THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLIV SEPTEMBER 1965 NUMBER 7

All-Weather Earth Station Satellite Communication Antennas

By J. S. COOK and A. J. GIGER

(Manuscript received May 27, 1965)

I. INTRODUCTION

The successful TELSTAR® and Relay satellite experiments have led to consideration of the equipment and techniques that might be incorporated in future commercial and military satellite communication systems.

The experiments re-emphasized two facts that had been recognized for some time: (1.) that worthwhile improvement in the earth station receiver sensitivity could be obtained by elimination of the antenna radome, and (2.) that an important practical improvement would be brought about by the location of the communication and tracking equipment on a platform that does not move with the antenna.

These improvements can be achieved by special design of the antenna itself, as indicated in the subsequent papers in this issue. Here the major considerations which motivated that work are reviewed.

II. RADOME

The advantages of having radome protection for the antenna are well known. Foremost is the elimination of wind and weather on the antenna. Second, and also important, is the convenience it permits in antenna construction and maintenance. The disadvantages are also twofold: (1.) extra noise is radiated and reflected from the radome into the re-

ceiver, particularly when the radome is wet by rain or snow, and (2.) interference with local microwave communications can result from extraneous radiation caused by the radome. The interference problem is already significant, and the future promises to make it still more serious.

2.1 *Transmission Degradation*

Measurements† conducted at the Andover Satellite Station during the past three years show that in rain or snow the characteristics of the communications system can be seriously degraded. Measurements of the same nature made by other workers at Bell Telephone Laboratories on antennas without radomes have shown substantially less adverse effects of rain and snow. It can be clearly concluded that the water layer on the radome surface significantly degrades the otherwise good electrical characteristics of the radome.

This degradation occurs in two ways. First, the thermal noise level in the receiving system increases. Temperatures of 160° Kelvin have been measured repeatedly in Andover under wet weather conditions as compared with about 30° Kelvin in dry weather. An uncovered antenna would have increased only about 30° Kelvin under the same conditions. Second, signal loss increases at both the receiving and transmitting frequencies. The loss at the receiving frequency is actually greater than would be expected from the measured increase in the noise temperature. This is because part of the signal is back-scattered into the cold sky and therefore does not give rise to an increase in operating noise temperature. Signal losses of 5 db at 4 gc and even more at a frequency of 6 gc are possible during periods of heavy rain or slushy snow. The losses in case of the uncovered antenna would be substantially lower and entirely predictable from the increase in system noise. The radome-related degradations can make a satellite communications system, which is designed to operate close to the threshold of detection, inoperative for certain periods of time. It is fortunate that in many locations, including the Andover site, these outages are quite rare. Clearly, frequent degradations of reduced severity are also undesirable. A quantitative study of the ground station requirements shows that elimination of the radome will enable the system designer to meet CCIR performance specifications* with a smaller ground antenna than would otherwise be needed.

---

† Giger, A. J., 4-gc. Transmission Degradation Due to Rain at the Andover, Maine, Satellite Station, B.S.T.J., this issue, p. 1528.

* CCIR covers the percentage of time a certain noise level should not be exceeded in a telephone channel.

2.2 *Interference*

The susceptibility of an antenna to interfering signals outside the main beam is primarily a function of its side- and back-lobe level. The horn-reflector antenna is known for its extremely low-back lobes which go down to 50 db or more below the isotropic level of the antenna. Such a feature is desirable because it eases the problem of working together with other microwave systems operating in the same frequency band. Site selection for a satellite ground station is simplified since the separation from the next microwave station can be reduced.

The presence of a radome alters the radiation characteristic of an antenna, especially in the low side-lobe region. Although this effect is not appreciable during dry weather for the thin inflatable radome used at the Andover station, it is significant when the radome surface is wet. The degradation of the side-lobe pattern depends on the particular geometrical relation between antenna aperture, radome and surrounding terrain. It is therefore difficult to predict in general.

III. ENVIRONMENTAL PROBLEMS

Exposure to the elements brings about a number of antenna problems. Probably the most serious is that of wind loading. Wind on the antenna structure can cause tracking and control problems, mechanical drive difficulties, and structural distortion. It is important that the antenna be of relatively compact configuration, both to reduce wind cross section and to permit structural rigidity. Ability to cope with the wind problem is a major consideration in the selection, design, and evaluation of all-weather antennas.

The thermal effects brought about by changes in ambient temperatures may be more serious than they are in a protected antenna. The effects of local heating by the sun also are important, therefore, the thermal radiating character and expansion characteristics of the reflector surfaces must be carefully considered.

Rain, snow, and ice problems must also be considered, but the nature of the possible solutions for them is such that they do not strongly influence the basic antenna design. A thin layer of rain water on an antenna reflector does not significantly change the antenna radiation characteristics.

IV. ANTENNA CONFIGURATIONS

At least two antenna configurations appear to be particularly well suited for all-weather use. These have evolved through a series of in-

vestigations of large-aperture, multiple-reflector, horn and cassegrain configurations. The two preferred approaches — the triply-folded horn and the open cassegrain — are discussed in the articles that follow. Each has certain advantages that may be inferred from the measurements, calculations, and practical considerations presented there. Both configurations are meant for operation without a radome, and each permits placement of the communication electronics on a stationary platform.

V. CONCLUSION

The following six papers present recent work that has been motivated by these concepts. If a conclusion can be drawn from the effort as a whole, it is that high-quality, practical, all-weather earth station satellite communication antennas lie comfortably within the boundaries of today's engineering technology.

Errata

A Precise Measurement of the Gain of a Large Horn-Reflector Antenna, D.C. Hogg, and R.W. Wilson, B.S.T.J., *44*, July-August, 1965, pp. 1019–1030.

On page 1023, replace Fig. 3 by Table III. On page 1025, replace 35.10 ± 0.3 db and 35.04 ± 0.03 db by 31.10 ± 0.3 db and 31.04 ± 0.03 db.

The Triply-Folded Horn Reflector: A Compact Ground Station Antenna Design for Satellite Communications

By A. J. GIGER and R. H. TURRIN

(Manuscript received May 20, 1965)

An antenna suitable for ground stations of satellite communications systems is described. The antenna has very good low-noise properties, high aperture efficiency, and excellent broadband characteristics. It can be operated without a radome and allows the location of all communications and tracking equipment in a stationary room on the ground. Called the "triply-folded horn-reflector antenna," it is derived from the well-known conical horn-reflector antenna by folding the horn three times to bring its apex into a stationary position on the ground. Plane reflectors are used in the folding process and the propagation in the antenna is based on the principles of geometrical optics.

The paper describes electrical tests on an antenna model at frequencies of 60 and 11 gc and presents results of hydrodynamic tests which were performed to study the behavior of the antenna in high wind.

I. INTRODUCTION

Among the antennas suitable for ground stations of a satellite communications system, the horn reflector is highly desirable since it unites in one design such electrical characteristics as high-aperture efficiency, low-noise temperature, immunity from interfering signals and extreme broadband capabilities. High-aperture efficiency results from guiding the fundamental mode electromagnetic energy from the focal point by means of the horn directly to the parabolic reflector. In this way the aperture field distribution is not strongly tapered and energy loss due to spillover is minimized. The low noise temperature and immunity from interference results mainly from the inherent shielding afforded by the horn-reflector structure.

Both the pyramidal and conical horn-reflector antenna designs have

been investigated and applied in the past few years.[1,2,3] Notable among the applications is the large conical horn-reflector antenna at the ground station for satellite communications near Andover, Maine. The conical horn structure was chosen for this application because of its structural and electrical advantages. While this antenna has performed well, a number of areas exist where improvements would be desirable. For instance, the need for carrying large amounts of equipment on the rotating structure of the antenna contributes to its high weight and cost. Furthermore, the radome which is necessary for operation of a large horn-reflector antenna under high wind conditions, degrades antenna performance. During periods of rain or snow the radome seriously increases the thermal noise and the signal loss in the communications system.

It is the purpose of this paper to describe a new antenna structure which eliminates the major disadvantages of high weight and need for a radome, without seriously degrading its electrical or operational characteristics. Results of scale model studies of electrical and wind loading characteristics are presented and other aspects of this new design are discussed.

II. THE TRIPLY-FOLDED HORN-REFLECTOR ANTENNA

2.1 *The Configuration*

The new antenna configuration evolved from the conical horn-reflector antenna shown in Fig. 1(a). It is apparent that by using the principles of geometrical optics, the length of the structure can be reduced considerably by introducing a plane reflector R_1 in the horn as shown in Fig. 1(b). In this way, the apex of the horn which coincides with the "folded" focal point F_2 of the paraboloid can be located near the lower plane of the structure. A large rotating joint as shown in Fig. 1(b) is now necessary to permit elevation rotation of the paraboloidal reflector section. This configuration gives the advantage of consolidating the terminal equipment at one level instead of the two levels shown in Fig. 1(a). However, all the equipment still would be required to rotate with the structure in azimuth with the associated penalty of inertia and weight loading, as well as the complexities of feeding power and signal leads through slip rings or cable wraps.

A further extension of this concept consists of introducing two additional plane reflectors R_2 and R_3 to permit the axis of the horn apex section to coincide with the azimuth axis of rotation (Fig. 1(c)). Thus, by introducing another rotating joint near the apex, the final feed section

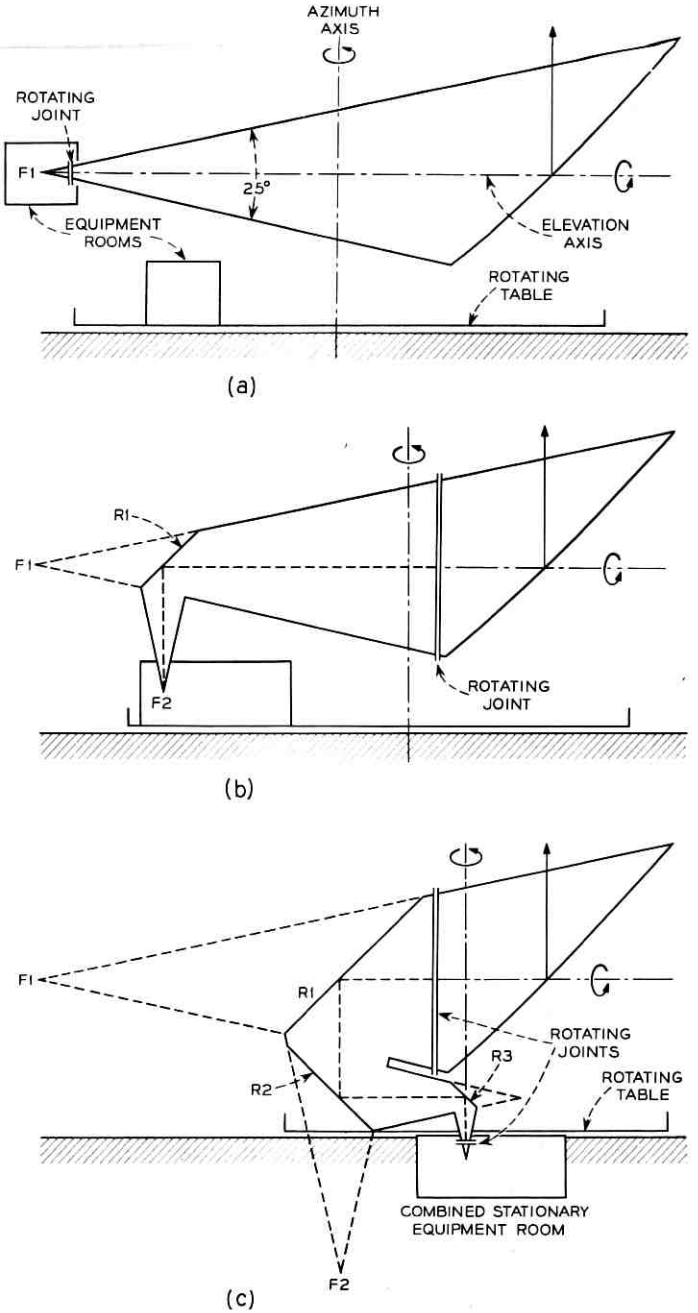

Fig. 1 — Development of the triply-folded horn geometry from the straight horn: (a) straight horn-reflector antenna; (b) single-fold intermediate stage; (c) triply-folded horn-reflector antenna design with horn apex coincident with the azimuth rotational axis.

can be made stationary and all of the terminal equipment mounted off the rotating antenna structure.

Although 90 degree reflections are shown in the triply-folded horn-reflector antenna of Fig. 1(c), the underlying principle of geometrical optics allows the use of reflection-angles other than 90 degrees.

2.2 *Operation Without Radome*

In order to eliminate the above mentioned transmission degradation during rain or snow, the triply-folded horn should be operated without a radome. The inherent compactness of the triply-folded horn antenna is a definite asset for achieving adequate satellite tracking during periods of high wind. A helpful by-product of the elimination of the detrimental radome effects is a reduction of the size of an uncovered antenna for given transmission requirements in the satellite system.

Several variations of the folded horn configuration were considered. As a result of hydrodynamic scale model tests, best resistance to high wind velocities was obtained by the configuration shown in Fig. 2. Here the azimuth axis is located to minimize the swept radius of the structure which includes an exterior fairing tightly fitted to the silhouette of the folded horn.

In environments such as Andover, Maine, means for melting snow and ice from the exposed parabolic reflector surface must be provided. One possible approach to this problem is the use of electrical deicing devices which can be zone controlled to localize heating effects. Another possibility consists of closing the aperture with a low-loss cover. While a wet aperture cover does not cause as great a transmission degradation as a radome, it should not be installed without some provisions for removing the water from its surface. Fig. 3 indicates a possible technique in which high-speed air is directed by nozzles tangentially over the cover to reduce or even blow away any possible water layer. A plenum is provided in the back of the cover where a slight over-pressure is produced by the blowers. Formed ducts, made of low-loss foam material which are designed according to aerodynamical principles and end in nozzles, accelerate the air to the high velocity required at the outside surface of the aperture cover.

2.3 *Mechanical Considerations*

The triply-folded conical horn-reflector antenna structure shown in Fig. 2 would be constructed of steel tubular members for the reflector panel supports and of standard rolled shapes of structural steel elsewhere.

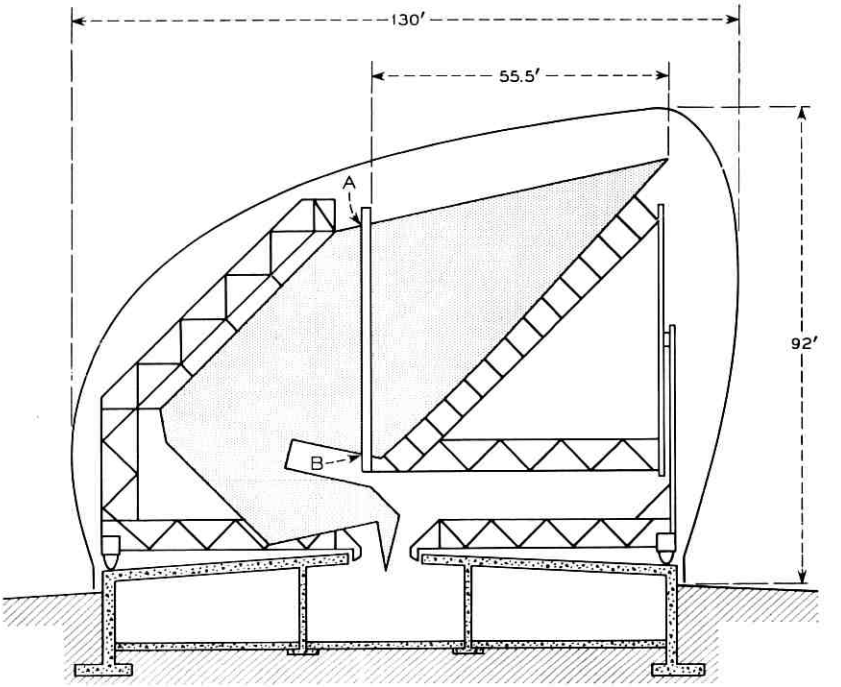

Fig. 2 — Cross sectional view of proposed triply-folded horn-reflector antenna showing asymmetrical outer fairing. The dimensions are for a 2250 square foot aperture size.

Aluminum would have disadvantages from a cost and fabrication viewpoint. When operating without an aperture cover, it is necessary for the parabolic reflector panels to be resistant to distortion under varying solar heating as the surface is exposed to rapidly varying sun and shadow. It also is necessary to melt ice and snow from the reflector. A stretch formed, single-skin aluminum panel would be suitable for this application rather than the aluminum honeycomb panels which were used at Andover. The honeycomb construction is vulnerable to serious warping if the outer skins are not at reasonably uniform temperatures and the core acts as an effective insulator against efficient heat transfer for snow and ice melting purposes. The single-skin aluminum panel would have to be about 50 per cent heavier than the equivalent honeycomb panel. The exterior fairing could be built of lightweight panels of aluminum or resin impregnated Fiberglas.

The entire structure rotates on a single circular track mounted on a

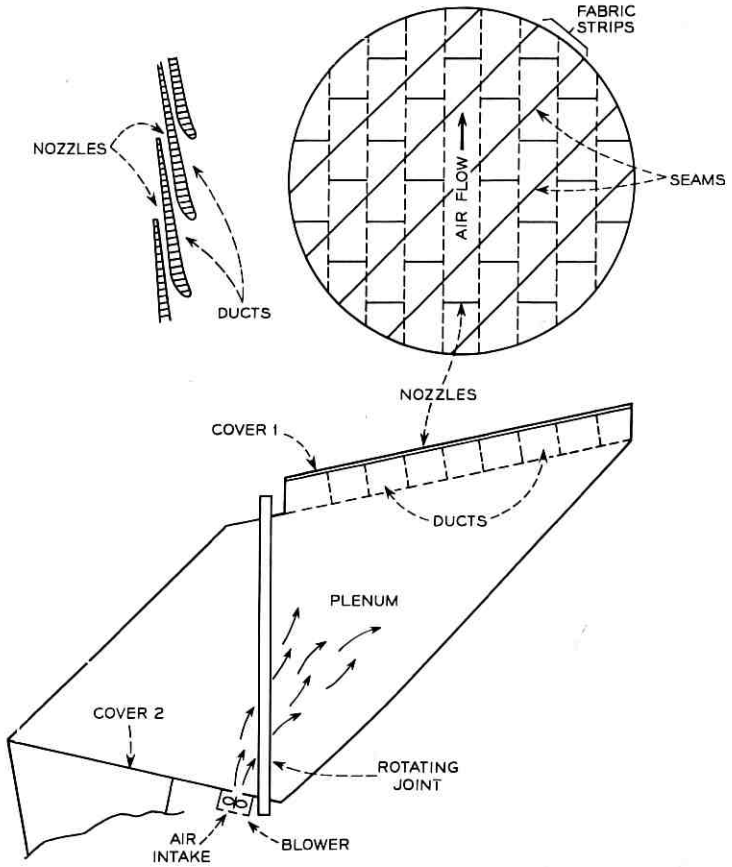

Fig. 3 — A method proposed to deflect precipitation from the aperture by means of high velocity air streams. Covers 1 and 2 might be Hypalon covered Dacron fabric similar to the Andover radome material but only 0.040 inch thick.

circular foundation. The foundation can incorporate enough living space to house all terminal equipment in an underground equipment area.

The overall RMS deflection of the four reflecting surfaces of the triply-folded horn should not exceed a value which depends on the maximum allowable gain degradation at the highest frequency of operation. The total allowable RMS deflection can be determined from formulas contained in Ref. 4. It is logical to divide the total RMS value among the reflecting surfaces such that the accuracy requirements increase from the largest to the smallest surface. A possible way of assigning the total RMS error, σ_{tot} among the four reflecting surfaces could be as follows: $\sigma_1 = 0.84\sigma_{tot}$, $\sigma_2 = 0.42\sigma_{tot}$, $\sigma_3 = 0.28\sigma_{tot}$, and $\sigma_4 = 0.21\sigma_{tot}$.

Aligning the reflector panels to the desired accuracy is a very important task which becomes relatively simple for the flat reflectors. The parabolic reflector can for instance be measured and aligned by triangulation using two telescopes mounted in positions A and B on the antenna. A small computer coupled to the telescopes quickly determines the deviation from the theoretical surface. A and B are supports whose location can be easily and accurately determined, and from which the parabolic reflector can always be seen equally well when rotated in elevation. Measurements of the parabolic surface can therefore be made for any elevation angle of the antenna, a very valuable feature unique to this antenna design.

III. ELECTRICAL SCALE MODEL TESTS

3.1 *The Antenna Model*

Analytic investigation of the triply-folded conical horn-reflector antenna is a formidable problem involving the solution of the propagation equations in the oversized conical waveguide represented by the folded horn. Electrical scale model testing, however, has been shown to be valuable in assessing the electrical characteristics of large antennas.[2] In the present case, the availability of a precision conical horn-reflector antenna model minimized the fabrication cost but dictated the size of the model.

In order to make direct comparison between folded and straight conical horn-reflector antennas, the model was constructed so that the precision parabolic reflector section could be attached to either the folded or the straight conical horn. Fig. 4 is a scaled cross section drawing of the folded conical horn-reflector model with pertinent dimensions. Shown by dotted lines is the position of the straight conical horn when attached to the reflector. The model has a focal length of 24 inches and a flare angle of 31.5 degrees. The RMS error of the parabolic surface of this model is about 0.002 inch which, according to Ruze[4] will not cause more than 0.05 decibels reduction in gain from the theoretical value at the highest measuring frequency of 60 gc.

The original conical horn-reflector model was fabricated of thin sheet brass while the folded conical horn was machined from aluminum castings. Flat precision aluminum plates were attached as fold reflectors. Each fold section may be replaced by a straight conical section thus permitting investigation of a single fold or combination of folds. Conical horn feeds for both the TE_{11} and TM_{01} modes were provided separately.

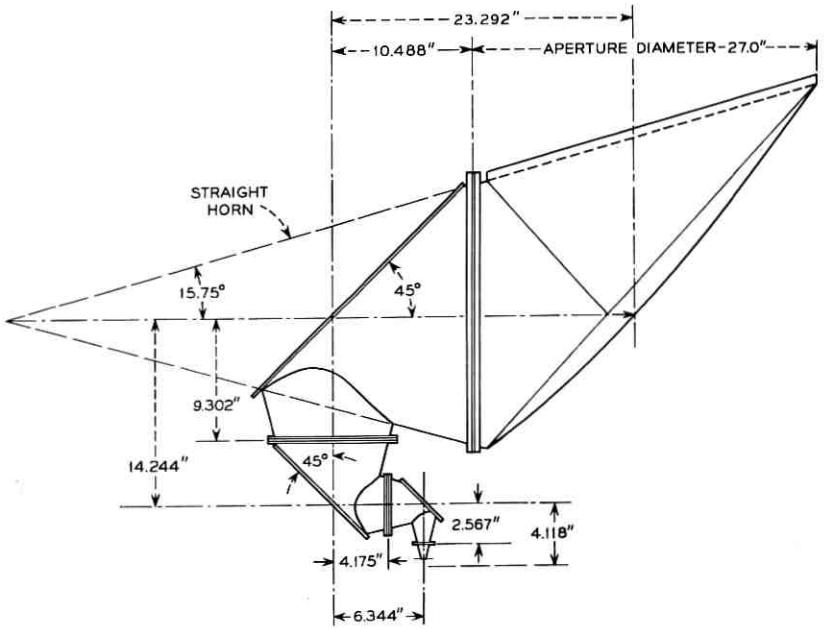

Fig. 4 — Conical horn-reflector antenna model employed for electrical tests showing pertinent dimensions of the inner surface.

Fig. 5 is a photograph of the triply-folded conical horn-reflector antenna model mounted for pattern measurements.

Aside from having a greater flare angle (31.5 vs 25 degrees) than the antenna in Fig. 2, the aperture diameter of the model is undersized by a factor of 1.6. This factor is based on the assumption that the ground station antenna of Fig. 2 would operate at 4 gc vs a 60 gc measuring frequency for the model. Both the larger flare angle and the undersized scale enhance the diffraction effects of the folds in the model compared with the antenna of Fig. 2. The model measurements should therefore give a somewhat pessimistic answer to the diffraction problems in the triply-folded horn.

4.2 *The Measuring Technique*

One purpose of the electrical measurements was to obtain far field radiation patterns having sufficient accuracy to be employed in the analytical determination of gain and noise temperature of the model. Sufficient accuracy is obtained by measuring at least 30 decibels below

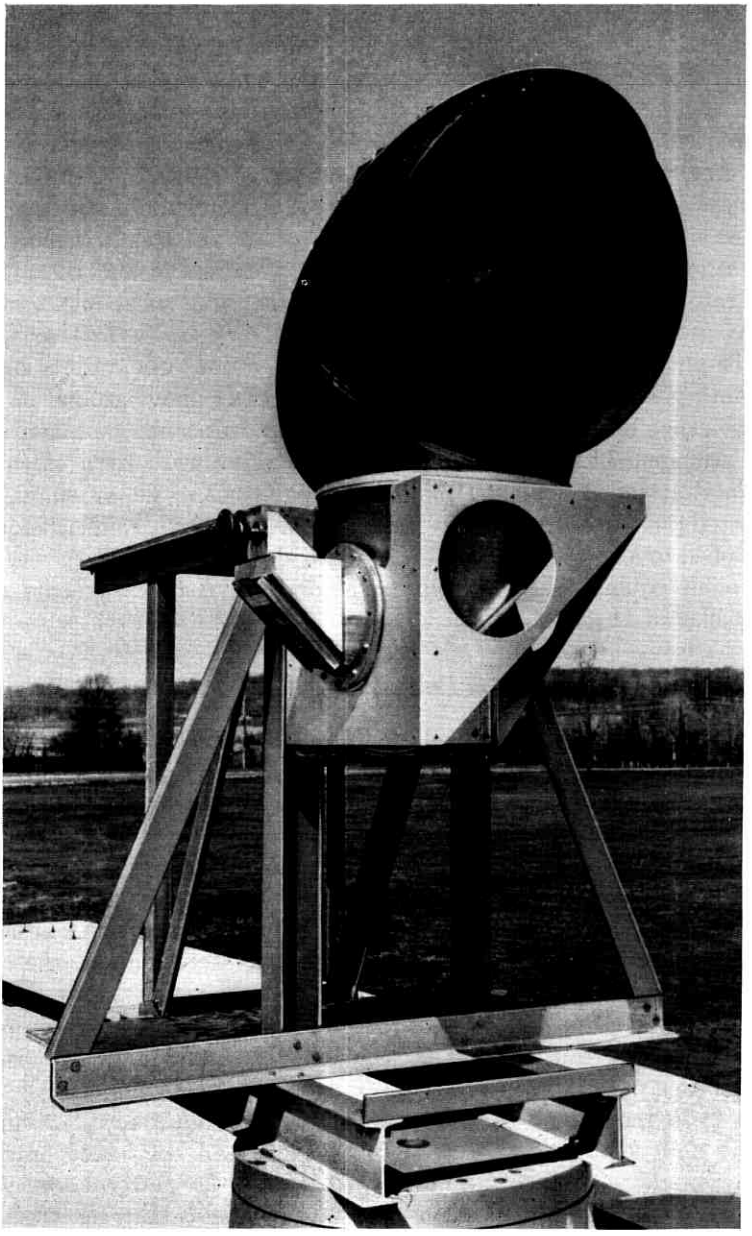

Fig. 5 — The antenna model mounted for radiation pattern measurements in the transverse plane and 0° elevation.

the isotropic level of the antenna. Since the gain of this antenna at 60 gc is about 51 decibels, an 81-decibel dynamic measuring range is required. This range is achieved through the use of a high-power short pulse technique. The short pulse technique allows both the elimination of spurious reflections from outside the direct signal path and a wide dynamic range. Otherwise, standard antenna measuring procedures are employed at the 1500-foot antenna range of the Holmdel Laboratory.

The signal source consists of a 10-kw peak power output magnetron operating at 60 gc. The pulse width is 0.2 microsecond and the repetition rate 1000 cps. A 30-decibel rectangular horn located about 12 inches above ground is used as a source antenna.

The antenna model can be mounted with various orientations on an azimuth positioner which is located about 30 feet above ground on top of the range building. Complete 360 degree patterns can be taken with this arrangement. A superheterodyne receiver followed by a video detector is attached to the output waveguide of the model antenna. The detected pulse is amplified in a narrow band 1000-cps amplifier before it is applied to a rectangular, logarithmic pattern recorder.

Similar equipment is also provided for pattern measurements at a frequency of 11 gc.

3.3 *The Measured Radiation Patterns*

All the pattern measurements for the TE_{11} mode were taken in the two principal planes (longitudinal and transverse) with linear field polarization in the same two planes. Cross polarized patterns were not taken. The planes are defined with respect to the straight horn-reflector antenna. The longitudinal plane contains the axis of the cone and the beam while the transverse plane contains the axis of the beam but is normal to the axis of the cone. The folded-horn antenna requires an additional designation to describe the rotation of the parabolic section about the cone axis. In practice this is the elevation angle of the beam with respect to the horizontal plane of Fig. 2. In the zenith or 90 degree elevation position, both the beam axis and the segments of the folded cone axis are in the same plane.

Over 30 radiation patterns were taken. These include all principal planes and polarizations of both the straight and the folded horn-reflector antennas at 60 gc and 11 gc. Only a few of the patterns are shown here. Fig. 6(a) shows a pattern for the straight horn-reflector antenna in the longitudinal plane with longitudinal polarization. The characteristic spillover lobe is clearly visible at about 70 degrees to the

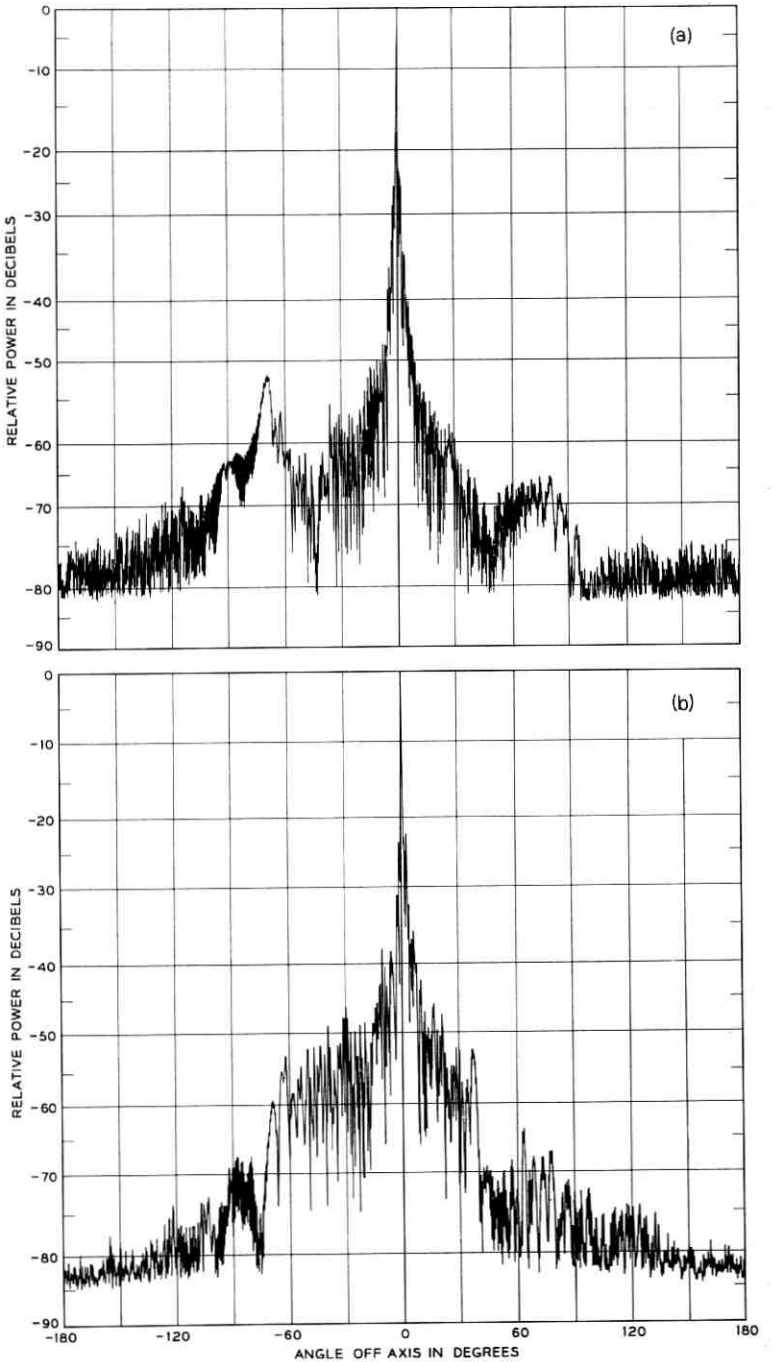

Fig. 6 — Measured radiation patterns for the TE_{11} mode in the longitudinal plane with longitudinal polarization at 60 gc: (a) straight horn (note the characteristic spillover lobe to the left of the main beam); (b) triply-folded horn in the zenith position.

left of the main beam, approaching the isotropic level of the antenna, 51 decibels. The same pattern was taken with the triply-folded horn antenna in the zenith position and is shown by Fig. 6(b). The spillover lobe has disappeared but other sidelobes closer to the main beam have become stronger. In Figs. 7(a) to 7(d), the patterns show a development of the radiation characteristics of the triply-folded horn from the straight horn. The single fold is the large area fold nearest the parabolic section. A strong spillover lobe which first becomes visible after the addition of the second fold appears at 62 degrees left of center and eventually reaches a level 6 decibels above isotropic. Of the many patterns taken, Fig. 7(d) shows the most pronounced effect of fold diffraction. In the horizon or 0 degree elevation position (not shown), the spillover lobe has practically disappeared. Fig. 8 shows an expanded section around the main beam of the pattern in Fig. 7(d). This expanded pattern has a first lobe level on the right side which is about 22 decibels below the peak. The one on the left side, however, has actually moved into the main beam. The computed pattern for the straight horn-reflector antenna gives the first sidelobe level as -24 decibels and a null depth of -26 decibels between the main beam and first sidelobe.[2]

For any satellite station application, the triply-folded horn must be equipped with an automatic tracking system similar to the one used at the Andover satellite station.[5] Since such a tracking system makes use of the TM_{01} mode, the TM_{01} radiation patterns of the triply-folded horn-reflector antenna model were measured. The results are presented in Fig. 9 together with the corresponding TE_{11} mode patterns. The peak levels of the two patterns were made equal, although the TM_{01} maxima were actually 5 decibels below the TE_{11} maxima. It is observed that the four TM_{01} mode patterns are quite symmetrical and that their nulls which are at least 28 decibels deep, align well with the peaks of the TE_{11} mode patterns.

In order to obtain a feeling for the amount of misalignment which could be tolerated in a triply-folded horn antenna, deliberate angular errors were introduced in the three axes segments of the folded conical horn. Pattern measurements indicate that errors in the amount of one beamwidth (0.5 degrees) in each segment do not cause serious degradation of the antenna characteristics. Aside from some expected beam-shifting, the main beam width of the TE_{11} mode patterns increases slightly while the TM_{01} mode patterns suffer from decreased null depth.

Pattern measurements at 11 gc, which are not included here, show increased diffraction effects due to the folds especially in the vicinity of the main beam. A general survey of all the available principal radiation

patterns indicates that a certain degradation of the antenna characteristics has taken place by introducing the three folds in the horn. This degradation can only be realistically judged by comparing the gain and noise temperature of the two types of antennas.

3.4 *Gain and Noise Temperature*

The accuracy of the standard horn comparison technique for gain determination was found to be inferior to the pattern integration method.

Both gain and antenna noise temperature were therefore computed by means of the following expressions which are derived in Appendix A:

$$G_0 = \frac{4\pi}{\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi} P_n(\theta,\phi)\ \sin\theta d\theta d\phi} \tag{1}$$

and

$$T_A = \frac{G_0}{4\pi}\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi} P_n(\theta,\phi)\,T(\theta,\phi)\ \sin\theta d\theta d\phi \tag{2}$$

where ϕ and θ are spherical coordinates with $\theta = 0$ coincident with the electrical axis of the antenna. Pattern measurements yield directly the normalized gain function $P_n(\theta,\phi) = P(\theta,\phi)/P(0,0)$. For temperature calculations, the temperature, $T(\theta,\phi)$, of the sphere surrounding the antenna has to be specified.

As explained in Appendix A, P_n consists of the sum of the principal and cross polarized radiation patterns. In the following computations, however, the cross polarized patterns were omitted, having not been measured. It is estimated that the inclusion of the cross polarized component would reduce the gain by about 0.1 decibel.

Equation (1) holds separately for the transverse and longitudinal linear polarizations used in the tests. For each polarization, the two patterns in the transverse and in the longitudinal planes are available for a total of four half patterns. Each half pattern is then assumed to encompass the 90 degree sector about the main beam axis and centered on the measured half pattern. This assumption permits the integration in ϕ to be carried out easily. The integration in θ is then performed using standard graphical techniques. Finally, the integrated patterns for transverse and longitudinal polarizations are averaged. Results of such computations are shown in Table I.

The theoretical gain values shown in Table I are calculated by the formula:

$$G = \eta(4\pi A/\lambda^2) \tag{3}$$

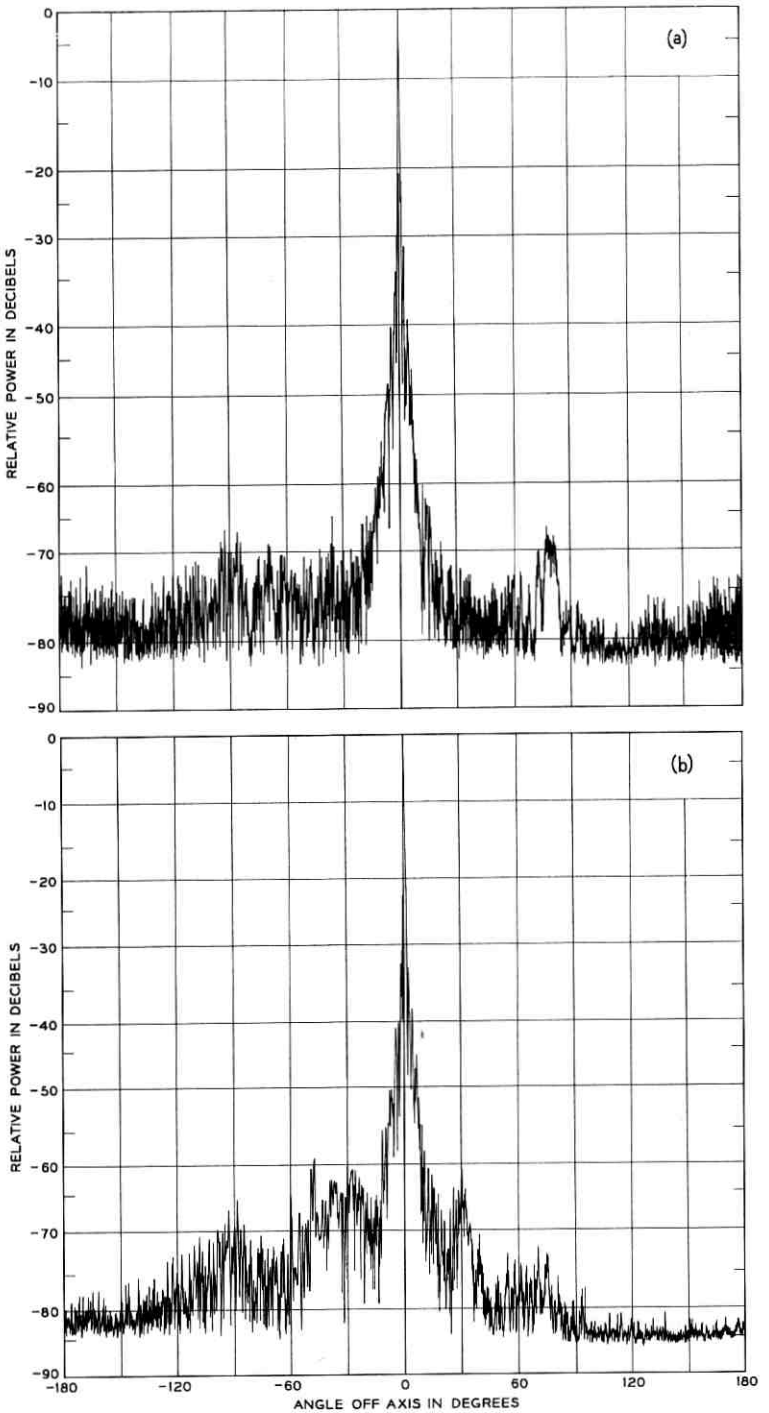

Fig. 7 — Measured radiation patterns for the TE_{11} mode in the longitudinal plane with transverse polarization at 60 gc showing the effect of progressive folding: (a) straight horn; (b) single large area fold; (c) double fold; (d) triple fold. The zenith position was used in all folded cases.

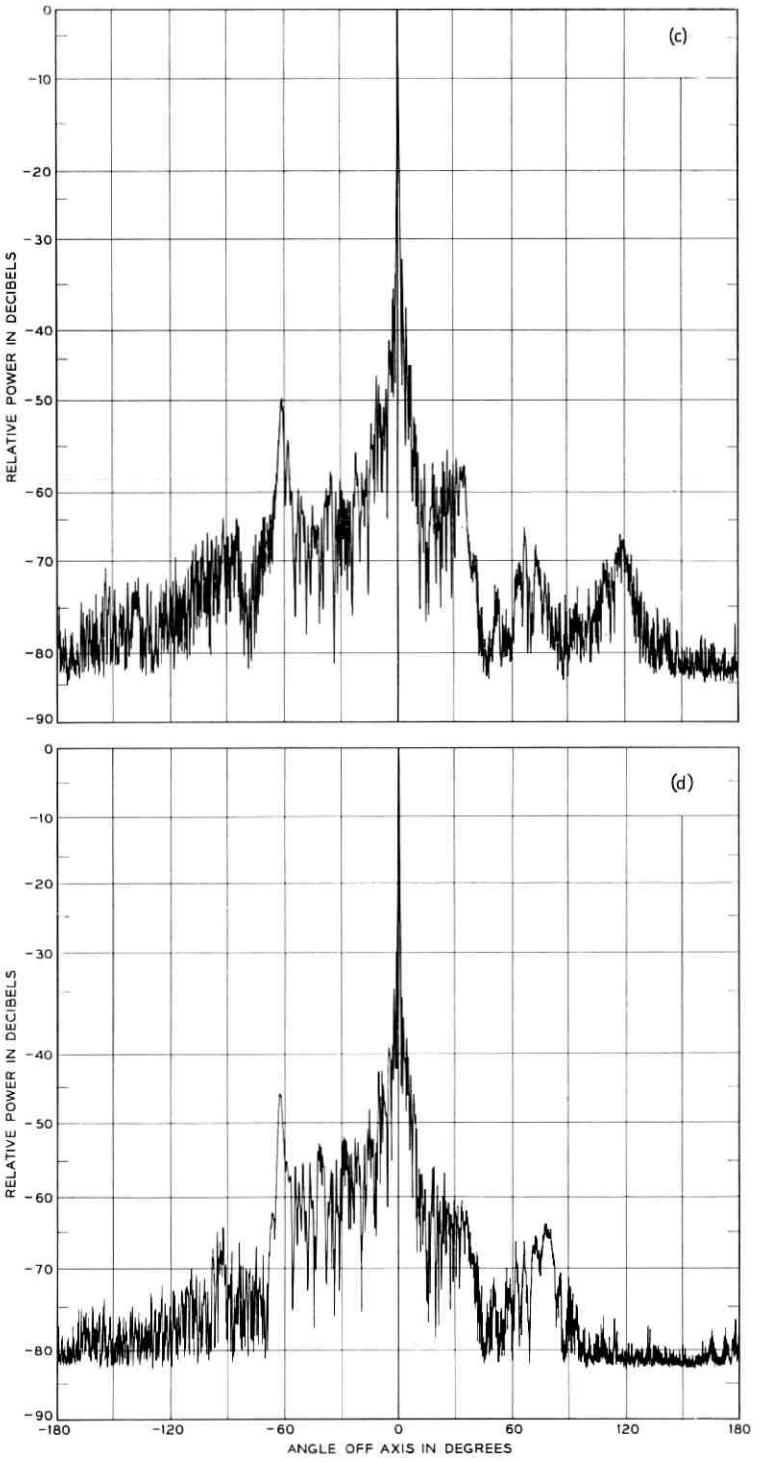
(c)
(d)
RELATIVE POWER IN DECIBELS
0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-180
-120
-60
0
60
120
180
ANGLE OFF AXIS IN DEGREES

FIG. 7 (cont.)

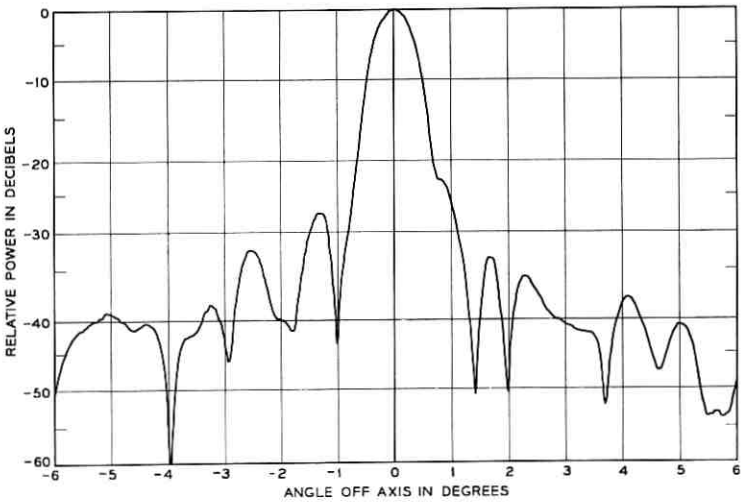

Fig. 8 — Expanded radiation pattern of Fig. 7(d) centered on the main beam and showing the first few side lobes.

where the efficiency η of the antenna aperture A is calculated to be 80.65 per cent for the conical horn-reflector antenna[2] if the effect of the characteristic spillover lobe is not taken into account. Spillover reduces the gain by approximately 0.2 decibel resulting in an efficiency of 77.1 per cent. Equation (3) and the gain values in Table I allow the calculation of antenna aperture efficiencies. For the triply-folded horn an efficiency of 60 $\pm$ 4 per cent at 60 gc is obtained practically independent of the elevation angle of the parabolic reflector. This is one decibel below the measured straight horn efficiency of 75.4 $\pm$ 5 per cent. It is expected that in an exact scaled model of the antenna of Fig. 2 the aperture efficiency would be higher and would come closer to the value of the straight horn antenna.

The antenna temperature, T_A, of (2) is computed for both the zenith position and an angle of 5 degrees above the horizon. For the zenith computation the temperature of the surrounding sphere can be assumed circularly symmetric about a normal to the earth surface. The assumed temperature distribution with elevation is:

$$T = \frac{2.2°K}{\cos \theta} \qquad 0 < \theta < 87.5° \tag{4}$$

and

$$T = 300°K \qquad 87.5° < \theta < 180°. \tag{5}$$

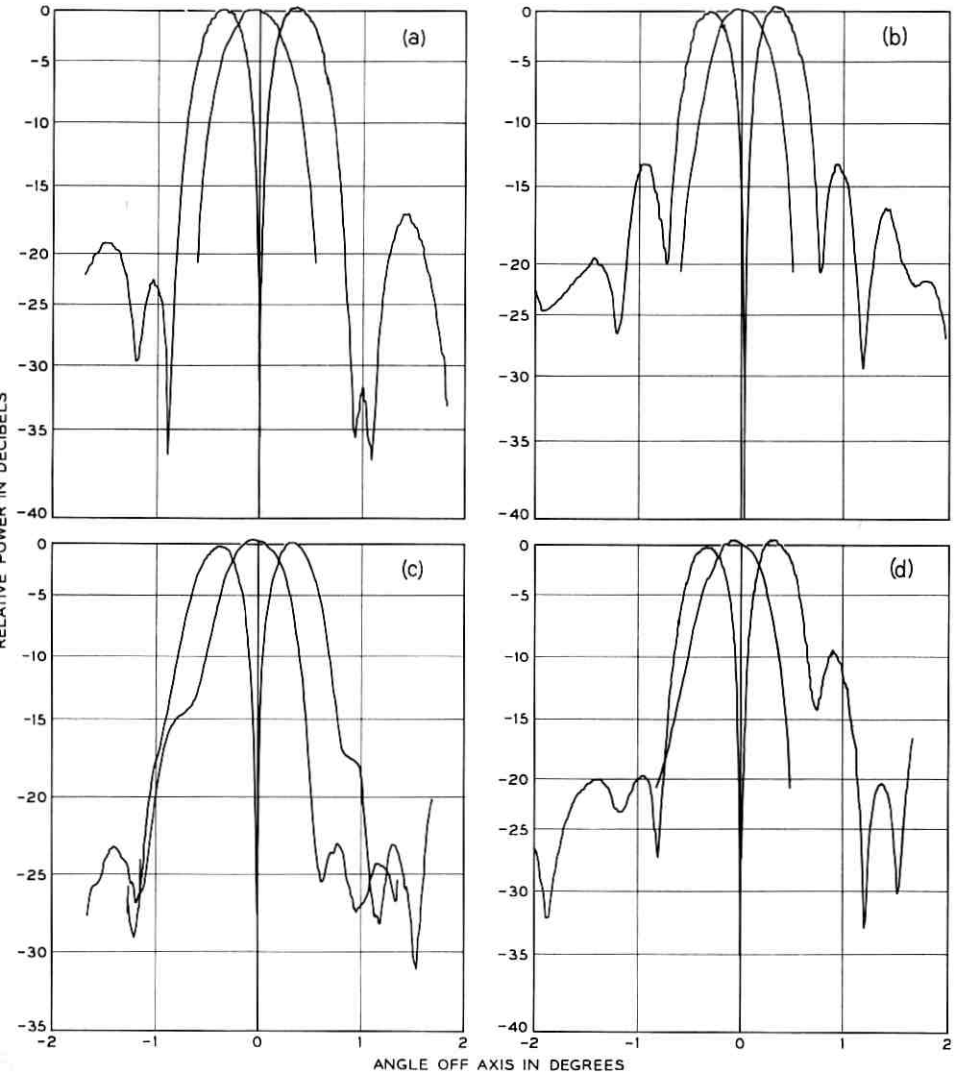

Fig. 9 — Measured radiation patterns of the triply-folded conical horn-reflector antenna model for the TM_{01} mode shown for comparison with the TE_{11} mode measured patterns. The pattern maximas have been normalized for display convenience. The measurement plane and polarization designations are: (a) transverse plane, transverse polarization and 0° position; (b) transverse plane, transverse polarization and 90° position; (c) longitudinal plane, longitudinal polarization and 0° position; and, (d) longitudinal plane, longitudinal polarization and 90° position.

TABLE I

GAINS OF THE ANTENNA MODEL DETERMINED BY INTEGRATION OF MEASURED PATTERNS

Antenna Elevation Degrees	f gc	Theoretical		Straight Horn		Triply-Folded Horn		Gain Difference Between Straight Horn and Triply-Folded Horn
		Gain db	Efficiency %	Gain db	Efficiency %	Gain db	Efficiency %	db
90	60.0	51.55	77.1	51.45	75.4	50.46	60.0	0.99
0	60.0	51.55	77.1	51.45	75.4	50.42	59.4	1.03
90	11.07	36.86	77.1	36.50	71.0	35.55	57.0	0.95
0	11.07	36.86	77.1	36.50	71.0	35.35	54.4	1.15
ONE SIGMA ERROR				±0.3	±5.0	±0.3	±4.0	±0.1

Equation (4) is a good expression for the atmospheric noise at a frequency of about 4 gc from zenith down to 2.5 degrees elevation. Below this elevation the beginning of the warm earth is assumed with a uniform temperature of 300 degrees Kelvin. This is a pessimistic assumption since the effective ground temperature will always be below 300 degrees K depending on the reflection coefficient of the earth's surface surrounding the antenna.

For an antenna elevation angle of 7.5 degrees (5 degrees above the physical horizon), the temperature distribution given by (4) and (5) was modified to simplify the integration of (2). This spherical temperature distribution can be described by two regions. The first is a spherical cap centered on the main beam described by $0 < \theta < 5°$ and consisting of atmospheric temperature only. This region is further divided into two semi-caps by a great circle arc tangential to the 7.5 degree elevation. The upper semi-cap has a temperature distribution:

$$T = \frac{2.2°K}{\cos\,(82.5° - \theta)} \quad \text{for} \quad 0 < \theta < 5° \qquad 0 < \phi < 180°. \tag{6}$$

The lower semi-cap extending below the 7.5 degree elevation line has a temperature distribution:

$$T = \frac{2.2°K}{\cos\,(82.5° + \theta)} \quad \text{for} \quad 0 < \theta < 5° \qquad 180° < \phi < 360°. \tag{7}$$

The second region may also be divided into two parts. The lower one represents the warm earth and is bounded by $5° < \theta < 180°$ and $180° < \phi < 360°$. The upper part extending toward the zenith completes the temperature sphere and is assumed to be at 0 degrees K. Results of temperature computations are shown in Table II in the form of excess antenna temperatures which are obtained by subtracting the atmospheric background temperature T_B given by (4) at the center of the main beam from T_A.

The zenith excess temperatures are very small in all cases considered, even for the triply-folded horn at 11 gc. At the lower elevation angle and 60 gc the triply-folded horn picks up about 3.9 degrees K more noise than the straight horn. At 11 gc the higher side lobe levels of the folded horn cause the noise to increase considerably at the lower elevation angle. Whereas the strong diffraction in the triply-folded horn at 11 gc did not manifest itself in a noticeable gain reduction, it shows up as a substantial increase in the noise temperature of the antenna at the lower elevation.

IV. HYDRODYNAMIC SCALE MODEL TESTS

The behavior of the antenna shown in Fig. 2 under conditions of heavy wind was determined by means of scale (1:96) model tests in water since wind-tunnel tests would have been less convenient. The measuring technique used at the hydrodynamic test facilities of the Davidson Laboratory of Stevens Institute of Technology, Hoboken, New Jersey is described elsewhere.[6] The most significant quantity measured was the azimuth wind torque coefficient C_w of the antenna. This quantity is defined by

$$T(t) = C_w V^2(t) \tag{8}$$

TABLE II

NOISE TEMPERATURES OF THE ANTENNA MODEL DETERMINED BY INTEGRATION OF MEASURED PATTERNS

Antenna Elevation Degrees	Antenna Angle Above Physical Horizon Degrees	f gc	Background Temp. °K	Straight Horn Excess Temp. °K	Triply-Folded Horn Excess Temp. °K	Excess Temp. Difference between Straight Horn and Triply-Folded Horn °K
90	87.5	60.0	2.20	0.46	0.46	0
7.5	5	60.0	16.86	4.20	8.08	3.88
90	87.5	11.07	2.20	0.45	0.79	0.34
7.5	5	11.07	16.86	13.64	89.49	75.85

where T = wind torque about the azimuth axis and V = wind velocity. C_w is a function of the wind direction and the antenna elevation. The highest value of C_w is obtained when the wind blows directly into the antenna aperture at 0 degrees elevation. This value is listed in Table III, for the triply-folded horn with and without aperture cover and for a 2250 ft.^{2} Andover-type horn-reflector antenna for comparison.

Wind induced torque about the elevation axis is small and presents no problem in tracking. The antenna overturning stability is described in terms of an overturning moment coefficient C_{wo} defined for an axis located in the base of the structure. The maximum value of C_{wo} for the most unfavorable orientation of the antenna is also given in Table III. Approximately 18×10^6 ft.-lbs. of torque is required to overturn the antenna. This estimate is based upon an estimated weight of 475,000 lbs and gives a safety factor of approximately 6 for 100 mph winds.

Calculations made by K. N. Coyne[6] give the wind speed at which the antenna stops tracking (worst case). This wind speed at stall is shown in Table III. The data is based on dual 25-hp hydraulic azimuth drives, which result in a maximum azimuth drive torque of 1.2×10^6 ft.-lbs. The table finally gives the wind velocity at which an RMS tracking error of 0.01° is reached. The assumption is made that a servo-system in the autotrack mode similar to the one used in the Andover satellite station is used.[7] The numbers indicate that negligible tracking errors are produced by winds up to 60 mph for the triply-folded horn and that the antenna can be stalled only by winds of about 80 mph or more. Comparing these numbers with those for an Andover-type antenna, one finds a very substantial improvement.

TABLE III

WIND LOAD CHARACTERISTICS OF ANTENNAS WITH 2250 FT.^{2} APERTURE

	Triply-Folded Horn		Andover Type Horn Reflector Without Radome**
	With Aperture Cover	Without Aperture Cover	
Maximum C_w in ft.-lbs./mph^2	132	184	500
Maximum C_{wo} in ft.-lbs./mph^2	156	309	—
Avg. Wind speed* for stall (at max. C_w) in mph	95	80	49
Avg. Wind speed* for 0.01° RMS autotrack error (at max. C_w) in mph	66	56	36.5

* Standard deviation of variable wind component assumed to be 30% of average velocity.

** Numbers based on wind tunnel tests.

V. SUMMARY AND CONCLUSIONS

The triply-folded conical horn-reflector antenna has been shown to be a compact antenna suitable for use in large ground stations for satellite communications. Measurements on an electrical model of the antenna indicate that the aperture efficiency of a full sized antenna should be higher than 60 per cent, which exceeds that of many existing low noise parabolic or cassegrain antennas and approaches that of the straight horn-reflector antenna. The zenith excess noise temperature of the folded horn due to side and back lobes is below 1 degree K, and, at an elevation angle of 7.5 degrees, the temperature does not exceed the very low value of the straight horn-reflector antenna by more than a few degrees. The noise temperature of a ground station containing a triply-folded horn will therefore be determined almost exclusively by circuit and atmospheric noise.

No degradation in the characteristics of the antenna is expected if the reflecting plates are aligned in angle to better than the half power width of the antenna beam. The required overall surface tolerance has to be distributed among the four reflecting surfaces. It is expected that the three flat reflectors can be easily built with high accuracy so that most of the available tolerance can be assigned to the parabolic reflector. The surface accuracy of the latter can be checked easily at any desired elevation angle by an optical triangulation method.

The compactness of the antenna and the rounded silhouette achieved by attaching lightweight fairings allows operation without a radome. Hydrodynamic model tests show that an antenna with a 2250 ft.^{2} aperture should be capable of tracking accurately up to wind velocities of more than 60 mph.

It can be said that the triply-folded horn-reflector antenna has achieved the desirable features of low-noise, high-gain, broadband operation and interference immunity because the signal path from the stationary equipment room to the parabolic reflector is entirely enclosed by a metallic sheath of considerable diameter and that this path is not folded back on itself. Some of the mentioned electrical characteristics can only be obtained in other antenna configurations by careful subreflector or beam (higher order modes) shaping. Broadband operation of such antennas is not achieved easily if at all. Construction of the triply-folded horn should also be simplified because it has only one curved reflector surface compared with at least two for cassegrain types.

VI. ACKNOWLEDGMENTS

We would like to extend our appreciation to Messrs. R. D. Peterson and T. B. Henry for measuring the radiation patterns and performing

the numerical integrations. Also to A. O. Schwarz and K. L. Warthman who modified the existing antenna model and built the folded sections. A. O. Schwarz's contributions in the mechanical area and K. N. Coyne's results of the hydrodynamic model testing have been used in writing this paper. The constant interest and support given by Messrs. I. Welber and H. P. Kelly is gratefully acknowledged.

APPENDIX A

Derivation of the Gain and Temperature Formulas

A.1 *The Antenna Gain*

According to Silver[8] the gain function of an antenna is defined by

$$G(\theta,\phi) = \frac{P(\theta,\phi)}{\frac{1}{4\pi} P_t} \tag{9}$$

where $P(\theta,\phi)$ = the power radiated per unit solid angle in direction θ,ϕ.
P_t = total power delivered to the (lossless) antenna in a single mode.

P_t obviously is the average of $P(\theta,\phi)$ integrated over the whole sphere.

$$P_t = \int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} P(\theta,\phi) \sin\theta \, d\theta d\phi. \tag{10}$$

The denominator of (9), $P_t/4\pi$, can be interpreted as being the power radiated per unit solid angle from an isotropic antenna which is fed the total power P_t. It is convenient to write $P(\theta,\phi)$ as the sum of two orthogonal (meaning completely decoupled) components:

$$P(\theta,\phi) = P_p(\theta,\phi) + P_c(\theta,\phi), \tag{11}$$

P_p being the principal and P_c the cross-polarized component. P_p and P_c could for instance, be horizontally and vertically or right- and left-hand circularly polarized components to mention only a few out of an infinite number of possibilities.

The definition given by (9) obviously assumes that all the power $P(\theta,\phi)$ can be usefully extracted by a distant receiving antenna. This means that it is capable of receiving both components P_p and P_c, or that its polarization is matched exactly to the incoming wave. In many practical applications this requirement is easily met, e.g., a horn-reflector antenna fed by a single TE_{11} mode radiates pure linear polariza-

tion on the beam axis (but not in all other directions) and a properly aligned linearly polarized receiving antenna can extract all the possible energy. If such a match is not provided, a polarization loss occurs by which the gain given by (9) will have to be reduced.

Of particular interest is the maximum value of the gain function. We assume it occurs at $\theta = 0$, $\phi = 0$ and is normally called the "gain" G_0 of the antenna. Inserting (10) in (9) we obtain

$$G_0 = \frac{4\pi P(0,0)}{\int\int P(\theta,\phi) \sin \theta d\theta d\phi}. \tag{12}$$

If we introduce the "normalized" radiation function

$$P_n(\theta,\phi) = \frac{P(\theta,\phi)}{P(0,0)} \tag{13}$$

we can write (12):

$$G_0 = \frac{4\pi}{\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi} P_n(\theta,\phi) \sin \theta d\theta d\phi}. \tag{14}$$

In general, $P_n(\theta,\phi)$ cannot be directly measured. It is much easier to determine the orthogonal components P_{np} and P_{nc} since they are the direct result of pattern measurements. We have

$$P_n(\theta,\phi) = P_{np}(\theta,\phi) + P_{nc}(\theta,\phi) \tag{15}$$

where

$$P_{np}(\theta,\phi) = \frac{P_p(\theta,\phi)}{P_p(0,0) + P_c(0,0)}$$

$$P_{nc}(\theta,\phi) = \frac{P_c(\theta,\phi)}{P_p(0,0) + P_c(0,0)}.$$

And finally, it would be easy to derive the following useful form for (9):

$$G(\theta,\phi) = G_0 P_n(\theta,\phi). \tag{16}$$

A.2 *The Antenna Temperature*

We assume that the antenna is fed in such a way as to produce the orthogonal far field intensities $P_p(\theta,\phi)$ and $P_c(\theta,\phi)$. The noise power radiated by the surrounding sphere in the matching orthogonal polarizations is proportional to the temperatures $T_p(\theta,\phi)$ and $T_c(\theta,\phi)$ existing

on the sphere. The noise power picked up by the antenna as a receiver is $W = kT_AB$, where

$$\begin{aligned} k &= \text{Boltzmann's constant} \\ B &= \text{noise bandwidth} \\ T_A &= \text{effective antenna temperature} \\ &= c\int_{\theta=0}^{\pi}\int_{\phi=0}^{2\pi} T_p(\theta,\phi)P_p(\theta,\phi)\sin\theta d\theta d\phi \\ &\quad + c\int_{\theta=0}^{\pi}\int_{\phi=\theta}^{2\pi} T_c(\theta,\phi)P_c(\theta,\phi)\sin\theta d\theta d\phi. \end{aligned} \tag{17}$$

Equation (17) immediately follows from the fact that the noise received by the antenna must be weighted by $P(\theta,\phi)$. The constant c can be easily determined if we assume the special case of a black-body enclosure, randomly polarized, at the constant temperature T_0, i.e.,

$$T_p = T_c = T_0$$

then we obtain from (17):

$$T_A = cT_0\int\int [P_p(\theta,\phi) + P_c(\theta,\phi)]\sin\theta d\theta d\phi. \tag{18}$$

Let us assume now that the antenna is perfectly matched into a termination and that it is perfectly insulated from any thermal heat sources. Then under thermal equilibrium the termination will accept the temperature of the black-body enclosure, T_0, and the noise powers coming from and flowing towards the antenna are both the same and equal to $W = kT_0B$. In practice it is impossible to find such an equilibrium between the antenna termination and the radiating enclosure. The termination can be warmer or colder depending on the input temperature of the receiver connected to the antenna. The noise powers flowing to and from the antenna are different in this case. We are only interested in the flow of noise energy coming from the antenna which for all practical purposes is still kT_0B. This means $T_A = T_0$ in (18) and we immediately find:

$$c^{-1} = \int\int [P_p(\theta,\phi) + P_c(\theta,\phi)]\sin\theta d\theta d\phi. \tag{19}$$

From (11) and (12), we find that expression (18) is also identical to

$$c^{-1} = \frac{4\pi P(0,0)}{G_0}.$$

Equation (17) now becomes:

$$T_A = \frac{G_0}{4\pi} \int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} [T_p(\theta,\phi) P_{np}(\theta,\phi) + T_c(\theta,\phi) P_{nc}(\theta,\phi)] \sin\theta d\theta d\phi. \quad (20)$$

This expression is useful in radio astronomy where radio sources are not always randomly polarized.

For the determination of the antenna noise in a satellite communications system, it is sufficient to assume randomly polarized surroundings. This is a simplification of the physical reality which will lead to a high (pessimistic) estimate of T_A. For random noise polarization, $T_p = T_c = T$, we obtain for (20) by using (15):

$$T_A = \frac{G_0}{4\pi} \int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} T(\theta,\phi) P_n(\theta,\phi) \sin\theta d\theta d\phi. \quad (21)$$

And with (16) we obtain the often used form:

$$T_A = \frac{1}{4\pi} \int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} T(\theta,\phi) G(\theta,\phi) \sin\theta d\theta d\phi. \quad (22)$$

REFERENCES

1. Crawford, A. B., Hogg, D. C., and Hunt, L. E., A Horn-Reflector Antenna for Space Communication, B.S.T.J., *40*, July 1961, pp. 1095–1116.
2. Hines, J. N., Li, Tingye, and Turrin, R. H., The Electrical Characteristics of the Conical Horn-Reflector Antenna, B.S.T.J., *42*, July, 1963, pp. 1185–1211.
3. Dolling, J. C., Blackmore, R. W., Kindermann, W. J., and Woodard, K. B., The Mechanical Design of the Conical Horn-Reflector Antenna and Radome, B.S.T.J., *42*, July 1963, pp. 1137–1186.
4. Annals of the New York Academy of Sciences, *Large Steerable Radio Antennas — Climatological and Aerodynamic Considerations, 116,* June 26, 1964, discussion by J. Ruze.
5. Cook, J. S., and Lowell, R., The Autotrack System, B.S.T.J., *42*, July 1963, pp. 1283–1307.
6. Coyne, K. N., Hydrodynamic Techniques for Study of Wind Effects on Antenna Structures, B.S.T.J. this issue, pp. 1339–1365.
7. Lozier, J. C., Norton, J. A., and Iwama, M., The Servo System for Antenna Positioning, B.S.T.J., *42*, July 1963, pp. 1253–1281.
8. Silver, S., *Microwave Antenna Theory and Design,* Rad. Lab Series, *12*.

The Open Cassegrain Antenna: Part I. Electromagnetic Design and Analysis

By J. S. COOK, E. M. ELAM and H. ZUCKER

(Manuscript received May 14, 1965)

The open cassegrain antenna combines an asymmetric cassegrain reflector system with antenna rotation about two non-orthogonal axes. The compact configuration provides well-controlled radiation with full hemispheric coverage.

A comprehensive analysis of the antenna geometrical and radiation characteristics has been made, and an experimental antenna with 40-inch aperture, operating at 60 gcs, has been constructed and measured electrically. Agreement was obtained between the computed and measured characteristics of the antenna and its components. By computation, it is found that the aperture efficiency of the experimental antenna is 70.4 per cent, the antenna efficiency (neglecting ohmic loss) is 65 per cent, and based on measured subreflector radiation patterns, the noise temperature due to spillover at the main reflector is less than 4°K.

I. INTRODUCTION

The open cassegrain antenna configuration is shown in Fig. 1. Its optical geometry is straight forward, consisting of hyperboloid and paraboloid surfaces, but it has the distinguishing characteristic that the axes of rotational symmetry of the sub- and main-reflector surfaces do not coincide. Fig. 2 shows a view of the antenna looking down the beam axis. The projected aperture is circular and no aperture blocking is introduced by the subreflector or its support structure (not shown in Fig. 2), hence the name open cassegrain.

Non-orthogonal beam steering axes are used. The antenna is directed at zenith in Fig. 1(a) and at its minimum elevation of $-5°$ in Fig. 1(b). The lowest elevation excursion is determined by the angle of the slant axis which is coincident with the "secondary optical" axis and is in-

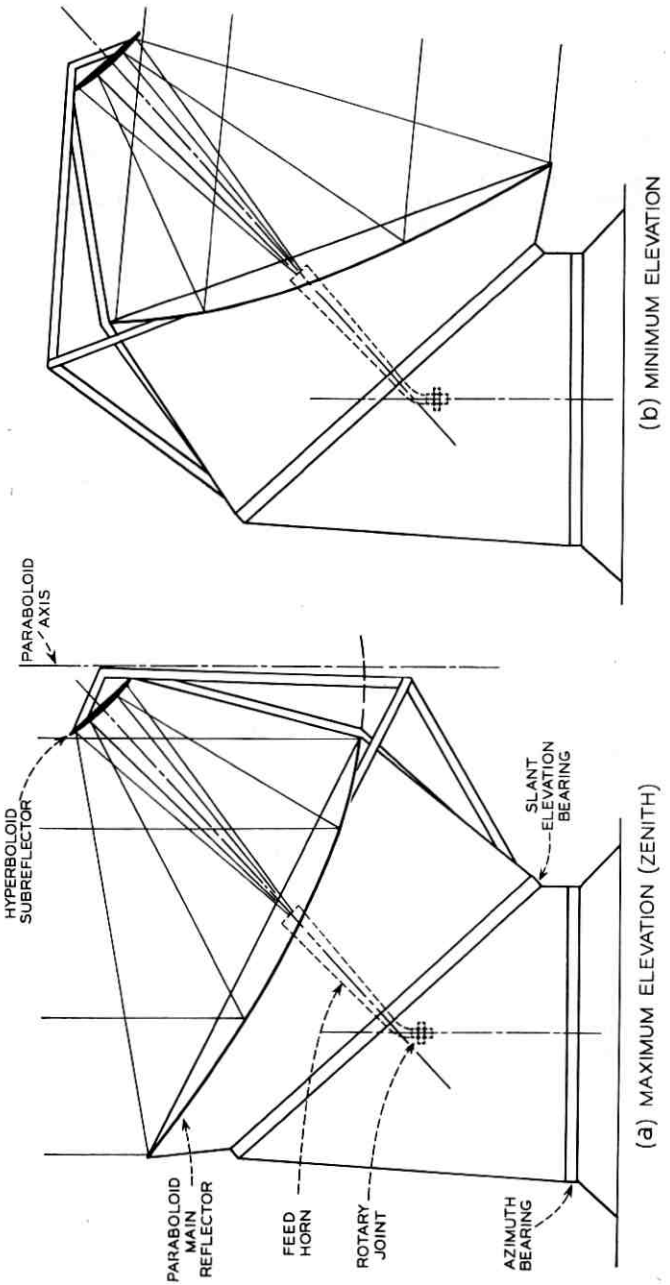

Fig. 1 — Open cassegrain antenna.

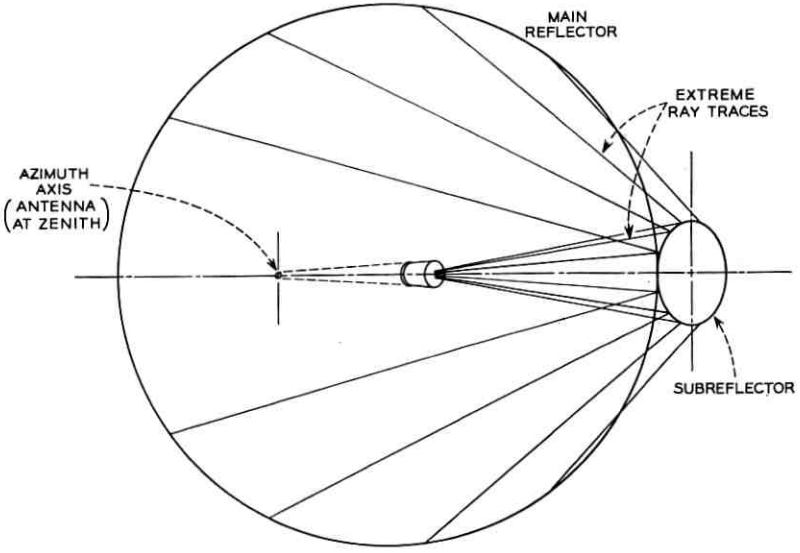

Fig. 2 — Antenna view from the direction of main beam.

clined 42.5° from the horizontal. The conical feed horn is attached to the azimuth unit and remains fixed when the antenna rotates about its slant axis. This unusual configuration is motivated by the desire to bring the antenna feed to a convenient equipment location as directly as possible, and to keep the structure small and easily enclosed so the antenna can be operated reliably without the protection of a radome.

Both electrical and structural characteristics of the open cassegrain have been studied. The particular antenna design that was used as a model for the studies was derived by balancing electrical against structural considerations to arrive (somewhat arbitrarily) at a configuration that would meet the general requirements of a wide band 4-gcs satellite communications link. A close interdependence exists between these two aspects of antenna design presented in this issue as Part I and Part II of the open cassegrain antenna study.

The following section of Part I considers the antenna system and the background against which antenna characteristics must be chosen. Section III presents the geometrical interrelationships of the reflector surfaces which set the ground rules for electrical and structural tradeoffs. The coordinates and geometrical expressions necessary for the electromagnetic analysis that follows are also presented in Section III. The analysis of the radiation characteristics is presented in Section IV,

followed in Section V by corresponding measurements of a 60-gc antenna. The concluding section contains a brief summary of antenna performance.

Dual-mode excitation of the circular-cone feed horn (TE_{11} and TM_{11} modes),[1] while not necessary for operation of the open cassegrain, brings about enough improvement in its characteristics over simple TE_{11} mode excitation that combined excitation has been assumed throughout the electromagnetic analysis. An improvement of roughly 1 db in signal-to-noise ratio is typical for a sensitive (maser) receiving system. Mode conversion techniques are investigated in a separate paper.[2]

II. ANTENNA SYSTEM CONSIDERATIONS

2.1 *Antenna Noise*

The essential low-noise advantage of the open cassegrain comes as a result of operating the antenna without a radome; but the *rf* configuration and the slant mount combine to make the antenna itself an inherently low-noise device.

The excess noise received by the more common symmetrical cassegrain (where the subreflector is centered on the paraboloid axis) comes mostly from three sources: (1.) scattering from the subreflector support (and subreflector blocking if it's excessively large), (2.) radiation past the edge of the subreflector, and (3.) radiation past the edge of the main reflector. The amount of extraneous radiation from these sources can be selectively minimized by using near-field subreflector illumination,[3] special subreflector supports, reflector skirts, etc., but it is difficult to avoid receiving a certain amount of extra thermal noise, particularly as the antenna swings down toward the horizon.

Since the open cassegrain subreflector is located outside the main beam, the extraneous scattering is minimized. Subreflector spillover is confined to an elevation of about 30° to 55° above the horizon regardless of the antenna position.

This leaves only the main reflector spillover as a major source of excess antenna noise, and it is essentially independent of elevation angle.

2.2 *Signal-to-Noise Optimization*

The quality of a low-noise antenna must be evaluated in terms of the system it serves. The receiver system quality is measured by its signal-to-noise ratio, which is proportional to the antenna gain and inversely proportional to the total system noise.

In clear weather, sky noise varies from 3°K at zenith to about 20°K at 7.5° elevation angle.[4] A maser receiver and its plumbing is likely to add 6 or 8°K to that. Thus, for a high-quality receiving system, the contribution to system noise exclusive of extraneous pickup from the antenna is about 10°K at zenith and increases to 27°K at the nominally minimum useful elevation. It is against this background that one must choose the antenna characteristics. One would not be willing to pay a great deal, for example, to improve the excess antenna noise from 5°K to 3°K since that would improve the signal-to-noise ratio only 0.6 db at zenith where it is needed least, and 0.3 db or less at the extremes of the satellite pass. On the other hand, one would like to have even that much improvement if it is easily achieved.

The open cassegrain lends itself to the optimization of the trade-off between aperture efficiency and excess noise. For a given projected aperture, the parameters to be optimized are: (1.) paraboloid focal length, (2.) subreflector diameter and surface shape (some improvement may be realized by slight modification of the hyperboloid), (3.) feed-horn aperture and taper, and (4.) TE_{11} to TM_{11} mode conversion.

Converting a certain portion of dominant mode in the horn to the higher-order TM_{11} mode broadens the subreflector illumination and controls the spillover at the same time.[1]

III. ANTENNA GEOMETRY

3.1 *Introduction*

The open cassegrain feed horn propagates both TE_{11} and TM_{11} modes. By suitable choice of mode amplitude ratio and relative phase a nearly circular radiation pattern can be obtained. The horn then radiates a nearly spherical wave front originating from a point designated as the phase center, which determines the location for one of the focal points of the subreflector hyperboloid. Based on geometrical optics the hyperboloid surface reflects the incident spherical wave as a wave effectively originating from the other focal point of the hyperboloid. In the open cassegrain, the latter focus coincides with that of the paraboloid surface of which the main reflector is an offset elliptic section. The main reflector converts the reflected spherical wave to a plane wave in the antenna aperture.

Several related coordinate systems were used for the antenna analysis, and a number of useful relationships were derived. Two sets of rectilinear coordinates are basic: x_p, y and z_p with origin at the paraboloid

focus and z_p lying along the axis of revolution as shown in Fig. 3, and x_s, y, z_s with the same origin but rotated so that z_s coincides with the secondary optics axis as shown in Fig. 4. These will be called the primary and secondary coordinate systems, respectively.

A spherical system is classically related to each of these rectilinear systems with poles on the z axis, θ measured from the $+z$ axis and φ measured counterclockwise from the x-z plane. y and r are common and carry no subscript. Certain convenient auxiliary coordinates are also used.

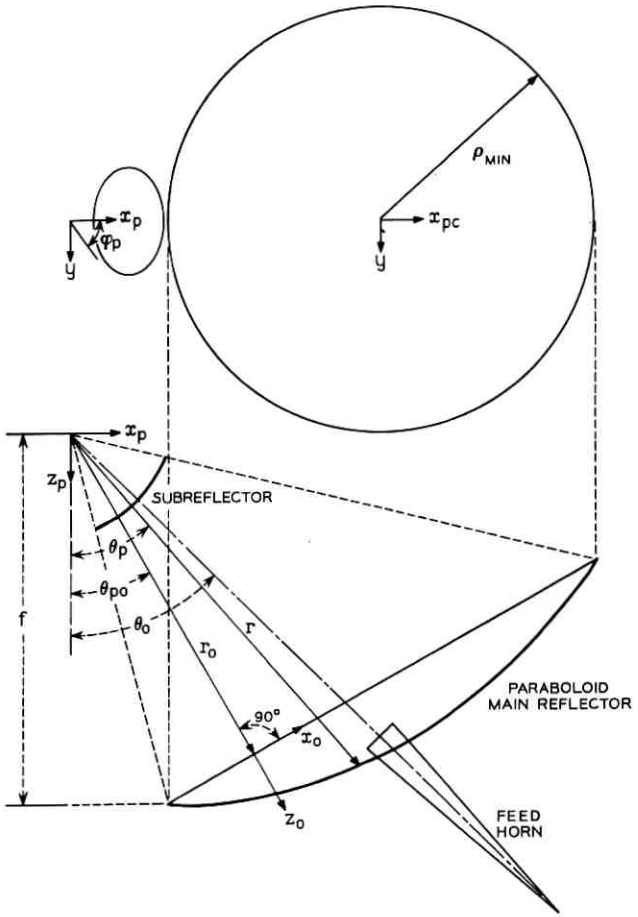

Fig. 3 — Antenna primary coordinates.

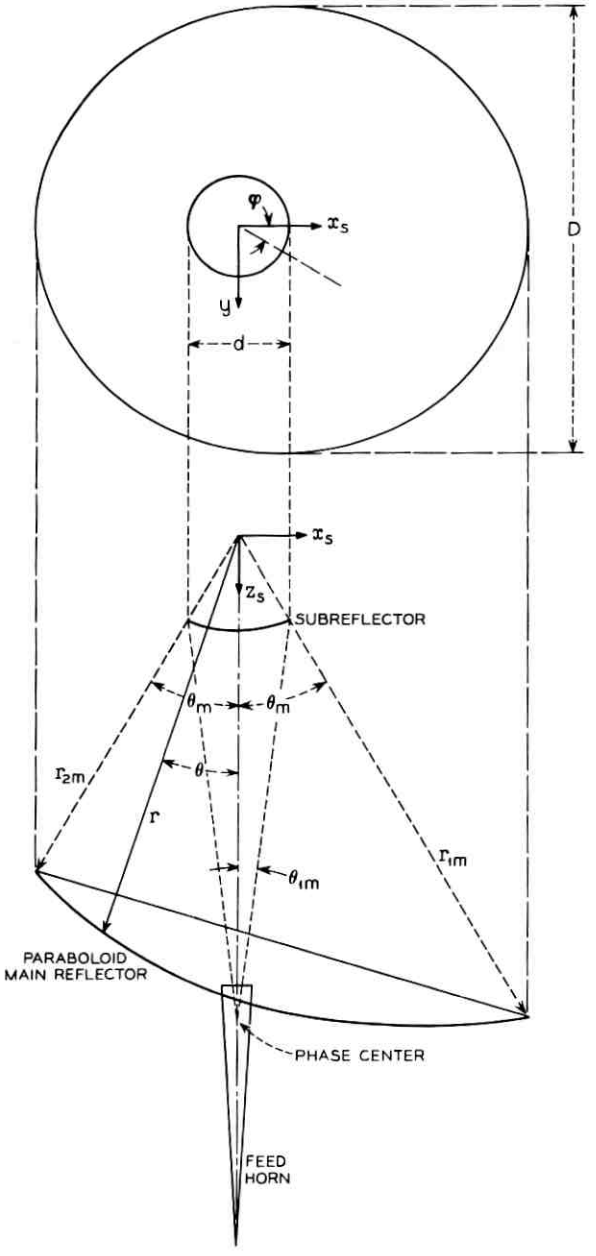

Fig. 4 — Antenna secondary coordinates.

3.2 *Paraboloid Reflector*

The paraboloid reflector is a section of a paraboloid of revolution about the z_p axis of focal length f. The equation for the paraboloid in primary spherical coordinates (aligned with the x_p, y, z_p coordinates) is:

$$r = \frac{2f}{1 + \cos \theta_p}. \tag{1}$$

In secondary spherical coordinates (aligned with the x_s, y, z_s coordinate system as shown in Fig. 4), the equation for the paraboloid surface is:

$$r = \frac{2f}{1 + \cos \theta \cos \theta_0 - \sin \theta \sin \theta_0 \cos \varphi}. \tag{2}$$

(The subscripts have been dropped from θ and φ secondary spherical coordinates for convenience.)

The curves of intersection of the paraboloid surface with cones, $\theta = \theta_c$, (constant) are ellipses and lie in planes perpendicular to the x_p, z_p plane. (Equations for the ellipses are presented in Appendix A.) The projections of these intersections onto the x_p, y plane are circles given by:

$$\left(x_p - \frac{2f \sin \theta_0}{\cos \theta_c + \cos \theta_0}\right)^2 + y^2 = 4f^2 \left(\frac{\sin \theta_c}{\cos \theta_c + \cos \theta_0}\right)^2. \tag{3}$$

The projections of the intersections of the planes $\varphi = \varphi_c$ with the paraboloid are also circular arcs given by:

$$(x + 2f \cot \theta_0)^2 + \left(y - \frac{2f \cot \varphi_c}{\sin \theta_0}\right)^2 = \frac{4f^2}{\sin^2 \theta_0 \sin^2 \varphi_c}. \tag{4}$$

The two sets of circles (3) and (4) are shown in Fig. 5 for $\theta_0 = 47.5°$. The sets of circles (3) and (4) are orthogonal; the projections of the intersection are therefore conformal. The special case, $\theta_0 = 90°$, corresponds to the projections of the horn-reflector antenna previously obtained by T. Li.[5]

3.3 *Hyperboloid Subreflector*

Referring to Fig. 6, the equation for the hyperboloid surface is given by the relationship:

$$|r_1| - |r_2| = b \tag{5}$$

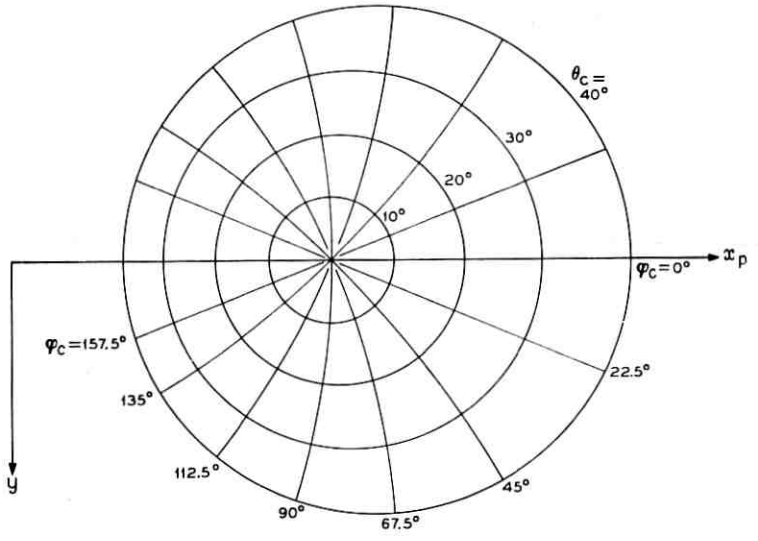

Fig. 5 — Projection circles of the paraboloid reflector.

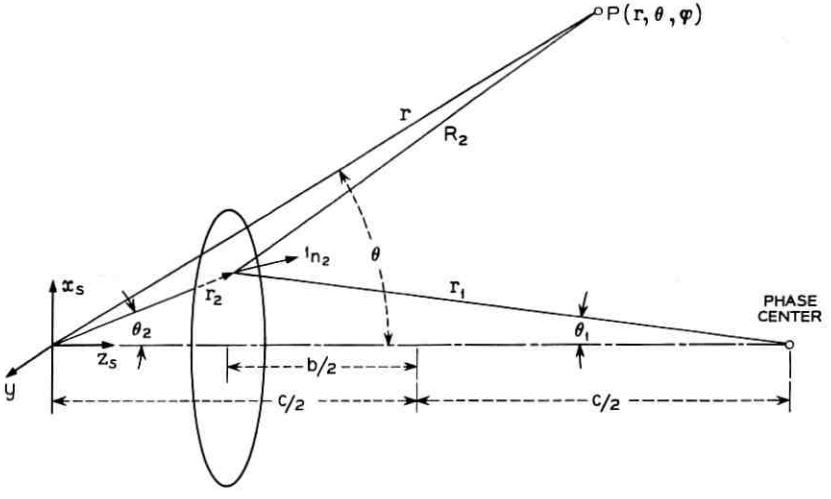

Fig. 6 — Subreflector coordinates.

where r_1 and r_2 are the distances from the foci and b is an arbitrary constant.

