THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XLI MARCH 1962 NUMBER 2

Resistivity of Bulk Silicon and of Diffused Layers in Silicon

By JOHN C. IRVIN

(Manuscript received July 25, 1961)

Measurements of resistivity and impurity concentration in heavily doped silicon are reported. These and previously published data are incorporated in a graph showing the resistivity (at $T = 300°\mathrm{K}$) of n- and p-type silicon as a function of donor or acceptor concentration.

The relationship between surface concentration and average conductivity of diffused layers in silicon has been calculated for Gaussian and complementary error function distributions. The results are shown graphically. Similar calculations for subsurface layers, such as a transistor base region, are also given.

I. INTRODUCTION

A diffused layer in silicon is generally characterized by four parameters: the concentration, C_s, of diffused donors or acceptors at the surface, the concentration, C_B, of acceptors or donors originally in the material (background concentration), the depth, x_j, of the resultant junction, and the sheet resistivity, ρ_s, of the layer. A knowledge of the relationship between these parameters is essential to the establishment of device processing recipes, the evaluation of diffusion techniques, and investigations of the thermodynamic properties of silicon.

The desired relationship may be readily calculated, given a knowledge of the distribution of the diffused impurities, the variation of the resistivity of n- and p-type silicon with donor or acceptor density, and a fast electronic computer. The results of such a computation were first made generally available three years ago, in the form of curves relating C_s to $1/\rho_s x_j$ for a given C_B, for n- and for p-type layers in silicon, and for several common distributions.[1] Recent calculations, however, based on new and more extensive silicon resistivity data, have indicated considerable error in the earlier results. Thus a comprehensive recomputation has been undertaken, the outcome of which is presented herewith.

A necessary adjunct to the calculation is an accurate knowledge of the resistivity of n- and p-type silicon with varying dopant concentration. To this end, most of the extant data have been reviewed and supplemented here and there with some new determinations. The results of this search are also presented here.

II. THE RESISTIVITY OF SILICON AS A FUNCTION OF IMPURITY CONCENTRATION

The variation of the resistivity of silicon at 300°K as a function of the concentration of acceptors or donors is shown in Fig. 1. This graph represents the author's judgment of a most reasonable compromise to

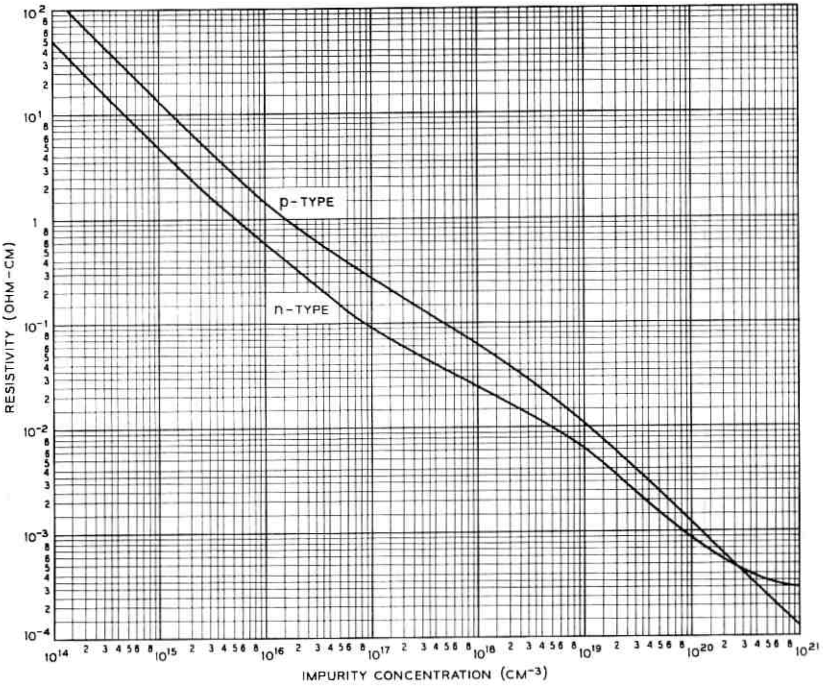


Fig. 1 — Resistivity of silicon at 300°K as a function of acceptor or donor concentration.

TABLE I — RESISTIVITIES AND IMPURITY CONCENTRATIONS IN SILICON (T = 300°K)

Resistivity (ohm-cm)	Impurity	Impurity Concentration (cm^{-3})	Carrier Concentration (cm^{-3})
0.00076	B	1.66×10^{20}	
0.00089	B	1.41×10^{20}	
0.0010	B		1.49×10^{20}
0.0010	B	1.12×10^{20}	
0.0012	B	1.04×10^{20}	
0.0011	B	1.12×10^{20}	
0.0014	B	9.23×10^{19}	
0.0013	B	8.84×10^{19}	
0.0067	B	1.43×10^{19}	
0.0073	B	1.43×10^{19}	
0.013	B	7.41×10^{18}	
0.014	B	7.03×10^{18}	
0.00095	As	1.80×10^{20}	
0.00094	As	1.86×10^{20}	
0.00094	As		1.1×10^{20}
0.00093	As	1.87×10^{20}	
0.00094	As	1.97×10^{20}	
0.00088	As	2.10×10^{20}	
0.00088	As	2.19×10^{20}	
0.00089	As		1.1×10^{20}
0.00083	As	2.30×10^{20}	
0.00083	As	2.20×10^{20}	
0.00080	As	2.46×10^{20}	
0.00082	As	2.44×10^{20}	

the mass of available and not altogether compatible data on the subject. These data include most of the previously published work (Refs. 3–12), recent, unpublished results kindly provided by other investigators,[2,13] as well as some measurements obtained expressly for the present study.

The last data are shown in Table I. The crystals involved were pulled from quartz crucibles, and hence can not be expected to be particularly low in oxygen content. After dissolution of the boron-doped crystals and separation of the dopant,[14] boron concentrations were determined by a photometric carmine technique essentially similar to published methods.[15] Arsenic concentrations were measured by gamma-ray spectrometry after pile neutron activation. Resistivity measurements were done with a four-point probe. In the case of a few samples, resistivity and carrier concentration were measured in Hall-effect apparatus (where it was assumed $\mu_H/\mu = 1$).

Drawing curves through these many points was accomplished by a succession of smoothing procedures, which were primarily visual. 75 per cent of the data points deviate less than 10 per cent from the curves thus obtained, both for the p-type and the n-type cases. The uncertainty is greatest in the degenerate region. For p-type silicon, suitable data become scarce at dopings greater than 10^{19} cm^{-3}, and none are available beyond 3×10^{20} cm^{-3}. For n-type material, there is an abundance of rather conflicting data representing donor concentrations between 10^{19} cm^{-3} and 6×10^{20} cm^{-3}. In this region a 10 per cent variation in the chosen line still includes 67 per cent of the data, however.

A single pair of curves obviously can not characterize with the same degree of accuracy all silicon material, regardless of dopant employed or degree of compensation. However, over the range 10^{14} cm^{-3} $\leqq N_I \leqq$ 10^{20} cm^{-3}, and subject to the limitations discussed below, Fig. 1 is considered to be within 10 per cent of reality. This graph refers specifically to uncompensated silicon containing a donor or acceptor impurity concentration, N_I, consisting of arsenic, phosphorus, or antimony for n-type, and aluminum, boron, or gallium for p-type material. (Actually, even among samples doped with the aforementioned impurities, small but consistent differences in carrier concentration and mobility, depending on the specific choice of donor or of acceptor, have been reported recently for silicon in the 0.001 ohm-cm region.[10,12]) In case of moderate compensation, the net impurity density, $|N_A - N_D|$, should be used for N_I. However, heavy compensation requires allowance for the added impurity scattering.

For impurity densities near or greater than 10^{20} cm^{-3}, Fig. 1 can not be considered very reliable. At such concentrations, impurity band conduction is prominent and its effects are apt to differ appreciably depending on choice of impurity. Even more serious are the degrees of impurity precipitation and lattice imperfection which occur in highly doped material and which furthermore vary with growth conditions and history of the crystal. It will be noted with some consternation that the p-type and n-type curves are shown to cross near $N_I = 3 \times 10^{20}$ cm^{-3}. The paucity of data, of course, casts considerable doubt on this result. However, for what they are worth, such are the indications. Perhaps this can be understood in light of the acceptor action of imperfections, especially vacancies, which are abundant in very highly doped material.

The calculations discussed in the remainder of this paper require a mathematical representation of Fig. 1. Straight-line approximations of the form $(1/\rho) = BN_I^{\alpha}$ have been obtained, which depart 10 per cent from the desired curve at the turning points and rapidly approach coincidence elsewhere. The parameters B and α are listed in Table II for the respective straight-line regions.

III. DIFFUSION PROFILES AND CALCULATIONS

The diffusion profiles of current practical interest are the complementary error function, $C_x = C_s \operatorname{erfc}(x/2\sqrt{Dt})$, and the Gaussian,

TABLE II — VALUES OF B AND α IN THE EQUATION $(1/\rho) = BN_I^{\alpha}$, REPRESENTING STRAIGHT-LINE APPROXIMATIONS TO THE ρ VS N_I CURVES OF n-TYPE AND p-TYPE SILICON (T = 300°K)

Region (cm^{-3})	B	α
n-type		
$2.35 \times 10^{20} \leqq N_D$	1.04×10^{-6}	0.456
$6.00 \times 10^{19} \leqq N_D \leqq 2.35 \times 10^{20}$	1.43×10^{-12}	0.744
$9.50 \times 10^{18} \leqq N_D \leqq 6.00 \times 10^{19}$	2.00×10^{-16}	0.940
$1.00 \times 10^{17} \leqq N_D \leqq 9.50 \times 10^{18}$	6.93×10^{-9}	0.543
$3.50 \times 10^{15} \leqq N_D \leqq 1.00 \times 10^{17}$	6.97×10^{-14}	0.837
$N_D \leqq 3.50 \times 10^{15}$	2.00×10^{-16}	1.000
p-type		
$1.50 \times 10^{19} \leqq N_A$	4.00×10^{-17}	0.966
$2.40 \times 10^{18} \leqq N_A \leqq 1.50 \times 10^{19}$	1.47×10^{-14}	0.832
$1.50 \times 10^{16} \leqq N_A \leqq 2.40 \times 10^{18}$	3.30×10^{-11}	0.650
$N_A \leqq 1.50 \times 10^{16}$	7.20×10^{-17}	1.000

$C_x = C_s \exp(-x^2/4Dt)$. In these expressions, x, D, and t are the depth, diffusion coefficient (assumed independent of impurity density), and time, respectively. C_x is the concentration of the diffused impurity at depth x and C_s, that at the surface. The former distribution is expected when diffusion takes place with the surface concentration C_s held constant; the latter when the total impurity diffusing is constant. Unfortunately it must be admitted that the accuracy of these expectations is open to question in some situations.[2,16] Also, precipitation and compensation of impurities near the surface may further distort the distribution. However, it is still useful to solve the problem under these assumptions, leaving corrections for later determination.

The "average conductivity" of a diffused layer (which throughout this paper is assumed to be diffused into a silicon slice of opposite conductivity type and uniform doping C_B) is given by the expression

$$\bar{\sigma} = 1/\rho_s x_j = (1/x_j) \int_0^{x_j} q\mu C \, dx$$

where q is electronic charge, μ the carrier mobility typical of a total ionized impurity density of $C_x + C_B$, $C = r(C_x - C_B)$ is the density of carriers, r being the fraction of uncompensated diffused impurity atoms which are ionized, and C_x the total density of diffused impurity atoms at depth x. (Possible variation of the mobility as a function of the proximity of the surface is a hazard which should be recognized in passing but is otherwise ignored in the present calculation.) Multiplying and dividing within the integrand by r' $(C_x + C_B)$, where r' is the ionized fraction associated with an uncompensated dopant density of $(C_x + C_B)$, and writing

$$q\mu r'(C_x + C_B) = \sigma_{(Cx+CB)} = B(C_x + C_B)^\alpha$$

the average conductivity becomes

$$\bar{\sigma} = (1/x_j) \int_0^{x_j} (r/r')(C_x - C_B)B(C_x + C_B)^{\alpha-1}\, dx.$$

Now (r/r') represents the ratio of degrees of ionization corresponding to $C_x - C_B$ and $C_x + C_B$ respectively. This ratio is very nearly unity unless C_x and C_B are comparable in magnitude. Such is the case only for the lamina nearest the junction, which contributes negligibly to the conductance of the whole layer. Hence, (r/r') may be justifiably taken as equal to unity, and writing $C_x = C_S f(x)$, where $f(x)$ depends on the profile of interest,

$$\bar{\sigma} = (1/x_j) \int_0^{x_j} [C_S f(x) - C_B] B[C_S f(x) + C_B]^{\alpha-1}\, dx.$$

A program for the evaluation of this expression has been devised previously by others and employed in the analysis of diffused layers in germanium.[17] With slight additions to facilitate automatic plotting, the same program has been used in the present work. Computations were performed on an IBM 704, and plotting of points was carried out with an Electronic Associates Variplotter.

IV. PRESENTATION OF RESULTS

Of frequent interest in transistor design and in the analysis of diffused layers, are the characteristics of a "subsurface" layer such as illustrated in Fig. 2. This layer, bounded on one side by the junction and on the other by a plane paralleling the junction at depth x, may be characterized by an average conductivity

$$\bar{\sigma} = 1/[\rho_S'(x_j - x)] = \frac{1}{(x_j - x)} \int_x^{x_j} q\mu C\, dx$$

where ρ_s' is the sheet resistance of the subsurface layer. It will be recognized that the base region of a diffused-base, alloyed-emitter transistor is an example of a subsurface layer. Another example is that portion of a diffused layer remaining after removing the top strata of depth x. Here, however, it must be remembered that the value of C_s specifying this layer pertains to the original surface at $x = 0$.

