make a satisfactory building unit is inherently nonperiodic and so has a continuous frequency spectrum. However the fact that the fundamental units have continuous spectra does not of itself complicate the results. If for example, the wave to be analyzed is periodic, the sum of the spectra of the various pulses must reduce to a discrete frequency spectrum. In the cases of interest here, when the pulse train under analysis is repetitive, combinations of identical pulses will be found to occur with the same fundamental period, and generally the first step in the summation of such spectra is to group the series of pulses into periodic waves with discrete spectra.

Manipulations of Single Pulses

In its use, the single pulse may be varied in amplitude, in length, and in position with respect to time. These changes have independent effects on the frequency spectrum. A variation in the amplitude of a pulse does not change its spectrum, except to increase proportionately the magnitudes of all components. A change in position of a pulse with time does not change the amplitude vs. frequency characteristic of the spectrum, but it does shift the phase of each component by an amount proportional to the product of the frequency and the time interval through which the pulse was shifted. A change in the length of a pulse will change the shape of the amplitude vs. frequency characteristic of the spectrum. Figure 3 shows this effect. However, if the center point of the pulse is not shifted in time, the relative phases of the components are not affected by such changes in length.

The single pulse can also be modulated to aid in the resolution of more complicated wave forms. This process is based on the use of the pulse as a function having a value of unity over a chosen time interval and a value of zero at all other times. Thus, to show a part of a sinusoidal wave, we need only multiply this wave by a pulse of the correct length and proper phase with respect to the sinusoid to show only the desired piece of the wave. In this simple case it is not difficult to derive the spectrum because what are produced are the sum and the difference products of the modulating frequency with the spectrum of the pulse. This gives two single pulse spectra shifted up and down in frequency by the frequency of the modulation. An example of this is shown in Fig. 4, where the spectrum of a single half cycle is determined.

Pulse Position Modulation

For the first example, a simple form of pulse position modulation will be analyzed. The pulse train in this case is made up of pulses spaced T seconds apart and the width of each pulse is a very small part of the spacing T. Such a pulse train is shown on Fig. 5. The pulse train is modulated by advancing or retarding the position (time of occurance) of the pulses by an amount proportional to the instantaneous amplitude of the signal at sampled instants T seconds apart. Figure 5 also shows the signal, in this case a sine wave of frequency $1/10T$, and the resulting modulated pulse train. The peak amplitude of the modulating sine wave is assumed to shift the position of a pulse by $1/4T$. The length and the amplitude of the pulses are the same since neither is affected in this type of modulation.

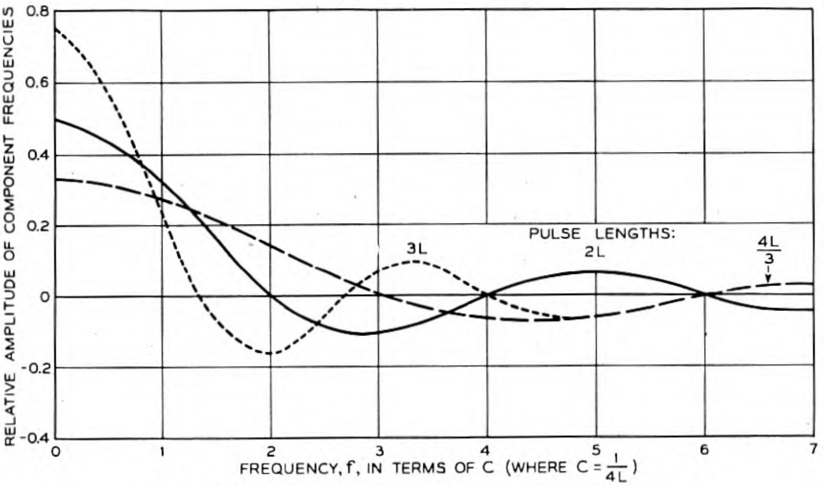


Fig. 3—Change in frequency spectrum with pulse length.

The first step in the analysis is to determine the spectrum of the pulse train before modulation. Each pulse contributes a spectrum of the form shown on Fig 1. Now the phase of each component in such a spectrum is so arranged that the spectrum forms a series of cosine terms all of which have zero phase angle at the center of the pulse. From successive pulses T seconds apart, the component at any given frequency will have the same amplitudes, but the relative phases will be $2\pi fT$ radians apart. It is apparent that frequencies for which $2\pi fT$ is 2π or some multiple of 2π radians

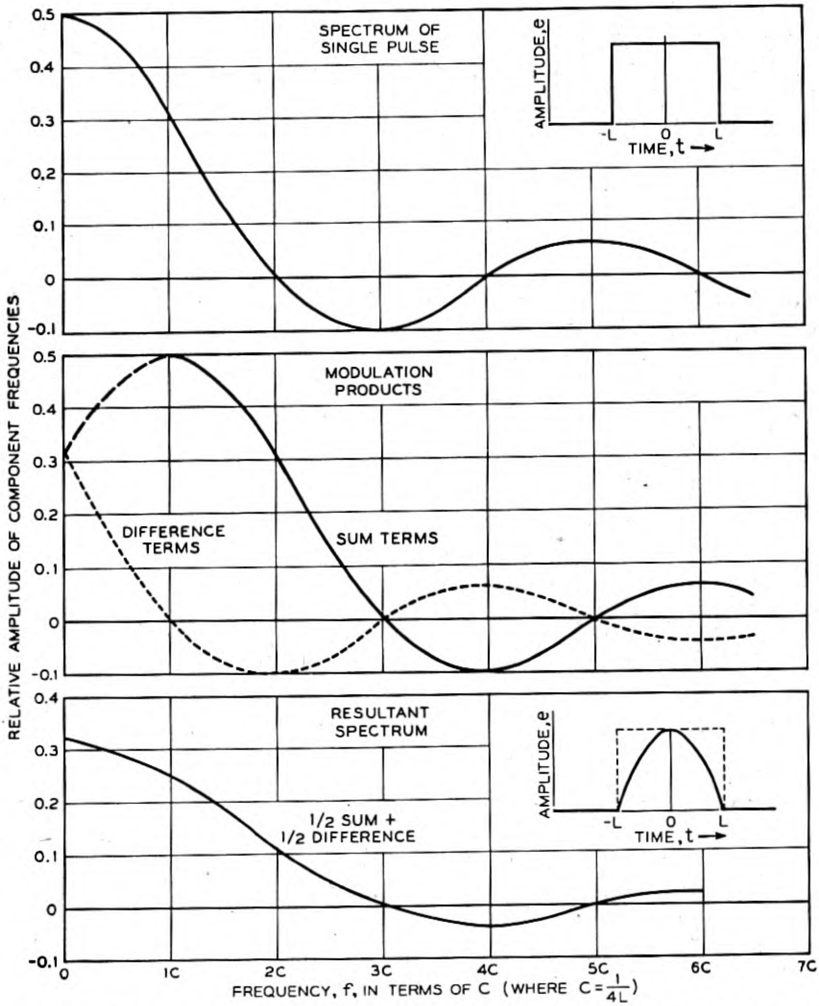


Fig. 4—Determination of spectrum of single half sine wave by modulation of single pulse spectrum with cos $2\pi ct$.

apart, the contributions from all pulses add in phase. These are the frequencies nc, where $n = 1,2,3$ and $c = \frac{1}{T}$. It is also apparent that at frequencies for which the phase differences between the components are not an exact multiple of 2π radians apart, the contributions from enough pulses must be spread in phase over an effective range of 0 to 2π radians in such a way as to cancel one another. For example, take the particular frequency for which the difference in phase between pulses is 361° instead of 360°.

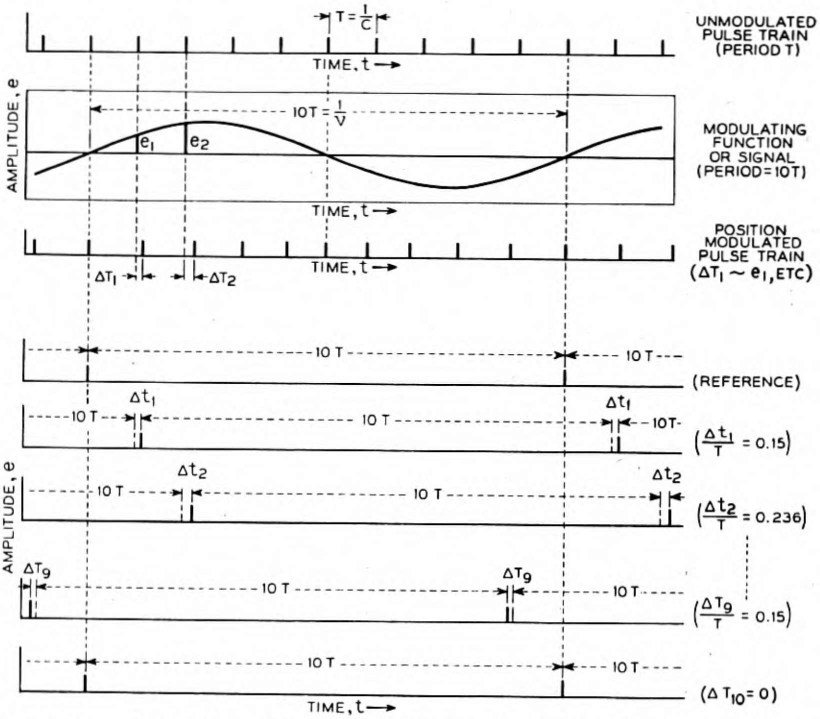


Fig. 5—Formation of pulse position modulated pulse train and its resolution into subsidiary unmodulated pulse trains.