From (5), the equation for the hyperboloid surface in terms of θ_1 is:

$$r_1 = \frac{c}{2} \cdot \frac{(1 - \beta^2)}{\cos \theta_1 - \beta} \tag{6}$$

where

$$c = \text{distance between the foci of the hyperboloid}$$

and

$$\beta = b/c. \tag{7}$$

Similarly, in terms of θ_2,

$$r_2 = \frac{c}{2} \cdot \frac{(1 - \beta^2)}{\beta + \cos \theta_2}. \tag{8}$$

The relations between the angles θ_1 and θ_2 are

$$\cos \theta_1 = \frac{(1 + \beta^2) \cos \theta_2 + 2\beta}{1 + \beta^2 + 2\beta \cos \theta_2} \tag{9}$$

and

$$\cos \theta_2 = \frac{(1 + \beta^2) \cos \theta_1 - 2\beta}{1 + \beta^2 - 2\beta \cos \theta_1}. \tag{10}$$

The representation of the hyperboloid surface in the two coordinate systems is useful for the computations of the radiation patterns from the horn and subreflector.

3.4 *Relations Between the Antenna Parameters*

From the antenna geometry, and the restriction that the subreflector must not block the aperture, certain relationships between the antenna parameters can be established.

The relationship between the diameter D of the projected circle of the paraboloid surface in the x_p, y plane, the geometrical illumination angle θ_m, and the focal length of the paraboloid f is:

$$\sin \theta_m = \frac{\tau[\cos \theta_0 + \sqrt{1 + \tau_2 \sin^2 \theta_0}]}{1 + \tau^2} \tag{11}$$

with

$$\tau = D/4f. \tag{12}$$

The relationship between the diameter d, of the hyperboloid, and the diameter of the projected circle is:

$$\frac{d}{D} = \frac{\sin(\theta_0 - \theta_m)\sin\theta_m}{\tau[1 + \cos(\theta_0 - \theta_m)]\sin(\theta_0 + \theta_m)}. \tag{13}$$

The angle θ_{0m} subtended by the hyperboloid reflector at the intersection of the feed horn axis and paraboloid is:

$$\tan\theta_{0m} = \frac{\tau(1 + \cos\theta_0)\tan\theta_m}{\frac{D}{d}\tan\theta_m - \tau(1 + \cos\theta_0)}. \tag{14}$$

The ratio of minimum to maximum distance from the focal point of the paraboloid to the elliptical main reflector surface is:

$$\frac{r_{2m}}{r_{1m}} = \frac{1 + \cos(\theta_0 + \theta_m)}{1 + \cos(\theta_0 - \theta_m)}. \tag{15}$$

The relations (11), (13)–(15) are shown in Fig. 7 as a function of $D/4f$ for $\theta_0 = 47.5°$.

3.5 *The Antenna Dimensions*

The antenna gain is closely related to the diameter D, of the projected circle of the paraboloid reflector. For a specified antenna gain and assuming a reasonable aperture efficiency (60–70 per cent) the diameter D is also specified.

It is evident from Fig. 7 that a large focal length paraboloid for a specified D (i.e., τ small) has several advantages. By using a larger f, a larger subreflector can be used without introducing aperture blocking. The angle θ_{0m} subtended by the subreflector is larger and hence a smaller horn aperture can be used to illuminate the subreflector. Also, a large subreflector will produce less radiation beyond the geometrical illumination angle and hence less spillover at the main reflector. On the other hand, a large subreflector is difficult to support mechanically.

The aperture of the horn feed should be as small as possible to avoid blocking of the main reflector area, but large enough to provide efficient illumination of the subreflector. A uniform phase (plane phase front) aperture feed would be most suitable since it provides the minimum beamwidth for a given aperture size. Such an aperture is not readily

realizable without the use of lenses, but a narrow angle horn provides a suitable alternative. The choice of the horn length and position was also influenced by mechanical considerations. The final horn dimensions were selected by computing the radiation patterns from a horn of fixed length and different angles, and maximizing the power intercepted by the subreflector.

The antenna dimensions are:

Paraboloid offset angle	$\theta_0 = 47.5°$
Focal length of paraboloid	$f = 152\lambda$
Diameter of the subreflector	$d = 40\lambda$
Horn length	$l = 100\lambda$
Horn angle	$\alpha = 3.25°$
Main reflector geometrical illumination angle	$\theta_m = 30.5°$
Subreflector geometrical illumination angle	$\theta_{1m} = 7.5°$.

The above dimensions and the location of the feed horn phase center which is 88.85λ from the vertex of the horn completely specify the antenna geometry.

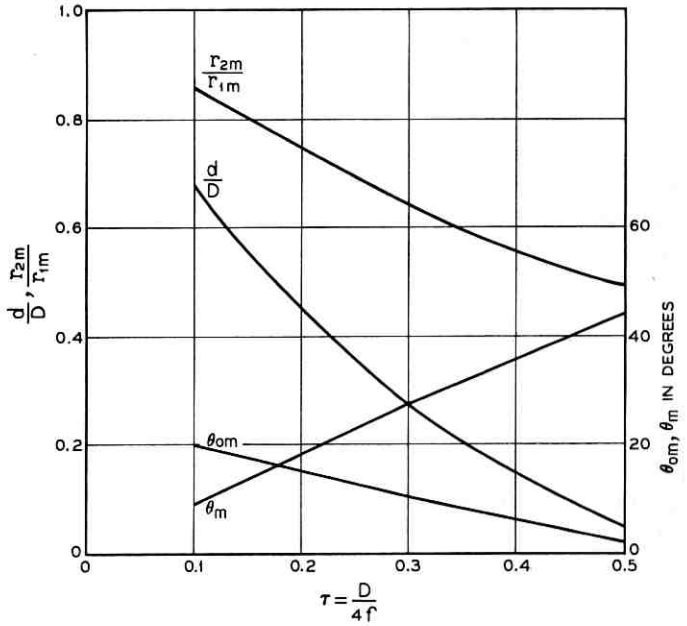

Fig. 7 — Antenna parameters.

IV. RADIATION PATTERN COMPUTATIONS

4.1 *Horn Radiated Pattern*

4.1.1 *Radiation Integrals*

The objectives of the horn pattern computations were: (1.) to determine the phase center and phase front of the radiation pattern, (2.) to optimize the horn angle for a specified horn length for maximum power interception by the subreflector, and (3.) to determine the TM_{11} to TE_{11} mode ratio for phase equalization of the radiation pattern in the two principal planes.

The radiation pattern from the horn has been computed by using the Kirckhoff approximation to the aperture radiation. Based on the above, the electric field, $\bar{E}_p$, at a distance of at least a few wavelengths from the aperture is:[6]

$$\bar{E}_p = \frac{jk}{4\pi} \int_s \int [\bar{E}_a(1 + 1_n \cdot 1_R) - \bar{E}_a \cdot 1_R(1_n + 1_R)] \frac{e^{-jkR}}{R} ds \quad (16)$$

where

s = horn aperture area
$k = 2\pi/\lambda$ = propagation constant
λ = wavelength

and 1_n and 1_R are unit vectors in the normal and R direction respectively as shown in Fig. 8. $\bar{E}_a$ is the field in the horn aperture, assumed to be the

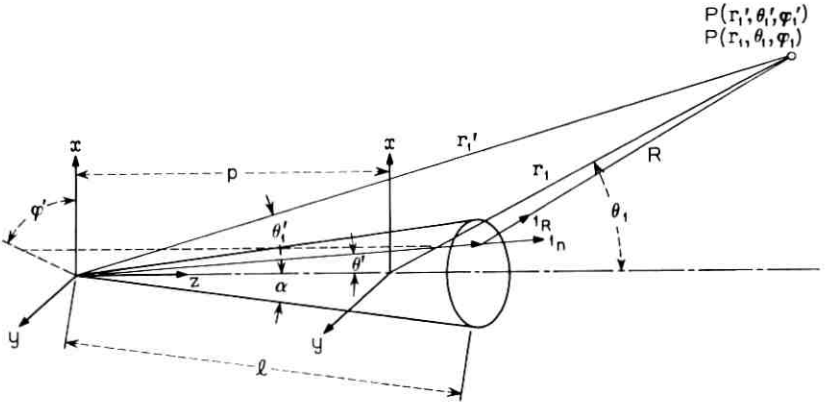

Fig. 8 — Coordinates for horn pattern computations.

same as for circular waveguide propagating TE_{11} and TM_{11} modes, but with spherical phase fronts.

The aperture field $\bar{E}_a$ has both TE_{11} and TM_{11} mode components designated by:

$$\bar{E}_a = \bar{E}_{a\mathrm{TE}} + \bar{E}_{a\mathrm{TM}}. \tag{17}$$

The aperture fields used in the computation are given in Appendix B.

Two different observation coordinate systems have been used to compute the radiation integral (16). The initial computation was performed in the coordinate system θ_1', φ_1' with origin at the vertex of the horn. From this computation, the phase center of the radiation pattern was determined. The subsequent integration was performed in the θ_1, φ_1 coordinate system with origin at the phase center.

Only the first term of the integral (16) has been evaluated, since it has been shown[7] that even for a much wider angle horn than considered here the contribution of the second term is negligible. Furthermore, it is shown in Appendix B that it is sufficient to evaluate (16) for one rectilinear component of one polarization in the two principal planes $\varphi_1 = 0$ and $\varphi_1 = \pi/2$. The other components of the radiation pattern can be derived from that computation.

The following integral has been evaluated in both coordinate systems for the case when the TE_{11} and TM_{11} mode are in phase at the horn aperture.

$$E_{py} = A_y \frac{kl^2}{4\pi} \int_0^{\alpha} \int_0^{2\pi} \left[(E_{ay\mathrm{TE}})_y + \frac{B_y}{A_y} (E_{ay\mathrm{TM}})_y \right] \cdot \frac{e^{-jkR}}{R} (1 + 1_n \cdot 1_R) \sin\theta' \, d\theta' \, d\varphi'. \tag{18}$$

The subscript y indicates that the y components of (68) and (71) have been used.

The resulting radiation patterns for both polarizations are:

$$(\bar{E}_p)_x = 1_{\theta_1} E_{py}(\pi/2) \cos\varphi_1 - 1_{\varphi_1} E_{py}(0) \sin\varphi_1 \tag{19}$$

$$(\bar{E}_p)_y = 1_{\theta_1} E_{py}(\pi/2) \sin\varphi_1 + 1_{\varphi_1} E_{py}(0) \cos\varphi_1 . \tag{20}$$

The explicit φ_1 dependence of the radiation patterns simplifies considerably the subsequent computations of the radiation from the subreflector.

4.1.2 *Phase Centers*

The phase center designates the location of the center of curvature of a spherical surface which is tangential to the equiphase surface near the

maximum of the radiation pattern. The phase center is particularly useful when the equiphase surface is nearly spherical over a relatively wide angle. For such a radiation pattern the phase variation is minimized in a spherical coordinate system with its origin at the phase center. The existence of a well-defined phase center depends on the spatial dependence of the equiphase surface. However, through any principal plane of the radiation pattern the radius of curvature at the center of the equiphase curve can be determined. For a circular horn coherently excited in the TE_{11} and TM_{11} modes it is possible to adjust the ratio of the two modes such that the centers of curvature of the radiation in the two principal planes coincide.

The radii of curvature in the two principal planes can be determined from the integral for the radiation pattern (18).[8,9] An alternate method is to compute the radiation pattern of the horn in a spherical coordinate system with its origin at the horn vertex. From the phase variation of the radiation pattern in this coordinate system at a constant radius the radii of curvature can also be determined. The latter method was used here.

Referring to Fig. 8, and assuming that the radiation pattern has a spherical wave front over a certain angular range, the phase dependence, δ, of the electric field will be

$$\delta = -(kr_1 + \delta_0). \tag{21}$$

The phase of the field computed in a coordinate system with its origin at the vertex of the horn

$$\delta = -\left(k\sqrt{(r_1')^2 + p^2 - 2r_1'p\cos\theta_1'} + \delta_0\right). \tag{22}$$

The relative phase, δ_r with respect to the phase at $\theta_1' = 0$ is:

$$\delta_r = -k[\sqrt{(r_1')^2 + p^2 - 2r_1'p\cos\theta_1'} - (r_1' - p)]. \tag{23}$$

From (23), p can be determined since δ_r is known from the computation:

$$\frac{p}{\lambda} = \frac{\frac{\delta_r}{4\pi}\left(2\frac{r_1'}{\lambda} - \frac{\delta_r}{2\pi}\right)}{\left[\frac{\delta_r}{2\pi} - \frac{r_1'}{\lambda}(1 - \cos\theta_1')\right]}. \tag{24}$$

From (24) the location of the phase center can be determined as a function of θ_1'. For the horn radiation pattern under consideration p is constant over a wide angular range of θ_1'. The value used for determining p was $\theta_1' = 0.5°$.

The position of the phase center was computed initially for a TE_{11} mode in the H plane. Subsequently, the pattern in the E plane was

computed (in the coordinate system with its origin at the H-plane phase center) for different ratios of TE_{11} to TM_{11} mode. The TM_{11} mode has little effect on the radiation pattern in the H plane (the H plane, far-field radiation pattern of an open waveguide is zero), and the location of the H-plane phase center was unaffected by variations in the mode ratio. The ratio of the TM_{11} to TE_{11} mode was selected to make the phase dependence in the two principal planes similar. The amplitude and phase of the radiation patterns in the E and H planes are shown in Fig. 9 for the following parameters:

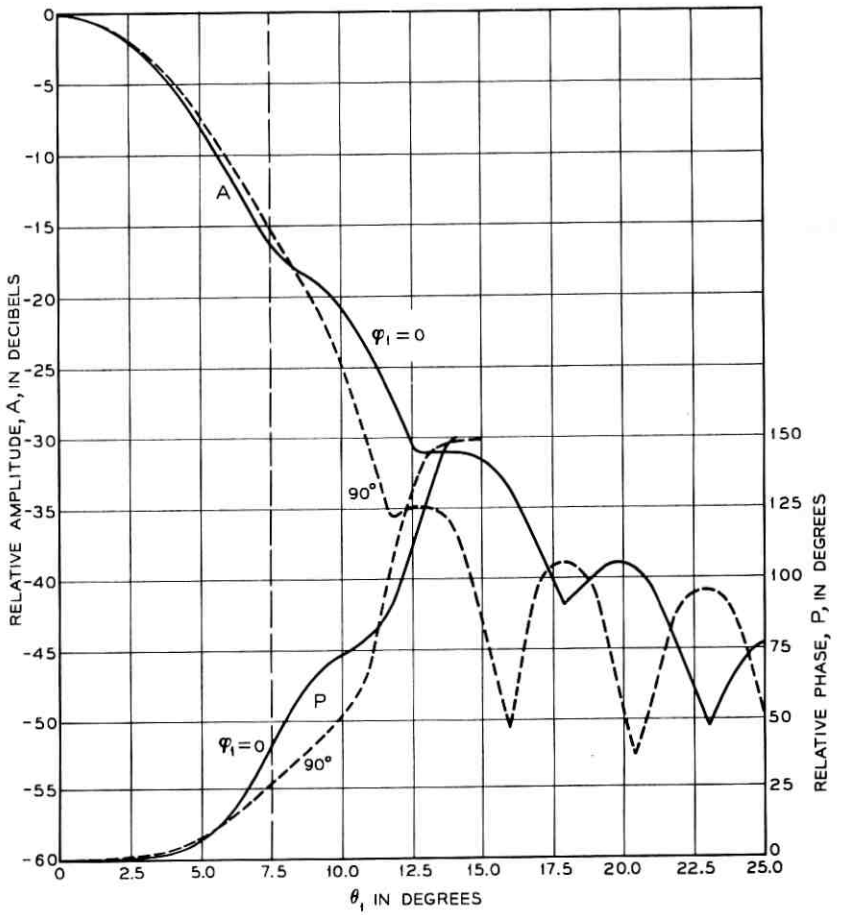

Fig. 9 — Amplitude and phase of feed horn radiation.

$l = 100\lambda$
$\alpha = 3.25°$
TM_{11} to TE_{11} mode ratio $(B_y/A_y) = 0.51$*
$p = 88.85\lambda$
$r = 149.84\lambda$

The subreflector intercepts the power over a 7.5° angle in θ_1. Over this range, the phase deviations are 26° and 39° in the E and H planes, respectively. As subsequently discussed, the subreflector has been compensated for the average phase deviation in the two planes. The power intercepted by the subreflector is 94.5 per cent of the total power radiated by the horn. This value has been obtained by integration of the computed radiation patterns.

4.1.3 *Subreflector Compensation*

The computed radiation pattern from the horn shows that its phase deviates from a spherical wave front. Based on geometrical optics, compensation of the hyperboloid subreflector is necessary to obtain a reflected wave with a spherical wave front. The theorem by Malus states that a surface exists which will convert an incident equiphase surface into a reflected equiphase surface and also satisfy Snell's law of reflection.[6] In Appendix C, a method for the construction of a reflector surface which converts an equiphase surface into a reflected spherical surface is presented. However, for a small phase deviation of the incident wave from a spherical phase front, a simpler method can be used.

Let

$$\Delta\delta_r = (2\pi/\lambda)r - (2\pi/\lambda)r_e \tag{25}$$

where

$\Delta\delta_r$ = phase deviation
r = geometrical distance from the phase center to the hyperboloid surface
r_e = apparent geometrical distance.

The equation for the reflector surface is, from (5),

$$r_e - r_2 = b. \tag{26}$$

Solving for r_2 considering that $\Delta\delta_r$ is small yields:

$$r_2 = r_{2h}(\theta_2) - \frac{\lambda\Delta\delta_r(\theta_2)}{4\pi} \cdot \frac{(1 + \beta^2 + 2\beta\cos\theta_2)}{(\cos\theta_2 + \beta)^2} \tag{27}$$

where $r_{2h}(\theta_2)$ is one equation for the hyperboloid surface given by (8).

* The corresponding TM_{11} to TE_{11} power ratio is 0.173.

4.2 *Radiation Patterns from the Subreflector*

To compute the radiation pattern from the subreflector it is necessary to determine the current in the subreflector due to the incident wave from the horn. Without solving an integral equation for the current, it can only be determined approximately. A good approximation for subreflectors large in comparison to the wavelength is to assume that the subreflector is locally plane. With the above assumption, and the assumption that the interaction of the feed with the subreflector is negligible, the surface current density $\bar{J}$ at the subreflector is:[6]

$$\bar{J} = \frac{2}{\eta} 1_n \times (1_{r_i} \times \bar{E}_i) = \frac{2}{\eta} 1_n \times (1_{r_r} \times \bar{E}_r) \tag{28}$$

where

1_n = unit normal to the subreflector surface
1_{r_i}, 1_{r_r} = directions of propagation for the incident and reflected waves, respectively
$\bar{E}_i$, $\bar{E}_r$ = incident and reflected electric fields, respectively.

The reflected field is related to the incident field by

$$\bar{E}_r = -\bar{E}_i + 2(1_n \bar{E}_i \cdot 1_n) . \tag{29}$$

It is preferable to express the current density in secondary spherical coordinates due to the subsequent integration which will be performed to obtain the subreflector radiation pattern.

For a hyperboloid or compensated hyperboloid subreflector

$$1_{r_r} = 1_{r_2} \tag{30}$$

and

$$1_{n_2} = \frac{1_{r_2}(\cos\theta_2 + \beta) - 1_{\theta_2} \sin\theta_2}{\sqrt{1 + \beta^2 + 2\beta\cos\theta_2}} . \tag{31}$$

For an incident field given by (19) or (20) namely

$$\bar{E}_i = 1_{\varphi_1} E_{\varphi_1} + 1_{\theta_1} E_{\theta_1} . \tag{32}$$

It follows that

$$\bar{E}_r = -1_{\varphi_2} E_{\varphi_1} - 1_{\theta_2} E_{\theta_1} . \tag{33}$$

The radiated electric field, $\bar{E}_s$ from the subreflector due to the current distributions (28) is, with reference to Fig. 6,

$$\bar{E}_s = \frac{j}{\lambda}\int_0^{\theta_m}\int_0^{2\pi} 1_{R_2} \times \left[\left(\bar{E}_r - \frac{1_{r_2}1_{n_2}\cdot\bar{E}_r}{1_{n_2}\cdot 1_{r_2}}\right)\times 1_{R_2}\right] \frac{e^{-jk(R_2+r_2)}}{R_2} r_2^2 \sin\theta_2\, d\theta_2\, d\varphi_2 \tag{34}$$

In (34), E_r represents the amplitudes of the reflected field components. The first term in the brackets of the equation is similar to the Kirckhoff approximation to the radiation from apertures; the second term is due to the radial subreflector currents.

In the computation of the radiation patterns 1_{R_2} has been approximated by 1_r. An actual computation showed that in the principal planes this approximation has a negligible effect on the radiation patterns.

The θ and φ component of radiated field from the subreflector with the above assumption is:

$$\bar{E}_s\cdot 1_{\theta,\varphi} = \frac{-j\eta}{2\pi}\int_0^{2\pi}\int_0^{\theta_m} \bar{J}\cdot 1_{\theta,\varphi}\frac{e^{-jk(R_2+r_2)}}{R_2}\frac{r_2^2\sin\theta_2}{1_{n_2}\cdot 1_{r_2}}\, d\theta_2\, d\varphi_2 . \tag{35}$$

From (28)–(33)

$$\bar{J}\cdot 1_\theta = \frac{2}{\eta\sqrt{1+\beta_2+2\beta\cos\theta_2}}\{E_{\theta_1}[\cos\theta\cos(\varphi_2-\varphi)\cdot(1+\beta\cos\theta_2)+\beta\sin\theta\sin\theta_2] - E_{\varphi_1}\cos\theta\sin(\varphi_2-\varphi)(\cos\theta_2+\beta)\} \tag{36}$$

and

$$\bar{J}\cdot 1_\varphi = \frac{2}{\eta\sqrt{1+\beta^2+2\beta\cos\theta_2}}\{E_{\theta_1}\sin(\varphi_2-\varphi)(1+\beta\cos\theta_2) + E_{\varphi_1}\cos(\varphi_2-\varphi)(\cos\theta_2+\beta)\}. \tag{37}$$

The distance R_2 is

$$R_2 = \sqrt{r^2+r_2^2-2rr_2\cos\gamma_2} \tag{38}$$

with

$$\cos\gamma_2 = \sin\theta\sin\theta_2\cos(\varphi-\varphi_2)+\cos\theta\cos\theta_2 . \tag{39}$$

The φ_2 dependence of E_{θ_1} and E_{φ_1} for x and y polarization is given by (19) and (20) respectively with $\varphi_1 = \varphi_2$.

It is sufficient to evaluate (34) for one polarization only in the two principal planes $\varphi = 0$ and $\varphi = \pi/2$, to obtain the φ dependence of the

radiation pattern on a spherical surface centered about the focal point of the hyperboloid. Furthermore, for y polarization, the φ component of (35) may be evaluated at $\varphi = 0$ and the θ component at $\varphi = \pi/2$. The φ dependence of the radiation pattern is the same as for the horn (19) and (20).

The φ and θ components of the electric field radiated by the subreflector, due to an incident field polarized in the y direction has been computed in the H and E plane, respectively. Figs. 10 and 11 show the amplitude and phase of the radiation pattern in the H plane. Values are shown along a circle whose center is at the focus of the hyperboloid and which passes through the intersection of the subreflector axis and the main reflector. A sample computation at distances corresponding to the location of the paraboloid surface showed an inverse distance relationship as would be expected from a spherical wave. Fig. 12 shows the E-plane pattern at the location of the paraboloid surface in the illumination region, and at a constant radius in the shadow region corresponding to the radial distance to the edge of the paraboloid reflector. The phase dependence of the E-plane and H-plane patterns are similar.

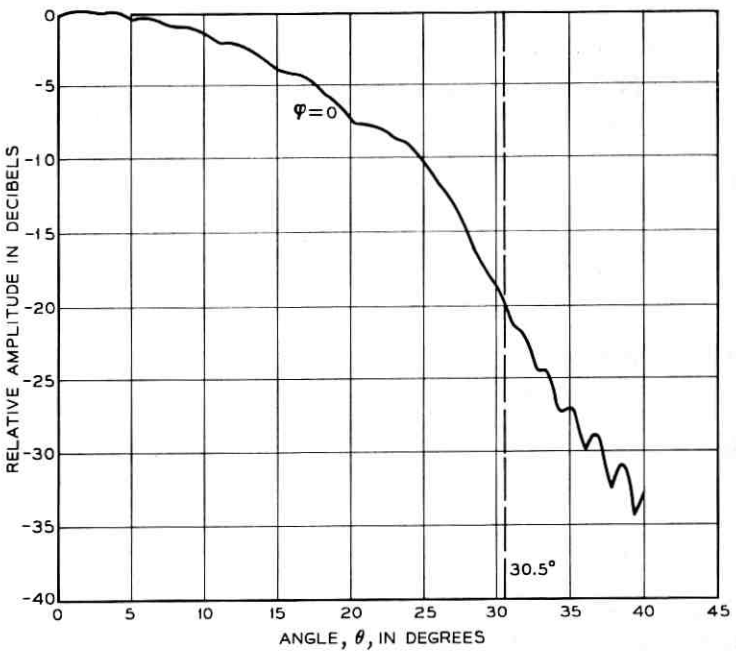

Fig. 10 — Computed subreflector radiation pattern (H plane).

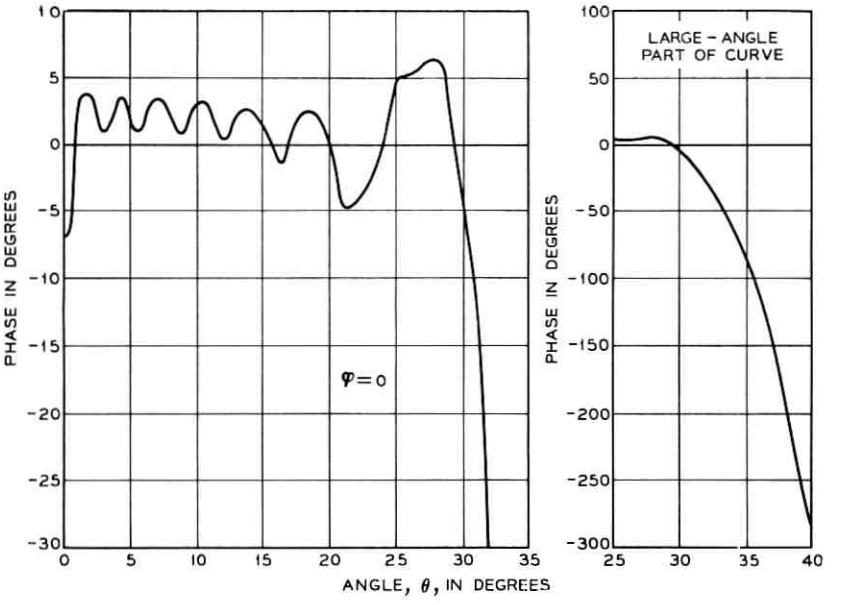

Fig. 11 — Computed phase of subreflector radiation pattern (H plane).

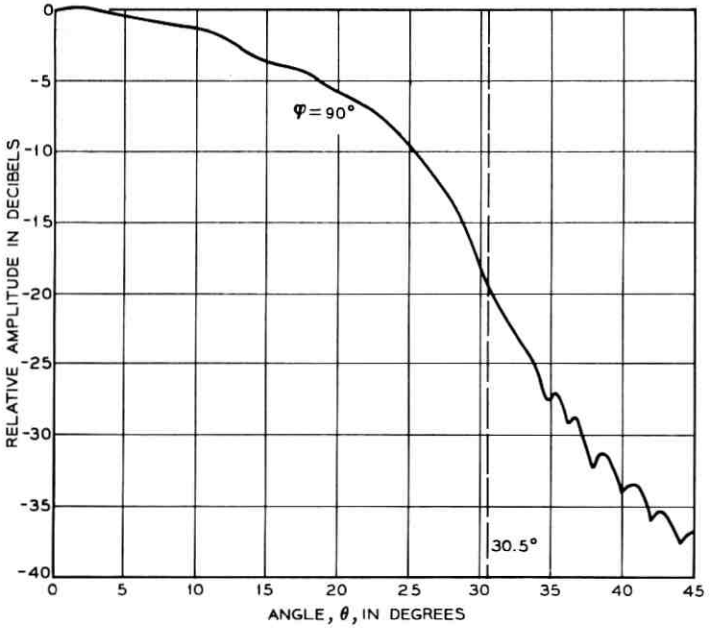

Fig. 12 — Computed subreflector radiation pattern (E plane).

There is a distinct difference in radiation patterns in geometrical illumination and shadow regions. In the illumination region the radiation is essentially spherical. In the shadow region (θ larger than 30.5°) the phase varies rapidly, and amplitude variations also occur.

The pattern outside the geometrical illumination region is of utmost interest since it largely determines the antenna noise temperature. The radiated power of the subreflector has been computed as a function of θ by integration of the radiation patterns. The radiated power normalized with respect to the radiated power in angular range 0–45 degrees is shown in Fig. 13 for the $\varphi = 0, \pi/2$ and π radiation patterns. Fig. 13 also shows the antenna noise temperature due to spillover at the main reflector as a function of the angle subtended by the main reflector. The computed antenna noise temperature assumes earth radiation temperature of 300°K.

4.3 *Radiation Pattern from the Main Reflector*

The radiation pattern from the main reflector can be determined by the same method as the radiation pattern from the subreflector. However, the radiation from a paraboloid reflector is primarily determined by the reflected electric field (i.e., the longitudinal currents can be neglected).

The reflected electric field is related to the incident field by (29). The

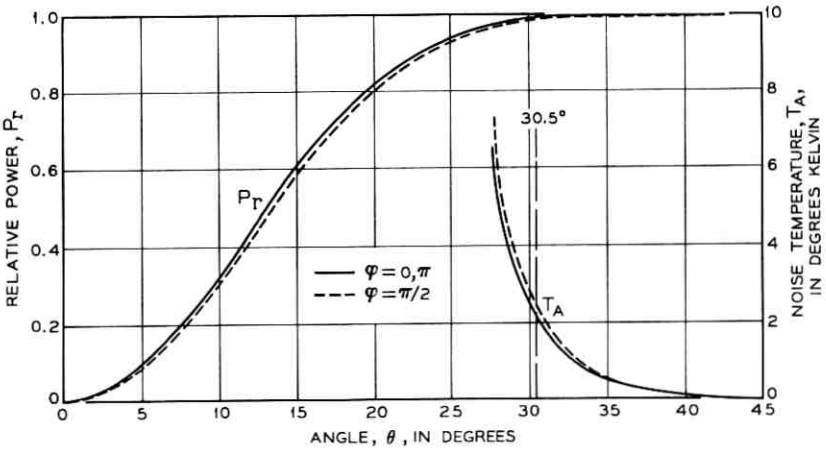

Fig. 13 — Antenna noise temperature and power radiated by the subreflector.

far field of the antenna is given by:

$$\bar{E}_r = \frac{j}{\lambda} \cdot \frac{e^{-jkr_p}}{r_p} \int \bar{E}_{rs} \exp\left[jk \sin\theta_a (x_{pc} \cos\varphi_a + y \sin\varphi_a)\right] ds. \quad (40)$$

The integration is to be performed over the projected area of the paraboloid in the x_p, y plane, and x_{pc} is the coordinate in the x_p direction from the center of the projected circle of the paraboloid. It is preferable to perform the integration in the subreflector coordinate system, since the incident field is already expressed in terms of the secondary coordinate. The vector dependence of the reflected field has to be determined in terms of its components x_p and y. The vector dependence of the reflected field can be obtained from (29), but more directly by noticing the φ and θ components of the incident field transform by reflection into the circles (3) and (4). The unit normal to these circles gives the directions of the reflected θ and φ components respectively. The relationships between the points on the paraboloid surface and their projections onto the x_p, and y plane are

$$x_p = \frac{2f(\cos\theta_0 \sin\theta \cos\varphi + \sin\theta_0 \cos\theta)}{1 + \cos\theta \cos\theta_0 - \sin\theta_0 \sin\theta \cos\varphi} \quad (41)$$

$$y = \frac{2f \sin\theta \sin\varphi}{1 + \cos\theta \cos\theta_0 - \sin\theta \sin\theta_0 \cos\varphi}. \quad (42)$$

In terms of x_p and y components, the electric field is:

$$\bar{E}_{rs} = \frac{\begin{bmatrix} 1_{xp}\{[\sin\theta_0 \sin\theta - \cos\varphi\ (1 + \cos\theta_0 \cos\theta)]E_\theta \\ \quad + \sin\varphi\ (\cos\theta_0 + \cos\theta)E_\varphi\} \\ -1_y\{\sin\varphi\ (\cos\theta_0 + \cos\theta)E_\theta \\ \quad - [\sin\theta \sin\theta_0 - \cos\varphi\ (1 + \cos\theta \cos\theta_0)]E_\varphi\} \end{bmatrix}}{1 + \cos\theta_0 \cos\theta - \sin\theta_0 \sin\theta \cos\varphi} \quad (43)$$

where E_φ and E_θ are the incident components of the electric field at the paraboloid.

It remains to determine the surface element ds in θ, φ coordinates. The surface element can be determined from the Jacobian of (41) and (42). However, a simpler method is available by considering the properties of the paraboloid. The surface element ds is

$$ds = \left(\frac{r^2 \sin\theta\ d\theta\ d\varphi}{1_{np} \cdot 1_r}\right) 1_{np} \cdot 1_{zp} \quad (44)$$

1_{np} = unit normal to the paraboloid surface. But for a paraboloid

$$1_{np} \cdot 1_r = 1_{np} \cdot 1_{zp}. \tag{45}$$

Therefore,

$$ds = r^2 \sin \theta \, d\theta \, d\varphi \tag{46}$$

where r is the equation for the paraboloid (2).

The on-axis antenna gain, G_M, has been computed for y polarization from the relation:

$$G_M = \frac{4\pi}{\lambda^2} \frac{\left| \int_0^{2\pi} \int_{\theta_h}^{\theta_m} E_{rsy} \, r^2 \sin \theta \, d\theta \, d\varphi \right|^2}{\int_0^{2\pi} \int_0^{\theta_m} (E_{rsy}{}^2 + E_{rsx}{}^2) \, r^2 \sin \theta \, d\varphi}. \tag{47}$$

The angle, θ_h, subtended by the horn at the focal point of the subreflector is 1.9°. (About 2 per cent of the power radiated by the subreflector is incident on the feed horn.)

The aperture efficiency, g, has been obtained from the relation

$$g = G_M / G_0 \tag{48}$$

where G_0 is the gain of a uniformly illuminated aperture.

From (3), G_0 is:

$$G_0 = \left[\frac{4\pi f \sin \theta_m}{\lambda (\cos \theta_m + \cos \theta_0)} \right]^2. \tag{49}$$

Equation (47) has been evaluated using the computed fields of the subreflector and assuming a uniform phase illumination. The results are tabulated in Table I.

The radiation patterns from the antenna have been computed from (40) for y polarization and for an illumination angle of 30.5°. The patterns in the plane of antenna symmetry, $\varphi_a = 0$ (x_p, z_p plane), and in the plane perpendicular to the plane of antenna symmetry, $\varphi_a = \pi/2$, are shown in Figs. 14 and 15 together with the measured patterns. Fig. 14 shows also the computed cross-polarized radiation pattern in the plane $\varphi_a = \pi/2$. The cross-polarized component is zero in the plane of antenna symmetry.

TABLE I

ANTENNA GAIN AND APERTURE EFFICIENCY

θ_m	G_M	g
30.5°	54.46 db	70.4%

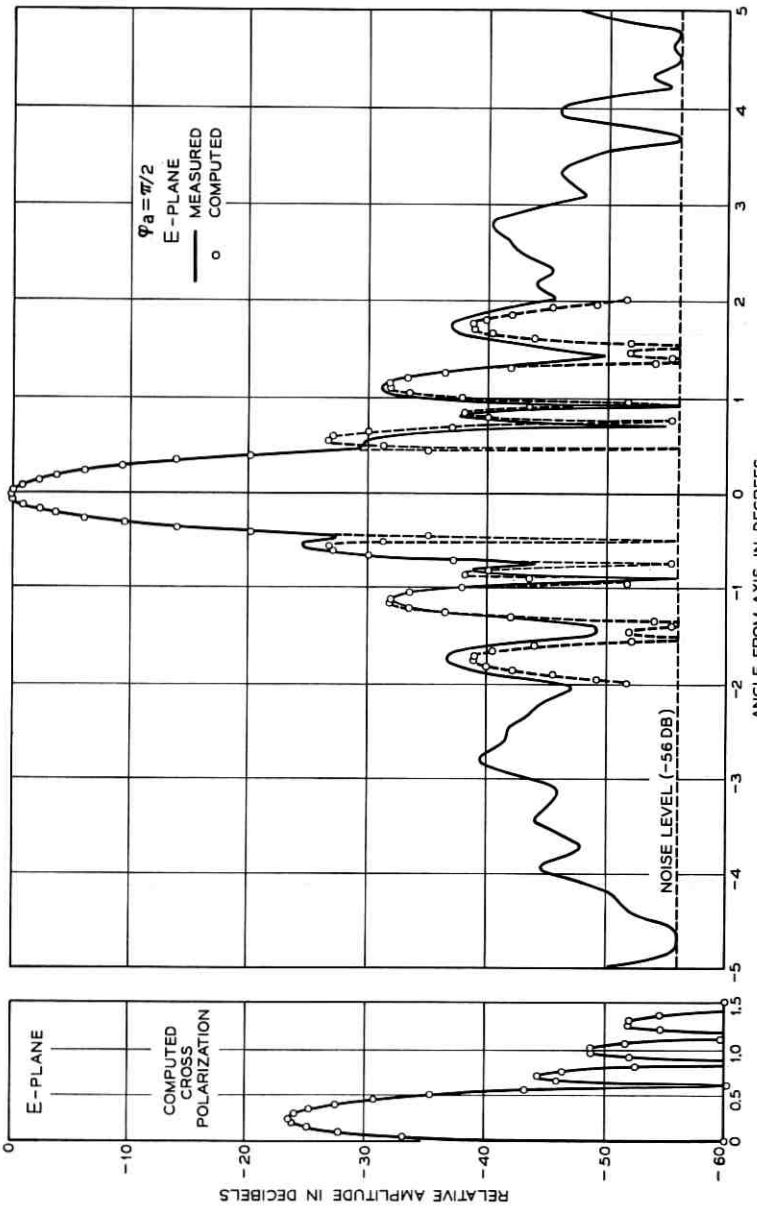

Fig. 14 — Antenna radiation pattern in the plane of asymmetry.

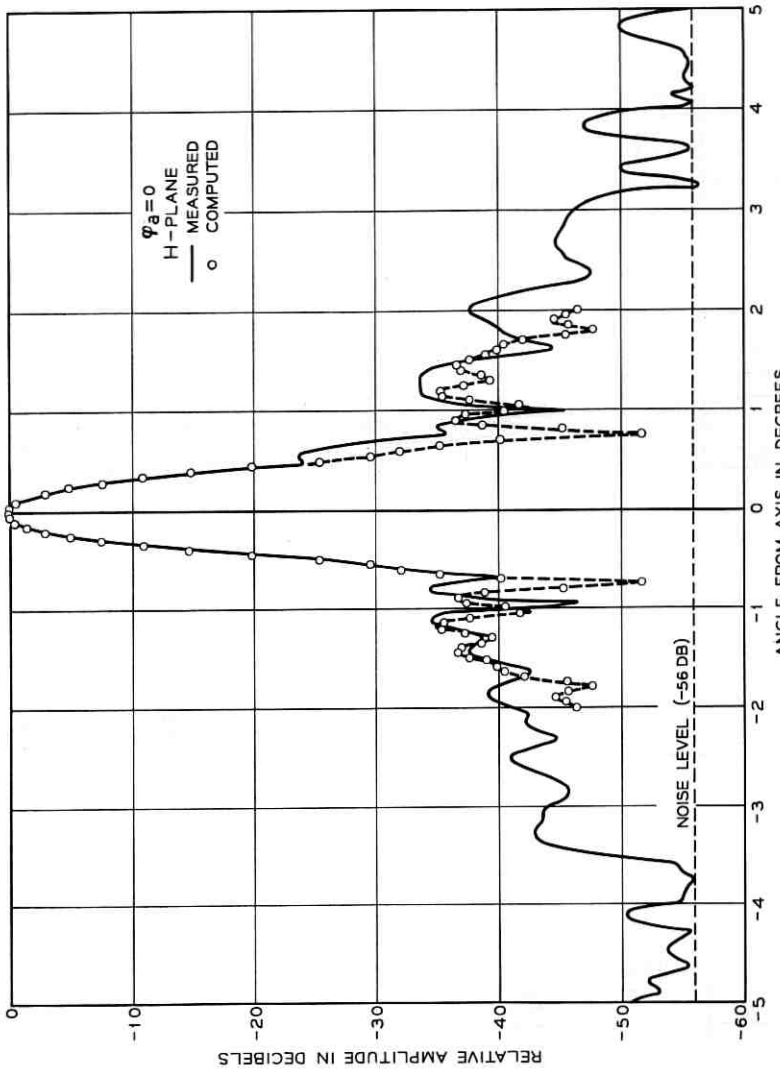

Fig. 15 — Antenna radiation pattern in the plane of symmetry.

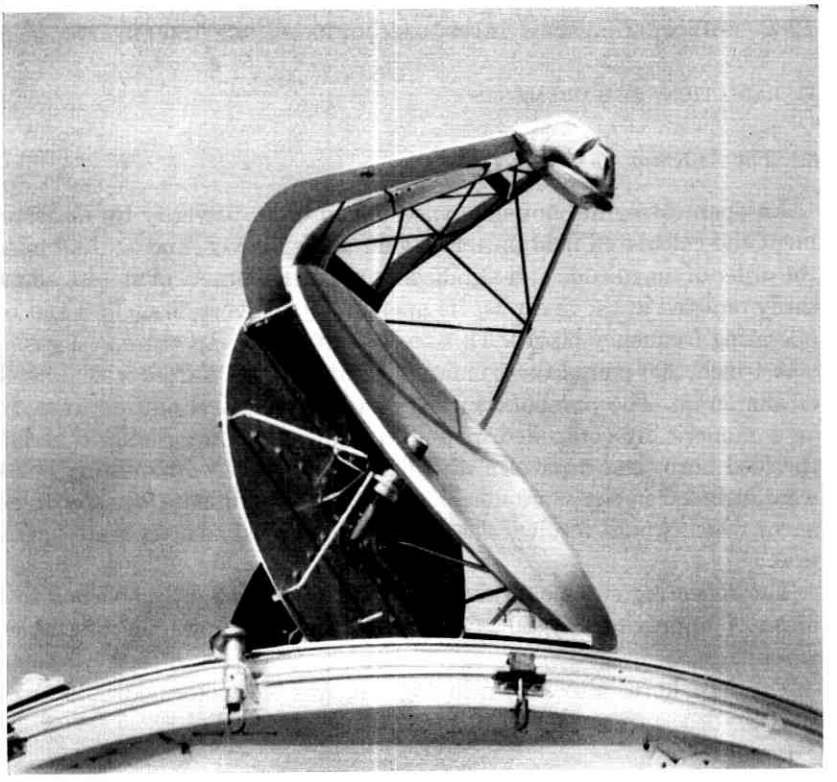

Fig. 16 — Antenna assembly.

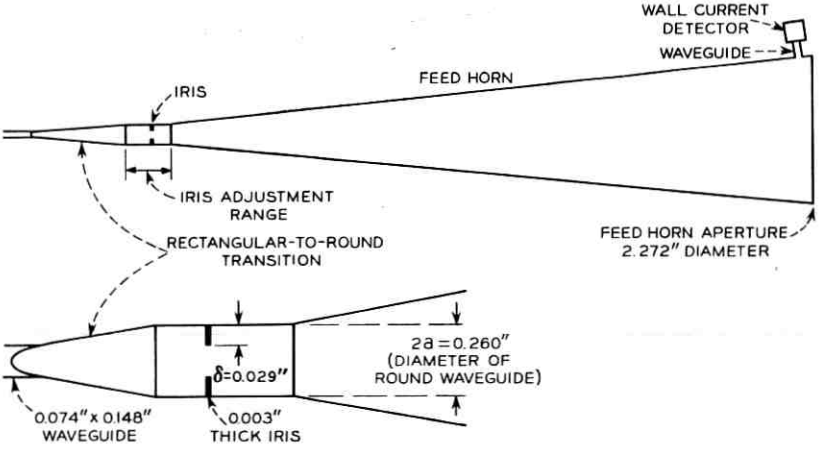

Fig. 17 — Mode converter.

V. RADIATION MEASUREMENTS

5.1 *The Antenna*

An open cassegrain antenna was built in order to verify by measurement the calculated field patterns, gain and spillover, and to determine the order of magnitude and significance of certain practical factors necessarily omitted in the analysis. Its aperture diameter is 40 inches, and its operating frequency 60 gcs. This scales to a 50-foot aperture at 4 gcs.

A 1-inch, 200-pound thick main reflector was milled from a solid block of aluminum. The parabolic surfaces and the elliptical outline were cut on a numerically-controlled milling machine. A slanting hole to accept the feed horn, and a pattern of drilled and spot faced mounting holes, were included in the same machine program to eliminate transfer tolerances which would result if the holes were located as a separate operation.

The reflecting surface of the main reflector was lightly polished to a finish of approximately 50 micro-inches. Fig. 16 shows the assembled antenna.

The feed horn, also shown in Fig. 16, was electroformed, and thick

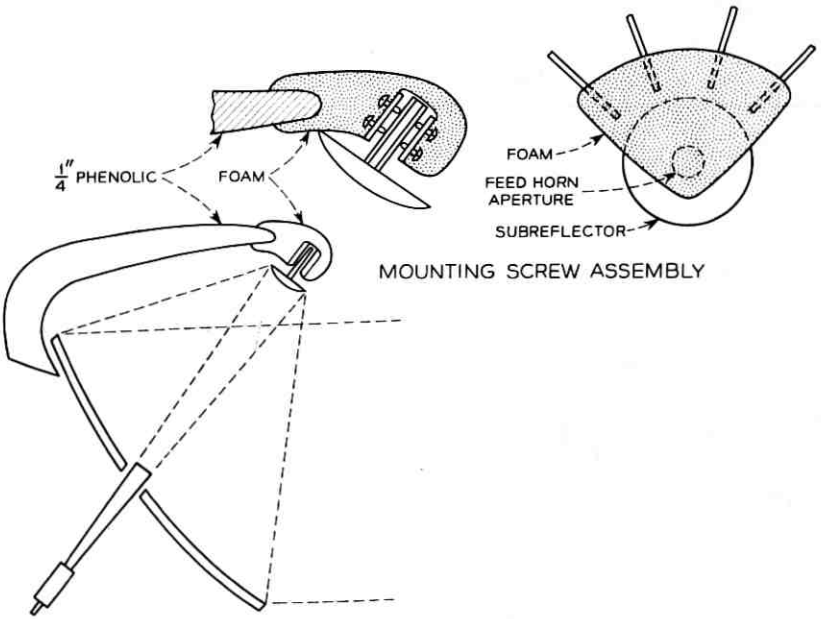

Fig. 18 — Subreflector support.

collars were added so that two sets of adjustable 3-point mounting screws could hold it in place without distorting the horn geometry.

A mode converter was mounted on the small end of the feed horn. The mode converter consists of a single iris chosen to give the required mode conversion when correctly placed with respect to the waveguide transition[2] (Fig. 17). The position of the iris along the horn axis was determined experimentally by monitoring the longitudinal wall current at the horn aperture until a minimum was observed, indicating that the contributions of the TE_{11} and TM_{11} modes were 180° out of phase.

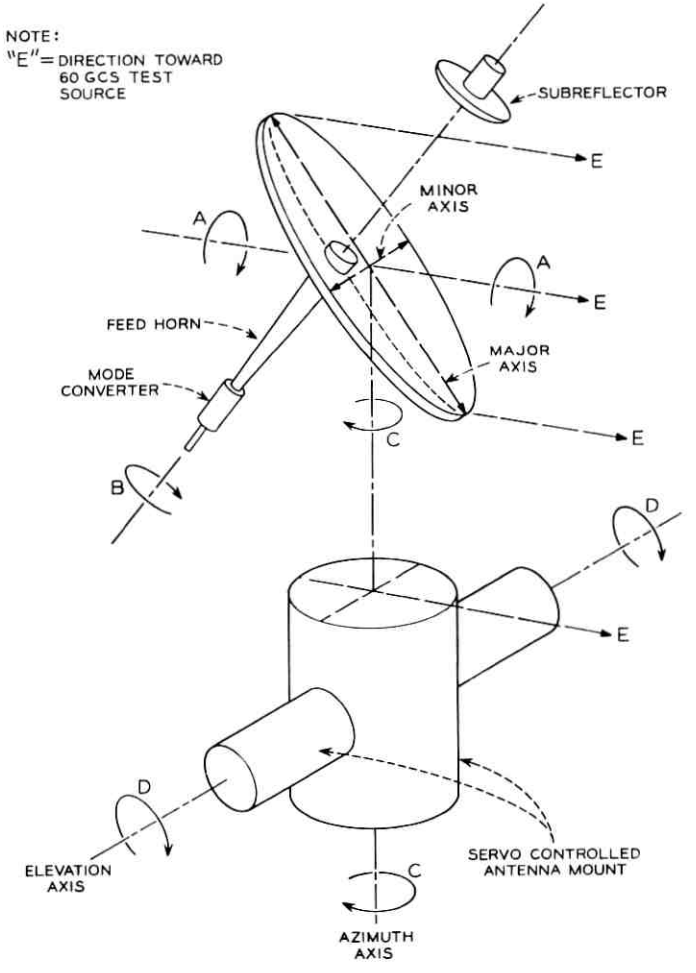

Fig. 19 — Antenna test mount.

The subreflector was turned from an aluminum disc, and as a convenient means of adjusting it, a stem was attached to its rear surface as shown in Fig. 18.

A framework of four $\frac{1}{4}$-inch thick legs of cotton fabric-filled phenolic material was attached to the rear surface of the main reflector. The small ends of the phenolic legs were then jointed together and to the subreflector mounting screw assembly by a block of polyurethane (isocyanate) foam. To eliminate the need for clamps and fasteners, the polyurethane was foamed in place around the items to be joined and supported. The phenolic legs stop short of the edge of the subreflector so that spillover radiation "sees" only the polyurethane foam as shown in Fig. 18.

5.2 *Mounting*

For convenience in electrical testing, the antenna was mounted on a stand which has a horizontal bearing oriented axially with the main beam. The complete antenna can be rotated by hand to any position around this axis "*A*" (Fig. 19), so that the subreflector may be located above, below, or at either side of the main reflector when viewed from the far field source.

The feed horn and mode converter assembly may be rotated by hand (with respect to the rest of the antenna) around its own axis "*B*".

Positioning around axes "*A*" and "*B*" permits *E*-plane and *H*-plane patterns to be measured in any geometrical plane of the antenna.

The stand supporting the antenna was erected on a two-axis servo-

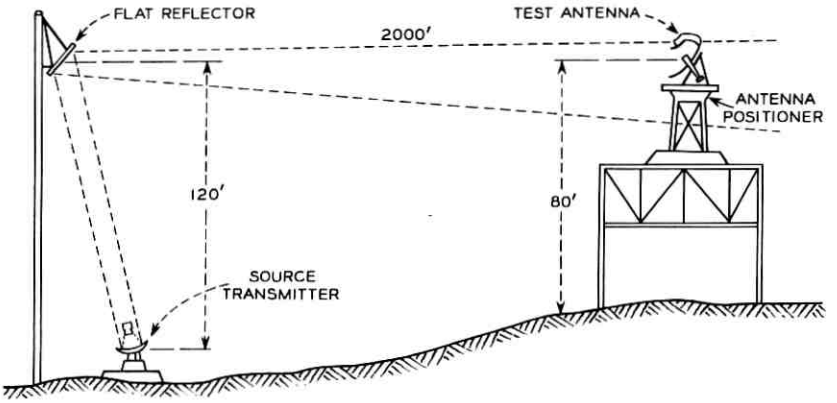

Fig. 20 — Antenna test range.

controlled antenna mount, with vertical axis "*C*" and a horizontal axis "*D*" (Fig. 19). The antenna was located with the intersection of the major and minor axes of the main reflector ellipse intersecting the mount vertical axis "*C*", so that the geometrical center of the main reflector remains stationary when the mount rotates in azimuth. All patterns were measured in the horizontal plane.

5.3 *Range Facility*

The test range configuration is shown in Fig. 20. The field incident on the aperture of the antenna is relatively flat, maximum at the center and between 0.10 and 0.15 db lower around the periphery.

The top of the source tower was stabilized by a long, trussed crossarm and guy wires, limiting movement of the flat, inclined reflector. In a 15-mph wind, the field amplitude variation at the test antenna aperture is approximately 0.1 db, and with 5-to 10-mph winds amplitude variations are not readily detectable.

5.4 *Patterns*

Patterns were measured with the complete antenna installed on the antenna mount of the 2000-foot range facility. *E*-plane and *H*-plane patterns are shown in Figs. 14 and 15.

5.5 *Short-Range Test Facility*

For measuring characteristics of the feed horn with its mode converter, and for measuring the subreflector pattern, a short-range test facility was established. Principal components of the facility are an azimuth antenna positioner (turntable) and three pedestals of rigid polystyrene foam, any one of which may be mounted on the turntable. The three foam pedestals support the feed horn with its mode converter, the subreflector, and either an 8-db or a 16-db pickup horn for sampling the fields.

5.5.1 *Feed Horn Patterns*

A polystyrene foam pedestal supporting the feed horn, with its attached mode converter, was mounted on the azimuth turntable. The feed horn was positioned so that its phase center was located on the vertical axis of rotation of the turntable (see Fig. 21).

For measurement of the patterns, a pickup horn with a $\frac{3}{16}$-inch square

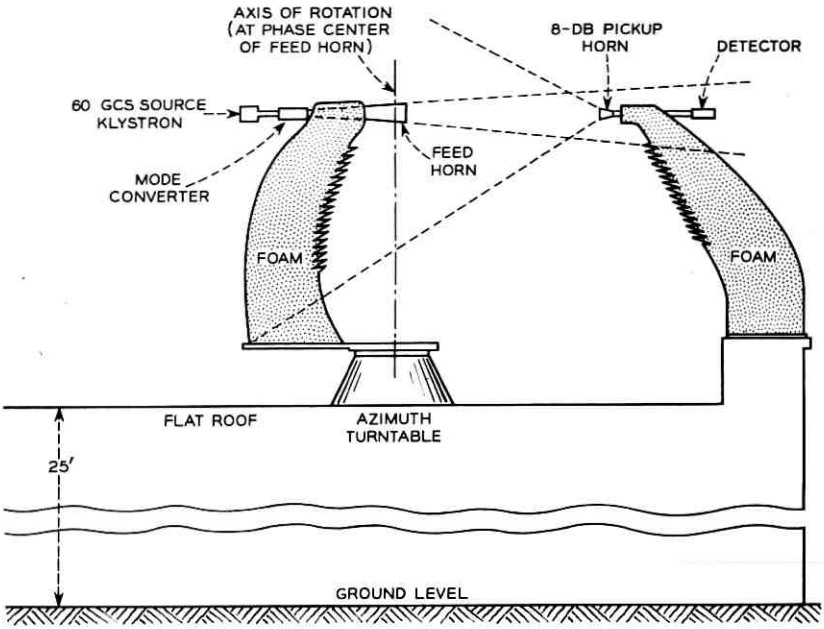

Fig. 21 — Test range for horn radiation measurements.

aperture was designed and fabricated. This horn met the following requirements:

(1.) Directivity was adequate so that no side reflections were detected.

(2.) Deviations of phase or amplitude in the field pattern of the feed horn are small over the angle (approximately 0.35 degrees) intercepted by the $\frac{3}{16}$-inch aperture of the pickup horn.

(3.) With the 100-milliwatt source, and the particular detecting and recording equipment used, the gain of the pickup horn (of the order of 8 db) resulted in a dynamic range, or maximum signal-to-noise ratio, of greater than 55 db.

Figs. 22 and 23 are the E-plane and H-plane measured patterns of the feed horn.

5.5.2 *Subreflector Patterns*

The feed horn with its mode converter was mounted on a polystyrene foam pedestal, and the subreflector was mounted on a second foam pedestal. A pickup horn, supported by a third foam pedestal, was mounted on the azimuth turntable (Fig. 24). The subreflector was located so that its focal point was on the vertical axis of rotation of the turntable.

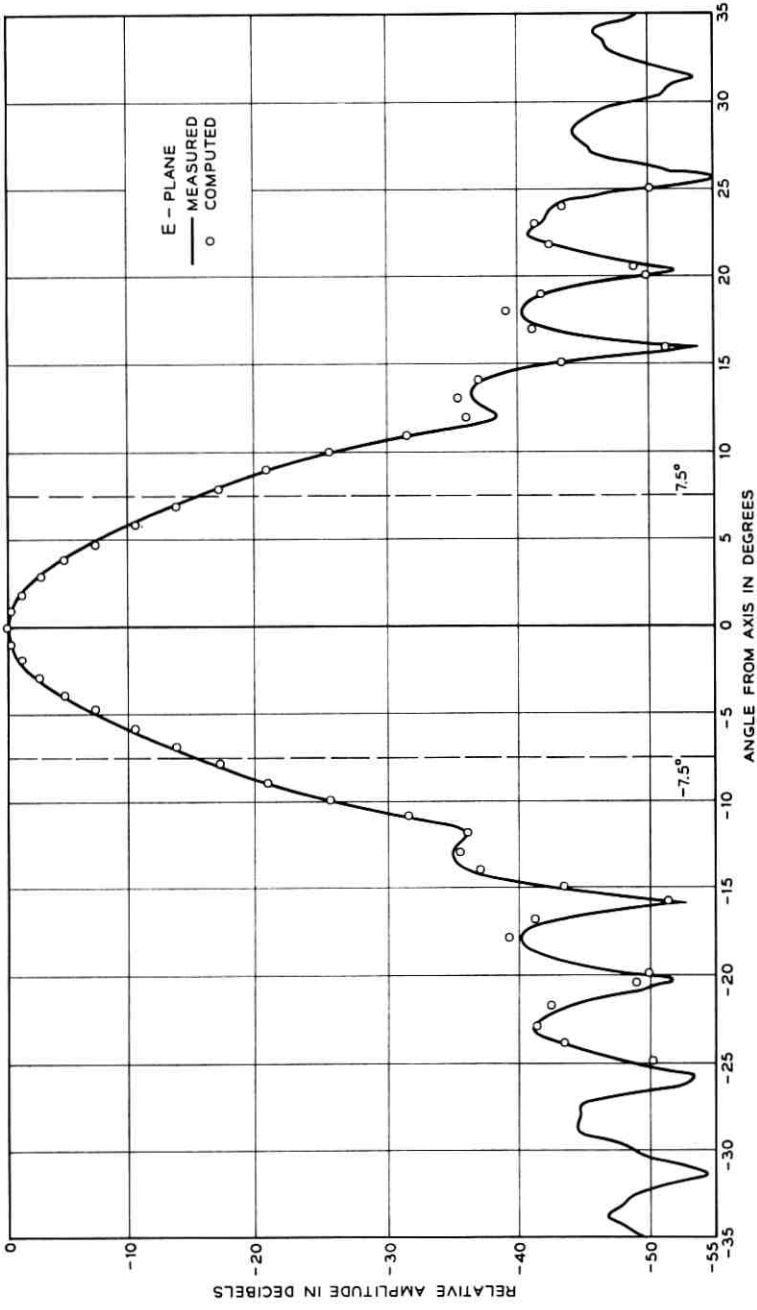

Fig. 22 — E-plane radiation pattern of feed horn.

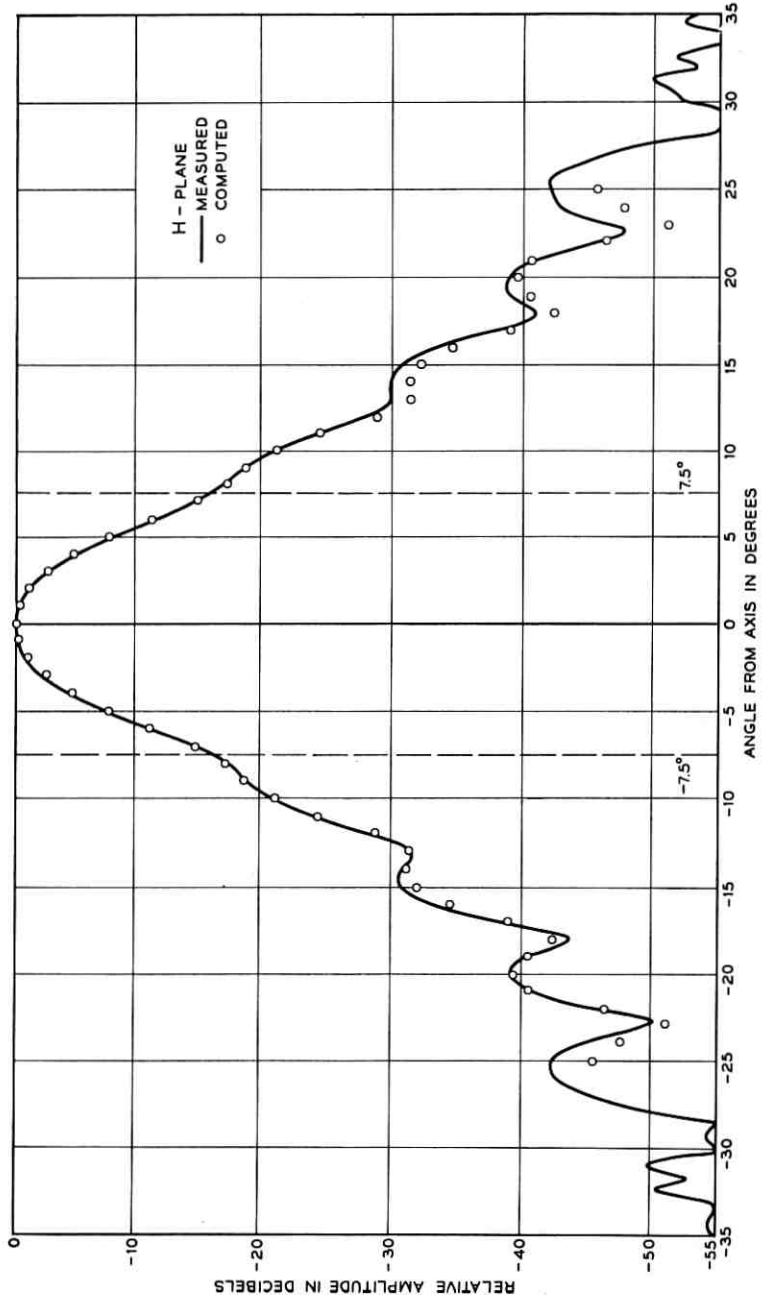

Fig. 23 — H-plane radiation pattern of feed horn.

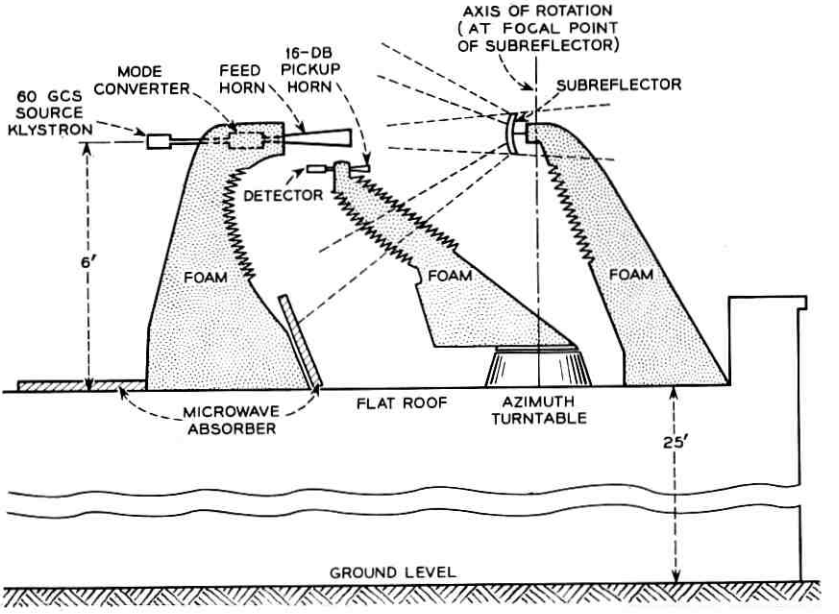

Fig. 24 — Test range for subreflector radiation measurements.

For the subreflector patterns, a 16-db horn was designed and fabricated, meeting requirements of directivity, gain, and pattern resolution somewhat more stringent than those for the feed-horn pattern measurements. In order to gain a little more dynamic range, a 500-milliwatt klystron was used as a 60-gcs source.

Because of the geometry of the antenna, the distance from the subreflector focus to different parts of the main reflector varies. Several subreflector patterns were made at different distances from the focus of the subreflector. The E- and H-plane patterns shown in Figs. 25 and 26 were taken at the distance corresponding to the intersection of the main reflector with the subreflector axis. Patterns at other distances were similar.

For clarity of illustration, Fig. 24 shows the pickup horn located directly below the feed horn. Actually, the feed horn-subreflector axis lies in the plane in which the pickup horn is rotated. Thus, at the closest distance to the subreflector focus, the pickup horn can rotate in front of the feed horn. At the other distances, the pickup horn is stopped short of boresight by 2 or 3 degrees. This missing portion of the center of the pattern is of little consequence, since it is intercepted by the aperture of the feed horn, and does not strike the main reflector.

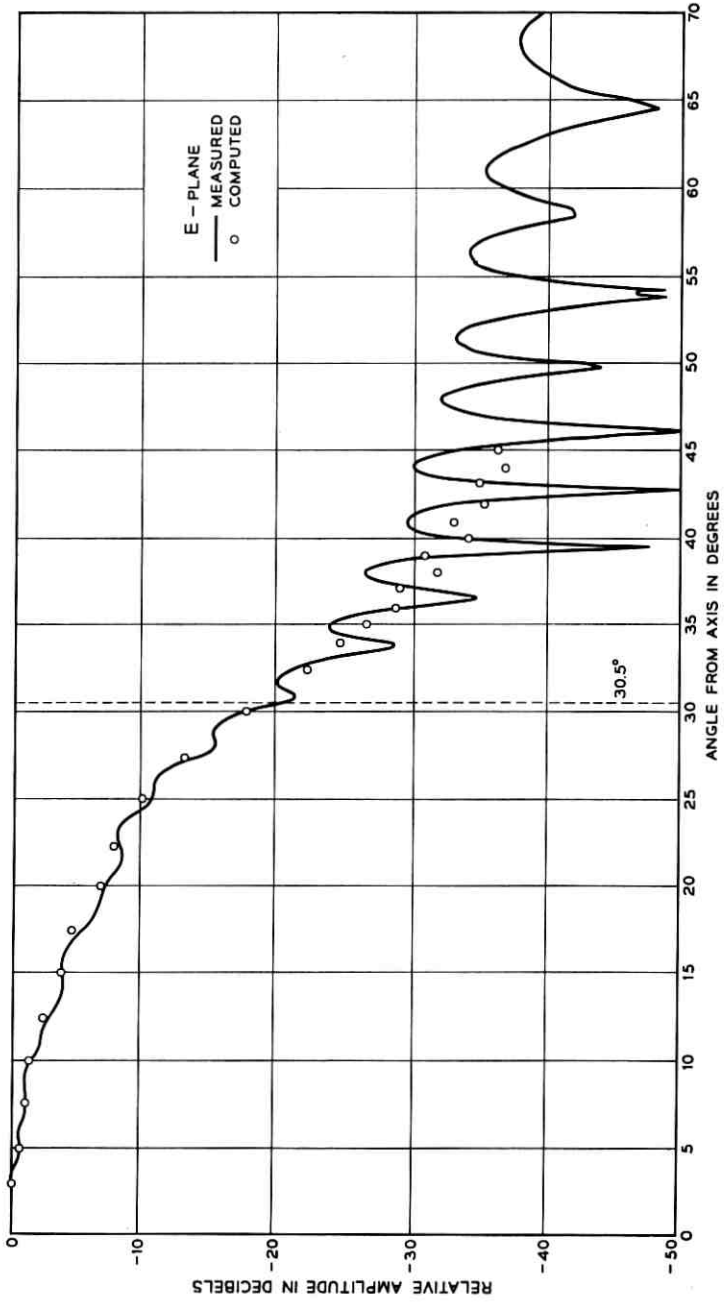

Fig. 25 — E-plane radiation pattern of subreflector.

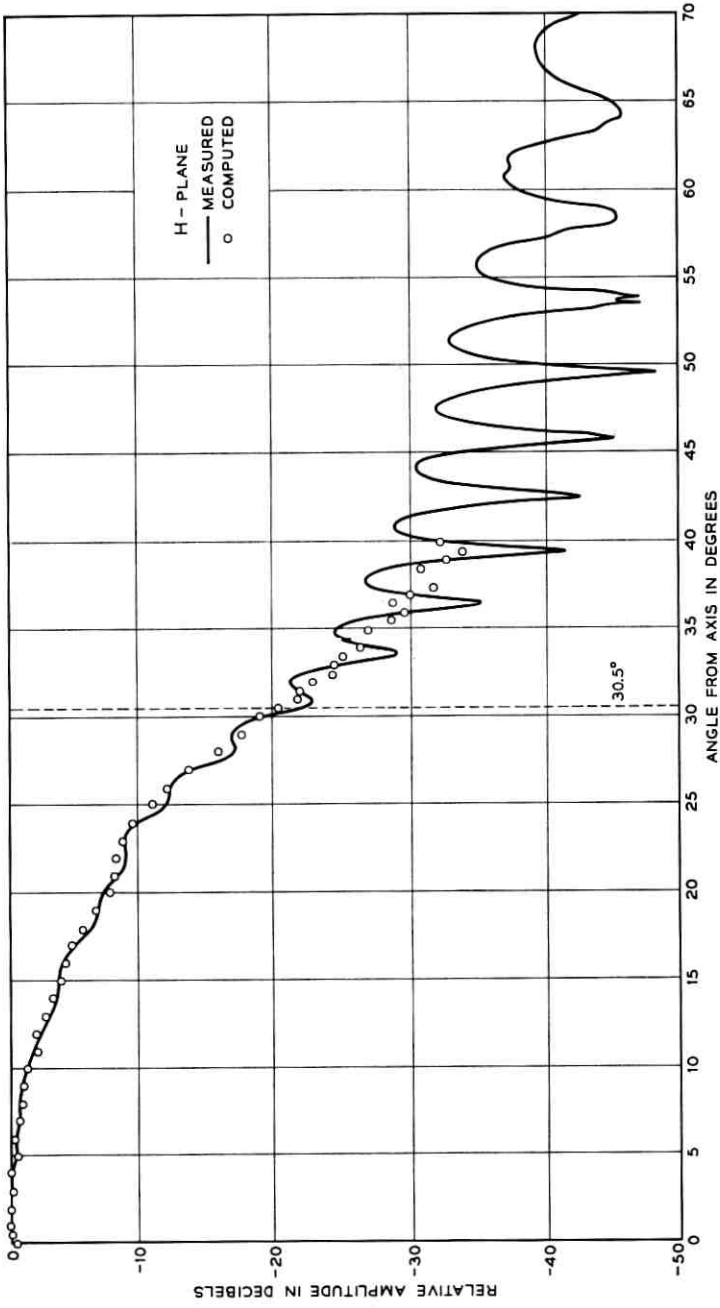

Fig. 26 — H-plane radiation pattern of subreflector.

5.6 *Comparison Between the Computed and Measured Radiation Patterns*

Very good agreement has been obtained between the computed and measured patterns of the feed horn. The slight asymmetry in the measured patterns has been demonstrated to be caused by eccentricity of the mode converter.

A comparison between the computed and a measured subreflector pattern is shown in Figs. 25 and 26. Good agreement has been obtained in the geometrical illumination region. In the shadow region the amplitudes and periodicity of the measured side lobes deviate somewhat from the calculated ones. Possible causes for the deviations are:

(1.) The computation is based on an approximate current distribution and does not include edge effects.

(2.) Higher order modes propagate in the feed horn.

(To determine the effect of the interaction of the feed horn with the subreflector, radiation patterns were measured by displacing the subreflector a quarter wavelength from its nominal position. Small changes were observed in the radiation patterns primarily in the geometrical illumination region. The patterns also showed a cyclic dependence on the displacement of the subreflector by a half wavelength, indicating that the changes were due to reflections from the horn.)

The noise temperature due to the spillover at the main reflector has been determined by integration of the measured subreflector radiation patterns. For an illumination angle of 30.5° the noise temperature is 3.3°K. This value is in good agreement with the computed noise temperatures shown in Fig. 13.

The measured main beam portions of the antenna patterns are in agreement with the computed patterns of the main beam. The measured side lobe patterns are in reasonable agreement with the computation in the E plane. In the H plane the pattern is non-symmetric. Asymmetry in this plane can only be caused by a field with non-uniform phase in the projected antenna aperture. Possible causes are:

(1.) Deformation of the main reflector during handling.

(2.) Slight mechanical misadjustments.

(3.) Small reflection from the subreflector support and from the protruding feed horn.

VI. SUMMARY

We have built an open cassegrain antenna which has a computed efficiency exceeding 65 per cent (including spillover and scattering losses but not including ohmic loss), a calculated noise temperature less than

4°K (when the antenna is appropriately mounted and assuming earth radiation of 300°K, but not including ohmic loss), and a very acceptable near sidelobe structure.

It is perhaps more significant that analytical and experimental tools have been developed which make possible the design of an open cassegrain antenna with predictable performance. These tools, the radiation analyses and mode coupler characteristics, taken together with a conceptual understanding of the geometrical tradeoffs that can be made, permit the development of antennas to meet a variety of specific requirements in a near optimum way.

ACKNOWLEDGMENTS

Following the TELSTAR® satellite experiments, people in diverse departments of Bell Telephone Laboratories studied many antenna configurations from as many different viewpoints. The evaluation of the open cassegrain was influenced by those people, as well as by work done elsewhere through the industry; but most directly by F. T. Geyling who was active in the study program and was party to the conversation wherein the concept crystallized.

The authors gratefully acknowledge the efforts of E. R. Nagelberg and J. Shefer who designed and evaluated the mode converter for the 60-gcs antenna; K. L. Warthman who aided in the design and construction of the antenna test mount; and L. H. Hendler who assembled the antenna test equipment, and assisted in the antenna measurements. Miss G. Fischbein and H. W. Lydiksen programmed the computation of the radiation pattern from the antenna and its components.

APPENDIX A

Geometry and Location of the Main Reflector Boundary

In assembling the antenna it is necessary to precisely position the subreflector with respect to the main reflector. It has been found convenient to position it with respect to the major and minor axes of the main reflector boundary ellipse.

This ellipse is a particular ellipse of the family of ellipses generated by the intersection of the paraboloid surface with planes perpendicular to the x_p, z_p plane (Fig. 3). The equation for the ellipses in the x_0, y plane are:

$$\cos^2 \theta_{p0} \left[x_0 - \tan \theta_{p0} \left(\frac{2f}{\cos \theta_{p0}} \right) \right]^2 + y^2 = \frac{4f\,(f - r_0 \cos \theta_{p0})}{\cos^2 \theta_{p0}} \qquad (50)$$

The curves of intersection of cones $\theta = \theta_c$ (constant) (Fig. 4) and the paraboloid surface are also ellipses. These ellipses coincide with the ellipses (50) and the following relations hold:

$$r_0 = \frac{2f \cos \theta_c}{\sqrt{1 + \cos^2 \theta_c + 2 \cos \theta_c \cos \theta_0}} \tag{51}$$

and

$$\tan \theta_{p0} = \frac{\sin \theta_0}{\cos \theta_c + \cos \theta_0}. \tag{52}$$

The major and minor dimensions of the ellipses in terms of θ_c and θ_0 are:

$$\rho_{\text{major}} = \frac{2f \sin \theta_c \sqrt{1 + \cos^2 \theta_c + 2 \cos \theta_c \cos \theta_0}}{(\cos \theta_c + \cos \theta_0)^2} \tag{53}$$

and

$$\rho_{\text{minor}} = 2f \frac{\sin \theta_c}{\cos \theta_c + \cos \theta_0}. \tag{54}$$

The radial distances from the origin, or focus, to the termini of the major and minor axes of the ellipses are:

$$r_{1,2\text{major}} = \frac{2f}{1 + \cos (\theta_0 \pm \theta_c)} \tag{55}$$

and

$$r_{\text{minor}} = \frac{2f (1 + \cos \theta_c \cos \theta_0)}{(\cos \theta_c + \cos \theta_0)^2}. \tag{56}$$

The φ coordinates of r_{minor} may be expressed:

$$\cos \varphi_{\text{minor}} = \frac{\sin \theta_c \sin \theta_0}{1 + \cos \theta_c \cos \theta_0}. \tag{57}$$

The φ coordinates of $r_{1,2\text{ major}}$ are 0 and π, respectively.

APPENDIX B

The Symmetries of the Radiation Fields of a Horn Excited with TE_{11} and TM_{11} Modes

Consider the function,

$$F = \frac{e^{-jkR}}{R} (1 + 1_n \cdot 1_R). \tag{58}$$

In the two coordinate systems, referring to Fig. 8, R and $1_n \cdot 1_R$ are: In the θ_1', φ_1' coordinate system,

$$R = \sqrt{(r_1')^2 + l^2 - 2r_1' l \cos \gamma'}, \tag{59}$$

with

$$\cos \gamma' = \sin \theta'_1 \sin \theta' \cos (\varphi_1' - \varphi') + \cos \theta_1' \cos \theta' \tag{60}$$

and

$$1_n \cdot 1_R = \frac{r_1' \cos \gamma' - l}{R}. \tag{61}$$

In the θ_1, φ_1 coordinate system,

$$R = \sqrt{r_1^2 + l^2 + p^2 + 2r_1 p \cos \theta_1 - 2lp \cos \theta' - 2r_1 l \cos \gamma_1}, \tag{62}$$

with

$$\cos \gamma_1 = \sin \theta_1 \sin \theta' \cos (\varphi' - \varphi_1) + \cos \theta_1 \cos \theta' \tag{63}$$

and

$$1_n \cdot 1_R = \frac{p \cos \theta' - l + r_1 \cos \gamma_1}{R}. \tag{64}$$

Function F is a periodic function of the variable $(\varphi' - \varphi_1)$ and furthermore

$$F(\varphi' - \varphi_1) = F(\varphi_1 - \varphi') \tag{65}$$

in both coordinate systems. The Fourier series expansion of F is

$$F = \sum_{n=0}^{\infty} C_n(\theta', \theta_1) \cos n (\varphi' - \varphi_1). \tag{66}$$

The rectangular components of TE_{11} and TM_{11} modes in a circular waveguide for x and y polarization are:[6]

$$\begin{aligned}(\bar{E}_{a\mathrm{TE}})_x = A_x\{1_x[J_0(k_{\mathrm{TE}}\rho) + J_2(k_{\mathrm{TE}}\rho) \cos 2\varphi'] \\ + 1_y J_2(k_{\mathrm{TE}}\rho) \sin 2\varphi'\}\end{aligned} \tag{67}$$

$$\begin{aligned}(\bar{E}_{a\mathrm{TE}})_y = A_y\{1_y[J_0(k_{\mathrm{TE}}\rho) - J_2(k_{\mathrm{TE}}\rho) \cos 2\varphi'] \\ + 1_x J_2(k_{\mathrm{TE}}\rho) \sin 2\varphi'\}\end{aligned} \tag{68}$$

with

$$J_1'(k_{\mathrm{TE}}a) = 0 \tag{69}$$

and

$$(\bar{E}_{a\mathrm{TM}})_x = B_x\{1_x[J_0(k_{\mathrm{TM}}\rho) - J_2(k_{\mathrm{TM}}\rho)\cos 2\varphi'] - 1_y J_2(k_{\mathrm{TM}}\rho)\sin 2\varphi'\} \tag{70}$$

$$(\bar{E}_{a\mathrm{TM}})_y = B_y\{1_y[J_0(k_{\mathrm{TM}}\rho) + J_2(k_{\mathrm{TM}}\rho)\cos 2\varphi'] - 1_x J_2(k_{\mathrm{TM}}\rho)\sin 2\varphi'\} \tag{71}$$

with

$$J_1(k_{\mathrm{TM}}a) = 0. \tag{72}$$

The subscripts x, y indicate x and y polarization respectively.

$J_n(_N)$ = Bessel function of order n

a = Radius of circular waveguide

For a narrow angle horn a linear relationship between δ and θ' may be assumed[7] given by

$$\rho = \alpha(\theta'/\alpha) \qquad \alpha = \text{horn angle.} \tag{73}$$

The y component of the first term of (16) is

$$(E_{py})_y = \frac{jkA_y l^2}{4\pi}\int_0^{\alpha}\int_0^{2\pi}\left[(E_{ay\mathrm{TE}})_y + \frac{B_y}{A_y}(E_{ay\mathrm{TM}})_y\right]\frac{e^{-jkR}}{R}(1 + 1_n\cdot 1_R)\sin\theta'\,d\theta'\,d\varphi' \tag{74}$$

where $(E_{ay\mathrm{TE}})_y$ and $(E_{ay\mathrm{TM}})_y$ are the y components of (68) and (71) respectively. Using (66)

$$(E_{py})_y = D_{y_0}(\theta_1) - D_{y_2}(\theta_1)\cos 2\varphi_1 \tag{75}$$

with

$$D_0(\theta_1) = \frac{jkA_y l^2}{2}\int_0^{\alpha} C_0(\theta',\theta_1)\left[J_0\left(k_{\mathrm{TE}}a\frac{\theta'}{\alpha}\right) + \frac{B_y}{A_0}J_0\left(k_{\mathrm{TM}}a\frac{\theta'}{\alpha}\right)\right]\sin\theta'\,d\theta' \tag{76}$$

$$D_2 = \frac{jkA_y l^2}{4\pi}\int_0^{\alpha} C_2(\theta',\theta_1)\left[\frac{B_y}{A_y}J_2\left(k_{\mathrm{TM}}a\frac{\theta'}{\alpha}\right) - J_2\left(k_{\mathrm{TE}}a\frac{\theta'}{\alpha}\right)\right]\sin\theta'\,d\theta'. \tag{77}$$

In the principal planes $\varphi_1 = 0$ and $\varphi_1 = \pi/2$

$$[E_{py}(0)]_y = D_0(\theta_1) - D_2(\theta_1) \tag{78}$$

and

$$[E_{py}(\pi/2)]_y = D_0(\theta_1) + D_2(\theta_1). \tag{79}$$

Due to the similarity of the aperture fields, by analogy

$$(E_{px})_y = D_2(\theta_1) \sin 2\varphi_1 \tag{80}$$

$$(E_{py})_x = D_2(\theta_1) \sin 2\varphi_1 \tag{81}$$

$$(E_{px})_x = D_0(\theta_1) + D_2(\theta_1) \cos 2\varphi_1 . \tag{82}$$

It is reasonable to assume that the components of radiated electric field have only components perpendicular to the radial direction. The nearly spherical wave front obtained in the coordinate system with its origin at the phase center partially justifies this assumption. In analogy to the electric field radiated from an open waveguide[6] it is assumed that

$$E_\theta = E_{px} \cos \varphi_1 + E_{py} \sin \varphi_1 \tag{83}$$

and

$$E_\varphi = E_{py} \cos \varphi_1 - E_{px} \sin \varphi_1 . \tag{84}$$

Using (75) through (84) the fields for both polarizations are (19) and (20).

Due to the φ' symmetries of the y component of the electric field it is sufficient to integrate (74) from 0 to π for $\varphi_1 = 0$ and from $-(\pi/2)$ to $(\pi/2)$ for $\varphi_1 = \pi/2$.

APPENDIX C

Construction of a Reflector Surface for the Conversion of an Equiphase Surface to a Spherical Surface

The theorem by Malus is derived for the case where the incident equiphase surface, S_i, is arbitrary and the reflected equiphase surface, S_s, is spherical. Based on the derivation, a method is presented for the construction of the reflector surface, S_r.

The surface S_i is assumed to be rotationally symmetric and consequently the surface S_r, and S_s are rotationally symmetric. The coordinates for the surfaces are shown in Fig. 27. The unit normal for S_i is 1_{r_0}.

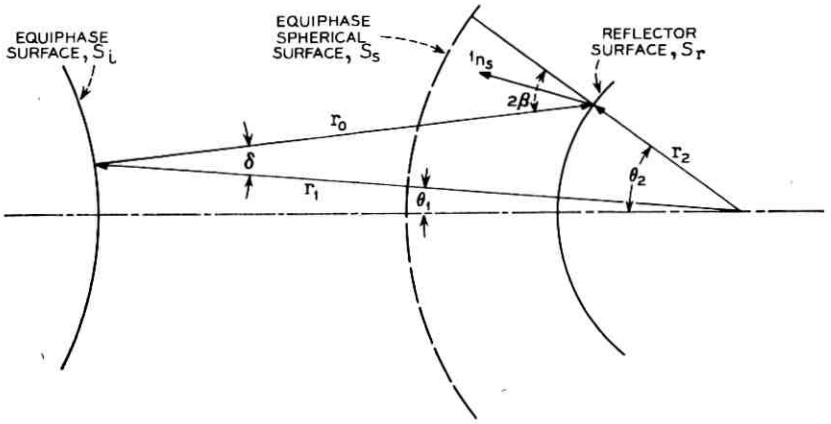

Fig. 27 — Equiphase and reflector surface coordinates.

From the geometry,

$$1_{r_1} r_1 - 1_{r_2} r_2 = -1_{r_0} r_0 . \tag{85}$$

To obtain a spherical equiphase surface it is necessary that:

$$r_0 = r_2 + a \tag{86}$$

where a is an arbitrary constant. It follows from (85) and (86) that

$$a^2 + 2r_2[r_2 \cos(\theta_2 - \theta_1) + a] - r_1^2 = 0. \tag{87}$$

The unit normal 1_{n_s} to the surface S_r is:

$$1_{n_s} = \frac{1_{r_2} - 1_{\theta_2} \frac{1}{r_2} \frac{\partial r_2}{\partial \theta_2}}{\sqrt{1 + \left(\frac{1}{r_2} \frac{\partial r_2}{\partial \theta_2}\right)^2}} \tag{88}$$

where 1_{r_2} and 1_{θ_2} are spherical vectors.