Since a subsurface layer becomes the entire diffused layer when $x = 0$, it is convenient to display the properties of both in the same plot by introducing the parameter (x/x_j). On pages 394 to 410 such graphs are presented for n- and p-type diffused layers of Gaussian and complementary error function profile. Each graph contains the family of ten curves $(x/x_j) = 0, 0.1, \cdots, 0.9$, and relates the average conductivity of

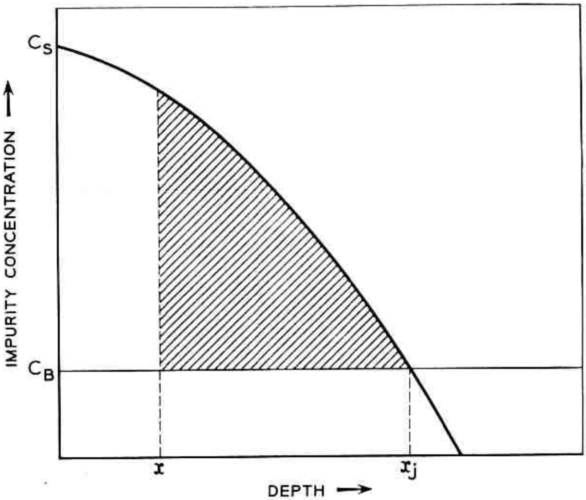


Fig. 2 — Profile of a diffused layer with subsurface layer shaded.

each layer to the surface concentration (at the *original* surface) for a given value of C_B. A separate graph is required for each value of C_B, which in the present work ranges from 10^{14} cm^{-3} to 10^{20} cm^{-3} at one-decade intervals. In each plot the range of surface concentrations spanned is from C_B to 10^{21} cm^{-3}. The so-called "Backenstoss" curve for a particular C_B is simply the right-most line ($x/x_j = 0$) in each graph.

The wiggle in the n-type average conductivity for diffusant concentrations near 10^{19} cm^{-3} is ascribable to the rather large change in slope occurring in the n-type resistivity plot at $N_I = 10^{19}$ cm^{-3}.

V. ACKNOWLEDGMENTS

It is a pleasure to acknowledge the contributions of many colleagues to this investigation. In particular, the boron determinations, devised and performed by C. L. Luke, were very valuable, as were the radioactive arsenic determinations of Miss K. Wolfstirn and Hall-effect measurements of R. A. Logan and R. L. Johnston. The author is indebted to D. Lassota for the growing of the boron-doped crystals, to D. B. Cuttriss, whose efforts brought about the preparation of the computer program, to Mrs. W. Mammel, for subsequent additions to it, to R. Lilienthal for various measurements, and especially to Mrs. M. S. Boyle for her indefatigable assistance in many measurements and the plotting of hundreds of curves.

REFERENCES

1. Backenstoss, G., B.S.T.J., **37,** 1958, p. 699.
2. Tannenbaum, Eileen, Solid-State Electronics, **2,** 1961, p. 123.

3. Backenstoss, G., Phys. Rev., **108,** 1957, p. 416.
4. Carlson, R. O., Phys. Rev., **100,** 1955, p. 1975.
5. Morin, F. J., and Maita, J. P., Phys. Rev., **96,** 1954, p. 28.
6. Ludwig, G. W., and Watters, R. L., Phys. Rev. **101,** 1956, p. 1699.
7. Long, D., Motchenbacher, C. D., and Meyers, J., J. Appl. Phys., **30,** 1959, p. 353.
8. Prince, M. B., Phys. Rev., **93,** 1954, p. 1204.
9. Wolfstirn, K., J. of Phys. & Chem. of Solids, **16,** 1960, p. 279.
10. Logan, R. A., Gilbert, J. F., and Trumbore, F. A., J. Appl. Phys., **32,** 1961, p. 131.
11. Esaki, L., and Miyahara, Y., Solid-State Electronics, **1,** 1960, p. 13.
12. Furukawa, Y., J. Phys. Soc. Japan, **16,** 1961, p. 577.
13. Carlson, R. O., private communication.
14. Luke, C. L., and Flaschen, S. S., Anal. Chem., **30,** 1958, p. 1406.
15. Hatcher, J. T., and Wilcox, L. V., Anal. Chem., **22,** 1950, p. 567.
16. Subashiev, V. K., Landsman, A. P., and Kukharskii, A. A., Soviet Physics-Solid State, **2,** 1961, p. 2406.
17. Cuttriss, D. B., B.S.T.J., **40,** 1961, p. 509.

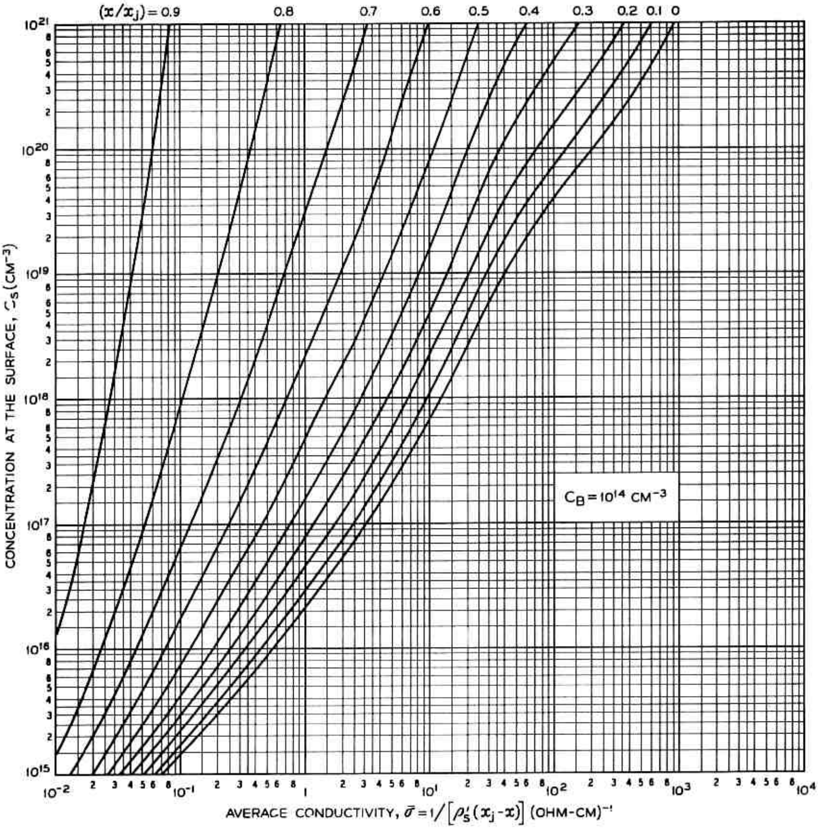


Fig. 3 — Average conductivity of n-type complementary error function layers in silicon.

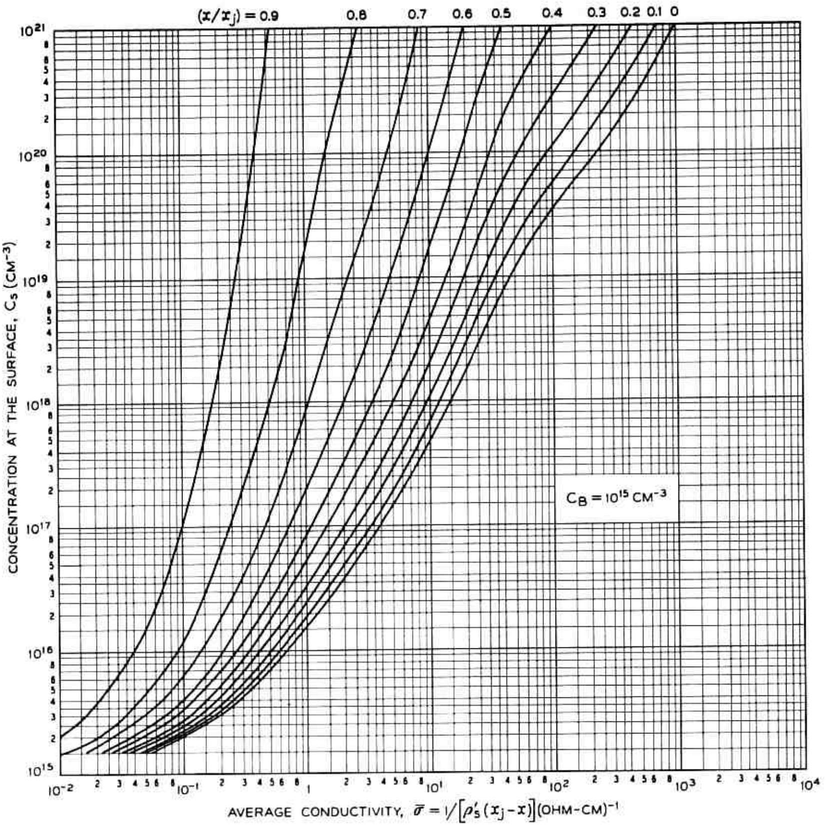


Fig. 3 (cont.) — Average conductivity of n-type complementary error function layers in silicon.

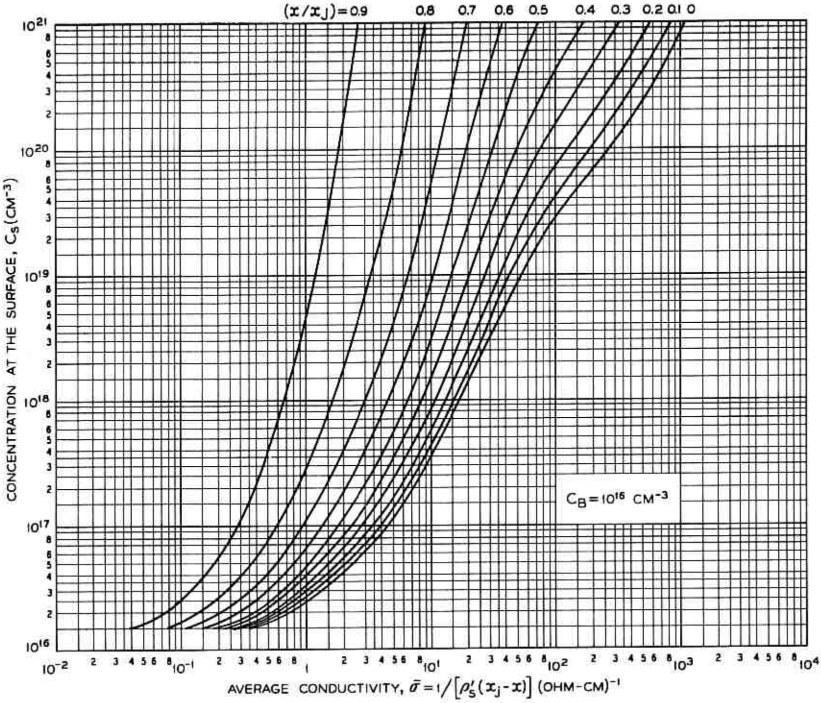


Fig. 3 (cont.) — Average conductivity of n-type complementary error function layers in silicon.

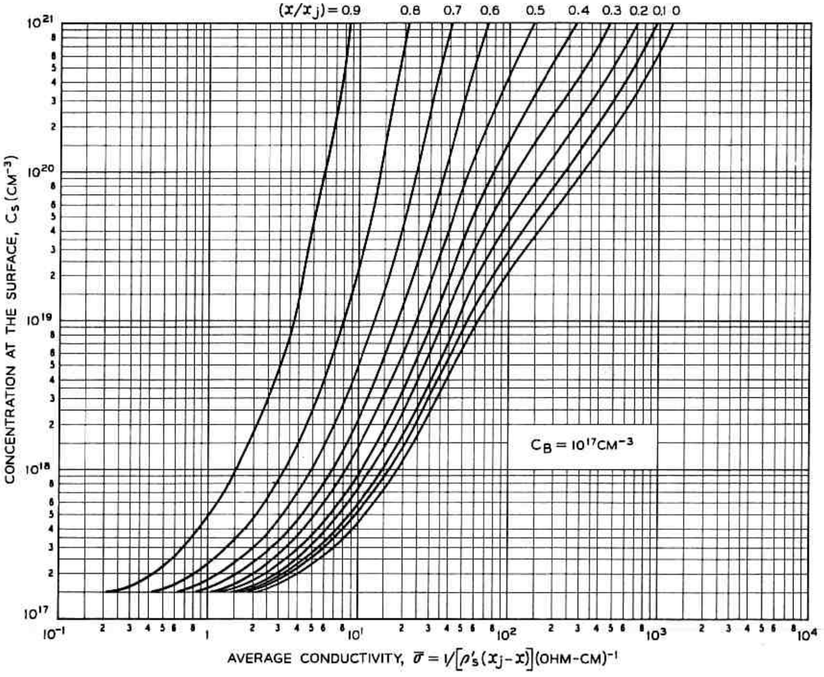


Fig. 3 (cont.) — Average conductivity of n-type complementary error function layers in silicon.

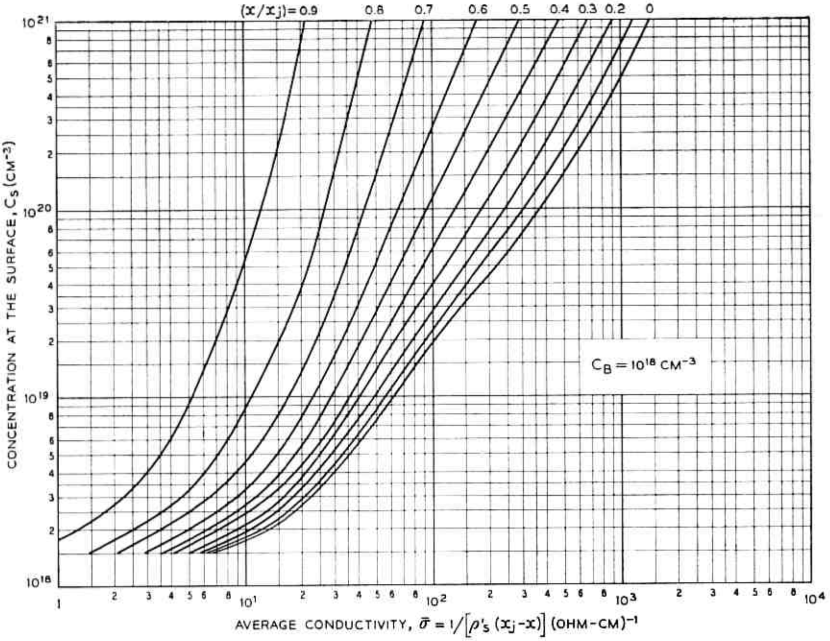


Fig. 3 (cont.) — Average conductivity of n-type complementary error function layers in silicon.