The contribution from each preceding pulse will be effectively advanced in phase 1° with respect to its successor, so that the contributions from pulses 180 periods apart will be exactly 180° out of phase. Therefore over a sufficient number of pulses, the net contribution is zero.

The spectrum of the unmodulated pulse train is thus made up of a dc term plus harmonics of the frequency $C = 1/T$. The dc term is the average, and therefore is equal to $E \times 2L/T$, where E is the magnitude of the pulse. All of the other components have the same relative magnitudes that they have in the single pulse spectrum. This gives a spectrum like that shown on Fig. 6. Figure 6 also shows for comparative purposes the spectrum of the subsidiary pulse wave consisting of every 6th pulse.

Thus in the unmodulated case, the pulses have a uniform recurrence rate and the resultant spectrum, found by adding those of the individual pulses, reduces to a train of discrete frequencies comprised only of the harmonics of the recurrence rate of the pulses. The fundamental frequency, corresponding to the recurrence rate, and its harmonics will be called the carrier frequencies of the pulse train. The effect of modulating the pulse train is to modulate each of these carriers, producing sidebands of the signal about them.

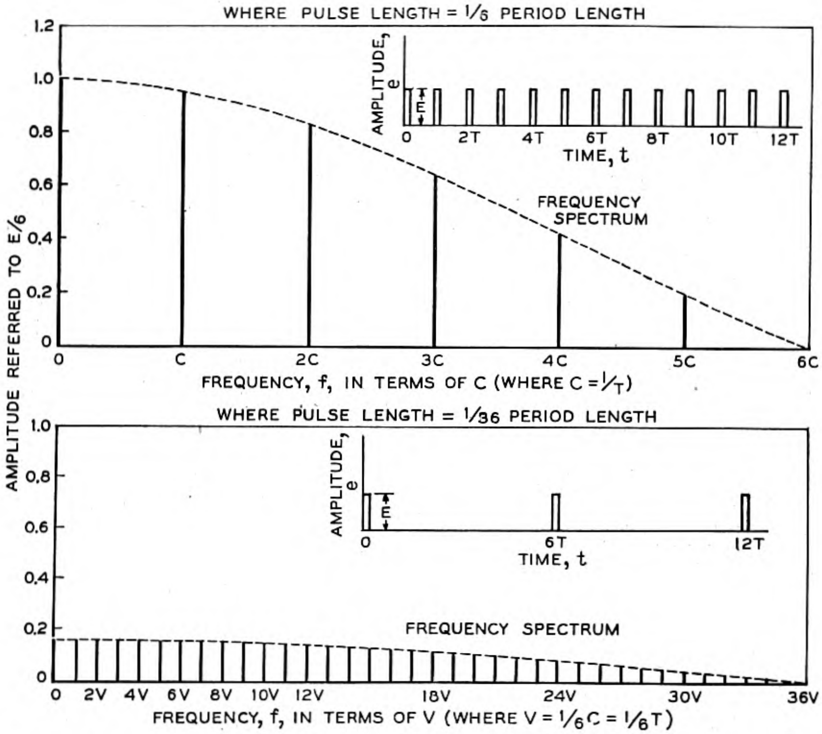


Fig. 6—Frequency spectrum of pulse trains where the spacing between the pulses is 6 and 36 times the pulse length respectively.

When the pulse train is position modulated, the pulses are shifted in position by an amount ΔT, corresponding to the instantaneous amplitudes of the modulating function. The spectrum of each pulse is unchanged, since the pulse length remains constant. However, components of successive pulses at the carrier frequency c and its harmonics will no longer add directly, because of the phase shifts that accompany the change in position. This phase shift is equal to ΔT, the shift in position, times the radian frequency of the component in question.

However, when the signal function is periodic, each pulse will have the same shift in position as any other pulse that occurs at the same relative instant in a later modulating cycle. Furthermore, when the carrier frequency is an exact multiple of the signal frequency i.e., $c = nv$, there will be a pulse recurring at the same relative instant in each cycle of v. Under these conditions, the pulse position modulated wave can be broken down into a group of unmodulated waves, each being made up of that series of pulses that recur at a given part of each modulating cycle, as shown in Fig. 5. These subsidiary waves are effectively unmodulated because, as each pulse recurs at the same instant in the modulating cycle, they are shifted to the same extent and hence will be uniformly spaced. This uniform spacing between pulses in a given wave is equal by definition to the period of the modulating function, and there will be as many of these unmodulated pulse trains as there are pulses in a single cycle. Thus, if $c = nv$, there will be n such pulse trains.

The reason for grouping the pulses into these unmodulated pulse tains is that unmodulated periodic trains have spectra of discrete frequencies. Since the pulse widths are all equal, and since the spacing between pulses is the same for each wave, the spectra of these unmodulated waves will all be identical. Furthermore, these spectra will be the same as that of the original carrier wave of pulses before modulation, except for two factors. First, the fundamental frequency is now v, corresponding to the modulating period, so that there are n times as many components as before. Secondly the amplitudes are reduced by the factor $\frac{1}{n}$ because there is only one pulse in these new waves to every n pulses in the original wave. Thus, instead of having a spectrum made up of the carrier frequency and its harmonics, we now have one made up of harmonics of v. Since $c = nv$, such frequencies as c, c, $\pm v$, $c \pm 2v$, etc., are included. An example of the spectra of both the subsidiary and original pulse waves is shown on Fig. 6, for the case where $n = 6$.

Thus the problem of finding the spectrum of such a pulse position modulated wave is reduced by this procedure to adding up the n equal components at each of the frequencies of interest, such as c and $c \pm v$, allowing for the phase difference between components corresponding to the position of one pulse with respect to that of the other n–l pulses in one modulating cycle. As an example, suppose $n = 10$ and the frequency to be computed is $c + v$. Now $c + v$ is 10% higher in frequency than c. Thus in the unmodulated case, when the n pulses are equally spaced, they are 360° apart at c and consequently 360° + 36 or 396° at $c + v$. Therefore in the unmodulated case, each component would be advanced in phase 36° with respect to the previous one, so that the diagram of the 10 components would form the

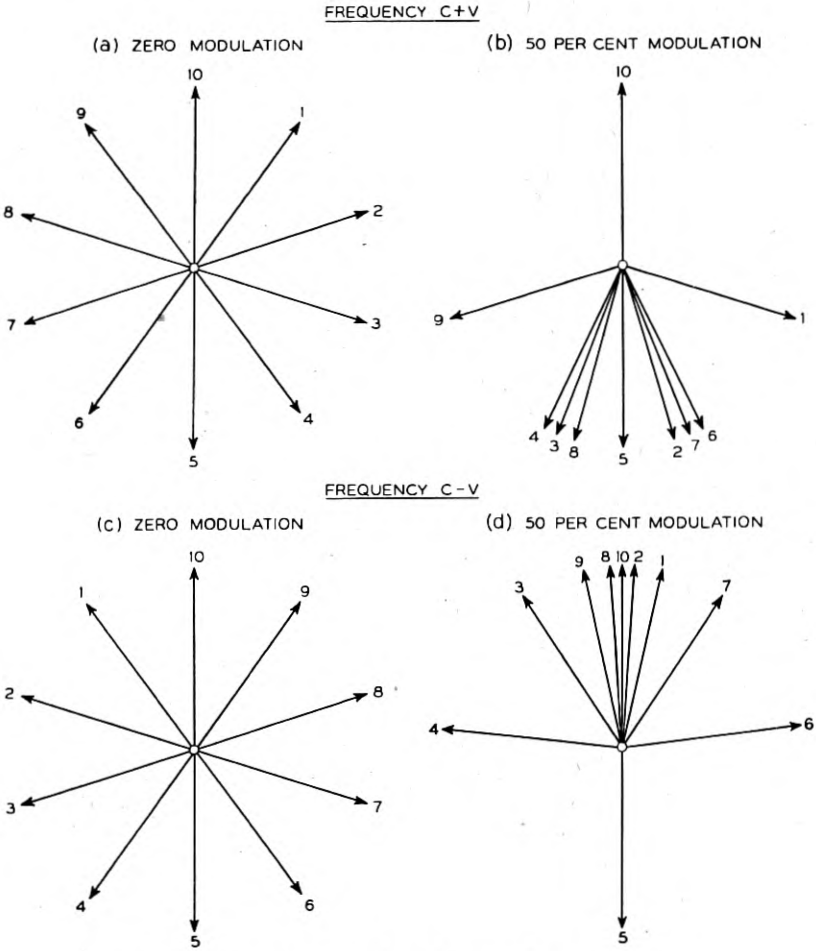


Fig. 7—Vector pattern of subsidiary pulse components.

vector pattern shown on Fig. 7A. The successive components are numbered 1 to 10. The sum in this unmodulated case is of course zero.