Snell's law for the surface S_r is:

$$1_{n_s} \cdot 1_{r_2} = -1_{n_s} \cdot 1_{r_0} . \tag{89}$$

From (88) and (89)

$$1_{r_2} \cdot 1_{r_0} - 1_{\theta_2} \cdot 1_{r_0} \frac{1}{r_2} \frac{\partial r_2}{\partial \theta_2} = -1. \tag{90}$$

From the geometry

$$1_{r_2} \cdot 1_{r_0} = -\cos 2\beta \tag{91}$$

and

$$1_{r_2} \cdot 1_{r_2} = \sin 2\beta. \tag{92}$$

With (91) and (92), (90) reduces to

$$\frac{1}{r_2}\frac{\partial r_2}{\partial \theta_2} = \tan \beta. \tag{93}$$

Also from the geometry and (87)

$$\tan \beta = \frac{r_1 \sin (\theta_2 - \theta_1)}{r_1 \cos (\theta_2 - \theta_1) + a}. \tag{94}$$

It remains to be shown (87) and (93) are satisfied simultaneously, this is done by differentiating (87) with respect to θ_2 and considering that

$$r_2(\theta_1) = r_2[\theta_1(\theta_2)]. \tag{95}$$

Differentiating (87) results:

$$\frac{1}{r_2}\frac{\partial r_2}{\partial \theta_2} = \frac{r_1 \sin (\theta_2 - \theta_1)}{r_1 \cos (\theta_2 - \theta_1) + a} + \frac{\left\{r_2 r_1 \sin (\theta_2 - \theta_1) + [r_2 \cos (\theta_2 - \theta_1) - r_1]\frac{\partial r_1}{\partial \theta_1}\right\}\frac{\partial \theta_1}{\partial \theta_2}}{r_2 [r_1 \cos (\theta_2 - \theta_1) + a]}. \tag{96}$$

Comparing (96) with (93) and (94) it follows that (87) and (93) are satisfied simultaneously provided that

$$\frac{1}{r_1}\frac{\partial r_1}{\partial \theta_1} = \frac{r_2 \sin (\theta_2 - \theta_1)}{r_1 - r_2 \cos (\theta_2 - \theta_1)}. \tag{97}$$

Since 1_{r_0} is a unit normal to the surface S_i by analogy to (93)

$$\frac{1}{r_1}\frac{\partial r_1}{\partial \theta_1} = \tan \delta = \frac{r_2 \sin (\theta_2 - \theta_1)}{r_1 - r_2 \cos (\theta_2 - \theta_1)}. \tag{98}$$

The latter equality follows the geometry. Therefore (87) and (93) are satisfied simultaneously.

The reflecting surface can be constructed as follows. Equation (86) is an equation for a hyperboloid with r_1 (the distance between the foci) and a as the defining parameters. The intersection of the hyperboloid with the line in the direction r_0, determines a point of the reflecting surface.

REFERENCES

1. Potter, P. D., A New Horn Antenna with Suppressed Side-lobes and Equal Beamwidths, Microwave J., *VI*, June 1963, pp. 71–78.
2. Nagelberg, E. R., and Shefer, J., Mode Conversion in Circular Waveguides, B.S.T.J., this issue, pp. 1321–1338.
3. Hogg, D. C., and Semplak, R. A., An Experimental Study of Near-Field Cassegrainian Antennas, B.S.T.J., *43*, Nov. 1964, p. 2677.
4. Hogg, D. C., Effective Antenna Temperatures due to Oxygen and Water Vapor
5. Hines, J. N., Li, T., and Turrin, R. H., The Electrical Characteristics of the in the Atmosphere, JAP, *30*, 1959, p. 1417.
   Horn-Reflector Antenna, B.S.T.J., *42*, July 1963, pp. 1187–1211.
6. Silver, S., *Microwave Antenna Theory and Design*, McGraw-Hill Book Co. Inc., New York, N. Y., 1949.
7. Li, T., and Turrin, R. H., Near Zone Field of the Conical Horn, IEEE Trans. *AP-12*, No. 6, Nov. 1964, pp. 800–802.
8. Bauer, K., The Phase Center of Aperture Rad, Arch. Elek, Ubertragung, *9*, 1955, p. 541.
9. Nagelberg, E., Fresnal Region Phase Centers of Circular Aperture Antennas, IEEE Trans. *AP-13*, May 1965, p. 479.

The Open Cassegrain Antenna: Part II. Structural and Mechanical Evaluation

By W. J. DENKMANN, F. T. GEYLING, D. L. POPE and A. O. SCHWARZ

(Manuscript received May 14, 1965)

The mechanical features of a preliminary concept for an open cassegrain antenna are discussed briefly. In the analysis, emphasis is given to the upper rotating structure, where the major problems are the provision for an efficient back-up structure for the main reflector and the selection of a suitable subreflector support structure. The philosophy and method of approach are described in detail. Representative deflection results are given for both gravity and wind loading. Other mechanical considerations pertinent to this configuration are discussed in general. The structural implications of exposed operation, in particular those due to wind, are considered at some length. The mechanical feasibility of this configuration is demonstrated by the current results.

I. INTRODUCTION

Various mechanical features of the present concept for an open cassegrain antenna are discussed below. The present concept is based on preliminary analysis of some of the problems posed by the unusual geometry and expected applications of the structure. These problems include the inherent asymmetry of the configuration, the need for rigid "external" supports for the subreflector, the design of the slant bearing and slant axis drives, and the desire to terminate the feed horn adjacent to the azimuth axis. The last requirement permits the horn to connect to the stationary communication equipment through a very short length of circular waveguide and rotary joint concentric with the azimuth axis.

An aperture diameter of 56′ was somewhat arbitrarily selected for this design study. An open cassegrain of that size would meet typical requirements for major satellite communication system earth stations.

Fig. 1 is a line drawing of the basic structural configuration that

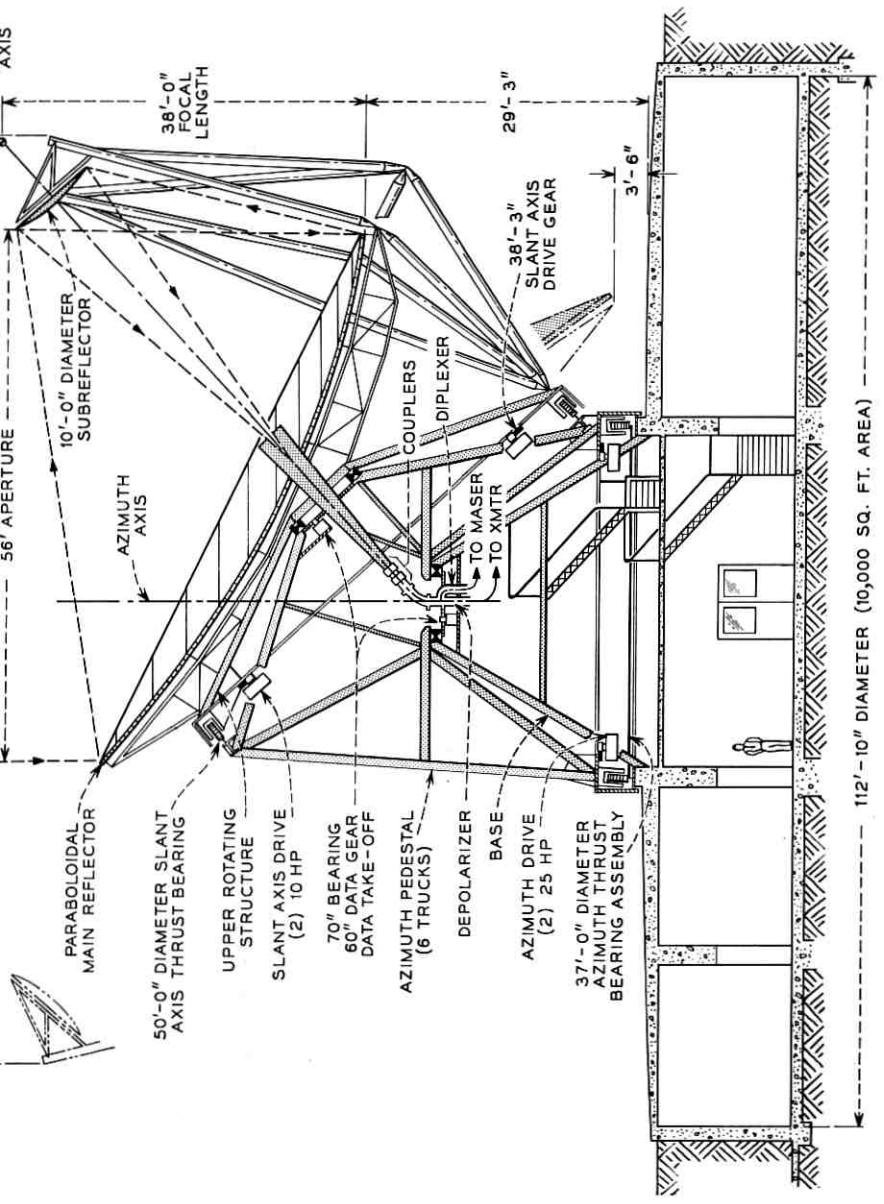

Fig. 1 — Basic layout for the open cassegrain antenna.

emerged from the study. Many of the terms used in the discussion of this configuration are defined in this figure. Fig. 10 further clarifies the geometrical relationship of the structural components and also defines two orthogonal coordinate systems that are referred to in the text.

Dimensions associated with this configuration are:

Aperture Diameter	56 feet
Aperture Area	2460 square feet
Focal Length of Paraboloid	38 feet
Subreflector Diameter	10 feet
Slant Bearing Diameter	50 feet
Azimuth Bearing Diameter	37 feet
Height of Structure	68 feet
Swept Diameter of Structure	96 feet

The slant axis makes an angle of 47.5° with the vertical. This value was chosen to place the minimum elevation 5° below the horizon, which permits easy tracking of the mount at small positive elevation angles. At minimum excursion azimuth and slant-elevation motion are redundant.[1]

Figs. 2, 3, and 4 are photographs of a scale model constructed to aid in the visualization of this concept. This model was also useful in providing a gross understanding of the structural problems and mass distribution. It was discovered, for instance, that rotational stability of the subreflector support structure was far more difficult to achieve than translational stability. The comparatively complex nine-member support grew out of that discovery.

The initial estimates of structural performance were obtained by assuming structural members similar to those used in the horn-reflector antenna at Andover, Maine. The weight of the upper rotating structure is estimated to be 75 tons. The azimuth structure has a weight of approximately 80 tons for a total rotating weight of 155 tons. The polar moment of inertia of the upper rotating structure about the slant axis (I_{zz}) is estimated to be 2.4×10^6 slug-ft^2. The product of inertia of the upper rotating structure (I_{xy}) is approximately 5.2×10^5 slug-ft^2. The polar moment of inertia of the azimuth pedestal alone about the azimuth axis ($I_{z'z'}$) is estimated to be 1.5×10^6 slug-ft^2. Compliances relating the input torque about each axis to the angular response of the rotated structure were also estimated. For the rotation of the upper rotating structure about the slant axis, this figure is 1.5×10^{-10} rad/ft-lb. For the rotation of the total structure about the azimuth axis, the compliance was determined to be 1.9×10^{-10} rad/ft-lb. The

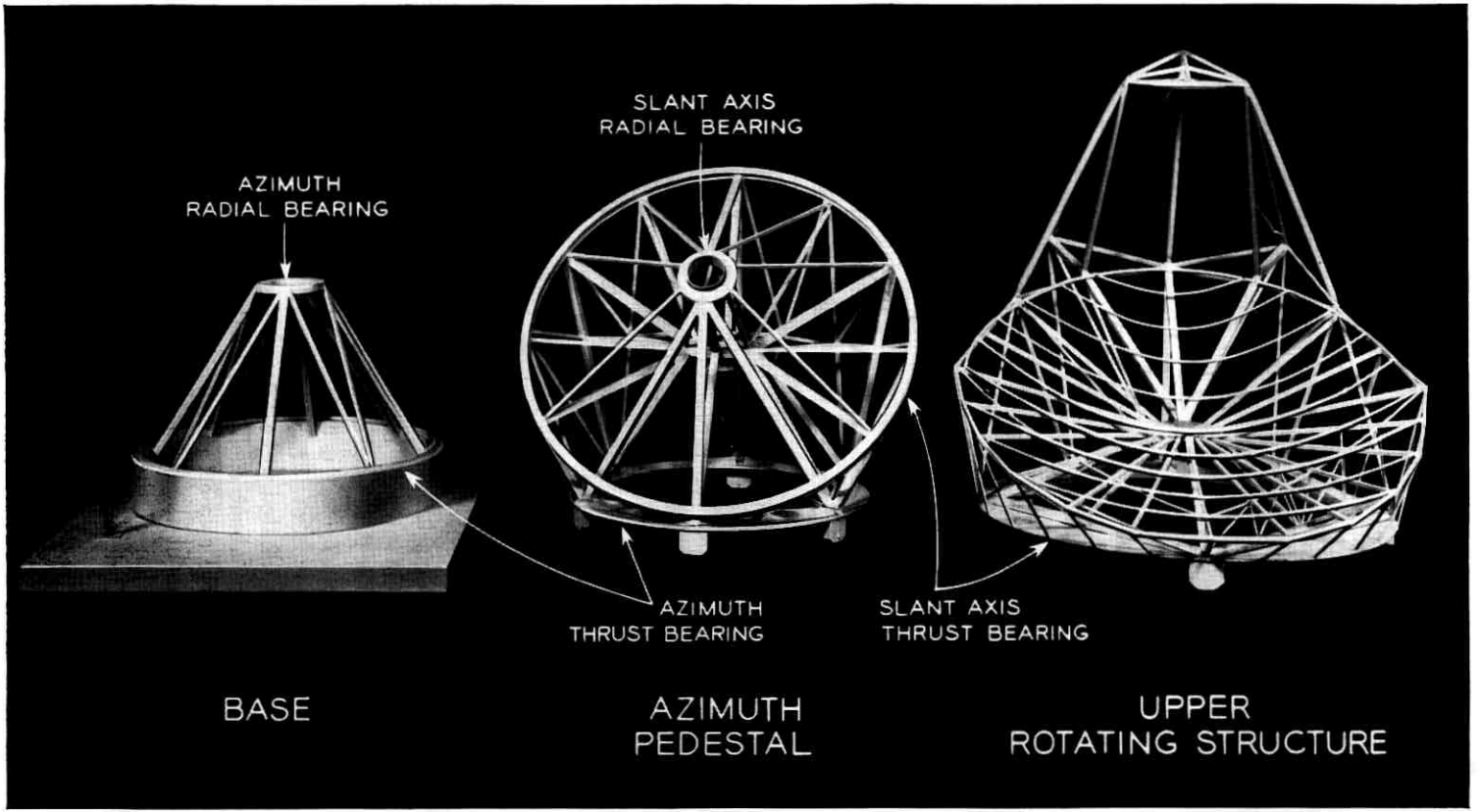

Fig. 2 — Structural components for the open cassegrain antenna.

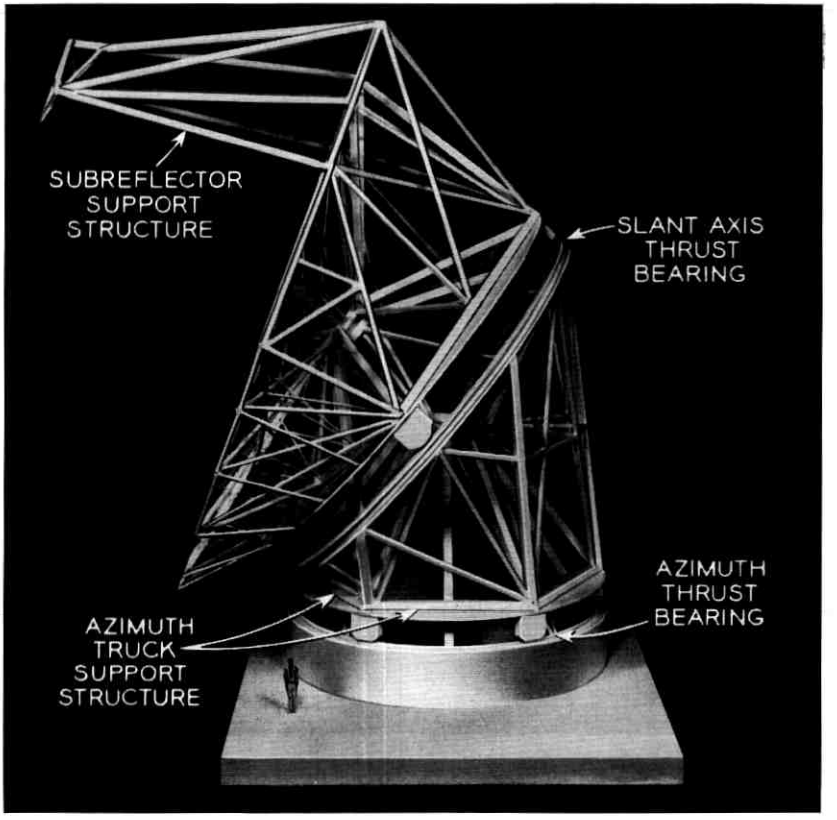

Fig. 3 — Structural components for the open cassegrain antenna assembled.

values of inertias and compliances cited here are necessary for the antenna control system analysis described elsewhere.[1]

II. STRUCTURAL DESIGN

2.1 *Structural Philosophy and Approach*

The open cassegrain configuration has a number of structurally appealing features. For example, the large slant axis bearing is located reasonably close to the reflecting surface. The reflector back-up structure can thus be designed to provide adequate rigidity without excessive weight. The azimuth structure is also compact and lends itself to inherently rigid conventional construction techniques. Provision of

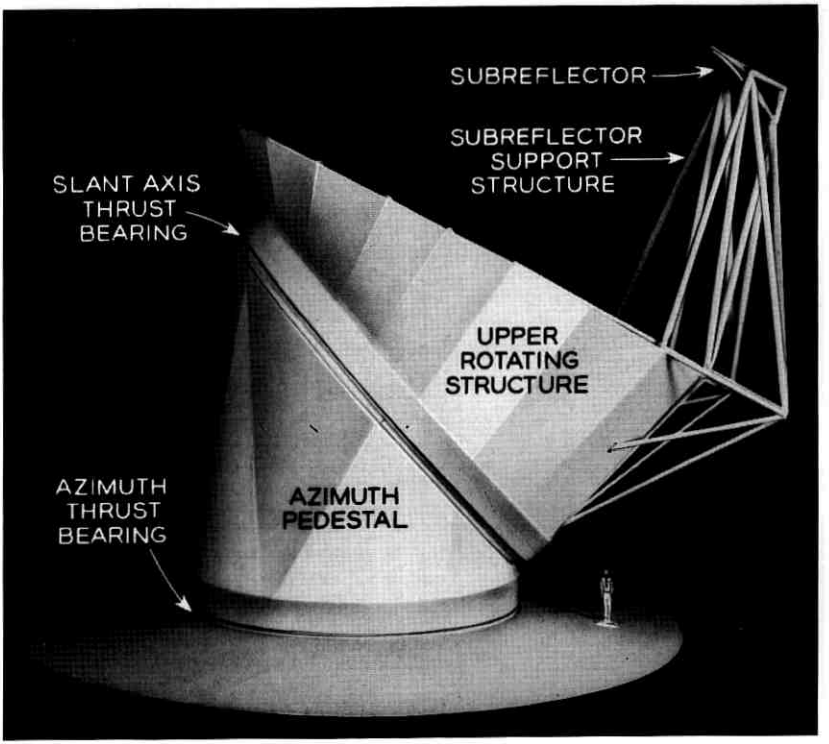

Fig. 4 — Final appearance of the open cassegrain antenna.

radial bearings well above the planes of the corresponding thrust bearings for both axes as shown in Figs. 2 and 5 increases the structure's ability to resist environmental loads. The favorable distribution of resisting forces obtainable with this bearing arrangement accounts for this improvement.

The main structural members of the azimuth pedestal are assumed to be built-up box beams, approximately one-foot square. These are stiffened and interconnected by lighter members. The azimuth pedestal is inherently rigid due to its robust construction and internal structure.

The 37-foot diameter azimuth bearing track supports the weight of the upper rotating structure as directly as possible at the azimuth trucks in an effort to avoid bending loads in this portion of the structure. The azimuth truck support is hexagonal, and is stiffened by a structural ring in the same manner as the slant axis truck support discussed below.

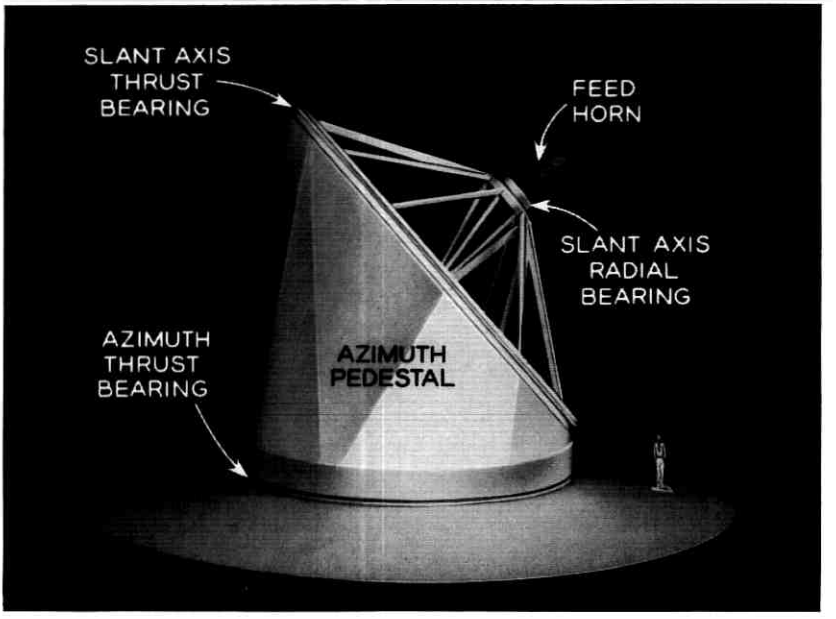

Fig. 5 — Azimuth pedestal of the open cassegrain antenna illustrating elevated radial bearing.

Fig. 3 shows the arrangement of these elements. The structural analysis of the azimuth pedestal, while important, is conventional to the extent that it can be safely assumed that the design can be accomplished in a routine manner at the appropriate time.

The basic structural element in the upper rotating structure is a square frame with bearing trucks at each corner, stiffened by an annular ring member (Fig. 6). The annular ring is shown as a solid section in the model. In practice, the ring would probably be built up in an appropriate manner to provide ample stiffness and backing for bull-gear segments. The slant axis bearing diameter of 50 feet was selected to provide good outboard support for the main reflector.

The reflector back-up structure is assumed to be a pin-jointed space truss for the purposes of analysis. The structural members are thin-walled steel tubes, two to six inches in diameter as the application dictates. The assumption that the joints are ideal, while not realistic, is both conservative and conventional at this stage of the analysis. The curved members crossing the main reflector surface, seen in Figs. 2 and 6, provide points of support for the reflector panels. These members

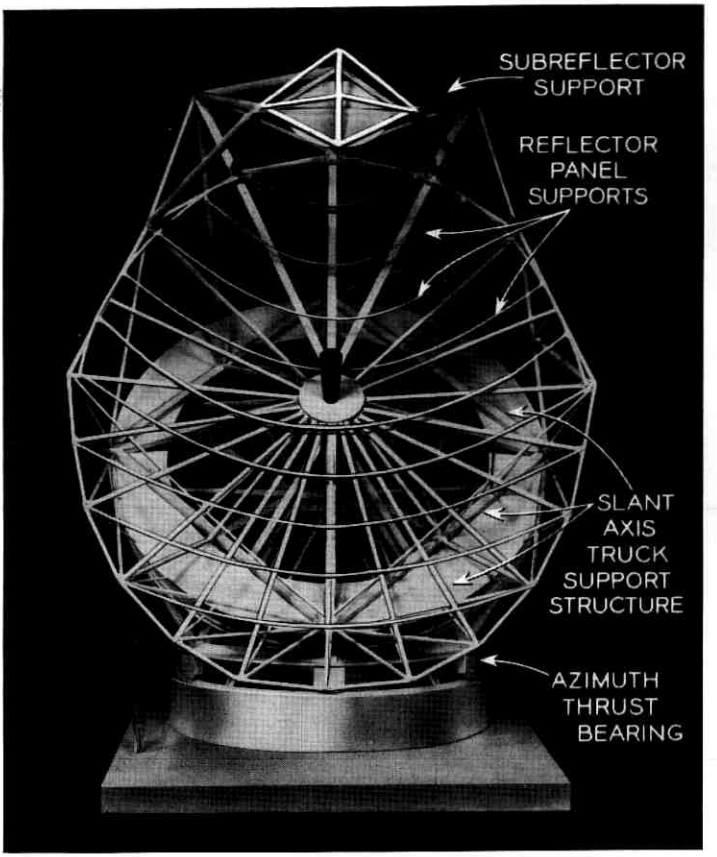

Fig. 6 — Assembled structure of the open cassegrain antenna illustrating slant axis truck support structure and reflector panel supports.

lie along lines of intersection of the reflector surface when the antenna points at zenith and appropriately spaced horizontal planes.

The two major structural problems faced in the configuration are the provision of a sufficiently rigid base for the subreflector support structure, and the design of an efficient back-up structure for the main reflector. These problems were considered early in the study, and a preliminary analysis was made on a variety of suggested structural arrangements. The results of these preliminary analyses were of considerable value for the selection of the structural configuration under discussion.

It is emphasized that the structural configuration, as represented by

the illustrations of the model, is a functional rather than a detailed concept. The described results simply provide a basis for the detailed structural analysis which must follow in the development of a design for construction, and are only indicative of the expected performance of the final structure.

2.2 *Structural Analysis*

Fig. 2 indicates how the structure may be thought of in terms of components. The base, attached to the foundation, provides support for the azimuth bearings and equipment decks. The lower rotating structure, or azimuth pedestal, is composed of two rings, two truncated cones, internal diaphragms, and an enveloping conoidal surface. The upper rotating structure is more complicated, as Fig. 2 shows. The base and the azimuth pedestal are sufficiently simple and straightforward that no unusual problems are anticipated in their design. Hence, subsequent consideration will be given only to the upper rotating structure. For purposes of discussion, this structure is assumed to have rigid support in the plane of the slant axis bearing.

The upper rotating structure is assumed to be composed of a series of radial support trusses extending from the inner cone supporting the slant axis radial bearing outward to the periphery of the main reflector surface. This type of internal stiffening for the main reflector surface is similar to that often used for symmetrical antennas, and proves to be quite efficient in spite of the lack of symmetry of this design. The provision of an adequate support for the subreflector structure can also be incorporated into this arrangement with a minimum of additional complication.

The interaction of the subreflector support structure and main reflector back-up structure has considerable significance for the open cassegrain antenna. Since the subreflector support structure is considerably heavier than a conventional cassegrain support, and since it is located on the edge of the back-up structure rather than centrally, the provision of sufficiently "hard" support points for this structure, is a difficult problem. The elastic behavior of the main reflector back-up structure must be considered in determining subreflector deflections for a given loading. In fact, the deflections of the points of attachment of the subreflector support structure have greater influence on subreflector deflections than the compliance of the support structure itself. For example, preliminary analyses of the subreflector support structure have been carried out with the assumption of complete fixity at the points of attachment. When the compliance of the main reflector

back-up structure was included, the deflections had increased by an order of magnitude over those found in the previous analysis for the same loading. Hence, it was clear that the main reflector back-up structure and the subreflector support structure had to be analyzed as a unit. It was therefore necessary to expand the capacity of existing numerical routines prior to attempting a detailed analysis of the configuration shown.

The structural analysis was done using a general purpose computer program for the solution of three-dimensional trusses developed by one of the authors (W.J.D.). The program is based on the matrix displacement method,[2] suitably modified,[3] to provide sufficient capacity to analyze the reflector back-up structure and the subreflector support structure simultaneously.

Some members in the upper rotating structure carry loads primarily in bending. This is particularly true of the four beams connecting the truck support points. In order to analyze these members with the three-dimensional truss program, the beams were broken up into several hypothetical truss members, and appropriate springs inserted at the nodes to simulate the behavior of the beam. This procedure permits the incorporation of such members into the truss program and predicts their response with satisfactory accuracy.

The loads considered were the dead weight of the structure, including the main reflector panels and the subreflector. The gravity field can be resolved into a component normal to the slant bearing plane and a component tangential to this plane. The effect of the normal component is independent of the rotation about the slant axis and hence complete compensation for these deflections can be accomplished by the final alignment procedure. The changes of the deflections of the reflecting surfaces as the antenna rotates create the difficulties in maintaining surface accuracy at all antenna pointing angles. In this configuration, these changes are due entirely to the component of gravity tangential to the slant bearing plane. Since this component is only 70 per cent of the full gravity load, it should be possible to achieve a reduction in structural weight for the same surface tolerances budgeted to gravity, as compared with a conventional antenna in which the full gravity load influences these changes in deflection.

The deflections of the subreflector and of a number of points on the main reflector surface due to gravity were calculated for both the minimum elevation and the zenith position of the antenna. Knowledge of the deflections due to gravity in two different positions of the antenna is not sufficient to permit their calculation in any position by

simple superposition. The deflections must also be calculated for a third independent position before this procedure can be applied.

Surface deflections were also calculated for a hypothetical wind load calculated from the results of hydrodynamic model testing.[4] A reasonable pressure distribution was assumed to act over the surface of the main reflector. The free parameters of the pressure distribution were adjusted to match statically the torques and resultant forces measured in the laboratory. These pressures were then converted into equivalent static forces at the nodes of the structure, and the resulting deflections calculated. Only one wind loading position was considered, for the antenna pointed at the zenith. A steady 40-mph horizontal wind at an azimuth aspect angle of 60°, (see Fig. 10), was blowing into the concave side of the main reflector. The deflections for this case are plotted in Fig. 7, a contour map of the main reflecting surface as seen looking toward the reflector along the main beam. The

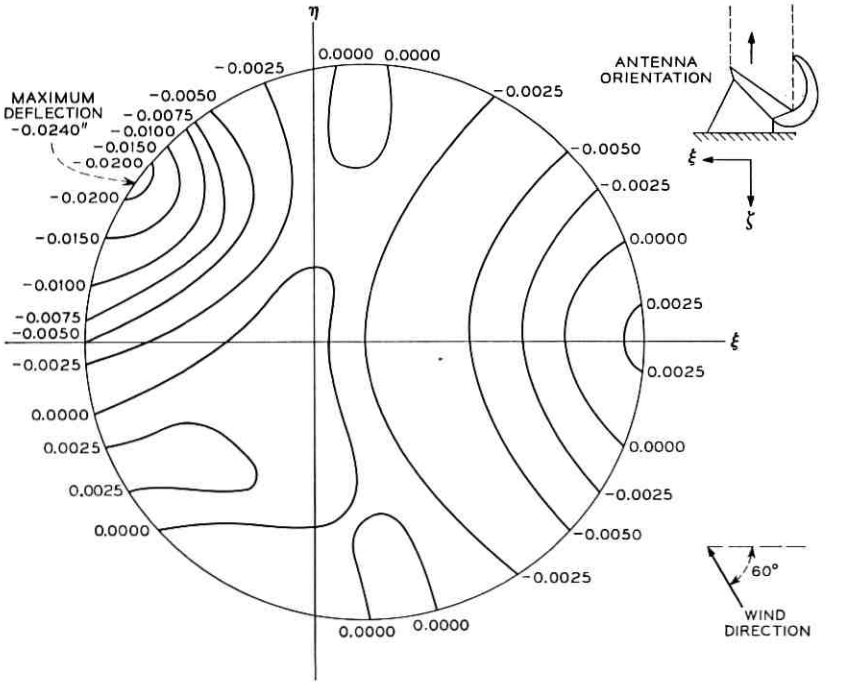

Fig. 7 — Deflections of the main reflector surface in the ζ direction due to a horizontal wind of 40 mph at a wind azimuth aspect angle of 60° (deflections in inches).

deflections shown are parallel to the ζ axis as shown in the figure. The normal and transverse deflections are available, but difficult to present graphically. Fig. 8 is a similar plot of the deflections parallel to the ζ axis due to gravity when the antenna points at zenith. Fig. 9 also shows deflections parallel to the ζ axis due to gravity, but for the antenna at its minimum elevation of $-5°$. The results of these studies reveal a "soft" spot on the periphery of the main reflector. This is immediately evident in Figs. 7, 8, and 9. In subsequent designs, appropriate steps would have to be taken to improve structural rigidity in this region.

Similar results can be superimposed to determine deflections due to combined effects. This would be a task of considerable magnitude. If all aspects of the wind loading are considered, there are four independent variables; the antenna pointing angle (two variables), the wind aspect angle, and the wind velocity. A preliminary study would probably be concerned only with some predetermined maximum operational wind velocity. If the investigation was further restricted to consideration of only the gross aspects of the deflection pattern, such as maximum deflection, or perhaps RMS deflection, it would be well within the scope of current capabilities. "Worst case" combinations of wind and gravity loading for subsequent design purposes could be obtained rapidly using computer techniques. An understanding of the interaction of such loadings would also be obtained.

The selection of an appropriate subreflector support was difficult. This component must be sufficiently rigid for all loading conditions,

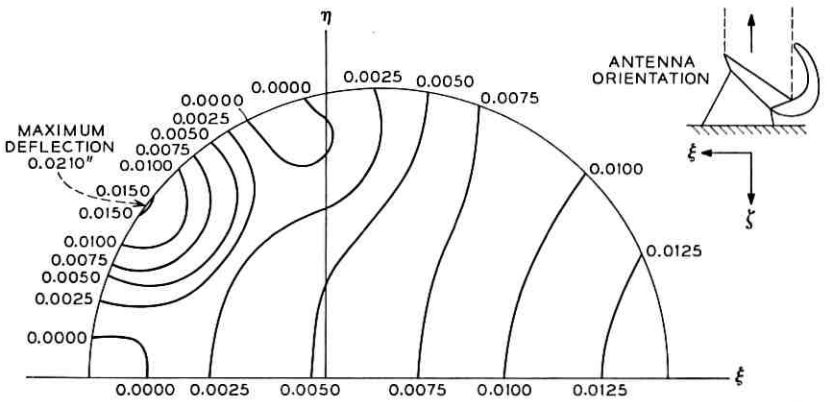

Fig. 8 — Deflections of the main reflector surface in the ζ direction due to gravity loading (deflections in inches).

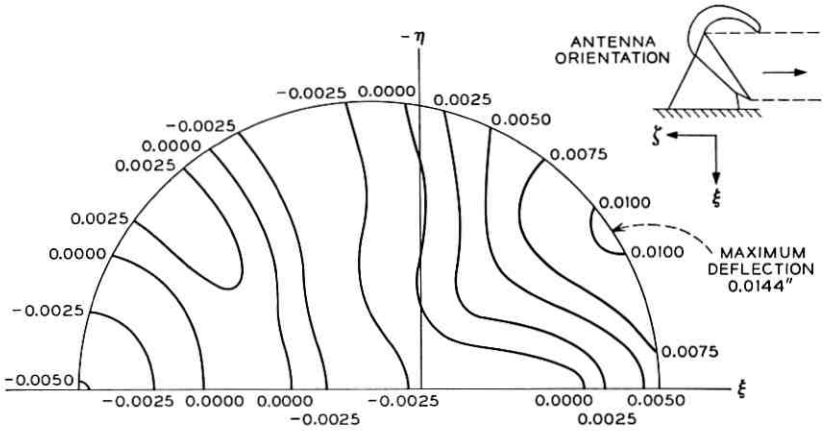

Fig. 9 — Deflections of the main reflector surface in the ζ direction due to gravity loading (deflections in inches).

but still be as light as possible due to its location in the structure. Many of the preliminary concepts studied provided adequate translational rigidity but were weak in torsion about an axis parallel to the axis of the antenna aperture. The structure shown in the figures overcomes this problem, but is considerably heavier than originally anticipated. It weighs about $6\frac{1}{2}$ tons, exclusive of the subreflector.

Regardless of the structural complications introduced by the need to support the subreflector at the edge of the main reflector and the unavoidable compliance of such a geometry, the absolute deflections of the subreflector caused by gravity and a steady wind are reasonably small. Representative values for these deflections in the zenith looking position due to gravity alone are shown in Table I. The coordinate system is the same as that shown in Fig. 8. The right hand screw rule determines the sense of the rotations. The wind deflections of the subreflector must also be considered. Moments about the base of the subreflector support structure induced by a steady 40-mph horizontal wind were also estimated. The wind direction relative to the antenna is the same as used above for the calculation of surface deflections. These moments were reduced to equivalent static loads by the procedure outlined earlier in connection with the main reflector surface. The absolute deflections of the subreflector due to this wind loading in the zenith looking position are also shown in Table I. These values include the effect of simultaneous wind loading on the main reflector surface.

TABLE I

DEFLECTIONS OF THE SUBREFLECTOR IN THE ZENITH LOOKING POSITION

Deflection or Rotation	Gravity	40 mph Steady Wind 60° Aspect Angle
Δ_ξ	$+3.9 \times 10^{-2}$ in.	$+2.0 \times 10^{-2}$ in.
Δ_η	0	$+1.1 \times 10^{-3}$ in.
Δ_ζ	$+2.3 \times 10^{-2}$ in.	$+9.0 \times 10^{-3}$ in.
θ_ξ	0	-1.99×10^{-5} rad.
θ_η	-2.7×10^{-6} rad.	$+1.20 \times 10^{-7}$ rad.
θ_ζ	0	$+1.00 \times 10^{-4}$ rad.

2.3 *Bearings and Azimuth Pedestal*

The slant axis bearing requirements are not significantly different than those for the azimuth bearing except that a smaller load is carried by the slant axis thrust bearing, and a component of the gravity load is continually applied to the slant-axis radial bearing. The wheel and track bearing is a relatively inexpensive and reliable low-friction device for large diameter applications and would be well suited for the open cassegrain antenna. Commercially available roller bearings are expected to meet the requirements for the radial bearing on each axis. The trucks of the slant axis bearing may be recessed into the back-up structure in order to keep the center of gravity of the upper structure as low as possible. Air-gap labyrinth seals are expected to provide environmental protection for both thrust bearings.

The structural function of the azimuth pedestal is to transmit the loads from the slant axis bearings to the azimuth axis bearings. As shown (Fig. 2), this can be done in a straightforward manner by means of a robust primary structure connecting the two bearing circles, together with an appropriate secondary stiffener system. The conical surfaces support the radial bearings in a natural way. Especially heavy members are used to carry the gravity component acting on the slant axis radial bearing directly to the azimuth trucks. The azimuth radial bearing is supported by a diaphragm which connects it laterally to the azimuth pedestal structure, and by the cone upon which it rests. The cone improves the structural integrity and increases the inherent rigidity of the azimuth pedestal.

2.4 *Reflector Alignment*

The main reflector surface may have to be adjusted to bring all points on its surface within acceptable tolerances. An optical align-

ment procedure is practical for this purpose. Such a procedure was used at Andover, Maine to adjust the reflector surface to within a one-sigma value of 0.060″. A relatively simple computer program reduced observed data and calculated the necessary adjustments at the points of support.

Alignment would require favorable weather conditions, and would probably be carried out at night to eliminate solar effects. Under severe conditions it may be necessary to provide a temporary shelter to protect the structure during alignment. An air-supported fabric structure might be considered for this application. Such a shelter could be relatively inexpensive, especially if amortized over a number of antenna installations.

III. ENVIRONMENTAL CONSIDERATIONS

3.1 *Wind Effects*

The effects of wind on an exposed antenna must be considered from the standpoint of:

(1.) Wind induced tracking error and loss of antenna gain under operational wind conditions.

(2.) Structural loading under operating and extreme or "survival" wind conditions.

(3.) Antenna overturning stability under survival wind conditions.

In most cases, the requirement for extreme structural rigidity in antenna design is the controlling factor and stress levels under the most severe operating conditions seldom become critical. An exception to this rule is the main reflector panels where the wind loading situation must be carefully considered. Section 2.2 presents the surface deflections due to a typical wind load distribution.

Wind induced deflections of the reflector surface and subreflector support structure produce pointing errors and a decrease in gain due to defocusing. These effects are of a random nature. Consequently, the structure must be designed to limit the gain reduction to a fraction of a db and the variation in pointing direction to a small fraction of the antenna beamwidth.

In addition to the pointing error due to structural deformation, an error is caused by the dynamic wind-induced torque about the antenna rotational axes. The magnitude of this torque is given by

$$T_w(t) = C_w V(t)^2$$

where $V(t)$ is the wind speed, a function of time

C_w is an experimentally determined wind torque coefficient in units of foot-pounds/(miles per hour)^2.

C_w has been experimentally determined for the open cassegrain antenna as a function of wind azimuth aspect angle and slant axis rotation angle by an extensive series of hydrodynamic tests on scale models of the antenna. Details of these tests are reported elsewhere.[4]

The minimum azimuth stall torque due to wind loading has been estimated to occur at a horizontal wind velocity of 71 mph. This situation would occur under the conditions of a slant axis angle of 45° and a wind aspect angle of 270°, as defined in Fig. 10. This estimate is based on the use of drive gear ratios which permit acceptable tracking rates for near-zenith missions, and the use of two 25-hp hydraulic azimuth drive units.

For the consideration of items (1.) and (2.) above, wind moment

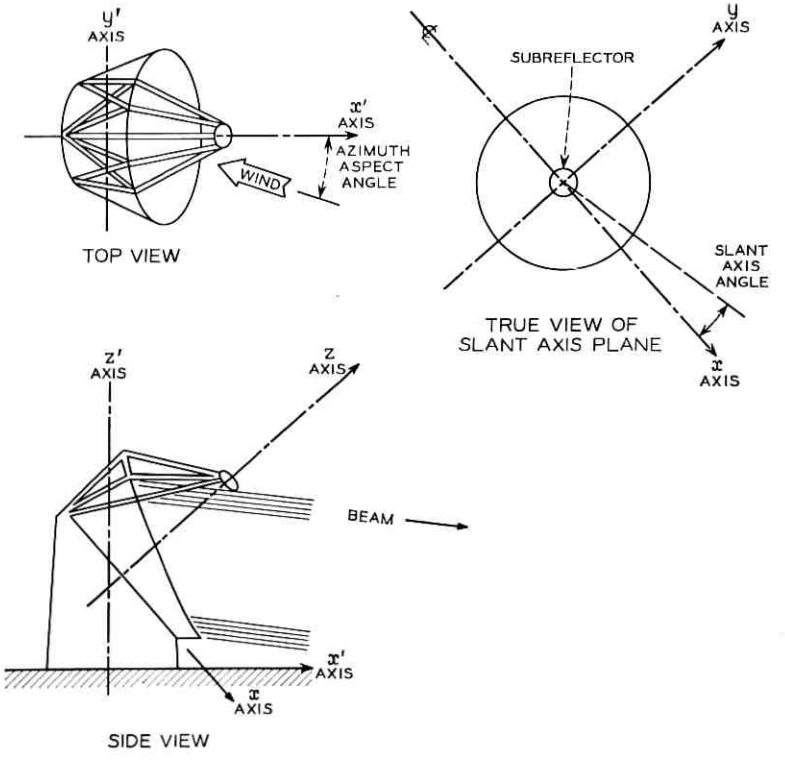

Fig. 10 — Orientation of axis of rotation (zero slant axis angle shown).

coefficients have been experimentally determined for pairs of axes in the planes of the slant and azimuth thrust bearings. The dynamic components of bearing and structural loading are calculated on the basis of these coefficients and the total drag coefficient. Table II exhibits maximum values of these coefficients obtained by hydrodynamic model tests from Ref. 4. These values indicate that the wind overturning moments at 100 mph are much less than the stability moments based upon the estimated weight and location of the center of gravity of the antenna. These conclusions support the contention that the open cassegrain is suitable for exposed operation.

The feasibility of friction drives for the two antenna axes has been investigated. Such drive systems would lead to reductions in cost and weight. However, such systems appear to be marginal in hypothetical worst case situations, with the possibility of slipping under certain circumstances. Hence, conventional ring-and-pinion gear drives are recommended for both axes. By mounting the slant-axis bull gear on the upper rotating structure, with the drive pinion on the azimuth pedestal, the necessity of carrying power beyond the slant axis bearing can be avoided.

3.2 *Thermal Effects*

A large structure exposed to full solar radiation may experience considerable distortion as portions of the structure are shadowed. If reflective paint finishes cannot be made to adequately minimize thermal deflections throughout the structure, especially in the long columns supporting the subreflector, it may be necessary to provide compensating mechanisms to maintain the required precision.

TABLE II

(ANGLES AS DEFINED IN FIG. 10)

Torque Axis	Maximum Wind Coefficient Ft-lbs/(mph)^2	Slant Axis Angle	Wind Azimuth Aspect Angle
* Vertical axis z'	58	45°	270°
* Slant axis z	12	135°	315°
Slant bearing axis x	56	135°	270°
Slant bearing axis y	40	180°	0°
Azimuth bearing axis x'	92	45°	90°
Azimuth bearing axis y'	160	0	0

* Determine drive torque requirements.

3.3 *Precipitation and Icing*

The geographical location of the antenna site will determine to a great extent the problems which must be faced to achieve reliable all-weather operation. The effects of snow and ice accumulation may have to be considered for most locations.

The steep incline of the main reflector surfaces at all pointing attitudes provides an inherent snow-shedding feature. It is proposed to fabricate the reflector surface of thin stretch-formed aluminum panels, which would make it practical to heat the rear of the panels to melt snow and ice accumulations. If electrical heating is used for this purpose, it has been estimated that a 250-kw capacity would be required to keep the reflector surface clear under nominal conditions. Surface treatment of panels and structure with anti-sticking flourocarbon resins may also be practical to minimize collection of ice and snow.

IV. CONCLUSIONS

It makes sense to think of the possibilities of using monocoque or semi-monocoque construction techniques for certain portions of the structure. Since the configuration of each component of the antenna can be described in terms of surfaces, there is no reason why portions of structural shells might not be suitable. This is particularly true of the azimuth pedestal and the conical surfaces supporting the radial bearings. Numerical analysis routines soon will be available to study such components.

Since the deflection patterns of the main reflector surface and the motion of the subreflector under various loading conditions are actually inputs to determine the decrease in system performance in terms of electrical parameters, it follows that the final mechanical criterion is based on electrical response characteristics. Hence it may be possible to formulate such a criterion directly and dispense with the obsolute tolerances on surface run-out, etc. that normally provide the standards for the mechanical designer. For example, deflections of the subreflector along the axis of the feed horn are much less severe than transverse displacements of the subreflector in terms of pattern degradation and tracking jitter. Also, surface deflections should be judged in terms of an auxiliary "best-fit" paraboloid which has its focal point coincident with the design paraboloid, rather than by reference to the original design paraboloid itself. All this suggests it would be meaningful to consider a structural optimization study in which the structure is designed on the basis of desired electrical performance rather than deflection tolerances.

Considerable future work must be done on the dynamic response characteristics of the open cassegrain configuration. In addition to the determination of natural frequencies and normal modes as inputs for control system design, the response of the structure to the random wind loading must be examined. The nature of the loading suggests that statistical variables might be very useful for such a description. There are few discussions of this type of structural response available in the literature. Those that appear are limited to the design of earthquake sensitive structures. An undertaking of this sort would be a prodigious effort, but might be extremely useful in the discussion of the behavior of exposed deflection-sensitive structures.

The various analyses conducted to date have demonstrated the mechanical feasibility of the open cassegrain configuration. The preliminary concept discussed in these pages has been shown to meet the structural requirements that are reasonable to impose at this stage. Detailed analyses have been directed only to certain specific problem areas where an obvious need for first-order quantitative information was recognized. These investigations have shown plausible solutions for such problems are available, and have provided better understanding of the overall structural behavior. In no sense should it be inferred that such calculations represent a design for the open cassegrain. Rather, they establish the configuration shown as a justifiable concept for such an antenna.

V. ACKNOWLEDGMENTS

K. N. Coyne and F. Brauns participated in many aspects of the structural analysis and their assistance proved invaluable. M. Lutchansky was involved in the early phases of the study and many of his ideas are reflected in the final structural configuration. H. W. Boschen demonstrated his skill and ingenuity in building the model of the structure.

REFERENCES

1. Nelson, W. L., and Cole, J. W., Autotrack Control Systems for Antenna Mounts with Non-Orthogonal Axes, B.S.T.J., this issue, pp. 1367–1403.
2. Gallagher, R. H., *A Correlation Study of Methods of Matrix Structural Analysis*, New York, The MacMillan Company, 1964.
3. Przemieniecki, J. S., Matrix Structural Analysis of Substructures, AIAA Journal, *1*, No. 1, Jan. 1963.
4. Coyne, K. N., Hydrodynamic Techniques for Study of Wind Effects on Antenna Structures, B.S.T.J., this issue, pp. 1339–1365.

Mode Conversion in Circular Waveguides

By E. R. NAGELBERG and J. SHEFER

(Manuscript received May 14, 1965)

Mode conversion from TE_{11} to TM_{11} modes in circular waveguides is investigated. It is found that an iris placed across the waveguide or an abrupt change in guide radius (step) will produce a wide range of mode conversion coefficients which can be used in most dual-mode feedhorn applications. The step discontinuity is found to produce relatively constant mode conversion over a wide band of frequencies.

I. INTRODUCTION

In order to obtain optimum illumination at the aperture of a dual-mode conical horn,[1] as in the open cassegrain antenna,[2] it is necessary to control the relative amplitudes and phases of the TE_{11} and TM_{11} modes. The mode combination is established (considering the antenna as a transmitter) by exciting the dominant mode (TE_{11}) in the circular waveguide that feeds the horn, and converting a part of that signal to propagate in the TM_{11} mode by introducing suitable transmission line discontinuities in the vicinity of the waveguide to cone transition (Fig. 1). The general requirement that a discontinuity be suitable for mode conversion as described above is that it perturb the incident transverse electric field in such a way as to produce a component of electric field in the direction of propagation. This may be done by introducing a conducting surface transverse to the direction of propagation, in which case the coupling arises directly from the fact that the component of electric field tangent to the surface must vanish.

To be generally useful, the mode conversion configuration must provide controlled mode conversion over a useful band of frequencies without introducing significant reflection of the input TE_{11} mode. It is further required that the converter discontinuity be circularly symmetric, both to prevent excitation of unwanted modes and to permit operation of the device for signals of arbitrary polarization.

Three basic schemes are possible, the iris, the groove, and the step, each of which is shown in Fig. 2.

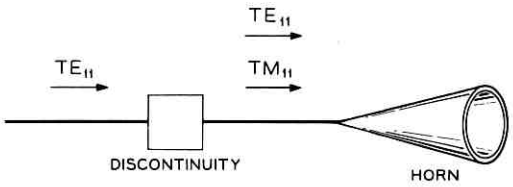

Fig. 1 — Horn and mode converter discontinuity.

It has been found that each of the configurations in Fig. 2 can provide adequate mode conversion with suitable adjustment. However, in cases (a) and (b), a single discontinuity excites both backward and forward propagating TM_{11} modes. This backward wave, after reflection from a transition (see Section 2.2) from standard to oversized waveguide, combines with the forward traveling wave with a relative phase which is highly frequency dependent, thus making a transition and single iris a frequency sensitive system. In addition, configuration (b) has an intrinsic dispersion due to resonance in the groove itself.

If the diameter of the input waveguide is chosen small enough so that it will not support a TM_{11} mode over the frequency band, configuration (c) behaves very well over a wide band. Such a discontinuity was used by Potter[1] in early experiments with dual-mode horn excitation.

II. THEORETICAL ASPECTS OF MODE CONVERSION

2.1 *General Considerations.*

The $TE_{11} \rightarrow TM_{11}$ mode conversion may be accomplished by introducing a circularly symmetric perturbation into a section of circular waveguide preceding the horn throat. Assume that the effect of this perturbation is to produce, at $z = 0$, a longitudinal component of electric

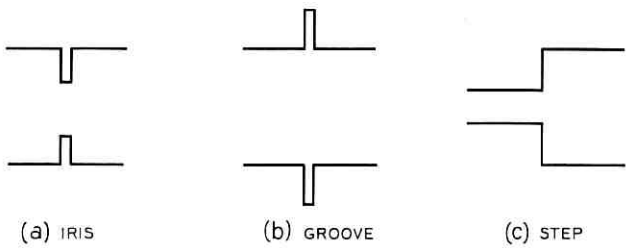

Fig. 2 — Three types of mode converters.

field with $\cos\varphi$ symmetry. A typical section along the z-axis will be as shown in Fig. 3. If the incident transverse electric field is given by

$$\mathbf{E}_t^{(\text{inc})} = \mathrm{E}_0\left[\mathbf{e}_\rho \frac{J_1(\gamma\rho/a)}{\rho/a}\cos\varphi - \mathbf{e}_\varphi J_1'(\gamma\rho/a)\sin\varphi\right]\exp(-j\beta_{\text{TE}}z) \quad (1)$$

where $\gamma = 1.841$, then interaction at the obstacle will produce a longitudinal electric field at $z = 0$ of the form,

$$E_z(z = 0) = E_0 f(\rho/a)\cos\varphi \quad (2)$$

where $f(\rho/a)$ will, of course, depend on the geometry. The resulting TM_{11} mode propagating in the $+z$ direction will then have the following components:

$$\begin{aligned} E_z^{\text{TM}} &= AJ_1(\chi\rho/a)\cos\varphi\exp(-j\beta_{\text{TM}}z) \\ E_\rho^{\text{TM}} &= \frac{-j\beta_{\text{TM}}a}{\chi}AJ_1'(\chi\rho/a)\cos\varphi\exp(-j\beta_{\text{TM}}z) \\ E_\varphi^{\text{TM}} &= \frac{j\beta_{\text{TM}}a}{\chi^2}\frac{AJ_1(\chi\rho/a)}{\rho/a}\sin\varphi\exp(-j\beta_{\text{TM}}z) \end{aligned} \quad (3)$$

where

$$\chi = 3.832$$

and

$$A = \frac{2E_0}{[J_0(\chi)]^2}\int_0^1 \zeta f(\zeta)J_1(\chi\zeta)\,d\zeta \quad (4)$$

where $f(\zeta)$ describes the radial dependence of E_z, as in (2). A convenient parameter, for both design and measurement purposes, is the ratio of magnitudes of the ρ components of the TM_{11} and TE_{11} electric fields, evaluated at $\rho = a$, $\varphi = 0$. This will be referred to as the conversion

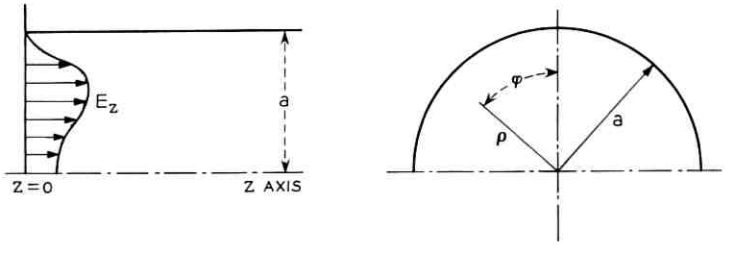

Fig. 3 — Effect of perturbation is to produce a longitudinal electric field at $z = 0$.

coefficient and is given by

$$C = \left|\frac{E_\rho^{\text{TM}}}{E_\rho^{\text{TE}}}\right| (\rho = a, \varphi = 0)$$

$$= \left|\frac{2\beta_{\text{TM}} a J_1'(\chi)}{\chi J_1(\gamma)[J_0(\chi)]^2} \int_0^1 \zeta f(\zeta) J_1(\chi\zeta)\, d\zeta\right| \quad (5)$$

$$= 2.2\beta_{\text{TM}} a \left|\int_0^1 \zeta f(\zeta) J_1(\chi\zeta)\, d\zeta\right|$$

which, in principle, may be calculated once the longitudinal electric field is known.

In the design of dual-mode conical horns, it is necessary to specify conditions at the radiating aperture, since it is the field distribution at this cross section which determines the feed characteristics. If we let the subscript 2 denote conditions at the aperture and the subscript 1 denote conditions at the cross section where conversion occurs, then the respective coefficients are related by

$$\frac{C_2}{C_1} = \frac{f}{f_c^{\text{TM}}} \left\{\left[\left(\frac{f}{f_c^{\text{TM}}}\right)^2 - 1\right]\left[\left(\frac{f}{f_c^{\text{TM}}}\right)^2 - 0.23\right]\right\}^{-\frac{1}{4}} \quad (6)$$

where f_c^{TM} is the TM_{11} cutoff frequency at the cross section where conversion occurs. The relationship of (6) is given in Fig. 4. To derive (6), the following assumptions are made:

(1.) The power transmitted in each mode is constant, i.e., the horn is lossless and no further conversion occurs.

(2.) The transverse electromagnetic fields at any cross section of a

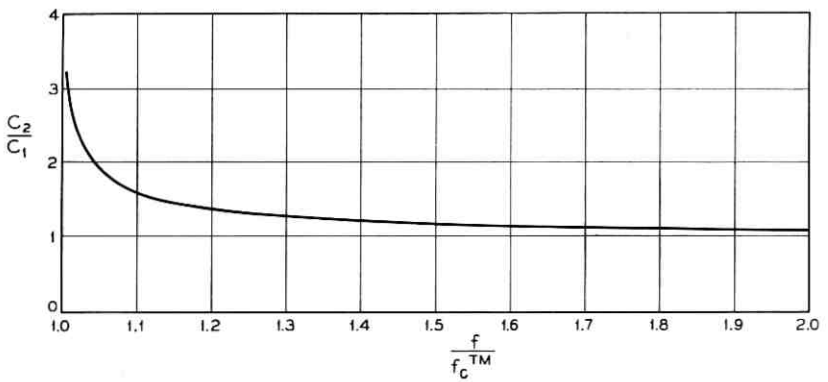

Fig. 4 — A graph of ratio of coefficients at horn aperture and converter.

narrow angle conical horn are approximately the same as in a circular waveguide with the same radius as the cross section.

(3.) The aperture diameter is much larger than the cutoff diameter for the TM_{11} mode, so that at the aperture the guide wavelengths of both modes are approximately equal to the free space wavelength.

Although the conversion coefficient as defined above refers only to the relative amplitudes of the two modes, one should recognize that it is equally important that their relative phases at the horn aperture be kept within tolerable limits over the prescribed frequency band. If we assume that the relative phase at the converter discontinuity is frequency independent, then the error is due to the difference in electrical length for the two modes as they propagate through the horn.

Consider the conical horn shown in Fig. 5, which has an aperture radius r_o and half angle α. It can be shown that the phase shift between the converter and aperture cross sections for a single mode is given approximately by

$$\theta_{12} = \frac{1}{\kappa_2 p}\sqrt{1-\kappa_2^2} - \kappa_2 \cos^{-1}\kappa_2 - \frac{1}{\kappa_1 p}\sqrt{1-\kappa_1^2} - \kappa_1 \cos^{-1}\kappa_1 \qquad (6a)$$

where

$$\kappa_{1,2} = \frac{2\pi z_{1,2}}{\lambda_0 p}$$

and p is equal to $\sin\alpha/1.84$ and $\sin\alpha/3.83$ for the TE_{11} and TM_{11} modes respectively.

Fig. 6 shows the result of calculating the differential phase shift $\theta_{12}^{TE} - \theta_{12}^{TM}$ at the aperture of a horn with $r_0 = 12\lambda_0$ and $\alpha = 3.25°$. The frequency range is 3.7–4.2 kmc and conversion is assumed to take place at that cross section with radius equal to 1.5 times the cutoff radius for the TM_{11} mode at 3.7 kmc.

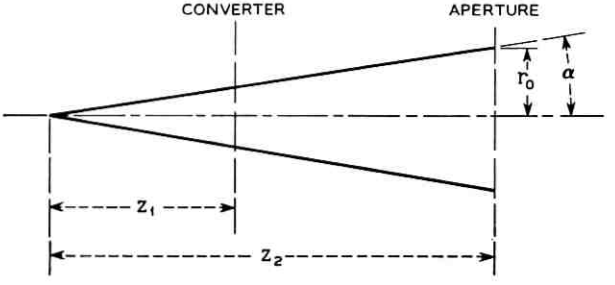

Fig. 5 — Conical horn.

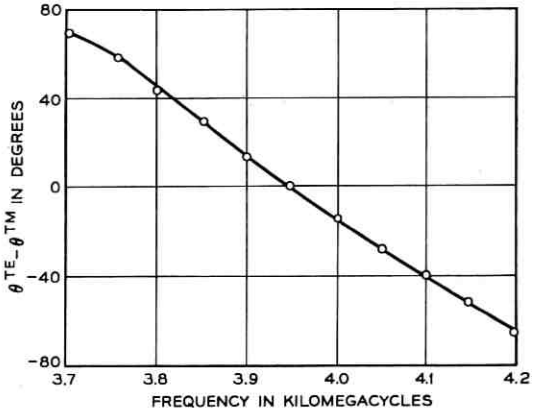

Fig. 6 — Differential phase shift as a function of frequency for $r_0 = 12\lambda_0$ and $\alpha = 3.25°$.

2.2 *Mode Conversion at an Iris*

To demonstrate the application of these formulas, we consider the example of mode conversion at an iris in an oversized circular waveguide, as shown in Fig. 7. The purpose of the standard input circular waveguide and subsequent transition is to prevent the uncontrolled excitation of a TM_{11} mode at the input of the system.

In order to calculate the conversion coefficient it is necessary to have an expression for the longitudinal component of electric field at $z = 0$. This may be determined, in principle, by solving the interior boundary value problem associated with the iris, i.e. by obtaining a solution to Maxwell's equations which has the following characteristics:

(1.) The component of electric field tangent to the waveguide, transition, and iris walls must vanish.

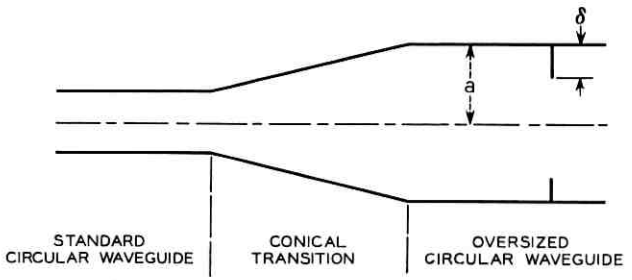

Fig. 7 — Mode conversion at an iris.

(2.) At $z = +\infty$ the field should consist of a linear combination of TE_{11} and TM_{11} modes propagating in the forward direction; at $z = -\infty$ the field should consist of a given TE_{11} mode propagating in the forward direction and a reflected TE_{11} mode propagating back toward the source.

(3.) The singularity at the edge of the iris should be such that the energy in any finite element of volume is finite.

Although the exact solution to the problem as stated is prohibitively complex, it is possible to make certain simplifying assumptions and still obtain useful information, which is in agreement with observation, particularly with respect to the variation of mode conversion with iris size and frequency. We approach the problem within the framework of perturbation theory, and assume that the incident TE_{11} wave is unperturbed at distances far from the iris, but is distorted at the iris walls in such a way as to produce a longitudinal component of electric field. An analogous two-dimensional electrostatics problem is the case of a uniform electric field terminated by a perfectly conducting plane with a thin protrusion, as shown in Fig. 8. If E_0 is the magnitude of the uniform field, then the x component of electric field in the plane $x = 0$ is given by

$$\begin{aligned} E_x &= \frac{E_0 y}{\sqrt{y^2 - \delta^2}} \qquad x = 0+, \qquad 0 \leqq y < \delta \\ E_x &= \frac{-E_0 y}{\sqrt{y^2 - \delta^2}} \qquad x = 0-, \qquad 0 \leqq y < \delta \\ E_x &= 0 \qquad\qquad x = 0, \qquad y > \delta . \end{aligned} \tag{7}$$

We are motivated, by analogy, to assume that the longitudinal electric field at the iris is given by a function of the form

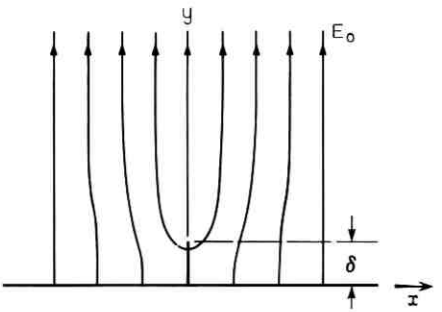

Fig. 8 — A uniform electrostatic field terminated by a perfectly conducting plane with a thin protrusion.

$$E_z(\rho,\varphi,z = 0+) = E_{\rho}^{\mathrm{TE}}{}_{(\rho=a)} \left(\frac{a-\rho}{\delta}\right) g(1-\rho/a)$$

$$= E_0 \cos\varphi \left(\frac{a-\rho}{\delta}\right) g\left(\frac{a-\rho}{\delta}\right) \qquad a \geqq \rho > a-\delta \quad (8)$$

$$E_z(\rho,\varphi,z = 0-) = -E_0 \cos\varphi \left(\frac{a-\rho}{\delta}\right) g\left(\frac{a-\rho}{\delta}\right) \qquad a \geqq \rho > a-\delta$$

$$E_z(\rho,\varphi,z = 0\pm) = 0 \qquad a-\delta > \rho \geqq 0.$$

Note that the uniform electric field of the electrostatic problem is replaced by the unperturbed ρ component of electric field of the TE_{11} mode, evaluated at $\rho = a$. In addition, an undetermined function $g(\zeta)$ has been introduced to account for the singularity at $\rho = a - \delta$. This function may be expected to have the general behavior indicated in Fig. 9, and must be square integrable on (0,1).

It was noted earlier that two TM_{11} waves, with equal amplitudes, propagate in opposite directions away from the iris. The forward wave will travel unperturbed; however, the backward wave is totally reflected from the transition since, by assumption, the waveguide to the left of the transition will not support a TM_{11} mode. This reflected wave will then combine with the forward wave excited at the iris. Since the reflected wave arrives back at the iris with a phase which is highly fre-

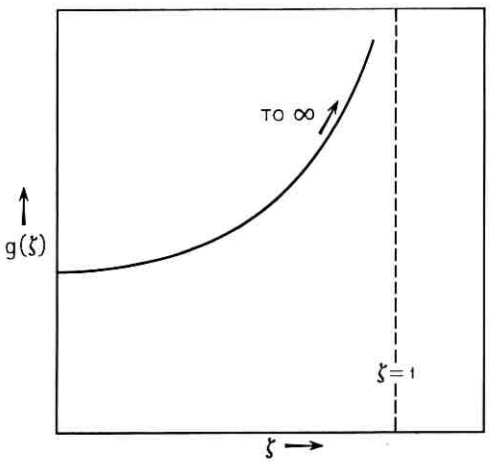

Fig. 9 — Expected behavior of the function $g(\zeta)$.

quency dependent, the amplitude of the sum varies rapidly with frequency.

Consider the wave initially excited in the forward direction. In order to calculate the conversion coefficient C_0, if only this wave were present, substitute the expression for $E_z(\rho,\varphi,z = 0+)$ as given in (8), into (5). This result is relevant even when the backward wave is taken into account, since $2C_0(\delta/a,ka)$ will then give the envelope of the combination as the frequency is varied. Under the assumption of small conversion

$$\begin{aligned} C_0 &\approx \frac{2\alpha\beta_{\mathrm{TM}}a}{J_1(\gamma)}\left(\frac{\delta}{a}\right)^2 + O\left[\left(\frac{\delta}{a}\right)^3\right] \\ &\approx \frac{2\alpha ka}{J_1(\gamma)}\frac{\lambda}{\lambda_g{}^{\mathrm{TM}}}\left(\frac{\delta}{a}\right)^2 \end{aligned} \tag{9}$$

where α is the second moment of the singularity function $g(\zeta)$,

$$\alpha = \int_0^1 \zeta^2 g(\zeta)\, d\zeta . \tag{10}$$

The value of α, assuming a square root singularity

$$g(\zeta) = \frac{1}{\sqrt{1-\zeta^2}} \tag{11}$$

as in the electrostatic case, would be $\alpha = \pi/4$, giving the "quasi-static" expression for the conversion coefficient

$$C_0 \approx \frac{\pi ka}{2J_1(\gamma)}\frac{\lambda}{\lambda_g{}^{\mathrm{TM}}}\left(\frac{\delta}{a}\right)^2 . \tag{12}$$

2.3 *Mode Conversion by a Step*

We now consider the problem of mode conversion at a simple discontinuity from a standard size waveguide, in which only the TE_{11} mode can propagate, to one which is oversized to support a TM_{11} mode as well. The configuration is shown in Fig. 10.

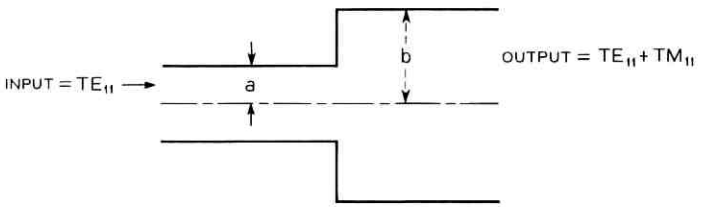

Fig. 10 — Mode conversion at a step.

The approach we use to determine the conversion coefficient in this case is somewhat different than the quasi-static approximation used for the iris. It may be shown that a knowledge of the transverse electric field at a cross section of waveguide uniquely determines the amplitudes of the TE and TM modes which comprise this field. In the present problem this implies that a knowledge of $E_t(\rho,\varphi)$ at $z = 0$ would permit the calculation, in principle, of the conversion coefficient for the step discontinuity. We have observed experimentally that the input standing wave ratio is very low for frequencies more than 5 per cent above the TM_{11} mode cut-off frequency for the oversized waveguide. This motivates the "perfect match" approximation for mode conversion by a step discontinuity, in which we calculate the modes propagating to the right by assuming that at $z = 0$ the transverse field in the common aperture is that due only to the unperturbed incident TE_{11} mode. Thus at $z = 0$

$$\begin{aligned} E_\rho &= E_0 a/\gamma\rho J_1(\gamma\rho/a)\cos\varphi && 0 \leqq \rho \leqq a \\ &= 0 && a < \rho \leqq b \\ E_\varphi &= -E_0 J_1'(\gamma\rho/a)\sin\varphi && 0 \leqq \rho \leqq a \\ &= 0 && a < \rho \leqq b. \end{aligned} \tag{13}$$

The components of the TM_{11} electric field to the right of the discontinuity will be given by

$$\begin{aligned} E_z^{TM} &= AJ_1(\chi\rho/b)\cos\varphi \exp(-j\beta_{TM}z) \\ E_\rho^{TM} &= \frac{-j\beta_{TM}b}{\chi} AJ_1'(\chi\rho/b)\cos\varphi \exp(-j\beta_{TM}z) \\ E_\varphi^{TM} &= \frac{+j\beta_{TM}b^2}{\chi^2\rho} AJ_1(\chi\rho/b)\sin\varphi \exp(-j\beta_{TM}z) \end{aligned} \tag{14}$$

where A is related to the transverse electric field at $z = 0$ by the expression

$$A^2 = -\frac{2\chi^2}{\beta_{TM}^2\pi b^4}\int_{\rho=0}^{a}\int_{\varphi=0}^{2\pi} \mathbf{E}_t\cdot\mathbf{E}_t^{TM}\rho d\rho d\varphi \tag{15}$$

where $\mathbf{E}_t$ is given in (13) and $\mathbf{E}_t^{TM}$ in (14), evaluated at $z = 0$. The calculation may be further simplified by assuming that

$$b = a(1 + \epsilon) \tag{16}$$

where $\epsilon \ll 1$, which is a case of practical interest. Under this additional approximation the conversion coefficient may be shown to be

$$C = 2\frac{b-a}{a} + O\left[\left(\frac{b-a}{a}\right)^2\right] \tag{17}$$

for the perfect match case.

III. METHOD OF MEASUREMENT

The TM_{11} mode propagating in the positive z direction will produce a spatial beat with the TE_{11} mode propagating with a different phase velocity, the net effect being a standing wave. A moving probe is used to measure the standing wave ratio, which in turn is directly related to the amplitude ratio of the two modes.

Suppose the radial electric field components of the two modes at $\rho = a$ and $\varphi = 0$ in Fig. 11 are given by

$$\begin{aligned} E_\rho^{TE} &= A \exp(-j\beta_{TE}z) \\ E_\rho^{TM} &= B \exp(-j\beta_{TM}z) \end{aligned} \tag{18}$$

then the current of a square law detector with an electric probe at $\rho = a, \varphi = 0$ is given by

$$I \approx |\exp(-j\beta_{TE}z) + C \exp(j\phi_1) \exp(-j\beta_{TM}z)|^2 \tag{19}$$

where

$$\frac{B}{A} = \left|\frac{B}{A}\right| \exp(j\phi_1) = C \exp(j\phi_1)$$

and the voltage standing wave ratio will be given by

$$R = \left(\frac{I_{max}}{I_{min}}\right)^{\frac{1}{2}} = \frac{1+C}{1-C} \tag{20}$$

or

$$C = \frac{R-1}{R+1}.$$

The distance between successive minima is

$$l = \frac{2\pi}{\beta_{TE} - \beta_{TM}} = \frac{\lambda_{gTE} \cdot \lambda_{gTM}}{\lambda_{gTM} - \lambda_{gTE}}. \tag{21}$$

Since there is no longitudinal wall-current flow at $\varphi = 0$ for the two modes, a slot and moving probe can be used, and the quantity C can be determined from the measured standing wave ratio R. It is also fairly easy to determine mode purity, i.e., the presence of higher order modes other than the TM_{11} mode, by plotting the probe-current distribution

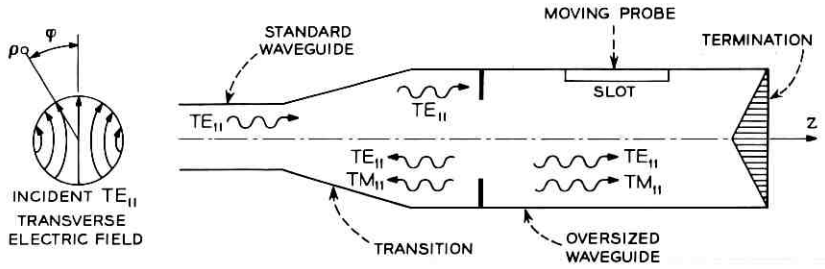

Fig. 11 — Measurement of conversion coefficient.

curve as we move between two minima. The harmonic variation of detector current with a periodicity given by (19) will be distorted if more than two modes are present.

This method of measurement requires that unidirectional traveling waves be present for the two modes, with no waves reflected from the termination. With a fairly extended load at the end of the oversized waveguide, a voltage reflection coefficient at the termination of less than 0.01 was obtained over a wide frequency band.

A practical limiting factor appeared when measuring mode content at frequencies that were far from the TM_{11} mode cut-off. As seen from (21), the minimum probe travel is proportional to $1/(\lambda_{gTE} - \lambda_{gTM})$ which becomes very large as we move away from cut-off. In an oversized circular waveguide with 2.8″ diameter and TM_{11} mode cut-off at $f_c^{TM} = 5120$ mc, it was difficult to make measurements at frequencies higher than 6500 mc.

A block diagram of the experimental arrangements is shown in Fig. 12.

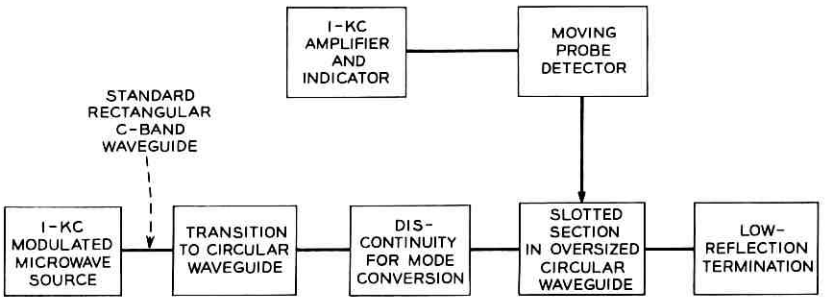

Fig. 12 — Experimental arrangement.

IV. EXPERIMENTAL RESULTS

Measured values of mode conversion coefficient C for iris and step discontinuities are shown in Figs. 13–17. The highly dispersive character of the transition-iris combination is evident in Fig. 13. The transition used was a standard transition from TD-2 rectangular waveguide to 2.8″ I.D. circular waveguide. A 12″ section of 2.8″ I.D. circular waveguide was placed between the iris and transition. A backward traveling TM_{11} wave is totally reflected at the transition between standard and oversized waveguides and is recombined with the forward traveling TM_{11} wave. The peak values correspond to points of positive interference, with amplitude $2C_0$ in (12). At the higher frequencies, the TE_{31} and higher order modes were observed whenever the backward traveling TM_{11} wave was reflected from a noncircular portion of the transition.

In Fig. 14, values of mode conversion are measured for different iris sizes at a frequency range where the reflected TM_{11} wave combines with the forward wave to give the peak value of conversion coefficient C =

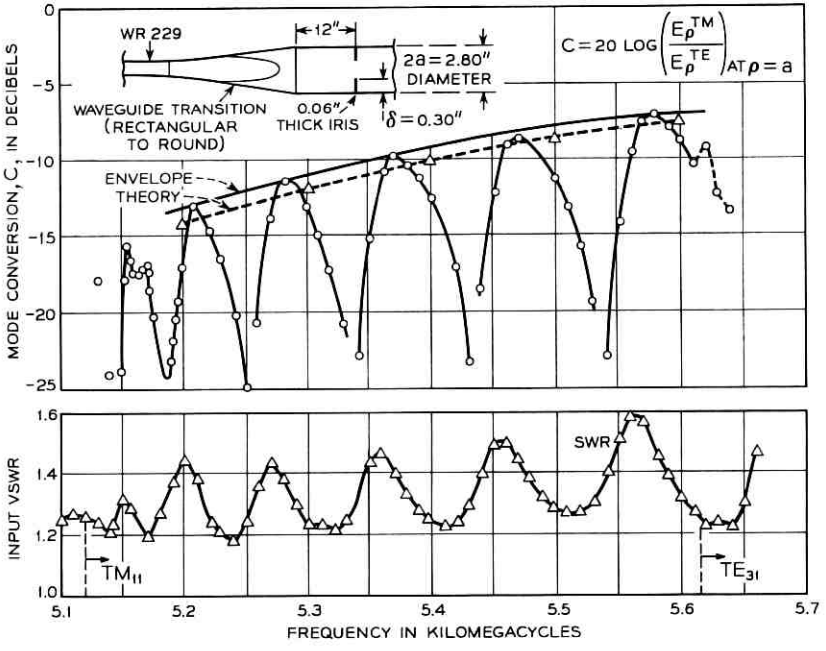

Fig. 13 — Mode conversion for transition-iris coupler, compared with quasi-static approximation.

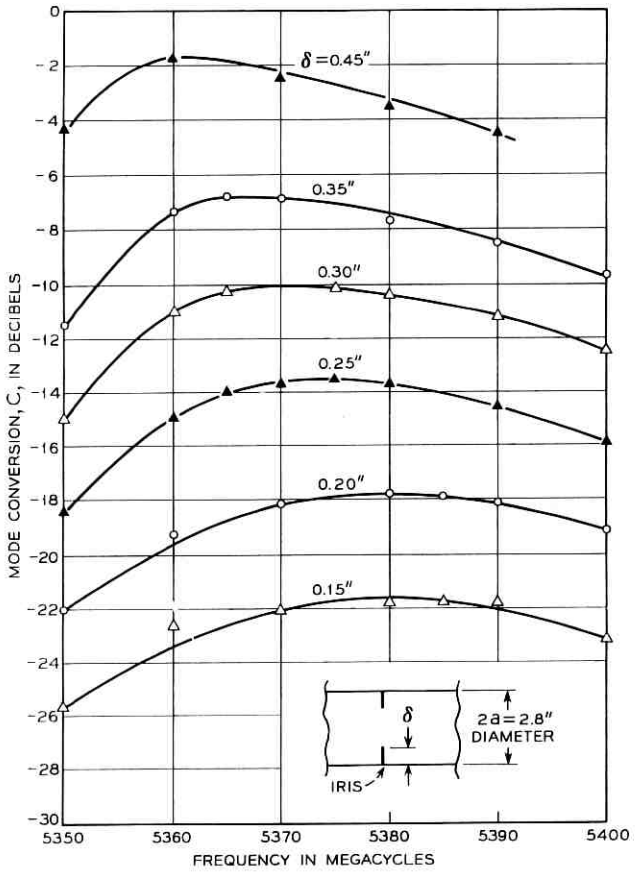

Fig. 14 — Mode conversion for transition-iris coupler.

$2C_0$. The derived values of C as a function of frequency, with the iris size as parameter, are shown in Fig. 15. The values are compared with the quasi-static approximation of (12). The overall reflection, measured in the standard C-band waveguide preceding the transition, is also shown. The theoretical approximation is seen to give good results, especially for the smaller discontinuities, where $\delta/a < 0.25$. The power reflected at the iris is then less than 0.05 of the incident power.

In Fig. 16, measured values of mode conversion are given for a step discontinuity. The waveguide to the left of the step is below cut-off for the TM_{11} wave. The mode conversion under these conditions is flat over

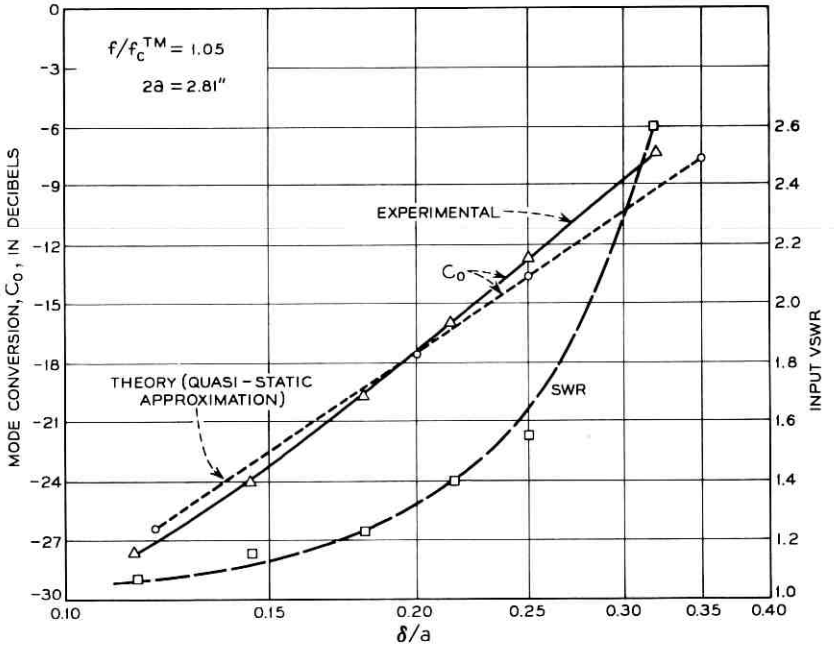

Fig. 15 — Mode conversion and overall input SWR for iris, (no reflected wave from transition).

a very wide range of frequencies. In the upper curve in Fig. 16, measured values of mode conversion are given for a step in combination with an iris, again resulting in a flat frequency response. The step-iris combination provides a means for making fine adjustments of mode conversion above the value provided by the step alone.

In Fig. 17, mode conversion at a step is plotted as a function of step size. It is of interest to note the linear dependence of C on step size, which bears out the prediction made by the theoretical approximation. The incident power reflection was small for step discontinuities, being less than -17 db in all cases, and less than -28 db for all steps measured at 6000 mc.

V. CONCLUSIONS

Any discontinuity in a circular waveguide which distorts the TE_{11} mode in such a way as to introduce a longitudinal component of electric field will couple the TE_{11} and TM_{11} modes of propagation. Prevention of

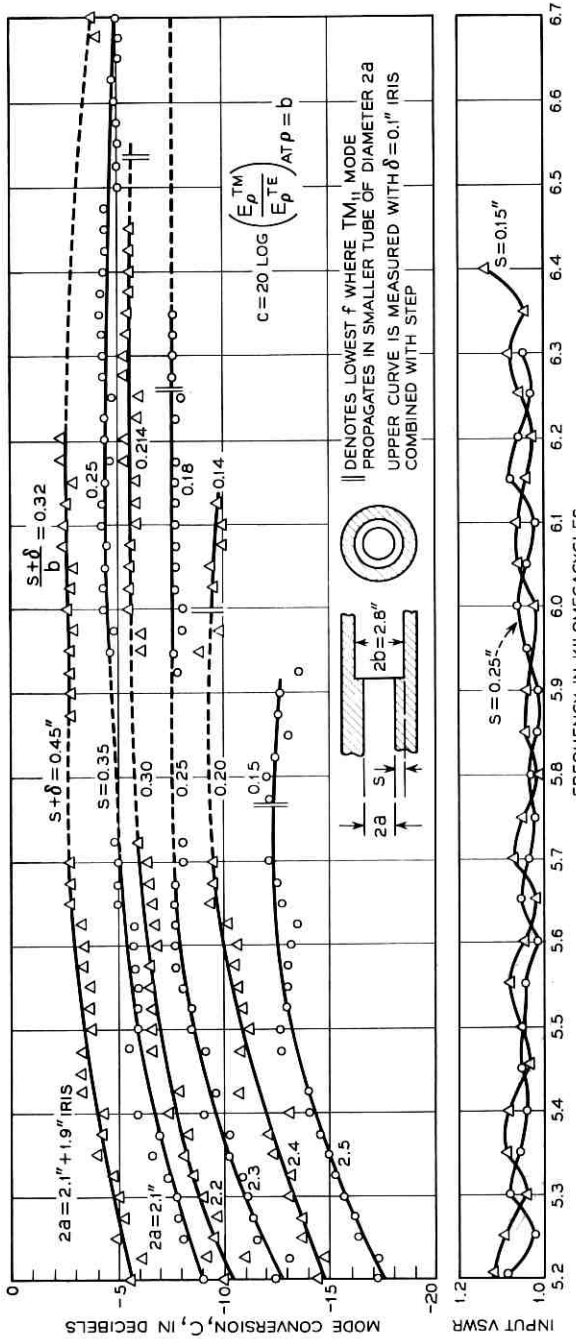

Fig. 16 — Mode conversion at step discontinuity in circular waveguide.

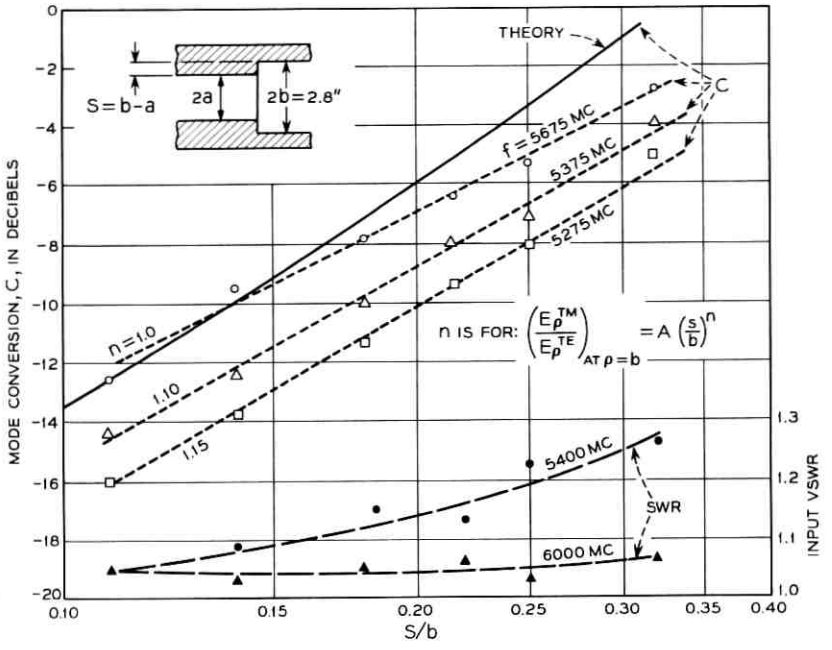

Fig. 17 — Mode conversion at step discontinuity.

coupling between these and other modes can be controlled by maintaining circular symmetry in the discontinuity and limiting the diameter of the waveguide. Two of the simplest configurations, the iris and the step, have been evaluated here and shown to have useful characteristics.