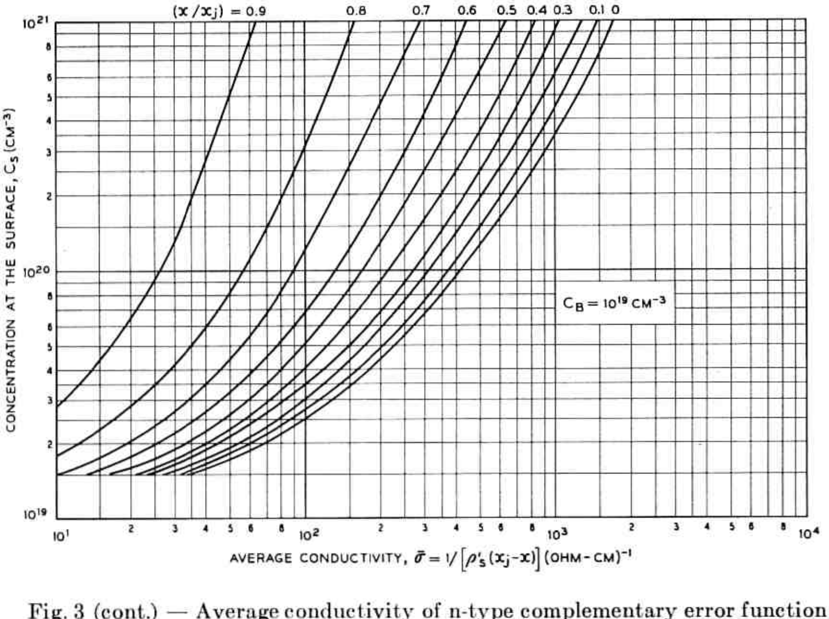


Fig. 3 (cont.) — Average conductivity of n-type complementary error function layers in silicon.

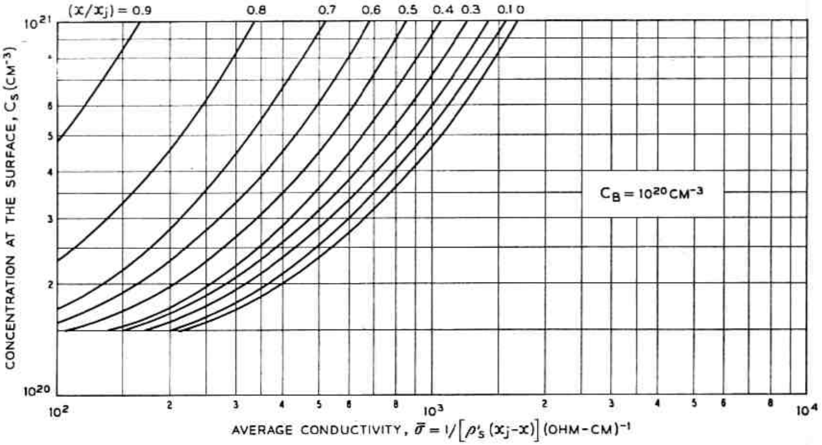


Fig. 3 (cont.) — Average conductivity of n-type complementary error function layers in silicon.

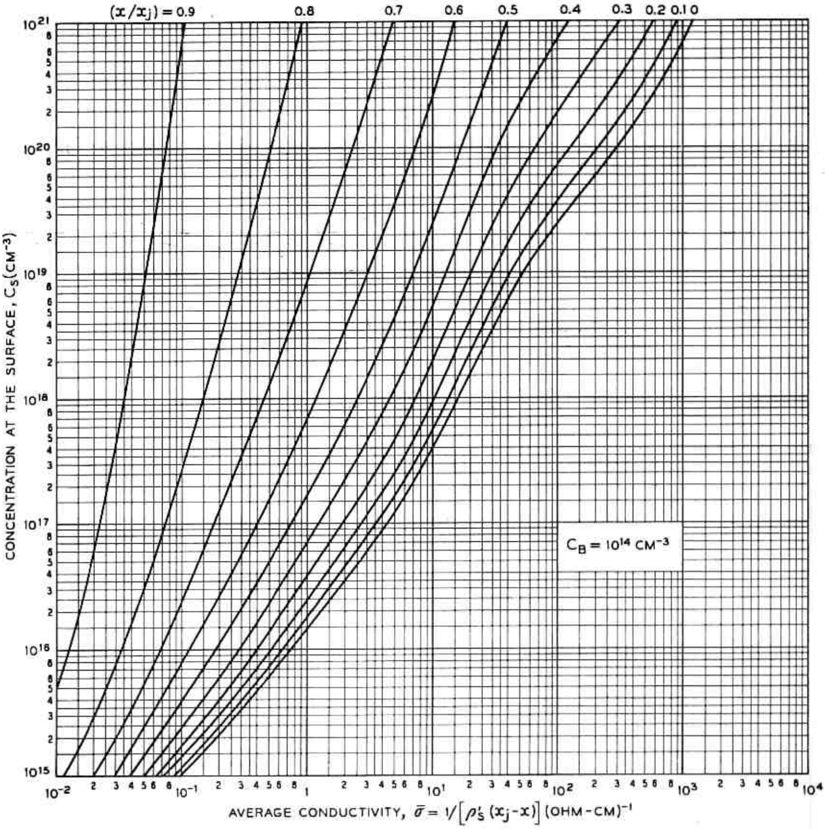


Fig. 4 — Average conductivity of n-type Gaussian layers in silicon.

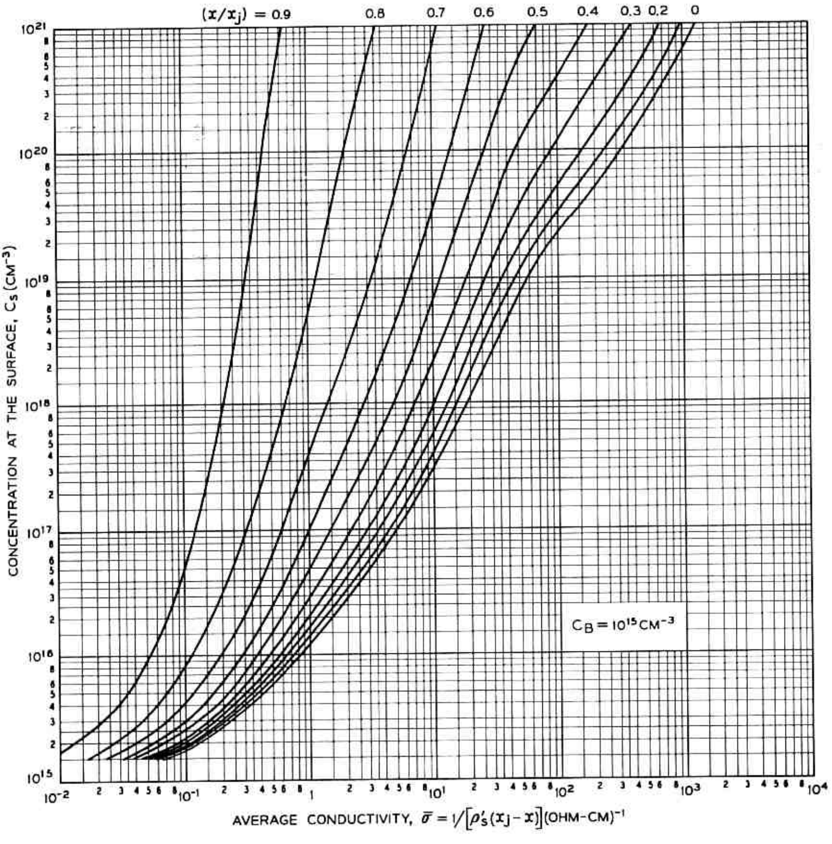


Fig. 4 (cont.) — Average conductivity of n-type Gaussian layers in silicon.

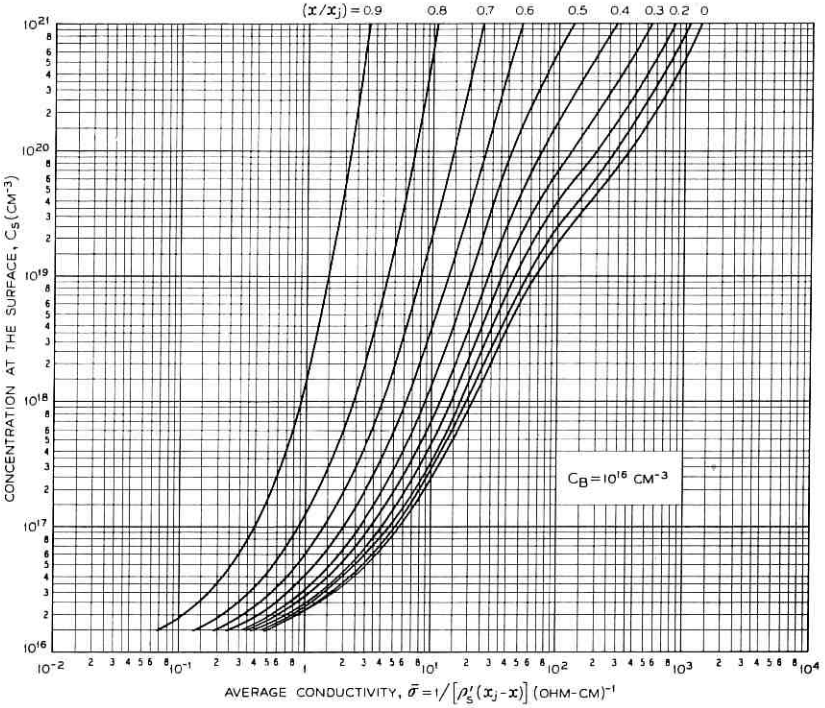


Fig. 4 (cont.) — Average conductivity of n-type Gaussian layers in silicon.

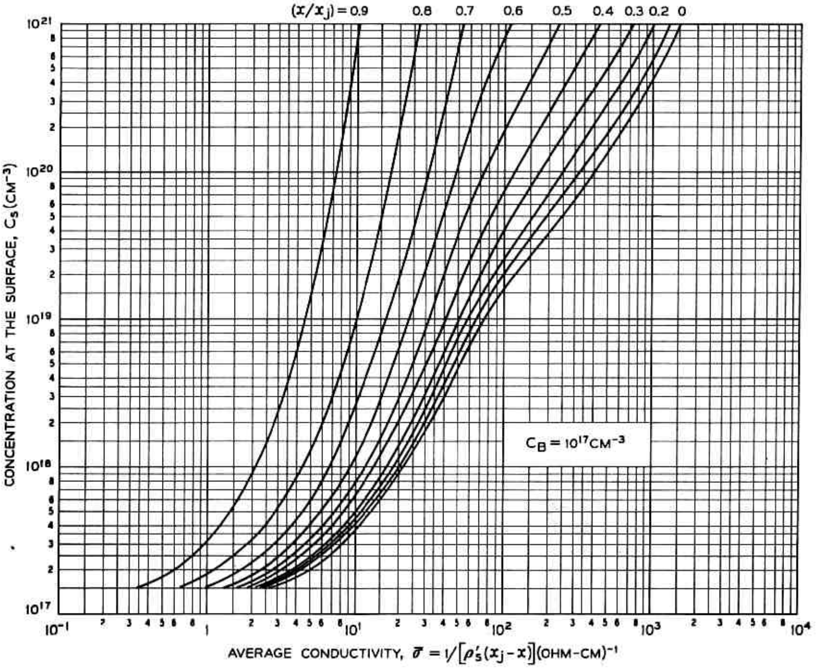


Fig. 4 (cont.) — Average conductivity of n-type Gaussian layers in silicon.

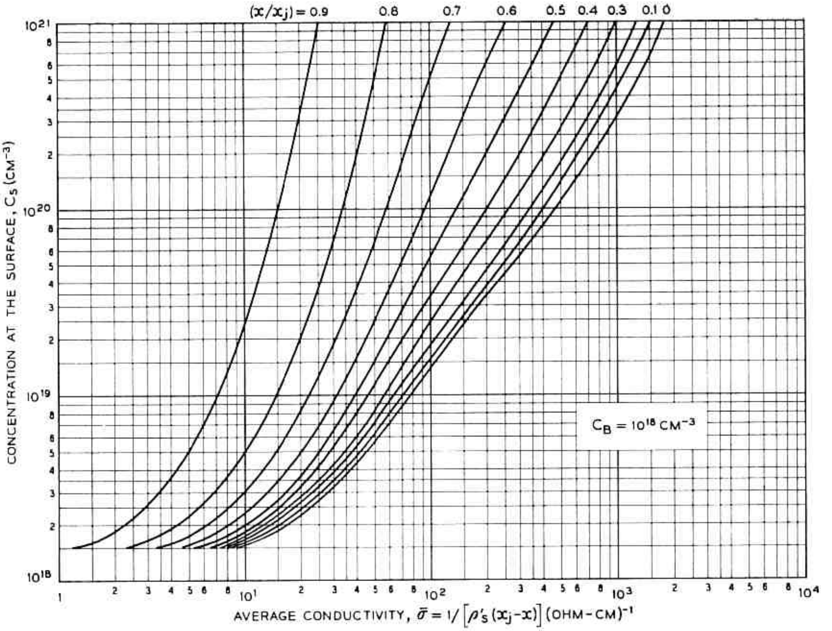


Fig. 4 (cont.) — Average conductivity of n-type Gaussian layers in silicon.