Now the effect of modulation is to shift the relative phases of these components by an amount determined by the shift in position of the corresponding pulses. When these relative phase shifts are such as to spoil the cancellation of the 10 components, a net component of this frequency is produced in the frequency spectrum of the pulse wave. Taking the example shown in Fig. 5, the 10 components in Fig. 7A would be shifted to the positions shown in Fig. 7B. These shifts in relative phase are determined in the following way. Figure 5 shows that the number 1 pulse is retarded an amount ΔT_1 equal to 15% of T, the normal spacing between pulses. Thus at the carrier frequency c, the phase shift between the component from this retarded pulse and the reference pulse is 15% more than 360° or 414°. Thus the component at the carrier frequency c from the first subsidiary pulse train is shifted 54° from its unmodulated position.

At $c + v$, since the frequency is 10% higher, the net shift is 10% more than at c or 59.5°. Thus the number 1 component on the vector diagram of Fig. 7B is rotated 59.5° clockwise from its unmodulated position shown on Fig. 7A.

Similarly pulses 2 and 3 are each shifted in position by equal amounts, ΔT_2 and ΔT_3. These shifts in position give 85° phase shift at the carrier frequency. Hence components 2 and 3 at $c + v$ are each rotated 10% more or 93.5° from their respective unmodulated reference positions shown on Fig. 12A. Component number 4 is shifted 59.5° clockwise just as number 1. Component 6 and 9 are also shifted 59.5° each, but in this case the modulating function has the reverse polarity so that the components are rotated counterclockwise. Similarly components 7 and 8 are rotated 93.5° counterclockwise.

The sum of these components in the vector diagram of Fig. 7B gives a resultant that is negative with respect to the reference direction and the magnitude that is 58% of the reference magnitude, where the reference magnitude and direction are those for the carrier c with no modulation.

This gives the relative magnitude and phase of the $c+v$ term produced by pulse position modulation for the case where the modulating function is a sine wave of frequency $v = c/10$ with a peak amplitude just large enough to shift a pulse by 1/4 of T, where T is the spacing between unmodulated pulses. A shift of this magnitude will be defined here as 50% modulation on the basis that 100% modulation should be $1/2T$, the maximum displacement that can be used without possible interference between pulses.

In the same way the other component frequencies in the spectrum such as c, $c - v$, $c \pm 2v$, etc., have been computed for the above case of 50% modulation, and for other peak amplitudes of the modulating sine wave giving 25%, 70% and 100% modulation. In all cases the frequency of the modulating function was held at $v = c/10$. This information is plotted on Fig. 8, showing v, c and the various components of the frequency spectrum that represent the sidebands about the carrier frequency c, as a function of the peak % modulation.

The above solution assumed a special case where c was an exact multiple of v. The purpose of this assumption was to simplify the problem to the extent that the periodicity of the modulated wave would be the same as that of the modulating function. There are two other possible cases. For one, the ratio of c to v could be such that a pulse would occur at the same instant of the modulating period only once every so many periods. The actual periodicity of the modulated pulse wave would be reduced accordingly because it would make the same number of periods of the modulating function before the modulated pulse train is repeated. This is a result of the fact that pulse modulation provides for a discrete sampling rather than a continuous measure of the modulating wave. The technique of spectrum analysis demonstrated above is just as applicable to this case as it was to the simpler one. However, there will be comparatively more terms to be handled. The other possible case is the one where c and v are incommensurate.[8] In this case, the resulting modulated wave is non-periodic. However, on the basis that the spectrum is practically always a continuous function of the signal frequency, this case has received no special attention here.

At frequencies for which c is very much greater than v, so that the number of component pulse trains becomes too numerous to handle conveniently in the above fashion, the sidebands about each carrier or harmonic of the switching frequency can be computed by the standard methods for phase modulation, as the next section will demonstrate. This result follows directly from the theorem that as the carrier frequency c becomes large with respect to v, pulse position modulation merges into a linear phase modulation of each of the carriers.

Pulse Position Modulation vs Phase Modulation

When a pulse, in a pulse position modulated wave, is shifted by 1/2 the spacing between pulses (100% modulation) it is apparent from the previous discussion that the component of the carrier in the frequency spectrum of the pulse is shifted by 180°. Therefore to compare the spectrum of a pulse position modulated wave like that on Fig. 8 with the equivalent spectrum of a phase modulated wave, what is needed is Fig. 9, showing the frequency spectrum of a phase modulated wave of the form $Cos(ct - k\, sin\, vt)$ as a function of k for values of k up to π radians or 180°. The computation of the frequency spectrum of such a phase modulated wave has been adequately covered elsewhere and all that is done here is to give the brief development shown in appendix B.

[8] Mr. W. R. Bennett has pointed out that this incommensurate case is the general one. It requires a double Fourier series, which reduces to a single series when the signal and carrier frequencies are commensurate. This analysis is based on the single Fourier series.

A comparison of the spectra on Figs. 8 and 9 shows that the sidebands have the same general pattern. However comparative sidebands are not quite equal in the two cases. In fact comparable upper and lower sidebands in the case of the pulse modulated wave shown on Fig. 8 are not

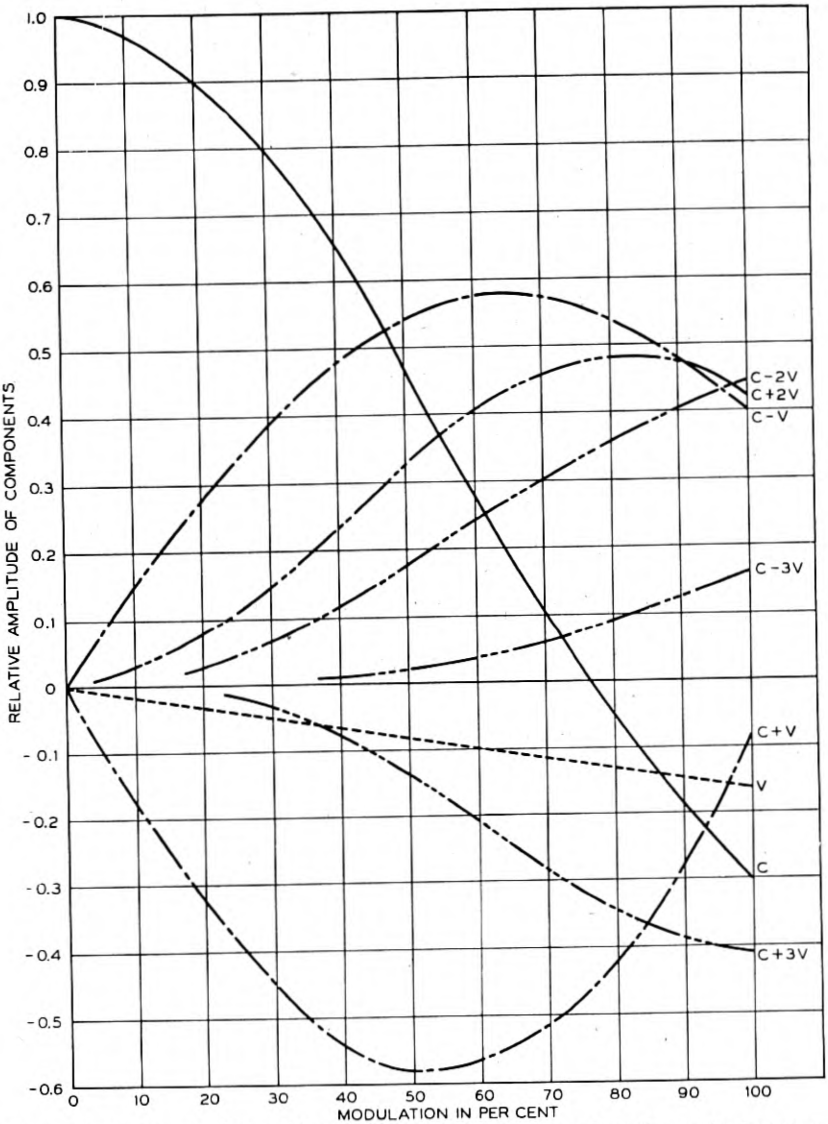


Fig. 8—Spectrum of pulse position modulated wave for case where the carrier frequency C is 10 times the signal frequency v.

equal in absolute magnitude to each other. This lack of symmetry is due to the fact that c is only 10 times v.