The iris was evaluated experimentally in conjunction with a waveguide transition, but the characteristics of the iris itself have been deduced and expressed in (12) and Fig. 15.

The bidirectional and reflective character of a single iris coupler (neglecting the transition) is not unlike that of a single hole coupling two waveguides. In fact, an analogy can be drawn between a single iris coupler and a single hole coupling two waveguides whose cutoff frequencies correspond with those of the TE_{11} and TM_{11} modes in the circular guide. The theory of iteratively and continuously coupled transmission lines may be used to design couplers consisting of multiple irises. Selective loading of the modes in the circular guide can provide a further control of the distributed iris coupling characteristics.

The very low-reflection and broadband coupling characteristics of the simple step converter make it particularly useful as a means for provid-

ing mode conversion. A fine adjustment of the conversion coefficient can be made by adding an iris at the step as indicated in Fig. 16.

These conversion techniques, as presented here or appropriately extended and combined, provide the means for achieving a wide range of characteristics.

ACKNOWLEDGMENTS

The authors wish to thank J. S. Cook and H. Zucker for many helpful suggestions. Also acknowledged are the efforts of J. R. Donnell and R. E. Pratt, who participated in the measurements, and E. M. Elam, who helped design the measuring apparatus.

REFERENCES

1. Potter, P. D., A New Horn Antenna with Suppressed Side-lobes and Equal Beamwidths, Microwave J., *6*, 1963, p. 71.
2. Cook, J. S., Elam, E. M., and Zucker, H., The Open Cassegrain Antenna: Part I. Electromagnetic Design and Analysis, B.S.T.J., this issue, pp. 1255–1300.

Hydrodynamic Techniques for Study of Wind Effects on Antenna Structures

By K. N. COYNE

(Manuscript received June 1, 1965)

The effects of random wind-induced torques and structural loading are described and evaluated in terms of antenna performance, pointing errors, and overturning stability. A model test theory is developed to utilize economical small scale models to accurately determine drag torque coefficients of complex asymmetrical structures, using water as the test medium. The technique utilizes towing of inverted instrumented models in a hydrodynamic test basin. Data are presented on both the open cassegrain and the triply-folded horn-reflector antennas.

I. INTRODUCTION

The feasibility of operating a narrow beam tracking antenna without the environmental protection of a radome is dominated by the consideration of antenna tracking accuracy under the influence of wind-induced random disturbance torque. Although the wind power spectrum at a point may be quite variable and may depend somewhat upon the locality, for our purpose it is sufficient to assume a characteristic resembling a first order low-pass filter with a cut-off angular frequency ω_c in the range 0.12 to 3 radians per second.[1,2,3] That is, we assume the variable component of wind velocity to have a two-sided power-density spectrum of the form:

$$\Phi_{vv}(\omega) = \frac{1}{\pi} \frac{\omega_c V_1^2}{\omega_c^2 + \omega^2} \tag{1}$$

where V_1 is the standard deviation of the variational component of the wind velocity.

It is assumed that a satellite communications antenna must survive winds of 100 mph velocity. For compressible flow, Bernoulli's theorem can be expressed as

$$\frac{dp}{\rho} = -V\,dV,$$

where p is pressure, ρ is fluid density and V is the fluid free stream velocity. For an adiabatic process,

$$p = K\rho^{\gamma}$$

where K and γ are constants, γ being the ratio of specific heats. For dry air, $\gamma = 1.405$. Then, denoting the state values of a gas at stagnation by the subscript o,

$$\int_{p_o}^{p} \frac{dp}{\rho} = K\gamma \int_{\rho_o}^{\rho} \rho^{\gamma-2}\,d\rho = \frac{K\gamma}{\gamma - 1}\left[\rho^{\gamma-1} - \rho_o{}^{\gamma-1}\right]$$

or

$$\int_{p_o}^{p} \frac{dp}{\rho} = \frac{\gamma}{\gamma - 1}\left(\frac{p}{\rho} - \frac{p_o}{\rho_o}\right).$$

Bernoulli's theorem then gives

$$\frac{\gamma}{\gamma - 1}\frac{p}{\rho} + \frac{1}{2}V^2 = \frac{\gamma}{\gamma - 1}\frac{p_o}{\rho_o}.$$

Solving for V^2 and using the adiabatic gas law

$$V^2 = \frac{2\gamma}{\gamma - 1}\frac{p_o}{\rho_o}\left[1 - \left(\frac{p}{p_o}\right)^{(\gamma-1)/\gamma}\right].$$

Noting that the speed of sound C is given by $C^2 = \gamma p/\rho$,

$$V^2 = \frac{2C_o^{\,2}}{\gamma - 1}\left[1 - \left(\frac{p}{p_o}\right)^{(\gamma-1)/\gamma}\right].$$

Solving for the pressure ratio p/p_o and expanding in a Maclaurin series;

$$\frac{p}{p_o} = \left[1 - \frac{\gamma - 1}{2}\left(\frac{V}{C_o}\right)^2\right]^{\gamma/(\gamma-1)}$$

$$\frac{p}{p_o} = 1 - \frac{\gamma}{2}\left(\frac{V}{C_o}\right)^2 + \frac{\gamma}{8}\left(\frac{V}{C_o}\right)^4 - \frac{\gamma(2 - \gamma)}{48}\left(\frac{V}{C_o}\right)^6 + \cdots.$$

The ratio of the third term to the second term in the series is $V^2/4C_o^{\,2}$. For a free stream velocity of 100 mph = 147 ft/sec, C_o is approximately 1.018 C_f where the subscript f refers to free stream conditions. Under standard temperature and pressure then, $V_f^{\,2}/4C_o^{\,2} = 0.0037$. For rounded bodies such as spheres and cylinders normal to the direction of flow the maximum pressure difference $|\ p - p_f\ | \approx 1.2\ (p_o - p_f)$ while

for bluff bodies $| p - p_f |$ may be locally as high as 1.7 $(p_o - p_f)$. Assuming the latter, $V^2/4C_o^2 = 0.010$; that is, for a free stream velocity of 100 mph at no point on the surface of the antenna does the third term in the series exceed one per cent of the second term. Since the series is convergent and alternating, the error in calculating dynamic pressure resulting from discarding all terms after the second must be less than 1 per cent. Then:

$$\frac{p}{p_o} = 1 - \frac{\gamma}{2}\left(\frac{V}{C_o}\right)^2 = 1 - \frac{1}{2}\frac{\rho_o}{p_o}V^2$$

or $p = p_o - \frac{1}{2}\rho_o V^2$. Thus the assumption of incompressibility leads to negligible error in calculating the forces and moments induced by winds up to 100 mph on a full scale antenna.

With a knowledge of the antenna structural and servo parameters, the pointing error resulting from a disturbing wind torque input can be predicted. In their generalized complex form, the angular displacement of the pointing vector, θ, can be related to the wind torque T by a system transfer function such that

$$K_{ij}(s) = \frac{T_i(s)}{\theta_j(s)} \tag{2}$$

where s is the Laplace transform variable and the indices refer to the antenna axes.

Little is known about the spatial distribution of instantaneous wind velocity, but if we can assume such variation to be negligible over the dimensions of an antenna; then, from Bernoulli's theorem for incompressible flow, we may write for the ith axis

$$T(t) = (\rho C_d A R/2)V^2(t) \tag{3}$$

where C_d is the drag coefficient, A is the antenna projected area, R is the length of the moment arm from the center of pressure to the antenna axis, all functions of the aspect angle φ, ρ is the density of air and V is the free stream wind velocity. Defining a wind torque coefficient for the ith axis

$$C_w = \rho C_d A R/2, \tag{4}$$

$T(t)$ may then be written

$$T(t) = C_w(\varphi)V^2(t). \tag{5}$$

$K_{ij}(s)$ is a complicated function of antenna and servo parameters, few of which have been evaluated for the antenna designs treated in

this paper. Consequently, we cannot calculate tracking error as a function of wind velocity for these antennas. However, because one of the most significant characteristics of an all-weather satellite communications antenna is its tracking performance in wind and because the values of C_w presented in this paper have little meaning unless related to antenna performance, it seems reasonable to make sufficient simplifying assumptions to reach an approximate functional relationship between C_w, V and tracking error such that on the basis of incomplete, preliminary design data, an estimate of wind performance can be made. We then assume:

(1.) Wind induced error about the vertical axis of the antenna is much greater than about the elevation or inclined axis. (This is a reasonable assumption because the wind torque and structural compliance are both much greater about the vertical axes.)

(2.) The structure is much more compliant about its rotational axes than about any other axis.

(3.) The foregoing is true and we need consider only the vertical axis and can neglect the coupling between axes; (2) can then be written

$$K_w(s) = [T(s)/\theta_a(s)] \tag{2'}$$

where $K_w(s)$ is the overall azimuth stiffness function

$\theta_a(s)$ is the azimuth pointing error.

Representing this simplified system as shown in the Appendix, the variance of the pointing error in the autotrack mode of operation is given by

$$\overline{\theta_a^2}(t) = \frac{8C_w^2V_o^2V_1^2}{\omega_c}\left[\frac{1}{2\pi j}\int_{-j\infty}^{j\infty}\frac{ds}{(1+s/\omega_c)K_w(s)(1-s/\omega_c)K_w(-s)}\right] \tag{6}$$

where V_o is the average wind velocity and V_1 is the standard deviation of the time variant component of wind velocity.

It is seen that the standard deviation of the antenna azimuth pointing error is proportional to the product $C_wV_oV_1$, but the integral cannot be evaluated because $K_w(s)$ involves unknown structural and servo parameters. Examination of the form of $K_w(s)$, however, reveals that the dominant term at low frequency is the azimuth structural stiffness K_a.

If $K(s)$ can be considered to be constant and equal to K_a over a frequency band extending well beyond the wind cut-off frequency, then the standard deviation of the antenna pointing error in the autotrack mode of operation is given by

$$\sigma(\varphi) = \frac{2V_oV_1C_w(\varphi)}{K_a} \tag{7}$$

This equation, though greatly oversimplified is nevertheless useful in preliminary antenna design and provides a basis for comparison of different configurations. Further discussion of this equation and an indication of its applicability appear in the Appendix.

The quantity C_w must be determined therefore, if the tracking performance in wind is to be estimated for a proposed antenna configuration. Similarly, the antenna overturning stability may be described in terms of an overturning moment coefficient $C_{wo}(\varphi)$ which may be defined for any convenient axes. The numerical values of C_w and C_{wo} for the most unfavorable orientation of an antenna may be used as a figure of merit for comparison of various antenna configurations.

II. MODEL TEST THEORY

Aerodynamic force and torque coefficients can be determined experimentally by measuring the forces and moments induced by fluid flow around a scale model. In such model tests it is important to maintain the same type of fluid flow as encountered in the full size antenna structure. The antenna configurations considered in this paper, particularly the triply-folded horn-reflector, have been designed to approximate a rounded body such as a sphere in so far as is practical. To the extent that this design effort was unsuccessful, the antenna is a bluff body; that is, for Reynolds Numbers N_{re} larger than about 1000, flow will separate at the largest cross section of the body and the drag coefficient C_d will be approximately constant at some fairly high value. However, to the degree that the faired contours of the antenna approximate a rounded body, a marked transition in the character of flow will be exhibited as the Reynolds Number is increased beyond some critical value, N_c. The critical Reynolds Number depends not only upon the shape of the body but also upon upstream turbulence. The greater the turbulence of the approaching fluid, the lower will be the value of N_c. For lack of better information we may assume for the present that N_c will be somewhere between 10^5 and 5×10^5, the approximate transition range for spheres and cylinders.

At fluid speeds in the approximate range of Reynolds Numbers from 10^3 to 10^5 the boundary layer on the upstream side of the body is laminar because of its extreme thinness and remains in contact with the surface of the body as far downstream as approximately the largest cross section of the body. Here the boundary layer fluid enters a region where the pressure increases in the direction of flow. The adverse pressure gradient forces the fluid away from the surface, i.e., the main stream separates from the body, creating a large wake area. As the

Reynolds Number is increased beyond N_c, the point of separation abruptly shifts farther back, reducing the wake area and thereby reducing the drag coefficient, C_d. The shift in separation point is due to the transition in the boundary layer from laminar to turbulent flow. The turbulent boundary layer, because of its increased momentum travels somewhat farther along the surface before the pressure causes it to separate again. There are then, two distinct regimes of flow, defined as

$$\text{subcritical} \quad -10^3 < N_{re} < 10^5$$

$$\text{supercritical} - N_{re} > 5 \times 10^5$$

where N_{re} is the Reynolds Number and the limits given above are estimates.

$$N_{re} = VD/\nu;$$

D is a characteristic linear dimension of the body and ν is the kinematic viscosity of the fluid.

Experiment has shown that the drag coefficient of a rounded or streamlined body in supercritical flow is relatively constant and is considerably lower than for subcritical flow. Air flow around a faired antenna structure of 2500 square foot aperture size is estimated to be supercritical at wind speeds higher than about 10 miles per hour. Since we are concerned with the effects of high wind velocity on antenna performance and survival, it is the values of C_w and C_{wo} determined for supercritical flow that are of interest.

Because the same type of flow encountered by the actual antenna must be reproduced in a scale model test and since the flow is in the supercritical region, i.e., $N_{re} > 5 \times 10^5$, the product of fluid speed and the model characteristic dimension must be >81 ft^2 per second if the test is to take place in air at 20°C. For a complex antenna shape, the dimension D, the smallest characteristic dimension of any area of the antenna whose drag force contributes significantly to the total induced forces or torques acting on the antenna is difficult to estimate accurately and may be much smaller than l the largest dimension of the antenna. Moreover, as indicated above, the critical Reynolds Number cannot be accurately estimated. Therefore, in selecting a model test facility, care must be taken to insure that the facility has the capability of providing flow velocity well in excess of the estimated requirement. The difficulty of testing in air may be shown by assuming a model size $l = 12$ inches and $l/D = 5$. The minimum required air velocity is then 275 mph or Mach .34 and the uncertainty in N_c and D makes the requirement for even higher speed a possibility. At these Mach

numbers, the third term in the power series expansion for the pressure ratio p/p_0 becomes significant with respect to the second term. That is, air at the Mach numbers likely to be encountered in wind tunnel testing must be considered a compressible medium and appropriate corrections must therefore be made to the test data before it can be applied to the full scale antennas.

If water is used as the test medium rather than air, the velocity requirement is reduced by a factor of 15 because of water's lower kinematic viscosity. The requirement for supercritical flow then becomes $VD >$ 5.4 ft^2 per second. Because air at a free stream velocity of 100 mph can be considered to be incompressible (with negligible error as demonstrated above) the scaling from air to water presents no difficulties. The hydrodynamic test facilities of the Davidson Laboratory of Stevens Institute of Technology, Hoboken, New Jersey, provide the required flow conditions, and were used in making the measurements to be described.

III. MODELS AND INSTRUMENTATION

Of the three models of all-weather satellite communications antenna concepts that were constructed and tested, two are reported in this paper. These are shown in Figs. 1 through 3. Model "A" represents the triply-

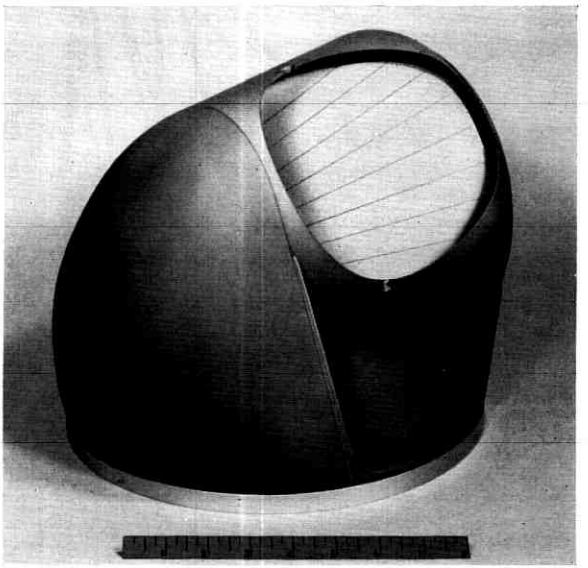

Fig. 1 — Model "A" with aperture cover removed.

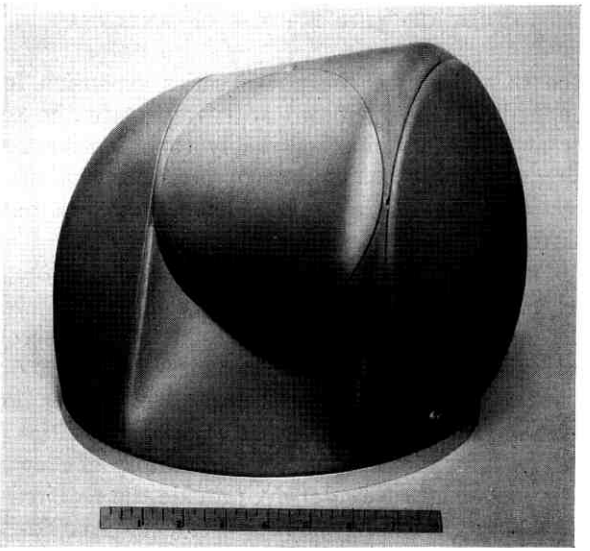

Fig. 2 — Model "A."

Fig. 3 — Model "B," the open cassegrain antenna.

folded horn-reflector antenna[5] and Model "B" is the open cassegrain antenna.[6] An aperture cover, of the same contour as the elevation drum, was installed on Model "A" in some of the test runs to determine the aerodynamic value of such a device.

Torque about the vertical axis, overturning moments about the antenna base and total drag were measured by means of a standard five component balance supplied by the testing facility. The inverted model mounted on the balance, was supported below the surface of the water. A large plate, simulating the ground plane at the surface of the water, was supported independently so that only the hydrodynamic forces acting on the model would be sensed by the balance. Fig. 4 shows a model mounted for testing.

The models also contained internal torque balances using constant stress cantilevers and linear differential transformers as seen in Figs. 5, 6, and 7. Fig. 8 defines the orientation of antenna axes with respect to flow direction.

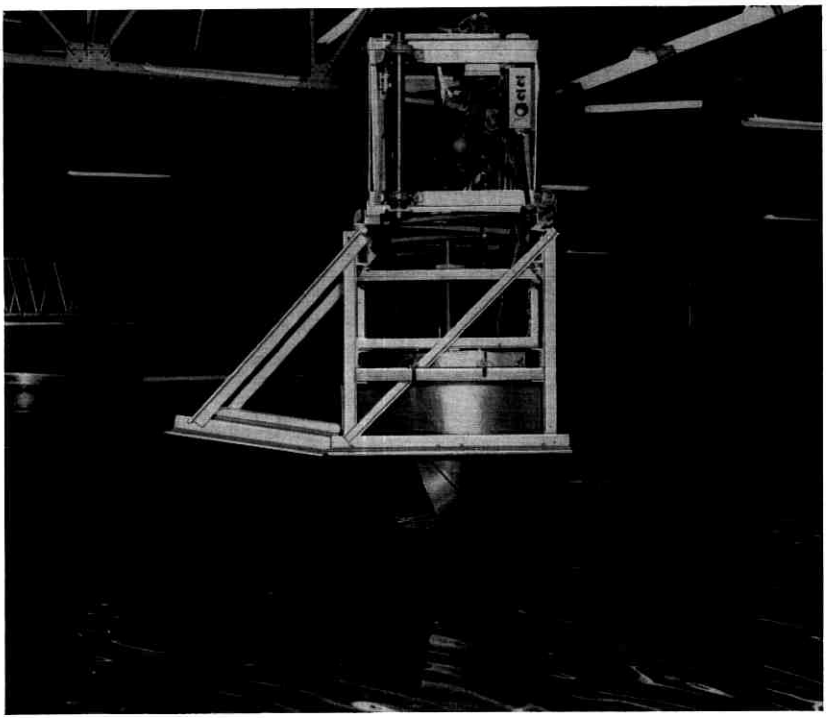

Fig. 4 — Model "B" about to be lowered into the tank for test run.

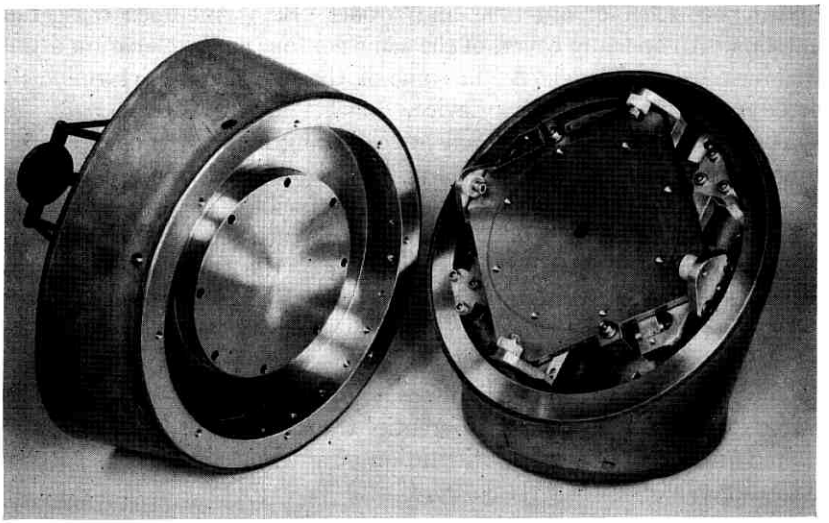

Fig. 5 — Model "B," showing internal uncoupled three component balance.

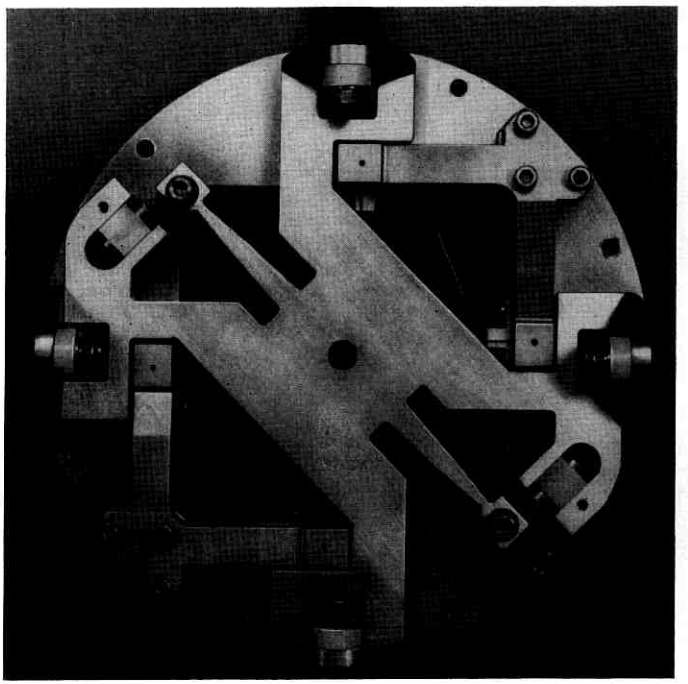

Fig. 6 — Uncoupled three component balance for Model "B."

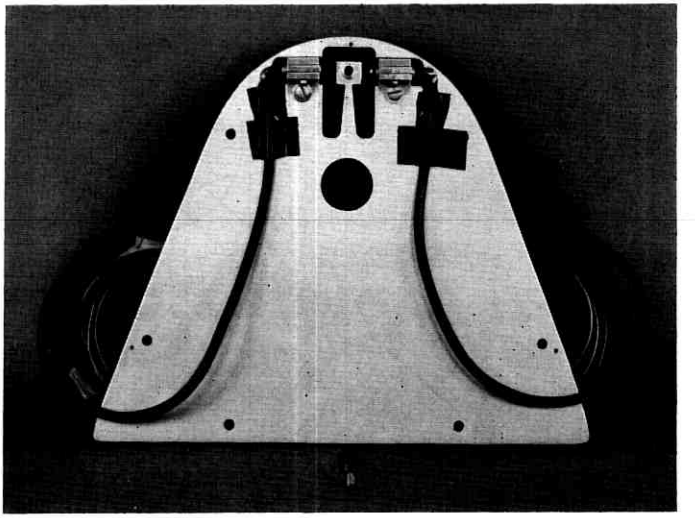

Fig. 7 — Elevation balance plate assembly for Model "A."

IV. TEST PROCEDURE

The required flow velocity was established by towing the model through the test tank at various constant velocities to obtain values for the drag coefficient, C_d . From a plot of C_d versus V, the transition speed was determined. The criterion used for establishing the appropriate test velocity for each model was a constant drag coefficient with increasing Reynolds Numbers.

Data runs were then made at the appropriate velocity for various orientations of the model. The outputs of the external and internal torque balances were recorded for each run. Fig. 9 shows a data run in progress.

V. TEST RESULTS

The experimental results for the two antenna concepts are presented in Fig. 10 through 19 as plots of wind torque coefficients versus wind azimuth aspect angle for various orientations of the elevation or inclined (slant) axis. The most significant quantity for both antennas is the wind torque coefficient about the vertical axis. Wind induced torque about the elevation (or inclined) axis of the two models is relatively small.

Precise values of weight and center of mass are not available during the preliminary design of an antenna, but on the basis of preliminary

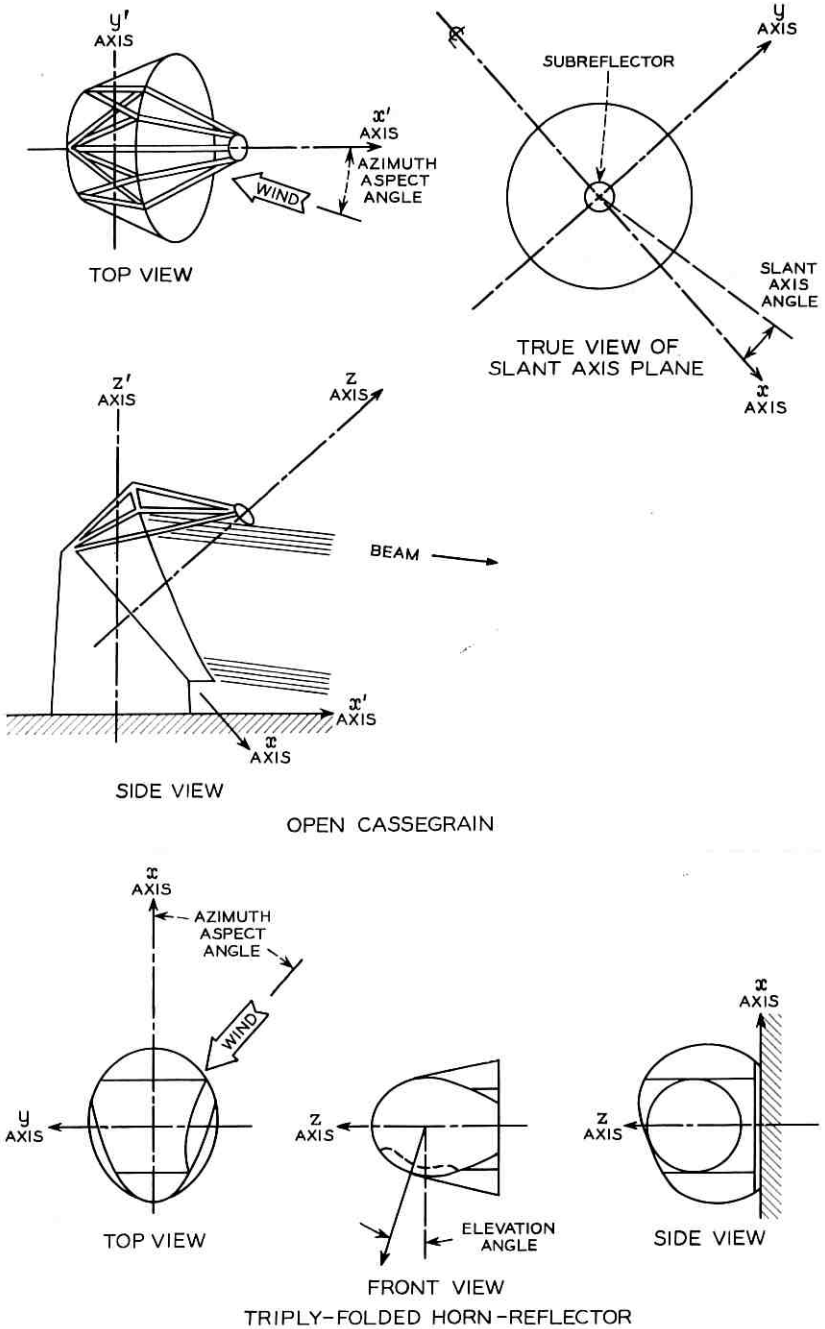

Fig. 8 — Axis orientation.

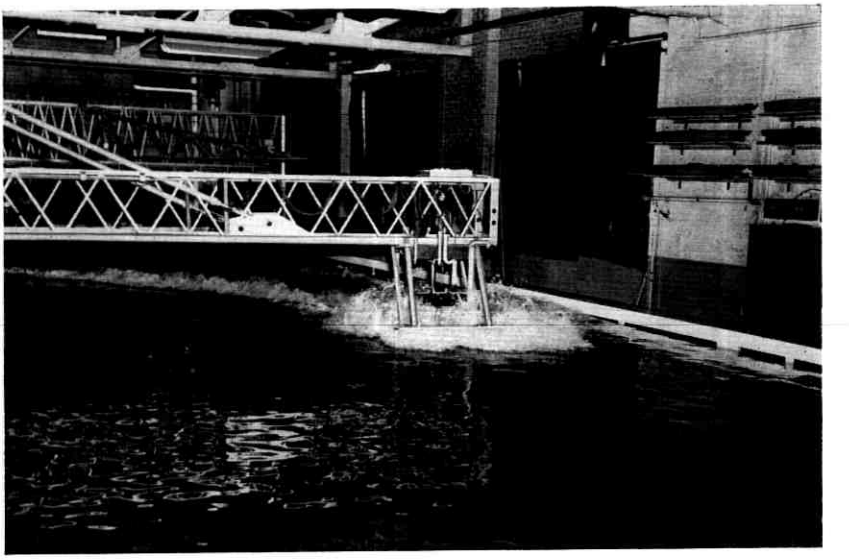

Fig. 9 — Data run.

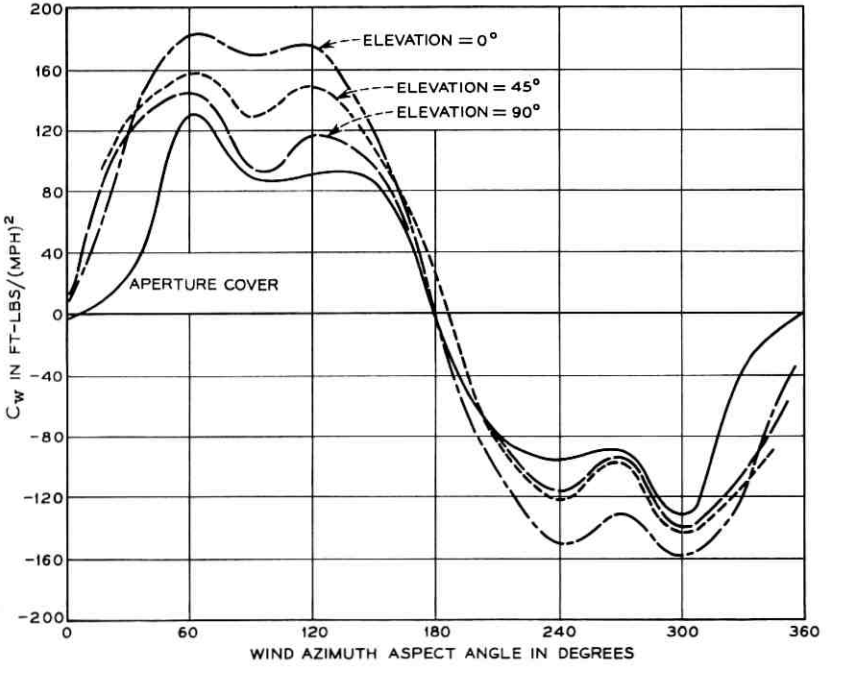

Fig. 10 — Azimuth torque — Model "A."

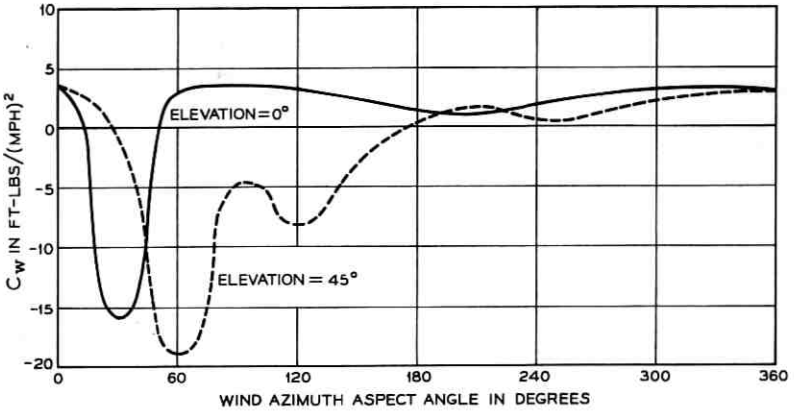

Fig. 11 — Elevation torque — Model "A."

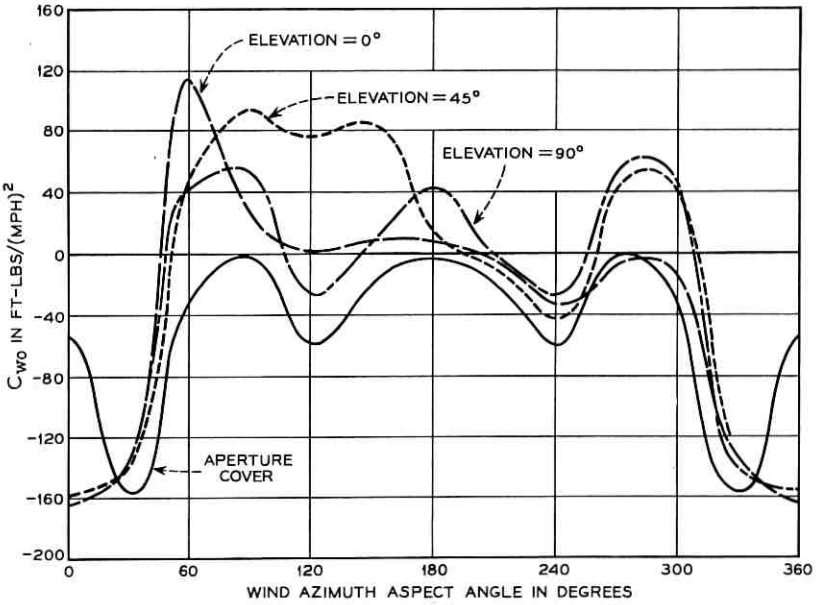

Fig. 12 — Overturning moment about transverse axis (Y) — Model "A."

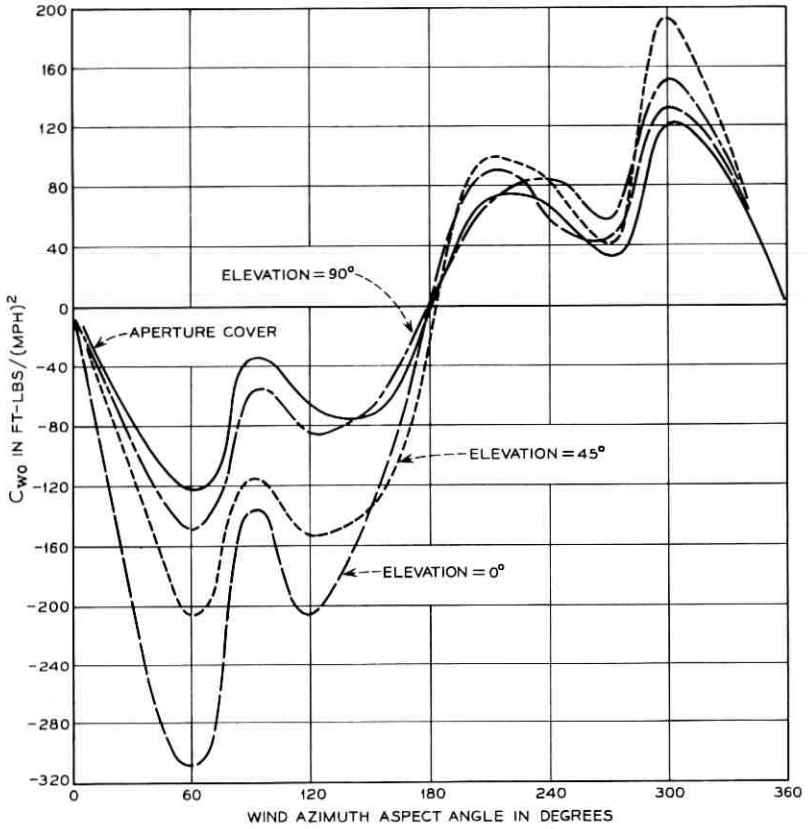

Fig. 13 — Overturning moment about longitudinal axis (X) — Model "A."

estimates, the maximum wind overturning moments and the resisting stability moments can be fairly accurately predicted. These have been calculated for the two models here considered and are listed in Table I for wind velocity of 100 mph.

To make the plotted test data meaningful in terms of antenna performance, some comparative values may be computed by assuming that the pointing error about the vertical axis is given by (7), that the standard deviation, V_1, of the wind velocity is 50 per cent of the average value, V_o, and by choosing a representative value of the stiffness, K_a. For the Andover Telstar antenna, $K_a = 2 \times 10^9$ ft-lbs per radian. It is conservative then to assume the same value for the relatively more compact and rigid structures considered in this paper, provided that a

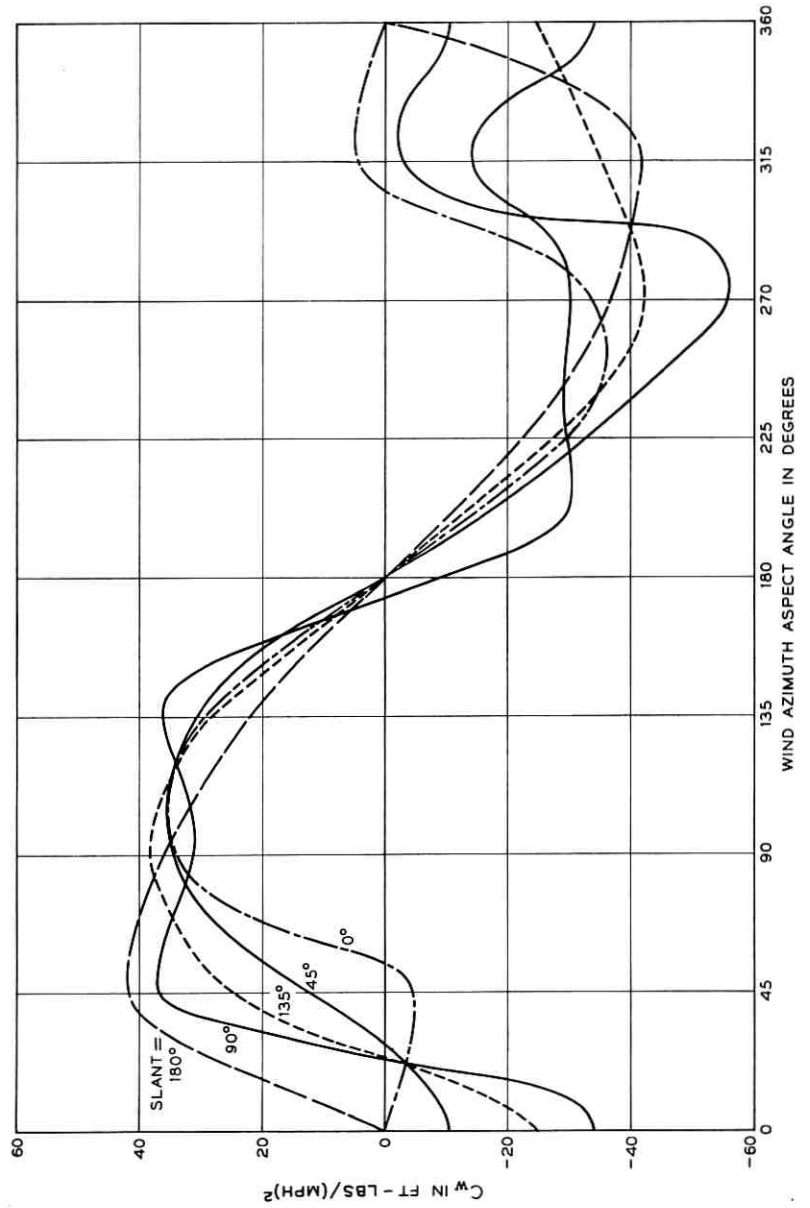

Fig. 14 — Open cassegrain — wind torque about vertical axis.

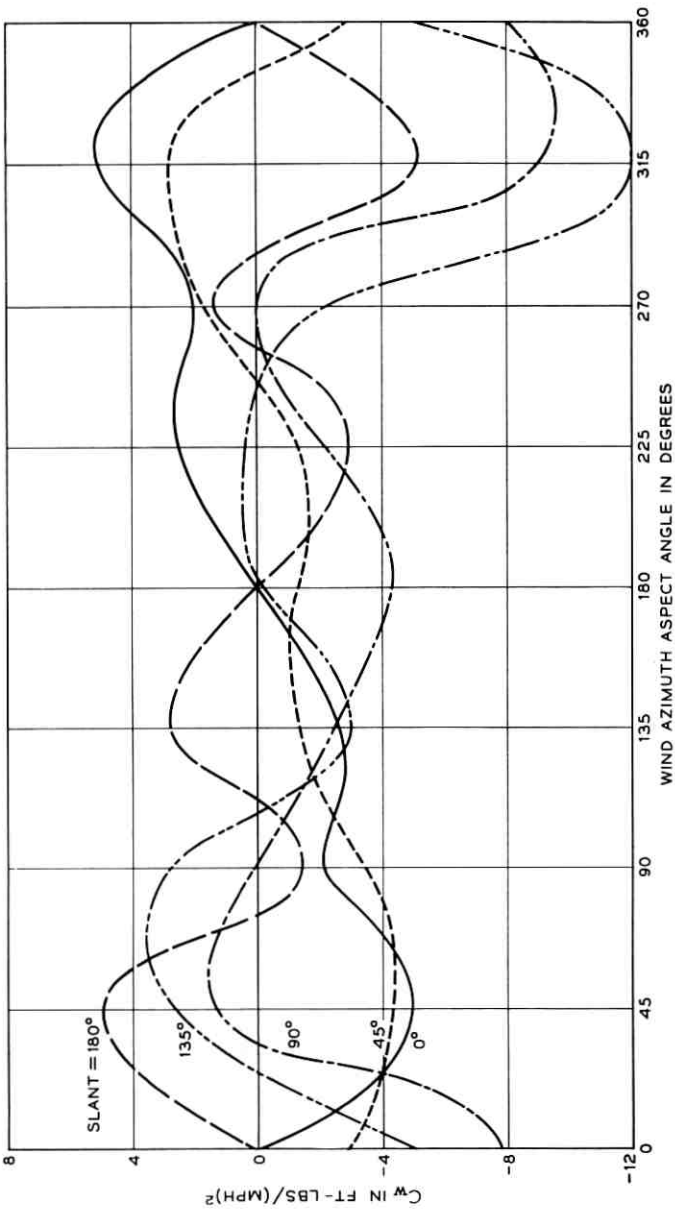

Fig. 15 — Open cassegrain — wind torque about slant axis.

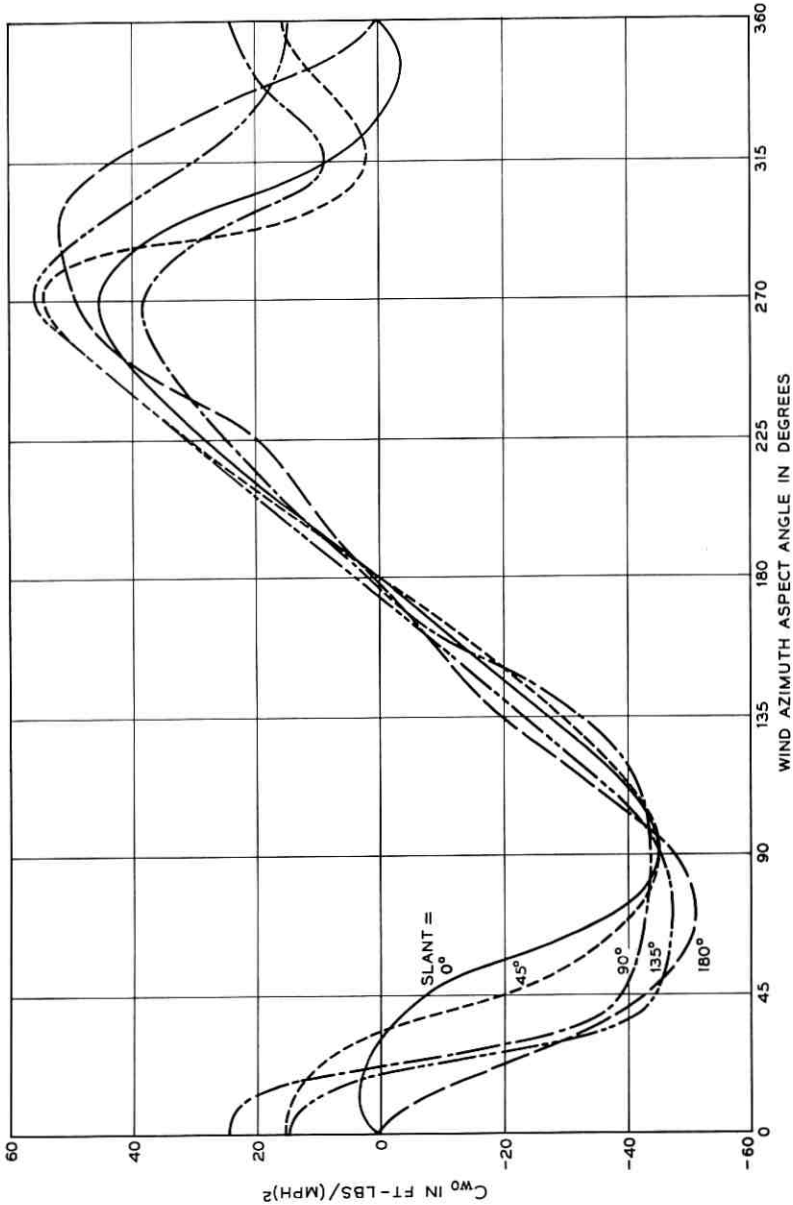

Fig. 16 — Open cassegrain — reaction moment on slant rail about x-axis due to wind on reflector section.

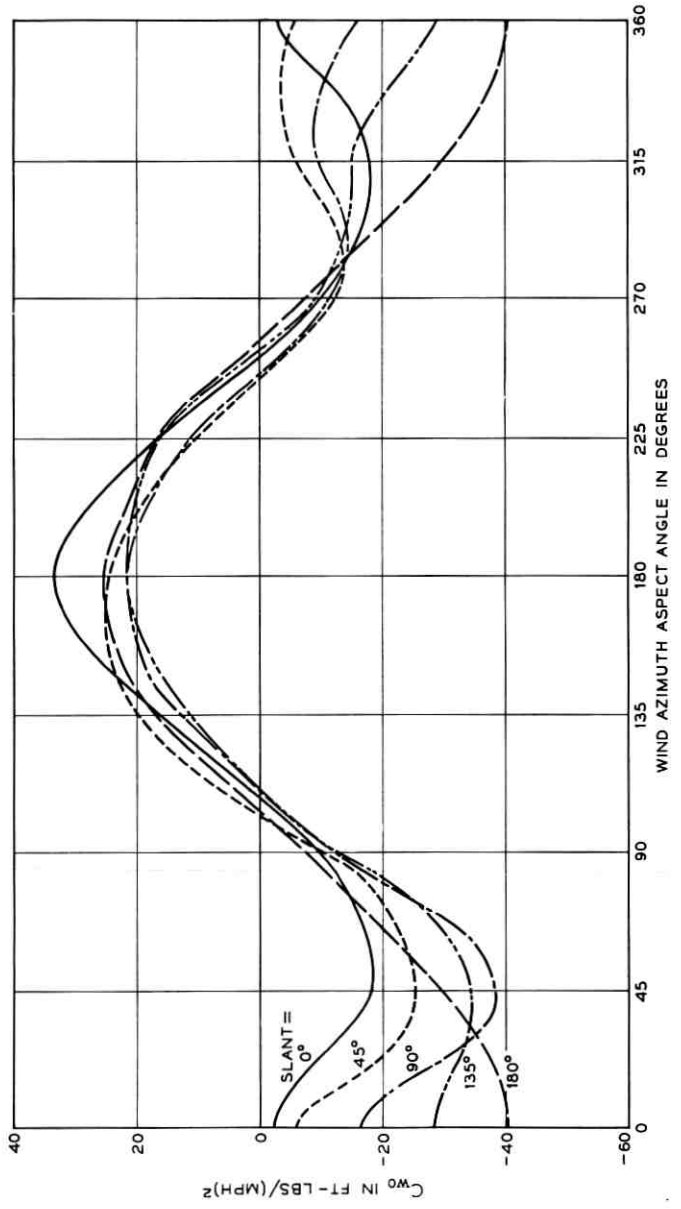

Fig. 17 — Open cassegrain — reaction moment on slant rail about y-axis due to wind on reflector section.

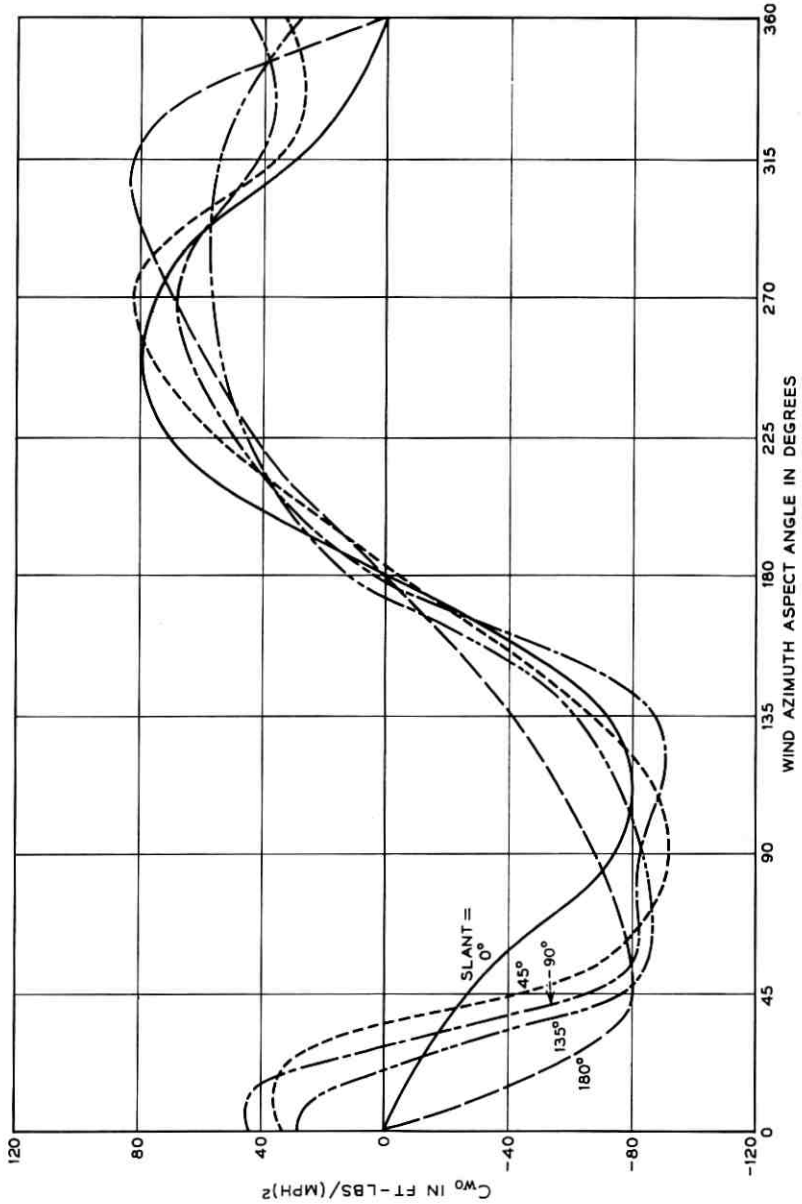

Fig. 18 — Open cassegrain — wind overturning moment about x'-axis in base.

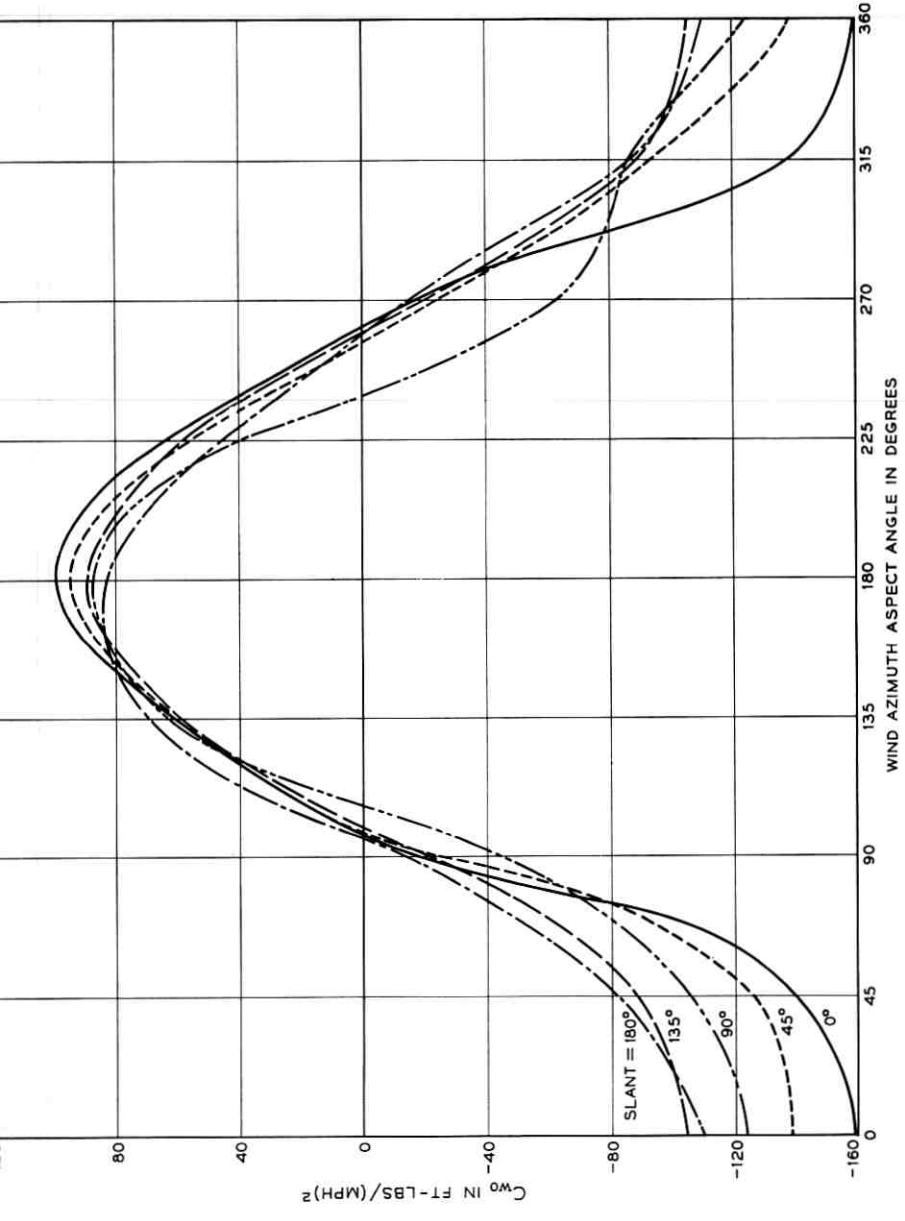

Fig. 19 — Open cassegrain — wind overturning moment about y'-axis in base.

TABLE I

WIND EFFECTS ON ALL-WEATHER SATELLITE COMMUNICATIONS ANTENNAS

		Overturning Moment at 100 mph ft-lbs	Stability Moment ft-lbs	Overturning Safety Factor	C_w for Vertical Axis ft-lbs/(mph)^2	Wind* Speed to Stall Antenna Drives, mph	Wind Speed‡ for 0.01° Pointing Error mph
Triply-folded horn-reflector antenna (2250 ft^2 aperture)**	With aperture cover or in "stow" position	1.56×10^6	24×10^6	15	132	95	$V_o = 51$ $V_1 = 25.5$
	Without aperture cover	3.1×10^6	19×10^6	6	184	80	$V_o = 43$ $V_1 = 21.5$
Open cassegrain antenna (2460 ft^2 aperture)**	Entire antenna	1.6×10^6	5.6×10^6	3.5	57.5	71	$V_o = 47$† $V_1 = 23.5$
	Reflector section only	$.685 \times 10^6$	1.7×10^6	2.5			

* Two 25-hp drives assumed. Available drive torque for triply-folded antenna is 1.2×10^6 ft.-lbs. Open cassegrain design assumes higher angular velocity, limiting available drive torque to about 2.9×10^5 ft.-lbs.

** The aperture sizes presented were arbitrarily selected and no electrical equivalence is indicated.

† See Appendix, p. 1365.

‡ V_o is the average wind speed; V_1 is the standard deviation of the variable component of wind speed.

correction is made to account for the open cassegrain subreflector support (see Appendix).

The wind velocity at which the standard deviation in antenna pointing error for the vertical axis equals 0.01° is given in Table I for the two antennas tested. The stall torque wind speed is also presented for each antenna. This is the wind speed at which wind induced torque about the vertical axis equals available drive torque and is based on the assumption that two 25-hp motors provide the azimuth drive. It will be noticed that the stall torque wind speed for the open cassegrain antenna is lower than might be expected from a comparison of wind torque coefficients. This results from the assumption of a lower drive gear ratio which would permit relatively higher antenna angular velocity and possible elimination of the zenith "tracking dead-zone" which is characteristic of all antennas with a vertical axis.

VI. CONCLUSIONS AND SUMMARY

The test results indicate that the wind torques induced about the rotational axes of the antennas are not excessive, and that with sufficient structural rigidity, tracking performance will be adequate. As pointed out above, however, the variation in instantaneous wind velocity over the dimensions of the antenna has been neglected. If an appreciable variation were found to exist the wind torque input to an antenna might be altered significantly, depending upon the correlation between the time and spatial variation in wind velocity.

Both antenna configurations are inherently stable in wind up to 100 mph.

VII. ACKNOWLEDGMENTS

The author wishes to thank Dr. P. Ward Brown of the Davidson Laboratory for his assistance in organizing the test program. The work of Dr. F. A. Russell and Dr. W. H. W. Ball on wind loading for the Andover antenna was an inspiration and aid in the hydrodynamic study. The success of this program was due in large measure to Messrs. J. H. Cave and H. W. Boschen whose assistance in the design and construction of the test models proved invaluable.

APPENDIX

Wind-Induced Pointing Error in the Autotrack Mode

The azimuth pointing error is given by

$$\theta_a(s) = [T(s)/K_w(s)]$$

where $T(s)$ is the wind torque disturbance, and

$K_w(s)$ is the overall azimuth system transfer function.

Representing the system in autotrack mode by a mechanical schematic (Fig. 20(a)) and an electrical analog (Fig. 20(b)) the transfer function can be written

$$K_w(s) = \left[R_1 + \frac{K_a K_g A F_s - K_a{}^2 R_3}{R_2 R_3 - K_g{}^2} \right] \tag{8}$$

where $R_1 = s^2 J_a + sB_a + K_a$

$R_2 = s^2 J_c + sB_c + K_a + K_g$

$R_3 = s^2 J_m + sB_m + K_g$

A is the servo torque gain

$F_s(s)$ is the servo shaping function

K, J and B are the stiffness, inertia and resistance respectively, and

the subscripts refer to the motion of the horn or reflector section (a), the motion of the cradle or pedestal (c), the motion of the gear boxes (g), and the motion of the drive motors (m).

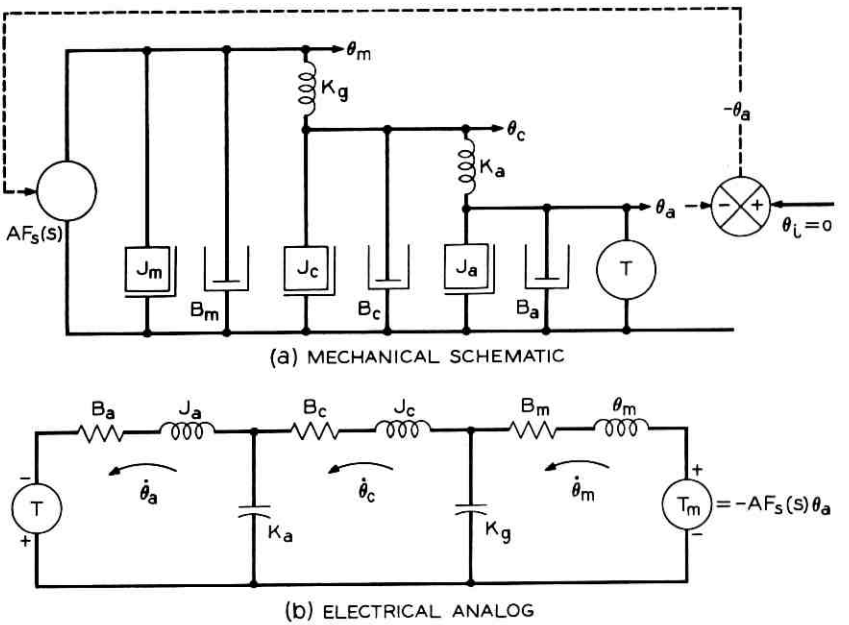

Fig. 20 — (a) Mechanical schematic, (b) electrical analog.

The time-variant input torque $T(t)$ is given by

$$T(t) = C_w V^2(t) = C_w[V_o^2 + 2V_o V_1(t) + V_1^2(t)]. \quad (9)$$

Noting that the constant term $C_w V_o^2$ produces no error in the auto-track mode, and assuming that $V_1^2(t) \ll 2V_o V_1(t)$, the transfer function from gust velocity to torque is $2C_w V_o$. The wind gust spectral power density is assumed to be of the form

$$\Phi_{vv}(s) = \frac{V_1^2 \omega_c}{\pi(\omega_c^2 - s^2)}. \quad (10)$$

For linear systems, the power density function of the output is equal to the power density function of the input multiplied by the square of the transfer function. The wind torque spectral density is therefore,

$$\Phi_{TT}(s) = 4C_w^2 V_o^2 \Phi_{vv}(s) \quad (11)$$

and the pointing error spectral density is

$$\Phi_{aa}(s) = \frac{4C_w^2 V_o^2 \Phi_{vv}(s)}{|K_w(s)|^2}. \quad (12)$$

The pointing error variance is then

$$\overline{\theta^2}_a(t) = 2\pi \frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \frac{\Phi_{TT}(s)}{|K_w(s)|^2} ds. \quad (13)$$

Substituting for $\Phi_{TT}(s)$,

$$\begin{aligned} \overline{\theta^2}_a(t) &= \frac{4C_w^2 V_0^2 V_1^2}{\pi\omega_c} 2\pi \left[\frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \frac{ds}{(1 - s^2/\omega_c^2)|K_w(s)|^2}\right] \\ &= \frac{8C_w^2 V_o^2 V_1^2}{\omega_c} \left[\frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \frac{ds}{(1 - s^2/\omega_c^2)|K_w(s)|^2}\right]. \end{aligned} \quad (14)$$

Since, for $s = j\omega$,

$$|K_w(s)|^2 = K_w(j\omega)K_w^*(j\omega) = K_w(j\omega)K_w(-j\omega)$$

$$|K_w(s)|^2 = [K_w(s)K_w(-s)],$$

we may write finally:

$$\overline{\theta^2}_a(t) = \frac{8C_w^2 V_o^2 V_1^2}{\omega_c} \left[\frac{1}{2\pi j} \int_{-j\infty}^{j\infty} \frac{ds}{(1 + s/\omega_c)K_w(s)(1 - s/\omega_c)K_w(-s)}\right]. \quad (15)$$

The standard deviation of the error is given by the square root of the variance.

The integral

$$\frac{1}{2\pi j}\int_{-j\infty}^{j\infty}\frac{ds}{(1+s/\omega_2)K_w(s)(1-s/\omega_c)K_w(-s)}$$

can be evaluated analytically on the basis of the residue theory, as well as either graphically, or numerically.[4] A computer program has been written and used for evaluating this integral with TSX-1 antenna servo parameters. The results obtained are peculiar to that antenna and are therefore not applicable to the antennas under discussion here. Having no precise values for the terms contained in $K_w(s)$, we can nevertheless obtain an estimate of the performance of an antenna in wind by assuming the overall antenna transfer function to be constant over the frequency band extending to perhaps a decade beyond ω_c.

Considering the transfer function, (8), and the schematics, one can see that the term R_1 is the contribution of the antenna compliance and the inertia of the reflector section. If $\dot{\theta}_c = 0$, $K_w(s) = R_1(s)$, and at low frequency, $R_1(s) \approx K_a$.

The magnitude of the other term in $K_w(s)$ depends upon the torque gain, A, and shaping function $F_s(s)$. With error integration in the shaping, AF_s is large for small ω and it is reasonable to assume that the real part of the term

$$\frac{K_aK_gAF_s - K_a^2R_3}{R_2R_3 - K_g^2}$$

is positive and the imaginary part is small compared with K_a over the desired frequency band. With these assumptions we conclude that K_w is positive-real and that $|K_w| \geqq K_a$ in the frequency range of interest for the calculation of wind induced pointing error.

For $K_w(s)$ then, we substitute K_a in (15), yielding

$$\overline{\theta^2}_a(t) = \frac{8C_w^2V_o^2V_1^2}{K_a^2\omega_c}\left[\frac{1}{2\pi j}\int_{-j\infty}^{j\infty}\frac{ds}{1-s^2/\omega_c^2}\right]$$

$$= \frac{4C_w^2V_o^2V_1^2}{K_a^2},$$

so that

$$\sigma_a = \frac{2C_wV_oV_1}{K_a}.$$

In the case of the TSX-1 antenna, this approximation coincides with the results obtained through computer evaluation of the integral (using

measured antenna parameters) for $\omega_c = 0.6$ rad per second and holds to ± 50 per cent for $0.125 \leqq \omega_c \leqq 1.85$. The correspondence can be expected to be at least as good for the relatively stiffer and lighter all-weather antennas discussed in this paper.

In the case of the open cassegrain antenna however we must take into account the compliance of the subreflector support structure when computing the wind induced pointing error. Flow around the tubular support members is in the transition region between 10 and 50 mph wind speed, so we conservatively assume the flow to be always subcritical. Structural analysis has shown that under these conditions, C_W/K due to subreflector motion is 50×10^{-9} radians per $(\text{mph})^2$ referred to the antenna pointing vector. Assuming for the rest of the structure $K_a = 2 \times 10^9$ ft-lbs per radian and $C_W = 56$ ft-lbs per $(\text{mph})^2$, the adjusted C_W/K_a is 78×10^{-9} radians per $(\text{mph})^2$.

REFERENCES

1. Titus, J. W., Wind Induced Torques Measured on a Large Antenna, NRL Report 5549, Dec., 1960.
2. Lumley, J. L., and Panofsky, H. A., The Structure of Atmospheric Turbulence, Interscience Monograph, 1964.
3. Barton, D. K., RCA Final Report, Instrumentation Radar AN/FPS-16(XN-1), Evaluation and Analysis of Radar Performance, ASTIA Report No. 212125, March 19, 1959.
4. Newton, G. C., Gould, L. A., and Kaiser, J. F., *Analytic Design of Linear Feedback Controls*, Wiley, 1957.
5. Giger, A. J., and Turrin, R. H., The Triply-Folded Horn Reflector: A Compact Ground Station Antenna Design for Satellite Communications, B.S.T.J., this issue, pp. 1229–1253.
6. Denkmann, W. J., Geyling, F. T., Pope, D. L., and Schwarz, A. O., The Open Cassegrain Antenna: Part II. Structural and Mechanical Evaluation, B.S.T.J., this issue, pp. 1301–1319.

Autotrack Control Systems for Antenna Mounts with Non-Orthogonal Axes

By W. L. NELSON and W. J. COLE

(Manuscript received May 17, 1965)

The use of non-orthogonal, or conic, mounts for steerable antennas introduces some control system design problems not present in the more conventional orthogonal mounts. These problems result from both the geometrical and the mechanical cross-coupling which occurs between the two non-orthogonal axes of motion.

This paper presents a general analysis and design of the control system for the open cassegrain antenna which can be readily applied to other non-orthogonal antenna structures. The form of the feedback controller for approximately non-interacting control of each axis is developed. Also described is a supplementary control strategy for providing tracking near the zenith region without excessively high slewing rates.

A computer simulation of the system has verified the basic control strategy for non-orthogonal mounts and established the feasability of operating compact antenna structures such as the open cassegrain design under severe wind conditions without a radome.

I. GENERAL SYSTEM DESCRIPTION

While the general control system design methods developed in this study apply to any antenna mount using two non-orthogonal axes, the specific structure considered throughout this paper is the slant-mounted open-cassegrain antenna.[1] In this structure, the antenna beam tracks the target by controlled rotational motion about the inclined and the vertical axes shown in Fig. 1. While motion about the vertical axis produces true azimuth motion of the beam, the motion about the inclined axis generates a combination of azimuth and elevation motion of the beam. Further cross-coupling between the azimuth and elevation tracking channels is introduced by unavoidable mechanical coupling of motion between the two drive axes.

Fig. 2 is a general block diagram of the tracking control system. The

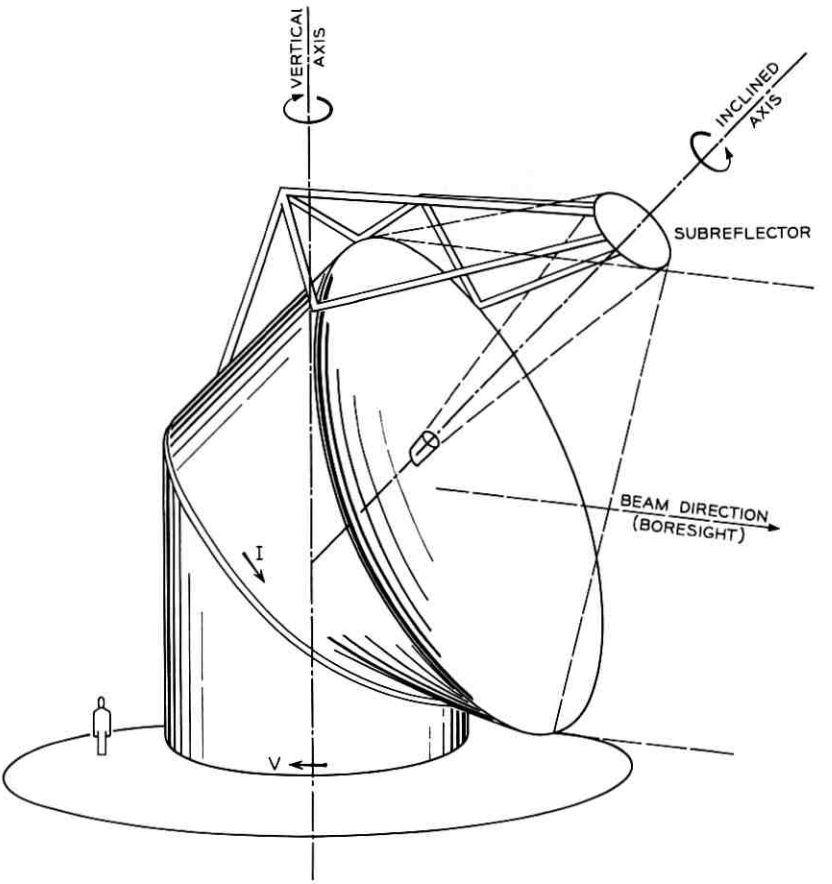

Fig. 1 — Open cassegrain antenna with two-axis conic mount structure showing simplified subreflector structure.

pointing error is resolved for convenience into the standard azimuth and elevation angle errors (this is a conceptual, not physical, portion of the system). In the tracking of active repeater satellites, the error signals are derived from the waveguide mode detector receiving the satellite beacon signal.[2] The horizontal and vertical error signals, ϵ_h and ϵ_v, at the output of the error detector are related to the azimuth and elevation errors by

$$\begin{aligned} \epsilon_h &= K_h \, (\cos E)(A_r - A) \\ \epsilon_v &= K_v(E_r - E) \end{aligned} \tag{1}$$

where K_h and K_v are the detector gains, A_r and E_r are the reference azimuth and elevation angles, respectively, of the satellite, and A and E are the controlled azimuth and elevation angles of the beam axis (electrical boresight) of the antenna.

All of the equations of motion of the antenna drive system and physical structure (considered in detail in Section III) can be represented here by the single nonlinear differential equation,

$$\frac{d\mathbf{W}}{dt} = \mathbf{F}(\mathbf{W},\mathbf{u}) \tag{2}$$

where $\mathbf{W}$ is the state vector containing as components those variables necessary to adequately represent the dynamics of the antenna system, and $\mathbf{F}$ is the vector-valued function relating the time derivative of the state to itself and to the control vector, $\mathbf{u}$ (with components u_1 and u_2).

From the error signals (1), as well as from feedback signals derived from the components of the state vector, $\mathbf{W}$, the controller must generate the two antenna drive signals, u_1 and u_2, which control the antenna angles, V and I, in such a way as to reduce the tracking error and keep the antenna beam automatically "locked-on" to the satellite. Because of the complex, nonlinear, multivariate nature of this system, the design of the controller cannot be achieved by conventional analytic design methods. A preliminary controller design, based on a linear approximation of the system dynamics, together with the appropriate coordinate transformations and supervisory control logic, is discussed in Section IV. The final design will be evolved from this preliminary design through an

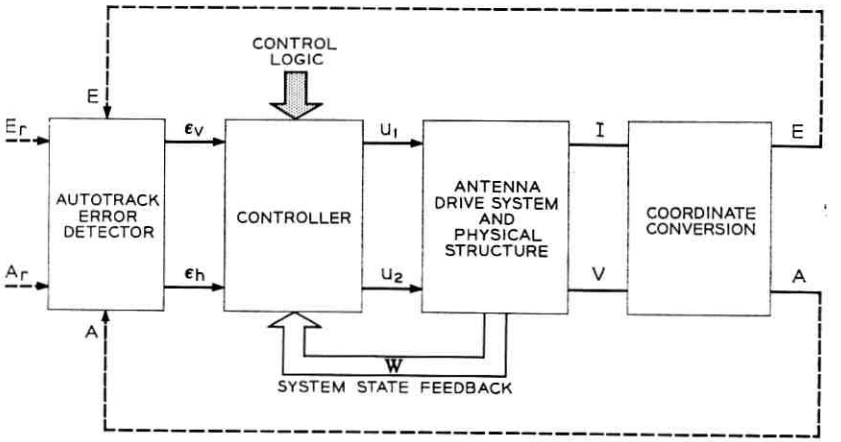

Fig. 2 — Preliminary block diagram for antenna control system.

accurate computer simulation of the overall system, its inputs, and its environment.

II. CONIC MOUNT CHARACTERISTICS

To gain an initial understanding of the tracking requirements on this system, the basic transformations between the antenna angles (I,V) and the tracking angles (E,A) are needed. Some of this is similar to a study on conic mounts by Norton,[3] and his notation is used here.

Fig. 3 shows the geometry of the conic mount, in particular the structural design angles (α,β), the inclined- and vertical-axis angles (I,V), and the elevation and azimuth angles (E,A).

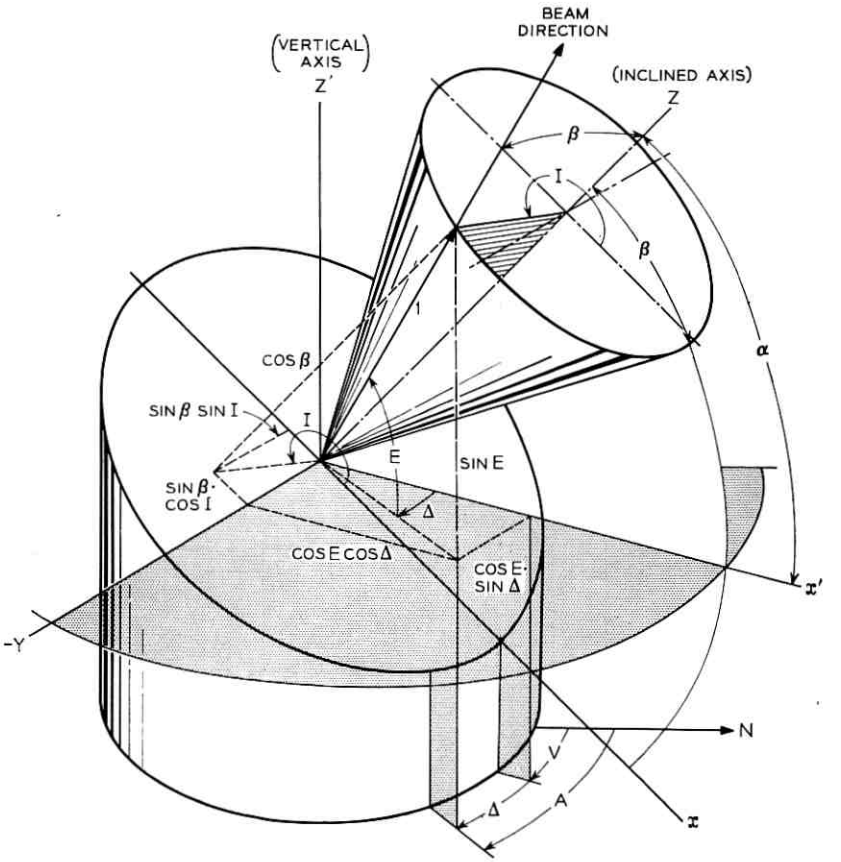

Fig. 3 — Conic mount geometry.

Since V produces only azimuth motion, the only coordinate conversions needed are those giving the azimuth angle Δ and the elevation angle E produced by I. Fig. 3 indicates that these angles are related by rotation about the y-axis through an angle $(90° - \alpha)$, which corresponds to the rectangular coordinate transformation:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \sin\alpha & 0 & \cos\alpha \\ 0 & 1 & 0 \\ -\cos\alpha & 0 & \sin\alpha \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{3}$$

For a unit radius, the spherical coordinate equivalents of these rectangular coordinates are,

$$\begin{pmatrix} \cos E \cos\Delta \\ -\cos E \sin\Delta \\ \sin E \end{pmatrix} = \begin{pmatrix} \sin\alpha & 0 & \cos\alpha \\ 0 & 1 & 0 \\ -\cos\alpha & 0 & \sin\alpha \end{pmatrix} \begin{pmatrix} \sin\beta\cos I \\ \sin\beta\sin I \\ \cos\beta \end{pmatrix}. \tag{4}$$

Multiplying out the right-hand side of (4), the coordinate conversions can be expressed as

$$\begin{aligned} E &= \sin^{-1}(\sin\alpha\cos\beta - \cos\alpha\sin\beta\cos I) \\ \Delta &= \tan^{-1}\left[\frac{-\sin\beta\sin I}{\cos\alpha\cos\beta + \sin\alpha\sin\beta\cos I}\right] \\ A &= V + \Delta. \end{aligned} \tag{5}$$

In order to have complete coverage of the zenith region, it is necessary that $E = 90°$ when $I = 180°$. From (5) this occurs if and only if $\alpha + \beta = 90°$. All relationships from here on assume this zenith condition. In particular, since $\alpha + \beta = 90°$, let us define,

$$\left.\begin{aligned} a &\equiv \sin\alpha = \cos\beta \\ b &\equiv \sin\beta = \cos\alpha \end{aligned}\right\} a^2 + b^2 = 1 \tag{6}$$

so the elevation-azimuth expressions (5) reduce to*

* The antenna mount for the open cassegrain design[1] has $\alpha = 42.5°$ and $\beta = 47.5°$ which gives a $-5°$ to $90°$ range in E. Therefore $\sin\alpha = \cos\beta = 0.67559$, and $\sin\beta = \cos\alpha = 0.73728$, but for simplicity of notation, and somewhat more generality, we continue to use a and b.

$$\begin{aligned} E &= \sin^{-1}(a^2 - b^2 \cos I) \\ \Delta &= \tan^{-1}\left[-\frac{1}{a}\tan\left(\frac{I}{2}\right)\right] \\ A &= V + \Delta. \end{aligned} \tag{7}$$

The inverse of these expressions gives the required antenna mount angles, I and V, to produce a given azimuth and elevation of the beam axis:

$$\begin{aligned} I &= \cos^{-1}\left(\frac{a^2 - \sin E}{1 - a^2}\right) \\ V &= A - \Delta \\ \Delta &= -\sigma(I)\tan^{-1}\left(\frac{1}{a}\sqrt{\frac{1 - 2a^2 + \sin E}{1 - \sin E}}\right) \end{aligned} \tag{8}$$

where we define

$$\sigma(I) \equiv \operatorname{sgn}(\sin I) \equiv \frac{\sin I}{|\sin I|}. \tag{9}$$

These relationships between the angles I, $\Delta = A - V$, and E are plotted in rectangular form in Fig. 4, and in polar form in Fig. 5. It is apparent from these figures that there are two pairs of drive angles (I,V) corresponding to every pair of tracking angles (E,A).* However, in continuous tracking of a moving target, the choice of which pair (I,V) to use in pointing the antenna to the given (E,A) is arbitrary only at the beginning of the track, since instantaneous switchover to the opposite pair is not possible. To switch from one pair to the other may be necessary or desirable in certain applications (see Section 4.2), but it can be accomplished only by moving the antenna boresight off the target for some finite period of time. The exception to this is the unique case of the target track which passes precisely through the zenith, at which instant the two (I,V) pairs coincide, so that instantaneous switchover can be made.

Finally, the coordinate conversions between the conic mount angular velocities $(\dot{I},\dot{V})$ and the tracking angular velocities $(\dot{E},\dot{A})$ are of interest in determining the control system requirements. Using (6), (8), and (9) we get

* For example, for the tracking angles ($E = 40°$, $A = 80°$), there are the two equivalent pairs of drive angles ($I_1 = 110°$, $V_1 = 145°$) and ($I_2 = 250°$, $V_2 = 15°$).

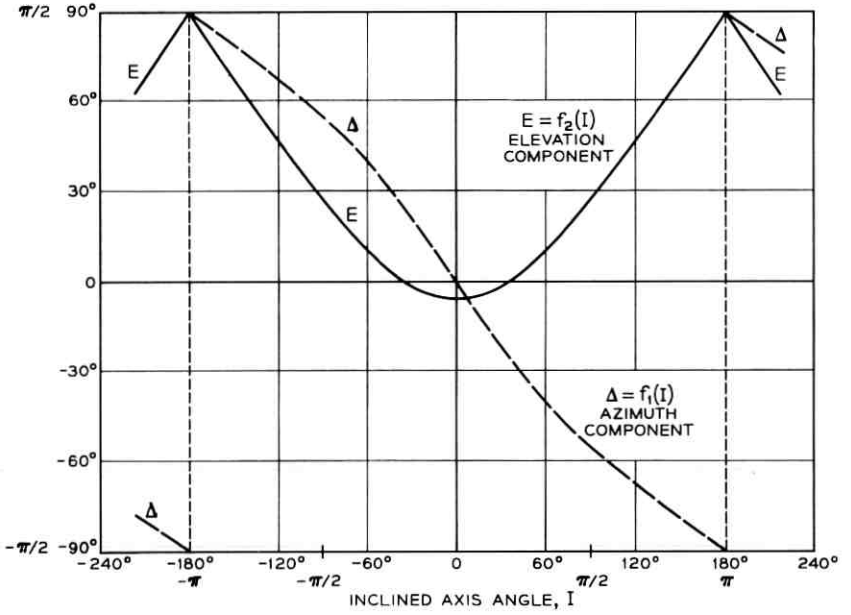

Fig. 4 — Resolution of inclined axis angle into elevation and azimuth components for antenna parameters: $\alpha = 42.5°$, $\beta = 47.5°$.

$$\dot{I} = \sigma(I)\left(1 - \frac{2a^2}{1 + \sin E}\right)^{-\frac{1}{2}} \dot{E}$$

$$\dot{\Delta} = -\frac{a\dot{I}}{1 + a^2 - b^2 \cos I} + \pi\delta\left(\cos\frac{I}{2}\right) \tag{10}$$

$$\dot{V} = \dot{A} - \dot{\Delta}$$

where δ $(\cos I/2)$ is the Dirac delta function of $(\cos I/2)$, representing the derivative of the step discontinuity in Δ which occurs whenever $\cos(I/2) = 0$, i.e., whenever $I = \pm(2n + 1)\pi$, $n = 0,1,2, \cdots$ (see Fig. 4).