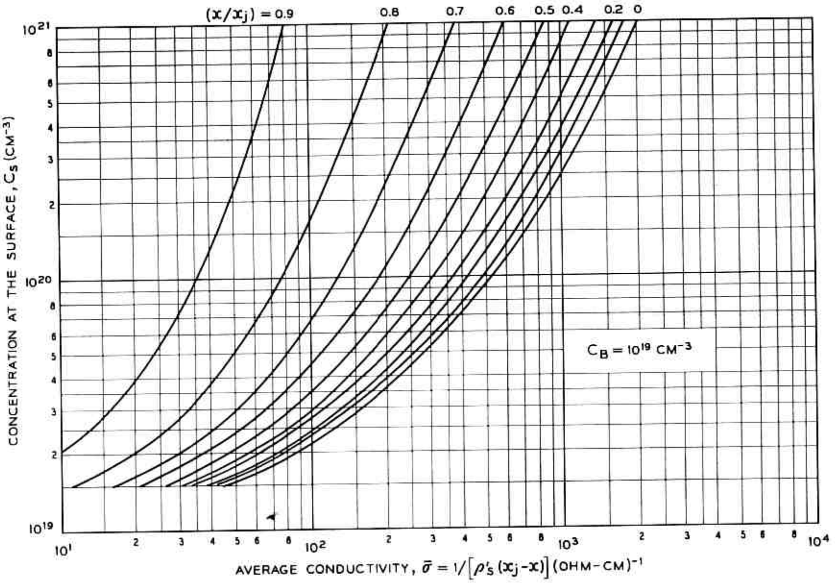


Fig. 4 (cont.) — Average conductivity of n-type Gaussian layers in silicon.

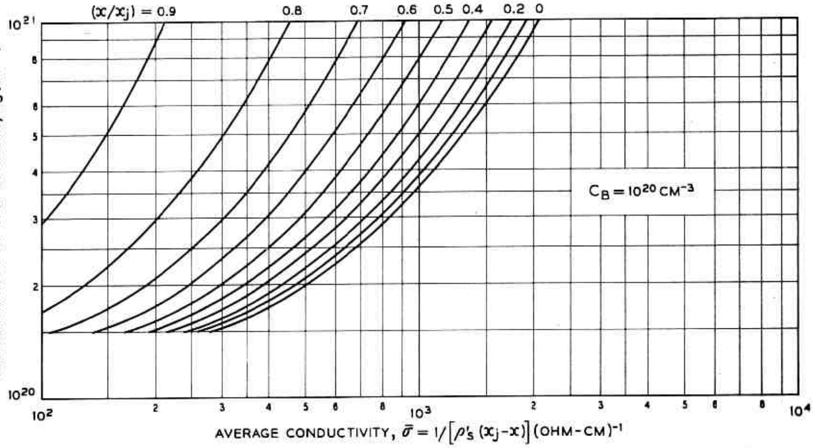


Fig. 4 (cont.) — Average conductivity of n-type Gaussian layers in silicon.

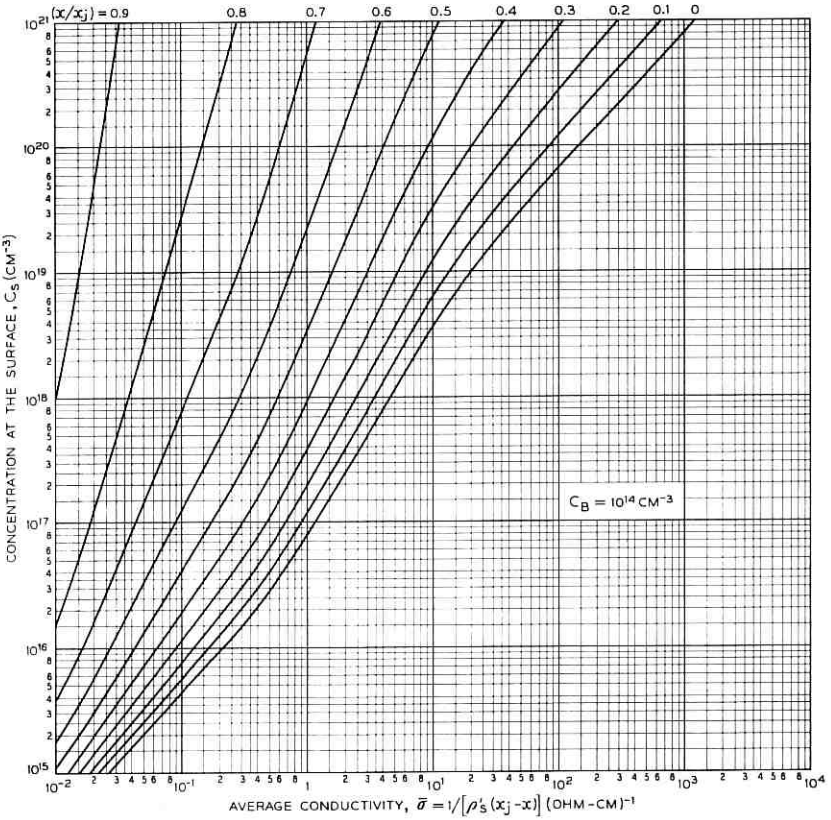


Fig. 5 — Average conductivity of p-type complementary error function layers in silicon.

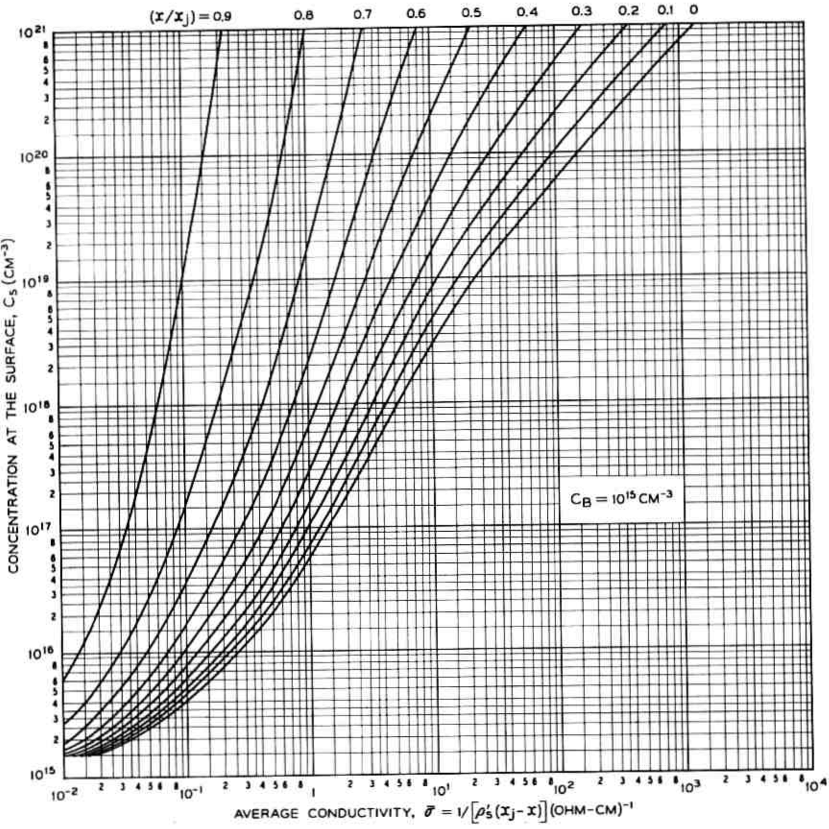


Fig. 5 (cont.) — Average conductivity of p-type complementary error function layers in silicon.

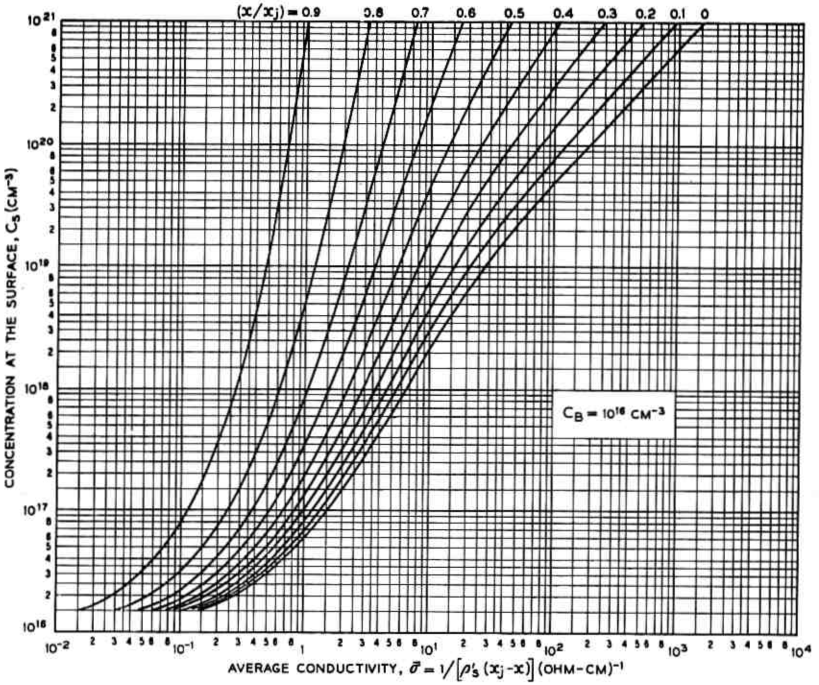


Fig. 5 (cont.) — Average conductivity of p-type complementary error function layers in silicon.

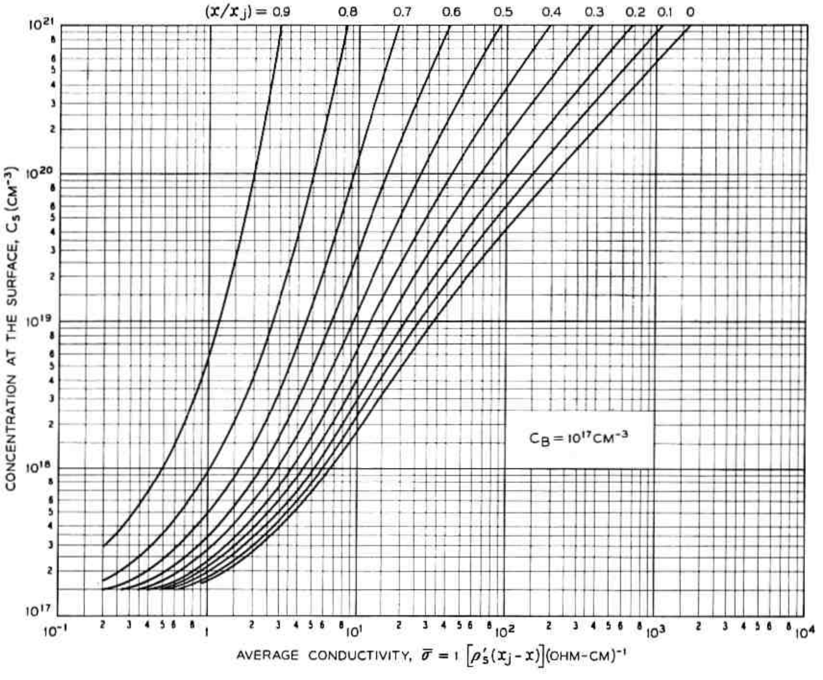


Fig. 5 (cont.) — Average conductivity of p-type complementary error function layers in silicon.

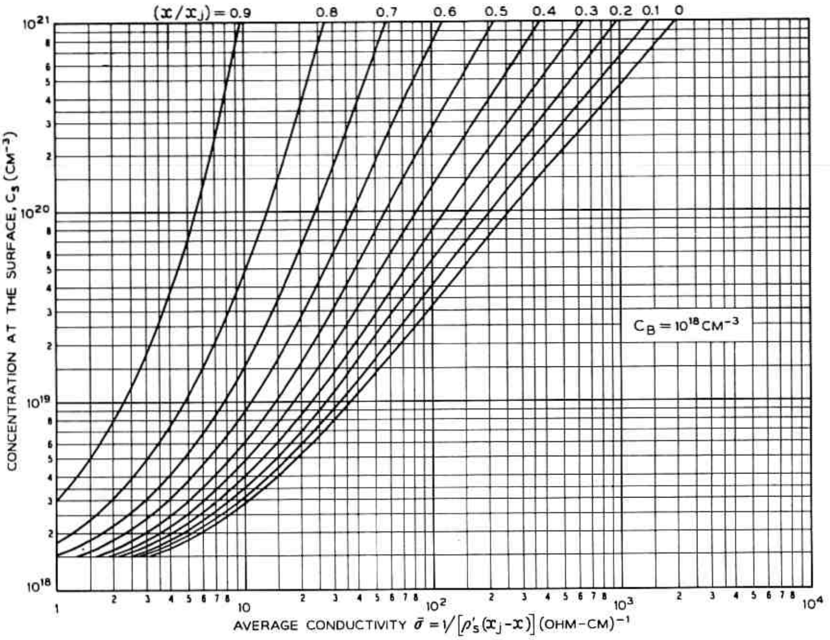


Fig. 5 (cont.) — Average conductivity of p-type complementary error function layers in silicon.

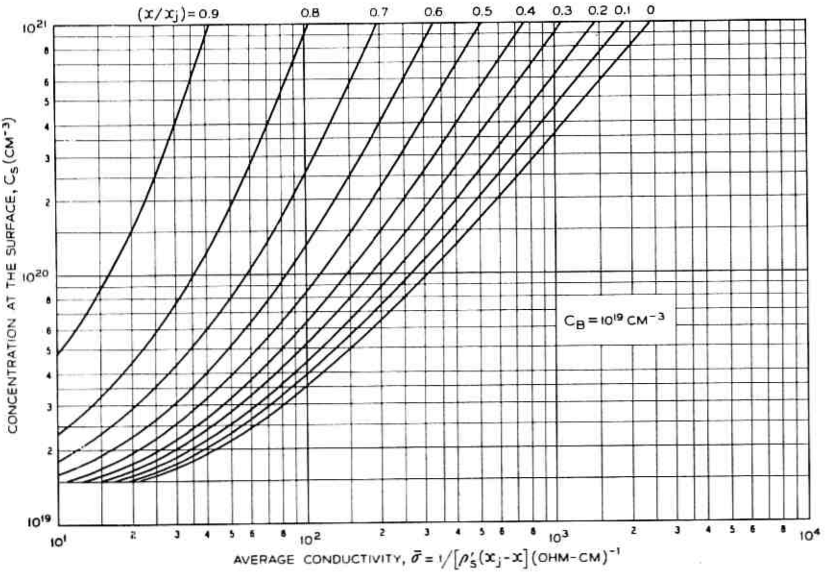


Fig. 5 (cont.) — Average conductivity of p-type complementary error function layers in silicon.