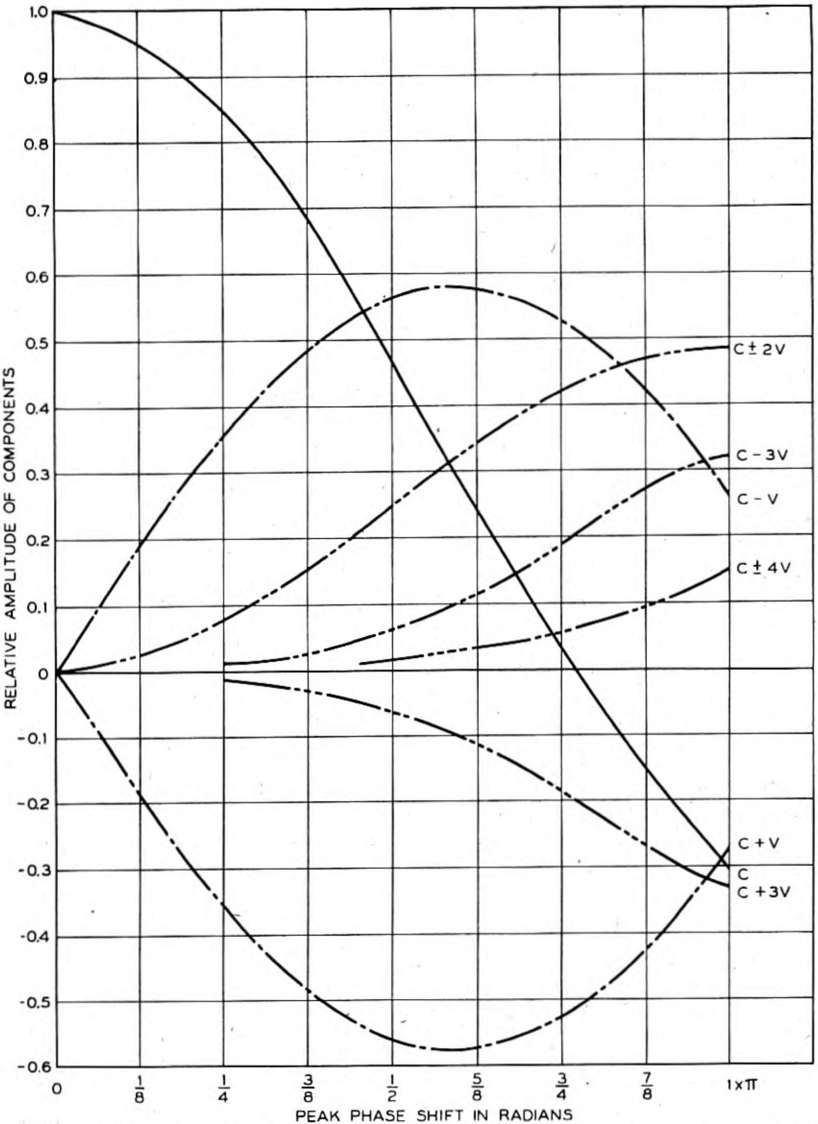


Fig. 9—Spectrum of phase modulated wave cos $(ct + k \sin vt)$ as function of peak phase shift k for values of k up to π radians.

One way of proving this is to go through the process of computing the $c - v$ term in this pulse modulated wave just as the $c + v$ term was computed earlier. Since the frequency $c - v$ is 10% less than c, the unmodulated pattern of the 10 subsidiary components, as shown on Fig. 7C, is the mirror image of that for $c + v$ in 7A, for the first component is now 360° less 10% or 324°, and subsequent components are each retarded 36° with respect to the previous one. When the pulse train is modulated the effect is similar to the case for $c + v$ and, for the same per cent modulation, the Vector pattern of Fig. 7D is formed. The resultant in this case differs from that of 7B in sign as well as in magnitude. The difference in sign comes from the fact that, since component 1 in 7A corresponds to component 9 in 7C and component 2 in 7A to component 8 etc., the modulation in the case of $c - v$ rotates these corresponding components in opposite directions. The difference in magnitude is due to the fact that since $c - v$ is an appreciabley lower frequency than $c + v$ in this case (approx. 20%), the phase shift corresponding to a given shift in pulse position is proportionately less. Thus the corresponding Vector components are not shifted the same number of degrees. Thus the absolute magnitudes of $c + v$ and $c - v$ are not equal in this case.

It is apparent that this difference in magnitudes of $c + v$ and $c - v$ becomes smaller as the carrier frequency c becomes larger with respect to v. In the limiting case of c very much greater than v, $c + v$ and $c - v$ would each be shifted the same number of degrees as c itself. If this more or less compromise shift of c is used to compute the $c \pm v$, $c \pm 2v$, and c $\pm$ $3v$ terms, then the resulting frequency spectrum is that of the phase modulated carrier on Fig. 9.

The higher harmonics of c in the pulse position wave are similarly phase modulated and the interesting point is that $2c$ is modulated through twice as many degrees phase shift and $3c$ 3 times as many degrees, etc. Thus a single pulse position modulator could be designed to produce a harmonic of c with almost any desired degree of phase modulation. This is a useful method for obtaining a phase modulated wave, or with a 6 db per octave predistortion of the signal, a frequency modulated wave.

Figure 8 also shows a term in v itself, which has been neglected so far in the discussion. It is apparent that the components at v contributed by the 10 subsidiary unmodulated waves must form the same kind of vector pattern as those of $c + v$ in Fig. 7. However, in this case $c + v$ is eleven times v in frequency, so that the components of v are rotated only one eleventh as much for a given pulse diplacement. Thus the magnitude of v at 100% modulation is equal to that of $c + v$ at approximately 9% modulation. For different frequency ratios of c to v the relationship of the v term to $c + v$ will vary, and it is apparent that for c very much greater than v, the v term will vanish. The relationship is such that the amplitude of the v component out of the modulator at a given per cent modulation is directly proportional to its own frequency v for all frequencies less than approximately one quarter of c, and the phase is 90° with respect to the input. Thus the modulator puts out a signal component that is the derivative of the input signal.

To summarize the case of pulse position modulation, the frequency spectrum may be determined by the methods based on subdividing the modulated pulse train into a series of unmodulated ones when the ratio of c to v is small, and by treating each harmonic of the carrier as a phase modulated wave of the form $Cos\, n\, (ct + \theta)$, where θ is the modulating function, when the ratio of c to v is large. In the case treated here, the modulating function was a simple sinusoidal wave. Of course the analysis holds for more complicated wave shapes having frequency spectra of their own. In this event however the restriction on the relative magnitudes of the frequencies v and c should be taken as one on c and the highest frequency in the modulating spectrum. The complexity of the modulating function does not affect the analysis when it is done by this technique of subdividing the pulse train, since all that need be known is how much each pulse is shifted, and this can be done graphically. The analysis given here has neglected the length of the individual pulses. This was done when it was assumed that the individual contributions from the various pulse trains had the same amplitude at all frequencies. For any finite pulse width, the relative magnitudes of the various components must be modified by the $\frac{sin\ x}{x}$ factor of the single pulse, as shown on Fig. 6.

As mentioned in the introduction, a complex wave could be analyzed by multiplying its magnitude-time characteristic by unit sinusoids at each frequency in question, sampling the product at a sufficient number of points uniformly spaced over a cycle of the envelope of the complex wave, and then averaging the values of the product thus obtained. This technique is particularly applicable to the analysis of pulse position modulated waves since, by taking the centers of the pulses of the modulated wave as the sampling instants, it is possible, with a finite number of samples (same as the number of pulses) to get the same results as though a very much greater number of uniformly spaced samples were taken. The interesting thing to note here is that the actual computations that would be involved in applying this sampling method of analysis to a pulse position modulated wave are almost identically the same calculations as required by the technique of resolving the pulse train into unmodulated subsidiary pulse trains used here.

Pulse Width Modulation

Pulse Width Modulation as defined here could also be termed "pure" pulse length modulation. The pulse train in the reference or unmodulated condition is a recurrent square wave, and the lengths of the pulses will be varied by the modulation without changing the position of the centers of the pulses. The term "pure" pulse length modulation is applicable to this special case where the phase relationship between spectra of adjacent pulses does not change with modulation because the centers of the pulses are not shifted by the modulation. The conventional form of pulse length modulation, where one end of the pulse is fixed in position, combines both this pulse width modulation and the pulse position modulation previously analyzed. The interest in this case of pulse width modulation arose in connection with the analysis of "hunting" servomechanisms, and the analysis provides a basis for a general solution of the response of a two-position switch or ideal limiter to various forms of applied voltages.