To predict the tracking requirements for the two-axis conic mount, it is clear from (8) and (10) that the elevation and azimuth angles (E,A), and rates $(\dot{E},\dot{A})$ are needed. In the tracking of communication satellites, these are functions of the orbit parameters of the satellites, the locations of the tracking station, and the time reference chosen. Computer routines for providing such data are available.[4] For the purpose of analysis and preliminary design, however, it is desirable to have

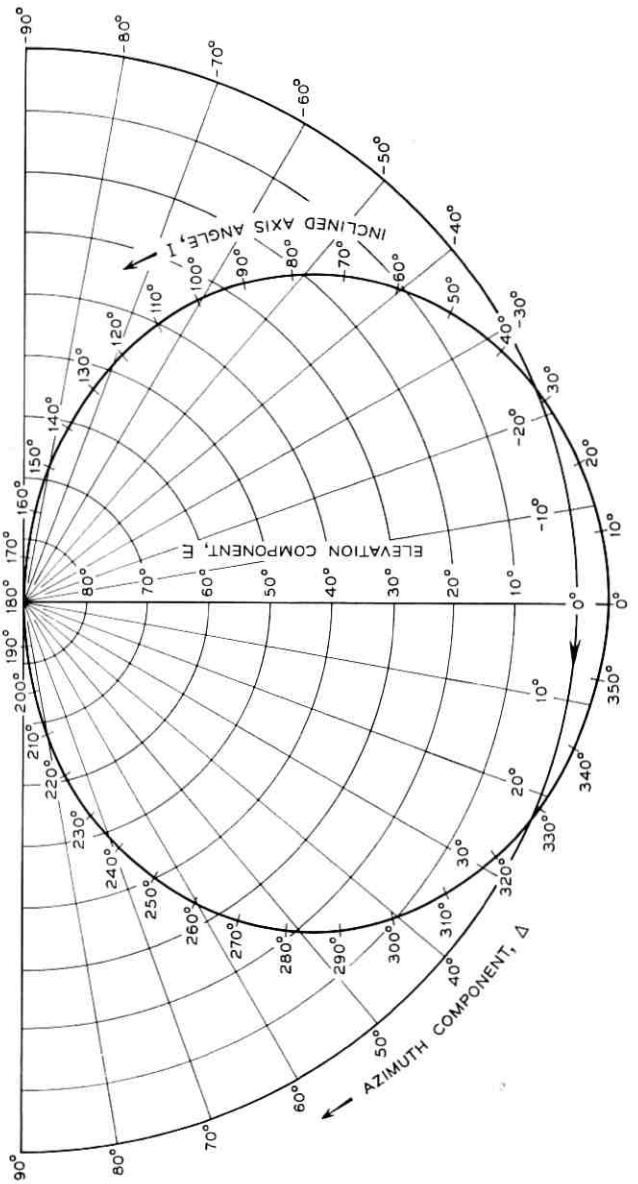

Fig. 5 — Polar plot of azimuth and elevation components of the inclined axis angle for antenna with parameters $\alpha = 42.5°, \beta = 47.5°$.

analytic expressions for the azimuth and elevation angles and rates for circular orbits. These expressions are derived in Appendix A.

III. DYNAMICAL MODEL OF ANTENNA AND DRIVE MOTORS

As an essential step in the design of the control system, we consider now the mathematical model of the antenna mechanical system and the drive motors. Although the model must realistically represent the output angle response (I,V) to input signals, (u_1, u_2), some simplification will be made to ease computer simulation.

The model of the antenna mechanical system, based on a study by Coyne,[5] which was considered to adequately represent the essential dynamics, is illustrated in Fig. 6.

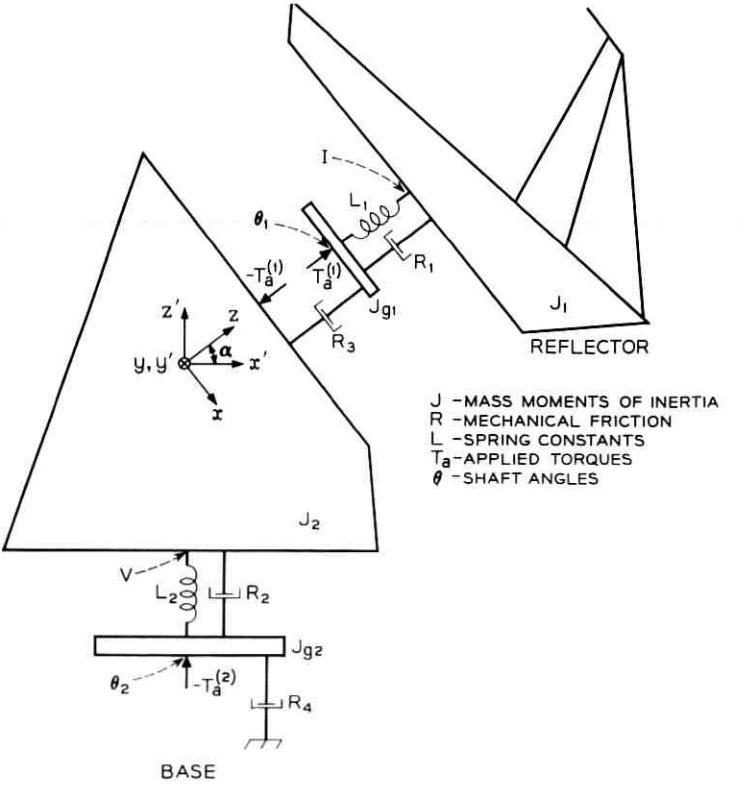

Fig. 6 — Model of antenna mechanical system.

Using the notation in Fig. 6, the equations of motion are

$$T_a^{(1)} = J_{g1}\ddot{\theta}_1 + L_1(\theta_1 - I) + R_1(\dot{\theta}_1 - \dot{I}) + R_3\dot{\theta}_1 \tag{11}$$
$$L_1(\theta_1 - I) + R_1(\dot{\theta}_1 - \dot{I}) = J_{1zz}\ddot{I} - m_1,$$

for the reflector structure, and

$$T_a^{(2)} = J_{g2}\ddot{\theta}_2 - L_2(V - \theta_2) - R_2(\dot{V} - \dot{\theta}_2) + R_4\dot{\theta}_2 \tag{12}$$
$$-L_2(V - \theta_2) - R_2(\dot{V} - \dot{\theta}_2) = J_{2z'z'}\ddot{V} - aT_a^{(1)} + m_2,$$

for the base structure, where

$$\begin{aligned} m_1 &= -bJ_{1xz}(\ddot{V}\cos I - \dot{I}\dot{V}\sin I) + aJ_{1zz}\ddot{V}, \\ m_2 &= b^2 J_{1yy}\sin I(\ddot{V}\sin I + \dot{I}\dot{V}\cos I) \\ &\quad + b\cos I[J_{1xx}b(\ddot{V}\cos I - \dot{I}\dot{V}\sin I) - J_{1xz}(a\ddot{V} - \ddot{I})]. \end{aligned} \tag{13}$$

The product mass moments of inertia of the reflector section are defined with respect to the x-y-z coordinate system. The $J_{2z'z'}$ is measured with respect to the vertical (z') axis.

Due to the large speed variations required in tracking, and because a stiff drive is needed to cope with disturbance torques, it is expected that hydraulic transmissions similar to the units used to drive the Andover horn-reflector antenna[6] will be employed as the drive units. In addition, direct gearing is assumed.

The differential equation which describes the hydraulic drive unit[6] is

$$K_u u = K_m\dot{\theta}_m + K_L P + K_C\dot{P} \tag{14}$$

where u, θ_m, and P represent the drive signal from the controller, motor shaft angle, and hydraulic pressure. The torque delivered by the hydraulic unit is proportional to the pressure, or

$$T_D = K_T P. \tag{15}$$

Equation (14) is valid provided the pressure, P, is less than the maximum pressure, $P_{\max}$, which is allowed in the transmission. This condition is shown in Fig. 7 where the lines AB and CD represent (14) with $u = \pm\bar{v}_{\max}$ and AD and BC represent $P = \pm P_{\max}$. The line EOF represents the operating line or static load line of the hydraulic transmission. To illustrate the dynamic operating of the hydraulic transmission, suppose the operating point is at G with $u = \bar{v}_2$ and a large change in drive signal to $u = \bar{v}_1$ occurs. The motor velocity cannot change instantaneously, so that the pressure in the transmission unit becomes larger. If the drive signal change is large enough, the P may exceed $P_{\max}$ and the

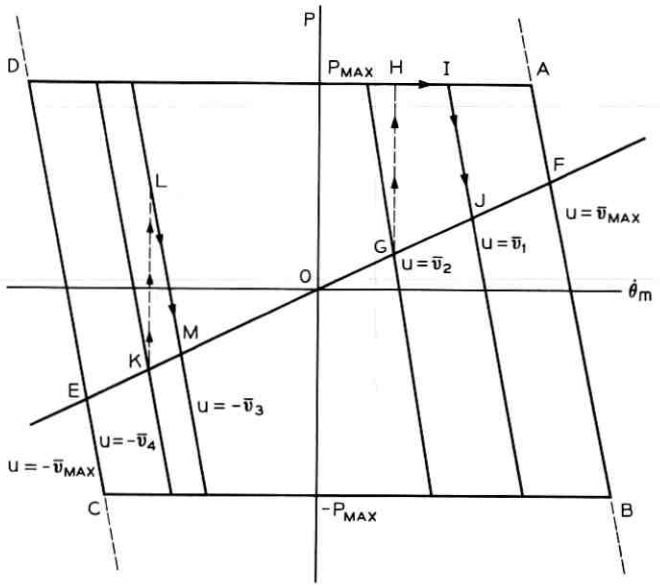

Fig. 7 — Dynamic operation of hydraulic transmission.

new operating point jumps to point H in Fig. 7. At this point, a relief valve is actuated which maintains the pressure at $P_{\max}$ and the torque at $K_T P_{\max}$ until the shaft velocity increases to the value at point I. Then, (14) is again valid with $u = \bar{v}_1$ and the shaft velocity will increase to the value at J. If the drive signal change is such that $P < P_{\max}$, the dynamic path is similar to KLM. As a result, in the system simulation, the value of P must be monitored and, if it exceeds $P_{\max}$, we must set $P = P_{\max}$, i.e., a limiting function.

A counter-torque arrangement of the motors was assumed to eliminate hysteresis effects in the drive systems.[6] This permits the linear gear train equations,

$$\begin{aligned} T_a &= N_G T_D , \\ \theta_m &= N_G \theta , \end{aligned} \tag{16}$$

to apply, where T_a, θ, and N_G are the torque applied to each axis in Fig. 6, the shaft angle at the gearbox output, and the gear ratio, respectively.

The mathematical model for the computer simulation of the antenna structure and the drive motors, represented by (11) through (16) and Fig. 7, is shown in detailed block diagram form in Figs. 8 and 9.

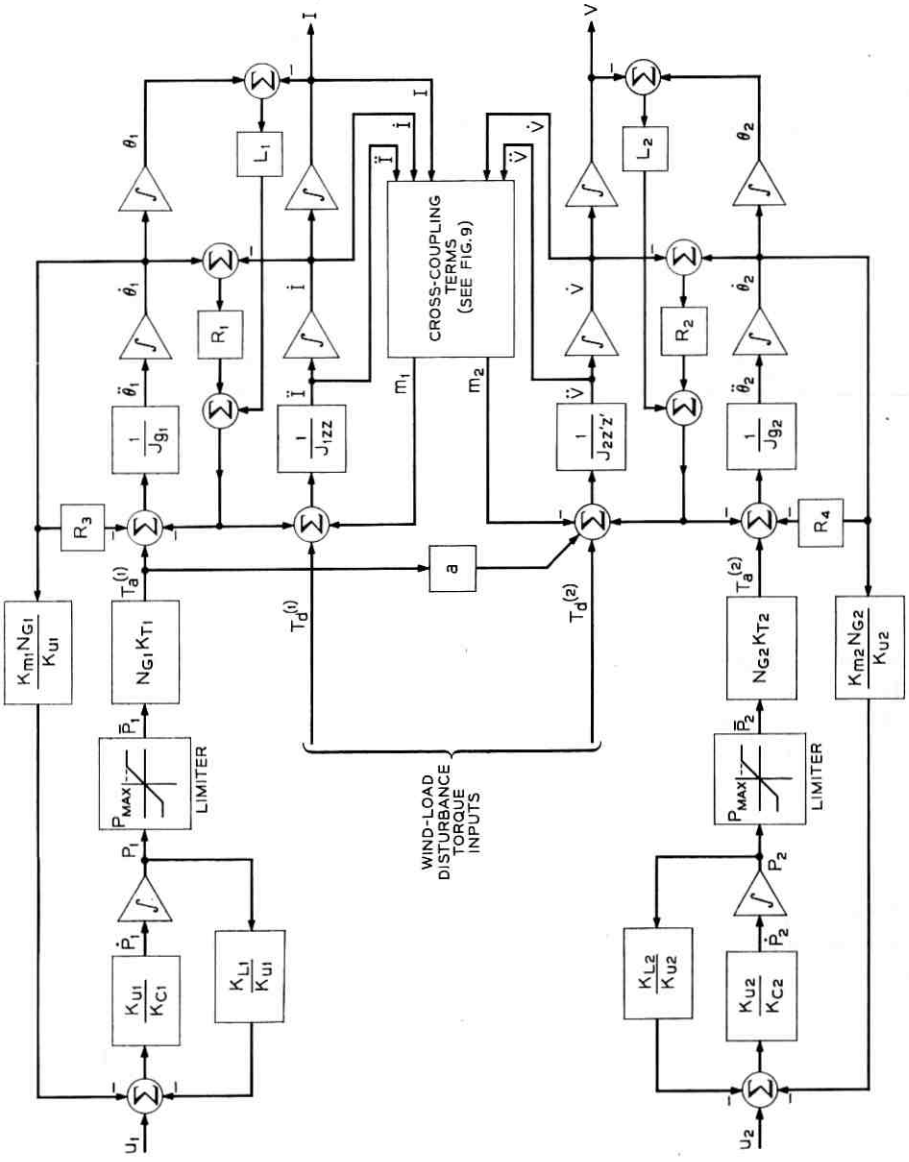

Fig. 8 — Block diagram of equations of motion of antenna structure in response to drive torques.

IV. CONTROLLER DESIGN

4.1 *Approximately Non-Interacting Control*

In order to design the controller for approximately non-interacting control of each channel, it is necessary first to represent the drive motors and antenna structure in terms of a linear system which is a good approximation of the actual system in the small error-signal case and normal operating conditions. We assume first that the drive motors are unsaturated, i.e. $P < P_{\max}$. Next we assume that the torsional spring constants of the reflector and base structures are sufficiently large that $\theta_1(t) \doteq I(t)$ and $\theta_2(t) \doteq V(t)$.

Although these assumptions linearize and simplify the drive and self-coupling portion of each channel in Fig. 8, some major complexities and nonlinearities of the system remain in the cross-coupling portion shown in Fig. 9. There are, unfortunately, no reasonable or standard assumptions to apply to this portion, only educated guesses. After examining the relative magnitude of the various terms in Fig. 9 for typical satellite tracks, the simplified linear system representation shown in Fig. 10 was chosen for the preliminary design study.

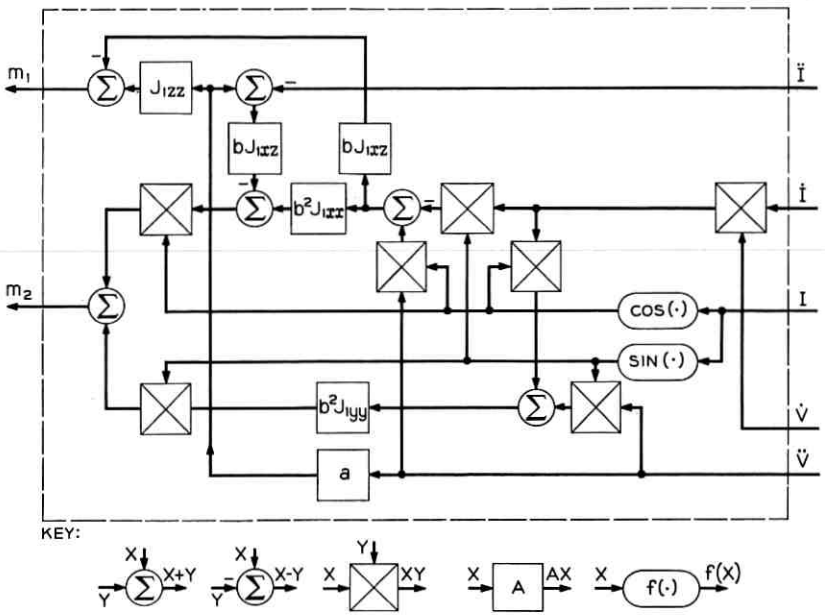

Fig. 9 — Cross coupling portion of Fig. 8.

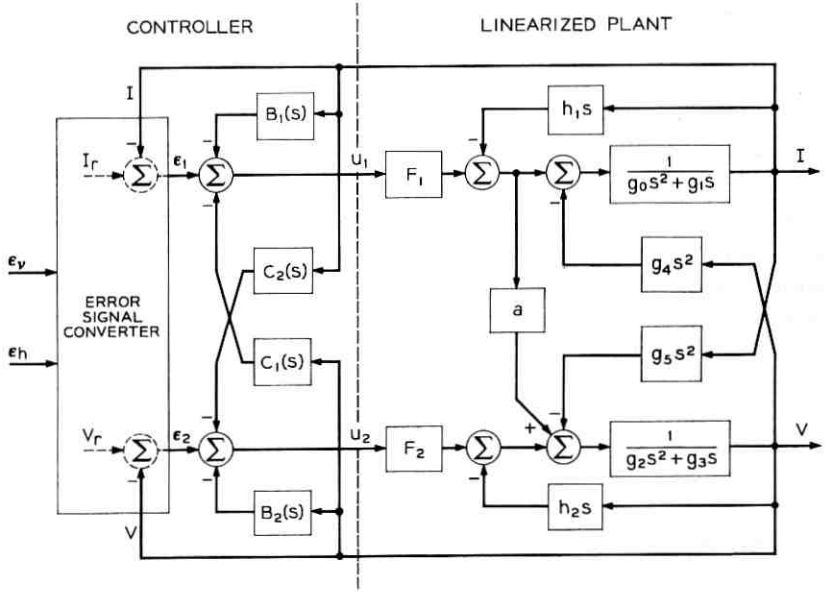

Fig. 10 — Simplified linear system representation for deriving basic non-interacting controller.

Fig. 10 also shows the elements of the basic controller design: the error signal conversion unit and the feedback compensation units.

The channel error signals, which are conceptually shown in Fig. 10 as the difference between the desired, or reference, angle and the output angle for each channel, are obtained as physical signals from the conversion of the autotrack error signals, given in (1). This conversion, derived in Appendix B under the assumption of small errors, is given by,

$$\begin{aligned} \epsilon_1 &\doteq \frac{1}{K_v}\left[1 + a^2 \cot^2 \frac{I}{2}\right]\epsilon_v \\ \epsilon_2 &\doteq \left[\frac{a}{2K_v} \csc^2 \frac{I}{2}\right]\epsilon_v + \left[K_h b^2 \sin^2 I \left(1 + a^2 \cot^2 \frac{I}{2}\right)\right]^{-1} \epsilon_h , \end{aligned} \tag{17}$$

where K_v and K_h are the detector gains in (1), and the constants a and b are defined in (6).

The controller outputs, u_1 and u_2, are derived from these error signals plus additional compensation through the feedback networks $B(s)$ and $C(s)$. The purpose of these networks is to improve the response to errors in the same channel, and to eliminate the undesired response to errors in the opposite channel. For the development of this non-interacting

design it is convenient to use matrix notation. We define the input, output and control vectors,

$$\varphi_r = \begin{pmatrix} I_r \\ V_r \end{pmatrix}, \qquad \varphi = \begin{pmatrix} I \\ V \end{pmatrix}, \qquad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}.$$

The overall input-output relationship is denoted by

$$\varphi(s) = D\varphi_r(s) \tag{18}$$

where, for non-interacting channels, the transfer matrix must be diagonal. Therefore, we require D to have the form

$$D = \begin{pmatrix} D_1(s) & 0 \\ 0 & D_2(s) \end{pmatrix}. \tag{19}$$

From the block diagram of Fig. 10, we obtain the intermediate relationships,

$$\begin{aligned} G\varphi(s) &= F\mathbf{u}(s) \\ \mathbf{u}(s) &= \varphi_r(s) - C\varphi(s), \end{aligned} \tag{20}$$

where

$$C = \begin{pmatrix} 1 + B_1(s) & C_1(s) \\ C_2(s) & 1 + B_2(s) \end{pmatrix}, \qquad F = \begin{pmatrix} F_1 & 0 \\ aF_1 & F_2 \end{pmatrix},$$

and (21)

$$G = \begin{pmatrix} g_0 s^2 + (g_1 + h_1)s & -g_4 s^2 \\ g_5 s^2 + a h_1 s & g_2 s^2 + (g_3 + h_2)s \end{pmatrix}.$$

Then, from (18) and (20), it follows that the controller matrix, C, is given by

$$C = D^{-1} - F^{-1}G. \tag{22}$$

Upon substitution of the matrices in (19) and (21) into (22), we obtain the required transfer functions of the controller units:

$$\begin{aligned} B_1(s) &= \frac{1}{D_1(s)} - 1 - \frac{1}{F_1}[g_0 s^2 + (g_1 + h_1)s] \\ B_2(s) &= \frac{1}{D_2(s)} - 1 - \frac{1}{F_2}[(g_2 + a g_4)s^2 + (g_3 + h_2)s] \\ C_1(s) &= \frac{1}{F_1}(g_4 s^2) \\ C_2(s) &= \frac{1}{F_2}[(a g_0 - g_5)\, s^2 + a g_1 s]. \end{aligned} \tag{23}$$

The only design objective incorporated in these controller functions is that of non-interacting channel control. This has fixed the cross-coupling controller units $C_1(s)$ and $C_2(s)$ in terms of the linear plant parameters, the output angles (I,V), and their derivatives. The self-coupling controller units, $B_1(s)$ and $B_2(s)$, however, depend not only on these plant parameters and state variables, but also on the choice of the channel input-output transfer functions, $D_1(s)$ and $D_2(s)$.

Since both $B_1(s)$ and $B_2(s)$ have the same functional form, we will use the subscript $i = 1,2$ to refer to either channel. From (23), $B_i(s)$ is written in the simplified form,

$$B_i(s) = \left(\frac{1}{D_i(s)} - 1\right) - \frac{s^2}{c_i} - \frac{s}{f_i}, \tag{24}$$

where for $i = 1$: $c_1 = F_1/g_0$, $\quad f_1 = F_1/(g_1 + h_1)$,

and for $i = 2$: $\quad c_2 = F_2/(g_2 + ag_4)$, $\quad f_2 = F_2/(g_3 + h_2)$.

Accuracy, fast response, stability, and a practical feedback structure are the general objectives which should be mutually satisfied to the extent possible in the choice of the channel transfer function, $D_i(s)$. More specifically, we consider the following requirements:

i. Each channel should have no steady-state error for step and ramp inputs.

ii. The feedback synthesis of $B_i(s)$ should employ signals proportional to the output angle, its velocity, and acceleration, but no higher derivatives.

iii. The error for a sinusoidal input, having an angular velocity no greater than ν_i and an acceleration no greater than γ_i, should not exceed the allowable value, ρ_i. (Appropriate numerical values of ν, γ, and ρ to be specified for each channel.)

iv. The choice of $D_i(s)$ should achieve a good compromise between the competitive aims of fast transient response and small noise bandwidth.

To satisfy requirements (i) and (ii), the channel transfer functions must be of the form

$$D_i(s) = \frac{c_i s + d_i}{s^3 + b_i s^2 + c_i s + d_i} \tag{25}$$

where the coefficient c_i is the same as in (24), but b_i and d_i are available as design parameters. Using (25) in (24) yields

$$B_i(s) = \frac{-1}{c_i s + d_i}\left[\left(\frac{d_i}{c_i} + \frac{c_i}{f_i} - b_i\right) s^2 + \frac{d_i}{f_i} s\right] \tag{26}$$

where c_i and f_i are defined below (24) in terms of the linear plant constants shown in Fig. 10.

The error transform for either channel is

$$\epsilon_i(s) = [1 - D_i(s)]\,\varphi_r(s) = \frac{s^2(s + b_i)\varphi_r(s)}{s^3 + b_i s^2 + c_i s + d_i} \tag{27}$$

where φ_r is the channel reference input (either I_r or V_r). Considering requirement (iii), φ_r is a sinusoidal signal,

$$\varphi_r(t) = \varphi_i(\omega) \sin \omega t,$$

where the peak amplitude is given by

$$\varphi_i(\omega) = \begin{cases} \nu_i/\omega, & \omega < \gamma_i/\nu_i \\ \gamma_i/\omega^2, & \omega \geqq \gamma_i/\nu_i . \end{cases} \tag{28}$$

The steady-state error for this input will not exceed the allowable value, ρ_i, provided

$$| 1 - D_i(j\omega) | \leqq \rho_i/\varphi_i(\omega) . \tag{29}$$

Both sides of (29) are shown in the log amplitude-log frequency plot of Fig. 11, using (28) and the asymptotic straight-line approximation of (27). Since at low frequencies,

$$| 1 - D_i(j\omega) | \doteq b_i\omega^2/d_i , \tag{30}$$

it follows that requirement (iii) imposes the design constraint (see Fig. 11),

$$b_i/d_i \leqq \rho_i/\gamma_i \tag{31}$$

which shows that for a given tolerable error, ρ_i, the limit on the parameter ratio is imposed by the peak acceleration, γ_i, rather than the peak velocity, ν_i.

In considering the transient response of each channel, we make use of a computer study made by J. F. Kaiser on the step and ramp response of the equivalent system

$$D(s) = \frac{\eta\omega_0^2 s + \omega_0^3}{s^3 + \mu\omega_0 s^2 + \eta\omega_0^2 s + \omega_0^3} \tag{32}$$

which has a noise bandwidth related to the coefficients by,

$$\mathfrak{B} \equiv \int_{-\infty}^{\infty} | D(j\omega) |^2 \, df = \frac{\eta^2 + \mu}{2(\eta\mu - 1)}\,\omega_0 . \tag{33}$$

From this study, the parameter values which seem to give the best com-

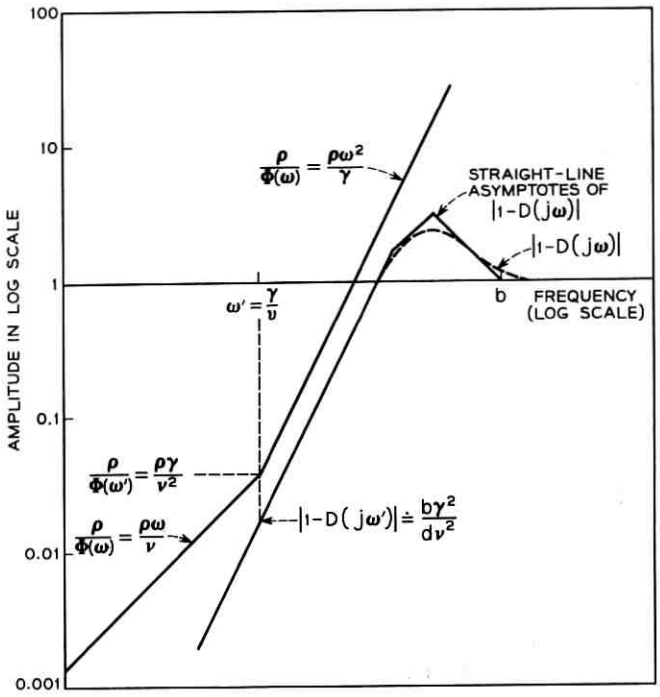

Fig. 11 — Amplitude-frequency sketch of error functions in relation (29).

promise between good transient response and small noise bandwidth lie in the region,

$$\left.\begin{matrix} 2.3 < \eta < 2.7 \\ 1.8 < \mu < 2.3 \end{matrix}\right\} \tag{34}$$

within which the noise bandwidth varies from about 0.9 ω_0 to 1.1 ω_0. For small bandwidth, ω_0 should be chosen small. However, since $b_i = \mu\omega_0$ and $d_i = \omega_0^3$, the lower bound on ω_0 to satisfy (iii) is, from (31),

$$\omega_0^2 \geqq \mu(\gamma_i/\rho_i). \tag{35}$$

Furthermore, the choice of ω_0 is also linked to the physical system constants, since $c_i = \eta\omega_0^2$.

The detailed choice of the numerical values in the controller design cannot be made until all the values of the system constants and the operating requirements are known. However, the specific structure of the non-interacting controller designs can be given in terms of the channel gains F_1 and F_2, the plant parameters g_0 through g_5 (see Fig. 10), and

the variable parameters b_1, d_1, b_2 and d_2. This controller design for the normal autotrack mode of operation is shown in Fig. 12. The design of supplementary control action for special tracking modes is considered in the following section.

4.2 *Near Zenith Control Modes*

As noted in Section II, there are two possible tracking modes for a given satellite track. For later reference, let "mode 1" (I_1, V_1) be the tracking mode where $0 \leqq I \leqq 180°$ and "mode 2" (I_2, V_2) be the tracking mode where $180° \leqq I \leqq 360°$. These two modes are illustrated in Fig. 13 as a function of time for the tracking of a satellite in a circular equatorial orbit of 6000 miles altitude with the antenna site located at 2° latitude. As can be seen from this figure, the maximum vertical axis angular velocity required to stay on track, which we shall call $\dot{V}_{max}$, occurs at the point of maximum azimuth tracking rate and maximum elevation. As the maximum elevation angle approaches 90°, $\dot{V}_{max}$ will eventually exceed the maximum vertical-axis velocity of the antenna, $\hat{S}$.

Fortunately, there is a factor which reduces vertical axis tracking requirement. The antenna has an "on-track" beamwidth,* 2ξ, so that it is possible to track without the antenna pointing directly at the target. We can utilize this "on-track" beamwidth in the following manner. For any satellite path which passes within ξ degrees of zenith, it is possible to switch tracking modes without any interruption in communications. This situation is illustrated in Fig. 14, where V_1, V_2 tracking modes near one of the switchover points are plotted as a function of time for an assumed beamwidth of 0.2 degrees, a maximum satellite elevation angle, $E_{max} = 89.9°$, and a satellite altitude of 6000 miles (circular equatorial orbit). The solid lines show the vertical axis angle if the satellite is on boresight. The two pairs of dashed lines represent the allowable variation in the vertical axis pointing angle due to the beamwidth of the antenna. If the vertical axis angle is anywhere within the area between the dashed lines, communications can be maintained with the satellite. Since the tracking areas have a point of intersection at $V = V_s$, a smooth transition between tracking modes is possible without a lapse in communications. For satellite tracks where $E_{max} > 89.9°$, the two tracking mode areas will have a large area of intersection rather than a single point of intersection as in the limiting case discussed above.

This operation of switching tracking modes will be referred to as the "switch-mode." If the switch mode is employed, the vertical axis is to

* That is, the beamwidth for which the signal-to-noise ratio at the receiver is considered sufficient for communication objectives of the system.

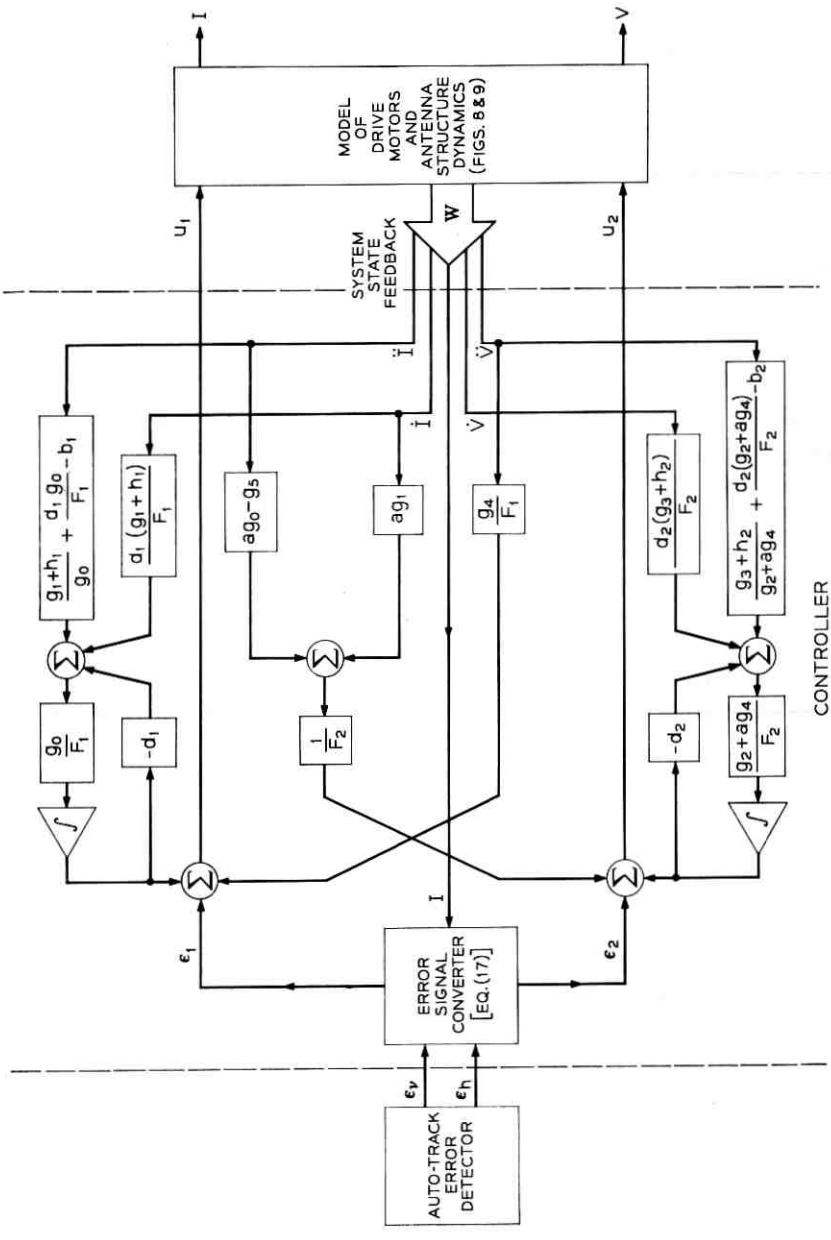

Fig. 12 — Detail structure of basic controller design for normal autotrack mode.

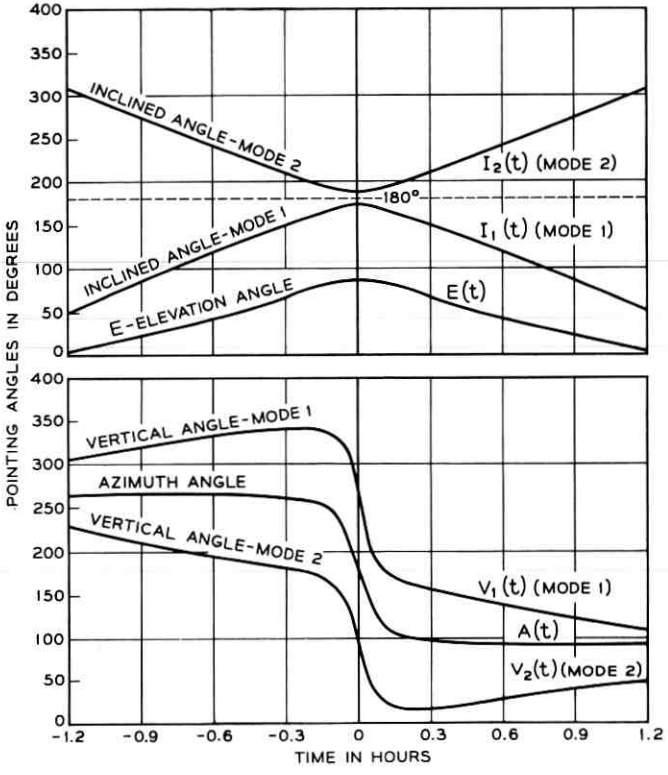

Fig. 13 — Antenna pointing angles for satellite in circular equatorial orbit of 6000 mile altitude with antenna station at 2° latitude.

track keeping the satellite on boresight until, at time t_{s1} in Fig. 14, the vertical axis angle of the tracking mode equals the angle at which switchover from one mode to the other will occur. This switchover angle, V_s, is equal to the azimuth angle at the maximum elevation point and is computed in advance from the predicted satellite orbit. At t_{s1}, a vertical error signal $(V_s - V)$ is employed to keep the vertical axis angle at V_s. This error signal $(V_s - V)$ is maintained until, at time t_{s2} as in Fig. 14, the vertical axis angle for boresight tracking of the second tracking mode is equal to V_s. At this point the switch mode vertical axis error $(V_s - V)$ is replaced by the boresight tracking error $(V_r - V)$ and tracking is continued using the second tracking mode.

Graphical investigation of this switchover process indicates that, depending on the satellite track, a particular tracking mode should be

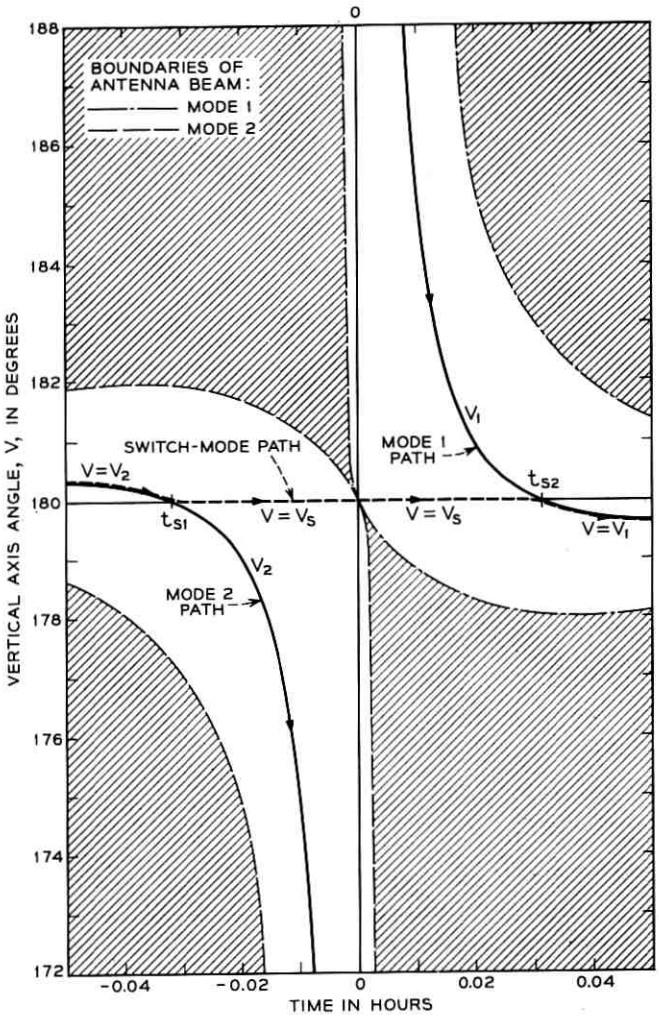

Fig. 14 — Tracking angles of vertical axis near one of the switchover points for $E_{max} = 89.9°$, $\xi = 0.1°$, and 6000 mile circular equatorial orbit.

chosen for tracking from the horizon, where the satellite is first acquired, to the maximum elevation point and the other mode from the maximum elevation to the horizon. The choice is made by determining, from orbit predictions, the initial vertical axis angular velocity for each tracking mode when the satellite initially appears at the horizon. If the initial vertical axis tracking velocity of mode 1, $\dot{V}_{H1}$, is greater than the initial

tracking velocity of mode 2, $\dot{V}_{H2}$, then mode 1 is used first and the switch-over is made to mode 2. The opposite procedure is employed if $\dot{V}_{h2} > \dot{V}_{h1}$. Although the tracking velocity requirement is greater at the horizon using this procedure, the tracking velocity requirements near zenith are reduced.

For satellite tracks where $E_{max} < 90° - \xi$, one cannot switch tracking modes without losing communications for a brief period. Therefore, one desires to track the satellite by remaining in the same tracking mode. However, instead of pointing directly at boresight when tracking the satellite, one can again utilize the "on-track beamwidth" to point off boresight and still be within the antenna beam as shown in Fig. 15. By following the tracking path, $\tilde{V}$, illustrated in Fig. 15, it is possible to significantly reduce the peak vertical axis velocity requirement from the velocity required for boresight tracking. This procedure will subsequently be referred to as the "slant-through" mode.

In the design of the antenna drive system, an upper limit is needed on the maximum vertical axis angular velocity required to follow the slant-through path, $\tilde{V}$, for all circular satellite orbits of a given altitude, i.e. the worst case slant-through speed, $\hat{S}$. Since it is expected that the com-

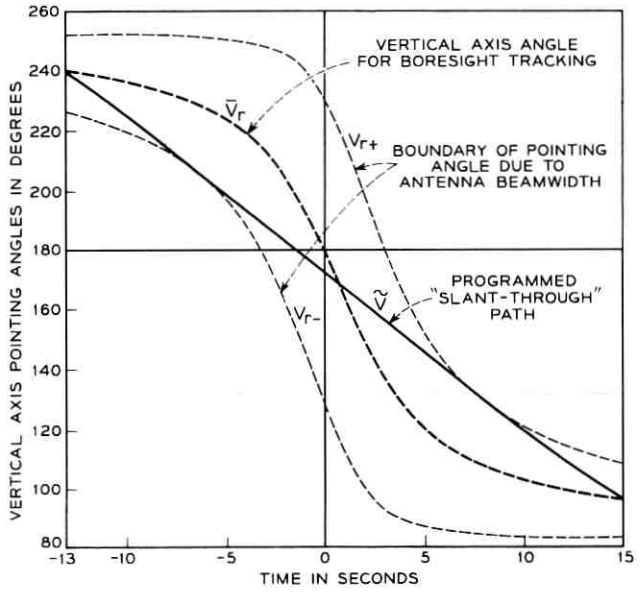

Fig. 15 — "Slant through" path for 6000 mile altitude, circular polar orbit with $E_{max} = 89.893°$.

munication satellites to be tracked will be launched in the same direction as the earth's rotation, the worst case will occur when the inclination angle, α, equals 90° (a polar circular orbit) and $E_{max} = (90° - \xi)$. (This elevation is chosen since if E_{max} is larger than $(90° - \xi)$ the switch mode will be employed.) An approximate expression for $\hat{S}$ is derived in Appendix C. Shown in Fig. 16 is a graph of $\hat{S}$ as a function of satellite altitude, $x = (r - 1)$, for antenna half-beamwidths of 0.1° and 0.2°. For example, the continuous tracking of satellites in nearly circular orbits of 6000 miles altitude with an antenna half-beamwidth of 0.1° would require a vertical axis slewing capability of approximately 1 rpm. A major factor affecting this capability is the gear ratio used in the drive system, the choice of which depends also on other important considerations such as tracking performance at very low speeds, immunity to disturbance torques, minimization of reflected load inertia, and drive power requirements.

This tracking capability should not be difficult to achieve for medium and high altitude satellite systems. For low altitude systems, continuous tracking may be achieved with the same speed capability by temporarily

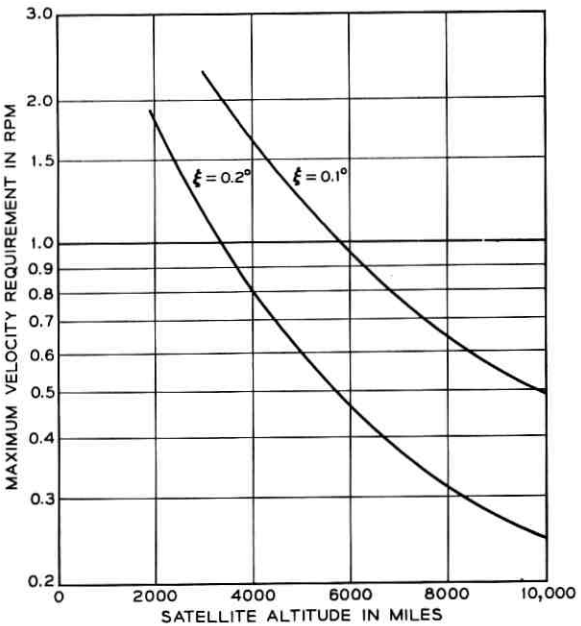

Fig. 16 — Vertical axis velocity requirement as a function of satellite altitude for antenna half-beamwidths $\xi = 0.1°$ and 0.2°.

broadening the antenna beam during the near-zenith portion when signal strength conditions are favorable.

The slant-through mode can be performed by biasing the vertical axis error signal with another signal $(\bar{V} - \bar{V}_r)$ between $t_1 < t < t_2$ as shown in Fig. 15. $\bar{V}$ is the computed vertical axis angle which will program the "slant-through" path within the antenna beamwidth. $\bar{V}_r$ is the computed vertical axis angles for boresight tracking determined from orbit predictions. Since $\bar{V}_r \doteq V_r$, the resultant error signal looks like $(\bar{V} - V)$.

The above discussion indicates the following tracking strategy. From orbit predictions of the satellite path, the elevation values (E) as a function of time are scanned. If the maximum elevation, $E_{\max}$, is greater than or equal to $(90° - \xi°)$ the "switch" mode will be employed as explained above. If $E_{\max} < (90° - \xi)$, but $(\dot{V}_{\max}) > \hat{S}$, the slant-through mode is employed at time t_1 in Fig. 15. If neither of these special modes is required, normal tracking will be employed. For the normal and slant-through modes, it is desirable to use the tracking mode which has the smaller vertical axis velocity required when initially acquiring the satellite at the horizon. The inclined axis velocity requirement at satellite acquisition is the same for either mode. If $\dot{V}_{H2} > \dot{V}_{H1}$, then mode 1 will be used for tracking and the opposite choice will be used if $\dot{V}_{H1} > \dot{V}_{H2}$.

The mode tracking strategy is shown in flow chart form in Fig. 17. Note that no special control signals are needed for the inclined axis. The initial inclined and vertical pointing angles when the satellite appears at the horizon, I_H and V_H, can also be determined from orbit predictions and (8).

V. DESIGN RESULTS AND CONCLUSIONS

The second phase of this design study consists of a simulation and design evaluation program on a hybrid analog-digital computer facility.[7] The major objectives of this program are to verify, improve, and if possible, simplify the basic controller strategy developed in Section IV.

The analog computer portion of the facility is being used for the simulation of the controller and the antenna dynamics. The experimental results discussed here were based on a simulation of an open cassegrain antenna with a 56-foot aperture, an overall height of about 70 feet, and a total weight of about 100 tons,* using two 25-hp hydraulic motors for the vertical axis drive and two 10-hp motors for the inclined axis drive.

The closed-loop response of the antenna drive system using the basic controller design shown in Fig. 12 was tested using step and ramp in-

* Subsequent design modifications have changed the weight and the antenna dynamics somewhat, but not enough to significantly affect the simulation results.

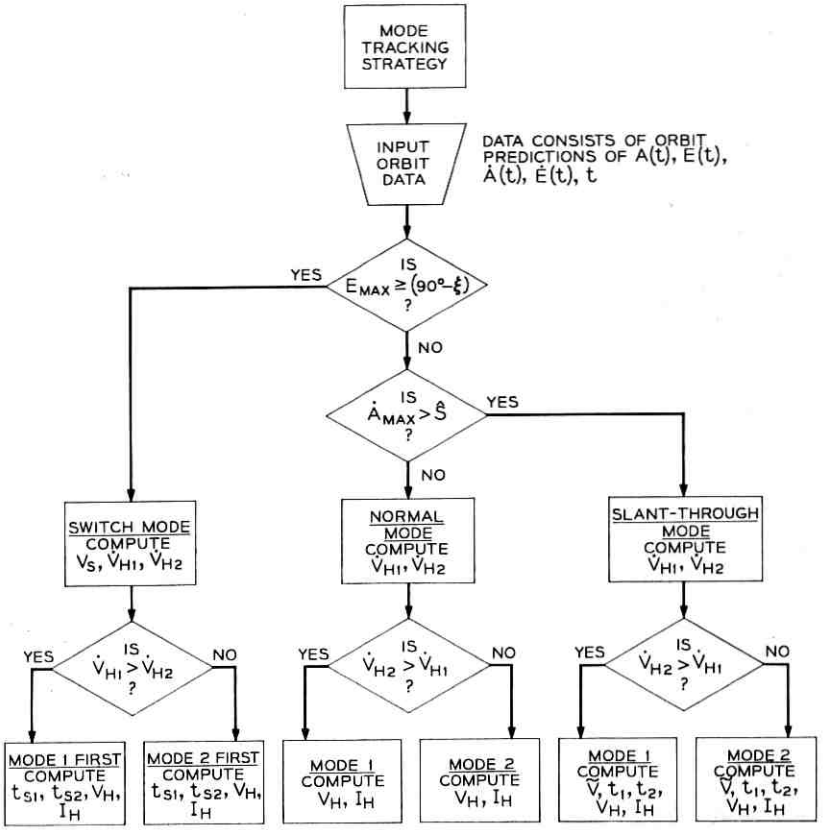

Fig. 17 — Flow chart of tracking strategy.

puts, I_r and V_r, as well as wind disturbance torque inputs. The design for each channel was based on the approximate model transfer function (32) with parameter values, $\eta = 2.4$, $\mu = 1.9$, and a range of values of ω_0 from 5 to 40 sec.$^{-1}$. Because of the low accelerations required for the expected satellite tracks, the constraint (35) was easy to satisfy with the available range of loop gains; therefore the major considerations in the choice of ω_0 were good transient response, steady-state accuracy, and immunity to wind disturbances. Satisfactory performance with respect to the objectives of zero steady-state error in tracking constant velocity inputs and minimum channel interaction was achieved in the experimental design with state feedback gain adjustments close to the nominal values computed for the expressions in Fig. 12.

The experimental data obtained from this design study has established

the feasability of operating the open cassegrain antenna under severe wind conditions without a radome. A series of tests using simulated wind loads corresponding to a 40-mph gale with gusts exceeding 80 mph have indicated that the control system is capable of maintaining the antenna beam on-track with both a mean and rms error less than 0.002 degrees, which is about 1/100th of the nominal beamwidth of the antenna.

The future test program in this design study will employ the digital computer portion of the hybrid computer facility to simulate satellite tracking data, autotrack error detector signals, and the necessary coordinate conversions for resolving the error signal inputs and the angular outputs of the non-orthogonal axes of motion. This will provide a complete simulation of the overall autotrack system shown in Fig. 2, and will allow testing and evaluation of the overall control strategy for non-orthogonal mounts, including the near zenith control modes discussed in Section 4.2.

VI. ACKNOWLEDGMENTS

The authors wish to acknowledge and thank Messrs. J. S. Cook, K. N. Coyne, J. Chernak, J. F. Kaiser, and J. A. Norton for their advice and contributions to this work.

APPENDIX A

We consider the polar coordinates $[r(t),\theta(t)]$ of the satellite in the orbital plane to be given, and then make the sequence of transformations necessary to relate the tracking coordinates (A,E) to these orbit parameters:

A.1 *Transformation from orbit plane to equatorial plane*

We define rectangular coordinates XYZ and $X'Y'Z'$ as shown in Fig. 18, with the X and X' axes coincident with the line formed by the intersection of the orbit and equatorial planes. The satellite moves into the northern hemisphere at the $+X$-axis. We define the following (see Fig. 18):

$r(t)$ = range from center of earth to satellite at time t.
$\theta(t)$ = angle which $r(t)$ makes at time t, measured from the X'-axis.
α = inclination angle of orbit plane with equatorial plane.*

* Should not be confused with antenna incline-angle, α, defined in Section II, since they will not be used in the same context.

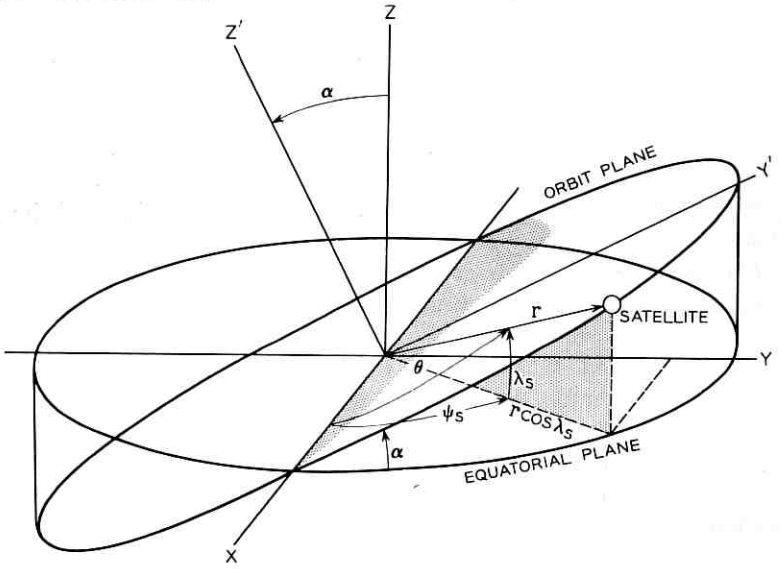

Fig. 18 — Satellite coordinate conversion — Orbit plane to equatorial plane.

$\psi_s(t)$ = longitude angle of satellite eastward from X-axis.
$\lambda_s(t)$ = latitude angle of satellite northward from equator.

The transformation from the orbit plane to equatorial plane corresponds to a rotation about the X-axis through an angle α, so that

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} \tag{36}$$

Expressing the rectangular coordinates in the equivalent spherical polar coordinates, we have

$$\begin{pmatrix} r\cos\lambda_s\cos\psi_s \\ r\cos\lambda_s\sin\psi_s \\ r\sin\lambda_s \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} r\cos\theta \\ r\sin\theta \\ 0 \end{pmatrix} . \tag{37}$$

A.2 *Transformation from equatorial plane to horizon plane at Antenna Station*

The rectangular coordinates (xyz) are located with the origin at the antenna site as shown in Fig. 19, with the $+x$-axis pointing northward. The notation used here is as follows:

R = radius of the earth (assumed constant)
ψ = longitude angle of antenna site, measured eastward from X-axis
λ = latitude angle of antenna site, measured northward from equator
ρ = slant range to satellite from anetnna site
A = azimuth angle measured CW from x-axis (North)
E = elevation angle measured up from horizon (xy-plane)

The transformation from XYZ coordinates to xyz coordinates can be

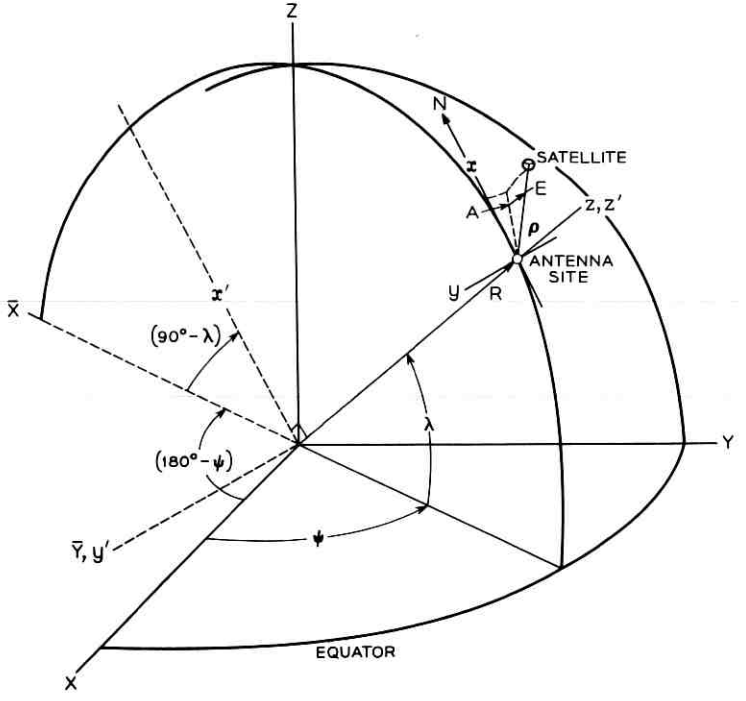

Fig. 19 — Diagram for the transformation of coordinates from the earth-centered system (XYZ) to the local antenna site coordinates (xyz) and pointing angles (A,E).

written as a sequence of two rotations and a translation, as shown in Fig. 19. The first is a CW rotation about the Z-axis through an angle $(180° - \psi)$, which can be written as

$$\begin{pmatrix} \bar{X} \\ \bar{Y} \\ \bar{Z} \end{pmatrix} = \begin{pmatrix} -\cos\psi & -\sin\psi & 0 \\ \sin\psi & -\cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}. \tag{38}$$

The second is a CW rotation about the $\bar{Y}$-axis through an angle $(90° - \lambda)$ as indicated in Fig. 19:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \sin\lambda & 0 & \cos\lambda \\ 0 & 1 & 0 \\ -\cos\lambda & 0 & \sin\lambda \end{pmatrix} \begin{pmatrix} \bar{X} \\ \bar{Y} \\ \bar{Z} \end{pmatrix}. \tag{39}$$

Thirdly, a simple translation along $+z'$-axis a distance R gives,

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ R \end{pmatrix}$$

or, in terms of the spherical coordinates at the antenna site

$$\begin{pmatrix} \rho\cos E\cos A \\ -\rho\cos E\sin A \\ \rho\sin E \end{pmatrix} = \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ R \end{pmatrix}. \tag{40}$$

Finally, combining (37) through (40), we have

$$\begin{aligned} \rho\cos E\cos A &= -r(\sin\lambda\cos\psi\cos\theta + \cos\alpha\sin\lambda\sin\psi\sin\theta \\ &\quad - \sin\alpha\cos\lambda\sin\theta) \\ -\rho\cos E\sin A &= r(\sin\psi\cos\theta - \cos\alpha\cos\psi\sin\theta) \\ \rho\sin E &= r(\cos\lambda\cos\psi\cos\theta + \cos\alpha\cos\lambda\sin\psi\sin\theta \\ &\quad + \sin\alpha\sin\lambda\sin\theta) - R. \end{aligned} \tag{41}$$

Separating these variables gives the desired transformations:

Slant range to satellite:

$$\rho = [r^2 + R^2 - 2rR(\cos\lambda\cos\psi\cos\theta + \cos\alpha\cos\lambda\sin\psi\sin\theta + \sin\alpha\sin\lambda\sin\theta)]^{\frac{1}{2}} \tag{42}$$

Azimuth angle to satellite:

$$A = \tan^{-1}\left(\frac{\sin\psi\cos\theta - \cos\alpha\cos\psi\sin\theta}{\sin\lambda\cos\psi\cos\theta + \cos\alpha\sin\lambda\sin\psi\sin\theta - \sin\alpha\cos\lambda\sin\theta}\right) \tag{43}$$

Elevation Angle to satellite:

$$E = \sin^{-1}\left\{\frac{r(\cos\lambda\cos\psi\cos\theta + \cos\alpha\cos\lambda\sin\psi\sin\theta + \sin\alpha\sin\lambda\sin\theta) - R}{[r^2 + R^2 - 2rR(\cos\lambda\cos\psi\cos\theta + \cos\alpha\cos\lambda\sin\psi\sin\theta + \sin\alpha\sin\lambda\sin\theta)]^{\frac{1}{2}}}\right\} \tag{44}$$

where $-\pi/2 \leqq E \leqq \pi/2$.

In these expressions the constant factors are the orbit inclination angle, α, the antenna site latitude, λ, and the earth radius R. The antenna site longitude angle, ψ, can be considered constant if the effect of the earth's rotation is neglected, but more generally it will have the form

$$\psi(t) = \psi_0 + \Omega t \tag{45}$$

where Ω = earth's angular velocity = $(\pi/12)$ rad/hr. To simplify the expressions we shall measure distance in units of earth radius (e.r.) so that

$$R = 1 \text{ e.r.} \tag{46}$$

Further, to emphasize the constants, let

$$\left.\begin{aligned} C_\alpha &\equiv \cos\alpha, & C_\lambda &\equiv \cos\lambda \\ S_\alpha &\equiv \sin\alpha, & S_\lambda &\equiv \sin\lambda \end{aligned}\right\} \tag{47}$$

Using the notation of (45), (46) and (47), the tracking angles can be written

$$A = \tan^{-1}\left\{\frac{\sin\psi\cos\theta - C_\alpha\cos\psi\sin\theta}{S_\lambda(\cos\psi\cos\theta + C_\alpha\sin\psi\sin\theta) - S_\alpha C_\lambda\sin\theta}\right\} \tag{48}$$

$$E = \sin^{-1}\left\{\frac{r(C_\lambda(\cos\psi\cos\theta + C_\alpha\sin\psi\sin\theta) + S_\alpha S_\lambda\sin\theta) - 1}{[1 + r^2 - 2r(C_\lambda(\cos\psi\cos\theta + C_\alpha\sin\psi\sin\theta) + S_\alpha S_\lambda\sin\theta)]^{\frac{1}{2}}}\right\} \tag{49}$$

where r, θ, and ψ are in general varying with time. The denominator in (49) is the range from antenna to satellite in units of earth radii. The numerator in (49) gives the necessary condition for the satellite to be above the horizon at the antenna site, namely:

$$r[C_\lambda (\cos\psi \cos\theta + C_\alpha \sin\psi \sin\theta) + S_\alpha S_\lambda \sin\theta] \geqq 1 \Rightarrow E > 0 \qquad (50)$$

where, of course, r is the radius vector of the satellite in units of earth radii.

During the period when (50) is satisfied, the $A(t)$ and $E(t)$ given above, as well as their time derivatives, give the required information to evaluate the antenna drive angles and rates, from (8) and (10) in Section II. For non-circular orbits, the expressions for (r,θ) as functions of time must still be determined, and in general a computer routine[4] would be used. The main usefulness of the analytical expressions (48) and (49) is in estimating tracking requirements for particular orbits where complete data are not needed.

A.3 *Tracking angles and rates for circular orbits*

For a given inclination angle, α, of the circular orbit plane (see Fig. 18), and a given latitude, λ, of the antenna site, the tracking angles $A(t)$ and $E(t)$ can be obtained from (48) and (49) as explicit functions of time by substituting the time functions $\psi(t)$, given in (45), and

$$\theta = \omega t + \theta_0 \qquad (51)$$

where $\theta_0 = \theta(t)$ at arbitrary time reference, $t = 0$, and the constant angular velocity of the satellite is given by

$$\omega = kr^{-3/2} \qquad (52)$$

where

$$\begin{aligned} r &= 1 + h/R, \\ h &= \text{satellite altitude} \\ R &= \text{earth radius} \\ k &= (g/R)^{\frac{1}{2}} \\ g &= \text{accel. due to gravity} \end{aligned}$$

or

$$k \doteq 4.47 \text{ hr}^{-1} \doteq 1.24 \times 10^{-3} \text{ sec}^{-1}.$$

The particular case of the *circular equatorial orbit* yields the simplest expressions for azimuth and elevation angles and rates. Letting $\alpha = 0$,

we obtain from (48) and (49) the angle expressions:

$$[A(t)]_{\alpha=0} \equiv A^*(t) = \tan^{-1}\left\{-\frac{\tan\varphi(t)}{\sin\lambda}\right\} \tag{53}$$

$$[E(t)]_{\alpha=0} \equiv E^*(t) = \sin^{-1}\left\{\frac{r\cos\lambda\cos\varphi(t) - 1}{[1 + r^2 - 2r\cos\lambda\cos\varphi(t)]^{\frac{1}{2}}}\right\} \tag{54}$$

where

$$\varphi(t) = (\omega - \Omega)t + \theta_0 - \psi_0 . \tag{55}$$

Taking the time derivative of (53), the azimuth *rate* for circular equatorial orbits is

$$[\dot{A}(t)]_{\alpha=0} \equiv \dot{A}^*(t) = \frac{-\sin\lambda}{\sin^2\varphi(t) + \sin^2\lambda\cos^2\varphi(t)}\dot{\varphi} \tag{56}$$

where, from (55), $\dot{\varphi} = \omega - \Omega =$ constant *relative* angular velocity of the satellite with respect to the earth. The maximum azimuth *rate* occurs when $\varphi(t) = 0$, and is given by†

$$|\dot{A}_{max}^*| = \frac{(\omega - \Omega)}{\sin\lambda} \tag{57}$$

where, from (45) and (52),

$$\dot{\varphi} = (\omega - \Omega) \doteq 4.47\, r^{-3/2} - (\pi/12)\ \text{rad/hr} \tag{58}$$

the maximum elevation, E, also occurs when $\varphi(t) = 0$, and has the value,

$$E_{max}^* = \sin^{-1}\left\{\frac{r\cos\lambda - 1}{[1 + r^2 - 2r\cos\lambda]^{\frac{1}{2}}}\right\} . \tag{59}$$

Differentiating (54), the elevation *rate* for the case of circular equatorial orbits is,

$$\dot{E}^*(t) = \frac{-r\cos\lambda\sin\varphi(t)\,(r - \cos\lambda\cos\varphi(t))\dot{\varphi}}{(1 - \cos^2\lambda\cos^2\varphi(t))^{\frac{1}{2}}(1 + r^2 - 2r\cos\lambda\cos\varphi(t))} , \tag{60}$$

where $\varphi(t)$ and and $\dot{\varphi}$ are given by (55) and (58), respectively. We are interested in the maximum value of $\dot{E}$ when the satellite is visible, i.e. when $r\cos\lambda\cos\varphi(t) > 1$. Once the parameters r and λ have been specified, this can be determined from (60).

† If $I = 180°$, then $\dot{V}_{max} = \dot{A}_{max}$.

APPENDIX B

We derive the inclined and vertical channel error signals as functions of the waveguide error signals and the controlled antenna angles (I,V) in this appendix. The desired error signals are

$$\begin{aligned} \epsilon_1 &\triangleq I_r - I \\ \epsilon_2 &= V_r - V . \end{aligned} \tag{61}$$

Since we desire to make $\epsilon_1(t)$ to be within some small specified tolerance, we assume

$$\sin \tfrac{1}{2}(I_r - I) \doteq \epsilon_1/2$$

and (62)

$$\sin \tfrac{1}{2}(I_r - I) \doteq \sin I.$$

Using (62) and the trigonometric identity,

$$\sin \tfrac{1}{2}(A + B) \sin \tfrac{1}{2}(A - B) = -\tfrac{1}{2}(\cos A - \cos B),$$

we can write

$$\epsilon_1 = \frac{(\cos I - \cos I_r)}{\sin I} . \tag{63}$$

Using (8) and (63), one obtains

$$\epsilon_1 \doteq \frac{(\sin E_r - \sin E)}{b^2 \sin I} . \tag{64}$$

Using (1), the trigonometric identity for the difference between $\sin E_r$ and $\sin E$, and assumptions similar to those used above

$$\epsilon_1 = \frac{(E_r - E) \cos E}{b^2 \sin I} = \frac{\epsilon_v \cos E}{b^2 K_v \sin I} . \tag{65}$$

From (7) and some algebraic manipulation, we can write (65) as

$$\epsilon_1 = \frac{\epsilon_v}{K_v}\left[1 + a^2 \cot^2\left(\frac{I}{2}\right)\right]. \tag{66}$$

The vertical-axis error signal, ϵ_2 is found using a similar procedure. From (61) and (8),

$$\epsilon_2 = (V_r - V) = A_r - \Delta(I_r) - A + \Delta(I). \tag{67}$$

From (1), we have

$$\epsilon_2 = \frac{\epsilon_h}{K_h \cos E} + \Delta(I) - \Delta(I_r). \tag{68}$$

Using the trigonometric identity,

$$\tan \Delta(I_r) - \tan \Delta(I) = \frac{\sin (\Delta(I_r) - \Delta(I))}{\cos \Delta(I_r) \cos \Delta(I)} \tag{69}$$

and, assuming that for normal tracking $(\Delta(I_r) - \Delta(I))$ is small, we obtain

$$\begin{aligned} \Delta(I_r) - \Delta(I) = \tfrac{1}{2}[\tan \Delta(I_r) - \tan \Delta(I)] \\ \times \{\cos [\Delta(I_r) + \Delta(I)] \\ + \cos [\Delta(I_r) - \Delta(I)]\} \end{aligned} \tag{70}$$

using (69) and the trigonometric identity for the product of two cosines. Using (7) and assuming that

$$\begin{aligned} -\frac{1}{a}\left[\tan\left(\frac{I_r}{2}\right) = \tan\left(\frac{I}{2}\right)\right] &\doteq -\frac{\epsilon_1}{a[1 + \cos I]} \\ \cos [\Delta(I_r) - \Delta(I)] &\doteq 1 \\ \cos [\Delta(I_r) + \Delta(I)] &\doteq \cos 2\Delta(I) \end{aligned} \tag{71}$$

(70) becomes

$$\Delta(I_r) - \Delta(I) \doteq -\frac{\epsilon_1}{2}[1 + \cos 2\Delta(I)][1 + \cos I]^{\frac{1}{2}}. \tag{72}$$

After some additional algebraic manipulation using (7) and trigonometric identities, one obtains

$$\Delta(I_r) - \Delta(I) \doteq -a\epsilon_1[1 - \sin E]^{-1}. \tag{73}$$

Using (68) and (7), (73) becomes

$$\epsilon_2 = \frac{\epsilon_h}{K_h}\left\{b^2 \sin^2 I\left(1 + a^2 \cot^2 \frac{I}{2}\right)\right\}^{-1} + \frac{a\epsilon_v}{2K_v} \csc^2\left(\frac{I}{2}\right). \tag{74}$$

APPENDIX C

An approximate expression for $\hat{S}$ is derived in this appendix. Referring to Fig. 20, one can say

$$\sin [(\pi/2) - \xi] = [r \cos \psi_0 - 1][(r \cos \psi_0 - 1)^2 + (r \sin \psi_0)^2]^{-\frac{1}{2}} \tag{75}$$

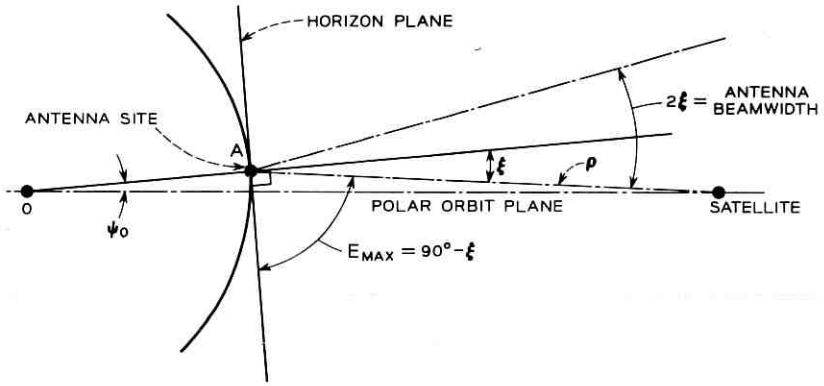

Fig. 20 — Tracking geometry for worst case slewing speed calculation.

Since ξ and ψ_0 are small angles, (75) can be written as

$$1 - \frac{\xi^2}{2} = \left[1 - \frac{r\psi_0^2}{2(r-1)}\right]\left[1 + \frac{r\psi_0^2}{(r-1)^2}\right]^{-\frac{1}{2}}. \tag{76}$$

Since $r\psi_0^2/(r-1)^2 \ll 1$, then (76) can be expressed in the form

$$\xi^2 = \frac{r\psi_0^2}{(r-1)} + \frac{r\psi_0^2}{(r-1)^2} + \cdots \tag{77}$$

where the remaining terms of the series are negligible compared to the first two terms in (77). Solving (77) for ξ,

$$\xi \doteq \frac{r\psi_0}{(r-1)}$$

or

$$\psi_0 \doteq \frac{(r-1)}{r}\,\xi. \tag{78}$$

Using (48), (49), (8) and (78) for the special case of a circular polar orbit of 6000 mile altitude ($r = 2.5$ e.r.) and $\xi° = 0.1°$, the vertical axis angle pointing directly at the satellite and boundaries on either side of the vertical axis angle which are allowable for tracking of the satellite were graphically determined and plotted in Fig. 15. The maximum velocity requirement, $\hat{S}$, was graphically determined by picking the minimum slope as shown in Fig. 15. For the particular example plotted,

$$\hat{S} \doteq 1 \text{ R.P.M.} \tag{79}$$

The velocity requirement for "boresight tracking," $\tilde{S}$, is obtained by taking the time derivative of (48) with $S_\alpha = C_\lambda = 1$ and evaluating at $t = 0$. The result is

$$\tilde{S} = | \dot{V}_{\max} |_{\alpha=90^\circ} \doteq \omega/\psi_0 \tag{80}$$

for small ψ_0 angles. Using equation (78), one can write

$$\tilde{S} = \frac{r\omega}{(r-1)\xi}. \tag{81}$$

Using (81) for a 6000 mile polar circular orbit, the comparable slewing requirement for "boresight tracking" is 2.85 rpm. Graphical plots for other orbit inclinations, altitudes, and antenna beamwidth angles indicate that tracking on the "edge of the antenna beamwidth," as illustrated in Fig. 15, reduces the required slewing capability for a given circular orbit by approximately $\frac{1}{3}$. Therefore, an approximate expression for $\hat{S}$ is

$$\hat{S} \doteq \tfrac{1}{3}\tilde{S}. \tag{82}$$

Using (81) and (52), (82) can be expressed as

$$\hat{S} \doteq \frac{2.3 \text{ R.P.M.}}{r^{\frac{1}{2}}(r-1)\xi^\circ} \tag{83}$$

where r is in earth radii and ξ°, the half-beamwidth angle, is in degrees.

REFERENCES

1. Cook, J. S., et al., The Open Cassegrain Antenna: Part I, Electromagnetic Design and Analysis, B.S.T.J., this issue, pp. 1255–1300.
2. Cook, J. S., and Lowell, R., The Autotrack System, B.S.T.J., *42*, July 1963, pp. 1283–1308.
3. Norton, J. A., Tracking Characteristics of a Conic Mount, Part 1, (Structural Analysis Series, SA-15), Unpublished work.
4. Claus, A. J., et. al., Orbit Determination and Prediction, and Computer Programs, B.S.T.J., *42*, July 1963, pp. 1357–1382.
5. Coyne, K. N., Open Cassegrain Antenna — Mathematical Model of Mechanical System, Unpublished work.
6. Lozier, J. C., Norton, J. A., and Iwama, M., The Servo System for Antenna Positioning, B.S.T.J. *42*, July 1963, pp. 1253–1281.
7. Semmelman, C. L., Description of Computer Facilities at Murray Hill, Proc. Eastern Simulation Meeting, Bell Telephone Laboratories, Murray Hill, N.J., June 15, 1965.

The T1 Carrier System

By K. E. FULTZ and D. B. PENICK

(Manuscript received May 11, 1965)

T1 carrier provides 24 voice channels by time division multiplexing and pulse code modulation (PCM). Each voice channel is sampled 8000 times a second and each sample is coded into a 7-digit binary word. Provision for signaling and synchronization raises the pulse repetition rate on the repeatered line to 1.544 × 10^6 pulse positions per second. The bipolar pulse train out of the terminals is transmitted over pulp, paper or plastic insulated paired cables by the use of regenerative repeaters. For 22-gauge cable pairs, repeaters are normally located at 6000-foot intervals.

The system has been designed for low cost and is being widely applied on many trunks interconnecting switching units within metropolitan areas. Western Electric Company manufacture of T1 began in 1962 and about 100,000 channels are now in service throughout the Bell System.

I. INTRODUCTION

The rapid expansion in the telephone network that has occurred since 1950 has stimulated a thorough investigation of methods for reducing the cost of additional trunk facilities. The desire to improve the quality of telephone service has given additional emphasis to studies of improved trunking arrangements. One way to obtain additional trunks for growth is to increase the utilization of existing conductors by using them to transmit more than one voice signal. For such an arrangement to be economical, the savings from the more efficient use of the transmission line must more than offset the cost of the terminal equipment required to multiplex a number of voice channels. On trunks between cities, carrier systems (systems transmitting a number of voice channels) have been economical for many years. The lower terminal costs achieved in the T1 carrier system

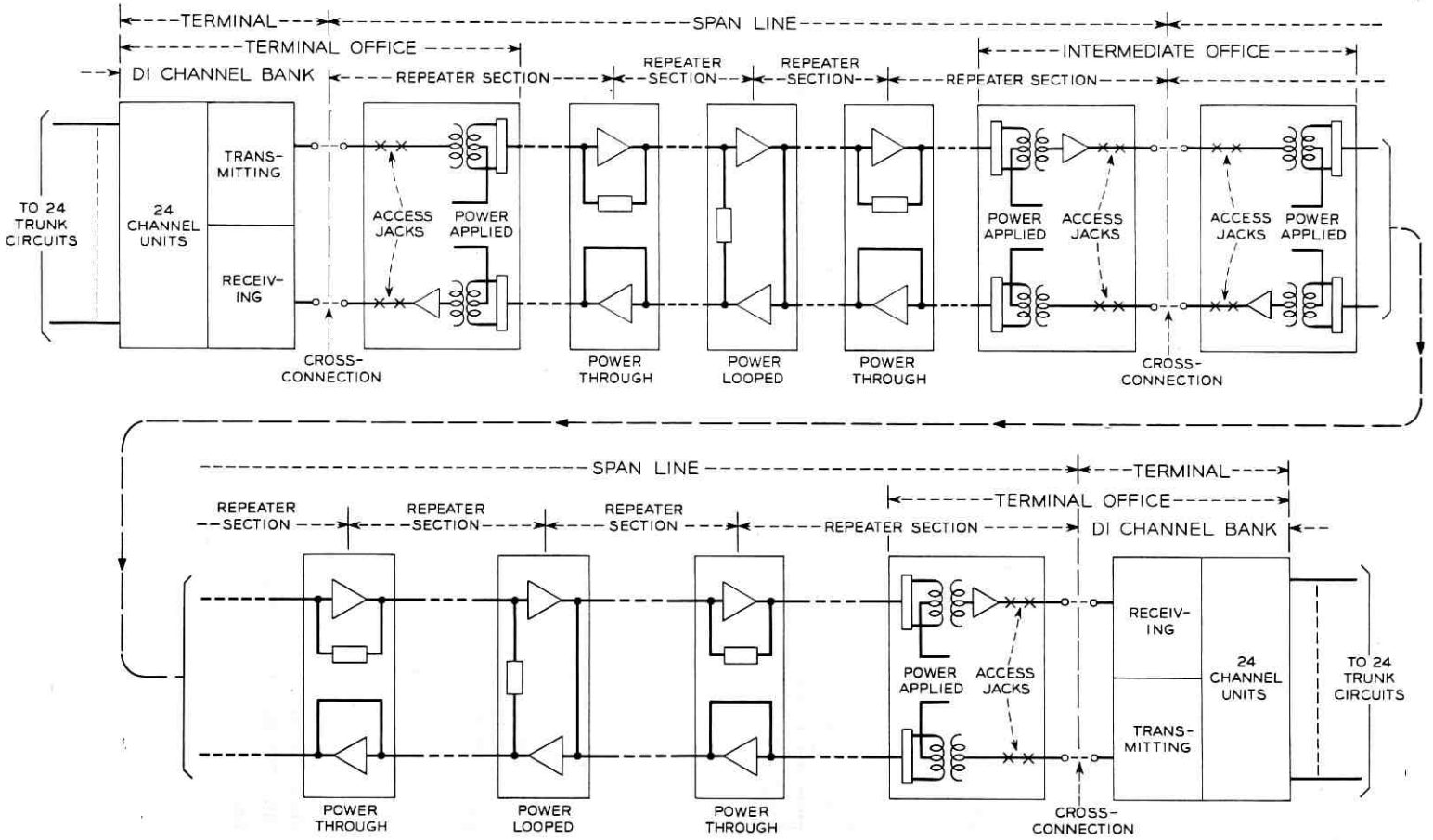

Fig. 1 — Typical T1 carrier system.

have made carrier systems economically attractive for the longer trunks between local offices within a city. In a large number of situations the T1 carrier system will prove-in over voice frequency circuits for distances longer than 10 to 12 miles. Satisfactory performance is achieved over lengths up to 50 miles, and the performance over longer lengths is being evaluated.

A major contributor to the low terminal costs in T1 is the economy with which the signaling information required to control the switching equipment can be transmitted in a digital system. In most carrier systems the digital signaling information is converted into analog tones for transmission. In a digital system the signaling information can be added directly to the coded speech samples with the saving in digital-to-analog conversion of the signaling information. Additional economies are achieved by an instantaneous compandor shared by a number of channels rather than individual channel syllabic compandors as used in some carrier systems.

The T1 carrier system now being manufactured by the Western Electric Company is a refinement of the experimental PCM system described in the January, 1962, issue of this Journal.[1-5] The basic system plan and the fundamental circuit approaches remain unchanged.

It is convenient to consider a PCM system as being composed of two parts — a PCM terminal and a digital transmission line. For regular telephone trunks, the PCM terminal for the T1 system is the D1 channel bank. The D1 channel bank combines 24 voice channels in a time division multiplex and encodes them in a scale of 127 quantized amplitude levels (63 steps positive and 63 steps negative from zero) into a single pulse train. In the receiving direction, it reconstructs the analog speech signals from the incoming pulse stream. Other terminal arrangements are being provided which prepare wideband data signals for transmission over T1 repeatered lines. These terminals are discussed in a companion paper.[6]

The T1 repeatered line consists of cable pairs equipped with regenerative repeaters at appropriate spacings. At the end offices, and at intermediate offices along the route, each repeatered line passes through an office repeater which provides a regenerator for the incoming signal, powering circuits for the line repeaters, access jacks for patching, monitoring jacks, and cross-connection points for route flexibility. A block schematic of a typical T1 carrier system is shown in Fig. 1.

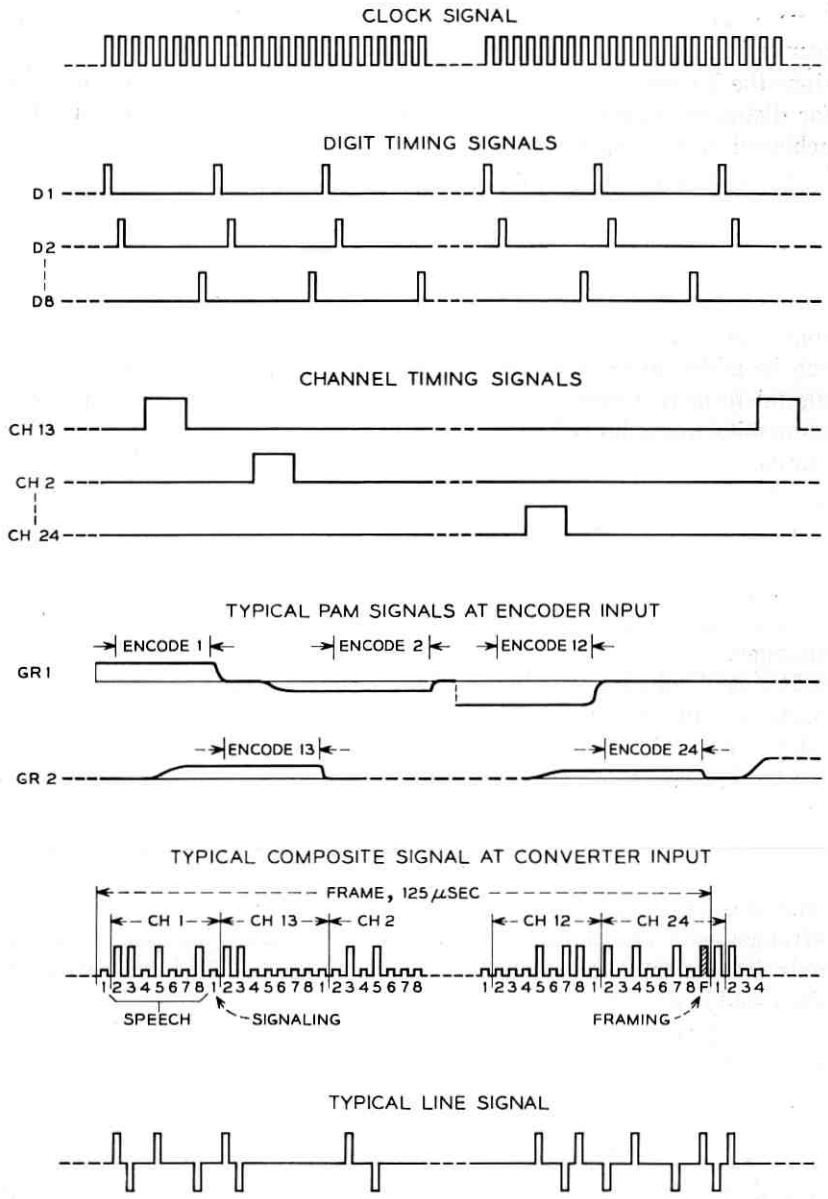

Fig. 2 — D1 bank pulse trains used in multiplexing and encoding.

II. D1 CHANNEL BANKS

2.1 *Group and Channel Circuits*

Most of the transmission functions in a D1 channel bank are performed in a block of circuits shared by a number of voice channels. These group circuits may be divided into two sections — transmitting and receiving. The transmitting group equipment samples the incoming voice signals for each channel, multiplexes the sample in time division, compresses and encodes the samples, combines the encoded sample with signaling information, and prepares the pulse train for transmission over the line. Fig. 2 shows the more important pulse trains involved in this process. The receiving group equipment accepts the incoming pulse stream, separates the signaling information from the coded samples, decodes and expands the speech samples, demultiplexes them, and reconstructs the voice signal. Thus, the group equipment provides 24 voice channels plus 24 signaling channels in each direction. Each signaling channel has a theoretical capacity of 8 kilobits/second. In some situations — revertive pulsing and foreign exchange lines — additional signaling capability is obtained by using the least significant* speech digit when speech would not usually be present.

The channel units shown in Fig. 1 are used to match the voice and signaling paths provided by the group equipment to the requirements of the individual switching circuits to which each channel is connected.

A block schematic of the group circuits is shown in Fig. 3. Consider first the transmitting direction shown in the upper half of the schematic. The transmission circuits in heavy lines come in at the left side from 24 plug-in channel units not shown. Six channels connect to each of four transmitting gate and filter plug-in units. Each gate and filter unit contains six low-pass filters and six sampling gates. The four gate and filter units are arranged in two pairs, a pair for each of two 12-channel groups. The sampling times of the two groups are interleaved so that group 1 channels are sampled at odd-numbered sampling times and group 2 channels at even-numbered times. Thus the channels appear in the PAM (pulse amplitude modulated) pulse train in the order: 1, 13, 2, 14, $\cdots$ 11, 23, 12, 24, 1, 13, 2, 14, $\cdots$.

The common output of each group of twelve gates connects to its own compressor, which reduces a wide range of input amplitudes to a

* The seventh digit of a seven-digit binary code is the least significant since it affects the coded amplitude by only 1 part in 128. The first digit affects the amplitude by 64 parts in 128, the second by 32 parts, etc.