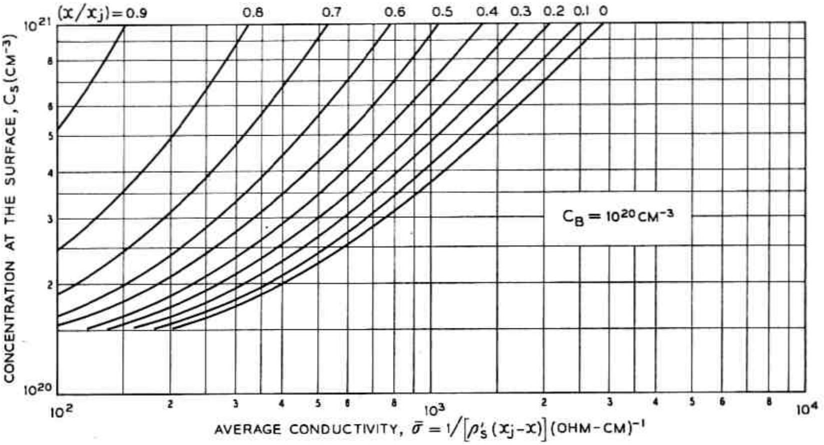


Fig. 5 (cont.) — Average conductivity of p-type complementary error function layers in silicon.

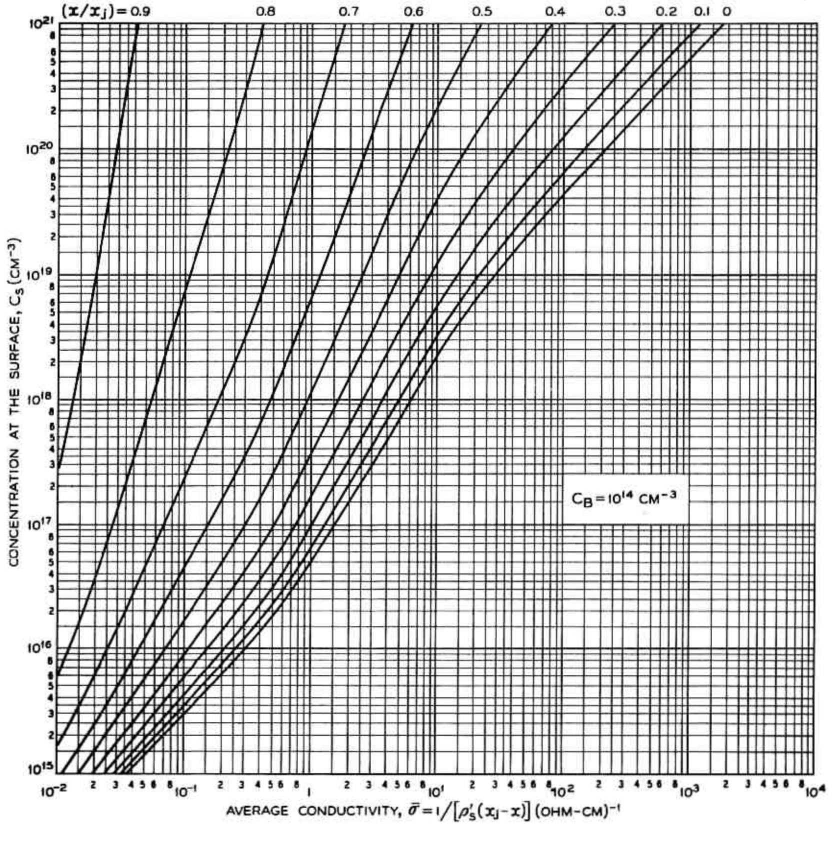


Fig. 6 — Average conductivity of p-type Gaussian layers in silicon.

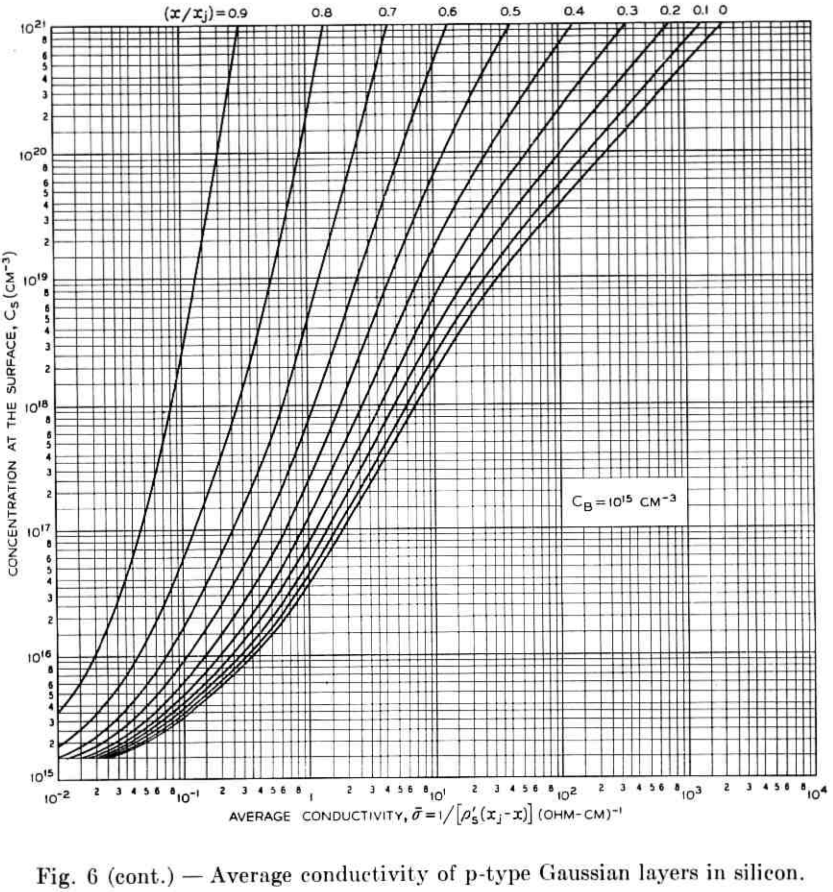


Fig. 6 (cont.) — Average conductivity of p-type Gaussian layers in silicon.

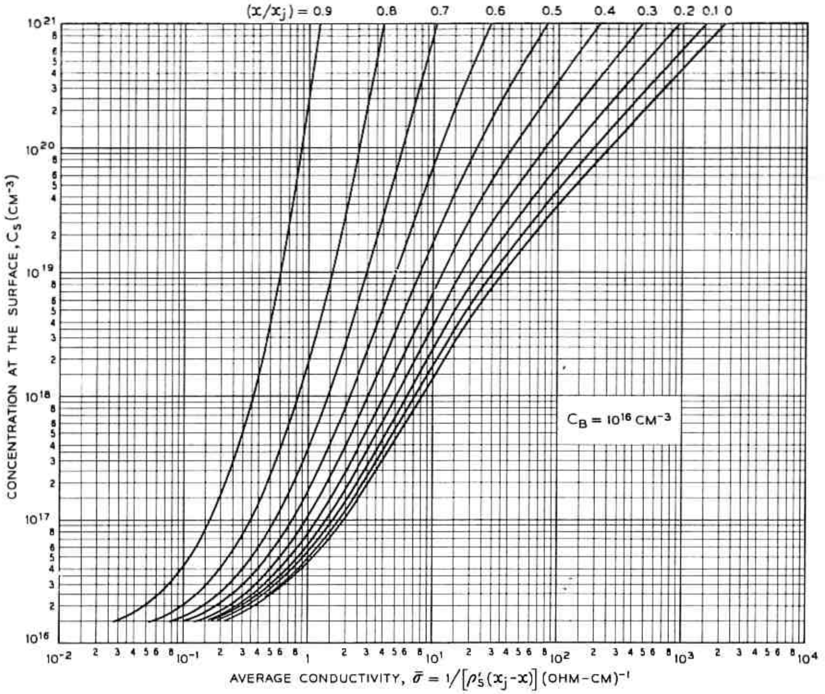


Fig. 6 (cont.) — Average conductivity of p-type Gaussian layers in silicon.

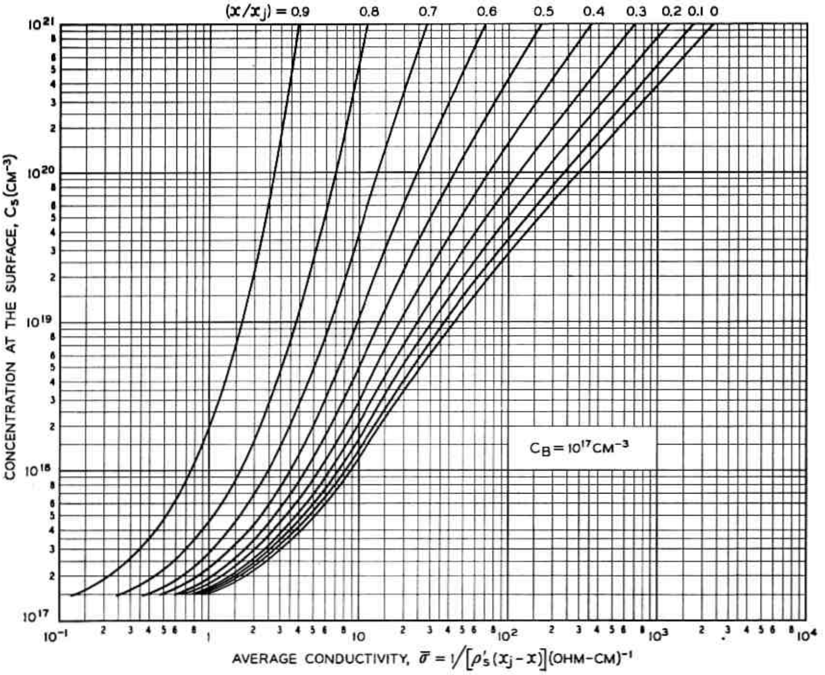


Fig. 6 (cont.) — Average conductivity of p-type Gaussian layers in silicon.

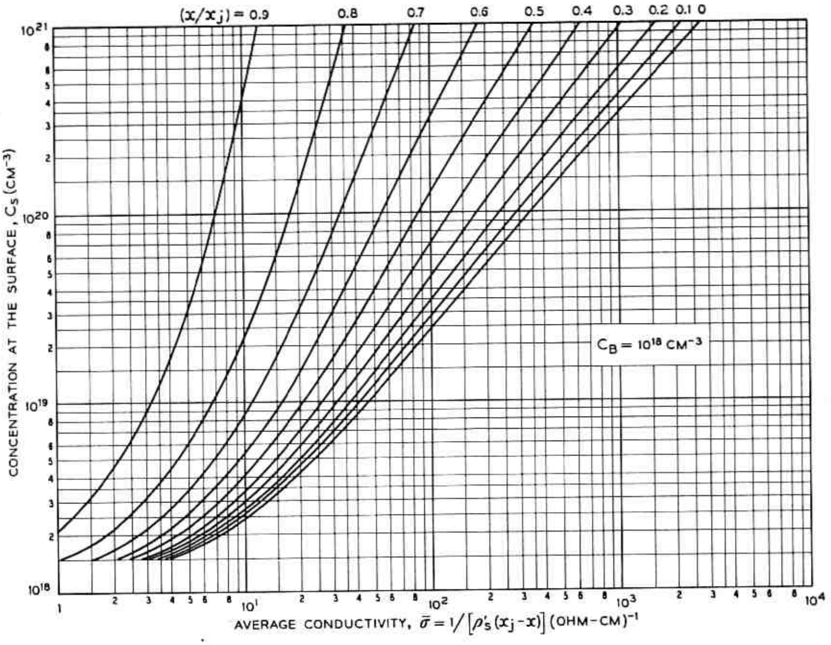


Fig. 6 (cont.) — Average conductivity of p-type Gaussian layers in silicon.

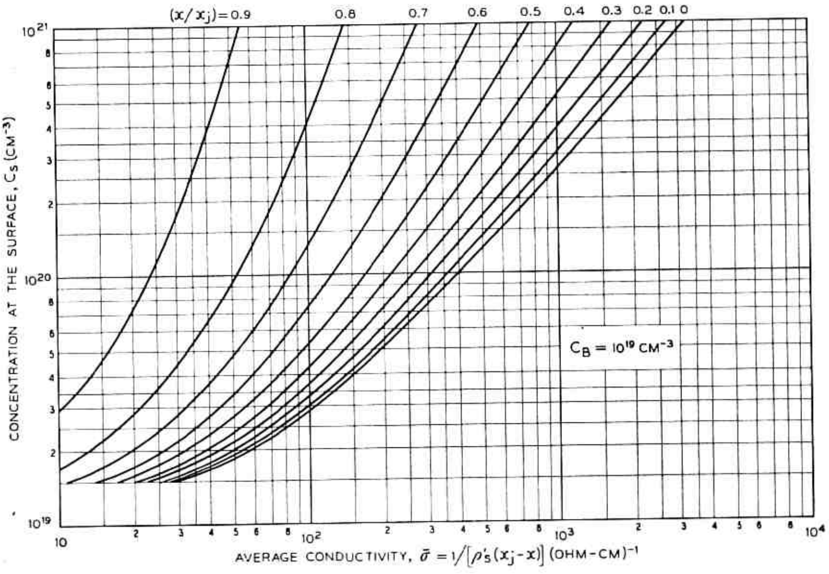


Fig. 6 (cont.) — Average conductivity of p-type Gaussian layers in silicon.

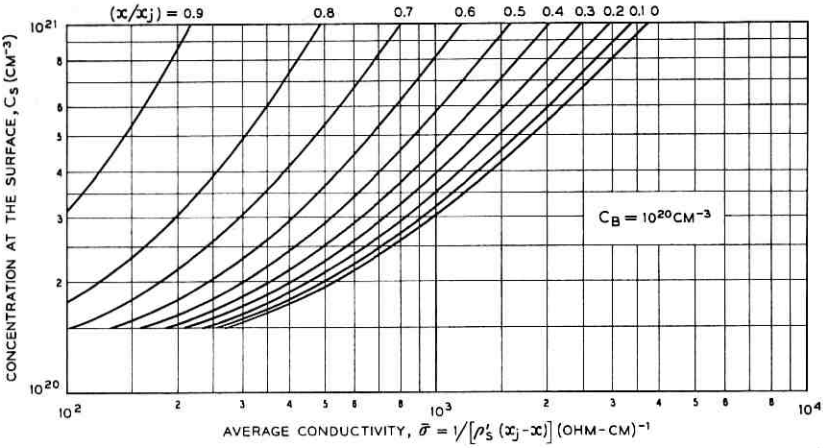


Fig. 6 (cont.) — Average conductivity of p-type Gaussian layers in silicon.