Since the unmodulated wave is a square wave with pulses of length $2L$ recurring at intervals of $T = 4L$, it has the familiar square wave spectrum including a d-c term, a fundamental term or carrier of frequency $c = 1/T$, a 3rd harmonic with a negative amplitude 1/3 that of the fundamental, etc. Figure 10 shows clearly that this spectrum is the sum of single pulses of width $2L$ spaced $T = 4L$ seconds apart. In the summation, all frequencies cancel except harmonics of c and, since they all add directly in phase, the component frequencies in the resultant spectrum have the same relative amplitudes as they have in one single pulse.

When this pulse train is modulated, the width of each pulse becomes $2(L + \Delta L)$, where the magnitude of ΔL depends in some specified way on the magnitude of thhe modulating function at the instant corresponding to the center of the pulse. For simplicity, the case will be taken where ΔL is proportional to the magnitude of the modulating function. For 100% modulation, ΔL will be assumed to vary from $-L$ to $+L$. Figure 3 shows how the relative amplitude of the components of the frequency spectrum of a pulse vary for 3 different values of ΔL , along with the equation that governs these amplitudes.

If the modulating function has a periodicity v such that $c = 10v$, then every 10th pulse, recurring at the same instant in each modulating cycle, will be widened to the same extent and so can be formed into a subsidiary unmodulated pulse train, as was done on Fig. 5 for the pulse position modulated wave.

Again vector diagrams like those in Fig. 7 may be formed showing the contribution of each of these subsidiary pulse trains at various frequencies such as c, $c + v$ and $c - v$. When the waves are unmodulated, the vector diagrams for the same frequencies will be the same as those for the pulse position modulated case, except for the absolute amplitudes of the components, as long as $c = 10v$ in each case. When the pulse width system is modulated, however, the modulation does not rotate the individual vector components as in the pulse position case since the spacing between pulses is not changed. What the pulse width modulation does is to change the length of the individual component vectors exactly as it does in the case of the single pulses shown on Fig. 3. This change of magnitude, of course, can spoil the cancellation of the ten unmodulated components at some frequency like $c + 2v$ just as effectively as rotating them did in the case of the pulse position modulated wave, thus producing a spectrum component at that frequency.

As an example, the case will be taken where the modulating function is a sinusoid of frequency v. Then the change in width with modulation is given by the formula

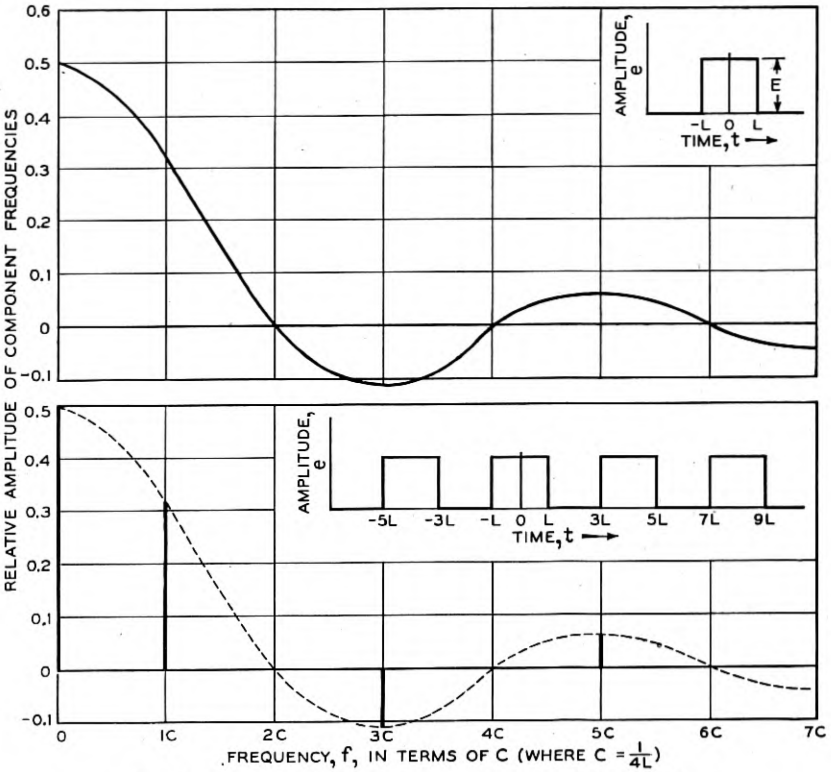


Fig. 10—Comparative spectra of square wave and single pulse.

$$\frac{\Delta L}{L} = k \sin vt.$$

Since $c = 10v$, the successive subsidiary pulse trains will be modulated an amount $\left(\frac{\Delta L}{L}\right)_m = k \sin\left(2\pi\frac{m}{10}\right)$ as m takes on the values from 1 to 10. Thus the spectra of these subsidiary pulse trains with pulses of length $2(L +$

ΔL_m) recurring every $1/v$ seconds will be a Fourier series of harmonics of v. The amplitude of the nth term of this series will be

$$B_n = \frac{2E}{10\pi n} \sin\left[\frac{\pi n}{2}\left[1 + k \sin\left(\frac{2\pi m}{10}\right)\right]\right].$$

This expression may be found from appendix C, equation (5a). Combining the 10 such components at each frequency, as shown on Fig. 7 for the case of the pulse position modulated wave, the spectrum for this case of Pulse Width Modulation on Fig. 11 is produced. This spectrum is comparable to that on Fig. 8 for the pulse position modulated case.

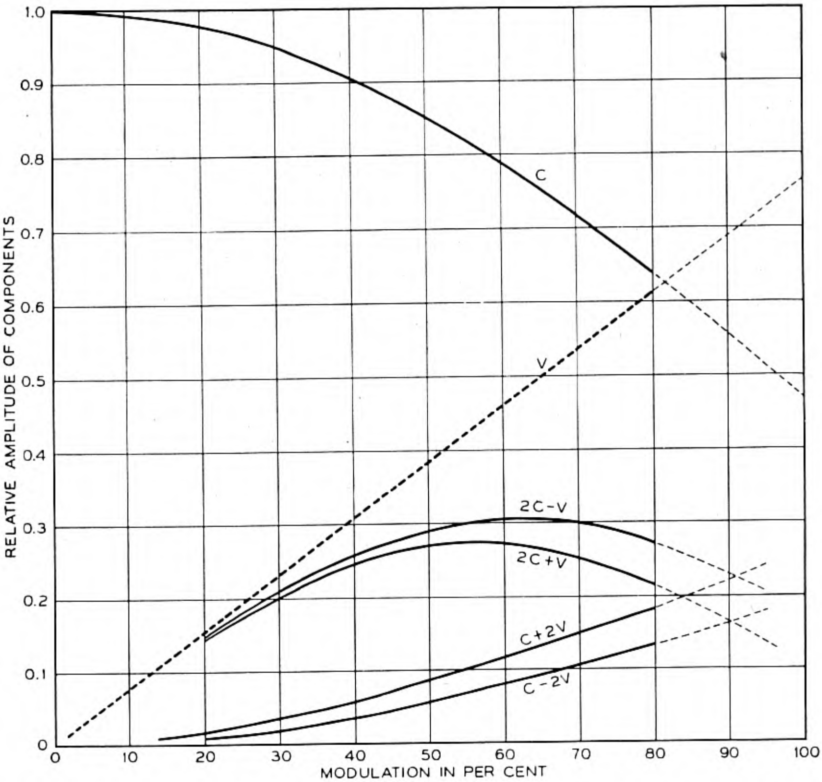


Fig. 11—Spectrum of pulse width modulated wave for case where carrier frequency C is 10 times the signal frequency v.