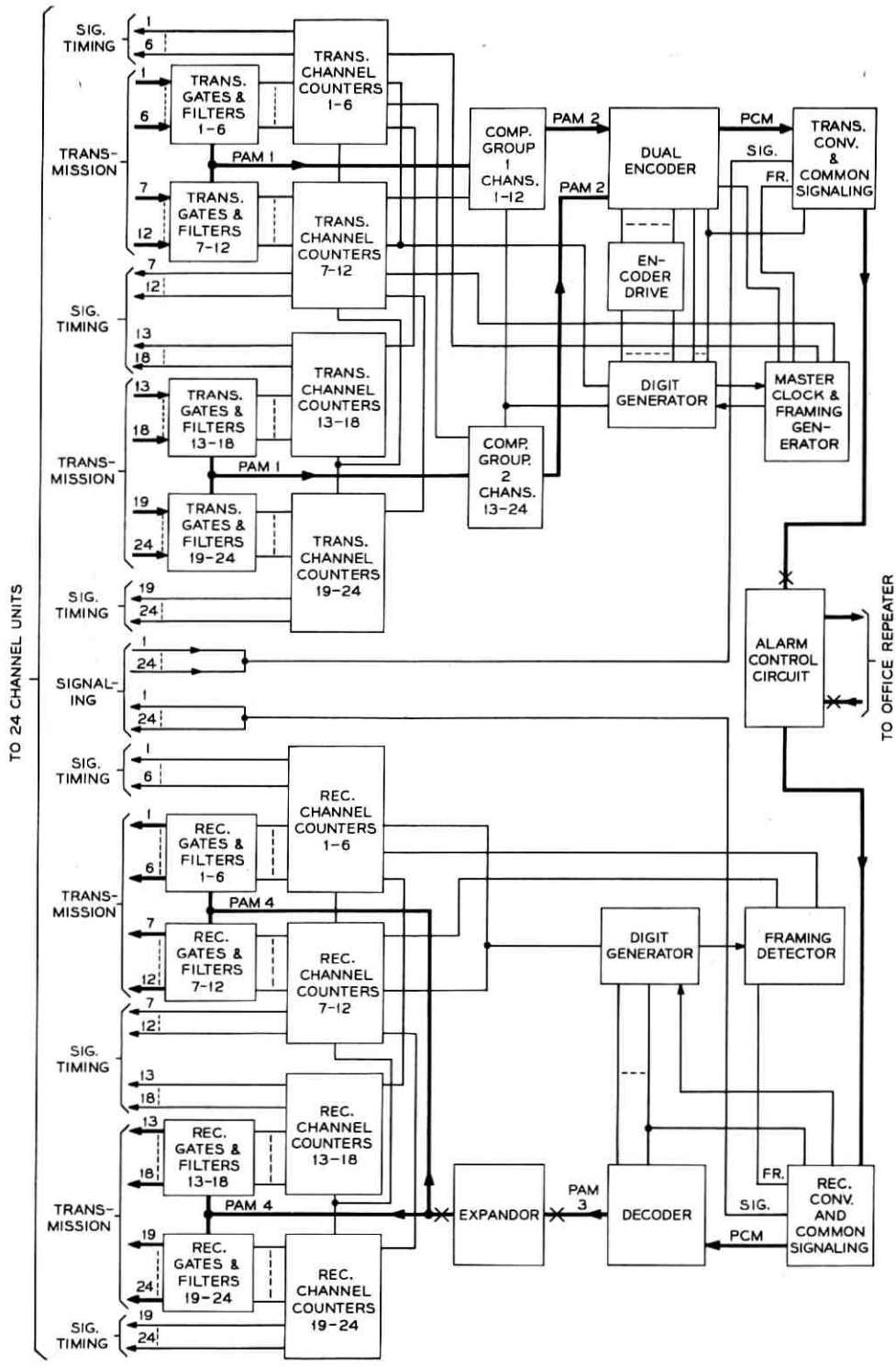

Fig. 3 — D1 bank group circuits.

smaller range of output amplitudes in an almost logarithmic relationship. An input range of approximately 60 decibels (1000-to-1 amplitude ratio) is reduced to an output range of 63-to-1 amplitude ratio in a modified logarithmic-to-linear conversion, so that the output variation in volts amplitude is approximately proportional to the input variation in decibels over most of the range. The two compressor outputs are connected to the dual encoder. Its two summing amplifiers and comparison networks, under the logic control of the encoder drive unit, encode the two pulse trains alternately into a single stream of PCM pulses.

This unipolar PCM signal, occupying seven of the eight pulse positions assigned to each channel sample, is one of three signals fed into the transmitting converter and common signaling unit. A second signal is processed by the common signaling portion of this unit, which accepts a signaling pulse from the scanning gate of each of the 24 channel units in turn, reshapes each one, and times it to interleave with the PCM pulses in the unipolar train. A third signal entering this unit is a framing signal from the framing generator, occupying a single pulse position per frame. These three signals, added together in the converter, form a combined pulse train of 193 pulse positions per frame. This number, multiplied by the frame repetition rate of 8000 per second, yields the basic pulse repetition rate of 1,544,000 per second.

As a final step in the converter processing before the pulse train is sent out over the repeatered line, each pulse is regenerated and alternate pulses, when they appear, are inverted to form a bipolar signal. This signal, then, is transmitted to the line by way of the alarm control unit located on the receiving shelf.

Timing for the signal processing circuits is derived from a crystal-controlled oscillator, a part of the master clock. The oscillator output, shaped into square-topped pulses, each occupying about one half of its allotted time interval, drives a digit generator which is basically a ring counter composed of blocking oscillators. Each of eight stages sends out one of eight successive digits on a lead per digit for use as required in encoding and other timing functions. A second lead per digit is also provided for digit pulses of opposite polarity. A ninth stage provides a ninth pulse at the end of each frame for framing control in conjunction with the framing generator included in the master clock unit.

Digit pulses, in turn, drive a set of channel counters which provide timing for both voice sampling gates and signaling scanning gates. As in the case of the digit generator, the counter stages are blocking

oscillators. They are turned on in rotation by one digit pulse and turned off by another. Each counter unit accommodates six stages, so that for a completely equipped D1 bank, four units are required. The circuits are arranged, however, so that the two units associated with the group 1 channels form a 12-stage ring counter which is self-sustaining. The two units for the group 2 channels are separately driven from the ring, and may be omitted in a partially equipped bank without disturbing the group 1 operation. Some of the functions of the group 1 counters are not required for group 2, so a separate network code, simpler and less expensive, is provided for group 2 only. The group 1 counter will also operate in group 2 positions, and therefore is conveniently used as a spare.

The interconnections of the plug-in units which make up the receiving portion of the D1 bank are shown in the lower half of Fig. 3. The combined pulse train from the distant terminal, transmitted over the repeatered line and through the local office repeater, is received by the alarm control circuit at the right side of the schematic. Reduced by a pad to a convenient amplitude, it is sent into the receiving converter and common signaling unit. At this point the pulse train is reconverted to unipolar form, regenerated, and impressed simultaneously on framing, signaling, and PCM circuits. These circuits time-select appropriate pulses from the combined pulse train for further processing.

The PCM circuit connects to the decoder, which scans the seven pulse positions allocated to each sample and synthesizes from the code the compressed sample amplitude for the corresponding PAM pulse. The resulting train of PAM pulses passes through the expandor, which restores the original, uncompressed amplitudes and transmits them to the bank of receiving gates. The gates, operating one at a time in rotation, route each PAM pulse through an individual low-pass filter to the receiving branch of its associated channel unit.

The signaling pulse associated with each seven-pulse code at the converter output is selected by the common signaling timing, is amplified to a suitable pulse amplitude and duration, and is passed to the bank of receiving signaling gates in the channel units. The gate in the appropriate channel unit transmits the individual pulse in each frame to its corresponding amplification and reconstruction circuit, also in the channel unit, and reproduces the signaling state corresponding to that which was scanned at the distant terminal for that channel.

Timing for the receiving circuits is very similar to that for the transmitting circuits except that the clock signal, instead of originating in a crystal-controlled oscillator, is derived from the incoming pulse

train itself acting on a tuned circuit resonant at the expected bit rate. The dissipation of the tuned circuit is low enough so that oscillation of the slave clock is maintained over moderately long blank periods in the incoming pulse train. The clock signal, produced in the converter as part of the pulse regeneration process, also drives a digit generator, a duplicate of the one in the transmitting circuit.

Digit pulses, as in the transmitting circuit, drive channel counters which time both the transmission gates and signaling receiving gates associated with the individual channels. Also, as in the transmitting circuit, a framing pulse is produced at the end of each frame as determined by the state of the channel counters. Thus, the bit rate, digit pulse rate, channel rate, and frame rate are identical with those in the transmitting circuit. Synchronism, once achieved, is therefore maintained indefinitely as long as the incoming pulse train is not interrupted.

Restoration of phase synchronism, or framing, after an interruption is accomplished under the control of the framing detector. This unit receives the framing signal generated in the receiving timing circuits and compares it with the corresponding signal in the incoming pulse train. The framing signal is a fixed pattern consisting of alternating ones and zeros in every 193rd pulse position, a pattern seldom duplicated for more than two or three frame intervals at a time in any other pulse position. When the framing detector comparison indicates a number of rapidly occurring differences between the received pattern and the local framing signal, a logic circuit starts a hunting action by inserting an additional pulse per frame in the local signal, thus comparing the local framing signal with each pulse position in turn of the incoming signal until the framing position is reached. When the two patterns match, the system is in frame and the hunting action ceases.

As noted earlier, the function of the channel units is to match the 24 sets of voice and signaling paths to the 24 individual trunk circuits to which they are connected at each end. A channel unit may provide a 4-wire terminating set and signaling converters for connecting conventional dc signaling to the carrier derived signaling channels or may connect the voice paths directly to a 4-wire trunk circuit. Instead of making numerous cross-connections at intermediate distribution frames to interconnect the specific terminating equipment required to implement a circuit order, a channel unit with the appropriate functions is selected and inserted in the carrier bay.

The use of channel units solely for matching the conditions on the

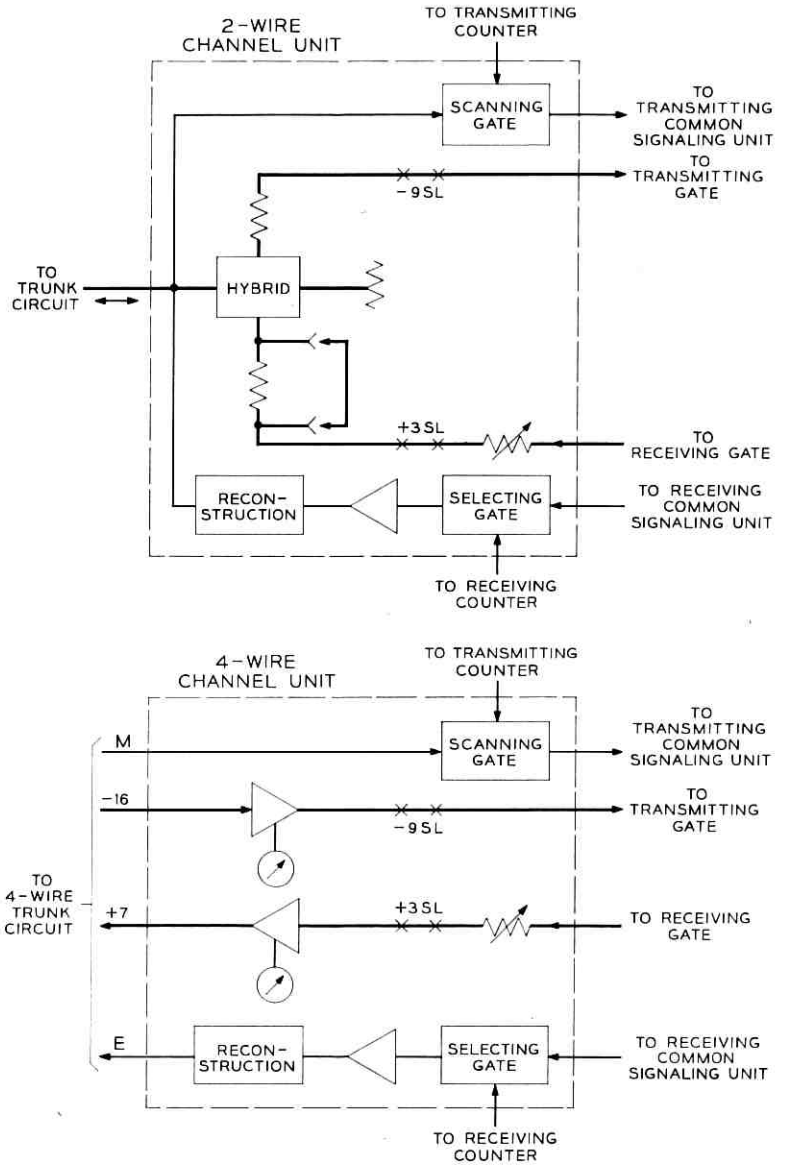

Fig. 4 — Typical D1 bank channel circuits.

voice frequency inputs to the carrier channels is quite different from the function of channel units in most frequency division multiplex (FDM) carrier systems. FDM channel units usually include filters which are different for each channel in a system. There is no difference in the T1 carrier channel units with respect to their position in the time division cycle. The different types of channel units required to meet local circuit needs may be intermixed in a channel bank in any order.

The two major types of channel units are the two-wire and the four-wire types as shown in the block schematic of Fig. 4. The two-wire channel units include a hybrid coil used as a terminating set. They also include transmitting and receiving access jacks for lineup use, a level adjusting pad, and two fixed pads, one of which may be strapped out. These elements constitute the transmission circuit and are the same for all two-wire units.

The four-wire unit transmission circuit does not require a hybrid coil, but provides an amplifier and an access jack in each direction of transmission, as well as a level adjusting pad in the receiving direction. The amplifier gains are adjustable over a range of about 1.5 db each for overcoming office wiring losses and are arranged to provide the nominal levels of −16 db and +7 db respectively, within 0.2 db, at the channel unit when the gain adjustments are turned to minimum.

The basic signaling functions for all channel units are the same. At the transmitting end in each direction, a scanning gate monitors the signaling state presented to it and converts it to a stream of corresponding signaling pulses, off or on, for transmission to the receiving end. There, a selecting gate recognizes the pulses, amplifies each one, and operates a reconstruction circuit which produces the signaling state corresponding to that scanned at the transmitting end. The differences between channel units lie in the methods required to translate the varying signaling states to pulses and reconstruct them again from pulses.

The most commonly used types of trunks in the exchange plant, which T1 carrier is designed to provide, are one-way trunks with either dial pulse or revertive pulse signaling and reverse battery supervision. The dial pulse signaling functions are quite straightforward. Loop closures are transmitted in the originating-to-terminating direction and battery reversals in the terminating-to-originating direction. In both directions, the digit 1 position in the train of eight pulses per channel is used to transmit the required information. Since the scan-

ning gate requirements and relay requirements are quite different for the two directions, it is convenient and economical to use different designs for the originating and terminating channel units.

The same design basis applies also for revertive pulse signaling. Here the loop closures and loop opens in the originating-to-terminating direction represent start and stop signals, respectively. In the terminating-to-originating direction, it is necessary to transmit both battery reversals for supervision, and loop closures for the revertive ground pulse during dialing periods. The second signal in this direction requires an additional scanning gate in the terminating unit and an additional selecting gate, amplifier, and reconstruction circuit in the originating unit.

It also requires another signaling state, provided by "borrowing" another digit in addition to the digit 1 normally provided. Since the added digit is not needed for signaling during the normal talking period, digit 8, the least significant of the 7 PCM digits, is used and is returned to the PCM function as soon as the called customer returns the normal supervisory signal. One result of this arrangement is that operator connections, or others which do not return supervision, will have only 6 PCM digits available for transmission. These added functions, of course, require two additional channel unit designs, one each for originating and terminating units.

A demand for foreign exchange trunk service over T1 has inspired the design of two more channel units, which are now available. They connect the line circuits at the serving office end and customer end, respectively. All three available signaling states are used in both directions of transmission. In the serving office-to-customer direction, a tip ground signal and a ringing signal are transmitted. In the customer-to-serving office direction, a loop closure signal and a ring ground signal are transmitted.

The four-wire channel unit is designed for symmetrical two-way trunks with identical signaling in the two directions. In either direction, ground and battery on the M lead at the transmitting end become open and ground, respectively, on the E lead at the receiving end. Thus, the same design of channel unit is used at both ends of such trunks.

The four-wire channel unit may also be used with existing trunk converter circuits to connect to any of a large number of other types of trunks for which specific channel units have not been provided.

It is also feasible to use a four-wire channel unit at one end of a T1 carrier circuit and a two-wire unit at the other end to avoid the use of a converter, which in some cases would otherwise be required.

2.2 *Bay*

The basic channel bank bay is 11 feet, 6 inches high and 23 inches wide. It mounts three D1 banks, each associated with one 24-channel system, with their associated power supplies. Fig. 5 is a photograph of a typical installation showing an unequipped bay and a working bay filled with plug-in units. The unequipped bay consists of a supporting framework, die-cast metal shelves for the plug-in units, multipin connectors, and terminal strips. The terminal strips and connectors, including special screw connectors for hanger-mounted power supply panels, are prewired and fully tested at the factory. A 9-foot bay mounting two D1 banks and a 7-foot double bay mounting three D1 banks are also available.

The cost of these unequipped bays is comparable to the cost of engineering and installing them. Since the engineering and installation costs per bay are lower when a number of bays are installed at a time, it is economical to install more bays than are required immediately. At installation, the voice frequency connections are wired to the distributing frame, the 1.544-megabit digital leads are extended to a 1.544-megabit cross-connect field on either the office repeater bay or a separate bay, and the −48 volt power leads are connected to the battery supply. When traffic requirements materialize, the more expensive plug-in units may be inserted in the carrier bays. At that time, office personnel will cross-connect the additional carrier-derived voice channels to switching trunk circuits and cross-connect the outgoing and incoming digital circuits to the appropriate repeatered line.

2.3 *Plug-In Units*

The active circuits are of three general classes: power supply, group timing and processing circuits, and channel units. The power supply consists of a dc-to-dc converter and regulators. It provides well regulated voltages of −24 v, +24 v, −42 v and +48 v from the −48-volt office battery. In general, the lower voltages supply the digital circuits and the higher voltages supply the dc stabilized analog circuits. The dc-to-dc converter and the analog voltage regulators serve all three D1 banks on a bay, but each D1 bank has its individual regulator for the digital circuit voltages.

The group equipment plug-in units are eight inches high and eight inches deep. One shelf, mounting fifteen units, is devoted to transmitting and a second shelf, mounting fourteen units, is used for receiving. Thus the group equipment for one D1 bank consists of 29 plug-in units mounted in about 16 inches of vertical bay space.

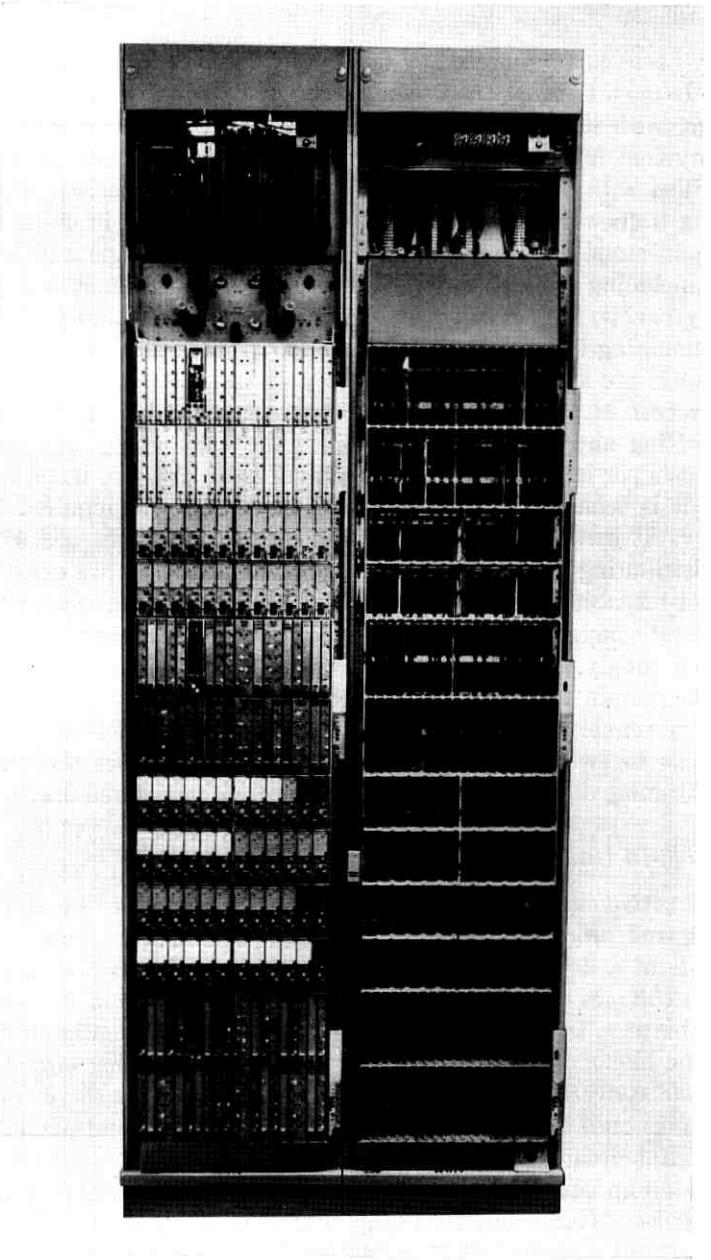

Fig. 5 — Typical D1 bank installation.

The 24 channel units in each D1 bank are mounted in two rows of twelve. Each channel unit is about 6 inches high, 8 inches deep and 1¾ inches wide. A photograph of a set of group and channel plug-in units in place for one D1 bank is shown in Fig. 6. Three typical plug-in units are shown in Fig. 7.

While the advantages of compactness are recognized, no major compromises were made in the equipment design for extreme miniaturization. Emphasis was placed instead on high reliability, design for

Fig. 6 — Group and channel units in place for one D1 bank.

mechanized assembly, mass soldering capability, and accessibility for inspection and repair. The book-case arrangement of units, however, uses the available volume efficiently.

One of the early choices in the arrangement of circuits and equipment was the consideration and rejection of the use of larger numbers of small, universal circuit packages such as gates, flip-flops, and blocking oscillators. The prospect of high production rates for relatively few types of basic modules is economically attractive. Analysis of the circuits showed, however, that each type of circuit block required so many variations for different points of application that the economies inherent in high production rate could not be realized. In the current design, the circuit packages are made large enough to include all of the specialized circuit blocks required to perform a larger circuit function. A digit generator, for example, uses nine similar blocking oscillators which operate in rotation. Several variations in these stages, however, would preclude using nine identical blocking oscillator packages unless each included all of the variations.

III. REPEATERED LINE

3.1 *Span Complements*

The digital transmission line, or repeatered line, extends from terminal to terminal of a system, and consists of two cable pairs equipped with repeaters for the two directions of transmission. The administrative line unit is a span line extending between office repeaters. A span line is composed of a number of repeater sections permanently connected in tandem at repeater apparatus cases mounted in manholes or on poles along the span. A span is defined[7] as the group of span lines which extend between two office repeater points. The repeatered line of the typical system of Fig. 1 is composed of two span lines, each of which happens to have four repeater sections.

Span lines are engineered, cable pairs assigned, and repeater mounting arrangements provided in multiples of 25-line complements. In one-cable installations (both directions of transmission in the same cable sheath), each set of apparatus cases along the cable serves a 25-line complement. In two-cable operation (the two directions of transmission in separate cable sheaths), two sets of apparatus cases, one for each cable, serve two 25-line complements. A single shop-wired repeater bay provides mounting arrangements for office repeaters for one end of three 25-line complements.

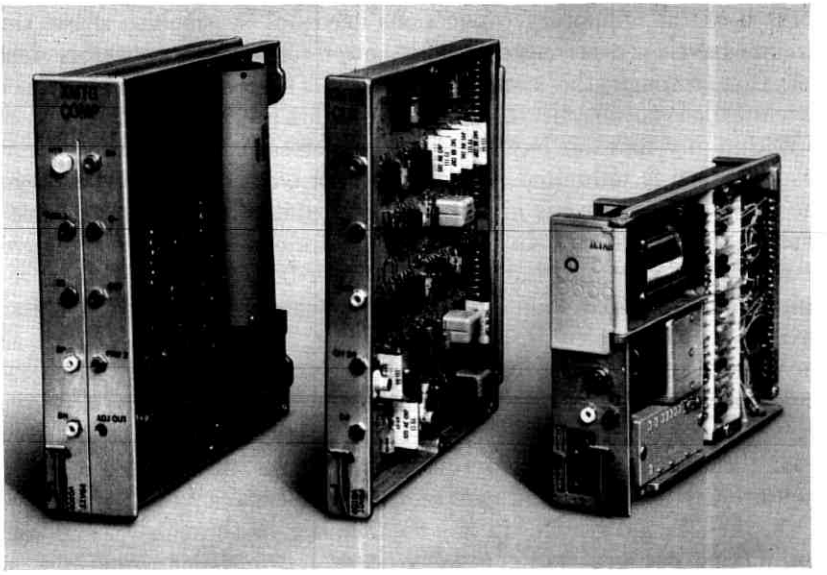

Fig. 7 — Typical D1 bank plug-in units.

Messrs. Crater and Cravis[7] have discussed the factors involved in the selection of cable sheaths and the spacing of apparatus cases.

The functions of the office repeater are to feed power to the repeatered line, regenerate the low-level incoming signal, provide jack access to the lines for patching and monitoring, and to provide a cross-connect field for connecting span lines to other span lines or to terminals.

3.2 *Cross-Connection Between Spans*

A cross-connect field is an integral part of the office repeater mountings. The 1.544-megabit connections to D1 banks and data terminals appear on this cross-connect field as well as the incoming and outgoing repeatered lines. All of these access points are designed to operate at the same signal level — namely, 3-volt pulses. To deliver a 3-volt pulse at the cross-connect field, D1 banks produce a 6-volt pulse which is padded down by 6 db in the combination of an equalizer on the D1 bank bay plus the cable from the D1 bank bay to the cross-connect field on the office repeater bay. With no level difference involved, all cross-connections can be made with unshielded twisted pair with no requirement for spacing or segregation. Similarly, patch

cords used for temporary connections need not be shielded. Also, the circuits have been arranged so that power for the line repeaters does not pass through the cross-connections. Thus, cross-connection or patching does not disturb the line power within a span.

Within any repeater bay, flexibility for cross-connecting any circuit to any other is unlimited. Where repeater bays are side by side in the same lineup, there is also complete flexibility for inter-bay cross-connections. Where repeater bays are separated, judicious assignment of routes can minimize the number of inter-bay cross-connections required, and a limited number of such cross-connections can be handled by means of tie cables. In general, however, flexibility is restricted. In such cases, complete flexibility can be gained by installing a central cross-connect field. A field is available which mounts in two bay spaces, has all terminals for cross-connecting placed less than 7 feet from the floor, and can be provided with demountable doors where they are desired for appearance reasons. The capacity of a single unit of this field is 450 systems in any combination of through systems and terminating systems, and can be extended indefinitely in multiples of 450 systems by adding more units side by side. All office repeaters, data terminals, and D1 Banks to be cross-connected are wired to this field. The only present restriction to the use of this or any other cross-connecting method is that for through systems, the total length of office wiring between the two interconnected span terminating repeaters may not exceed 150 feet.

3.3 *Line Repeaters*

The design of an experimental line repeater and the engineering of the repeatered line have been described in earlier issues of this Journal.[2, 7] The line repeater developed for manufacture is substantially the same as the experimental unit with some refinements. Fig. 8(a) shows the configuration of a typical repeater and Fig. 8(b) is a block schematic of a regenerator. Each repeater contains two regenerators, mountings for two line build-out networks, a powering circuit to provide a regulated voltage to operate the active transmission circuits, and four option screws, L, L, and T, T, used in pairs for either looping the line power or connecting it through as required.

Each regenerator contains a preamplifier, a threshold bias circuit, a clock rectifier, a clock signal processing circuit, timing gates, and a pulse generator. The preamplifier amplifies and equalizes the incoming signal, reshaping each pulse to reduce its dispersion into adjacent time slots. Its output drives not only the regenerating circuit but also the

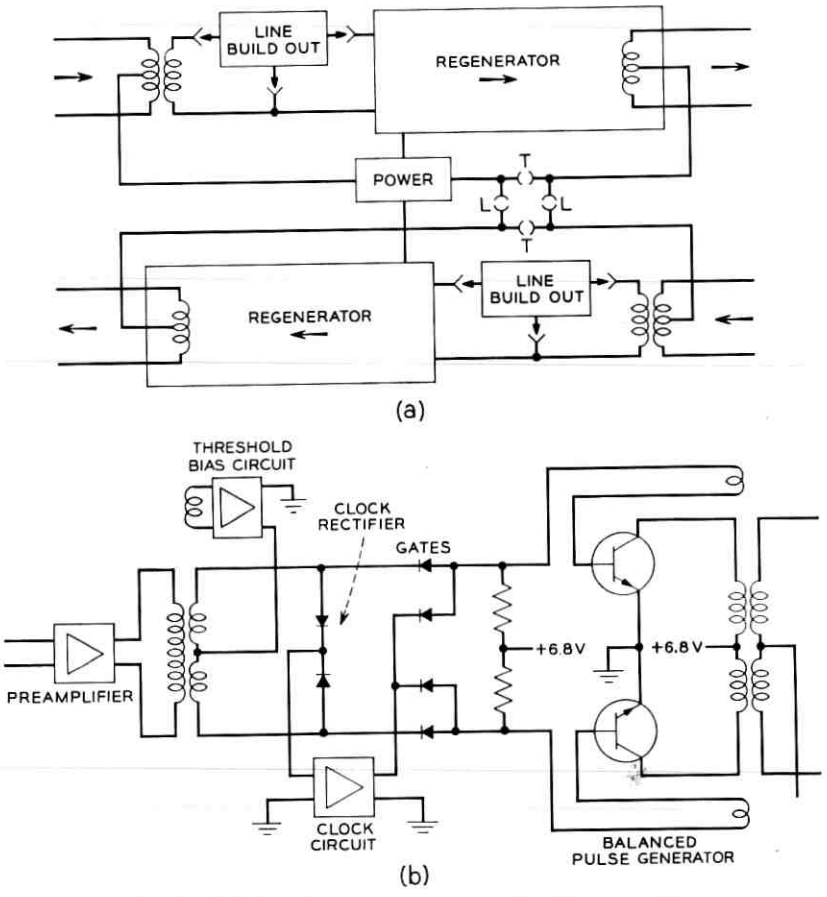

Fig. 8 — (a) 201A repeater configuration; (b) repeater regenerator.

clock circuit and the threshold bias circuit. The threshold bias circuit sets the decision level which determines for each time slot whether or not a pulse is to be regenerated, and optimizes it over a moderate range of variation of incoming signal level. The clock rectifier converts the incoming bipolar signal into a unipolar pulse train which contains a strong component of energy at the original repetition rate of 1.544 megacycles.* This 1.544-megacycle component is selected by a tuned

* In order to limit the time interval during which the clock must sustain the pulse repetition rate, the all-zeroes code of the transmitted signal is not used. This reduces the number of available codes from 128 to 127 and assures that in each eight-digit word there will be at least one pulse.

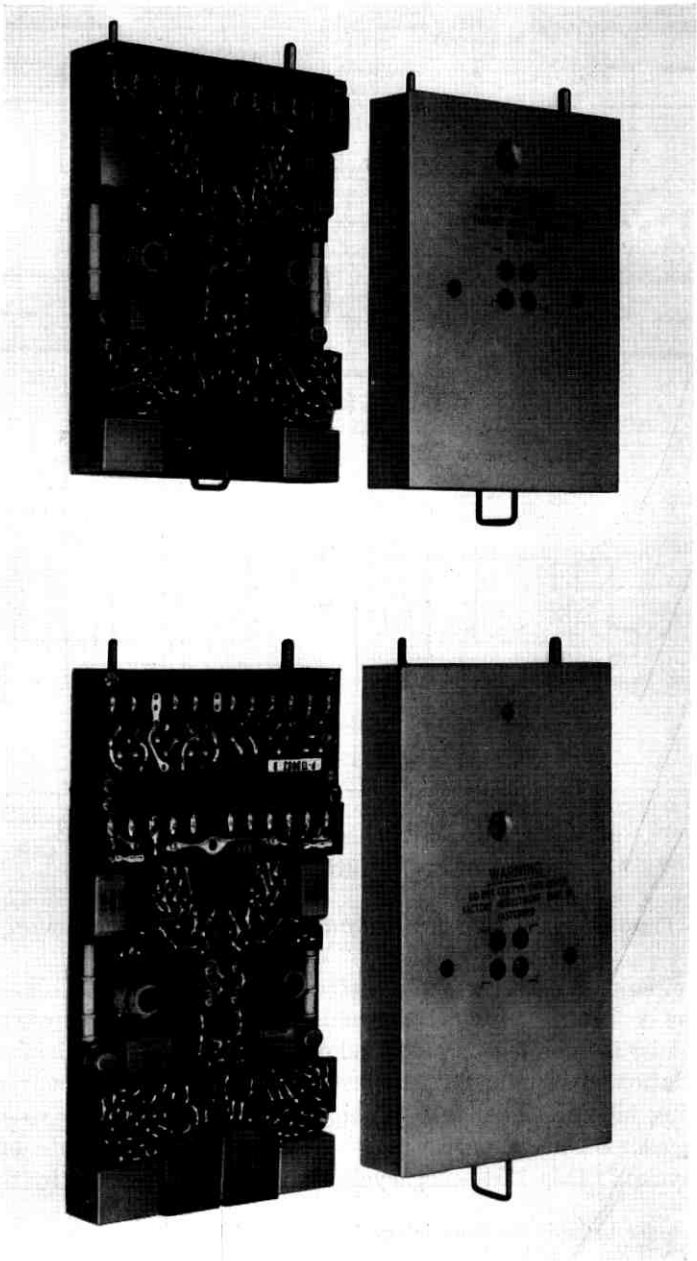

Fig. 9 — 201A and 205A line repeaters.

circuit, amplified, and shaped in the clock circuit to provide a turn-on and turn-off timing pulse for the beginning and end, respectively, of each pulse position. Two gates are provided, one for each polarity of signal pulse, and corresponding blocking oscillator pulse generators. Whenever an incoming signal pulse of either polarity coincides in time with the turn-on timing pulse, the pulse generator of the same polarity is triggered to send out a new pulse. With a normal, error-free signal, the gates and pulse generators operate alternately.

The production units are coded as 201A and 201B repeaters without surge protection and 205A and 205B repeaters with secondary surge protection. Fig. 9 is a photograph of 201A and 205A repeaters with and without covers. The A codes are for use in one-cable installations (both directions of transmission in the same cable sheath) and the B codes are for two-cable installations (the two directions of transmission in separate cable sheaths). The A and B codes are indentical except for the wiring of the connector and the power looping options. With a single cable-splicing pattern for all repeater cases, bi-directional or uni-directional operation in the cable may be chosen by filling all cases in a span with one code or the other.

Repeaters in underground cable systems without aerial exposure do not require lightning surge protection. If there is aerial exposure, however, protection must always be provided. The primary protection is the standard carbon block type which limits the maximum longitudinal or metallic surge to 600 volts peak. In addition, the 205-type repeaters contain secondary protection which consists of a series string of parallel oppositely poled diodes bridged across each incoming pair with 5.6 ohm current-limiting resistors in series with each wire on the line side of the diode string. This arrangement effectively limits the surge to 50 amperes, a tolerable value.

Line repeaters are normally mounted in manholes or on poles in complements of 25 in cylindrical, hermetically sealed apparatus cases. The repeater positions are arranged in five columns of five repeaters each. The 25 connectors, placed in a rectangular pattern at the back of the supporting structure with its 25 guide slots, are wired in the factory to a stub cable which is spliced by normal techniques into the main cable. Any pairs in the apparatus case not used for T1 carrier may be used for voice transmission by plugging into the appropriate positions either through connectors or voice loading coils as required.

Two sizes of case are available. The 466-type, about 10 inches in diameter, is used for 201-type repeaters, and the 468-type, about 2 inches larger in diameter, for the larger 205-type repeaters. The 468-

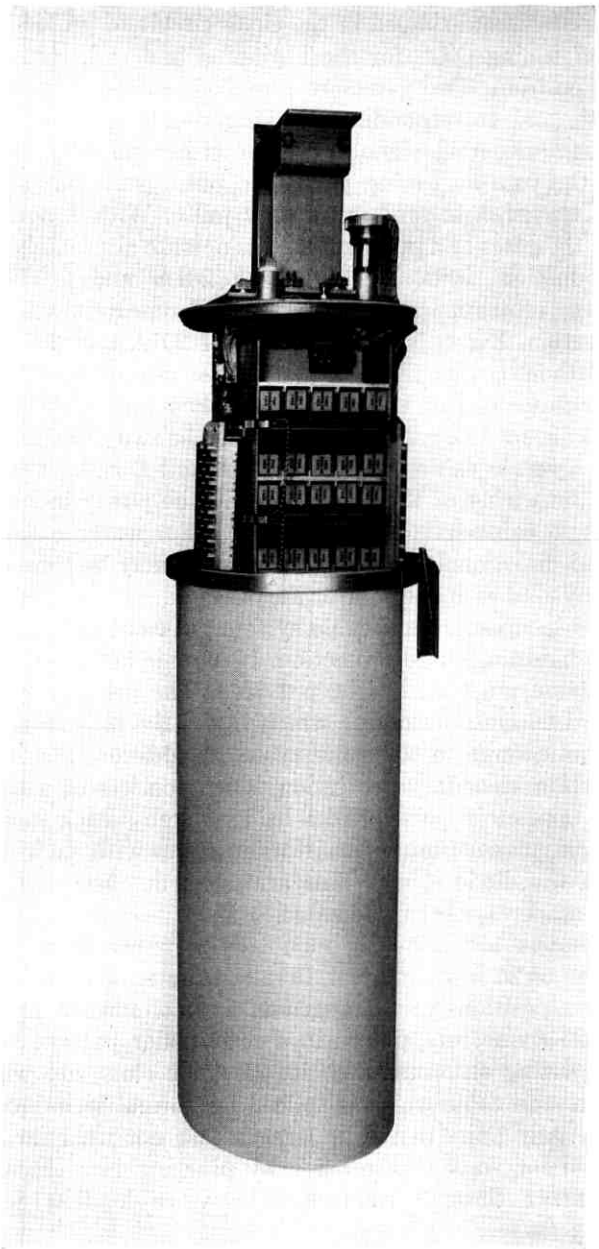

Fig. 10 — 468A apparatus case with cover partly removed.

type case also provides carbon blocks, one on either side of the repeater structure. Both cases provide a fitting for order wire terminals and a shelf for mounting a fault-locating filter which connects to all repeater positions in parallel. Fig. 10 is a photograph of a 468A apparatus case with the cover partly removed to show two rows of repeaters. For cramped locations, either the 466 or 468-type case is available with the cover in two sections fitted together with a gasket and O-ring.

For special service routes not expected to grow beyond four lines plus one spare, a shorter case of each diameter is being standardized to accommodate a single row of 5 repeaters. All other features of the 466A and 468A cases are also provided in the shorter cases.

Power for a string of repeaters is provided from a central office over a simplex loop consisting of the two pairs which feed through each of the repeaters. Looping may be done at any repeater by using the screw options provided, or may be done at a distant central office. Looping at a repeater is illustrated in the system block schematic of Fig. 1. The line current also serves as sealing current for unsoldered splices. Each repeater has a nominal voltage drop of 10.6 volts, and the voltage drop due to the resistance of a normal line section is of the same order of magnitude, depending on its length and wire size. The power is supplied from available office batteries, usually −48 v, +130 v, or −130 v as required. For the longest loops, two batteries of opposite polarity may be used for the two ends of the power loop. Line current is maintained at 140 milliamperes by a current regulator in each office repeater. In older installations the line current is adjusted by a fixed value building-out resistance at the central office so that the nominal line current at the highest expected line temperature is 140 milliamperes. At lower temperatures, the current is larger and may be as high as 200 milliamperes at −40 degrees for aerial cable. The repeaters are designed to operate over a temperature range of −40 degrees to +140 degrees Fahrenheit. This is the range expected to be encountered in pole-mounted repeater cases. Manhole temperatures range from about 25 degrees to 85 degrees F.

3.4 *Office Repeaters*

Two physical arrangements exist for mounting office repeaters: a shop-wired office repeater bay, and an installer wired assemblage of repeater mounting panels. The shop-wired office repeater bay has only recently been placed in manufacture. Most of the present office re-

peater bays consist of combinations of bank terminating assemblies and span terminating assemblies. Both arrangements provide the functions indicated above for the office repeaters, and both arrangements are capable of the nominal assignment and cross-connect philosophy presented above. The line capacity, jack access points and physical arrangement of regenerators in the shop-wired bay are more convenient than in the older arrangement. However, the two arrangements are fully compatible in the same office.

The outstanding difference between the two arrangements is that the plug-in units for the new bay are specialized office repeaters. These office repeaters each contain a single regenerator, an individual line current regulator, transformers for simplexing the power onto the line, and access jacks. The old span terminating assemblies mount regular 201B line repeaters and a separate control circuit for administering the line powering. The jacks are part of the mounting assembly. Since the 201B repeater contains two regenerators and regenerators are used only on incoming span lines, two incoming span lines are coupled together physically in a common office repeater. In two-cable operation, this coupling is not too objectionable because it also occurs in all line repeaters, but it has greatly complicated the administration of one-cable installations.

Operationally, the other major difference is that both incoming and outgoing jacks for each line are provided in the new bay where only incoming jacks were provided before. With the older arrangement, the only access to an outgoing line is at the circuit to which it is cross-connected. In addition to simplifying spare line patching, the additional jacks allow temporary cross-connections to be made via patch cords and eliminate the need for the access jacks on the bank terminating assembly.

The new repeater bay terminates 75 span lines (three complements of 25) while the older bay only terminates 72. Both bays are arranged for connection of any or all of these span lines to D1 banks or data terminals. Thus they possess a cabling capacity for connecting to 1.544-megabit office equipment of 150 or 144 pairs, respectively (half incoming and half outgoing). When a central cross-connect field is used, this cabling capacity is used to extend the span lines to the central cross-connect field, and D1 banks or data terminals are terminated directly on the cross-connect field.

IV. PERFORMANCE

Prototype models of T1 equipment were field tested between Newark and Passaic, New Jersey. Trunks between various combinations of

switching units in the Newark and Passaic buildings were routed over T1 carrier channels. Commercial service was provided over these trunks for about one year beginning July 25, 1961. These field tests, which employed an underground cable, were followed by field tests over an aerial cable in Akron, Ohio, during the last half of 1962. Early production terminals and repeatered lines were used in Ohio.

Measurements made during the New Jersey trial, the Ohio trial, and at a number of the early installations indicate that the following performance is representative of properly functioning T1 carrier systems:

4.1 *Channel Frequency Characteristic*

Fig. 11 shows the frequency response of a typical 2-wire T1 channel. Low-frequency loss has been deliberately added to improve the low-frequency stability of the channel and to control the generation of quantizing noise in the channel resulting from the quantization of input low-frequency noise such as power line hum. A rather gradual roll off at the top edge of the band has been accepted as a compromise in lieu of higher filter costs. For completeness, the transmitter and receiver contributions to the over-all frequency characteristic are also shown in Fig. 11.

Except for input signals above +3 dbm at 0 db system level (SL), the channel frequency characteristic is essentially independent of input signal level.

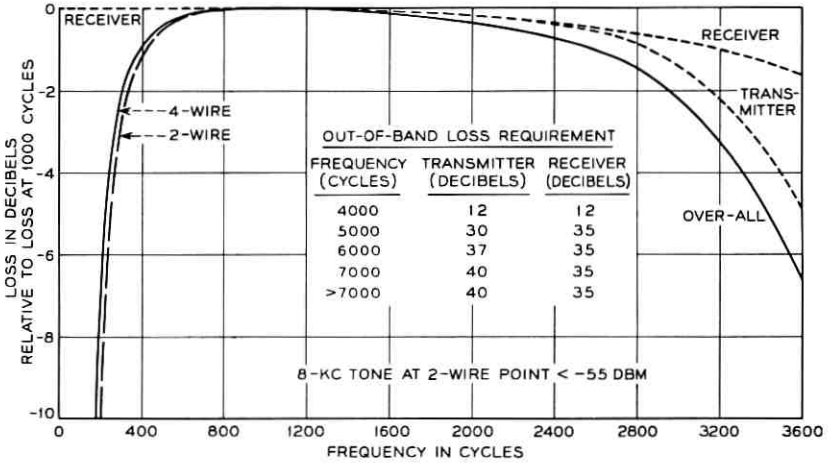

Fig. 11 — Frequency characteristic of a typical T1 carrier channel.

4.2 *Load Characteristic*

The peaks of a +3 dbm sine wave applied at the outgoing switch of a trunk composed of a T1 carrier channel will just cover the full amplitude range of the coder. Larger voltage peaks will be clipped. T1 carrier limits the maximum input signal at a lower level than has generally been common for carrier systems. In frequency division carrier systems the primary problem in achieving load capacity is the generation of intermodulation products in the common amplifiers. If a large number of channels are multiplexed together, the average talkers contribute the primary portion of the load on the common amplifier rather than the occasional high peaks of a few loud talkers. Therefore, a common amplifier designed for the bulk of the talkers will pass occasional high peaks in a individual channel without significant penalty. Provision of the dynamic range in the channel equipment is not expensive. Consequently, rather high overload capabilities have usually been provided. With individual channel coded PCM, the consequences of the choice of full load signals are more specific. Increasing the overload by 6 db costs another digit in the code group, if other performance characteristics are to be maintained; or, looking at the choice in a more practical way, for a fixed compandor characteristic and a fixed number of digits in the code group, a direct exchange may be made between full load signal amplitude and noise in the absence of signal. The subjectively most satisfactory choice is to reduce the noise in the absence of speech to a low level. Therefore, both the full load signal and the noise in the absence of signal are somewhat lower for T1 carrier than for many other carrier systems.

The general shape of the load characteristic is given by Fig. 12. Individual channels will display fine ripples on this basic characteristic caused by the particular match of the input signal to the quantizing grid. Also, if the channel net loss is measured at a frequency very close to a submultiple of the sampling frequency, time variations in the net loss will be observed as the phase relation between the input signal and the sampling time varies.

4.3 *Noise in the Absence of Signal*

Subjective evaluations of telephone transmission are made on the basis of articulation and naturalness of the received speech when speech is being transmitted and on the basis of noise and crosstalk when no speech is being received. It has long been recognized that the noise may be allowed to be higher in the presence of speech than in its

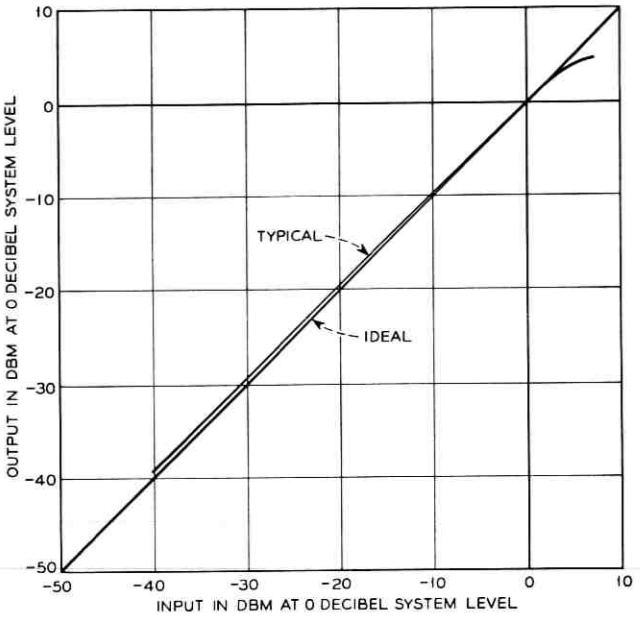

Fig. 12 — Load characteristic of a typical channel.

absence. It also appears that if the articulation and naturalness of the circuit are reasonably good, practically all of the subjective evaluation is made during the absence of speech. Therefore, this characteristic is very important in the subjective evaluation of most modern telephone systems. Because of its importance, it was concluded that the noise in the absence of speech should be comparable to the noise on existing interoffice trunks provided by cable pairs and type E repeaters.

When no direct signal is being applied at the input to a PCM system, the quiescent input voltage to the coder will be within a half quantum step of the decision point between two code levels. Ideally, the quiescent voltage would always be exactly a half step from the decision point, but in practice it may be much closer. If spurious voltages cause the coder to switch back and forth between two codes, the decoder will generate a square wave with a peak-to-peak amplitude equal to the step size.

It was decided that a reasonable size for the quantizing step at the origin was about 70 db smaller than the full coder amplitude range. Since a +3 dbm sine wave encompasses the full dynamic range, a

−67 dbm sine wave will just fill the voltage range from code 63 to code 64 (the first code step from zero). If the coder is switching between 63 and 64, the power in the square wave at the decoder output will be −64 dbm because the power in a square wave is 3 db larger than the power in a sine wave with the same peak-to-peak amplitude. If the transitions between code 63 and 64 occur at random time intervals, the output will approximate random noise. Noise at the system input may trigger these coder transitions when the quiescent input voltage is very close to a decision level. In this situation the output noise will be approximately 24 dbrn at zero level C-message weighting. If the quiescent input voltage is nearly midway between decision levels, input noise may not switch the coder, and the input noise is suppressed. In this situation the major contributor to channel noise is the receiving amplifiers and gates.

If crosstalk from test tones or other sources of periodic signals should happen to control the switching between codes, a periodic wave may appear at the channel output. If the fundamental of this wave is in the middle of the audio band, the weighted noise output may rise to 26 dbrn C-message.

Fig. 13 gives a distribution of the observed output noise at 0 SL, with the opposite end of the channel terminated, as measured by a Western Electric 3A noise measuring set with C-message weighting. Since the amount of noise in a channel depends upon the relative position of the decision level and the quiescent input to the channel, it is a function of individual channel gates and biases. This gives rise to the variation between channels in a system. Since these biases will also change with time, the noise in a particular channel will change with time. Recordings of noise on the Newark-Passaic systems show that the noise power on a particular channel was quite stable with time, indicating that these bias voltages change significantly only over a period of several days.

4.4 *Signal-to-Distortion Ratio*

Fig. 14 shows the signal-to-distortion ratio for a channel as a function of input signal level. This curve is the result of a measurement of the total distortion power introduced by the channel when a sine wave of 1100 cycles is applied at the input. The distortion power is the sum of the quantizing error power for the compression characteristic employed and system imperfections.

Subjective evaluations, using regular telephone sets, of the received

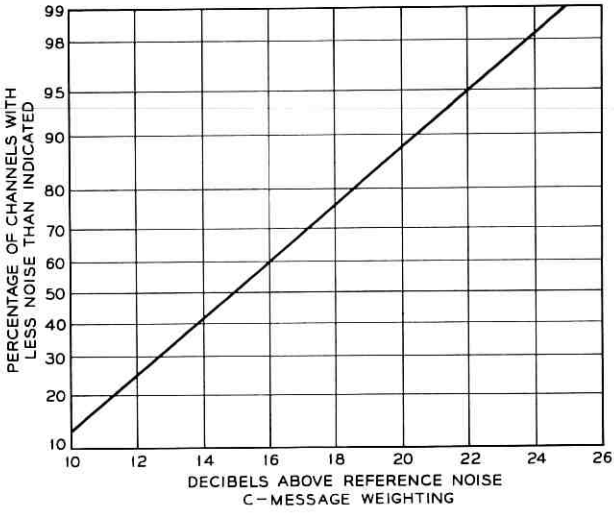

Fig. 13 — Noise in the absence of signal, referred to 0 db system level (far end of channel terminated).

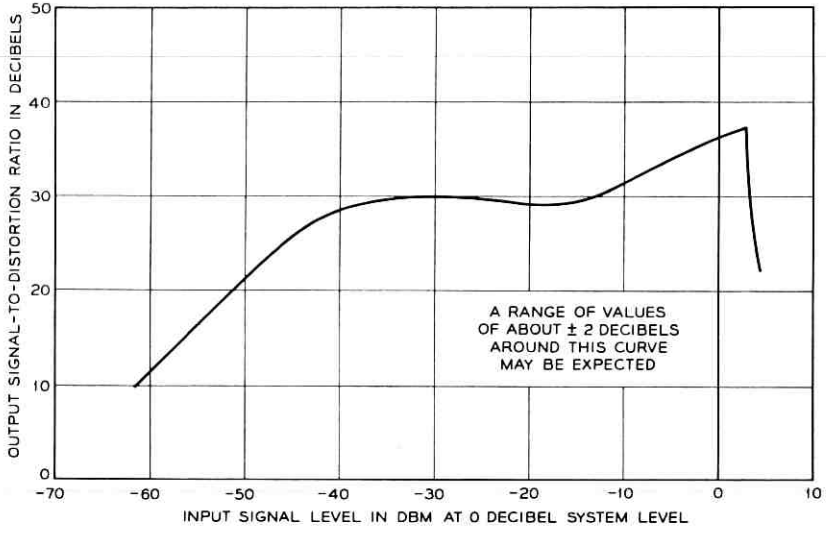

Fig. 14 — Signal-to-distortion ratio for a typical channel.

quality of speech that has been transmitted over a T1 channel have shown that a single T1 link is generally indistinguishable from a physical circuit with 20 dbrn of white noise and the same loss and bandwidth for speech levels between −40 and −10 VU. Clipping distortion is discernible at higher speech levels and critical observers experienced in the sound of quantizing distortion can discern the effects of quantization at lower speech levels. However, the quality is satisfactory over the speech volume range from 0 to −50 VU.

V. SYSTEM ALIGNMENT AND MAINTENANCE

In T1 carrier systems, as in most other multi-channel transmission systems, the techniques required for aligning and maintaining the transmission medium are quite different from those required for terminals. Lines and terminals may even be administered by separate organizations in the operating company.

When a line and its terminals are connected together and put into service as a working system, still another class of problems arises. One problem is the need for immediate indication of a system failure and rapid identification of the defective part so that service may be restored quickly. Another problem, where channels connect to switching systems, is the sudden load imposed on the switching equipment when many channels fail simultaneously and as a result of the failure send false signals interpreted by the switching mechanism as multiple demands for service.

Treatment of these considerations in the T1 carrier system will be discussed in the order mentioned: lines, terminals, system failures, and trunk processing.

5.1 *Repeatered Line Maintenance*

The stub cable which connects the repeater apparatus case to the main cable contains four pairs in addition to the transmission pairs required for the 25 repeater positions. One pair in the stub cable connects the fault-locating filter in the case to the fault-locating pair in the main cable as described in another paragraph in this section. One pair is a spare. The other two pairs are used, at least in the first apparatus case at each repeater location, to carry through a voice order wire which connects, through an access position on the office repeater bay at one end of the span, to office battery and to the switching circuits in the office. A lineman can bridge the order wire terminals on the outside of the case with a portable telephone set and dial any de-

sired number to request assistance. He can also establish a talking circuit with another lineman on the same order wire and drop off the office switching selectors for the duration of his usage of the line.

Each repeater is adjusted in the factory to an optimum slicing level for an incoming signal which has been attenuated by a normal line having 31 db loss at 772 kilocycles. A threshold bias circuit in the repeater described earlier maintains near optimum slicing level over a range of signal variation of ±4 db from this nominal. For optimum repeater performance, each line must be built out to within a few db of the nominal 31 db loss by installing, in the associated repeater, one of a series of 836-type line build-out networks, which are available in sizes from zero to 24 db in 2.4-db steps. To select the proper size, measurement of the transmission loss of each pair from apparatus case to apparatus case is made using a 113A test set at each end. The test set is a small, battery-powered unit containing a crystal-controlled oscillator with preset output amplitude and a calibrated detector marked to indicate directly the code of the repeater build-out network needed for the repeater connected to the pair being measured. Access to the pairs is through a fixture which is inserted into a repeater position and is connected to the set by a flexible cord. The test frequency is 650 kilocycles, near the peak of the normal signal energy distribution characteristic, but differing from it to avoid interference into working systems. A photograph of a 113A test set is shown in Fig. 15.

Before repeaters are installed in a line, they are given a pre-installation test using a repeater test set together with a fault-locating set and an error-detecting set. These three sets are shown in Fig. 16. The combination of sets applies a repetitive pulse train at a nominal input level to the repeater under test, adds a controlled amount of interfering signal, monitors the transmission of pulses, and tests for bipolar violations. The repeater test set also makes a voltage breakdown test of the insulation between repeater circuit and case, and tests for transmission of pulses after insertion of the selected line build-out networks.

When the line has been powered, ready for use, its operating capability is established by two types of test. The first type, which can be made at any time without disturbing the transmitted signal, is a bridging test to determine that pulses are being transmitted and are free of errors. The transmitted signal may be derived either from a working terminal or from a test set. The presence of pulses and of errors, if there are any, is indicated by the error-detecting set. This set, which is powered by the 48-volt office battery, may be bridged across the output of any terminal or office repeater by patching into a moni-

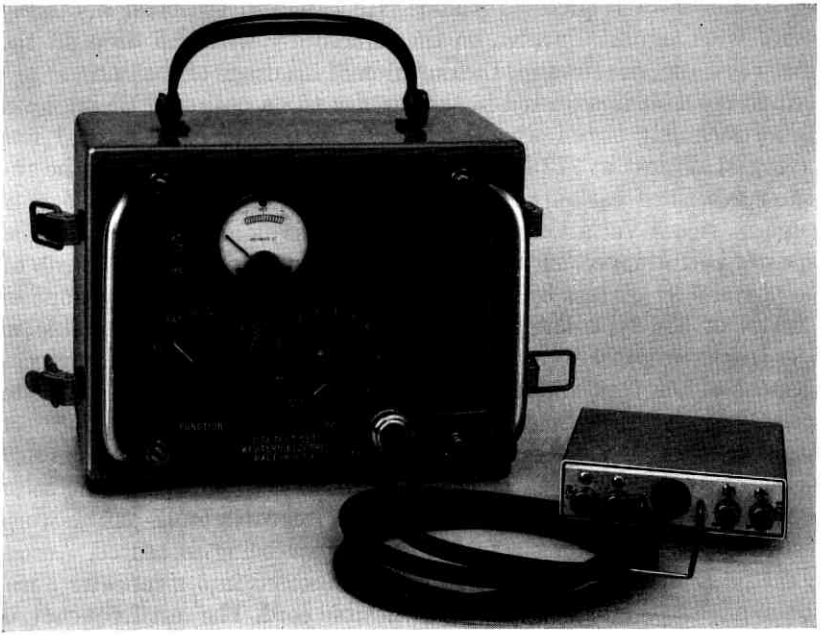

Fig. 15 — 113A test set.

tor jack provided for each span or bank output. A small indicator lamp on the test set panel flashes for each bipolar violation occurring in the transmitted signal. When a panel switch is operated, the lamp lights continuously to indicate the presence of pulses. Since a single error always produces a bipolar violation and since errors rarely occur in pairs or longer sequences, each bipolar violation always represents at least one error, and usually only one.

The second type of test can be made only when the line is not in service. It consists of introducing a test signal containing deliberate bipolar violations of adjustable violation density whose violation polarity is reversed at an adjustable audio frequency rate. This signal is produced by the fault-locating set. The output of each repeater driven by such a signal contains an audio frequency component whose frequency is the reversal rate and whose amplitude is roughly proportional to violation density, within the range of repeater capability. This audio frequency output is used to identify defective repeaters in the span by a test at a span terminating office. The fault-locating filter in each apparatus case along the line is a narrow-band selective

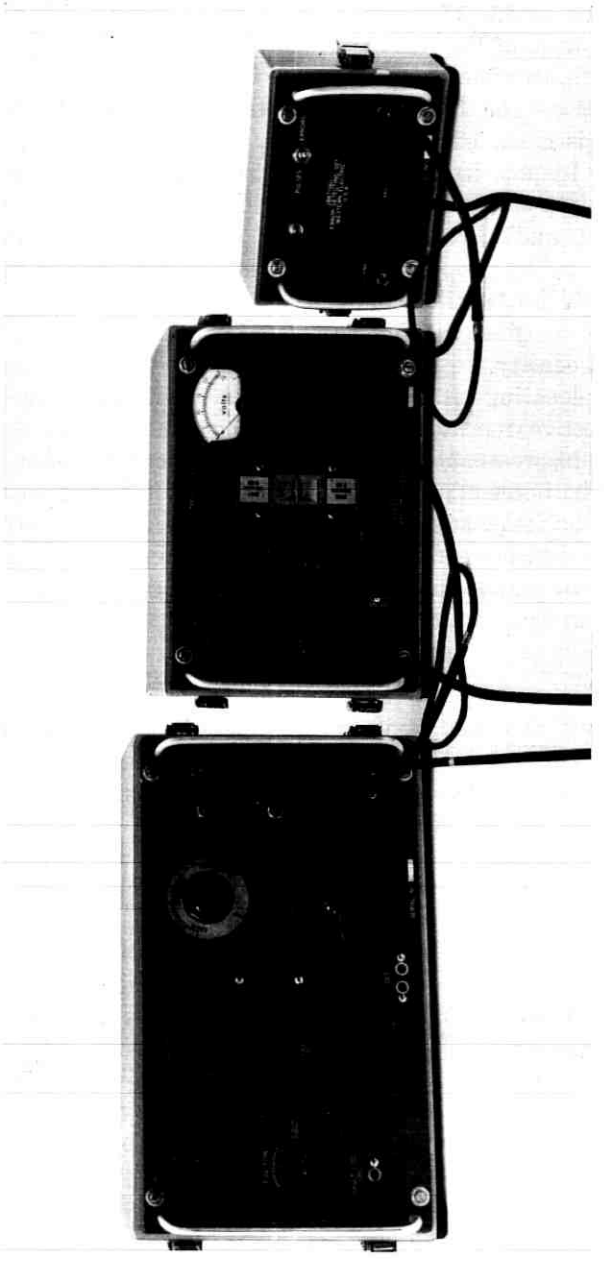

Fig. 16 — Repeater test set, fault-locating set, and error-detecting set.

filter centered at one of a series of 12 audio frequencies within the adjustment range of the test signal generator. The filter input is multipled through isolation networks to the outputs of all repeaters in the apparatus case. The filter output is bridged on the fault-locating pair, which is a loaded voice frequency line, together with filters of other center frequencies at other repeater points, and carried back to the transmitting office to be measured with a noise meter. In a normal working line, audio frequency outputs can be observed from each repeater in turn by properly adjusting the reversal rate of the signal. A repeater which has failed completely will not return an audio signal. One which is marginal will return a signal which is not proportional to the violation density.

This fault-locating test is useful not only in determining the location of defective repeaters but also in confirming that a properly working line has reasonable margins. The fault-locating set is arranged to provide the necessary pulse patterns for the test and to filter the returning audio frequency signal, thereby improving the effective fault-locating signal-to-noise ratio of the fault-locating pair. The set can also provide bipolar signals of adjustable pulse density for use in the preinstallation testing of repeaters in the repeater test set.

5.2 *Terminal*

The variety and complexity of the plug-in units composing the D1 channel bank render its alignment and maintenance a problem of considerable magnitude. Detailed analysis of defects in such circuits requires high-speed oscilloscopes with sophisticated time-scale features and highly trained specialists to operate them. Since application of such methods on a large scale in plant operation seems impractical at the present time, other approaches must be used.

The greatest maintenance simplification, of course, is that the nature of a defect within a plug-in unit need not be identified. A defective plug-in unit is normally replaced by a spare and returned to a repair center where equipment is available to analyze defects in detail.

As another aid to alignment and maintenance, several specialized test units, shown in Fig. 17, have been designed to plug into one of three access points in the circuit, normally filled by through-connectors. These access points, marked by crosses in Fig. 3, are in (*i*) the PAM 3 lead, which carries pulse amplitude modulated signals from the decoder to the expandor, (*ii*) the PAM 4 lead, which connects the expandor output to the receiving gates, and (*iii*) the transmitting and receiving line leads at the alarm control circuit.

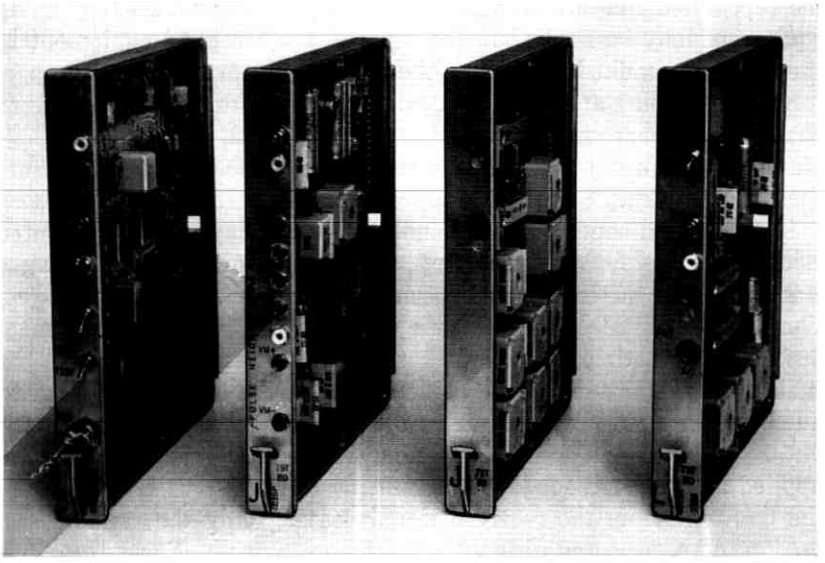

Fig. 17 — Test units used for alignment and maintenance testing of D1 banks.

One of these test units contains a code generator and a code detector. The code generator produces a digital code, timed by the master clock in the transmitting circuit, which represents a 2000-cycle sine wave signal of maximum amplitude in one channel. This signal is impressed on the receiving D1 bank, passed through the receiving converter, decoder, and expandor, and is finally converted into its equivalent voice frequency tone in another special test unit called a PAM detector. The amplitude of this tone is measured by a standard voice frequency transmission measuring set and is adjusted to its proper value by the expandor gain control, the only alignment required in the receiving group equipment.

The code detector portion of the code generator and detector is a device which converts a train of pulses to a direct current ouput, except that it does not respond to the code for zero signal, which is a single pulse at D2 time. Thus, the transmitting D1 bank is adjusted for zero signal by setting the encoder biases for a null reading of the code detector direct current output. The gain of the transmitting D1 bank is adjusted by using the previously adjusted receiving bank as a digital detector to indicate that the proper digital code is being transmitted when a standard voice frequency input is impressed.

Since the normal signals in the two directions of transmission on

the repeatered line are similar, the output of the transmitting group equipment may be applied to the input of the receiving equipment in the same D1 bank. Insertion of the code generator and detector performs this looping and also phases the receiving timing circuits so that channel 1 is connected to channel 7, 2 to 8, 6 to 12, 13 to 19, etc. Regular voice frequency transmission measuring sets are used to measure the voice frequency gain of the looped circuit.

In the looped condition, compandor tracking may also be measured using the transmission measuring set, and idle circuit noise using a standard noise meter. A third special test unit, a one-kilocycle rejection filter, facilitates measurement of distortion due to misalignment and to the quantizing noise which results from reconstructing the actual signal amplitude only to the accuracy of the nearest of a limited number of discrete levels. Patched between the receiving channel output and a noise meter while a harmonic-free 1000-cycle test tone is applied at the corresponding transmitting channel, this filter removes the test tone from the received signal, leaving only the residual distortion to be measured directly with the noise meter. These three test units, together with standard transmission testing equipment, are adequate for normal D1 bank alignment and evaluation procedures.

A fourth test unit is provided, however, to aid in identifying a defective plug-in unit. This is the trouble location network, which has three functions. The first is to indicate pulse rate, and the second is to indicate pulse height. The test unit converts these parameters to dc signals measurable by a voltmeter. The third function is to loop the analog output of the compressor to the input of the expandor, eliminating from the loop the encoder, decoder, and converters.

Another necessary maintenance adjunct is a matching network unit for interconnecting the 600-ohm balanced circuits of transmission measuring equipment to the 2500-ohm unbalanced circuits available at the channel unit access jacks. It mounts separate transmitting and receiving circuits which may be used simultaneously for loop measurements. It contains only passive elements such as jacks, matching transformers, resistance pads, and level adjusting keys, and is permanently mounted at a convenient working level at the side of one of each group of four D1 bank bays.

5.3 *System*

An alarm control unit, one of the normal complement of plug-in units in each D1 bank, is specifically designed both to indicate trouble conditions and to aid in identifying the part of the system affected.

The trouble condition indicator is a circuit which continuously monitors the incoming framing signal, the only unique signal always transmitted in a normally working system. The amplified framing signal normally holds an alarm relay operated. If the signal fails, the relay releases and operates audible and visual alarms in the receiving office. At the same time, it forces the transmission of a special signal in the opposite direction. The special signal consists of the normal bit stream with all pulses in digit one and digit eight positions inhibited. The forced absence of these pulses is detected in a second alarm control unit circuit at the far-end D1 bank. This circuit also operates audible and visual office alarms. Thus, if transmission fails in one direction, alarms are operated almost simultaneously at both ends of the system. Different colored lamps on the alarm control unit indicate whether the alarm is due to the primary framing signal failure or the forced failure of digits one and eight.

Sectionalization of the failure is accomplished by manually operating a looping key on the alarm control unit at each end. If both looped terminals indicate normal framing signals, the system failure must have been in the repeatered line. If either looped terminal is out of frame, that terminal must be defective. Since the direction of transmission of the system failure is known from the original alarm indication, it is also known whether the transmitting or receiving portion of a defective terminal is at fault.

If a system failure is due to a defective line, a spare line may be quickly patched in to replace the defective section. Thus, service over the system may be restored at once, thereby reducing the urgency of the time-consuming fault-locating procedures and replacement of defective repeaters.

If the defect is in a terminal, established trouble-location procedures are followed until the defective plug-in unit is identified and replaced. Most such troubles are quickly found by the straightforward application of simple rules. As an aid in identifying the occasional obscure case of trouble, a list of known trouble symptoms is provided, and under each heading a list of the types of units which have been found in manufacturing testing to produce each symptom, arranged in order of frequency of occurrence.

5.4 *Trunk Processing*

Since T1 carrier, like most other carrier systems, sends a continuous signal in the signaling path of each channel when it is idle, a cessation of the signal constitutes a demand for service. If a system fails

while many of its 24 channels are idle, the simultaneous cessation of signal in those channels imposes an abnormal load on the switching equipment, which may temporarily impair its responses to normal service demands immediately thereafter. In the case of busy channels, a system failure may affect only one direction of transmission and may not send the proper supervisory signals to stop charges on toll calls, even though the circuits are useless. The system may also continue to accept calls which cannot be completed because of the failure.

To minimize the impact of a system failure on the switching equipment and on customer service, a carrier group alarm circuit, as shown in Fig. 18, is normally provided as an optional accessory. Such a circuit at each end of a system makes or breaks relay contacts as required in each channel during a system failure to produce the least service disruption. Enough contact arrangements are provided, by screw-down options, to process on an individual channel basis any of the types of trunk circuits which may be transmitted over T1 carrier. The carrier group alarm circuit is actuated by the alarm control circuit of the D1 bank with which it is associated, and because of the design of the T1 carrier office alarm system already described, operates almost simultaneously at both ends of the system for a failure in either direction of transmission. Both the system alarm and the carrier group alarm are self restoring so that when a system failure is cleared, the trunks are automatically put back in service. To guard against rapidly recurring service interruptions due to possible intermittent system failures, the carrier group alarm restoral is normally delayed by approximately ten seconds after the system is performing satisfactorily. Since

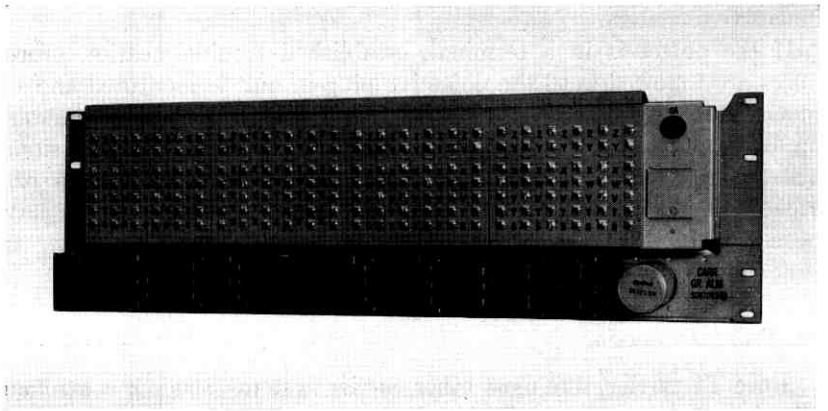

Fig. 18 — Carrier group alarm.

the amount of delay is not precisely controlled, an additional interlock feature, using the signaling path of one of the channels, ensures that the restoral will be simultaneous at both ends of the system.

Twelve carrier group alarm units can be mounted on an 11-foot 6-inch bay, 23 inches wide. The bay is usually placed in the middle of a group of four D1 bank bays, and hand-wired by the installer between the 12 D1 banks and the trunk appearances for these banks on the office distributing frame.

VI. MANUFACTURING TESTING

6.1 *D1 Bank Units*

Of the 15 types of plug-in units which compose the group equipment of a D1 bank, 9 are primarily digital circuits and pose few obscure or unusual problems of inspection testing or adjustment in the factory. Requirements for pulse amplitude, pulse width, and phase are not usually critical, and normal inspection procedures are found adequate to ensure correctness of unit assembly. This is largely true also of the transmitting and receiving gate circuits, which include analog transmission paths as well as digital circuits. They do not require unusually high precision except for the maintaining of high balance in the transmitting gates, which is sometimes a problem in device manufacture. The four remaining units, which do require a high degree of care in assembly and precision in testing, are those involved in analog-to-digital conversion. They are the compressor, encoder, decoder, and expandor circuits.

The compressor and expandor circuits include several amplifiers which must maintain high gain over a wide frequency band with a high degree of stability in spite of aging or variations in temperature and battery voltage. Success in meeting these requirements, however, depends less on advanced manufacturing and testing techniques than on adequate design of feedback circuits and proper choice of stable components. A few routine measurements of gain with and without feedback are sufficient to ensure high performance.

The point where unusual care and precision are required is the measurement and adjustment of the nonlinear network elements. The compressor and expandor nonlinear networks are designed to have identical current-voltage characteristics when measured in the circuits in which they operate. The required inverse characteristics are obtained by their method of connection into the circuits, as illustrated in Fig. 19. For the compressor, a current proportional to the signal is injected

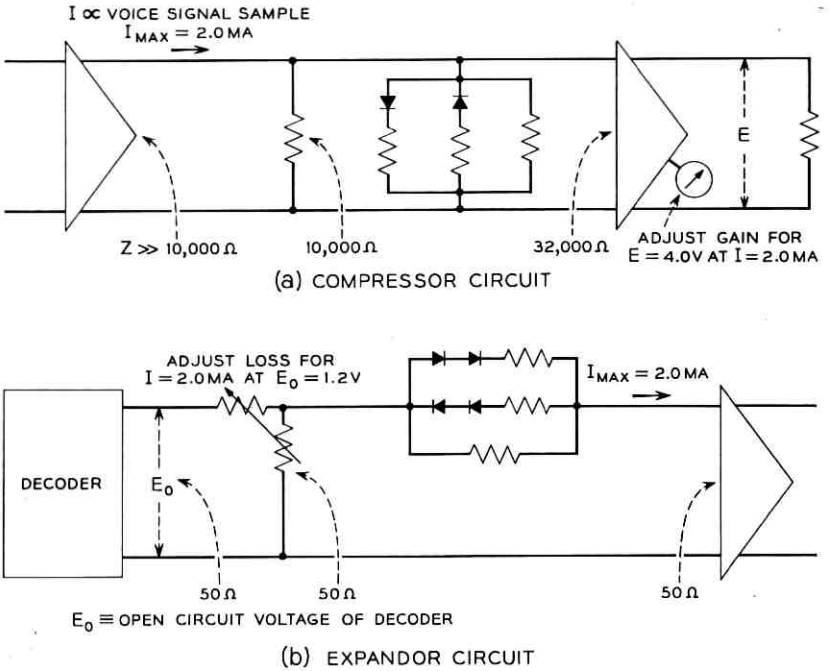

Fig. 19 — Basic operating circuits of nonlinear networks in compressor and expandor.

into the network, and the corresponding voltage is read out and transmitted through the system to the expandor. For the expandor, the voltage is impressed on the network, and the corresponding current is read out as the received signal.

A practical objective for the accuracy of duplication of nonlinear networks is that at any network current within the usable range, the voltage difference between any two networks should not exceed one half code step, one part in 126 of maximum positive or negative signal. This ensures that the maximum error in any signal sample due to mistracking is no greater than the maximum error due to quantizing. Implicit in this statement of the objective is the consideration that absolute magnitudes of nonlinear network voltages are not critical as long as the normalized characteristic shape meets the objective. Normalization is easily accomplished in the circuit by properly adjusting associated circuit gains or losses to match particular nonlinear networks.

In the expandor, a two-element pad, adjustable in small steps, builds out the precisely measured dc voltage at the maximum signal current of 2 milliamperes, to the normal maximum voltage, 1.2 volts, delivered by the decoder. The pad is assembled as part of the expandor network. In the compressor, the gain of an amplifier following the nonlinear network is adjusted until the peak output signal voltage is 4 volts, its normal maximum, for an impressed sine wave input signal of 2 milliamperes peak value.

Assurance of the desired characteristic shape is attained in two steps by precision dc measurements of the voltage of each diode at three values of current corresponding to 0 db, −20 db, and −40 db with respect to maximum signal current of 2 milliamperes.* In the first step, at the device factory, diodes at the normal operating temperature of 120 degrees C are paired within close voltage limits at the middle current for use in the opposite polarity legs of the same network. The absolute voltages also must fall within broad limits, as must the voltage ratios for upper and middle currents and for middle and lower currents. In the second step, at the network assembly factory, the diodes are installed in their operating environment, again a nominal temperature of 120 degrees C. Building-out series and shunt resistors are selected by test and permanently installed so that voltage ratios of the over-all network for the same pairs of test currents are within close limits about the nominal values which produce the desired shape of characteristic.

Table I gives in abbreviated form the translation between input signal and digital code on the line. For a well aligned system, the quiescent input voltage to the coder is about half-way between the decision voltages for codes 63 and 65. A sinusoid as large as about −67 dbm must be applied at 0 db System Level before through transmission will occur. Table I gives the largest sinusoid that may normally be applied to a channel without exciting code levels farther from the origin than the range indicated.

Close control of the diode operating temperature is maintained by careful mechanical arrangement of the diodes, a sensitive, vacuum-sealed, bimetal thermostat, and a heating coil within a reflectorized, vacuum-insulated, cylindrical glass container. An assembled network and its major components are shown in Fig. 20. The mechanical construction of the core is designed to provide thermal low-pass filtering so that the wide temperature excursions of the heating element result in thermostat temperature variations of about 1 degree and diode tem-

* These procedures are discussed in greater detail in Ref. 5, pp. 187–189.

TABLE I

Input Signal Power at 0 db System Level Point	Range of Quantizing Steps Transcended by Peak-to-Peak Excursion of Input Sinusoid
−67.0 dbm	64 min — 64 max
−58.5	63 min — 65 max
−52.9	62 min — 66 max
−46.3	60 min — 68 max
−41.3	57 min — 71 max
−37.0	53 min — 75 max (check point)
−33.5	49 min — 79 max
−30.1	45 min — 83 max
−26.5	41 min — 87 max
−21.7	36 min — 92 max
−17.0	31 min — 97 max (check point)
−12.5	26 min — 102 max
−8.4	21 min — 107 max
−4.6	16 min — 112 max
−1.6	11 min — 117 max
+0.9	6 min — 122 max
+3.0	1 min — 127 max (check point)

perature variations well within the objective of ±0.1 degree C. The design allows for a small overshoot in the thermostat temperature in both directions of temperature change so that just before contact closure or just after contact opening, the relative contact velocity is a maximum. Noble metal contact material is used to prevent contact sticking. Transistor amplifier control of the heating element reduces the contact current and voltage to about 5 milliamperes and 5 volts, respectively.

Many of these nonlinear networks, six per system, have been operating successfully for more than three years, and increasing numbers for shorter periods, with few observed changes in characteristics. The thermostats, cycling at an average rate of about once per minute, give almost no evidence of deterioration.

The encoder and decoder also are high-precision circuits and include logic switching elements which must be unerring in their operation. It is possible to test the separate elements of these circuits by conventional means and to make over-all measurements of the 127 individual code levels using precision equipment. This procedure, however, is laborious and time-consuming. For quantity production, specialized test equipment has been developed which greatly simplifies over-all evaluation and also aids in identifying logic circuit errors where they occur.

The encoder test set consists of a linear generator which provides a uniformly varying encoder input, a word generator which produces in

sequence the binary 7-digit codes expected at the encoder output, and a comparator which counts the disagreements between the actual encoder output and the word generator. The ramp generator is adjusted to cover the encoding range from maximum negative to maximum positive signal at the rate of one code step per eight words at the normal encoding rate. The word generator repeats each word eight times before proceeding to the next code. Thus the total number of errors divided by 127, the number of codes, gives the average encoding error in eighths of a code step. An oscilloscope display of the ramp and corresponding code patterns affords an effective means of localizing defective logic elements.

The decoder test set is the inverse of the encoder set. A word generator, of the type used in the encoder set, sends a sequence of digital codes into the decoder. The output, ideally a linear ramp, is displayed on an oscilloscope and is evaluated visually. Here again, the nature of the display aids in identifying logic errors.

6.2 *Over-all Group Test*

When the 29 plug-in units which compose the group equipment of a D1 bank have been manufactured and tested individually, a final test

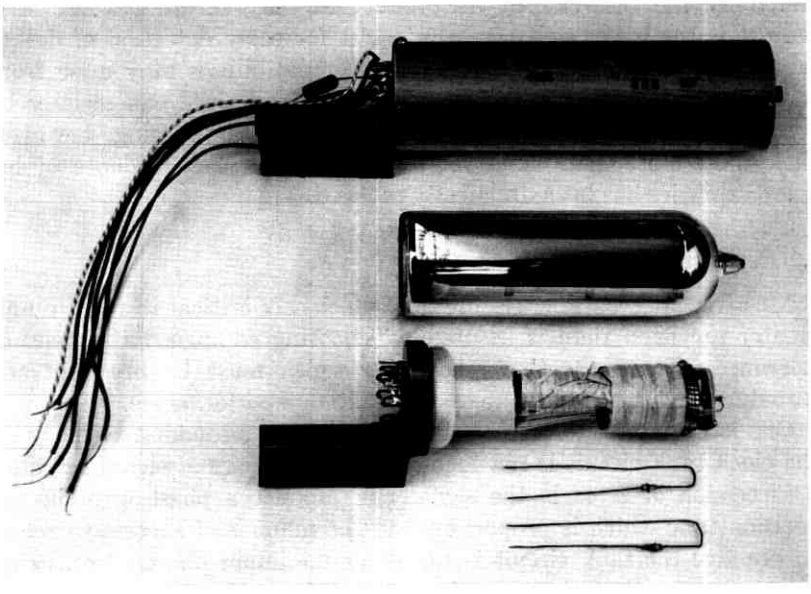

Fig. 20 — Nonlinear network for compressor.

is made of the group assembly. A standard, fully wired D1 bank framework is used as a test position, and tests of the group parallel those which will be made by the operating company in the final installed position.

This type of test has several valuable features. The first, and perhaps most important, is purely statistical. Experience shows that individual plug-in units of this type which have been through normal manufacturing, inspection, and testing routines, have about a 1.6 per cent probability of being defective in a second test. An assembly of 29 such units, therefore, would have almost a 50 per cent probability of a defect. A D1 bank which has successfully met its requirements as an assembly, however, has only a 4 per cent probability of a defect on a second test. A second feature of the over-all test is that it permits individual unit tests to be abridged with impunity. For example, many units have multiple outputs from one transformer. Considerable effort can be saved in testing individual units by measuring only one of these outputs, with the assurance that defects in the other output circuits will be recognized in the over-all test. Similarly, human errors in identifying units for the assembled bank are recognized when the assembly is tested. Experience has also shown that misadjustment of an individual unit test set may sometimes be detected by the behavior of its product in the over-all test. A final feature is the opportunity in such a test for early detection of design incompatibilities among units. Such incompatibilities may arise from any of a wide variety of causes, but once recognized can usually be corrected easily. The value of this feature is diminishing, however, since after more than three years of production the incidence of incompatibilities has been drastically reduced.

6.3 *Repeater Adjustments*

Most repeater parameters are not sufficiently critical to require any greater precision than is assured by choosing components of normal tolerances. There are three, however, which must be precisely adjusted for each regenerator to assure maximum performance.

One is the tuning of the clock circuit. If the resonant frequency of the clock tank circuit is not identical with the normal signal bit rate, a succession of zeros in the signal will produce a phase error in the decision time which is proportional to the number of successive zeros. In practice the tank circuit is tuned for minimum phase jitter, using a test signal adjusted as closely as possible to the nominal bit rate. The pulse pattern of the test signal is alternately sparse and dense,

and minimum jitter is marked by the sharpest zero crossing of the superimposed pulse patterns as observed with an oscilloscope at the clock circuit output. Deviations of the clock tuning due to environmental or other influences may be measured by varying the test signal bit rate for comparison with the nominal.

A second critical parameter is the adjustment of the automatic threshold bias circuit. This circuit normally holds the threshold decision amplitude at the center of the eye over a range of about eight db of input pulse amplitude variation. The adjustment is made by sending a normal minimum-amplitude signal of random pulse pattern, adding an interfering signal of such amplitude that one error per second is produced, and varying the bias until the interfering signal, holding the error rate constant, is a maximum. Experience shows that the precise shaping of the input signal pulse is very important. No network of lumped constants has been found which shapes a regenerated pulse properly. For shop testing, therefore, a special cable in a hermetically sealed container is used to simulate the characteristics of the most frequently encountered exchange cables.

The third adjusted parameter is the energy of the regenerated output pulse. The amplitude of the received pulse depends primarily on the energy of the output pulse and the loss of the line. In order to leave as much margin as possible for environmental variations of line loss and for inaccuracies due to the 2.4-db step size for line build-out networks, it is desirable to keep the output energy as uniform as possible. This energy is approximately proportional to the product of pulse height and pulse width. The pulse height is proportional to the power supply voltage, which may vary ± 10 per cent from nominal because of manufacturing variations of the regulator diode which determines it. The pulse width depends to some extent on the switching time of the blocking oscillator transistors, but is adjustable in the clock spike generator circuit. This circuit turns on each pulse to be regenerated with a positive spike and turns it off with a negative spike after an interval determined by the circuit parameters. The interval is adjusted until the amplitude of the output pulse, attenuated by a standard artificial line, has its nominal value. The duration of the positive and negative pulses is matched by pairing transistors at the device factory for matched switching time.

6.4 *Environmental Test*

A major consideration in repeater manufacture is the need for maintaining exceptionally high quality of product. In large population

centers, repeaters are mounted in hermetically sealed cases in manholes, often in busy streets of high traffic density. Since it is expensive and time-consuming to open a manhole to replace a repeater, every effort is made to ensure that each unit is in good operating condition when it is put in place and that its failure rate in service is as low as possible.

Each dual repeater contains 145 electrical components and approximately 500 soldered connections. If a quality objective is set to allow no more than one defective repeater in 100, this translates into one defective soldered connection in 50,000 or one defective component in 14,500. Shop experience in the manufacture of similar types of equipment indicates that higher failure rates than these are to be expected with normal manufacturing and inspection methods.

Early in the preparatory stages of manufacture, therefore, it was decided that two extraordinary precautions would be taken to improve as far as possible the probability that a repeater when installed would operate properly. One was the provision already described, for making a pre-installation test of each repeater just before inserting it in a system. The other was the provision for temperature cycling of the adjusted, inspected product followed by a final inspection test. Records are kept of the test results and defect analysis, and continuous quality control is exerted to impound any lots which appear to be substandard. The temperature cycling continues for 20 hours and includes five complete cycles from room temperature down to −40 degrees, up to +150 degrees F, and back to room temperature, holding for at least one hour at each extreme temperature in each cycle. At the high-temperature point of the last cycle, the repeaters are removed from the chamber a few at a time and tested while hot to ensure operation at the design maximum temperature of +140 degrees F. After they have stabilized at room temperature, detailed final tests are made to ensure that changes due to temperature cycling have not been appreciable. About five per cent of the product is cycled in a special chamber equipped to monitor operation while the units are held at the extreme temperatures, in this case −40 and +140 degrees F.