A Miniature Tuned Reed Selector of High Sensitivity and Stability

By L. G. BOSTWICK

(Manuscript received August 23, 1961)

This paper describes a selective contacting device that is responsive only to sustained frequencies in a discrete narrow band and is insensitive to speech and noise interference. It is of small size suitable for use in a pocket-carried radio receiver and is sufficiently stable to permit 33 discrete resonant frequencies, spaced 15 cycles apart, in less than an octave between 517.5 and 997.5 cycles per second. It has a threshold sensitivity of about 35 microwatts and other operating characteristics that are essential in large capacity systems.

I. INTRODUCTION

Tuned reed selectors used as selective receivers in multifrequency systems involving large numbers of individual selections, such as personal radio signaling,[1] must operate within close and specifiable limits in order to avoid false signaling and to assure satisfactory performance under devious environmental and circuit conditions. In particular, three operating characteristics, or their equivalents, must be controlled, namely: the resonant frequency, the sensitivity (current or power needed at the most sensitive frequency), and the bandwidth (the frequency band in which contacting occurs with an input power twice that needed at the most sensitive frequency).

The permissible variation in these characteristics is much smaller than would seem necessary from first considerations. Resonant frequency changes that seem negligible compared to the frequency spacing between adjacent selectors often become important when other system requirements are considered simultaneously. For example, the frequency range over which contacting will occur depends upon the electrical input level and the selector bandwidth. Consequently, feasible limits for both of these latter quantities must be considered, and in determining allowable frequency deviations from nominal, the lowest probable input level and the narrowest bandwidth must be taken into account. On the other hand, excessively high input levels cannot be allowed even in those unusual instances where conserving power is unimportant, because this necessitates wider channel separations in order to avoid transient operation of adjacent selectors, particularly those having high sensitivities. Furthermore, high input levels result in longer decay times, which often cannot be tolerated. When these and other related factors are considered and the widest manufacturing tolerances are sought, it is found that the above three selector characteristics are closely interrelated, and one cannot be relaxed without making one or both of the others more stringent.

The tuned reed selectors described in this paper have factory adjustment provisions and sufficient structural stability to control in a practical manner the resonant frequencies, the sensitivities and the bandwidths within adequate and compatible limits. As a result, it is feasible to use 33 discrete resonant frequencies, 15 cycles apart, in less than an octave between 517.5 and 997.5 cycles. An available electrical power of 35 microwatts at each individual resonant frequency will just operate the contact, and a power of 100 microwatts will close the contact to a low resistance over 20 per cent or more of the reed period. These and other capabilities to be described distinguish these selectors from many others that are not adequate for reliable operation in large systems.

II. GENERAL DESCRIPTION

Fig. 1 is a photograph showing one complete reed selector with the outside shell removed. Fig. 2 is a partially exploded view showing the subassemblies and indicating how the parts are fitted together. The shell is formed from permalloy sheet; it serves as an effective shield from extraneous fields and as a high-permeability flux path for the internal magnetic circuit. All parts are electrically insulated from the shell. The complete selector weighs about 8 grams.

As shown in these photographs, a tuning fork formed from two reeds brazed to a base block serves as the resonant element. This balanced type of structure does not require a massive support as would a single cantilever reed in order to isolate it from extraneous influences, an important matter for a miniature device. This fork is freely supported within the shell by a compliant frame that further isolates any small residual vibration of the fork base from the rest of the selector, and yet is sufficiently stable to permit the vibrating contact on the end of the tuning-fork tine to be precisely positioned with respect to the stationary contact. This latter contact is carried by a loop of wire spot-welded to a rotatable stud that fits into a tapered hole in an insulating bushing in

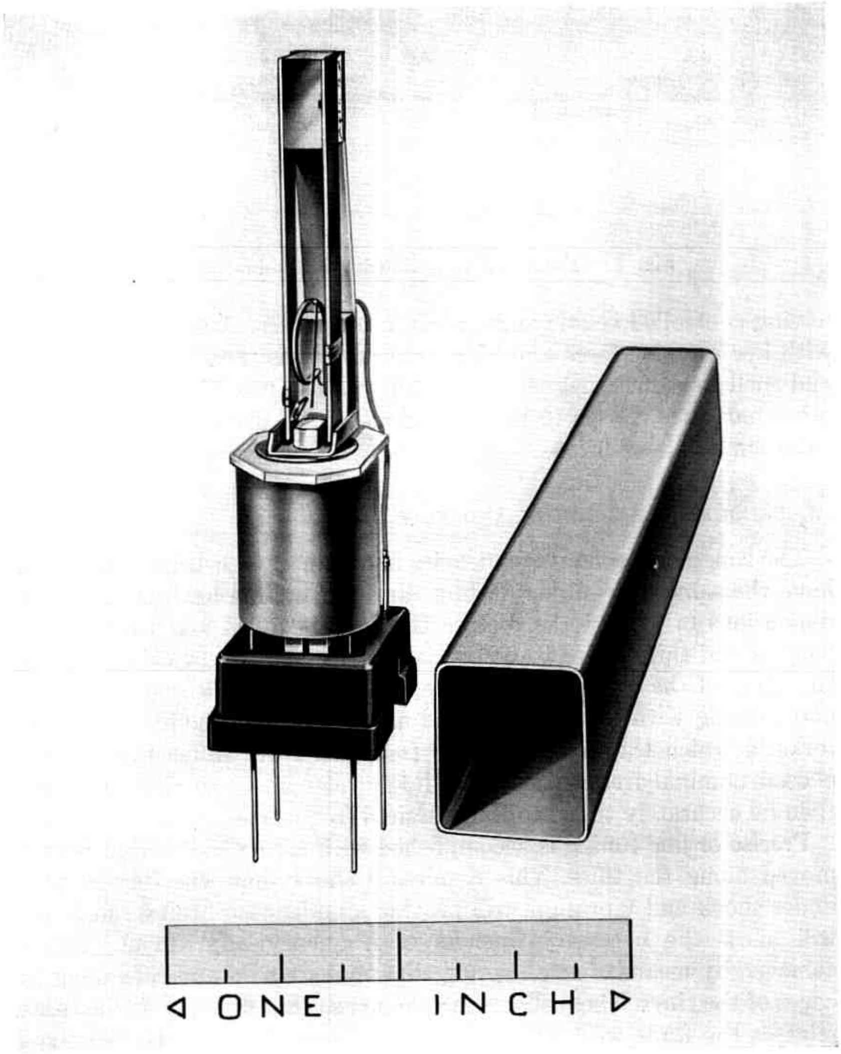


Fig. 1 — Tuned reed selector with shell removed.

the frame between the tines. A magnetic polepiece is positioned between the open ends of the tines, forming two equal gaps. Polarizing magnetic flux is set up in these gaps by a small permanent magnet attached to the opposite end of the polepiece. The energizing coil surrounds the center portion of the polepiece.

The tuning fork is made of a nickel-iron-molybdenum alloy[2] (vibralloy)

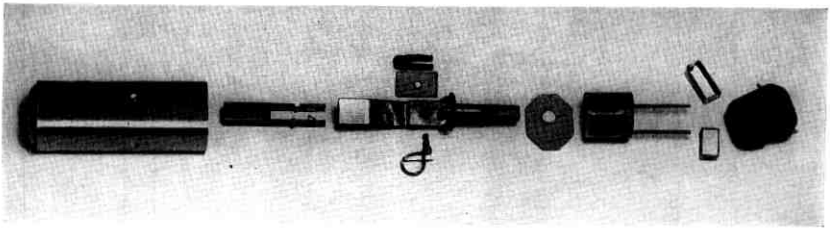


Fig. 2 — Exploded view showing individual parts.

having controlled elastic and magnetic properties. Annealed permalloy with low coercive force and high permeability is used for the polepiece and shell to reduce magnetic flux changes. The materials and shapes of other parts are chosen to minimize dimensional changes with time and environmental conditions.

III. FREQUENCY SELECTION AND FINE TUNING

The range of resonant frequencies is obtained with tuning forks that have the same over-all length but varying free tine lengths. The small dimensions of these forks require the brazing fillets and the free reed lengths and thicknesses to be precisely controlled. By special attention to rolling of the reed stock, precise jigging of the reeds and base block, and brazing with minimum fillet dimensions, it is feasible to produce forks in which the individual tine frequencies are sufficiently close to chosen nominal frequencies spaced 15 cycles apart so that they may then be accurately tuned to these desired frequencies.

Precise or fine tuning is accomplished with spring sliders that may be moved along the tines. This requires a slider that will stay in place under shock and vibration, will provide an adequate tuning range, and will allow the necessary fineness of frequency adjustment. This is achieved by means of small spring clips that snap on and ride along the edges of the tines. These sliders are shaped so that pressure at the center releases the force with which the slider seizes the reed and permits it to be moved. Each slider has a mass of about 1 milligram and provides a tuning range of about 10 cycles on forks near 500 cycles and of about 25 cycles on forks near 1000 cycles. The sliders may be moved in increments less than a thousandth of an inch, permitting the resonant frequencies to be readily set to a desired value within ±0.05 cycle. The seizure forces are large so that shock and vibration acceleration in excess of 1500 G are required to move the sliders.

IV. CONTACT FACILITY AND SENSITIVITY ADJUSTMENT

The sensitivity is adjusted in manufacture by changing the contact gap separation. A fine rhodium wire having a resonance frequency above the frequency range of the tuning forks is supported by a loop of larger wire that may be rotated on a tapered stud through the frame. The fine wire is pretensioned with a prescribed force against the loop wire to form a lift-off type of contact that is accurately positioned and will follow large tine excursions without objectionable interference with the tine motion. This construction[3] results in a contact that makes to a low resistance with the vibrating contact on the reed for intervals of time that may be 25 per cent or more of the reed period, depending on the applied power. The operating sensitivity of the selector is precisely set by rotating the loop on the stud axis and thereby causing the end of the contact wire to move toward or away from the reed contact. The point of contact is close to the axis of rotation so that a fine control of the contact gap may be achieved.

Bandwidth Control

The bandwidth or sharpness of the resonance curve is determined primarily by three dissipative factors, namely: internal frictional losses in the reed material, viscous losses in the air surrounding the reeds, and eddy-current losses in electrically conducting parts. The last factor has been chosen as the adjustment or control means for bandwidth. A copper washer is placed around the polepiece and where flux changes due to motion of the reeds induce eddy currents in the copper. By selecting the proper washer thickness and diameter and by setting the magnet strength to yield the proper flux density, eddy currents are developed when the tines vibrate that absorb energy and reflect into the system as an effective mechanical resistance that broadens the resonance curve by the desired amount.

V. VIBRATING SYSTEM PARAMETERS

Tabulated in Table I are some measured and derived data that show the magnitudes of the more important vibrating system constants of two selector samples with resonant frequencies nearly an octave apart. These are typical values that will be of interest to those concerned with the vibrational mechanics, electromechanical coupling, and other analytical design factors.

TABLE I

	Nominal Frequency 517.5 cps	Nominal Frequency 997.5 cps
Reed dimensions — length	1.4 cm	1.01 cm
thickness	0.015 cm	0.015 cm
width	0.254 cm	0.254 cm
Effective reed stiffness	1.45×10^5 dynes/cm	3.88×10^5 dynes/cm
Resonant frequency as brazed	560 cps	1068 cps
Resonant frequency with contact	530 cps	1011 cps
Resonant frequency with slider as tuned	517.5 cps	997.5 cps
Effective reed mass as brazed	0.0118 grams	0.0087 grams
Effective reed mass with contact	0.0130 grams	0.0096 grams
Effective reed mass with slider as tuned	0.0138 grams	0.0099 grams
Electrical impedance at resonant frequency	$478 + j231$	$448 + j430$
Electrical blocked impedance at same frequency	$220 + j277$	$235 + j485$
Electrical motional impedance at same frequency	$258 - j46$	$213 - j55$
Current to just close contact	0.275 milliamps	0.275 milliamps
Bandwidth	1.1 cycles	1.3 cycles
Effective mechanical resistance of fork at resonance	0.19 mechanical ohms	0.16 mechanical ohms
Electromechanical coupling factor	$2.24 \times 10^5 \,\overline{\lvert 5^\circ}$ dynes/abamp	$1.88 \times 10^5 \,\overline{\lvert 7.2^\circ}$ dynes/abamp
Effective magnetic gap stiffness (each gap — from frequency shift measurements)	-0.02×10^5 dynes/cm	-0.02×10^5 dynes/cm
Corresponding gap flux density	200 gauss	200 gauss
Maximum tine flux density (assuming fringe flux equal to gap flux)	4000 gauss	4000 gauss

VI. PERFORMANCE OBJECTIVES

Consideration of the over-all system operating requirements for personal radio signaling pertaining to such factors as the needed number of individual selections, practical radio receiver power levels, calling rates, and environmental conditions, led to the following objectives for the performance of the reed selectors:

1) Nominal frequency range — 517.5 to 997.5 cycles.

2) Nominal frequency separation — 15 cycles.

3) Frequency deviation limits — ±0.3 cycle, including adjustment tolerances, aging, shock, magnetic changes, and all other instabilities except those due to temperature changes.

4) Temperature-frequency deviation limits — ±0.2 cycle over temperature range of 35°F to 110°F (2°C to 43°C).

5) Nominal bandwidth — 1.0 cycle.

6) Bandwidth deviation limits — 0.8 to 1.4 cycles resulting from temperature changes and all other causes.

7) Nominal current to just operate contact — 0.25 milliamps for a nominal 500-ohm coil impedance at resonance.

8) Just-operate current deviation limits — ±3.0 db resulting from temperature changes and all other causes.

These objectives are mutually consistent in that the limits given in each case are as large as can be tolerated without reducing the limits on some other factor. There are other important design considerations that must not be neglected, such as weight, size and shape, contact life, shock tolerances, corrosion resistance, magnetic interaction and so forth, and with respect to which the selectors must, of course, be adequate. However, the above-tabulated characteristics are the most significant from an operating standpoint and are sufficient under marginal conditions to assure positive operation and avoid false signaling.