Pulse Width vs Amplitude Modulation

That pulse width modulation is a form of amplitude modulation of the carriers of the unmodulated pulse train is shown mathematically by Equa-

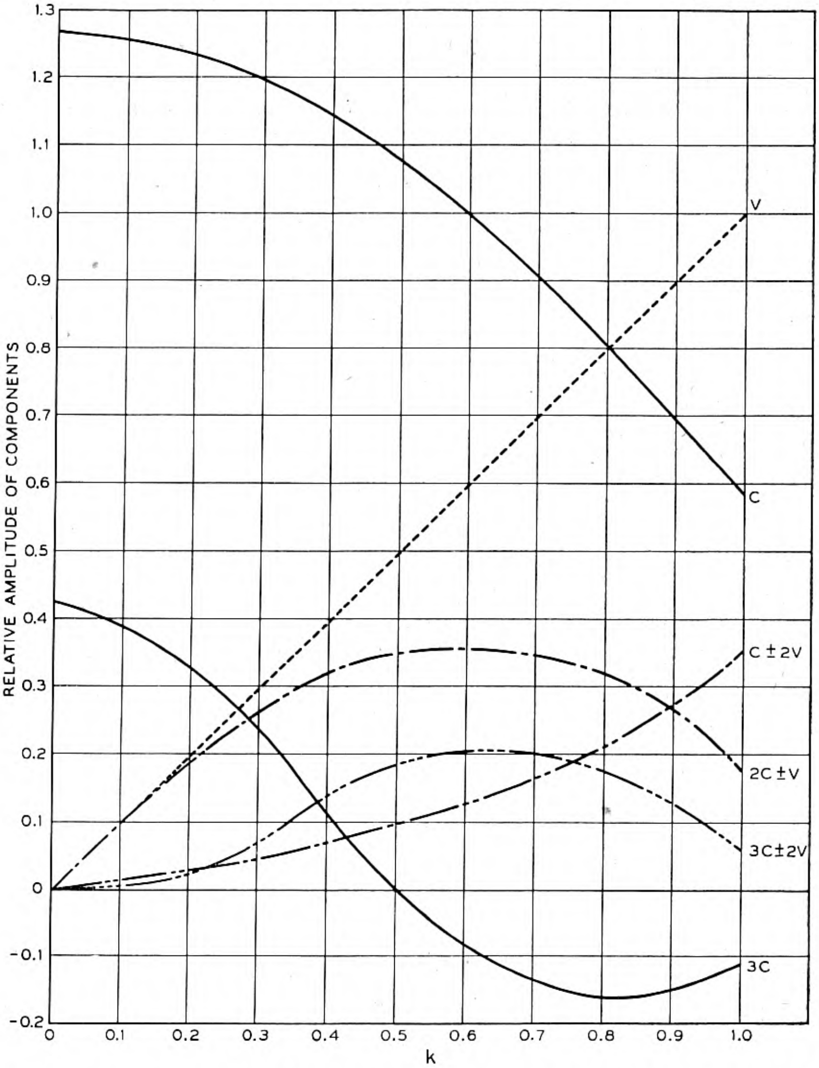


Fig. 12—Response of ideal limiter to simultaneously applied isosceles triangle wave and sine wave inputs. *k* is the ratio of the peak amplitudes of sinusoidal and triangular waves at the input.

tion (8) in Appendix C, where the spectrum is developed as a Fourier series in harmonics of the pulse rate *c* with the modulation affecting only the amplitude of the coefficients.

This mathematical analysis is continued in Appendix D where the frequency spectrum is determined for $\frac{\Delta L}{L} = \text{k} \sin vt$. The spectrum thus computed is shown in Fig. 12.

An example of this type of pulse modulator is given by a two position switch or ideal limiter when the signal to be modulated is applied simultaneously to the limiter with an isosceles triangle wave as carrier. The carrier should have a higher peak amplitude than the signal and a recurrence rate based on the desired carrier frequency. Figure 12 is arranged to show the output spectrum for such a limiter in terms of k, when k is the ratio of the peak amplitudes of the sinusoidal signal and triangular carrier wave inputs.

A comparison of this spectrum with that on Fig. 11 shows that the two spectra have almost the same form. c and v have the same amplitude characteristics in each case. The $c \pm 2v$ and $2c \pm v$ terms have differences that are like those found before in comparing the pulse position modulated wave on Fig. 8 and the phase modulated carrier on Fig. 9. As in that case, when c becomes very much greater than v the differences vanish.

Application of Pulse Width Modulator

Practical interest in this case lies in the fact that the signal is present in the output spectrum with a linear characteristic that makes such a modulator a linear amplifier. The "on-off" or "hunting" servomechanism is based on a modified form of such an amplifier in which the carrier is supplied by the self oscillation of the system. The term modified form is used because the self oscillations in general are more nearly sinusoidal than triangular in form and so do not give a linear change in pulse length over as wide a range of input amplitudes as does a triangular carrier. No attempt will be made to analyze such a system here since it has been handled elsewhere.[9] However the above method is applicable to such problems regardless of the shape of the carrier or the signal.

Other Forms of Pulse Modulation

Another form of pulse modulation of interest is that of pulse length modulation in which either the start or the end of each pulse is fixed, so that the centers of the pulses vary in position with the length. This is a combination of both the pulse position and the pulse width modulations described above and can be analyzed by a combination of the methods developed.

These same methods are also applicable to the analysis of frequency and phase modulated waves after they have been put through a limiter, as they generally are before detection.

[9] See L. A. Macall, "The Fundamental Theory of Servomechanisms" D. Van Nostrand Company, 1945.

APPENDIX A

Fourier Transforms For Single Pulse

The amplitude $g(f)$ of the component of frequency f in the spectrum of the Complex Magnitude-time function $e(t)$ is given by the d-c component of the Modulation products of $e(t)$ and *cos* $2\pi ft$, found by averaging the product over the period of the complex wave.

Thus, for non-periodic waves, where the period is from $-\infty$ to $+\infty$, the amplitude of the spectrum at f is

$$g(f) \cong \int_{-\infty}^{\infty} e(t) \cos 2\pi ft \, dt. \tag{1}$$

For the single pulse, where $e(t) = E$ for $-L < t < L$ and $e(t) = 0$ for all other values of t, equation (1) reduces to

$$g(f) \cong \int_{-L}^{L} E \cos 2\pi ft \, dt. \tag{2}$$

Integrating,

$$g(f) \cong \frac{F}{2\pi f} \sin 2\pi ft \Big|_{-L}^{L}$$

or

$$g(f) \cong \frac{E}{\pi f} \sin 2\pi fL. \tag{3}$$

Equation (3) is the expression for $g(f)$ plotted on Fig. 1.

Similarly, in the case of the single pulse, each increment in frequency df contributes a factor proportional to $g(f) \cos 2\pi ft \, df$ to the composition of $e(t)$, so that

$$e(t) = \int_{-\infty}^{\infty} g(f) \cos 2\pi ft \, df. \tag{4}$$

Substituting in (4) the expression for $g(f)$ given by equation (3), this becomes

$$e(t) \cong \frac{E}{\pi} \int_{-\infty}^{\infty} \frac{\sin 2\pi fL}{f} \cos 2\pi ft \, df. \tag{5}$$

APPENDIX B

Frequency Spectrum Of Phase Modulated Wave

The Phase Modulated Wave in this case is given by

$$\cos (ct - k \sin vt) = \cos (ct) \cos (k \sin vt) + \sin (ct) \sin (k \sin vt)$$

Now $\cos (ct) \cos (k \sin ct) = J_0 (k) \cos (ct)$
$\qquad + J_2 (k) \cos (c - 2v) t$

$$\begin{array}{ll}
 & + J_2 (k) \cos (c + 2v)\, t + \cdots \\
\text{and } \sin (ct) \sin (k \sin ct) & = J_1 (k) \cos (c - v)\, t \\
 & - J_1 (k) \cos (c + v)\, t \\
 & + J_3 (k) \cos (c - 3v)\, t \\
 & - J_3 (k) \cos (c + 3v)\, t + \cdots \\
\therefore \cos (ct - k \sin vt) & = J_0 (k) \cos (ct) \\
 & + J_1 (k) \cos (c - v)\, t \\
 & - J_1 (k) \cos (c + v)\, t \\
 & + J_2 (k) \cos (c - 2v)\, t \\
 & + J_2 (k) \cos (c + 2v)\, t \\
 & + J_3 (k) \cos (c - 3v)\, t \\
 & - J_3 (k) \cos (c + 3v)\, t + \cdots
\end{array}$$

APPENDIX C

In this Appendix the spectrum of a train of rectangular pulses of length $2(L + \Delta L)$ recurring every T seconds, will be found from the spectrum of a single pulse of this train.