VII. CONCLUSION

Many T1 carrier systems have been in successful operation for more than three years. Many more are being installed and put into service in metropolitan areas all over the United States. In general, acceptance has been good. Although new concepts and sophisticated

circuitry are involved, the operating company craftsmen have been able to align and maintain these systems with evident success. Equipment prices were slightly lower from the beginning than the preliminary estimates, and have since been reduced further. The demand for systems has been greater than anticipated so that the manufacturing program has been increased each year over earlier long-range forecasts.

As T1 carrier usage in the plant increases, it is to be expected that needs for connection to other types of trunks will develop. As such markets materialize, the design of additional types of channel units will be undertaken.

Looking to the future, it seems certain that the next few years will see increasing use of the T1 carrier line as a transmission facility for high-speed digital data. Its inherent capability for pulse transmission at the 1.5 megabits per second rate coupled with its relatively low cost make it particularly attractive. This subject is discussed in a companion paper. The expected growth of the T1 carrier trunk network in coming years will greatly facilitate these data transmission applications.

VIII. ACKNOWLEDGMENTS

Contributions to the development of the T1 carrier system have been made by the authors of papers listed as references, their associates, members of Bell Telephone Laboratories development and engineering groups, and members of Western Electric Company and American Telephone and Telegraph Company engineering groups. Particularly valuable have been the earlier work of O. L. Williams, A. C. Longton, and N. E. Lentz, and the continuing efforts of F. H. King, E. J. Anderson, and C. D. Donohoe.

REFERENCES

1. Davis, C. G., An Experimental Pulse Code Modulation System for Short-Haul Trunks, B.S.T.J., *41*, January, 1962, p. 1.
2. Mayo, J. S., A Bipolar Repeater for Pulse Code Signals, B.S.T.J., *41*, January, 1962, p. 25.
3. Aaron, M. R., PCM Transmission in the Exchange Plant, B.S.T.J., *41*, January, 1962, p. 99.
4. Shennum, R. H., and Gray, J. R., Performance Limitations of a Practical PCM Terminal, B.S.T.J., *41*, January, 1962, p. 143.
5. Mann, H., Straube, H. M., and Villars, C. P., A Companded Coder System for an Experimental PCM Terminal, B.S.T.J., *41*, January, 1962, p. 173.
6. Travis, L. F., and Yaeger, R. E., Wideband Data on T1 Carrier, B.S.T.J., to be published.
7. Cravis, H., and Crater, T. V., Engineering of T1 Carrier System Repeatered Lines, B.S.T.J., *42*, March, 1963, p. 431.

A Statistical Basis for Objective Measurement of Speech Levels

By PAUL T. BRADY

(Manuscript received March 15, 1965)

First-order probability distributions of speech amplitudes are studied to establish a theoretical basis for obtaining a measure of speech level. The logarithm of the long-term waveform of the speech envelope is found to be approximately uniformly distributed above a threshold. The average peak level (apl) is obtained by taking the time average of the log of the envelope waveform and deriving from it the peak of the log-uniform distribution which would have produced the same average. A theoretical analysis of various properties of the apl indicates that, within certain bounds, the apl satisfies a postulated set of requirements of an "ideal" speech level measure. A critical requirement is that the measure remain independent of the value of a threshold employed by a speech detector in the measuring device. It appears that variation in the threshold can typically change the apl by about one db.

The Digital Speech Level Meter is described as an instrumentation of the technique used to obtain the apl. Measurements made with this meter are easily obtained and very repeatable, and are in general agreement with theoretical predictions.

I. INTRODUCTION

1.1 *Object of Study*

The goal of this study is determining a speech level measure ideally having the following properties:

(1.) It is objective, and is not based on the judgment of an observer.

(2.) It is based on measurements made only while speech is present and is not influenced by long silent intervals.

(3.) It is expressible as a single number.

(4.) It varies on a db-for-db basis with attenuation or amplification of the voice signal.

(5.) It is not a function of an arbitrary convention used to take the measurement, such as the value of a threshold in a meter.

(6.) It is not influenced by singular loud transients on the voice circuit.
(7.) It is easily and reliably obtained.

The specification of the *level* of a signal implies a description of certain physical properties of the amplitude of the waveform. The *loudness* of a signal is a measure of the volume of a sound as perceived by a listener. Although it may be possible to correlate level measurements with loudness, no attempt will be made to do so in this study.

1.2 *Outline of Report*

In seeking a measurement satisfying the above requirements, an analysis is made of the statistics of speech levels as they appear above a threshold. This analysis, appearing in Section II, shows the logarithms of these levels are nearly uniformly distributed.

Section III indicates that the peak amplitude occurring in a speech sample satisfies most of the requirements listed above. It may, however, be due to some isolated event (such as coughing or a circuit transient on the voice circuit) which is not characteristic of the general speech process.*

A different measure, the *average peak level* (apl), is therefore proposed in lieu of the sample peak. The apl is a parameter of the postulated uniform level distribution of the speech sample. It is shown that if speech actually is "log-uniformly" distributed, as seems to be the case for some speech samples, the apl is equivalent to the peak. For other speech samples, it will still satisfy some of the requirements stipulated above, and will approximately satisfy the others. Since it is a measure taken over the entire sample, the apl has an advantage over the peak in that it is relatively uninfluenced by singular loud events.

Section IV shows the apl to be a better objective measure of speech levels than the volume unit (VU) presently measured, since the latter exhibits significant observer bias and variability.

Section V describes the Digital Speech Level Meter, an instrument which demonstrates a technique used to obtain the apl. Some of the measurements made with the meter are included in Section VI. These measurements are easily obtained and highly repeatable. It is emphasized that the instrument described here is only an experimental model and may be subject to many revisions before it is suitable for general use.

* The second highest peak, the average of the first and second highest peaks, the third highest peak, etc., also satisfy most of the requirements, but are also influenced by extraneous events. In addition, as more peaks are involved in the measurement, the mathematics for a theoretical analysis becomes intractable. Complex functions of several peaks will therefore not be considered in this paper.

II. DISTRIBUTION OF SPEECH LEVELS

2.1 *Density and Cumulative Functions*

In this report, upper case X will be used to denote a random variable and x will denote a particular value which X can assume. Only continuous functions will be studied. The *density function* will be denoted $p(x)$, and the *cumulative function*, $P(x)$. The cumulative function will be Prob. $(X \geqq x)$, the complement of the definition normally used in mathematics, but which is commonly used in speech literature.[1,2,3,4]

2.2 *Establishing a Threshold*

In order to measure speech levels only during the time when speech is "actually present," we must establish an objective indicator of intervals over which the speech waveform is to be observed. Ideally, this indicator should mark off intervals *which would retain their pattern regardless of the level at which the speech sample is played.* If such a pattern could be established then some simple statistic, such as the rms voltage (V_{rms}) measured and averaged only over the prescribed intervals, could satisfy all of the requirements in Section 1.1.

Because of the wide dynamic range of speech, it is virtually impossible to establish the required level-invariant speech patterns if noise is present. A previous study[5] dealt with this problem in some detail, however, and it was shown that on a special simulated toll circuit, a threshold of -40 dbm re OTL* is sufficiently sensitive for detecting most of the speech while avoiding noise operation. Such a threshold detection incorporates no hangover and therefore differs from a conventional speech detector. Expressed mathematically, let X be a random variable such that

$$x = 10 \log \frac{(1000)\,(v^2)}{600} \tag{1}$$

where v is the voltage representing the speech waveform and x is the equivalent level in dbm. Then the speech considered for analysis in this study will be such that

$$\text{Prob}(X \geqq -40 \text{ dbm}) = 1 \text{ (or 100 per cent)}. \tag{2}$$

* Zero dbm equals 0.775-volts rms across a 600-ohm resistor and will thereby cause one milliwatt to be dissipated. Although dbm implies a power measurement, it is often used to specify a voltage without regard to power or resistance, as is done here. Zero dbm is about 2.22 db below one volt (zero dbv).

The zero transmission level (OTL) point is a point to which all level points in a telephone toll system can be referred. It is analogous to citing altitude by referring to height above sea level.

Having thus adopted an arbitrary threshold criterion, the task of this study will be to specify the level of a speech sample with a measure which will be relatively insensitive to the threshold value.

2.3 *Source Material*

All of the measurements in this study were made with 8 recorded conversations involving 4 pairs of men and 4 pairs of women. Each conversation was about 7 minutes long except for one which lasted only 3.3 minutes. The recordings were made at the OTL point of a simulated toll circuit. In addition, a "continuous speech" tape was produced by manually editing out conversational pauses, thus condensing each person's speech from 7 minutes to about 1 minute. (A more detailed description of the conversations may be found in Ref. 5.)

2.4 *Instantaneous Level Distribution*

The instantaneous level of speech is interpreted here as the absolute magnitude of the speech waveform at a particular instant of time, expressed in dbm. Shown in Fig. 1 is the computer-obtained cumulative function for the levels occurring in a 67 second sample of continuous speech from subject AD. The speech was 4-kc low-pass filtered and was

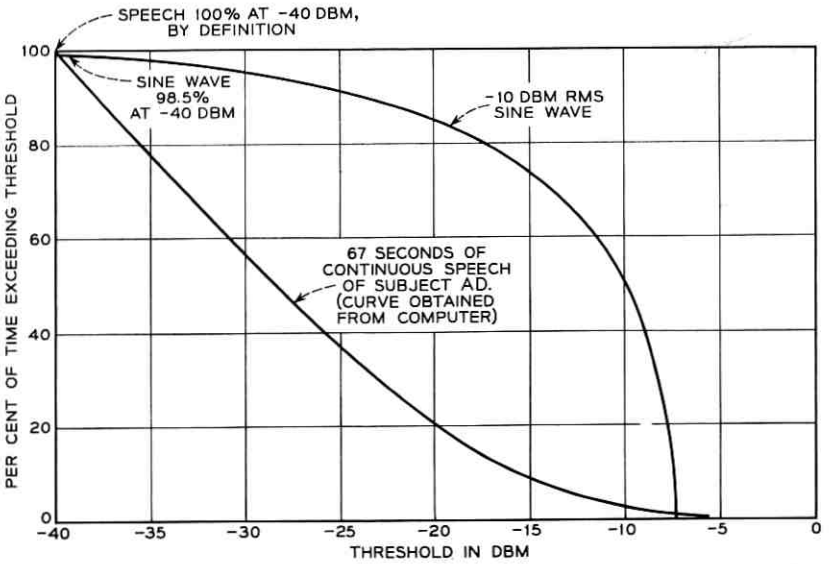

Fig. 1 — Instantaneous cumulative functions of speech and of a sine wave.

sampled at 10 kc for analog to digital conversion. Continuous speech was used because it was more economical for computer work; the original speech contained too many pauses.

The speech curve of Fig. 1 starts at 100 per cent for −40 dbm, by convention stated in (2), and then decreases almost linearly (for a major part of its range) toward a cutoff point near −10 dbm. The approximate linearity of the speech curve is of crucial importance in this study, since the level measuring technique to be described later depends on this property.

To illustrate the contrast between the speech distribution and a sinusoidal distribution, the cumulative function for the instantaneous levels of a full-wave rectified −10 dbm rms sine wave is also included in Fig. 1.

A few of the speech level distributions appearing in the literature are plotted in Fig. 2. The conversion of the original thresholds to the dbm scale is accomplished simply by transferring the shape of the literature data onto the author's graph, ignoring the absolute values of the literature thresholds. This conversion is valid since only the shapes, and not the absolute values, of the different curves will be compared. The curves of Fig. 2 are taken from Sivian,[1] Dunn and White,[2] Davenport,*[3] and Shearme and Richards.[4]

2.5 *The Log-Uniform Distribution as an Empirical Formula*

Figs. 1 and 2 indicate that all of the speech data are very similar, and that the cumulative functions can be approximately drawn as straight lines over much of their range. If the cumulative function were truly linear, then the density function would be uniform over its whole range with value 1/[peak − (−40)], and would be zero outside of this range. This distribution will be called the *log-uniform distribution* since the logarithm of the amplitude is uniformly distributed. In Section III certain properties of the log uniform distribution will be investigated. It is worthwhile first, however, to examine the distribution of the speech wave envelope.

2.6 *The Envelope Distribution*

Let speech be played into a full-wave rectifier, whose output in turn is applied to an RC filter having approximately equal rise and decay times. (Such a circuit is shown in Fig. 11.) The waveform at the filter output will be considered as the speech envelope. It was chosen for use

* Davenport presents *density* functions of measured speech levels, and establishes an empirical formula for the distribution. Plotted in Fig. 2 of the present report is the cumulative function of the empirical formula.

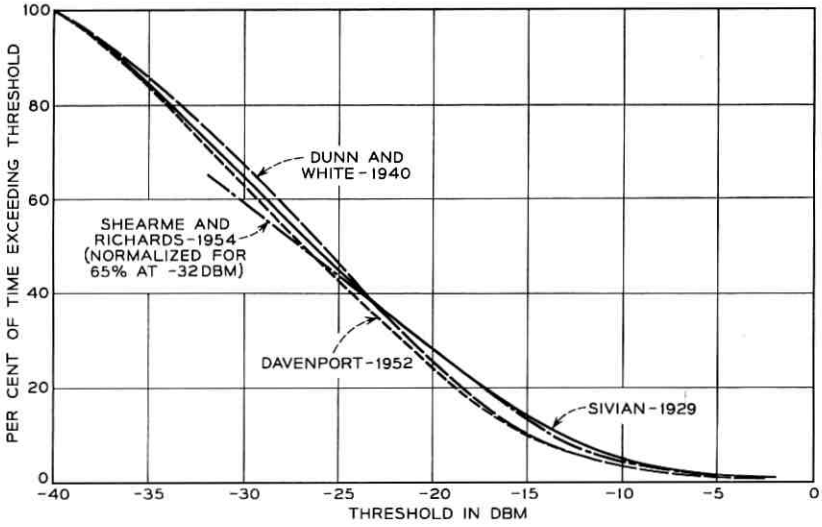

Fig. 2 — Instantaneous speech distributions from four earlier studies. The absolute levels of each curve were adjusted to the DBM scale so that the levels would be roughly equivalent to those in the present study.

in this study since it varies at a lower rate than does the original waveform and thus leads to simpler instrumentation in sampling circuits. It is shown in the next section that the envelope level distribution is similar in shape to that of the instantaneous waveform, although they differ in absolute values. Because of the similarity of distributions, the technique developed later in this paper for measuring levels would work equally well with either the envelope or the original speech waveform.

2.7 *Choice of the Time Constant*

The distribution of a speech wave envelope will depend on the choice of the time constant of the RC filter. A family of unnormalized cumulative functions for the 67 second continuous speech sample of subject AD with time constant as the parameter is shown in Fig. 3. Also included is the computer-obtained cumulative function of the instantaneous amplitudes.

With large values of RC, the speech amplitude peaks become smeared out in time, and more low level energy is evident. This is shown in Fig. 3, which is in agreement with the data of Shearme and Richards.[4] With an RC of 2.5 msec, the envelope levels are spread over almost as great a range as are the instantaneous levels. In Section III it will be shown

that the level measurement should be taken with the threshold fairly close to the lower end of the linear range of the distribution. Fig. 3 shows that large values of RC compress the distribution, narrowing the allowable threshold range. An RC of 2.5 msec is chosen to avoid this difficulty.

The VU meter (discussed more fully in Section IV) is constructed so that the needle follows the speech level at a "syllabic rate" and has a time constant of about 140 msec.* The longer time constant is necessary to allow an observer to follow the meter movement; he would be at a loss to keep track of the needle if the time constant were 2.5 msec. The

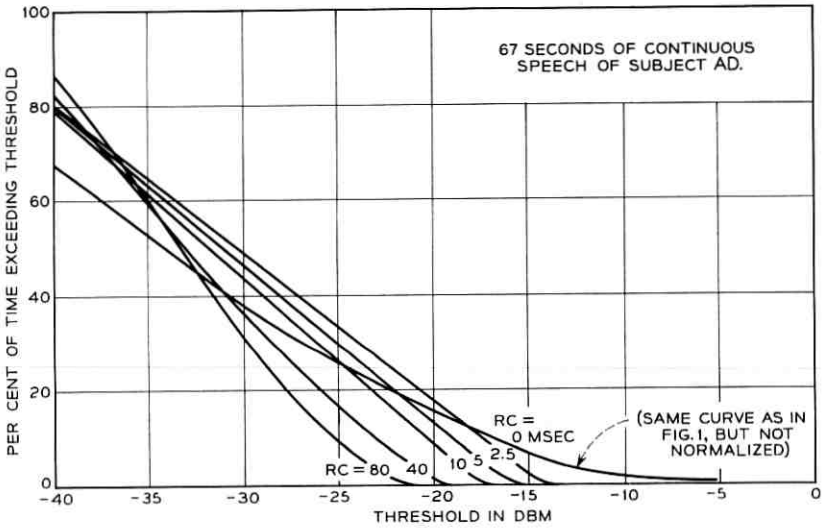

Fig. 3 — Effect of changing time constant on envelope distribution.

smaller value can be used in this study because the human observer limitation is not present.

2.8 *Results of Envelope Distribution Measurements*

The cumulative distribution of the envelope of the combined speech of all 16 talkers is shown in Fig. 4. This represents about 25 minutes of speech exceeding the −40 dbm threshold, with a total elasped "real time" of about 103 minutes. This distribution is again linear over most of the range, but a better approximation is to use two straight lines, as

* Based on measurements made of the 63 per cent rise and decay times for three different VU meters.

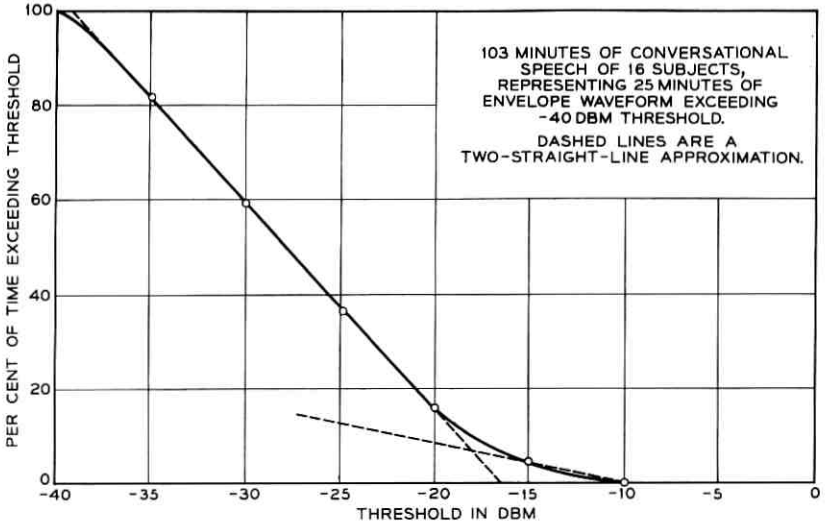

Fig. 4 — Envelope distribution of combined speech of all subjects.

shown in the figure. (The two-line approximation will be discussed in Section 3.3.)

Regarding the distributions of the individual speakers, the curves can be placed into three categories, as shown in Fig. 5. The curves of half the speakers were distributed log-uniformly; an example is the speech of NS. The speech of seven others had curves which seemed to be composed of two log-uniform functions; similar to that of BS. Most of the break points occurred very close to the bottom of the curve; the one illustrated is in fact the most pronounced case of a double valued distribution. One speaker, JM, had a noticeable downward break point near the top of the curve. This effect was also present to a very small degree in two or three of the other speakers. It will be shown in Section 3.4 that such low level break points have little effect on level measurements, and for this reason this distribution will not be considered in subsequent analysis.

2.9 *Length of Sample*

The statistical speech level measure to be proposed in this study is based on the assumption that the speech sample has an approximately log-uniform distribution. It is therefore of interest to learn: (1.) what length speech sample is required to yield a log-uniform distribution,

and (2.) what length is required before the sample is representative of the long term distribution of a particular speaker. To answer these questions, the *continuous* speech tapes of four men and four women were analyzed as follows: The distribution of a one-second (real time, not time over a threshold) sample of speech for a subject was obtained. Then, a two-second sample was analyzed such that the two-second sample included the one-second sample. This was done in like manner for 4, 8, 16 and 32 seconds. The whole process was repeated for each subject. Fig. 6 is an example of the cumulative function obtained with this technique for subject MB.

For six of the subjects, practically every cumulative function was a straight line. An exception was the one-second sample which was occasionally curved. For the two other speakers, the 8 second sample was the shortest sample which appeared linear. This represents between four and five seconds of speech exceeding the −40 dbm threshold. It appears therefore that at least four or five seconds of "over the threshold" speech is required to achieve a log-uniform distribution.

In general, no conclusion could be drawn regarding a desirable sample length for a representative result because the data are inconsistent on this point. For example, for some speakers the one-second segment happened to be loud, while for others, it was quiet. Thus as the sample became longer, for some speakers the distribution settled downward,

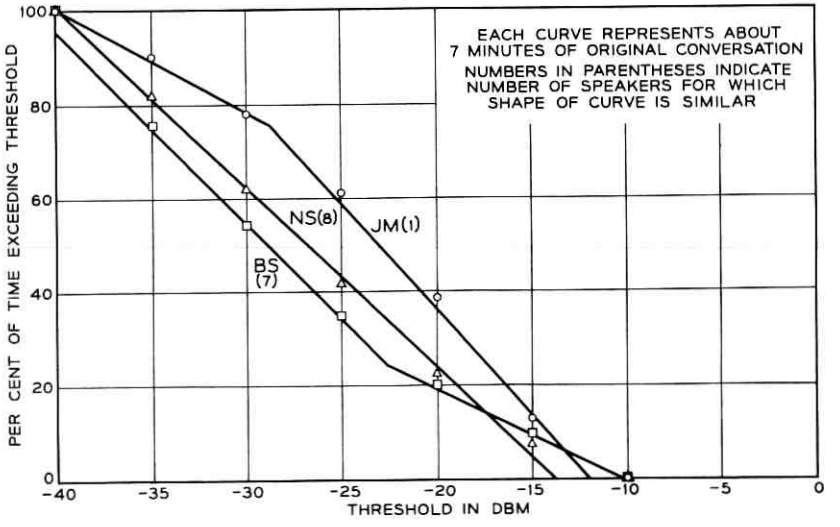

Fig. 5 — Representative speech envelope distributions for individual talkers.

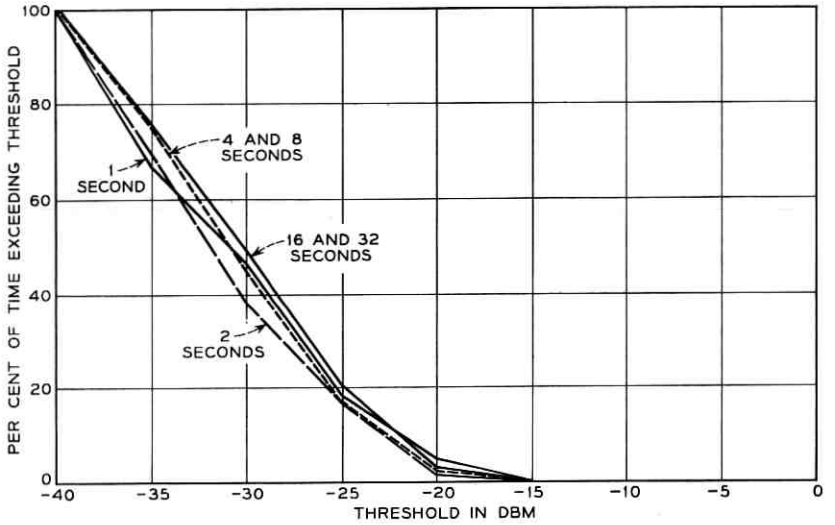

Fig. 6 — Envelope distributions for various sample lengths of continuous speech of subject MB. Data points are 5 db apart.

for others it went up, and for a few it fluctuated. Some distributions were "stabilized" at 8 seconds (that is, the 8 second function was the same as that for 16 and 32 seconds), others never stabilized.*

III. ESTABLISHING A MEASURE OF SPEECH LEVEL

3.1 *Properties of the Simple Log-Uniform Distribution*

Let X be a random variable, already defined by (1), which is uniformly distributed between a and b, where a and b are expressed in dbm. The peak value of X, at b, will be denoted X_{peak}. The density function is

$$p(x) = 1/(b - a) \tag{3}$$

and is shown in Fig. 7, along with the cumulative function. The lower limit, a, could be considered the threshold for a log-uniform speech distribution. This distribution, having a single constant value for $p(x)$, will be called the *simple log-uniform distribution* to distinguish it from the composite distribution, which will be defined in Section 3.3.

* The instability of the level distribution of a speaker is called *speech variation* and is further treated in Section 6.4.

The mean, or average value of X, is equal to

$$X_{\text{ave}} = (a + b)/2. \tag{4}$$

This quantity may be measured by obtaining the time average of X sampled and averaged only over those time intervals where X exceeds the threshold a.

The above-the-threshold rms voltage, denoted V_{rms}, is shown in Appendix B to be equal to

$$V_{\text{rms}} \text{ (in dbm)} = 6.38 + 10 \log_{10} (\Delta\text{mw}) - 10 \log_{10} (b - a) \tag{5}$$

where

$$\Delta\text{mw} = \log^{-1} \frac{b}{10} - \log^{-1} \frac{a}{10}. \tag{6}$$

That is, Δmw is the difference in milliwatts between the end points of the log-uniform distribution.

The average absolute voltage, again measured above the threshold, is denoted V_{ave} and is given by (see Appendix B)

$$V_{\text{ave}}(\text{in dbm}) = 10 \log \frac{(1000)\,(\bar{V})^2}{600} \tag{7}$$

where

$$\bar{V} = \left(\frac{8.686}{b - a}\right) (0.775) \left(\sqrt{\log^{-1} \frac{b}{10}} - \sqrt{\log^{-1} \frac{a}{10}}\right). \tag{8}$$

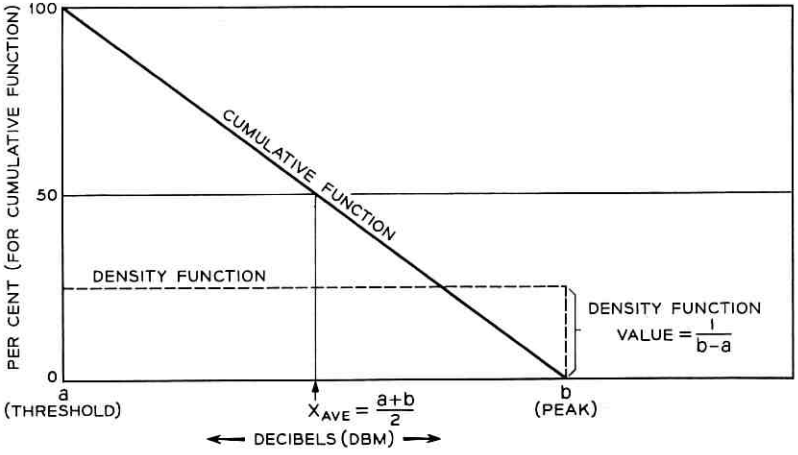

Fig. 7 — The log-uniform distribution.

(The quantity (V) $(b - a)/(8.686)$ is the difference, in volts, between the voltages to produce b and a dbm.)

Consider now what would happen to the density function of the random variable shown in Fig. 7 if the threshold a were raised (moved to the right) while the level of the pre-threshold random variable were held fixed (i.e., b remains fixed). The density function would increase in height, since $(b - a)$ would be smaller, but it would still be uniform over the range from threshold to peak. Although the peak b does not change, the quantities X_{ave}, V_{rms}, and V_{ave} do, as shown in Fig. 8. (The curves in the figure were calculated from (4), (5), and (7).)

It is clear from Fig. 8 that X_{peak} is the only quantity shown which is not dependent on the threshold setting. In fact, the other quantities vary so strongly with threshold that they would be completely meaningless were the threshold not specified.

Fig. 9 is also a plot of X_{peak}, X_{ave}, V_{rms}, and V_{ave} except that in this case the threshold is held fixed and the pre-threshold level is varied, in effect changing the value of b. The peak is seen to be the only quantity which varies on a db-for-db basis with level changes.

3.2 *The Average Peak Level*

If it were guaranteed that the levels in every speech sample were log-uniformly distributed, then our search for an ideal level measure

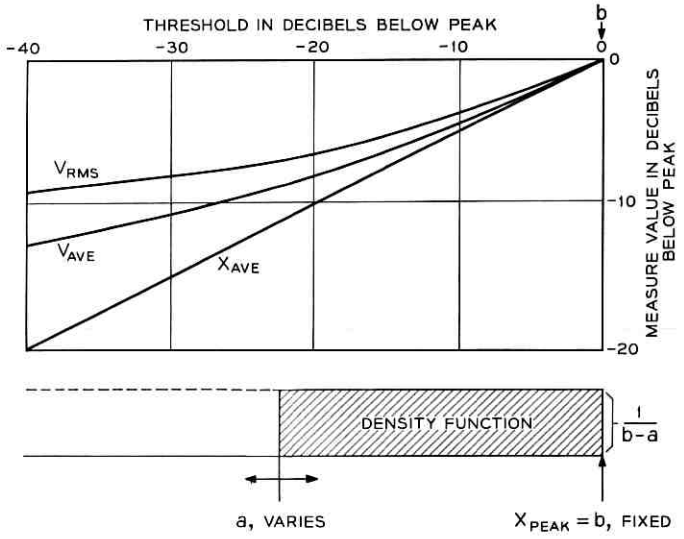

Fig. 8 — Measures of the log-uniform distribution as a function of varying threshold. Pre-threshold signal level remains unchanged.

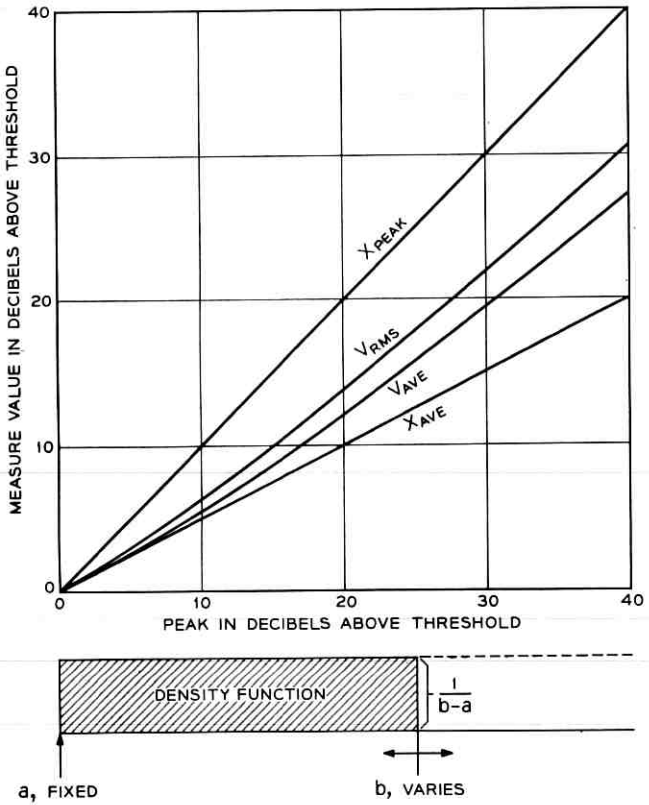

Fig. 9 — Measures of the log-uniform distribution as a function of varying pre-threshold signal level. Threshold is fixed.

would indeed be over. We would simply record the peak voltage (or peak envelope voltage) which occurs in a speech sample, and use this quantity to specify the level of the sample. It was already noted, however, that the peak is generally unsatisfactory as a level measure because it is too sensitive to isolated disturbances. Another approach to the problem is evident if we rewrite (4), solving for b:

$$b = a + 2(X_{\text{ave}} - a). \tag{9}$$

The peak can now be obtained by building a device to measure X_{ave} above some threshold, a, and then substituting X_{ave} in (9).* X_{ave} will of course be a function of a, but this is unimportant since the a dependence

* The Digital Speech Level Meter, described later in this paper, is actually constructed to measure the quantity $(X_{\text{ave}} - a)$ and then substitute this difference into (9).

will cancel out upon solving for b. We shall denote the quantity obtained applying (9) as the *average peak level*, or apl:

$$\text{apl} \equiv a + 2(X_{\text{ave}} - a). \tag{10}$$

The apl is the peak of a hypothetical simple log-uniformly distributed variable which would have produced the same X_{ave} as was actually obtained. If the speech sample levels are in fact log-uniformly distributed, the apl will be equivalent to the sample peak and will possess all the properties of the peak. If the distribution is log-uniform except for some loud extraneous sound, the apl may deviate slightly from some of the stipulated requirements of an "ideal" measure, but it will be fairly immune to the extraneous sound since the measurement is taken over the entire speech sample.

The peak can also be obtained from V_{rms} or V_{ave}. Assume a device is built to measure V_{rms} above a known threshold a. Once V_{rms} is known, (5) might be solved for b, but this is a rather difficult task. A simpler method would be to read the peak from the V_{rms} curve of Fig. 9. The resulting measure would be the peak of a simple log-uniform distribution which would yield the same V_{rms} as was actually measured.

In this study, the peak is computed with X_{ave} rather than V_{rms} or V_{ave} because the apl has a simple, linear relationship to X_{ave} (10), whereas one must resort to graphs, tables, or involved computation with the other measures. The instrumentation required to apply (10) is straightforward, as will be shown in Section V.

3.3 *The Composite Log-Uniform Distribution*

Certain speech samples have cumulative distribution functions which are markedly different from a single straight line and are therefore not from a log-uniform density function. They can, however, be approximated by log-uniform functions in the following way. Consider a process in which a random variable X_1, log-uniformly distributed between a and b_1, is observed for five minutes, and is followed by X_2, log-uniformly distributed between a and b_2, for 10 minutes. (This discussion will be restricted to two variables, but any number of X_i could be considered.) To obtain the distribution for the entire 15 minute process, one adds the two separate density functions, each weighted by a suitable factor. The fraction of time X_1 is present will be called θ_1, and θ_2 will be the fraction of time for X_2.

The structure of the composite distribution is shown in Fig. 10. The density functions are in the upper drawing and the lower drawing shows the composite cumulative function. Given the cumulative function, one

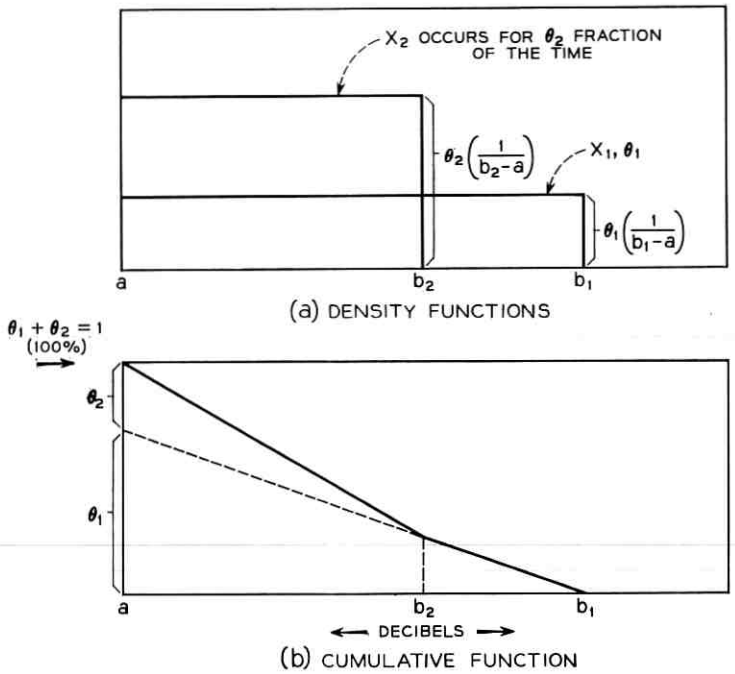

Fig. 10 — Distribution of a random variable X which is a composite of X_1 (b_1, θ_1) and X_2 (b_2, θ_2).

can immediately obtain b_1, b_2, θ_1, and θ_2. The manner in which this is done is shown on the drawing.

Routine calculation shows that the mean of the composite distribution is

$$X_{\text{ave}} = (X_1, X_2)_{\text{ave}} = \frac{a + \theta_1 b_1 + \theta_2 b_2}{2}. \tag{11}$$

This is the same mean which would result from a simple log-uniform distribution which has a peak at $\theta_1 b_1 + \theta_2 b_2$. The apl of such a distribution would therefore be a weighted average of the peaks of the log-uniform random variables which generate the composite distribution. That is,

$$\text{apl} = \theta_1 b_1 + \theta_2 b_2. \tag{12}$$

Unfortunately, the apl is no longer threshold-invariant, since θ_1 and θ_2 are themselves dependent upon the threshold. This can be seen by letting the threshold a in Fig. 10 approach b_2. The variable X_2 will become less

evident as it vanishes below the threshold, and θ_2 will approach zero while θ_1 approaches unity. The apl will therefore move towards b_1 from some point inbetween b_1 and b_2.

The amount of apl variation caused by changing the threshold will depend on the values of all the parameters involved. A theoretical analysis of this effect is included in Appendix C. It is shown that for most of the speech samples in this study, apl variations in the order of one db could occur if the threshold were allowed to become close to b_2. (If the threshold is too close to b_2, the variation is more severe.) Some experimental measures of this effect, included in Section 6.3, support the theoretical estimates.

3.4 *Suitability of Log-Uniform Approximation to Speech Levels*

Several of the speech samples analyzed here have cumulative functions which are quite linear. For a few others, the functions can be very well fitted by two lines, indicating a two variable composite distribution.

Now consider Fig. 4 which shows a distribution which slopes off gradually and for which the two line approximation introduces a noticeable error at the break point. This error can be reduced if a three line fit is made, and if ten lines are used, the error all but vanishes. The Fig. 4 curve can therefore be considered a composite of a large number of log-uniform distributions, all having a common threshold and having successively higher peaks.* In general, the composite distribution is valid if the cumulative function exhibits the following two properties:

(1.) For all points above the threshold, the curve cannot break downward (its second derivative cannot be negative).† This guarantees that the composite distribution contains no simple distribution which exists entirely above threshold.

(2.) If the curve breaks upward, the lowest break point (in dbm) must be above the threshold. Thus there are no density function peaks below threshold.

Every speech distribution noted by the author obeys both of the above rules if proper care is taken in determining the threshold. For the composite log-uniform distribution to be valid, the threshold may be set anywhere in the linear range of the curve between the downward and upward break points. This is generally a broad range of at least 15 db.

* The suitability of a Gaussian model is discussed in Appendix D.

† This rule is violated in Fig. 4 if the −40 dbm threshold is used. In making measurements from the curve, the threshold is raised until this rule is obeyed, as is done in Appendix A. For later reference, we might note here that the Digital Speech Level Meter uses a −30 dbm threshold, which is sufficiently high to clear all downward break points of the speech used in this study.

The threshold should, however, be set in the lower part of this range to minimize the threshold dependence of the apl.

IV. THE VU METER

4.1 *Technique of Using the VU Meter*

The VU meter is a widely accepted speech level measuring instrument. Its basic design consists of an amplifier, full-wave rectifier, and meter. The characteristics of the unit, especially the meter movement, are standardized and may be found in several references.[6] A standard procedure for reading the VU Meter (when monitoring a telephone conversation) has been adopted and is described by Carter and Emling:[7]

"The volume used by the party selected is the arithmetic average in VU of a series of individual volume measurements made on a selected party's speech throughout the conversation.

"An individual volume measurement provides a single figure based on a portion of speech several seconds in length (say 3 to 10 seconds). It is . . . the visual or inspection mean of the highest meter deflections, exclusive of the one or two very highest deflections, observed during the measuring period.

"[Typically], in a 5-second measuring interval, for example, there may be about 25 syllables, with a meter deflection or swing resulting from everyone of these. These swings can be divided roughly into two types: a large group of relatively small swings from the weaker syllables and a small group of high swings from the six or seven loudest syllables. It is on this second class of strong swings that the volume measurement is based; the highest one or two are excluded, however, since these may be somewhat special as to emphasis or accent and are not related closely to the five or six remaining strong swings."

One could regard the above process as a method of estimating the "average peak" of the meter response. Judging from the work reported in the previous sections of this paper, this measure is ideally a very good indication of speech level. In practice, however, it exhibits variability first because an observer's readings of the same speech sample are not repeatable, and secondly because different observers show different biases in reading the meter. Measurements of these variabilities were made by the author in a brief unpublished study. The standard deviation of a single observer was found to be as much as 1.5 VU (db) and a range of observer bias of almost 3 VU occurred among observers.

Shearme and Richards[4] report similar findings. They find that a

"trained observer will yield 5 per cent of readings as much or greater than 2 db away from the mean value." This corresponds to a standard deviation of 1 VU, obtained with our most experienced observer. Shearme and Richards also report that "even with trained observers a total range [of observer bias] of 4 db is encountered."

4.2 *Relationship of this Study to VU Measurements*

It is apparent that the VU meter yields imprecise readings when used in an attempt to make objective speech level readings. A major source of the variability is due to the human observer, and one way of removing this variation is to instrument the reading process. This could be done by constructing a device which would follow a set of rules similar to those stated by Carter and Emling.[7] But these rules were tailored for an observer and perhaps there could be a better measure of speech level when the human limitation is removed.

Indeed, the apl has been shown to have a direct relationship to the underlying speech level distributions. Principally for this reason, the author chose to construct a device which measures the apl and not the VU level of a speech sample. It will be shown that this device, called the Digital Speech Level Meter, yields readings of less variation than VU readings and is therefore potentially a more precise instrument for measuring levels.

This does not imply, however, that the Digital Speech Level Meter is a total replacement for the VU meter. The VU meter is generally adequate for setting a "good" recording level, and its readings are often considered to be an indication of subjective loudness. This is usually argued on the basis of the design of the needle movement. Further reasoning is based on the ground that the meter is read by an observer who himself has some ideas about the loudness of the signal. Thus the VU and apl readings reflect somewhat different properties of speech. They may be compared with reference to objective level measurements, but in other respects each measure must be judged on its own merits.

V. THE DIGITAL SPEECH LEVEL METER

5.1 *Obtaining the Log-Average Voltage*

It is shown in Section 3.2 that the apl is easily obtained once the average of the log-uniform distribution (X_{ave}) is known. Fig. 11 illustrates the technique used to obtain X_{ave}. Speech is full-wave rectified, filtered, and applied to a log voltage to frequency converter. That is, the output

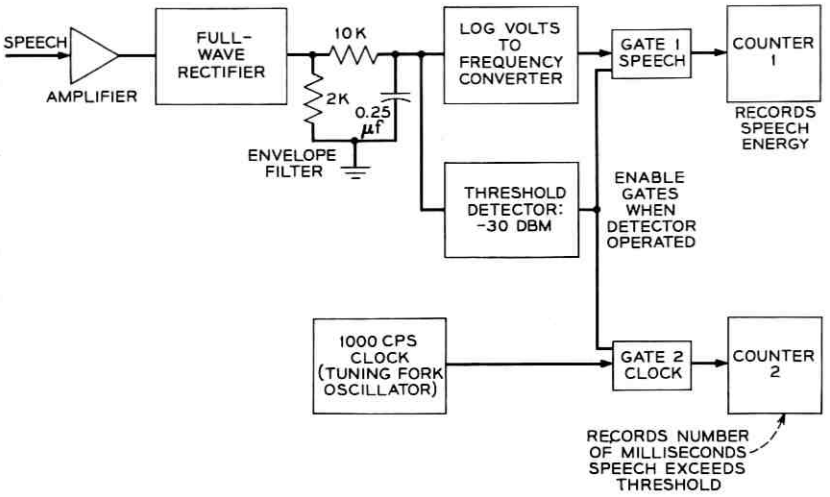

Fig. 11 — Basic design of digital speech level meter.

frequency exhibits uniform incremental changes with uniform changes in the decibel level of the input. The linearity extends over about a 30-db range.

The filtered signal is also applied to a speech detector consisting of solid state circuitry and having insignificant pickup and hangover times. The threshold of the detector is −30 dbm re OTL. This value, rather than −40 dbm, was chosen to keep the speech in the linear range of the detector. This threshold setting is still low enough to keep the apl nearly independent of the threshold for those speech samples which do not have a simple log-uniform distribution.*

The speech detector operates two gates. The "speech gate" sends pulses from the voltage controlled oscillator to counter 1, whose reading may be interpreted as the accumulated energy of the log voltage. The "clock gate" sends pulses from a 1000-cps clock to a second counter, whose reading specifies the amount of time the speech level has exceeded the threshold.

To obtain a level reading for a speech sample, the counters are first

* The statement that "the apl is nearly independent of threshold" does not imply that one may randomly vary the Speech Level Meter threshold without affecting the meter reading. Since the meter measures X_{ave}, which itself *does* depend on threshold (4), the threshold must be taken into account in solving for the apl (10). Thus consider two meters with thresholds of −35 and −30 dbm, respectively. If each is properly calibrated with respect to its own threshold, then each should read approximately the same apl for the same speech sample.

reset to zero and then the speech sample is played. When finished, the counter 1 reading is divided by the counter 2 reading to obtain an average frequency. This of course may be directly converted to average log voltage since the frequency is a linear function of the log voltage.

The instrumentation up to this point is very similar to the method used by P. D. Bricker to obtain a measure of speech level.[8] Bricker's circuit is almost identical to that of Fig. 11 except that his speech detector has a 200-msec hangover time and his first counter is driven by a *linear* voltage to frequency converter. Thus his average frequency obtained from the two counter readings is an approximate measure of V_{ave} rather than X_{ave}. Bricker's success in estimating VU readings with his technique provided considerable encouragement for the present study.

5.2 *Obtaining a Direct Reading of the APL*

The above procedure requires that the observer write down two numbers, divide them, and apply a conversion to yield the apl value. One way of instrumenting these operations is as follows. In Fig. 11, a flip-flop is installed on counter 2 in such a way as to shut down the whole device upon observing 1000 msec of speech. This automatically accomplishes the necessary division. Counter 1 is constructed to count toward zero starting from some negative number whose value depends on the calibration of the voltage to frequency converter.

The voltage to frequency converter is adjusted so that for a 1 db increment in the level of a sine wave input signal (thereby increasing the mean of the logarithm of its envelope by 1 db), the frequency converter changes by 2.0 cps. (This is actually 20 cps, but a decimal point is inserted in the read out.) Recall now that if *speech* has its over-all level increased by 1 db, the mean of its logarithm is increased by only 0.5 db. The converter, having a 2 to 1 "frequency to db" conversion, will exhibit a frequency change of 1.0 cps, correctly reflecting the change in speech level.

Because of the nature of the speech level meter calibration, it will not work properly in its present form if used to measure the levels of a tone or other signals which do not have a log-uniform distribution.*

The meter can be set to read speech over intervals of time other than

* Consider a random variable Y having a probability distribution such that the peak is linearly related to the above-the-threshold average (denoted $\bar{y}$) by $\bar{y} = a + [(b - a)/k]$ where a is the threshold, b is the peak, and k is a constant independent of a. The (log-) uniform distribution, in which $k = 2$ (see (4)), is one of a large class of such distributions. Although the technique described in this paper for measuring speech levels might be suitable for measuring other random variables, the speech level meter is calibrated for $k = 2$ and requires a uniform distribution.

one second without subsequent division. This is accomplished by inserting flip-flops just ahead of the counters. For example, if one flip-flop is placed in front of each counter, the counting rate will be halved and the meter will read directly for 2 seconds of speech.

Fig. 12 is a photograph of the speech level meter. To obtain a reading, the observer presses the reset key which turns off the display and starts the internal counters integrating the speech energy. When the lower counter reaches 1000 (this display is usually not illuminated), the upper display is turned on, the observer records the number, and again pushes the reset key if another reading is desired.

It is possible to modify the meter so it does not stop after a fixed time interval but continues counting in the manner described in the previous section. This and several other options are provided by various front panel controls.

VI. RESULTS OBTAINED WITH THE SPEECH LEVEL METER

6.1 *Scope of the Results*

The data presented here represent measurements made on 16 samples of telephone speech, each about 7 minutes long. All of the samples were

Fig. 12 — Digital speech level meter.

recorded on the same circuit. Any conclusions which are based on these data must be regarded as limited in scope and can be broadened only through further data acquisition. The data included here should, however, suggest the general limits of performance which can be expected.

6.2 *Measuring Technique*

All level measurements reported here were made by taking the average of a succession of 4 second readings. If the meter is reset immediately after each reading, this technique yields a result which is equivalent to that obtained by allowing both counters (Fig. 11) to run continually and forming a ratio at the end of the sample. The present method was adopted because it is easy for the observer to use, and reads directly, without conversion.

6.3 *Response of the APL to Changes in Level*

The requirement that the apl be invariant with threshold is equivalent to the requirement that it vary on a db-for-db basis with attenuation or amplification of the voice signal. This is true because a signal attenuation of, say, 5 db will yield the same shape probability density function as will raising the threshold by 5 db, although the resulting distributions will differ in absolute levels. The apl's of the two new distributions should ideally differ by 5 db.

The following experiment illustrates the effect of level changes on the apl. Four 7-minute samples of speech were each played through the speech level meter at three different levels, each 5 db apart. The readings were as follows:

TABLE I

EFFECT OF OVER-ALL LEVEL CHANGES ON APL READINGS

Level	Speaker			
	AD	JS	MH	CB
−5 db	−20.01	−20.57	−15.26	−17.57
	Δ = 5.28	Δ = 4.68	Δ = 5.48	Δ = 4.85
Normal	−14.73	−15.89	−9.78	−12.72
	Δ = 5.63	Δ = 6.06	Δ = 4.06	Δ = 4.61
+5 db	−9.10	−9.83	−5.72	−8.11

With one exception, the apl readings for each speaker reflect the speech level variations with an error of less than 1 db over the 5 db

increments, and all of the speakers are within 1 db for the 10 db increments. This is in general agreement with the theoretical results of Appendix C, namely, that the apl is threshold invariant to within about 1 db for most speech samples.

6.4 *Repeatability of Meter Readings*

The speech samples of three talkers were each played ten times to determine the variation which might be expected in obtaining repeated measurements. The estimations of the standard deviations of the levels for the samples were 0.080, 0.154, and 0.043 db. The meter readings are therefore highly repeatable.

A sample of the readings taken during one of the runs is shown in Table II. Only 5 of the original 10 columns are shown. These data are included to illustrate two very different sources of variation which occur in taking readings.

The first source is *speech variation*, which exists because the speaker varies his level as he talks. This variation is reflected in the range which exists in the numbers in a single vertical column. For example, one concludes from the data in the first column that the apl for the entire speech sample is -10.99 dbm, with an estimated standard deviation for any randomly chosen 4-second sample of 3.57 db.

Speech variation does not enter into the repeatability of measuring the level of a *particular* speech sample. In this case, the variation of this measure would be determined in part by the variability of reading the same 4-second sample and by the number of samples taken. A rough idea of the repeatability of a 4-second sample reading is found by reading across the top horizontal row of Table II. (Other rows are not suitable for comparison because of timing errors in resetting the meter. That is, the fifth reading may not be taken for exactly the same speech sample every time the tape is played.) A rough guess at the standard deviation of a particular sample is 0.3 db, based on cursory inspection of data taken on short speech samples (not included here).

If this value of 0.3 db is divided by $\sqrt{N}$, where N is the number of 4-second readings, one might expect to obtain the standard deviation of the average of the entire speech sample. For the data of subject SK, as shown in part in Table II, N equals 30, which would lead us to expect a σ of 0.055 db. The measured value of σ was 0.154 db. The data from the other two speakers having deviations of 0.080 and 0.043 db are more in line with the expected value of 0.06 db. For each of these speakers, $N \approx 25$.

TABLE II

Repeated measurements of a seven minute sample of speech of Subject SK. (Only 5 of the original 10 columns are shown. All readings are negative numbers.)

1	2	3	4	5
19.3	19.1	19.2	19.0	19.2
11.5	11.8	12.0	11.4	11.1
18.8	18.7	18.9	18.7	18.0
12.4	12.5	12.4	12.5	12.5
15.4	15.3	15.5	15.1	15.5
9.0	9.3	8.9	9.5	8.9
9.5	10.8	7.4	11.4	9.1
5.2	4.4	5.1	1.5	5.2
4.7	4.9	4.5	7.7	4.6
13.4	11.4	13.6	5.8	10.0
12.4	12.0	12.2	13.6	12.1
15.0	14.6	15.4	11.2	14.1
10.7	9.9	10.7	12.4	9.5
13.5	15.8	13.7	14.8	15.7
9.8	8.2	8.2	9.7	8.5
9.3	9.7	9.7	10.6	9.4
8.2	8.7	8.7	7.9	7.8
12.6	9.2	10.3	8.8	12.5
10.7	12.0	12.0	13.0	10.7
13.7	10.8	10.6	11.8	13.2
11.9	14.0	13.8	13.7	11.6
10.3	10.4	10.1	12.2	10.5
8.1	8.2	8.0	6.2	8.9
10.8	10.8	10.8	11.5	9.0
10.4	10.2	10.6	10.4	11.0
8.9	8.7	9.0	10.0	6.8
12.0	12.1	9.5	6.7	11.5
9.4	9.7	10.0	11.1	8.6
2.6	2.6	5.4	7.6	2.4
10.1	9.0	6.4	0.7	10.5
		Column Averages		
−10.99	−10.83	−10.75	−10.55	−10.61

The variation in any one column is predominantly due to speech variation. For example, σ = 3.57 db for column 1. Differences in the column averages are due to measurement variation. For all 10 columns, σ = 0.154 db.

6.5 *A Comparison of Different Types of Measurements*

The apl readings made by the speech level meter were compared with the apl estimates based on the graphical technique described in Appendix A. The results are shown in Table III. Also included in this table are the VU readings for the samples taken by Miss Kathryn L. McAdoo, an experienced VU reader.[9]

The meter and graphical apl levels are plotted against each other in Fig. 13. The linear least mean squares fit passes through the two averages

TABLE III

APL READINGS FOR ALL SPEECH SAMPLES*

Subject	Meter APL	Graphical APL	VU
RT	−19.56 dbm	−19.76 dbm	−24.68 vu
AD	−15.49	−17.5	−23.00
MB	−17.49	−18.25	−24.24
JS	−16.56	−19.0	−23.00
PF	−14.20	−15.0	−19.67
JM	−8.79	−11.75	−16.45
ES	−21.11	−21.75	−24.66
PR	−18.24	−19.55	−23.67
MH	−10.00	−12.3	−17.73
SR	−9.22	−11.07	−14.41
SK	−10.77	−12.85	−19.14
CB	−13.28	−14.0	−18.60
SS	−16.97	−20.64	−22.77
BS	−13.96	−15.56	−20.33
NS	−13.35	−13.7	−20.08
VB	−18.05	−21.76	−27.12
Averages	−14.82	−16.53	−21.22

Rank Order Correlations: Meter APL vs Graphical APL = 0.944
Meter APL vs VU = 0.949

* These readings should not be directly compared with those in Table I because of a difference in calibration used for the two sets of readings.

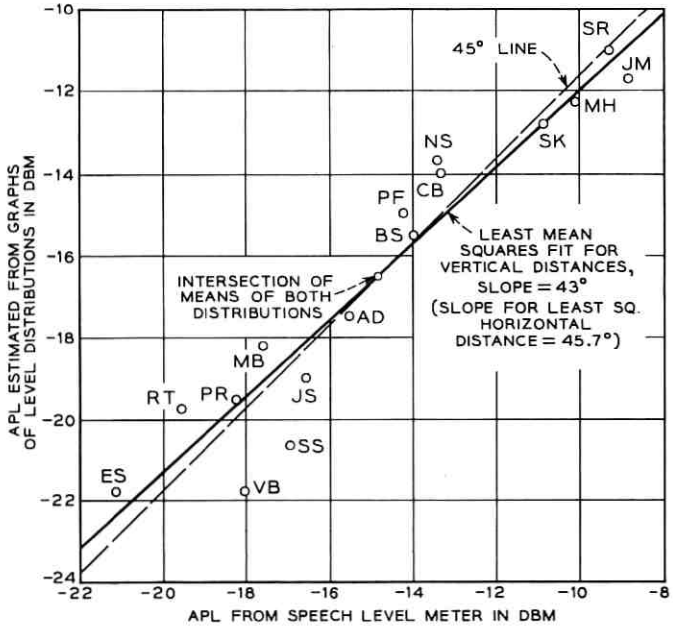

Fig. 13 — A comparison of graphical and meter APL readings.

of both distributions and has a calculated slope of 43°. Because the slope is so close to 45°, we conclude that on the average, the methods are consistent with each other in comparing relative levels among speakers.

The means of the distributions do not coincide, showing an over-all bias such that the meter reads about 1.7 db higher than the graphs, with a variation of about 1 db. This can be attributed to several factors, such as differences in the instrumentation used in the meter and in the equipment which generated the graphs, and the inadequacy of the two-line approximation used in the graphical analysis. Another factor is the threshold dependency of the apl; the meter had a threshold of −30 dbm while the threshold in the graphical analysis was closer to −40 dbm.

The VU readings and meter apl levels for the 16 speech samples are plotted against each other in Fig. 14. The slope of the linear least mean squares fit is 41.3°, showing that the apl and VU readings tend to differ by within 2 db of a constant over the range of the speech samples. (A 15-db change in meter level readings produces a 13 db change in the least mean squares fit.)

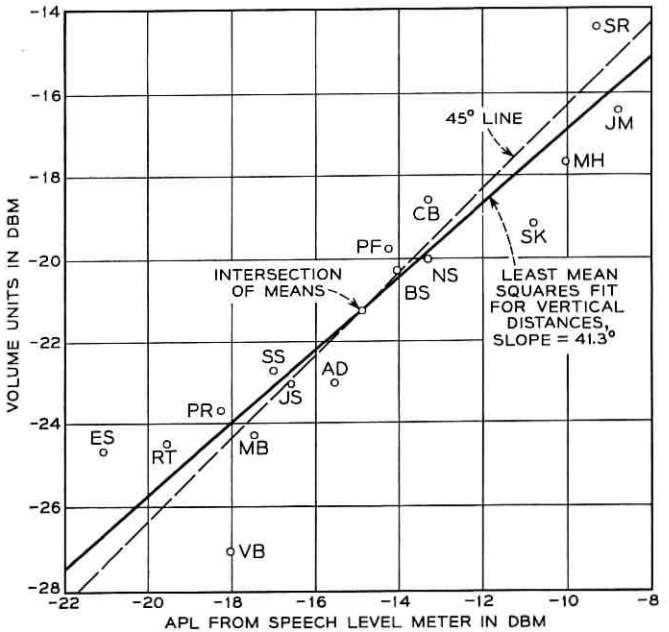

Fig. 14 — A comparison of VU and meter readings.

VII. CONCLUSION

In this study we have shown that the apl is a one-dimensional objective measure of speech level and that it satisfies, within certain bounds, a stipulated set of requirements of an "ideal" measure. The Digital Speech Level Meter is presently undergoing further tests which will help to determine more precisely the properties of the apl.

One unanswered question is that of determining a relationship between meter readings and subjective impressions of loudness. Other areas of further study include measuring levels of clipped or volume limited speech, high-fidelity speech (as opposed to telephone speech), and possibly other types of signals such as noise. Note that the demonstrated correspondence between level distributions of telephone and high-fidelity speech (Figs. 1 and 2) implies that the meter would work equally well with either type of speech.

The level measuring technique described here has many possible applications if further experimentation indicates the method to be suitable. The limited data already available show that the technique is promising.

VIII. ACKNOWLEDGMENT

I am most grateful to Miss Donna Mitchell for the many long and tedious hours she spent gathering and analyzing almost all of the data obtained for this study. I am also indebted to F. S. Fillingham and S. E. Michaels for their assistance in the design and construction of the experimental meter.

APPENDIX A

Procedure Used to Obtain Graphical APL Measurements

The cumulative function for the syllabic waveform of the speech of SR is shown in Fig. 15. This curve is chosen for demonstration because it has two break points, and obtaining the apl reading for it involves more steps than for most other graphs.

Notice that the break point near -35 dbm is below the -30 dbm threshold used in the speech meter and therefore does not affect the meter reading. This is the case for all speakers having such break points. The break point is therefore ignored and the curve is extended to 100 per cent as a linear extrapolation. In computing θ_1 and θ_2, the vertical

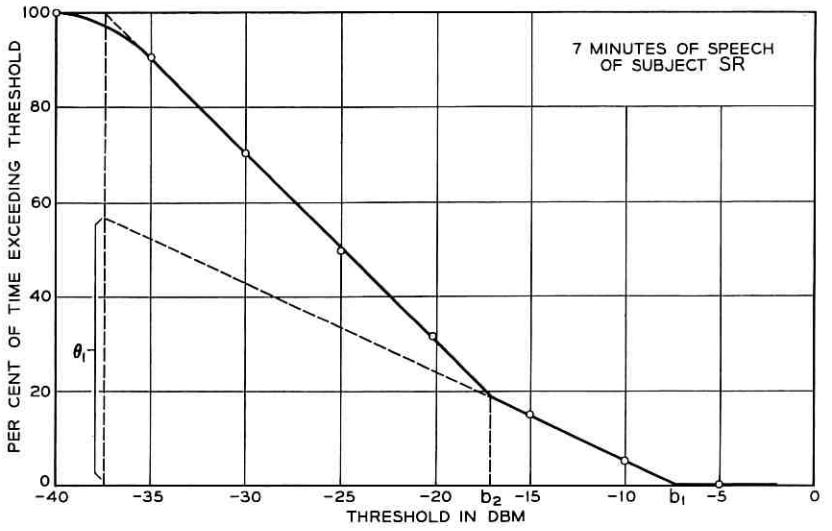

Fig. 15 — Estimating APL from a distribution having a low-level break point.

line from which θ_1 is chosen must be the line at which the cumulative function reaches 100 per cent. This no longer occurs at -40 dbm but rather at -38 dbm.

Having found θ_1, θ_2, b_1, and b_2, the apl is computed from (12). Table IV is a tabulation of these quantities for all of the 16 speakers.

TABLE IV

GRAPHICAL CALCULATION OF THE APL FOR 16 SPEAKERS

Speaker	b_1 (dbm)	θ_1	b_2	θ_2	Graphical APL
RT	−15.5	.31	−22	.69	−19.76
AD	−17.5	1	—	—	−17.5
MB	−18.25	1	—	—	−18.25
JS	−19.0	1	—	—	−19.0
PF	−15.0	1	—	—	−15.0
JM	−11.75	1	—	—	−11.75
ES	−21.75	1	—	—	−21.75
PR	−15.0	.30	−21.5	.70	−19.55
MH	−9.0	.50	−15.6	.50	−12.3
SR	−7.5	.58	−16.0	.42	−11.07
SK	−9.3	.52	−16.7	.48	−12.85
CB	−14.0	1	—	—	−14.0
SS	−13.5	.27	−23.3	.73	−20.64
BS	−10.2	.56	−22.4	.44	−15.56
NS	−13.7	1	—	—	−13.7
VB	−17.2	.57	−27.8	.43	−21.76

APPENDIX B

Derivation of V_{rms} and V_{ave} for the Log-Uniform Distribution

B.1 V_{rms}

Let X be a random variable uniformly distributed between a and b (Fig. 7) such that

$$x_{(\text{dbm})} = 10 \log \frac{1000(v^2)}{600}. \tag{13}$$

Let Y be a random variable which represents the power in milliwatts dissipated in a 600 ohm resistor. Then

$$x = 10 \log y. \tag{14}$$

From Fig. 7,

$$\text{Prob } X \leqq x = \int_a^x \frac{1}{b-a}\, dx = \frac{x-a}{b-a}. \tag{15}$$

Substituting (14) into (15),

$$\text{Prob } Y \leqq y = \frac{(10 \log y) - a}{b-a}. \tag{16}$$

Recall that for any variable Z,

$$\log_{10} Z = (0.4343) \ln_e Z. \tag{17}$$

This is used in differentiating (16),

$$\begin{aligned} p(y) &= \frac{4.343}{y(b-a)} \text{ for } y \text{ between } \log^{-1} \frac{a}{10} \text{ and } \log^{-1} \frac{b}{10} \\ &= 0 \text{ elsewhere.} \end{aligned} \tag{18}$$

Equation (18) tells us that the density function of the power in milliwatts is of the form of a hyperbola, not an exponential as might be guessed from the uniform distribution of log y.

To obtain V_{rms}, obtain the average power, that is, the expectation of y

$$E(y) = \int_{\log^{-1}(a/10)}^{\log^{-1}(b/10)} yp(y)\, dy = \int_{\log^{-1}(a/10)}^{\log^{-1}(b/10)} \frac{4.343}{b-a}\, dy. \tag{19}$$

Define

$$\Delta \text{ mw} = \log^{-1} \frac{b}{10} - \log^{-1} \frac{a}{10}. \tag{20}$$

Then, integrating (19),

$$E(y) = \frac{4.343}{b - a} \ (\Delta \text{ mw}) . \tag{21}$$

The rms voltage is the voltage required to produce this average power. Expressed in dbm,

$$V_{\text{rms}} = 10 \log E(y) = 6.38 + 10 \log (\Delta \text{ mw}) - 10 \log (b - a). \tag{22}$$

B.2 V_{ave}

Let V be a random variable representing the *absolute* voltage which would generate X. This voltage is monotonically related to the power by

$$y = \frac{(1000) (v)^2}{600} . \tag{23}$$

Taking logarithms,

$$10 \log y = 10 \log \frac{10}{6} + 20 \log v. \tag{24}$$

Substitute into (16),

$$\text{Prob } (V \leqq v) = \frac{\left(10 \log \frac{10}{6} - a + 20 \log v\right)}{b - a} . \tag{25}$$

Differentiating,

$$\begin{aligned} p(v) &= \frac{8.686}{(b - a)v} \text{ for } v \text{ between } 0.775 \sqrt{\log^{-1} \frac{a}{10}} \\ &\qquad \text{and } 0.775 \sqrt{\log^{-1} \frac{b}{10}} \\ &= 0 \text{ elsewhere.} \end{aligned} \tag{26}$$

Define

$$\Delta v = (0.775) \left(\sqrt{\log^{-1} \frac{b}{10}} - \sqrt{\log^{-1} \frac{a}{10}} \right) . \tag{27}$$

Then in an identical manner of obtaining (19),

$$\bar{V} = E(V) = \left(\frac{8.686}{b - a} \right) \Delta v. \tag{28}$$

Expressed in dbm,

$$V_{ave} \text{ (in dbm)} = 10 \log \frac{(1000)(\bar{V})^2}{600}. \tag{29}$$

APPENDIX C

Variation of the APL With Threshold

Fig. 10 shows a composite log-uniform distribution of two variables, X_1 and X_2, each occurring above threshold for θ_1 and θ_2 fractions of the total time, respectively. The apl equals $\theta_1 b_1 + \theta_2 b_2$, and since θ_1 and θ_2 vary with threshold, the threshold will also influence the apl.

Let the threshold be increased to some new a', which is somewhere between a and b_2. We define

$$\varphi_1 = \left(\frac{b_1 - a'}{b_1 - a}\right)\theta_1 \tag{30}$$

$$\varphi_2 = \left(\frac{b_2 - a'}{b_2 - a}\right)\theta_2. \tag{31}$$

The variables φ_1 and φ_2 represent the respective proportions of X_1 and X_2 which remain above threshold, each weighted by the original value of θ_i. Since $\varphi_1 + \varphi_2 \neq 1$, new values for θ_i are obtained by letting

$$\theta_1' = \frac{\varphi_1}{\varphi_1 + \varphi_2}, \qquad \theta_2' = \frac{\varphi_2}{\varphi_1 + \varphi_2}. \tag{32}$$

Knowing θ_1', θ_2', b_1, and b_2, it is possible to calculate a new apl$'$ and subtract from it the original apl to determine the variation produced by the threshold change. The general relationship between apl variation and threshold change is rather involved, and will be omitted here. From Fig. 10, however, one can see that if a is moved a short distance to the right, the effect upon the apl will vary, depending upon whether the move was made very near to b_2 or some distance from it. Assume, for example, θ_2 is very large (say 0.95), causing b_2 to dominate the apl for low thresholds. A 2 db change in a, if a is low, may hardly affect the apl, but if a is very near b_2, the 2 db change could eliminate the X_2 variable and cause the apl to shift rapidly toward b_1.

Calculations were made to determine what the graphical apl's in Table IV (Appendix A) would have been had a threshold of -25 dbm been used instead of -40 dbm. This is a severe test, as -25 dbm is a

TABLE V

VARIATIONS IN THE APL WITH RESPECT TO THRESHOLD

Speaker	−40 dbm apl	−25 dbm apl	Differences, db
RT	−19.76	−18.59	1.17
PR	−19.55	−18.42	1.13
MH	−12.30	−11.83	0.47
SR	−11.07	−10.36	0.71
SK	−12.85	−12.20	0.65
SS	−20.64	−17.31	3.33*
BS	−15.56	−12.48	3.08*
8 other speakers	No difference, since $\theta_2 = 0$		

* For a −30 dbm threshold (instead of −25 dbm), these differences are: SS, 0.93 db; BS, 1.24 db.

somewhat unreasonable threshold for these particular speech samples. (In fact, since the new threshold clears b_2 for subject VB, this sample is not considered in this comparison). The apl comparisons are as shown in Table V.

Based upon the results in the table, the statement is made that for most speech samples in this study, the apl is, within about 1 db, invariant with threshold.

It might be possible to reduce the apl threshold dependence by subtracting a small correction factor from fairly low readings, in which the peaks are close to the threshold. The value of the correction would taper off for higher apl's. Further study may determine whether such a procedure is advisable or feasible.

APPENDIX D

The Gaussian Distribution as a Speech Model

The Gaussian distribution, in addition to the log-uniform distribution, may serve as a model for the speech data. The cumulative speech function of Fig. 4 is replotted in Fig. 16, along with the (log-) Gaussian cumulative function with mean (μ) of −27.9 dbm and standard deviation (σ) of 8.3 db. The mean was set equal to the speech median, while σ was derived by setting $\sigma/2$ equal to the +19.2 per cent point above the median (−32.0 dbm).

The two curves are very similar* and might be even more so if the

* The fact that the cumulative function for all speakers is nearly Gaussian is *not* a consequence of the central limit theorem. The theorem states that the sum (or average) of n independently distributed variables will have a nearly Gaussian dis-

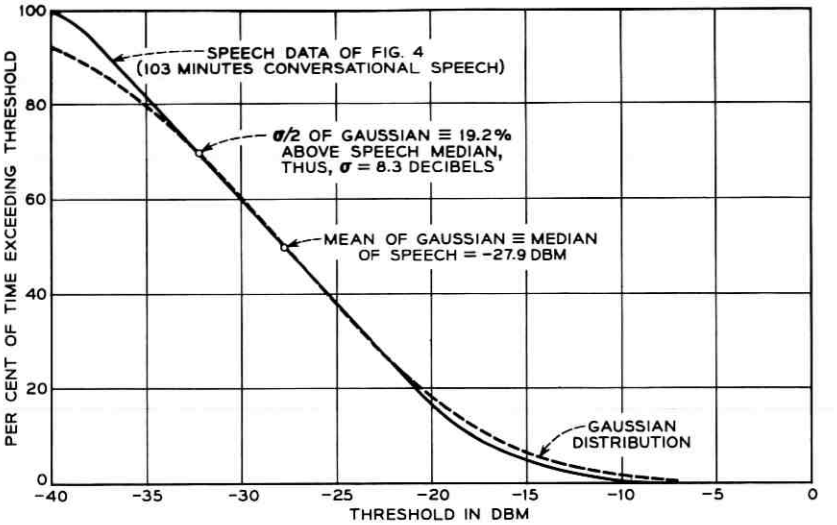

Fig. 16 — Speech distribution compared with Gaussian distribution.

speech function had not been normalized to 100 per cent at −40 dbm. (If the unnormalized speech curve were used, the Gaussian parameters would need readjustment.) The Gaussian model is, of course, most familiar and is of great help in analysis. For specifying speech levels, however, it is inferior to the log-uniform model for the following reasons:

(1.) It is not unidimensional; both μ and σ are required to specify one distribution curve.

(2.) There seems to be no clear-cut method of obtaining either μ or σ. In Fig. 16, μ was equated to the speech median, but the median is dependent upon the threshold. If the threshold were removed, the circuit would be under constant observation and the silent intervals would introduce data of uncertain significance. In Fig. 16, σ was obtained from a quantile point (19.2 per cent above median), and this also varies with threshold.

(3.) For some speech distributions, the simple log-uniform model is a better fit than the Gaussian model (see the NS curve of Fig. 5). And even when the Gaussian fit is better, as in Fig. 16, the composite log-

tribution when n becomes very large. The speech distribution in Fig. 16 is of *one* variable: the waveform of the envelope of a 103 minute speech sample. If each speaker had a simple log-uniform distribution with an apl of −10 dbm, then the 103 minute sample would have precisely that distribution. It may be that the overall level distribution for many speakers is approximately Gaussian, but this is a result of the nature of the speakers and not of a limiting theorem.

uniform model is still valid, if the threshold falls in the linear range of the cumulative function.

We are actually in the favorable position of not caring whether the distribution is Gaussian or log-uniform, as our only concern is that there exists a (quasi-) linear part of the cumulative function, and either of the above models provides for this. For this reason, the composite log-uniform distribution, which embraces both of these models, is used as a basis for specifying a unidimensional speech level.

REFERENCES

1. Sivian, L. J., Speech Power and Its Measurement, B.S.T.J., *8*, Oct., 1929, p. 646.
2. Dunn, H. K., and White, S. D., Statistical Measurements on Conversational Speech, JASA, *11*, Jan., 1940, p. 278.
3. Davenport, W. B., An Experimental Study of Speech-Wave Probability Distributions, JASA, *24*, July, 1952, p. 390.
4. Shearme, J. N., and Richards, D. L., The Measurement of Speech Level, Post Office Electrical Eng. J., *47*, 1954, p. 159.
5. Brady, P. T., A Technique for Investigating On-Off Patterns of Speech, B.S.T.J., *44*, Jan., 1965, p. 1.
6. Beranek, L. L., *Acoustic Measurements*, John Wiley, New York, 1949, p. 504.
7. Carter, C. W., and Emling, J. W., Unpublished report on making volume measurements on telephone message circuits, Bell Telephone Laboratories, 1950.
8. Bricker, P. D., A Technique for Objective Measurement of Speech Level, J. Acoust. Soc. Amer., Aug., 1965, letter.
9. McAdoo, K. L., Speech Volumes on Bell System Message Circuits 1960 Survey, B.S.T.J., *42*, Sept., 1963, p. 1999.

Some Extensions of Nyquist's Telegraph Transmission Theory

By R. A. GIBBY and J. W. SMITH

(Manuscript received April 21, 1965)

The conditions necessary to achieve undistorted transmission of a pulse signal over a channel of finite bandwidth have been set down by Nyquist. These conditions are extended in this paper to eliminate the bandwidth restrictions. Conditions on the real and imaginary parts of the overall system characteristic which lead to the elimination of intersymbol amplitude and pulse width distortion are found. These generalized constraints do not depend on any sharp band limitation and permit one to find ideal conditions for band pass and gradual cutoff systems. The application of Nyquist's conditions usually amounts to equalizing the transmission characteristics in order to approximate an overall linear phase and some sort of symmetrical amplitude roll-off. This paper shows that the principles of channel shaping for distortionless transmission are a good deal more flexible than this. The application of this more general interpretation of Nyquist's theory is illustrated by several examples.

I. INTRODUCTION

Nyquist's classic paper[1] considered the conditions necessary for digital data transmission without intersymbol distortion, and these conditions have provided the guides for system design for many years. However, Nyquist treated the case in which no energy is transmitted at a frequency above twice the signaling speed, although he mentioned the general case in passing. As a consequence, his results cannot be applied directly to cases in which the amplitude characteristics extend beyond twice the signaling speed (gradual cutoff systems) or baseband systems without low-frequency components (bandpass). In addition, Nyquist's theory has been incompletely exploited in practice. The usual application of the principles of channel shaping amounts to equalizing the phase to make it linear across the band, and equalizing the amplitude to produce a symmetric roll-off characteristic. This procedure is valid and consistent with the theory, but is only a special application of the theory.

This paper extends the previous results by showing that it is not necessary to restrict the bandwidth to arrive at an efficient description of the amplitude and phase constraints for distortionless transmission. In other words, transmission systems without a sharp cutoff frequency are considered and constraints on the system characteristics are obtained. The removal of the bandwidth limitation means that one can easily find the constraints for gradual cutoff and bandpass systems.

In addition, the applications of the principles developed here are extended to give a good deal of flexibility in the design of transmission networks. In particular it is shown that distortionless transmission can be achieved under conditions of nonlinear phase and nonsymmetrical roll-off in amplitude, provided the proper relationships between these two quantities exist.

Fig. 1 illustrates the general baseband system to be examined. We can

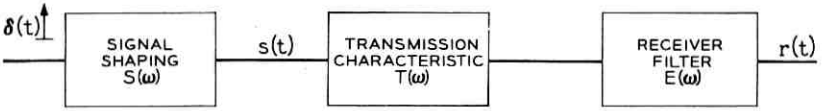

Fig. 1 — General digital transmission system.

assume, without loss of generality, that the information is contained in a random sequence of impulses at the input to the system. Thus a signal $s(t)$ having an amplitude of 0 or 1 is transmitted every T seconds. The system output, $r(t)$, with the Fourier transform

$$R(\omega) = S(\omega)T(\omega)E(\omega) \tag{1}$$

is used to decide whether $s(t)$ was transmitted with amplitude 0 or 1 at a particular time. The type of decision criterion used determines the constraints on $R(\omega)$. Decisions based upon pulse amplitude (usual PCM) at a fixed time and pulse width (telegraph) will be considered.

A sequence of input signals will, in general, produce a sequence of overlapping output pulses. To prevent intersymbol distortion at the output, either the pulse amplitude or the pulse width must be unaffected by the tails of adjacent signals. Fig. 2 illustrates the types of waveform which possess these characteristics. It should be noted that both waveforms require periodic zero crossings away from the main peak. These constraints on the time domain signals are translated into constraints in the frequency domain.*

* These constraints on channels are based on the preservation of periodic zero crossings in the output response. In the case where information is contained in the amplitude of a binary signal, this concept is straightforward. Complications arise, however, when information is associated with the pulse width (such as certain

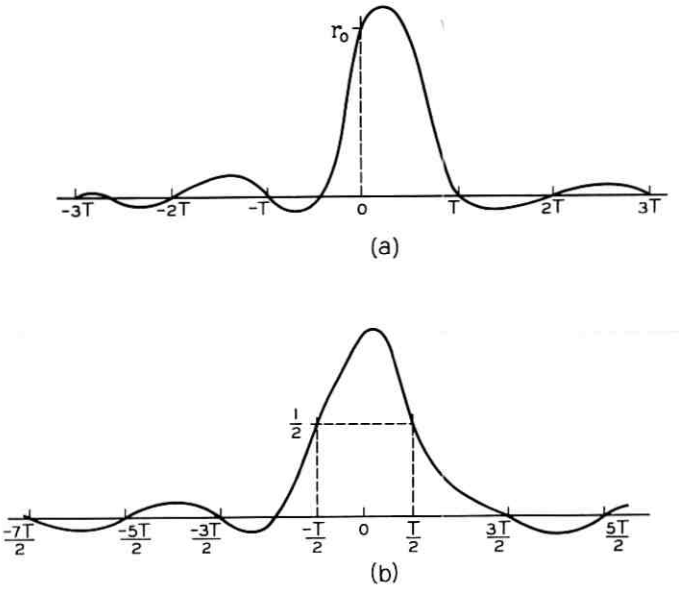

Fig. 2 — Undistorted system responses, (a) $r(t)$ with pulse amplitude undistorted by adjacent pulses; (b) $r(t)$ with pulse width undistorted by adjacent pulses.

II. SYSTEM CONSTRAINTS — UNDISTORTED AMPLITUDE TRANSMISSION

In this section decisions based upon pulse amplitude will be considered. From Fig. 2(a) the constraints on the output pulse may be written

$$r(kT) = r_k = r_0\delta_{k0}\,. \tag{2}$$

These sample values may be written in terms of the Fourier transform

$$r(t) = \int_{-\infty}^{\infty} R(\omega)\, e^{j\omega t}\, d\omega \tag{3}$$

$$r_k = \int_{-\infty}^{\infty} R(\omega)\, e^{j\omega kT}\, d\omega \tag{4a}$$

$$= \sum_{n=-\infty}^{\infty} \int_{\frac{\pi}{T}(2n-1)}^{\frac{\pi}{T}(2n+1)} R(\omega)\, e^{j\omega kT}\, d\omega \tag{4b}$$

types of telegraph transmission and systems involving timing recovery). In such cases there may occur troublesome excursions of the signal in between those points which are preserved by the constraints. Unless special apparatus is used in the detection (or timing recovery) process errors will result. The analysis of this problem, which is inherent in the original Nyquist work as well as in the present study, is very complicated and beyond the scope of this paper.

$$= \sum_{n=-\infty}^{\infty} \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} R\left(u + \frac{2n\pi}{T}\right) e^{jukT}\, du. \tag{5a}$$

Assuming that $\sum_n R[u + (2n\pi/T)]e^{jukT}$ is a uniformly convergent series, one obtains

$$r_k = \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \sum_{n=-\infty}^{\infty} R\left(u + \frac{2n\pi}{T}\right) e^{jukT}\, du. \tag{5b}$$

Notice that r_k is just the kth coefficient of an exponential Fourier series expansion of

$$\frac{2\pi}{T} \sum_{n=-\infty}^{\infty} R\left(u + \frac{2n\pi}{T}\right) \qquad -\frac{\pi}{T} \leqq u \leqq \frac{\pi}{T}.$$

The requirement that $r_k = r_0\delta_{k0}$ implies that only the *zero*th coefficient of the expansion of

$$\frac{2\pi}{T} \sum_{n=-\infty}^{\infty} R\left(u + \frac{2n\pi}{T}\right)$$

is not zero, and hence

$$\frac{2\pi}{T} \sum_{n=-\infty}^{\infty} R\left(u + \frac{2n\pi}{T}\right) = r_0. \tag{6}$$

By using the amplitude and phase characteristics

$$R(\omega) = A(\omega)e^{j\alpha(\omega)} \tag{7}$$

one gets

$$\sum_{n=-\infty}^{\infty} A\left(u + \frac{2n\pi}{T}\right) \exp\left[j\alpha\left(u + \frac{2n\pi}{T}\right)\right] = \frac{r_0 T}{2\pi}. \tag{8}$$

Separating (8) into real and imaginary parts one obtains

$$\sum_{n=-\infty}^{\infty} A\left(u + \frac{2n\pi}{T}\right) \cos\alpha\left(u + \frac{2n\pi}{T}\right) = \frac{r_0 T}{2\pi} \tag{9a}$$

and

$$\sum_{n=-\infty}^{\infty} A\left(u + \frac{2n\pi}{T}\right) \sin\alpha\left(u + \frac{2n\pi}{T}\right) = 0 \tag{9b}$$

for $-\pi/T \leqq u \leqq \pi/T$. Because of symmetry conditions [$A(\omega) = A(-\omega)$, $\alpha(\omega) = -\alpha(-\omega)$] the interval $0 \leqq u \leqq \pi/T$ is sufficient.

Fig. 3 illustrates the constraints for a characteristic that is limited to

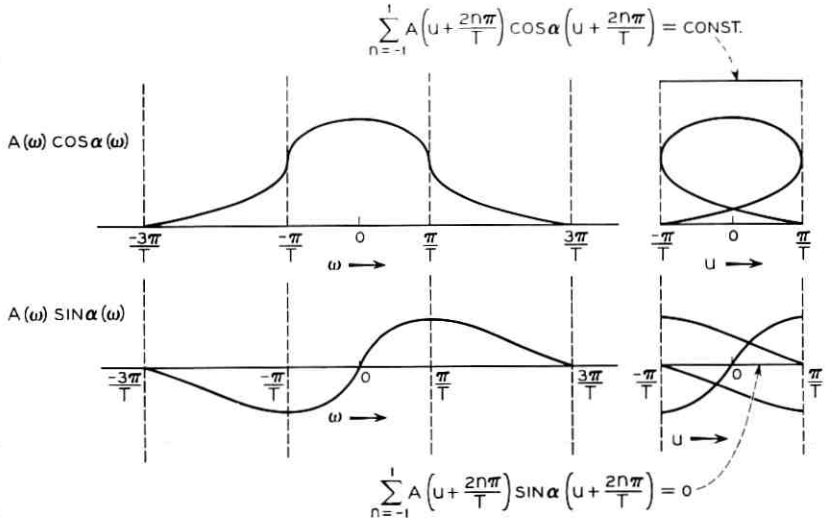

Fig. 3 — Constraints for no intersymbol amplitude distortion [$A(\omega) = 0 \mid \omega \mid \geqq 3\pi/T$].

$\mid \omega \mid < 3\pi/T$. There is, however, no reason for this limitation other than for clarity in the diagram. The only restriction on the frequency characteristic is an asymptotic one. The condition that $\sum_n R[u + (2n\pi/T)] \cdot e^{jukT}$ be a uniformly convergent series is satisfied if $A(\omega) \to 1/\omega^q$, $q \geqq 2$, as $\omega \to \infty$. This is a more realistic restriction than forcing $A(\omega) = 0$ for large ω.

One may also note that the constraints are more general than Nyquist's symmetry conditions because of the elimination of the cutoff requirements. These symmetry conditions may be obtained by limiting $A(\omega)$ to the region $-2\pi/T < \omega < 2\pi/T$. From Fig. 4 it is easily seen that

$$A(u)\cos\alpha(u) + A[u - (2\pi/T)]\cos\alpha[u - (2\pi/T)] = \text{Const.} \quad (10a)$$

and

$$A(u)\sin\alpha(u) + A[u - (2\pi/T)]\sin\alpha[u - (2\pi/T)] = 0 \quad \text{for } 0 \leqq u \leqq \pi/T \quad (10b)$$

which are Nyquist's conditions.

Consider, now, (9a) and (9b) and their ramifications. No longer is one confined to low-pass sharp cutoff systems. It is now possible to

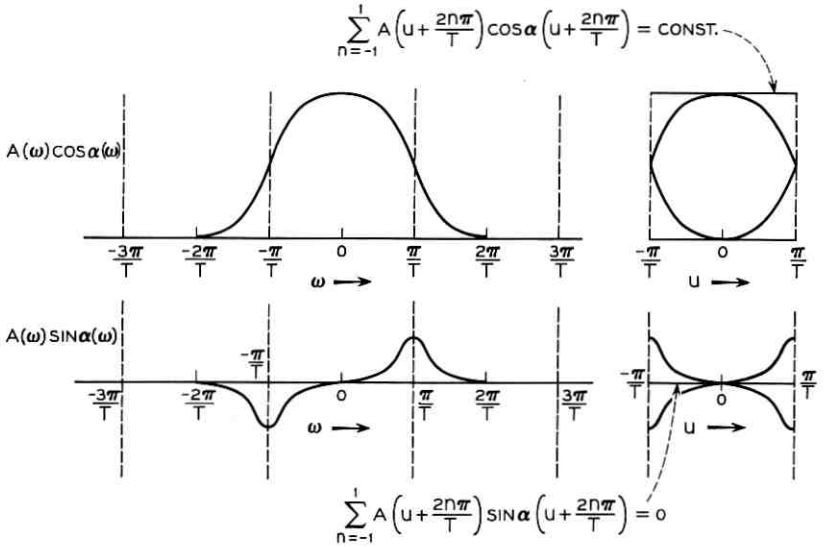

Fig. 4 — Constraints for no intersymbol amplitude distortion [$A(\omega) = 0 \mid \omega \mid \geqq 2\pi/T$].

express compactly the conditions for distortionless transmission for bandpass or gradual cutoff systems as well. Fig. 3 shows a gradual cutoff system and Fig. 5 illustrates an acceptable bandpass characteristic.