VII. TYPICAL MEASURED DATA

Presented below are measured data showing that the above-described reed selector meets these objectives. By means of the spring sliders, the two tine frequencies are made alike within a small fraction of a cycle and are given values that result in a combined fork frequency well within requirements. Attention is given in the assembly and adjustment procedure to magnetically and mechanically stabilize the whole structure. The magnet is stabilized well below its maximum remanence; the whole final assembly is subjected to a moderately high temperature to relieve residual stresses; and the tines are vibrated at a suitable level to bring them into a normalized magnetic state prior to final adjustment. The resulting selectors have resonant frequencies that will remain within ±0.3 cycle from their nominal frequencies at normal room temperatures and under reasonable conditions of mechanical shock and electrical overload. Negligible changes occur under shocks up to 1500 G (2 milliseconds duration) or with input levels 20 db above the just-operate values.

Frequency stability with temperature is achieved by making the forks of a nickel-iron-molybdenum alloy of such a composition that magnetic permeability changes are small and the temperature coefficient of Young's modulus is low and of a magnitude to compensate for dimensional changes with temperature. Operate current stability is realized by additional attention to the design geometry and materials so that changes in temperature cause variations in contact separation that are a small fraction of a mil-inch.

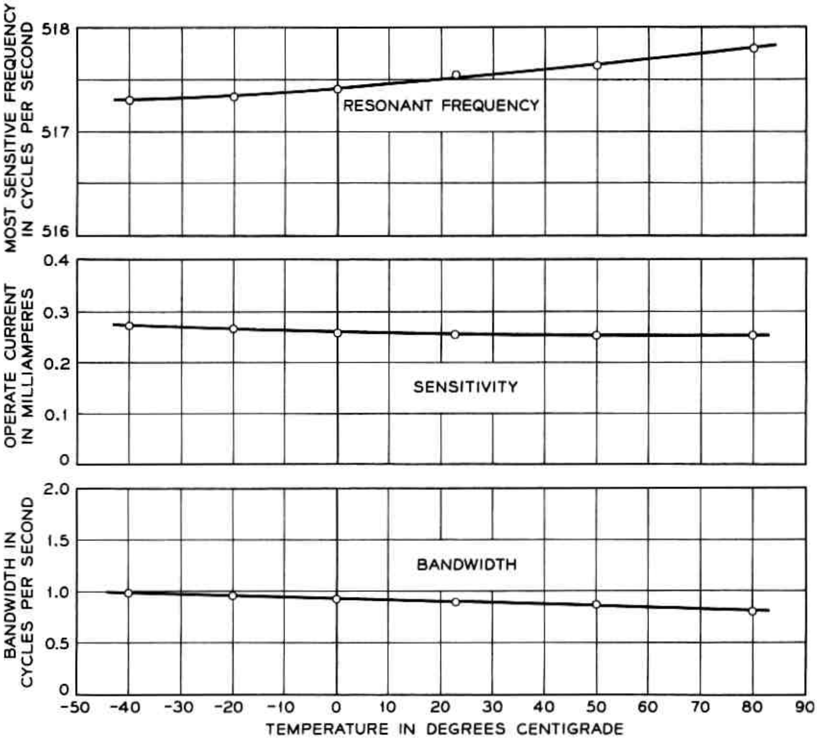


Fig. 3 — Variation with temperature in the operating characteristics of a typical lower-frequency tuned reed selector.

Fig. 3 and Fig. 4 are graphs of measured data showing variations with temperature in the resonance frequency, just-operate current and bandwidth of two typical samples, one at each end of the nominal frequency range. The range covered by these graphs is much wider than that required for most applications. In the more common temperature range of 35° to 110°F, the deviations are well within the limits tabulated above.

Fig. 5 and Fig. 6 are electrical impedance diagrams of the same two selector samples with resistance and reactance as coordinates and frequency as the variable parameter. This form of plot emphasizes the interesting values near resonance and may be used for analytical purposes.[4] From these graphs, it can be determined that the conversion of electrical to effective mechanical power is about 46 per cent and that the available electric power necessary to just operate the contact is about 33 microwatts.

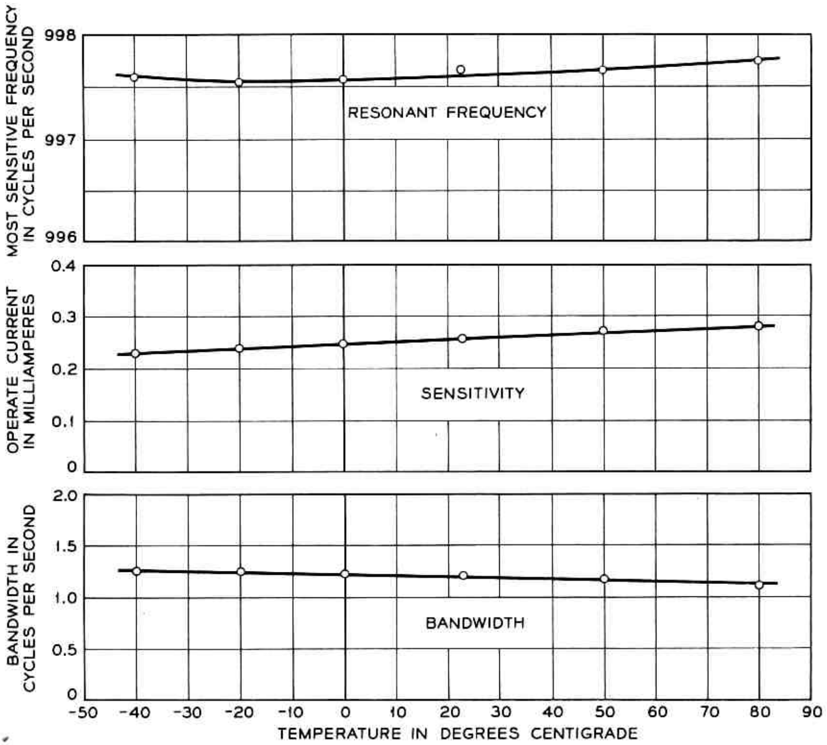


Fig. 4 — Variation with temperature in the operating characteristics of a typical upper-frequency tuned reed selector.

VIII. NOMINAL OPERATING LEVELS AND TIMES

The electrical power source supplying selectors in a system must have an available power capacity sufficient to cause dependable contacting under the worst temperature and adjustment conditions. These worst conditions obtain when the frequency deviation from nominal and the just-operate current are at their maximum values. Considering the limits permitted in these selectors and making allowance for contact quality and life with some statistical advantage taken of the small chance of all limiting conditions occuring simultaneously, it was determined that the minimum electrical input power should be 6 db above that needed to barely close the contact of a nominal selector. At this level, the time required to close the contact after energizing the coil is equal to the time needed for the reed amplitude to decay below contact-

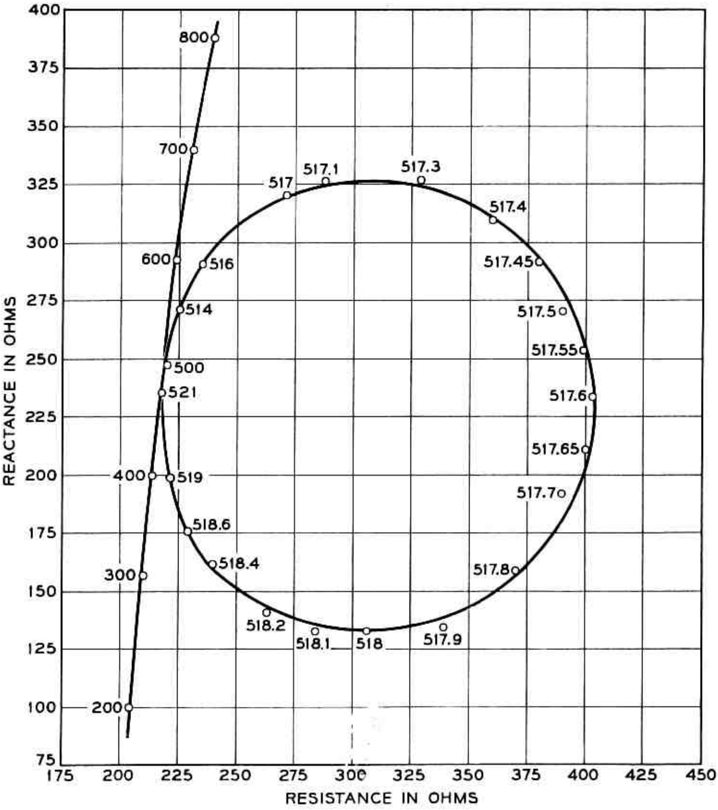


Fig. 5 — Vector impedance diagram of a typical lower-frequency unit.

ing amplitude after the coil current is stopped. For nominal selector constants, this time is approximately 225 milliseconds. Input levels higher than 6 db above just-operate will result in faster operating times and slower decay times, but the sum of the operate and decay times will increase less than 20 per cent up to input levels 12 db above the nominal just-operate value.

IX. CONTACT CAPACITY AND LIFE

The contact has greater capability than would at first seem likely. Such a light contact is most frequently used in circuits to change the potential on a tube or transistor and thereby trigger some desired signaling or switching function without the contact current exceeding a few

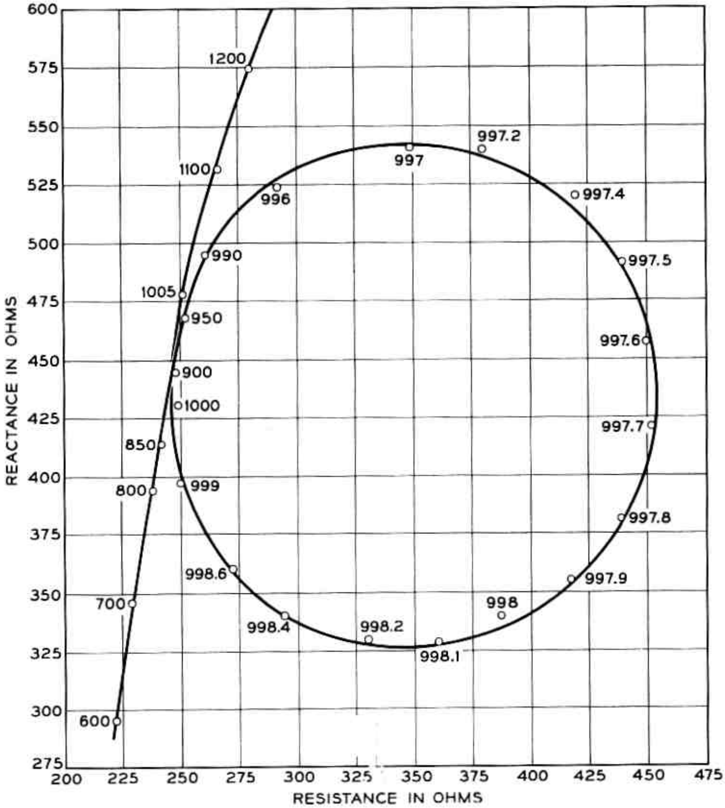


Fig. 6 — Vector impedance diagram of a typical upper-frequency unit.

milliamperes. The contact closure is intermittent at a rate corresponding to the frequency of the selector, and the duration of the individual closures is a small fraction of a millisecond, depending upon the frequency and input level. These short closures, however, occurring at a rate of several hundred times per second, may control current pulses that have an integrated or averaged power that is a substantial fraction of a watt.

The maximum power that can be controlled depends mostly upon the reactive elements in the contact circuit and the life needed from the selector. As an example of what may be expected, Fig. 7 shows changes that occurred in the resonance frequency and the sensitivity of a typical selector when operated continuously (except for a few minutes about every 100 hours during check test) over a period of 1500 hours. The

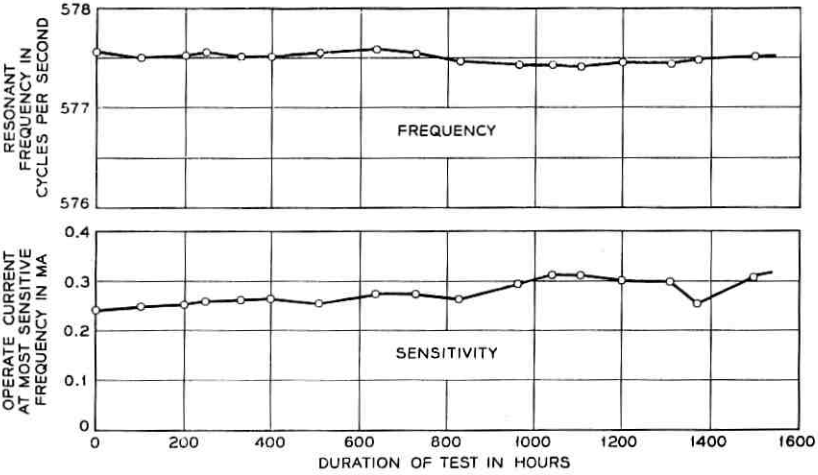


Fig. 7 — Variation with time in the sensitivity and frequency of a selector closing a 12-volt battery through a 240-ohm resistor.

electrical input was 9 db above the just-operate value, and the contact closed a 12-volt battery through a 240-ohm resistor, giving a closure current of 50 milliamperes. Throughout the test period the resonance frequency changed only slightly and the just-operate current increased about 20 per cent. This later change was due to erosion of the contact wire, which increased the contact gap. Erosion was minimized by connecting the fine contact wire to the negative side of the battery. At the end of the test, the diameter of the contact wire was approximately half its original value.

X. APPLICATIONS

The manner in which these selectors are used in the circuits of the BELLBOY Personal Radio Signaling system will be described in a paper to be published on the pocket radio receiver. In this system, three tuned reed selectors are operated simultaneously in the receiver, and these trigger a transistor oscillator that gives an audible signal. The power controlled by the contacts in this case is small.