For the single pulse at any frequency f,

$$g(f) \cong \frac{E}{\pi f} \sin 2\pi f(L + \Delta L). \tag{1}$$

For a series of such pulses recurring with a spacing $T = 1/c$, then the sum of spectra of the individual pulses form a Fourier series of harmonics of c. Thus

$$e(t) = A_0 + \sum_{n=1}^{\infty} A_n \cos 2\pi nct, \tag{2}$$

where A_n is the sum of an infinite number (one from each pulse) of infinitesimal terms $g(nc)$ and $g(-nc)$, shown in (1). Thus

$$A_n \cong 2\Sigma \frac{E}{\pi nc} \sin 2\pi nc(L + \Delta L) \tag{3}$$

Now to put an absolute value to the amplitudes g(f) shown in equation (1), it is necessary to average them over the recurrence period of the single pulse, making them infinitesimals. However, in the train of pulses recurring every $T = 1/c$ seconds, the amplitude of A_n can be determined by averaging the terms in (1) over an interval T. Then

$$A_n = \frac{2E}{\pi ncT} \sin 2\pi nc(L + \Delta L). \tag{4}$$

When $T = 4L = 1/c$, (4) reduce to

$$A_n = \frac{2E}{\pi n} \sin \frac{n\pi}{2} \left(1 + \frac{\Delta L}{L}\right) \tag{5}$$

For the example taken in the text, when the pulse train was subdivided into 10 subsiding pulse trains, the period $T = 1/v = 10/c = 40L$. Thus in this case, the Fourier coefficients of the harmonics of v are

$$B_n = \frac{2E}{10\pi n} \sin \frac{\pi n}{2} \left(1 + \frac{\Delta L}{L}\right). \tag{5a}$$

The expression for A_n in equation (5) can be put in simpler form by using the formula for the sin of the sum of two angles. In this way, we get

$$A_n = \frac{2E}{\pi n}\left[\sin\left(\frac{\pi n}{2}\right)\cos\left(\frac{\pi n}{2}\frac{\Delta L}{L}\right) + \cos\left(\frac{\pi n}{2}\right)\sin\left(\frac{\pi n}{2}\frac{\Delta L}{L}\right)\right]. \tag{6}$$

Now, for n odd, $\sin \frac{\pi n}{2}$ alternately assumes the value ±1 and $\cos \frac{\pi n}{2}$ vanishes, and for n even, $\cos\left(\frac{\pi n}{2}\right)$ alternately assumes the value ±1 and $\sin \frac{\pi n}{2}$ vanishes. The A_0 term, being the *d-c* average of the pulse train, is given by

$$\frac{E/2(L + \Delta L)}{T} = \frac{E}{2}\left(1 + \frac{\Delta L}{L}\right). \tag{7}$$

If the pulse train is transformed by shifting the zero so that it alternates between $\pm E/2$ instead of O and E, the first term in equation (7) vanishes and (2) becomes, from (6) & (7),

$$\begin{aligned} e(t) &= A_0 + A_1 \cos 2\pi ct \\ &\quad + A_2 \cos 2\pi\, 2ct + \cdots \end{aligned}$$

Where

$$\left.\begin{aligned} A_0 &= \frac{E}{2}\left(\frac{\Delta L}{L}\right) \\ A_1 &= \frac{2E}{\pi}\cos\left(\frac{\pi}{2}\frac{\Delta L}{L}\right) \\ A_2 &= \frac{2E}{2\pi}\sin \pi\left(\frac{\Delta L}{L}\right) \\ A_3 &= \frac{2E}{3\pi}\cos\frac{3\pi}{2}\left(\frac{\Delta L}{L}\right) \end{aligned}\right\} \tag{8}$$

etc.

APPENDIX D

The purpose of this section is to compute the spectrum of the carrier given by equation (8) in Appendix C as their amplitudes vary with $\frac{\Delta L}{L} = k \sin vt$.

For the *d-c* term,

$$A_0 = \frac{E}{2}\frac{\Delta L}{L} = \frac{E}{2} k \sin vt.$$

For the fundamental or c term,

$$A_1 \cos 2\pi ct = \frac{2E}{\pi} \cos\left(\frac{\pi}{2} k \sin vt\right) \cos 2\pi ct$$

Using the Bessel's expansion of cos (2 sin θ), we get,

$$A_1 \cos 2\pi ct = \frac{E}{2\pi}\begin{cases} J_0(k) \cos 2\pi c \\ +J_2(k) \cos 2\pi(c - 2v)t \\ +J_2(k) \cos 2\pi(c + 2v)t \\ + \cdots \text{ etc.} \end{cases}$$

In a similar fashion, the other terms can also be computed, giving the spectrum shown on Fig. 12, where $J_0(k)$ becomes the amplitude of c, $J_2(k)$ the amplitude of either $c + 2v$ or $c - 2v$, etc.

Abstracts of Technical Articles by Bell System Authors

Commercial Broadcasting Pioneer. The WEAF Experiment: 1922–1926.[1] William Peck Banning. *WEAF*, the radio call letters which for nearly a quarter of a century designated a broadcasting station famous for its pioneering achievements, ceased last November to have its old significance. *WNBC* are the new call letters. This book is an excellent record of the four years during which this station was the experimental radio broadcasting medium of the American Telephone and Telegraph Company.

The author indicates that the WEAF experiment aided the development of radio broadcasting in three ways:

First, in the scientific and technological field.

Second, in the emphasis of a high standard for radio programs.

Third, in determining the means whereby radio broadcasting could support itself.

When *WEAF* changed hands from the American Telephone and Telegraph Company to new ownership, public reaction to almost every type of broadcast had been tested, network broadcasting had been established and the economic basis upon which nationwide broadcasting now rests had been founded. A trail had been blazed that thereafter could be followed without hesitation.

In so far as radio broadcasting is concerned, this book is a significant chapter in communication history.

A Multichannel Microwave Radio Relay System.[2] H. S. Black, J. W. Beyer, T. J. Grieser, F. A. Polkinghorn. An 8-channel microwave relay system is described. Known to the Army and Navy as AN/TRC-6, the system uses radio frequencies approaching 5,000 megacycles. At these frequencies, there is a complete absence of static and most man-made interference. The waves are concentrated into a sharp beam and do not travel along the earth much beyond seeing distances. Other systems using the same frequencies can be operated in the near vicinity. The transmitter power is only one four-millionth as great as would be required with nondirectional antennas. The distance between sets is limited but by using intermediate repeaters communications are extended readily to longer distances. Short pulses of microwave power carry the intelligence of the eight messages utilizing *pulse position modulation* to modulate the

[1] Published by Harvard University Press, Cambridge, Massachusetts, 1946.

[2] *Elec. Engg., Trans. Sec.*, December 1946.

pulses and *time division* to multiplex the channels. The eight message circuits which each *AN/TRC*-6 system provides are high-grade telephone circuits and can be used for signaling, dialing, facsimile, picture transmission, or multichannel voice frequency telegraph. Two-way voice transmission over radio links totaling 1,600 miles, and one-way over 3,200 miles have been accomplished successfully in demonstrations.

Further Observations of the Angle of Arrival of Microwaves.[3] A. B. CRAWFORD and WILLIAM M. SHARPLESS. Microwave propagation measurements made in the summer of 1945 are described. This work, a continuation of the 1944 work reported elsewhere in this issue of the *Proceedings of the I.R.E. and Waves and Electrons*, was characterized by the use of an antenna with a beam width of 0.12 degree for angle-of-arrival measurements and by observations of multiple-path transmission.

The Effect of Non-Uniform Wall Distributions of Absorbing Material on the Acoustics of Rooms.[4] HERMAN FESHBACH and CYRIL M. HARRIS. The acoustics of rectangular rooms, whose walls have been covered by the non-uniform application of absorbing materials, is treated theoretically. Using appropriate Green's functions a general integral equation for the pressure distribution on the walls is derived. These equations show immediately that it is necessary to know *only* the pressure distribution on the treated surfaces to predict completely the acoustical properties of the room, such as the resonant frequencies, the decay constants, and the spatial pressure distribution. The integral equation is solved approximately using (1) perturbation method, and (2) approximate reduction of the integral equation to an equivalent transmission line. Criteria giving the range of validity of these approximations are derived. It was found useful to introduce a new concept, that of "*effective admittance*," to express the results for the resonant frequency and absorption for then the amount of computation is reduced and the accuracy of the results is increased. The absorption of a patch of material was found as a function of the position of the absorbing material and was checked experimentally for a convenient case, an absorbing strip mounted on the otherwise hard walls of a rectangular room. Particular attention is given to the case where the acoustic material is applied in the form of strips. The results may then be expressed in series which converge very rapidly and are, therefore, amenable to numerical calculation. Approximate formulas are obtained which permit estimates of the diffusion of sound in a non-uniformly covered room. In agreement with experience, these equations show that diffusion increases with frequency and with the

[3] *Proc. I.R.E. and Waves and Electrons*, November 1946.
[4] *Jour. Acous. Soc. America*, October 1946.

number of nodes on the treated walls. The "interaction effect" of one strip on another is shown to decrease with an increase of the number of nodes. The results are then applied to the case of ducts with non-uniform distribution of absorbing material on its walls. Results are given which permit the calculation of the attenuation per unit length of duct. The methods of this paper hold for any distribution of absorbing material and also if the admittance is a function of angle of incidence.

High Current Electron Guns.[5] L. M. FIELD. This paper presents a survey of some of the problems and methods which arise in dealing with the design of high current and high current-density electron guns. A discussion of the general limitations on all electron gun designs is followed by discussion of single and multiple potential guns using electrostatic fields only. A further discussion of guns using combined electrostatic and magnetic fields and their limitations, advantages, and some possible design procedures follows.