Note that (9a) and (9b) represent constraints on the real and imaginary parts of the characteristics and not upon the amplitude and phase. In general, these equations imply nothing about conditions on the amplitude and phase individually (the exception being the bandlimited case [$A(\omega) = 0, \mid \omega \mid > \pi/T$] where $A(\omega) = K$ and $\alpha(\omega) = 0$ are the conditions). Constraints on $A(\omega)$ are imposed only if $\alpha(\omega)$ is arbitrarily chosen or vice versa. The usual application of Nyquist's results (linear phase and symmetric roll-off) is just such an arbitrary choice.

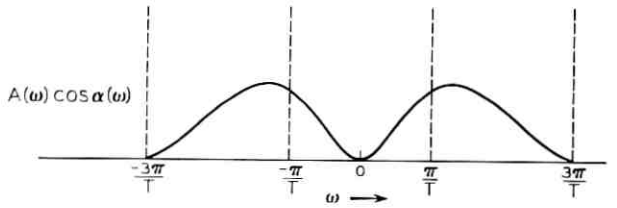

Fig. 5 — Distortionless bandpass characteristic.

teristics may be such that the required $A(\omega)$ in one of the intervals may approach infinity (if $\alpha_j(u) = \alpha_i(u) + n\pi$). With phase equalization, for $\alpha_i(u)$ or $\alpha_j(u)$ to be real phase angles it is necessary that

$$\{A_i(u) + A_j(u)\}^2 \geqq F^2(u) + G^2(u) \geqq \{A_i(u) - A_j(u)\}^2 \qquad -(\pi/T) \leqq u \leqq \pi/T. \tag{15}$$

This condition determines the intervals, if any, in which phase equalization may be applied. It may happen that, because of a poor choice of transmission speed or poor characteristics outside the i and j intervals, this type of equalization cannot be used. In most practical cases, however, the transmission rate can be judiciously chosen, and phase equalization is theoretically possible. It might be pointed out that (15) or its generalization (where phase equalization is allowed over the entire spectrum) can be used to determine the maximum rate for a fixed amplitude characteristic. The application and some ramifications of (15) are illustrated in Appendix A for the Nyquist problem of (10).

As a specific example of some of the concepts outlined, consider the usual Nyquist problem ($A(\omega) = 0$ for $\omega > 2\pi/T$) given in (10a) and (10b). One can obtain the constraints on either $A(\omega)$ or $\alpha(\omega)$ by letting

$$F(u) = K \tag{16a}$$

$$G(u) = 0 \tag{16b}$$

$$A_i(u) = A(u), \qquad \alpha_i(u) = \alpha(u) \tag{16c}$$

$$A_j(u) = A[u - (2\pi/T)], \qquad \alpha_j(u) = \alpha[u - (2\pi/T)] \tag{16d}$$

in (14a–d). The resulting equations become

$$A(u) = \frac{K \sin \alpha\left(u - \frac{2\pi}{T}\right)}{\sin\left[\alpha\left(u - \frac{2\pi}{T}\right) - \alpha(u)\right]} \tag{17a}$$

$$A\left(u - \frac{2\pi}{T}\right) = \frac{-K \sin \alpha(u)}{\sin\left[\alpha\left(u - \frac{2\pi}{T}\right) - \alpha(u)\right]}, \tag{17b}$$

$$\alpha(u) = \cos^{-1} \frac{K^2 + A^2(u) - A^2\left(u - \frac{2\pi}{T}\right)}{2KA(u)} \tag{17c}$$

and

$$\alpha\left(u - \frac{2\pi}{T}\right) = \cos^{-1} \frac{K^2 + A^2\left(u - \frac{2\pi}{T}\right) - A^2(u)}{2KA\left(u - \frac{2\pi}{T}\right)} \qquad (17d)$$

$$\text{for } 0 \leqq u \leqq \frac{\pi}{T}.$$

Equations (17a–d) form a relationship which must be satisfied for ideal transmission. In general, $\alpha(\omega)$ need not be linear and $A(\omega)$ need not have the usual symmetrical roll-off. All that is required is that the phase and amplitude satisfy the equations.

For an unequalized channel, with known $A(\omega)$ and $\alpha(\omega)$, this can be accomplished by either leaving the phase unchanged and computing the matching amplitude from (17a–b) or by leaving the amplitude unchanged and computing the matching phase by (17c–d). It is apparent that this gives a good deal more freedom and flexibility to one confronted with the task of equalizing a channel. Some examples of the use of equations will now be considered.

2.1 *Examples*

The amplitude characteristic $A(\omega)$ of a channel with some kind of resonant peaking is shown in Fig. 6(a) together with the minimum phase characteristic associated with $A(\omega)$. Since these channel characteristics do not satisfy ideal transmission conditions, the impulse response of the channel will be distorted. This is indicated in Fig. 6(b) in which the zero crossings of the response do not coincide with the sampling points. As stated before, there are several ways of equalizing the channel. Phase equalization may be achieved by substituting the value of $A(\omega)$ into (17c–d) and obtaining the matching phase. This is shown in Fig. 7(a) (with the original minimum phase shown dashed for comparison). The resulting impulse response, shown in Fig. 7(b), is seen to have zero crossings which are properly spaced, thus satisfying the condition for undistorted transmission.

It can be seen that equalizing a channel by means of (17c,d) offers considerable reduction in complexity over the method which requires a flat delay and symmetrically shaped amplitude. For example, in the above illustration it was necessary to alter only one of the characteristics instead of both. It was not required that the phase be linear but only that its shape be altered in a prescribed manner. An important practical factor stems from the fact that the delay for the equalized channel is not

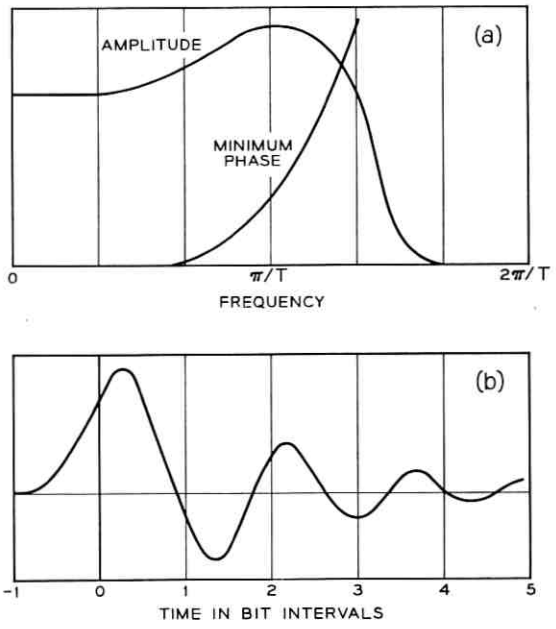

Fig. 6 — Initial system response, (a) transmission frequency characteristics; (b) impulse response.

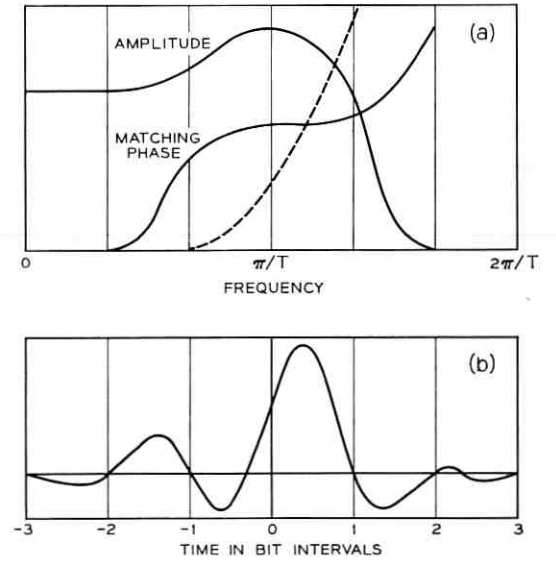

Fig. 7 — System response with phase correction, (a) transmission frequency characteristics; (b) impulse response.

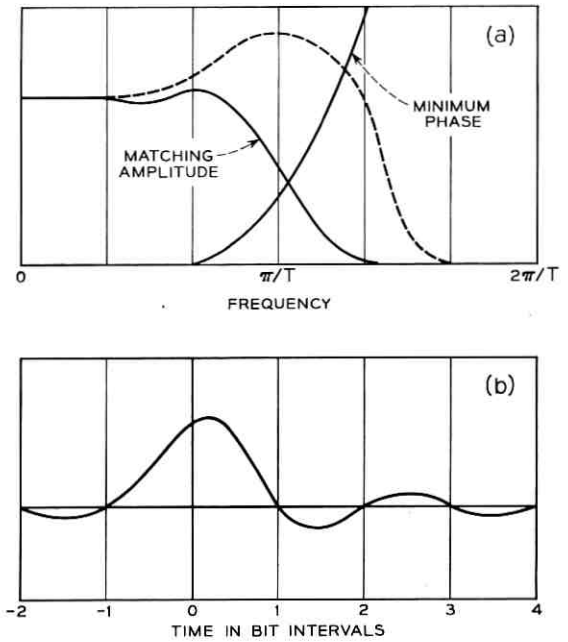

Fig. 8 — System response with amplitude correction, (a) transmission frequency characteristics; (b) impulse response.

flat. While it is usually thought desirable to have a channel with a flat delay, it is apparent that in this case linear phase across the band would degrade rather than improve transmission.

A second method of equalizing the channel of Fig. 6 is obtained when the equalized amplitude characteristic is obtained from the original minimum phase by (17a–b). The resulting $A(\omega)$ is shown in Fig. 8(a) together with the impulse response for the equalized channel in Fig. 8(b).

III. SYSTEM CONSTRAINTS — PULSE WIDTH UNDISTORTED

If the pulse width is to be undistorted by adjacent pulses, $r(t)$ must satisfy the conditions

$$\begin{aligned} r_k &= r\left(\frac{2k-1}{2}\,T\right) = 0 \qquad k \neq 0, 1 \\ r_0 &= r_1 = \frac{1}{2}. \end{aligned} \tag{18}$$

Again, writing these sample values in terms of the Fourier transform, one obtains (3)

$$r(t) = \int_{-\infty}^{\infty} R(\omega) e^{j\omega t}\, d\omega \tag{19a}$$

$$r_k = \int_{-\infty}^{\infty} R(\omega) e^{-j\omega(T/2)}\, e^{j\omega kT}\, d\omega$$

$$= \sum_{n=-\infty}^{\infty} \int_{\frac{\pi}{T}(2n-1)}^{\frac{\pi}{T}(2n+1)} R(\omega) e^{-j\omega(T/2)}\, e^{j\omega kT}\, dw \tag{19b}$$

$$= \sum_{n=-\infty}^{\infty} \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} R\left(u + \frac{2n\pi}{T}\right) \exp\left[-j\left(u + \frac{2n\pi}{T}\right)\frac{T}{2}\right] e^{jkTu}\, du. \tag{20a}$$

Assuming that $\sum_n R[u + (2n\pi/T)]e^{-jn\pi}e^{-ju(T/2)}e^{jukT}$ is a uniformly convergent series one obtains

$$r_k = \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \left\{ \sum_{n=-\infty}^{\infty} R\left(u + \frac{2n\pi}{T}\right) e^{-jn\pi} \right\} e^{-ju(T/2)}\, e^{jukT}\, du \tag{20b}$$

$$= \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \left\{ \sum_{n=-\infty}^{\infty} (-1)^n R\left(u + \frac{2n\pi}{T}\right) \right\} e^{-ju(T/2)}\, e^{jukT}\, du. \tag{20c}$$

The value of r_k is the kth coefficient of an exponential Fourier series expansion of

$$(2\pi/T) \sum_{n=-\infty}^{\infty} (-1)^n R[u + (2n\pi/T)]e^{-ju(T/2)}.$$

From (20c) it is seen that the expansion is

$$\sum_{n=-\infty}^{\infty} (-1)^n R[u + (2n\pi/T)]e^{-ju(T/2)} = (T/2\pi) \sum_k r_k e^{-jukT}. \tag{21}$$

Letting

$$G_R(u) + jG_I(u) = \left\{ \sum_{n=-\infty}^{\infty} (-1)^n R[u + (2n\pi/T)] \right\} e^{-ju(T/2)} \tag{22}$$

and using the conditions $r_0 = r_1 = \frac{1}{2}$ and $r_k = 0$, $k \neq 0,1$ one gets

$$G_R(u) + jG_I(u) = (T/2\pi)[\tfrac{1}{2} + \tfrac{1}{2}e^{-juT}]. \tag{23}$$

Separating the real and imaginary parts of the equation yields

$$G_R(u) = (T/4\pi)(1 + \cos uT) \tag{24a}$$

and

$$G_I(u) = (T/4\pi)(-\sin uT) \qquad \text{for } -(\pi/T) \leqq u \leqq \pi/T. \tag{24b}$$

Letting

$$R_R(u) = \text{Re}\{\sum_n (-1)^n R[u + (2n\pi/T)]\} \tag{25a}$$

and

$$R_I(u) = \text{Im}\{\sum_n (-1)^n R[u + (2n\pi/T)]\} \qquad \text{for } -(\pi/T) \leqq u \leqq \pi/T \tag{25b}$$

one gets

$$\begin{aligned} G_R(u) &= R_R(u)\cos(uT/2) + R_I(u)\sin(uT/2) \\ &= (T/4\pi)(1 + \cos uT) \end{aligned} \tag{26a}$$

and

$$\begin{aligned} -G_I(u) &= R_R(u)\sin(uT/2) - R_I(u)\cos(uT/2) \\ &= (T/4\pi)\sin uT \qquad \text{for } -(\pi/T) \leqq u \leqq \pi/T. \end{aligned} \tag{26b}$$

Solving these two equations, the constraints for no intersymbol interference become

$$R_R(u) = \text{Re}\{\sum_n (-1)^n R[u + (2n\pi/T)]\} = (T/2\pi)\cos(uT/2) \tag{27a}$$

and

$$R_I(u) = \text{Im}\{\sum_n (-1)^n R[u + (2n\pi/T)]\} = 0 \qquad \text{for } -(\pi/T) \leqq u \leqq \pi/T. \tag{27b}$$

Finally, writing

$$R(\omega) = A(\omega)e^{j\alpha(\omega)}$$

one obtains

$$\sum_n (-1)^n A[u + (2n\pi/T)]\cos\alpha[u + (2n\pi/T)] = (T/2\pi)\cos(uT/2) \tag{28a}$$

and

$$\sum_n (-1)^n A[u + (2n\pi/T)]\sin\alpha[u + (2n\pi/T)] = 0 \qquad \text{for } -(\pi/T) \leqq u \leqq \pi/T. \tag{28b}$$

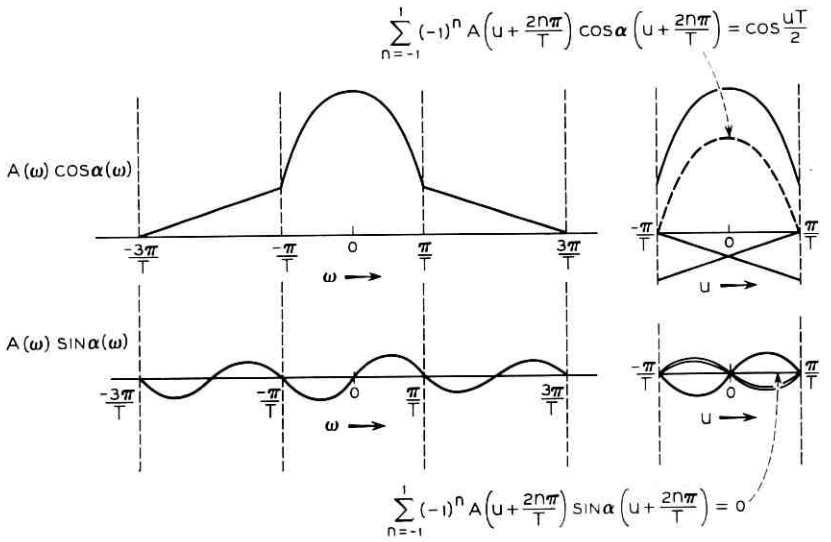

Fig. 9 — Constraints for no intersymbol pulse width distortion [$A(\omega) = 0 \mid \omega \mid \geqq 3\pi/T$].

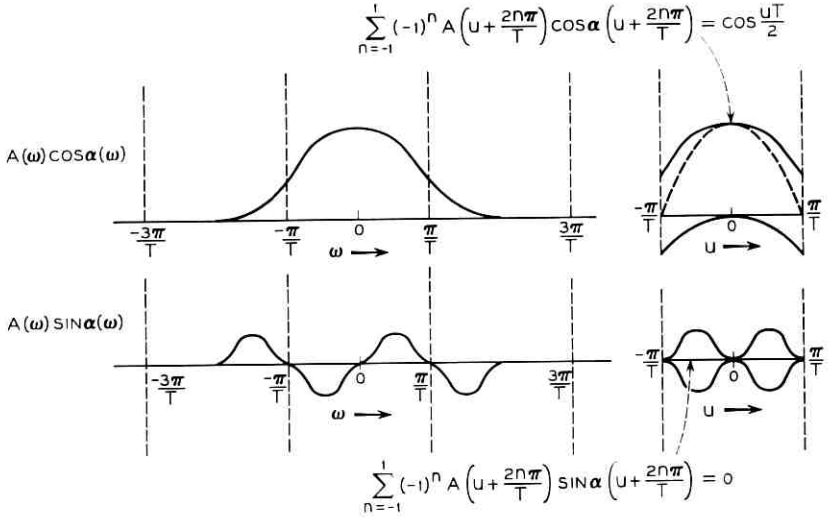

Fig. 10 — Constraints for no intersymbol pulse width distortion [$A(\omega) = 0 \mid \omega \mid \geqq 2\pi/T$].

Again, there is only an asymptotic bandwidth restriction on these constraints. Fig. 9 illustrates a satisfactory characteristic with $A(\omega) = 0, |\omega| \geqq 3\pi/T$ for clarity. With $A(\omega) = 0, |\omega| \geqq 2\pi/T$ the conditions become the familiar Nyquist results shown in Fig. 10.

The general statements of Section II about the implications of (9a–b) can be applied here to (28a–b). The specific results of Section II can be obtained by replacing

$$A[u + (2n\pi/T)] \text{ by } (-1)^n A[u + (2nT/\pi)]$$

and K by $K \cos(uT/2)$. For the specific case of the usual bandwidth limitation $[A(\omega) = 0, |\omega| \geqq 2\pi/T]$ one gets from (17a–d) the constraints

$$A(u) = \frac{K \cos \frac{uT}{2} \sin \alpha\left(u - \frac{2\pi}{T}\right)}{\sin\left[\alpha\left(u - \frac{2\pi}{T}\right) - \alpha(u)\right]}, \tag{29a}$$

$$A\left(u - \frac{2\pi}{T}\right) = \frac{K \cos \frac{uT}{2} \sin \alpha(u)}{\sin\left[\alpha\left(u - \frac{2\pi}{T}\right) - \alpha(u)\right]}, \tag{29b}$$

$$\alpha(u) = \cos^{-1}\left[\frac{K^2 \cos^2 \frac{uT}{2} + A^2(u) - A^2\left(u - \frac{2\pi}{T}\right)}{2KA(u) \cos \frac{uT}{2}}\right] \tag{29c}$$

and

$$\alpha\left(u - \frac{2\pi}{T}\right) = \cos^{-1}\left[\frac{K^2 \cos^2 \frac{uT}{2} + A^2\left(u - \frac{2\pi}{T}\right) - A^2(u)}{-2KA\left(u - \frac{2\pi}{T}\right) \cos \frac{uT}{2}}\right] \tag{29d}$$

$$0 \leqq u \leqq \frac{\pi}{T}.$$

IV. CONCLUDING REMARKS

This paper has extended Nyquist's work on transmission theory to eliminate bandwidth restrictions. The extension is important for a full understanding of data systems. In the past, incomplete results have been obtained from the imposition of arbitrary band limitations. For example, one paper[2] stated that only one waveform jointly satisfies the two criteria discussed here (pulse height and pulse width preservation). In Appendix B this is shown to be false in general but true if $A(\omega) = 0, |\omega| \geqq 2\pi/T$.

Although distortionless transmission has been the main consideration, it is possible, with the approach used in the paper, to obtain an estimate of system quality when the conditions of ideal transmission are not met. In Appendix C, a measure of the distortion (for systems which base decisions on pulse amplitude) is derived in terms of the frequency domain characteristics.

The discussion makes clear that the constraints are not obtained on the phase and amplitude characteristics individually, but only the real and imaginary parts of the transfer characteristics. Specific constraints on the amplitude and phase are the result of arbitrary design choices. Equalization requirements are thus less stringent than usually assumed. It is seen that equalization is only necessary over intervals of π/T or $2\pi/T$ (subject to the conditions discussed) and not over the entire band. Further, it may only be necessary to compensate either the amplitude or the phase but not both.

APPENDIX A

Realizability Conditions for Phase Equalization

In Section II, the question of equalizer realizability was briefly considered. This question is closely related to the choice of transmission rate which is of sufficient importance to discuss further at this point. Thus, it is possible to illustrate the realizability conditions for phase equalization by considering a transmission system with variable phase equalizer and determining the maximum signaling speed. By assuming that the system has a continuous sharp cutoff amplitude characteristic [$A(\omega) = 0, \omega \geqq \omega_c$] and that it is desirable that the signaling speed ($\omega_s = \pi/T$) be

$$\omega_c/2 \leqq \pi/T \leqq \omega_c \tag{30}$$

one has the usual Nyquist problem. Under these assumptions, the conditions for phase equalization (15) become

$$\{A(u) + A[u - (2\pi/T)]\}^2 \geqq K^2 \geqq \{A(u) - A[u - (2\pi/T)]\}^2 \tag{31a}$$

or

$$\begin{aligned} A(u) + A[u - (2\pi/T)] \\ \geqq K \geqq A(u) - A[u - (2\pi/T)] \geqq -K \end{aligned} \tag{31b}$$

or

$$A_+(u) \geqq K \geqq A_-(u) \geqq -K \text{ for } 0 \leqq u \leqq \pi/T. \tag{31c}$$

The more general condition would be used if the above assumptions are removed, but this example illustrates the concepts adequately. By using condition (30) and the fact that

$$A(\omega) = 0, \qquad \omega \geqq \omega_c$$

one obtains

$$\begin{aligned} A(u) + A[u - (2\pi/T)] \geqq A(0) \geqq A(u) \\ - A[u - (2\pi/T)] \geqq -A(0) \end{aligned} \tag{32}$$

and

$$A(\pi/T) \geqq \tfrac{1}{2}A(0), \tag{33}$$

and these must be satisfied for phase equalization.

By examining the amplitude characteristics graphically, it is easier to study some of the other implications of the equations. As an example, consider the problem of finding the maximum signaling speed for the amplitude characteristic shown in Fig. 11 (a). From the previous results, it is known that the maximum speed lies between $\omega_c/2$ and $z(A(z) = A(0)/2)$. Fig. 11 (b) shows $A_+(u)$ and $A_-(u)$ for

$$\omega_c/2 < \pi/T < z$$

and Fig. 11 (c) shows the same curves for

$$\pi/T = \omega_c/2.$$

Notice that $A_+(u) \not> A(0)$ for all u in Fig. 11 (b), and phase equalization cannot yield distortionless transmission. For $\pi/T = \omega_c/2$ the network can be phase equalized. Notice also that $\omega_c/2$ is the maximum signaling speed for distortionless transmission with phase equalization. In other words, any amplitude characteristic which is strictly decreasing $[A(\omega + \delta) < A(\omega)]$ cannot have undistorted signaling above $\omega_c/2$. Because $A'(0^+) \neq 0$, a slightly higher signaling speed would mean that both $A_+(u)$ and $A_-(u)$ would be identical and have a slope different from zero at $u \approx 0$. This situation would not satisfy (32).

The above observation can be generalized by noting that the upper signaling speed is limited by $\frac{1}{2}(\omega_c + \omega_l)$ where ω_l = lowest frequency at which $A'(\omega) \neq 0$. Fig. 12 illustrates this feature for a peaked amplitude response. Here

$$\pi/T = \tfrac{1}{2}(\omega_c + \omega_l)$$

and a slightly higher speed would again mean that $A_+(u) = A_-(u)$ at some point with $A_+'(u) = A_-'(u) \neq 0$.

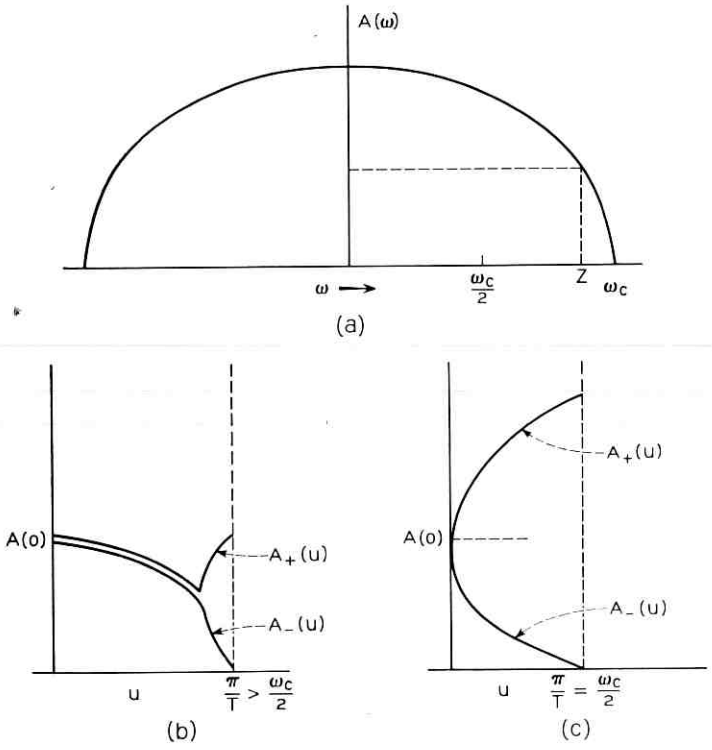

Fig. 11 — Maximum signaling speed when $A'(0^+) \neq 0$.

To show that $\frac{1}{2}(\omega_c + \omega_l)$ is only a limit and not the true maximum speed, consider the example in Fig. 13. It is apparent that the frequency $\frac{1}{2}(\omega_c + \omega_l)$ is too high and thus the true maximum is ω_l.

It is difficult to sum up in words all of the considerations in deciding whether equalization is possible or, equivalently, what is the highest signaling speed at which it is possible. Equation (32) contains all of the required information, and this section was intended to give some idea of its use.

APPENDIX B

Combination of the Two Cases

In the previous analysis, two types of undistorted transmission were treated independently. It will now be determined under what conditions these two cases can be realized simultaneously. The equations which

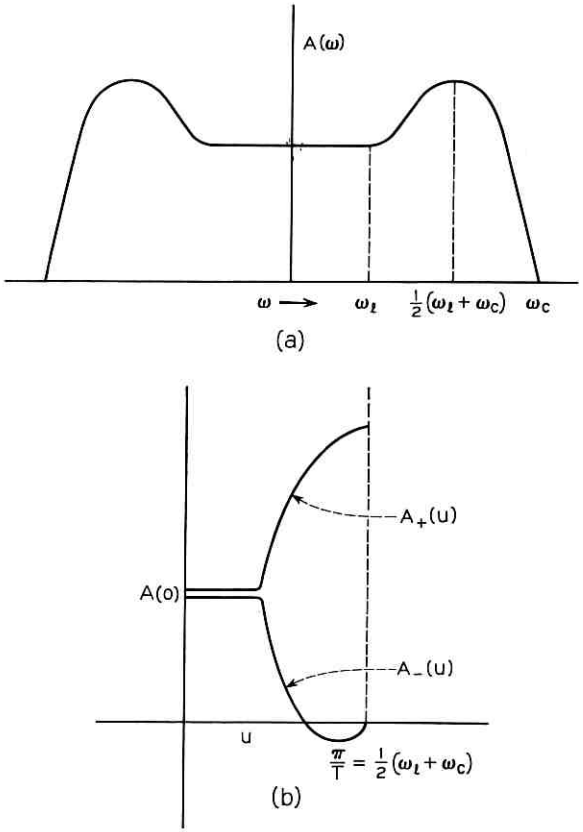

Fig. 12 — Maximum signaling speed for peaked amplitude response.

must be satisfied are:

equation (9a)

$$\sum_n A[u + (2n\pi/T)] \cos \alpha[u + (2n\pi/T)] = K,$$

equation (9b)

$$\sum_n A[u + (2n\pi/T)] \sin \alpha[u + (2n\pi/T)] = 0,$$

equation (28a)

$$\sum_n (-1)^n A[u + (2n\pi/T)] \cos \alpha[u + (2n\pi/T)] = K \cos (uT/2),$$

and equation (28b)

$$\sum_n (-1)^n A[u + (2n\pi/T)] \sin \alpha[u + (2n\pi/T)] = 0$$

$$\text{for } -(\pi/T) \leqq u \leqq \pi/T.$$

The simultaneous solutions to these equations are

$$\sum_{n \text{ odd}} A[u + (2n\pi/T)] \sin \alpha[u + (2n\pi/T)] = 0, \tag{34}$$

$$\sum_{n \text{ even}} A[u + (2n\pi/T)] \sin \alpha[u + (2n\pi/T)] = 0, \tag{35}$$

$$\sum_{n \text{ even}} A[u + (2n\pi/T)] \cos \alpha[u + (2n\pi/T)] = \tfrac{1}{2}K\{1 + \cos (uT/2)\} \tag{36}$$

and

$$\sum_{n \text{ odd}} A[u + (2n\pi/T)] \cos \alpha[u + (2n\pi/T)] = \tfrac{1}{2}K\{1 - \cos (uT/2)\} \quad \text{for } -(\pi/T) \leqq u \leqq \pi/T. \tag{37}$$

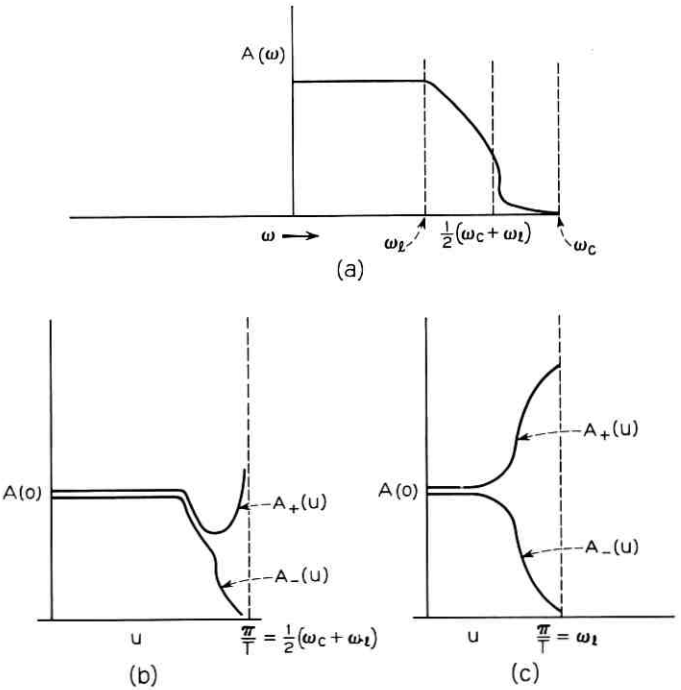

Fig. 13 — Maximum signaling speed equal to ω_l.

In general there will be many possible solutions to these four equations. For the particular case $A(\omega) = 0, |\omega| \geqq 2\pi/T$ each of the above summations reduces to one term and

$$\alpha(u) = 0, \tag{38}$$

$$\alpha[u - (2\pi/T)] = 0, \tag{39}$$

$$A(u) = (K/2)\{1 + \cos(uT/2)\}, \tag{40}$$

and

$$A[u - (2\pi/T)] = (K/2)\{1 - \cos(uT/2)\} \qquad \text{for } 0 \leqq u \leqq \pi/T. \tag{41}$$

Taken together, (40) and (41) define the amplitude characteristic across the band as the familiar[2] full cosine roll-off, which may be written by a single expression

$$A(\omega) = (K/2) + (K/2)\cos(\omega T/2) \qquad -(2\pi/T) \leqq \omega \leqq (2\pi/T). \tag{42}$$

This amplitude characteristic is shown in Fig. 14(a). The corresponding

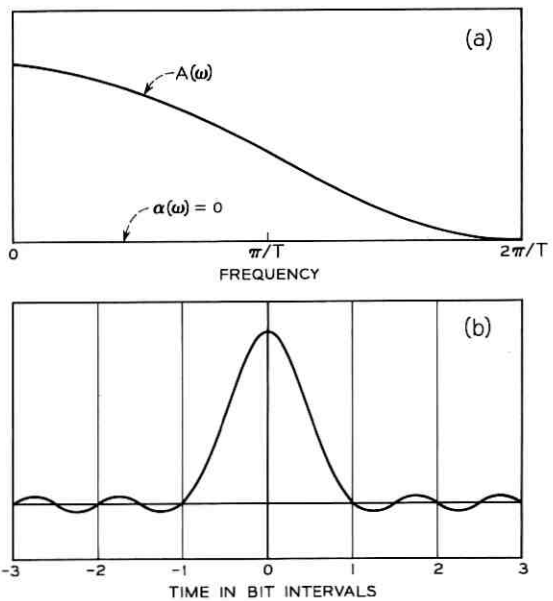

Fig. 14 — System response satisfying both critria, (a) transmission frequency characteristics; (b) impulse response.

impulse response in Fig. 14(b) satisfies both types of undistorted transmission, as expected.

APPENDIX C

A Distortion Measure

It is possible to use the results of the paper to obtain an estimate of system quality when the conditions of ideal transmission are not met. The variance of the intersymbol distortion distribution.

$$\sum_{k\neq 0} r_k^2$$

can be shown to provide an indication of transmission quality (for undistorted amplitude transmission). Since

$$r_k = \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \sum_n R\left(u + \frac{2n\pi}{T}\right) e^{jukT} du \quad \text{(equation 5b)}$$

one could write

$$\sum_n R\left(u + \frac{2n\pi}{T}\right) = \frac{T}{2\pi} \sum_k r_k\, e^{-jukT} \tag{43a}$$

or

$$\sum_n R\left(u + \frac{2n\pi}{T}\right) - \frac{r_0 T}{2\pi} = \frac{T}{2\pi} \sum_{k\neq 0} r_k\, e^{-jukT} \tag{43b}$$

and

$$\left[\sum_n R\left(u + \frac{2n\pi}{T}\right) - \frac{r_0 T}{2\pi}\right]^* = \frac{T}{2\pi} \sum_{k\neq 0} r_k\, e^{jukT} \qquad \text{for } -\frac{\pi}{T} \leqq u \leqq \frac{\pi}{T}. \tag{43c}$$

Multiplying (43b) and (43c) and integrating, one obtains

$$\frac{T}{2\pi} \sum_{k\neq 0} r_k^2 = \int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \left[\sum_n R\left(u + \frac{2n\pi}{T}\right) - \frac{r_0 T}{2\pi}\right]^* \cdot \left[\sum_n R\left(u + \frac{2n\pi}{T}\right) - \frac{r_0 T}{2\pi}\right] du \tag{44a}$$

or

$$\sum_{k\neq 0} r_k^2 = \frac{2\pi}{T}\int_{-\frac{\pi}{T}}^{\frac{\pi}{T}} \left\{\left[\sum_n A\left(u+\frac{2n\pi}{T}\right)\right.\right.$$
$$\cos\ \alpha\left(u+\frac{2n\pi}{T}\right) - \frac{r_0 T}{2\pi}\Bigg]^2 + \left[\sum_n A\left(u+\frac{2n\pi}{T}\right)\right. \qquad (44b)$$
$$\left.\left.\cdot \sin\alpha\left(u+\frac{2n\pi}{T}\right)\right]^2\right\} du \qquad \text{for } -\frac{\pi}{T} \leqq u \leqq \frac{\pi}{T}.$$

REFERENCES

1. Nyquist, H., Certain Topics in Telegraph Transmission Theory, AIEE Trans., *47*, April, 1928, pp. 617–644.
2. Brand, S., and Carter, C. W., A 1,650-Bit-Per-Second Data System for Use over the Switched Telephone Network, AIEE Trans. Pt. I (Communication and Electronics), *80*, 1962, pp. 652–661.

A Stable, Single-Frequency RF-Excited Gas Laser at 6328Å

By J. A. COLLINSON

(Manuscript received April 27, 1965)

A number of 6328Å RF-excited He-Ne lasers have been designed and constructed with special attention to maximizing CW power in a single longitudinal and transverse mode. A single-mode output power of 1.6 mw has been obtained. Some novel features of the cavity structure provide good intrinsic frequency stability. The details of design and operation are given.

I. INTRODUCTION

One of the reasons for the interest in lasers is their potentially very large information bandwidth. Realization of this potential requires, among other things, lasers which oscillate in a single frequency of much improved stability. Since a typical cavity dimension is thousands of wavelengths, lasers are inherently multimode devices. For this same reason, the cavity Q is orders of magnitude greater than the Q for a particular transition, and the laser frequency is determined principally by cavity dimensions. As a result, the precise frequency of a laser oscillator is extremely sensitive to cavity microphonics.

Several techniques have been used to obtain single-mode operation. Javan et al[1] obtained a single axial mode in the first gas laser by reducing gain per pass and maintaining it barely above threshold. Then, of the large number of allowed axial modes, oscillation occurred in only that mode closest to the maximum of the atomic transition. As Javan pointed out, this required very fine control of the excitation to prevent variations in gain. Three-mirror cavities, proposed by Kleinman and Kisliuk[2] and executed by Patel and Koglenik,[3] provide one axial mode even at excitation levels well above threshold. The third mirror varies the frequency dependence of cavity losses and thus aids in the discrimination against all but one mode. However, the addition of a mirror compounds the already serious problem of stabilization of the elementary two-mirror cavity. A sophisticated method for obtaining a single frequency from a

laser in which there is a large number of allowed axial modes was described by Massey et al.[4] A phase modulator within the cavity was operated to give an array of axial modes having the same amplitudes and phases as the sidebands of an FM signal. The output beam was then demodulated, giving a single frequency. They explained that, since this applied to the entire output from a high-power, multimode laser, the technique did not suffer the power loss inherent in other approaches. They obtained an output of 0.1 mw. Once again, however, the addition of components makes the stabilization problem more difficult.

The direct method is to build a two-mirror cavity of such geometry that only one mode exists within the spectral width of the gain even at saturated operation of the laser. This was done first by Gordon and White,[5] using a dc-excited He-Ne 6328Å tube in a cavity about 4 inches long. The gain per pass and axial mode separation in this laser are such that one axial mode at most can oscillate. When the laser containing a single neon isotope is tuned so that the oscillating mode is near the atomic line center, maximum power is reached in that mode. When the cavity length is then detuned by a quarter wavelength, oscillation stops. The present article describes a similar short, single-frequency laser which, however, is RF-excited. The design also incorporates some novel, frequency-stabilizing features.

II. DC VS RF EXCITATION

Dc excitation of gas lasers understandably is the most used technique. Coupling of power is direct; efficient ionization of the gas is sure; and the discharge is maintained without requiring any special attention on the part of the user. However, some dc discharges (although stable on a macroscopic scale) are known to be electrically noisy, and various workers now have observed noise in the output of dc lasers. The amount and character of the noise is variable, but Bolwijn et al[6] have observed noise power as high as 77 db above detector shot noise. Prescott and van der Ziel[7] have demonstrated a correlation between the laser noise and the dc discharge current noise. The variability of the noise is described by Bellisio et al,[8] who occasionally found conditions under which laser noise did not noticeably exceed detector shot noise. Bellisio[9] did not learn how to achieve this quiet condition in a controlled way. Although some workers may have learned ways to reduce the noise in dc lasers, there appears to have been no publication of any technique or theory.

In contrast to dc excitation, RF discharges are characteristically quiet. Both Bellisio et al[8] and Bailey and Sanders[10] found no laser noise significantly above detector shot noise when using RF excitation of the

laser. Paik et al[11] have found that application of RF signals to the anode of dc discharge tubes causes the noise to be replaced by "coherent, noise-free oscillations." The suitability of RF excitation for stable operation of lasers is further suggested by the fact that RF lasers can be operated closer to threshold than can dc lasers.

Since one objective of the present work is to obtain as stable a laser frequency as possible, RF excitation has been used. Capacitive coupling of RF power to the small bore ($\approx$1 mm) discharges required for single-mode operation poses a major technical problem, and its solution will be described.

III. LASER DESIGN: SINGLE FREQUENCY

The laser was designed to oscillate at maximum power in one frequency in a cavity of maximum intrinsic stability. The general approach was similar to that of Gordon and White,[5] and involved building a two-mirror cavity of such geometry that only one axial mode can oscillate. Maximum power, however, required analytical selection of cavity length and mirror transmission.

From the Fabry-Perot condition for resonance, the frequency interval between successive axial modes is $c/2\mu d$, where c is the velocity of light, μ is the index of refraction of the medium, and d is the physical length of the cavity. Clearly, making d arbitrarily small makes the mode separation arbitrarily great. However, as d diminishes, so does the gain per pass at the line center. We want that length which gives maximum absolute difference between the gain of the desired mode, presumed to be located near the line center, and the gain of the adjoining axial mode. Hence, for a low-gain transition (order of ten per cent per meter), we write for the gain of the Doppler-broadened line

$$g(\nu,d) = g_0 d \exp - \left[\frac{\nu - \nu_0}{0.6\Delta\nu_D}\right]^2$$

where g_0 is gain per unit length at line center, d is cavity length, $\nu - \nu_0$ is frequency interval from line center, and $\Delta\nu_D$ is the half-maximum Doppler width. The gain is evaluated at line center and at $\nu - \nu_0 = c/(2\mu d)$ and the difference is maximized. For the 6328Å line in Ne at 450°K, the resultant optimum cavity length is 15 cm. This assumes that gain is uniform from mirror to mirror. For a real cavity, part of whose length must be wasted, the mirror separation is made $16\frac{1}{2}$ cm and the length of gain is $13\frac{1}{2}$ cm.

It is well known that gain, in this type of gas laser, varies inversely with tube diameter. A diameter of 1.2 mm yields an expected g_0 of 20

per cent per meter,[5] or 2.7 per cent in $13\frac{1}{2}$ cm, which is well above dissipative losses. A tube diameter of 1.2 mm gives a Fresnel number of 0.7 in a cavity comprising a flat mirror separated $16\frac{1}{2}$ cm from a one meter-radius spherical mirror. This gives 0.4 per cent diffraction loss per pass in the lowest-order transverse mode, but at least 5 per cent loss per pass in all higher-order modes.[12] Thus, the cavity will support only one transverse mode.

For $d = 16\frac{1}{2}$ cm, $c/(2\mu d) = 900$ mc, and, with one axial mode on the line center, gain at the adjoining axial modes is about 1 per cent. In a cavity designed for minimum loss, these adjoining modes would oscillate. However, when maximum power is desired in an external beam, the procedure is then to increase transmission of one mirror until losses are too great for the adjoining modes. Bennett[13] has reported a calculation by Kompfner and Rigrod of the optimum mirror transmission for maximum coupling out of both (identical) cavity mirrors. They obtain

$$T_{\text{opt}} = \sqrt{GL} - L$$

where G is gain per pass and L is loss per pass. Since maximum power is desired in one beam, and since mirrors now can be made with total losses much smaller than other cavity losses,[14] one mirror is coated for maximum reflectivity. Such mirrors are estimated to reflect at least 99.8 per cent. In the present "round-trip" case, the above expression gives a T_{opt} for the output mirror of about 1 per cent or a little more for a realistic range of estimated losses. The output mirror therefore is coated for 1 per cent transmission.

This design gives a single-frequency laser only if the desired axial mode is close to the line center. If two adjoining modes are arranged symmetrically about the line center, each will have a gain of about 2 per cent and will oscillate. Thus, as the cavity changes length by quarter wavelengths, it drifts between one and two oscillating modes. As another option, the cavity can be made so short that only one axial mode can exist at any instant, but such a cavity would now drift between one and zero oscillating modes. Unless cavity dimensions are held constant to about a quarter wave, the only real choice is between a cavity that oscillates sometimes in two modes and a cavity that sometimes doesn't oscillate. The present approach is to design for maximum power and to stabilize at least well enough to maintain one frequency.

IV. LASER DESIGN: STABILITY

The degree of frequency instability is obtained from the Fabry-Perot condition for resonance, from which it follows that

$$-\frac{\Delta\nu}{\nu} = \frac{\Delta\mu}{\mu} + \frac{\Delta d}{d}$$

where ν is frequency, μ is the index of refraction of the medium, and d is physical length of the cavity. The size of $\Delta\mu/\mu$ depends on the construction of the laser. External-mirror lasers shift in frequency by tens of megacycles as the air within the cavity changes temperature and pressure. Thermal expansion of the frame changes d. A change in temperature of 0.1°C shifts the frequency of the 6328Å laser 1,200 mc on an aluminum frame, 40 mc on an Invar frame, and 20 mc on a frame of fused quartz. Jaseja et al[15] observed that lasers on Invar frames can shift 140 kc by magnetostriction in the earth's field.

Past work on frequency stabilization has centered on acoustic isolation of the laser[16] or the generation of an error signal which is fed back and which corrects the length of the cavity[17-20] against thermal drift. Much less attention seems to have been paid to the stability of the structure itself. The effort here has been to construct a laser frame of maximum intrinsic mechanical stability. The cavity, shown in Fig. 1, comprises a flat fused quartz mirror, a perforated tube of fused quartz $1\frac{1}{2}$ inches in diameter and $6\frac{1}{2}$ inches long, and a spherical fused quartz mirror of one meter radius. Each end of the quartz tube is relieved so as to leave three small studs equally spaced around the tube circumference. The lines connecting pairs of studs at opposite ends of the tube are parallel with the tube axis. The stud faces are slightly convex and optically polished. The margins of the mirrors are uncoated, and, when the mirrors are properly installed, optical contact can be observed be-

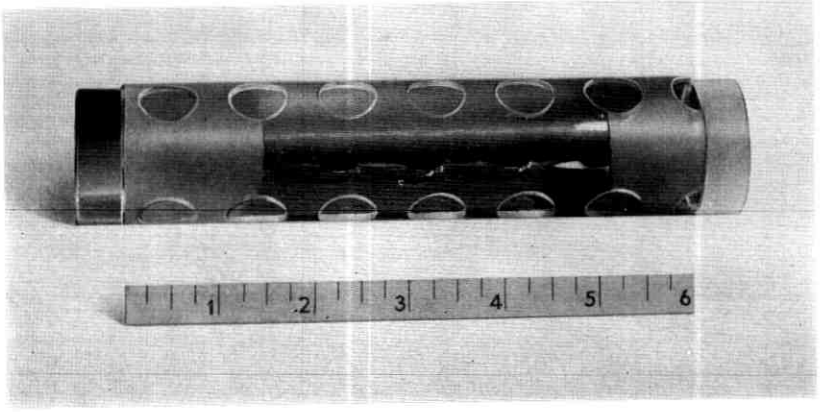

Fig. 1 — The laser cavity, comprising two mirrors and a fused quartz tube which alone determines mirror spacing.

tween mirrors and studs. (A small black spot appears, surrounded by Newton rings.) Ordinarily, mechanical structures contain microscopically rough interfaces. Disturbances of such structures produce jitter in the relative positions of the parts as the actual points of contact are shifted. Such jitter is precluded in the present design. Meissner[21] has observed that passive Fabry-Perot interferometers must follow this design if they are to maintain their adjustment. Some idea of the consequences of this for laser cavities can be gained from realizing that the best resolving power in Meissner's day was the order of 10^7. For an instrument with a resolution of 10^7 to lose its adjustment implies a change in optical frequency of at least tens of megacycles.

The mirrors are held against the ends of the quartz tube; there is no provision for mirror adjustment. The tube is simply made with sufficient care that the optical axis of the installed mirrors is sensibly collinear with the axis of the tube. The discharge tube, shown in Fig. 2, is mounted in the cavity and is then positioned laterally until it is aligned on the cavity axis. The discharge tube is then clamped into position with screws which are mounted on the quartz tube.

Each mirror is backed up by a spring which ensures steady contact between mirror and quartz tube. The spring also decouples the compression of the quartz tube from the supporting structure which uses ordinary, large thermal coefficient materials. The cavity is tuned by adjusting spring compression and hence compression of the quartz tube. Young's modulus and cross section of the tube are such that a force of about 4 pounds tunes the cavity through one axial mode spacing, effectively 900 mc.

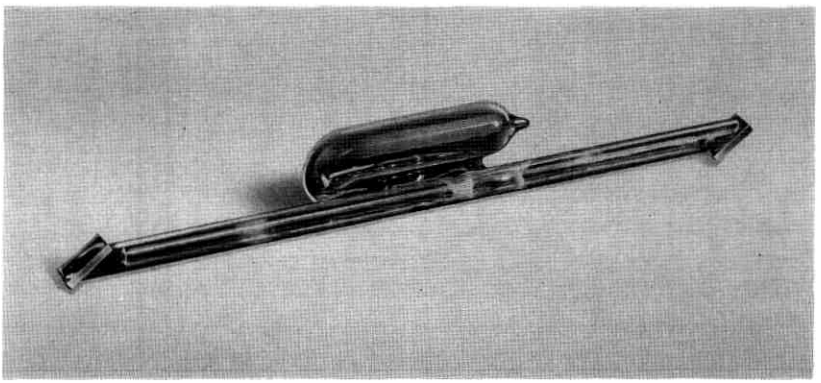

Fig. 2 — The fused quartz discharge tube. The windows are optically contacted. Length is 6¼ inches.

These are the essentials of the design. One embodiment of the design is shown in Fig. 3. Most of the materials of the supporting structure are plastic to avoid excessive eddy-current losses. The RF is coupled capacitively with $1\frac{1}{2}$-inch diameter copper rings which are far out of contact with the discharge tube. The rings are actually a snug fit to the inside of the quartz tube of the cavity. When the electrodes are wrapped directly on the outside of the discharge tube, the electric field in the discharge tube near the electrode is large. Bombardment damage is then rapidly produced on the inside wall of the tube, and tube life is much reduced. The life of RF-excited laser tubes appears to be determined by the rate of this damage process.

V. DISCHARGE TUBE

The discharge tube, shown in Fig. 2, is $6\frac{1}{4}$ inches overall. The tube-within-a-tube arrangement (1.2 mm bore inside a 4 mm bore tube) provides the small bore needed for high gain over most of the length, while at the same time leaving a large diameter space at each end which is easily ionized by the RF power. Once the ends are ionized, the small bore section then lights also. A straight tube of 1.2 mm bore can be ionized with capacitive coupling, but only with very large electric fields and hence a large rate of damage.

The body of the tube and windows are of fused quartz. Assembly is accomplished by optically contacting the windows to the optically-polished faces of the tube. Early research with RF-excited gas lasers showed that tube life was severely limited unless the tube was made of

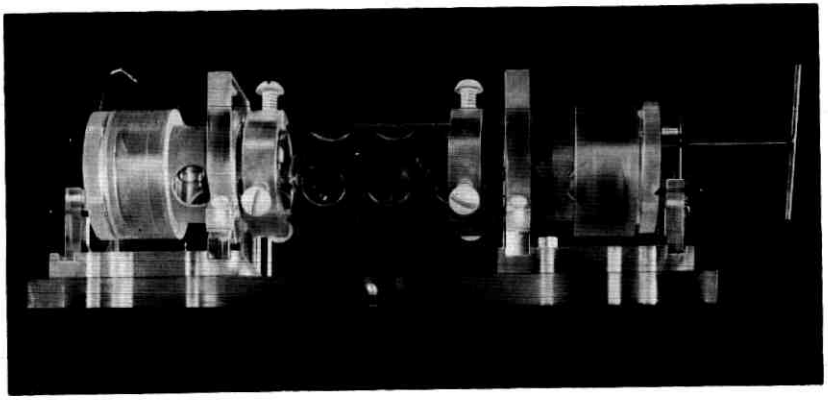

Fig. 3 — One form of the assembled laser. The supporting structure is mostly plastic to avoid eddy-current losses.

fused quartz.[22] The evidence in the literature[23-26] suggests that the reason for this is that the rate of damage in ordinary glasses is orders of magnitude greater than in quartz. Since the tube is entirely of quartz, the dominant source of contamination is the damage process cited above. Tubes of the type shown have had useful lifetimes of hundreds of hours. Getters now are used in the tubes, and greater lifetimes are anticipated.

The tubes are pumped and baked at about 450°C overnight. When the system has cooled and is valved off before filling, the pressure is in the 10^{-9} Torr range. The tubes are filled to 3.0 Torr with a 5/1 ratio of He-Ne gas. Natural abundance gases are used. These tubes, in the cavities described above, routinely radiate about 1 mw in one frequency in a single coherent external beam, and values as large as 1.6 mw have been obtained.

The frequency characteristics of these lasers are now being investigated. It is commonly known that the short-term frequency spread of the oscillation of external mirror gas lasers under laboratory conditions typically is tens, even hundreds, of megacycles. Our preliminary measurements show that this can be reduced to 100 kc or less with the present lasers — even when operated on an optical bench in a second-floor laboratory. Work is continuing to reduce the microphonic sensitivity of these lasers, and further improvements in the frequency spread are expected.

I wish to acknowledge the resourceful, energetic assistance of R. H. Delaney.

REFERENCES

1. Javan, A., Ballik, E. A., and Bond, W. L., J. Opt. Soc. Am., *52*, 1962, p. 96.
2. Kleinman, D. A., and Kisliuk, P. P., B.S.T.J., *41*, March, 1962, p. 453.
3. Kogelnik, H. W., and Patel, C. K. N., Proc. I.R.E., *50*, 1962, p. 2365.
4. Massey, G. A., Oshman, M. Kenneth, and Targ, Russell, Appl. Phys. Letters, *6*, 1965, p. 10.
5. Gordon, E. I., and White, A. D., Proc. IEEE, *52*, 1964, p. 206.
6. Bolwijn, P. T., Alkemade, C. Th. J., and Boschloo, G. A., Phys. Letters, *4*, 1963, p. 59.
7. Prescott, L. J., and van der Ziel, A., Appl. Phys. Letters, *5*, 1964, p. 48.
8. Bellisio, Jules A., Freed, Charles, and Haus, Hermann A., Appl. Phys. Letters, *4*, 1964, p. 5.
9. Bellisio, Jules A., private communication.
10. Bailey, R. L., and Sanders, J. H., Phys. Letters, *10*, 1964, p. 295.
11. Paik, S. F., Wallace, R. N., and McClees, H. C., Phys. Rev. Letters, *10*, 1963, p. 78.
12. Boyd, G. D., and Gordon, J. P., B.S.T.J., *40*, March, 1961, p. 489.
13. Bennett, W. R., Jr., Appl. Opt. Suppl., *1*, 1962, p. 24.
14. Perry, D. L., Mirror Coating Procedures for High-Power Gas Lasers, NEREM Conference, Nov. 4–6, 1964, Boston, Mass.
15. Jaseja, T. S., Javan, A., Murray, J., and Townes, C. H., Phys. Rev., *133*, 1964, p. A1221.
16. Jaseja, T. S., Javan, A., and Townes, C. H., Phys. Rev. Letters, *10*, 1963, p. 1965.

17. Rowley, W. R. C., and Wilson, D. C., Nature, *200*, 1963, p. 745.
18. Shimoda, K., Paper 2-1 Conf. Precision Electromagnetic Measurements, June 23, 1964, Boulder, Colorado.
19. Bennett, W. R., Jr., Jacobs, S. F., LaTourrette, J. T., and Rabinowitz, P., Appl. Phys. Letters, *5*, 1964, p. 56.
20. White, A. D., Gordon, E. I., and Labuda, E. F., Appl. Phys. Letters, *5*, 1964, p. 97.
21. Meissner, K. W., J. Opt. Soc. Am., *31*, 1961, p. 405.
22. Bennett, W. R., Jr., private communication.
23. Johnson, B., Lineweaver, Jack L., and Kerr, John T., J. Appl. Phys., *31*, 1960, p. 51.
24. Bills, D. G., and Evett, A. A., J. Appl. Phys., *30*, 1959, p. 564.
25. Donaldson, E. E., and Rabinowitz, M., J. Appl. Phys., *34*, 1963, p. 319.
26. Allen, F. G., Buck, T. M., and Law, J. T., J. Appl. Phys., *31*, 1960, p. 31.

Contributors to This Issue

PAUL T. BRADY, B.E.E., 1958, Rensselaer Polytechnic Institute; M.S.E.E., 1960, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1961—. His work in human factors engineering has been concerned with studies of speech and voice-operated devices, especially as applied to satellite communication circuits.

W. JAMES COLE, B.S.E.E., 1963, Lehigh University; M.S.E.E., 1964, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1963—. He has been working on the simulation and design of control systems for nonorthogonal antenna mounts. Member, IEEE, Tau Beta Pi, Eta Kappa Nu.

J. A. COLLINSON, A.B., 1950, Oberlin College; M.S., 1951, Yale University; Ph.D., 1954, Yale University; Bell Telephone Laboratories, 1962—. He has worked on gas lasers and placed emphasis on frequency characteristics and atmospheric transmission of laser beams. Member, American Physical Society, Sigma Xi, Phi Beta Kappa.

JOHN S. COOK, B.S., M.S., in E.E., 1952, Ohio State University; Bell Telephone Laboratories, 1952—. He has engaged in research on traveling wave tubes, electron focusing, coupled systems, parametric amplification, and more recently in phased-array radar study. He is now head of the antenna research department engaged in microwave antenna and transmission research and development, and laser research. Senior member, IEEE, Sigma Xi, Eta Kappa Nu, Tau Beta Pi.

KEVIN N. COYNE, B.S.M.E., 1959, Columbia University; M.S. (Engineering Mechanics) 1961, New York University; Bell Telephone Laboratories, 1959—. He has been engaged in the design and development of electro-optical equipment and has worked on structural design and analysis and erection of the Andover horn-reflector antenna. Mr. Coyne also has worked on the development of a precise digitally controlled mount for infrared radiometry and on mathematical studies for submarine detection systems. He is currently engaged

in design studies for future satellite communications antennas. Member, Tau Beta Pi.

W. J. Denkmann, B.S.M.E., 1961, M.S., 1963, State University of Iowa; Bell Telephone Laboratories, 1963—. Mr. Denkmann is concerned with the analysis of structures, particularly antennas, and with the computer applications of structural analysis techniques. Member, Tau Beta Pi, Pi Tau Sigma, ASME, ACM, SIAM.

Edward M. Elam, B.S.E.E., 1952, University of California, Western Electric Company, 1952–1962; Bell Telephone Laboratories, 1962—. He has worked on the TELSTAR satellite autotrack systems at Andover, Maine, and Pleumeur-Bodou, France. He is presently engaged in the development of radar and communications antennas. Member, IEEE.

K. E. Fultz, B.S.E.E., 1948, M.S.E.E., 1950, Kansas State University; Bell Telephone Laboratories, 1950—. He was first concerned with systems engineering studies of Bell System transmission systems and with studies of the Distant Early Warning line. He later worked on systems engineering of T1 carrier and in presently, Head, Pulse Transmision Studies Department. Member, IEEE, Tau Beta Pi, Eta Kappa Nu, Phi Kappa Phi.

Franz Th. Geyling, B.S. (Civil Eng.), 1950, M.S. (Civil Eng.), 1951, and Ph.D. (Eng. Mechanics), 1954, Stanford University; Bell Telephone Laboratories, 1954—. He has worked in areas involving photoelastic stress analysis, shell theory and satellite and space vehicle ballistics. He has written computer programs for the digital simulation of space flight missions. His other areas of responsibility include blast studies, the analysis of large antenna structures, and hypervelocity impact studies. Member, AIAA, ASME, International Association for Bridge and Structural Engineering, Tau Beta Pi.

Richard A. Gibby, B.S., 1949, M.S., 1950, University of Utah; Ph.D., 1955, Northwestern University; Bell Telephone Laboratories, 1955—. Engaged in data communication system analysis, he is presently in charge of a group concerned with this work. Member, IEEE, Eta Kappa Nu, Sigma Xi, Tau Beta Pi.

Adolf J. Giger, Diploma in Electrical Engineering, 1950, and Dr. sc. techn. 1956, Swiss Federal Institute of Technology; Bell Telephone

Laboratories, 1956—. Mr. Giger was first associated with the TH microwave relay system as a circuit designer and later as a leader of a group engaged in the development of the TH protection switching system. Later he was active in the field of satellite communications especially low noise receivers, antennas and waveguide circuits for ground stations. He is now supervisor of a group working on an all solid-state microwave radio system. Senior member, IEEE.

Elliott R. Nagelberg, B.E.E., 1959, City College of New York; M.E.E., 1961, New York University; Ph.D., 1964, California Institute of Technology; Bell Telephone Laboratories, 1964—. He has been concerned with problems involving microwave antennas and propagation. Member, IEEE, American Physical Society, Eta Kappa Nu, Sigma Xi.

Winston L. Nelson, B.S., 1950, University of Utah; M.S., 1953, Ph.D., 1959, Columbia University; Bell Telephone Laboratories, 1960—. He has been involved with studies of satellite tracking control systems, satellite attitude control systems, weak-signal detection techniques employing feedback, and system optimization techniques. He presently supervises a group engaged in communication and control systems research. Member, IEEE, SIAM, Sigma Xi, Tau Beta Pi.

Dixon B. Penick, B.S. in E.E., 1923, B.A. 1924, University of Texas; M.A. 1927, Columbia University; Western Electric Company, Engineering Department, 1924–25; Bell Telephone Laboratories, 1925—. After 12 years in the vacuum tube research area, he transferred to a carrier development group and has worked on C5 Carrier, 2B Pilot Channel, J2 Carrier, A2 Channel Bank, Carrier Program, L3 mastergroup, and P1 Carrier developments. In 1960 he was assigned to a PCM group and is now responsible for continuing development work on the T1 Carrier system. Member, Tau Beta Pi, senior member, IEEE.

Daniel L. Pope, B.C.E., 1953, Ph.D. (Mechanics), 1961, Cornell University; Bell Telephone Laboratories, 1960—. He has been concerned with defensive missile system analysis, and participated in the early phases of orbital mechanics studies for communication satellites. More recently he has worked on array radar structural analysis, study of advanced techniques for large complex structures and structural optimization. He presently supervises the Analytical Mechanics De-

partment's Continuum Mechanics group. Member, Chi Epsilon, SIAM, Tau Beta Pi.

ALFRED O. SCHWARZ, B.S.M.E., 1949, Lehigh University; Bell Telephone Laboratories, 1951—. He has worked on a variety of electromechanical devices associated with servo mechanisms, as well as a proximity fuze for mortars, and optical equipment for evaluating tracking performance of radar antennas. He performed engineering laision work for the design and installation of the horn antennas at Andover, Maine and Pleumeur-Bodou, France for the TELSTAR satellite project. He is currently working on the development of a coin telephone set.

JOSHUA SHEFER, B.S., 1948, Technion, Israel Institute of Technology; Ph.D., 1956, London University; research fellow, Harvard University, 1960–1962; Bell Telephone Laboratories, 1962—. He has been engaged in studies of microwave propagation and antennas, particularly in relation to surface wave guiding structures. Senior member, IEEE, associate member, IEE (London), member, Sigma Xi.

JAMES W. SMITH, B.E.S., 1956, Dr. Eng., 1963, Johns Hopkins University; Bell Telephone Laboratories, 1963—. He has been concerned with analysis problems in the areas of analog and digital data transmission. Member, IEEE, Tau Beta Pi, Sigma Xi, Eta Kappa Nu.

RICHARD H. TURRIN, B.S.E.E., 1956, Newark College of Engineering; M.S.E.E., 1960, New York University; Bell Telephone Laboratories, 1956—. He has been concerned with propagation and antenna work at micro- and millimeter wavelengths. He participated in the design of the TELSTAR satellite ground-station antennas. Member, IEEE, Eta Kappa Nu, Tau Beta Pi.

H. ZUCKER, Dipl.-Ing. 1950, Technische Hochschule, Munich, Germany; M.S.E.E., 1954, Ph.D, 1959, Illinois Institute of Technology; Bell Telephone Laboratories, 1964—. He has been engaged in the analysis and design of satellite communication antennas. Member, IEEE, Eta Kappa Nu, Sigma Xi.

B.S.T.J. BRIEFS

Planar Epitaxial Silicon Schottky Barrier Diodes

By D. KAHNG and M. P. LEPSELTER

(Manuscript received June 30, 1965)

It has been demonstrated that a metal-to-semiconductor rectifying junction can be designed as a fast computer diode.[1,2] It can also be designed as a high performance varactor.[3,4] There is increasing evidence that Schottky barrier diodes may, with suitable design modifications, outperform the conventional point contact diodes in the field of varistor applications.[5,6]

The earlier versions of ESBAR (Epitaxial Schottky Barrier) diodes had either a mesa-like structure, well-encapsulated to withstand the environmental influences, or a pseudo planar structure of doubtful passivation capability. Manufacturability of these diodes is greatly improved by taking advantage of the "planar" process of making diodes by the well-known photoresist masking. Such a structure is shown in Fig. 1.

The rectifying barrier in a planar ESBAR diode was obtained between a suitable metal having an appreciable amount of silicon in solution (for convenience, labeled as metal silicide in Fig. 1) and the epitaxial silicon. The metal silicide was formed by evaporating the metal over the silicon dioxide with appropriately sized windows, and allowing it to react with the exposed silicon at a suitable temperature (anywhere between 300°C and 700°C, depending on the metal used) for relatively short periods of about 30 minutes. The amount of Si depletion due to this solid-solid reaction can be controlled by the amount of metal available and the reaction temperature.

After removing unreacted metal by a suitable means, such as selective etching, which leaves silicon rich metal in the oxide window, a 0.5 μ thick Pt layer was deposited over the entire surface preceded by a thin film ($\sim$200 Å) deposition of Cr or Ti. The latter metals insure good adhesion of Pt layer on the oxide surface inhibiting lateral diffusion of ambient gases. The photoresist technique was again used to achieve a selective Au plating on Pt, overlapping the oxide window. The thickness of this plated Au is typically a few microns. This serves as an effective mask for the operation of Pt removal by some means,

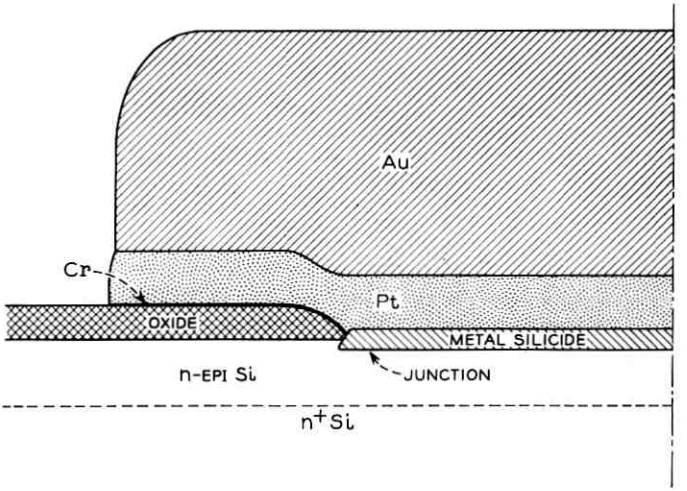

Fig. 1 — Structure of planar ESBAR.

such as backsputtering,[7] from surfaces other than the ones under the Au layer.

The major feature of this structure is the reasonable passivation of the rectifying barrier against the environment and suppression of leakage currents. For instance, the leakage current is much smaller in these diodes than the mesa type diodes. The n factor in the diode equation

$$I = I_s \exp\left(\frac{qV}{nkT} - 1\right)$$

is less than 1.1 as compared to 1.2–1.5 for the unprotected diodes of area of one mil diameter circle. Forward characteristics of Cu-Si diodes of diameter of 2, 1 and $\frac{1}{2}$ mils are shown in Fig. 2. Here $n = 1.03$, the theoretically expected value.[8] Sharp reverse breakdown is common for these diodes as opposed to the mesa type where it is a rarity.

These diodes, unencapsulated, have shown no degradation after 100 hours aging at 350°C in room air or at 300°C in one atmosphere of steam. As for the non-planar diodes, hermetically sealed encapsulation is mandatory in order to survive such stress agings.

The barrier height measured from the Fermi level of a metal silicide diode is significantly different from the metal-silicon barrier height. For instance, a Pt-Si diode has barrier height of larger than 1 volt while a Pt silicide-Si diode has the value of 0.87 volt. Cu-Si system

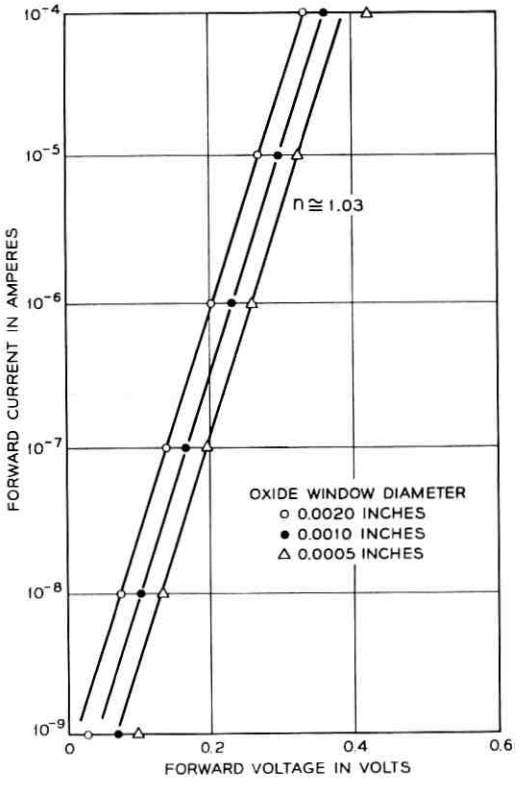

Fig. 2 — Forward characteristics of planar copper silicide, epitaxial silicon, surface barrier diodes.

gives the barrier height of 0.58 volt. Cu-Silicide-Si system, after 30 minutes heating at 350°C, has the barrier height of ~0.78 volts. If more Si is allowed to dissolve in the "silicide", by raising the reaction temperature, the barrier height increases to saturate at about 0.9 volts. This saturation occurs at various temperatures depending on the metals. For instance, Mo-Si barrier shows little change after heating up to 700°C. Similar results on W-Si system have also been reported.[9]

In addition to increased reliability, one obtains with the planar ESBAR structure, better control of diode geometry inherent with the photoresist techniques. This should result in much more uniform terminal characteristics of the diodes. One of the important parameters is the stray capacitance associated with the Pt overlay. One wishes, in general, to minimize the stray capacitance. It should be recognized

that minimization of the stray capacitance associated with the overlay is achieved only with decreased ambient protection due to the decrease in the overlay area.

REFERENCES

1. Kahng, D., and D'Asaro, L. A., Microwave Diode Research Report No. 12, U.S. Army Signal R&D Lab., Contract DA36-039 SC-89205, 10 June 1963, also published in B.S.T.J., *43*, 1964, pp. 225–232.
2. Krakauer, S. M., and Soshea, S. W., Electronics, *29*, 1963, pp. 53–55.
3. Irvin, J. C., Microwave Diode Research Report No. 13, U. S. Army Signal R&D Lab., Contract DA36-039 SC-89205, 20 Sept. 1963.
4. Kahng, D., B.S.T.J., *43*, 1964, pp. 215–224.
5. Herndon, M., and MacPherson, A. C., Proc. IEEE, *52*, 1964, pp. 975–976.
6. Vanderwal, N. C., Bell Telephone Laboratories, Inc., private communications.
7. Lepselter, M. P., Metalizing for Beam-Lead Devices, late news paper presented to ECS Meeting, San Francisco, California, May 13, 1965.
8. Sze, S. M., Crowell, C. R., and Kahng, D., JAP, *35*, 1964, pp. 2534–2536.
9. Crowell, C. R., Sarace, J. C., and Sze, S. M., Trans. Metallurgical Soc., AIME *233*, 1965, pp. 478–481.

4-gc Transmission Degradation Due to Rain at the Andover, Maine, Satellite Station

By A. J. GIGER

(Manuscript received July 22, 1965)

I. INTRODUCTION

The microwave link between a ground station and a communications satellite is normally very stable and essentially free from fading. Under conditions of rain or snow, however, the transmitted and received signals encounter extra attenuation and additional noise is introduced into the low-noise receiver on the ground. A good knowledge of such rain effects is important for the design of satellite ground stations which have to meet certain statistical requirements for transmission degradation. It is known that radome covered ground stations like Andover, suffer more degradation during rain than uncovered stations. Some analytical work has been done by D. Gibble[1] and B. C. Blevis[2] to determine the effects of a water layer on radomes. Their theoretical work has been supplemented by an experimental technique applicable at existing satellite ground stations and to be described in this brief report. It consists of measuring the reduction of the noise power received from the strong and stable radio star Cassiopeia A during periods of rain.

that minimization of the stray capacitance associated with the overlay is achieved only with decreased ambient protection due to the decrease in the overlay area.

REFERENCES

1. Kahng, D., and D'Asaro, L. A., Microwave Diode Research Report No. 12, U.S. Army Signal R&D Lab., Contract DA36-039 SC-89205, 10 June 1963, also published in B.S.T.J., *43*, 1964, pp. 225–232.
2. Krakauer, S. M., and Soshea, S. W., Electronics, *29*, 1963, pp. 53–55.
3. Irvin, J. C., Microwave Diode Research Report No. 13, U. S. Army Signal R&D Lab., Contract DA36-039 SC-89205, 20 Sept. 1963.
4. Kahng, D., B.S.T.J., *43*, 1964, pp. 215–224.
5. Herndon, M., and MacPherson, A. C., Proc. IEEE, *52*, 1964, pp. 975–976.
6. Vanderwal, N. C., Bell Telephone Laboratories, Inc., private communications.
7. Lepselter, M. P., Metalizing for Beam-Lead Devices, late news paper presented to ECS Meeting, San Francisco, California, May 13, 1965.
8. Sze, S. M., Crowell, C. R., and Kahng, D., JAP, *35*, 1964, pp. 2534–2536.
9. Crowell, C. R., Sarace, J. C., and Sze, S. M., Trans. Metallurgical Soc., AIME *233*, 1965, pp. 478–481.

4-gc Transmission Degradation Due to Rain at the Andover, Maine, Satellite Station

By A. J. GIGER

(Manuscript received July 22, 1965)

I. INTRODUCTION

The microwave link between a ground station and a communications satellite is normally very stable and essentially free from fading. Under conditions of rain or snow, however, the transmitted and received signals encounter extra attenuation and additional noise is introduced into the low-noise receiver on the ground. A good knowledge of such rain effects is important for the design of satellite ground stations which have to meet certain statistical requirements for transmission degradation. It is known that radome covered ground stations like Andover, suffer more degradation during rain than uncovered stations. Some analytical work has been done by D. Gibble[1] and B. C. Blevis[2] to determine the effects of a water layer on radomes. Their theoretical work has been supplemented by an experimental technique applicable at existing satellite ground stations and to be described in this brief report. It consists of measuring the reduction of the noise power received from the strong and stable radio star Cassiopeia A during periods of rain.

II. THE MEASURING TECHNIQUE

If the ground station antenna is pointed exactly at Cassiopeia A, the noise power received will be proportional to

$$T_{tot} = T_{SYS} + tT_A \tag{1}$$

where T_{SYS} = receiving system noise temperature referred to the input terminal of the low noise receiver (maser) if the antenna is pointed in the vicinity of the star but far enough away from it to make its contribution to the noise temperature negligible. T_{SYS} can be conveniently measured at the Andover station

T_A = temperature due to the radio star alone at the maser input under the assumption of zero loss in the transmission path (except for geometrical path loss)

t = power transmission coefficient due to the transmission path (atmosphere, rain, radome, waveguides) to the input of the maser.

Solving (1) for t we obtain:

$$t = \frac{T_{tot} - T_{SYS}}{T_A} = \frac{T_{SYS}}{T_A}\left(\frac{T_{tot}}{T_{SYS}} - 1\right). \tag{2}$$

Now we introduce $\Delta = T_{tot}/T_{SYS}$, a quantity which can be easily measured as a power ratio at the output of the receiver by moving the antenna on and off the radio star. Equation (2) becomes now:

$$t = (T_{SYS}/T_A)/(\Delta - 1). \tag{3}$$

A measurement during dry weather yields

$$t_o = (T_{SYS_o}/T_A)(\Delta_o - 1). \tag{4}$$

This allows the definition of an excess transmission coefficient

$$t_x = \frac{t}{t_o} = \frac{T_{SYS}}{T_{SYS_o}}\frac{\Delta - 1}{\Delta_o - 1} \tag{5}$$

giving the signal loss over the normal "dry" value. The degradation D of the signal-to-noise (SNR) ratio in the receiver is

$$D = \frac{\text{SNR}}{\text{SNR}_o} = \frac{t/T_{SYS}}{t_o/T_{SYS_o}} = t_x\frac{T_{SYS_o}}{T_{SYS}} = \frac{\Delta - 1}{\Delta_o - 1}. \tag{6}$$

Since rainfall occurs at rather unpredictable times, it was not feasible to drive the Andover antenna along the path of Cassiopeia A by computer tape. Instead, a computer printout was made available to the

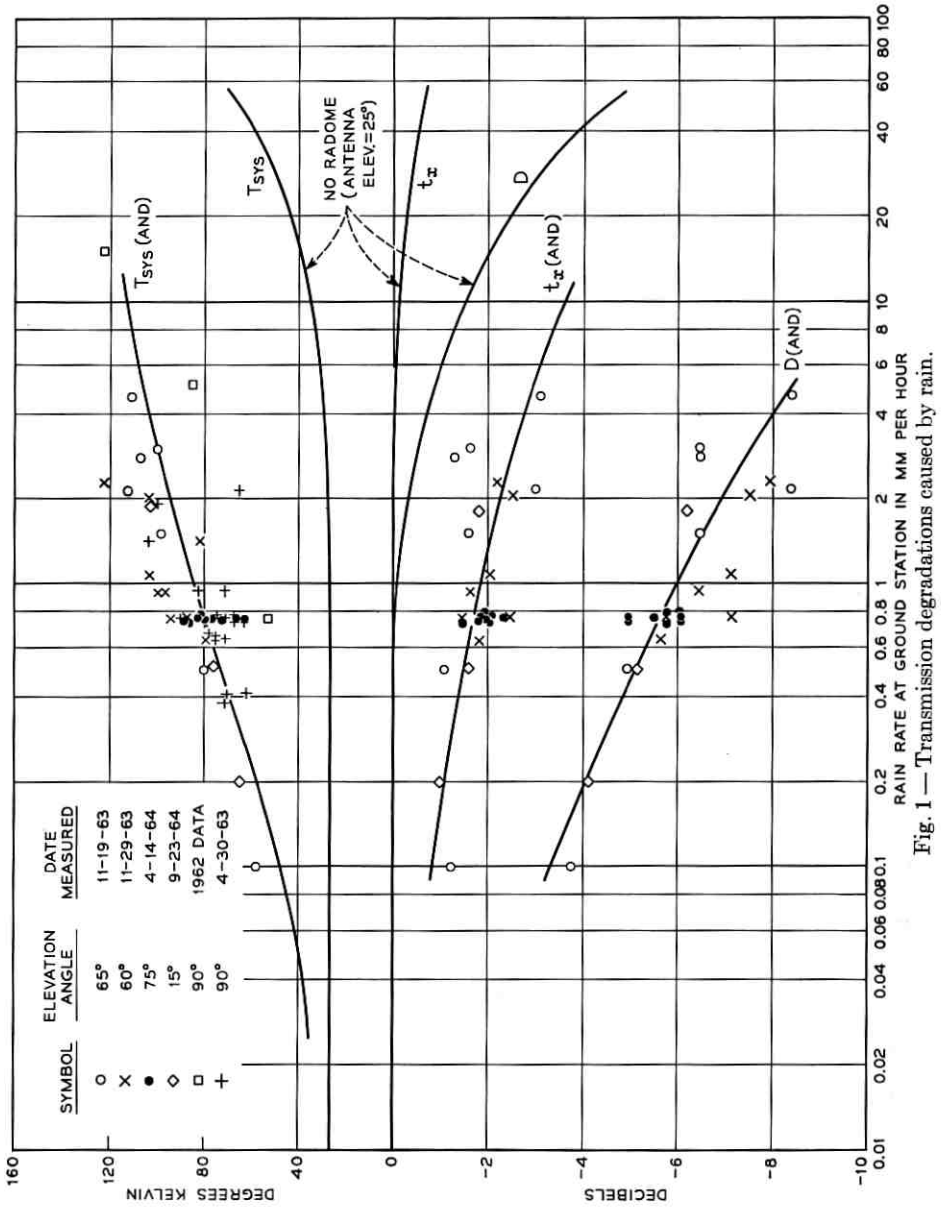

Fig. 1 — Transmission degradations caused by rain.

station operators in Andover containing the star's coordinates at intervals of 10 minutes for a period of several months. The antenna is then set manually to the calculated Cas A coordinates about 5 minutes ahead of time and the drift of the star through the antenna beam is observed on the IF power meter at the output of the receiver. The maximum noise level is recorded and the antenna is quickly swung away about 3° from the star. The noise power drop yields Δ. In addition, T_{SYS} is measured in this position. Measurements can be repeated every 10 minutes at any desired time since Cassiopeia A is always visible from Andover at elevation angles ranging from 13.3° to 76°. Equations (5) and (6) require the knowledge of T_{SYS_o} and Δ_o, two quantities which have to be measured on a dry day as a function of antenna elevation angle.

Measured values of T_{SYS}, t_x and D are plotted in a scatter diagram (Fig. 1) vs the rain rate existing at the Andover site. Because the transmission degradation is mainly due to the water layer on the radome, which is directly related to the local rain rate, a reasonable correlation is exhibited by the scatter plots. Since the radome is still wet, some transmission degradation is observed for about 30 minutes after the rain has stopped. Measurements taken during this period of time have been excluded from Fig. 1.

The measurements were made at the elevation angles of Cassiopeia A at the time of the rain, but the results do not seem to indicate any systematic dependence on elevation. This has been found before during measurements of T_{SYS} vs elevation during rain (see Fig. 16 of Ref. 3). It also checks with Gibble's theory which says that the thickness of the water layer on the radome is approximately the same everywhere.

As indicated by Fig. 1, the degradation in signal-to-noise ratio can be appreciable. Since it is the water layer and not the radome material which causes the degradation, results are expected to be similar with other materials. This assumes that water cannot penetrate the material and is preferably repelled by the radome surface. The inflated Andover radome, made of Hypalon coated Dacron fabric, has such desirable features. It is about 2-mm thick, the dielectric constant is 3.0, and the loss tangent 0.0155 under dry conditions.

III. ESTIMATED RESULTS FOR AN UNCOVERED ANTENNA

No measurements of the same nature are available for an uncovered antenna at Andover. It is important, however, to know the difference in the electrical characteristics of a covered and an uncovered antenna during rain. A simple model is proposed to estimate the rain characteristics of the uncovered antenna. It consists of a volume of rain with uniform

rate of precipitation and temperature T_R and a constant ceiling, h, of 5 km, extending indefinitely in lateral direction. This model should give a pessimistic answer for the transmission degradation because: (1.) precipitation normally stops at an altitude of about 3 km in temperate zones and at about 1 km in tropical zones with a heavy nimbo-stratus cloud mass extending to about 3 km, Ref. 4, and (2.) the average rain rate over an extended area is normally below the rain rate measured (during rainfall) at a fixed point[5] (the satellite ground station). We further assume that the antenna is directed at a fixed elevation of 25° which corresponds to the angle at Andover when operating with the stationary satellite Early Bird. This model gives a path-length in rain of $l = h/\sin 25° = 5/0.4226 = 11.8$ km. Signal attenuations per km, α, for various rain rates and frequencies are given by S. D. Hathaway and H. W. Evans.[6] The attenuation in the 11.8-km long path can therefore be easily calculated at a frequency of 4 gc for various rain rates. The rain attenuation at 4 gc is entirely due to absorption except for very heavy rainfalls where about 1 per cent of the energy is scattered. Neglecting scattering entirely, the rain loss αl (in decibels) can be directly related to the extra noise temperature T_a picked up by the antenna of the ground station,

$$T_a = T_R(1 - t_x) \tag{7}$$

where

$$t_x = 10^{-\alpha l/10}. \tag{8}$$

It should be noted that no such simple relation exists between loss and noise temperature for the radome covered antenna because substantial scattering is provided by the wet radome.

The total noise temperature of the uncovered system at Andover would then be

$$T_{SYS} = T_{SYS_o} \text{ (at 25° elevation)} + T_a = 26°\text{K} + T_a \tag{9}$$

and the degradation becomes

$$D = \frac{\text{SNR}}{\text{SNR}_o} = t_x \frac{T_{SYS_o}(25°)}{T_{SYS}} = \frac{t_x}{1 + \dfrac{T_a}{26°\text{K}}}. \tag{10}$$

The three quantities t_x, T_{SYS} and D for the uncovered antenna are shown in Fig. 1 for comparison.

The curves show that a radome covered ground station antenna is affected considerably more by rain than an uncovered antenna.

ACKNOWLEDGMENTS

The A.T.&T. personnel at the Andover satellite ground station gathered the experimental data used in this report and Mr. I. Welber suggested the use of Cassiopeia A for the measurements. Their contributions are gratefully acknowledged.

REFERENCES

1. Gibble, D., Effects of Rain on Transmission Performance of a Satellite Communications System. Paper presented at IEEE International Convention, New York, N.Y., March 23–26, 1964.
2. Blevis, B. C., Losses Due to Rain on Radomes and Antenna Reflecting Surfaces. IEEE Trans. Antennas and Propagation, Jan. 1965, pp. 175–176.
3. Giger, A. J., Pardee, S., Jr., and Wickliffe, P. R., Jr., The Ground Transmitter and Receiver, B.S.T.J., *42*, July, 1963, pp. 1063–1107.
4. Holzer, W., Atmospheric Attenuation in Satellite Communications, Microwave J., March, 1965, pp. 119–125.
5. Rainfall Intensity-Frequency Regime, Part III—The Middle Atlantic Region, Tech. Paper No. 29, Weather Bureau, U.S. Dept. of Commerce, July, 1958.
6. Hathaway, S. D., Evans, H. W., Radio Attenuation at 11 Gc and Some Implications Affecting Relay System Engineering, B.S.T.J., *38*, Jan. 1959, pp. 73–97.