Reflection of Sound Signals in the Troposphere.[6] G. W. GILMAN, H. B. COXHEAD, and F. H. WILLIS. Experiments directed toward the detection of non-homogeneities in the first few hundred feet of the atmosphere were carried out with a low power sonic "radar." The device has been named the *sodar*. Trains of audiofrequency sound waves were launched vertically upward from the ground, and echoes of sufficient magnitude to be displayed on an oscilloscope were found. Strong displays tended to accompany strong temperature inversions. During these periods, transmission on a microwave radio path along which the sodar was located tended to be disturbed by fading. In addition, relatively strong echoes were received when the atmosphere was in a state of considerable turbulence. There was a well-defined fine-weather diurnal characteristic. The strength of the echoes was such as to lead to the conclusion that a more complicated distribution of boundaries than those measured by ordinary meteorological methods is required in the physical picture of the lower troposphere.

A Cathode-Ray Tube for Viewing Continuous Patterns.[7] J. B. JOHNSON. A cathode-ray tube is described in which the screen of persistent phosphor is laid on a cylindrical portion of the glass. A stationary magnetic field bends the electron beam on to the screen, while rotation of the tube produces the time axis. When the beam is deflected and modulated, a continuous pattern may be viewed on the screen.

[5] *Rev. Mod. Phys.*, July 1946.
[6] *Jour. Acous. Soc. Amer.*, October 1946.
[7] *Jour. Applied Physics*, November 1946.

The Molecular Beam Magnetic Resonance Method. The Radiofrequency Spectra of Atoms and Molecules.[8] J. B. M. Kellogg and S. Millman. A new method known as the "Magnetic Resonance Method" which makes possible accurate spectroscopy in the low frequency range ordinarily known as the "radiofrequency" range was announced in 1938 by Rabi, Zacharias, Millman, and Kusch (R6, R5). This method reverses the ordinary procedures of spectroscopy and instead of analyzing the radiation emitted by atoms or molecules analyzes the energy changes produced by the radiation in the atomic system itself. Recognition of the energy changes is accomplished by means of a molecular beam apparatus. The experiment was first announced as a new method for the determination of nuclear magnetic moments, but it was immediately apparent that its scope was not limited to the measurement of these quantities only. It is the purpose of this article to summarize the more important of those successes which the method has to date achieved.

Metal-Lens Antennas.[9] Winston E. Kock. A new type of antenna is described which utilizes the optical properties of radio waves. It consists of a number of conducting plates of proper shape and spacing and is, in effect, a lens, the focusing action of which is due to the high phase velocity of a wave passing between the plates. Its field of usefulness extends from the very short waves up to wavelengths of perhaps five meters or more. The paper discusses the properties of this antenna, methods of construction, and applications.

Underwater Noise Due to Marine Life.[10] Donald P. Loye. The widespread use of underwater acoustical devices during the recent war made it necessary to obtain precise information concerning ambient noise conditions in the sea. Investigations of this subject soon led to the discovery that fish and other marine life, hitherto generally classified with the voiceless giraffe in noisemaking ability, have long been given credit for a virtue they by no means always practice. Certain species, most notably the croaker and the snapping-shrimp, are capable of producing noise which, in air, would compare favorably with that of a moderately busy boiler factory. This paper describes some of the experiments which traced these noises to their source and presents acoustical data on the character and magnitude of the disturbances.

Elastic, Piezoelectric, and Dielectric Properties of Sodium Chlorate and Sodium Bromate.[11] W. P. Mason. The elastic, piezoelectric, and di-

[8] *Rev. Mod. Phys.*, July 1946.
[9] *Proc. I.R.E. and Waves and Electrons*, November 1946.
[10] *Jour. Acous. Soc. America*, October 1946.
[11] *Phys. Rev.*, October 1 and 15, 1946.

electric constants of sodium chlorate ($NaClO_3$) and sodium bromate ($NaBrO_3$) have been measured over a wide temperature range. The value of the piezoelectric constant at room temperature is somewhat larger than that found by Pockels. The value of the Poisson's ratio was found to be positive and equal to 0.23 in contrast to Voigt's measured value of -0.51. At high temperatures the dielectric and piezoelectric constants increase and indicate the presence of a transformation point which occurs at a temperature slightly larger than the melting point. A large dipole piezoelectric constant (ratio of lattice distortion to dipole polarization) results for these crystals but the electromechanical coupling factor is small because the dipole polarization is small compared to the electronic and ionic polarization and little of the applied electrical energy goes into orienting the dipoles.

Paper Capacitors Containing Chlorinated Impregnants. Effects of Sulfur.[12] D. A. McLean, L. Egerton, and C. C. Houtz. Sulfur is an effective stabilizer for paper capacitors containing chlorinated aromatics, in the presence of both tin foil and aluminum foil electrodes. Sulfur has unique beneficial effects on power factor which are especially marked when tin foil electrodes are used. The value of R (Equation 4) can be used as an index of ionic conductivity in the impregnating compound. Diagnostic power factor measurements on impregnated paper are best made at low voltages. Electron diffraction studies give results in line with the previously published theory of stabilization. Several previous findings are reaffirmed: (a) the importance of all components of the capacitor in determining its initial properties and aging characteristics, (b) the superiority of kraft paper over linen, and (c) widely different behavior of capacitors employing different electrode metals.

A New Bridge Photo-Cell Employing a Photo-Conductive Effect in Silicon. Some Properties of High Purity Silicon.[13] G. K. Teal, J. R. Fisher, and A. W. Treptow. A pure photo-conductive effect was found in pyrolytically deposited and vaporized silicon films. An apparatus is described for making bridge type photo-cells by reaction of silicon tetrachloride and hydrogen gases at ceramic or quartz surfaces at high temperatures. The maximum photo-sensitivity occurs at 8400–8600A with considerable response in the visible region of the spectrum. The sensitivity of the cell appears about equivalent to that of the selenium bridge and its stability and speed of response are far better. For pyrolytic films on porcelain there are three distinct regions in the conductivity as a function of temperature. At low temperatures the electronic conductivity is given by the expression

[12] *Indus. & Engg. Chemistry*, November 1946.
[13] *Jour. Applied Physics*, November 1946.

$\sigma = Af(T)exp-(E/2kT)$. At temperatures between 227°C and a higher temperature of 400–500°C $\sigma = Aexp-(E/2kT)$, where E lies between 0.3 and 0.8 ev; and at high temperatures $\sigma = Aexp-(E/2kT)$, where $E = 1.12$ ev. The value 1.12 ev represents the separation of the conducting and non-conducting bands in silicon. The long wave limit of the optical absorption of silicon was found to lie at approximately 10,500A (1.18 ev). The data lead to the conclusion that the same electron bands are concerned in the photoelectric, optical, and thermal processes and that the low values of specific conductances found (1.8×10^{-5} ohm^{-1} cm^{-1}) are caused by the high purity of the silicon rather than by its polycrystalline structure.

Non-Uniform Transmission Lines and Reflection Coefficients.[14] L. R. WALKER and N. WAX. A first-order differential equation for the voltage reflection coefficient of a non-uniform line is obtained and it is shown how this equation may be used to calculate the resonant wave-lengths of tapered lines.

[14] *Jour. Applied Physics*, December 1946.

Contributors to this Issue

HARALD T. FRIIS, E.E., Royal Technical College, Copenhagen, 1916; Sc.D., 1938; Assistant to Professor P. D. Pedersen, 1916; Technical Advisor at the Royal Gun Factory, Copenhagen, 1917–18; Fellow of the American Scandinavian Foundation, 1919; Columbia University, 1919. Western Electric Company, 1920–25; Bell Telephone Laboratories, 1925–. Formerly as Radio Research Engineer and since January 1946 as Director of Radio Research, Dr. Friis has long been engaged in work concerned with fundamental radio problems. He is a Fellow of the Institute of Radio Engineers.

RAY S. HOYT, B.S. in Electrical Engineering, University of Wisconsin, 1905; Massachusetts Institute of Technology, 1906; M.S., Princeton, 1910. American Telephone and Telegraph Company, Engineering Department, 1906–07. Western Electric Company, Engineering Department, 1907–11. American Telephone and Telegraph Company, Engineering Department, 1911–19; Department of Development and Research, 1919–34. Bell Telephone Laboratories, 1934–. Mr. Hoyt has made contributions to the theory of loaded and non-loaded transmission lines and associated apparatus, theory of crosstalk and other interference, and probability theory with particular regard to applications in telephone transmission engineering.

W. D. LEWIS, A.B. in Communication Engineering, Harvard College, 1935; Rhodes Scholar, Wadham College, Oxford; B.A. in Mathematics, Oxford, 1938; Ph.D. in Physics, Harvard, 1941. Bell Telephone Laboratories, 1941–. Dr. Lewis was engaged in radar antenna work in the Radio Research Department during the war; he is now engaged in microwave repeater systems research.

J. C. LOZIER, A.B. in Physics, Columbia College, 1934; graduate physics student, Princeton University, 1934–35. R.C.A. Victor Manufacturing Company, 1935–36; Bell Telephone Laboratories, Inc., 1936–. Mr. Lozier has been engaged in transmission development work, chiefly on radio telephone terminals. During the war he was concerned primarily with the theory and design of servomechanisms.