The substantial power capacity of the contacts can be used to operate relays and other devices directly. Pulses of current from a battery at the selector frequency can be supplied to a smoothing or integrating capacitor, and the relatively constant voltage across the capacitor can be used to operate a sensitive dc relay. The battery may be at the loca-

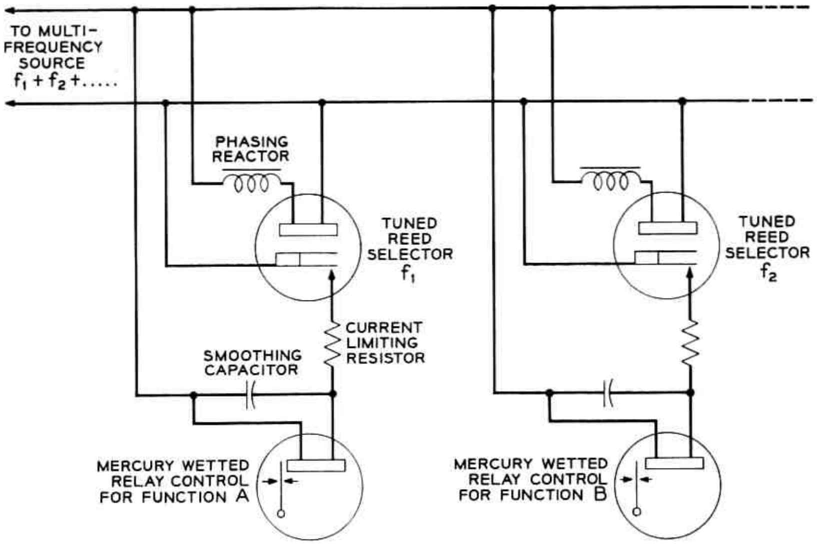


Fig. 8 — Reed selector actuated mercury relay for selective control of multiple functions requiring substantial powers.

tion of the reed selector or may be supplied by superposition over the same circuit used to transmit the selector frequency.

The contact may also be used as a synchronous rectifying means to generate dc from the same ac source that operates the selector, as shown in Fig. 8. When the source frequency corresponds to that of the reed selector, the contact of the selector closes in synchronism once each cycle to send unidirectional pulses to the capacitor and relay in parallel. The capacitor smoothes the pulses and gives a nearly constant current in the relay winding. For maximum sensitivity it is desirable that the contact closures occur near the peaks of the supply voltage wave, and this is accomplished by connecting a large reactance (either inductive or capacitative) in series with the selector winding. This reactance also serves to attenuate the supply voltage applied to the selector winding to avoid overdriving the reeds, because a supply voltage large enough to operate a relay is ordinarily many times that needed to operate the reed selector. Combination circuits using reed selectors and mercury-wetted contact relays provide a simple means of selectively controlling substantial powers to perform a multiplicity of functions over a single pair of wires.

When operated just below the contacting level, these selectors have a

Q (resonant frequency-to-bandwidth ratio) in the range of 500 to 1000 and therefore may be used effectively in a selective bridge or filter circuit as described in a previous paper.[5] The use of such a selective circuit in the feedback loop of a single transistor oscillator results in an attractively simple source of frequency having a precision corresponding to that of the selector.

XI. ACKNOWLEDGMENTS

Original suggestions regarding the construction of this selector and skilled model work were contributed by the late R. L. Guncelle. Essential refinements in the design and in the fabrication techniques were made by K. F. Bradford and D. H. Wenny. E. J. Kasello carried out most of the adjusting and testing. The successful outcome of this development is due in no small measure to their efforts.

REFERENCES

1. Mitchell, D., and Van Wynen, K. G., B.S.T.J., **40,** Sept., 1961, p. 1239.
2. Fine, M. E., and Ellis, W. C., Jour. Metals, **3,** Sept., 1951, p. 761; also U. S. Patent 2,561,732.
3. Alternatives disclosed in U. S. Patent 2,877,319.
4. Hunt, F. V., *Electroacoustics*, Harvard University Press, John Wiley and Son, New York, 1954.
5. Keller, A. C., and Bostwick, L. G., Trans. A.I.E.E., **68,** Pt. I, 1949, p. 383, and paper by Pruden, H. M., and Hoth, D. F., p. 387; also see U. S. Patents 2,583,542 and 2,630,482.

An X-Ray Diffraction Study of the Structure of Guanidinium Aluminum Sulfate Hexahydrate

By S. GELLER and H. KATZ†

(Manuscript received March 21, 1961)

The Busing-Levy IBM 704 least squares program has been applied to three-dimensional X-ray diffraction data from crystals of guanidinium aluminum sulfate hexahydrate taken with the Bond-Benedict single-crystal automatic diffractometer. Indications of interactions between parameters were evident in the early stages of refinement and were not removed in the subsequent cycles. Strong interactions were subsequently corroborated by large values of many of the correlation coefficients of pairs of parameters. In this case these interactions prevent refinement. The correctness of the general features of the structure as given in a previous paper on the gallium isomorph is nevertheless corroborated by the present investigation.

To enable those who have had similar difficulties to compare results, a fairly detailed account is given of the course of the attempt to refine the structure. The effects of highly correlated parameters are emphasized.

I. INTRODUCTION

The purposes of the investigation to be described were manifold. An approximate structure of the isomorphous gallium compound has already been reported.[1] The gallium compound with the heaviest metal atom among the isomorphs appeared to be best for establishing the general features of the structure.[2] However, in the hope of finding a closer relation between the structure and its electrical properties, it appeared that a refinement of the structure would be very worthwhile. In such a case, one would wish to have all of the atoms of more nearly the same scattering power; thus the guanidinium aluminum sulfate hexahydrate (G.A.S.H.) compound seemed most suitable for this pur-

† The contribution of H. Katz to this work was made during a period of employment at Bell Telephone Laboratories in the summer of 1959.

pose. Furthermore, this crystal would have the lowest linear absorption coefficient for all practical radiations; the importance of this feature will be discussed later. But probably most important, it was anticipated that the aluminum compound would be the one on which most measurements of various sorts would be made. This has indeed been the case.

While our earlier paper[1] was in press, a note[3] appeared in *Kristallografiia* which gave an approximate structure for G.A.S.H. and its isomorphs which differed from that reported by us. A check with our data indicated that the structure reported by Varfolomeeva *et al.*[3] was incorrect,[2] but this did not mean that the structure reported by us was necessarily correct. We had to face the question as to whether the correct structure might lie between the two structures or as mentioned in our first paper, perhaps some subtle disorder existed in the structure. In any case the appearance of the other result gave additional impetus to completion of work that had been started several years ago.

There is a further importance of this work. The quantitative X-ray data were taken with the Bond-Benedict single-crystal automatic diffractometer.[4] It is the only crystal so far studied with this equipment and perhaps is the first X-ray structure analysis to be based on three-dimensional data collected automatically. Thus at least a small part of this paper will be devoted to an assessment of this equipment and suggestions as to future plans.

Perhaps the most frustrating experience encountered is to find indeterminate a problem which has taken considerable expenditure of time and effort of various sorts. One such reported problem in the field of X-ray crystallography is that of the determination of the structure of tetragonal $BaTiO_3$; this problem was found by Evans[5] to be indeterminate by X-ray analysis, at the very least on the basis of the data collected. The results of the work on the three-dimensional data of G.A.S.H. indicate that the structure as originally reported by us is essentially correct. But we find that although a low discrepancy factor and standard error of fit are obtained by the least squares method of refinement, the structure cannot be refined; that is, convergence is not attained: there are parameter oscillations in each least squares iteration; some improbable interatomic distances and large error estimates are obtained. The cause appears to be strong interdependence of many of the parameters.

In this investigation the correlation matrix is used to demonstrate the existence of the strongly interacting parameters. The importance of this approach has also been demonstrated by a recent investigation described in a paper written by one of us (S.G.).[6]

TABLE I — LATTICE CONSTANTS OF GUANIDINIUM ALUMINUM SULFATE HEXAHYDRATE

Investigators	a,Å	c,Å
Wood	11.77 ± 0.04	8.98 ± 0.03
Ezhkova, *et al*	11.737 ± 0.002	8.948 ± 0.002
This work	11.75 ± 0.02	8.94 ± 0.01

II. CRYSTAL DATA

Guanidinium aluminum sulfate hexahydrate, $C(NH_2)_3Al(SO_4)_2 \cdot 6H_2O$, is isostructural with the previously reported[1] gallium compound. The morphology and unit cell dimensions have been reported by Wood.[7] Lattice constants have also been reported by Ezhkova *et al.*[8] The central values of our lattice constants, obtained from careful measurement of Buerger precession camera photographs, differ from those reported in both of the aforementioned papers, but are in better agreement† with those of Ezhkova *et al.*[8] For purposes of comparison, the variously reported values are listed in Table I.

As described earlier,[1] the most probable space group to which the crystal belongs is P31m and the unit cell contains three formula units. The molecular weight of the Al compound is 387.29, the volume of the unit cell is 1,069 Å^3, and the X-ray density is 1.804 g/cc.

III. DETERMINATION OF THE STRUCTURE

The determination of the structure has been described in the paper on the gallium compound. The evidence for the correctness of the general features of the structure described in that paper, including the orientation of the guanidinium ions, is conclusive as will be shown subsequently.

IV. EXPERIMENTAL

The Bond-Benedict single-crystal automatic diffractometer[4] was used to collect the three-dimensional data. Some changes from the original design of the instrument and in the electronics were made before the final data were taken. A detailed description of these changes must be left to the original authors. However, it should be mentioned that for these particular data (which were taken in 1956), a proportional counter replaced the Geiger counter and the "back-set" correction[4] was virtually

† Dr. E. A. Wood and Mrs. V. B. Compton have informed us that their recent measurements of lattice constants of G.A.S.H. give values which agree more closely with those of Ezhkova *et al.*[8] and of the present work.

eliminated by circuitry changes. Also, the internal geometry of the collimator was changed to square cross section.

The need for a collimator with square cross section derived from the mechanics of the instrument. The "back-setter" produces a jarring of the goniometer head which could at times translate the crystal very slightly out of the original alignment in the X-ray beam. If the beam has a circular cross section, slight deviation from coincidence of crystal cylinder and rotation axes causes significant differences in intensity when the diameter of the crystal is large relative to the beam cross section. This is not true of a beam with a more or less square cross section.

Of course, one would not have to worry too much about this if small crystals were being used. However, for this instrument and the use of the usual type of sealed X-ray tube, it is necessary to use large crystals to obtain the data. (This will be discussed further later.)

Two cylindrical crystals were used to obtain the data attainable by this instrument with Cu$K\alpha$ radiation and a pentaerythritol monochromator. The crystal aligned along the c-axis had a diameter of 0.67 mm; the crystal aligned along the $[20\cdot1]$ direction (orthohexagonal A-axis) had a diameter of 0.54 mm. With a linear absorption coefficient for Cu$K\alpha$ radiation of 48.7 cm^{-1}, the values of μR for these crystals are 1.64 and 1.32 respectively.

As described in the paper by Bond,[4] the single-crystal automatic diffractometer works on a principle similar to that of the equi-inclination Weissenberg camera. With Cu$K\alpha$ radiation, seven levels were obtainable about the c-axis and fifteen about the orthohexagonal A-axis.

Data from a particular level n were collected as follows: The alignment of the crystal was checked. This was done in two ways whenever possible. A microscope could be used to align the crystal cylinder axis with the rotation axis of the instrument. The equi-inclination angle was calculated and the crystal set to this angle. The arrangement of the counter of the instrument is always set so that the diffracted beam is incident perpendicularly to the window. Thus the counter is actually moved to twice the angle of the crystal from the zero level situation. If a particular reflection (for example, $00\cdot l$ on the lth level about the c-axis) was observable when the counter angle was equal to zero degrees for a given layer, this reflection was used to readjust crystal and counter.

To obtain the weak intensities, the diffraction unit settings were usually 40 kv and 20 ma. To obtain the stronger reflections, proper settings of the voltage and tube current were made so as to record enough moderate reflections to establish a scale between the two patterns.

Integrated intensities, crystal angles and counter angles for each level were recorded automatically by the Leeds-Northrup two-pen recorder as described in the papers by Bond and Benedict.[4] As indicated above, resetting was made manually for each new level.

Following the collection of the data by the recorder, it was necessary to index the data: This was the most time-consuming (i.e., on a man-hour basis) part of the data processing required to obtain the observed amplitudes. The indexing was carried out with the use of the plotting device.[4] (The indexing problem will be discussed further later.)

Following the indexing of all the data, the usual absorption, Lorentz-polarization and Tunell[9] rotation factors† were applied to extract the relative $|F_o|^2$. (The polarization correction is for monochromatized radiation.) The calculation was programmed for the IBM 704 by R. G. Treuting. The corrections calculated were based on the formulae‡ given by Bond and the tables used for the absorption corrections are those given in Bond's paper.[10] The program written by Treuting put the resultant $|F_o|^2$'s or $|F_o|$'s out on cards as well as on a print-out. The individual Lorentz-polarization, absorption and Tunell rotation factors were also printed out for each reflection for each layer on which it appeared.

Having extracted the $|F_o|^2$'s for each layer about each of the two axes, the next step involved an iterative cross-calibration process to bring the values to the same basis. An IBM 704 program written by W. R. Romanow allowed us to apply constant factors to the sets of $|F_o|^2$ put out by the intensity correction program. Romanow's program also put out new cards so that we could apply a different constant to the new values if necessary.

When we felt we had arrived at the best values, it was decided to carry out the subsequent least squares refinement on the basis of the $|F_o|$ values. Using a short program written by Romanow, square roots were taken of all the $|F_o|^2$'s and put out on cards. Those that came from layers about the orthohexagonal A-axis were then sorted on the values of l for ease in setting up the data for the least squares refinement.

As described in the Bond-Benedict papers, some reflections do not get entirely into the counter; thus, in order to be sure that all are ob-

† The proportional counter employed had a linear response to counting rates of over 20,000 cps. Because for even the strongest reflections, observed counting rates over 10,000 cps gave integrated intensities which went off scale on the recorder, no dead-time correction[4] was necessary for any of the reflections.

‡ The formula for P_l on p. 380 of Bond's paper should read

$$P_L = T \sin 2\theta \Bigg/ \left\{1 + \left(\frac{q - \sin^2 \nu}{(1+q)\cos^2 \nu}\right)(1 + \cos 2\theta)^2 - \frac{2q}{1+q}(1 + \cos 2\theta)\right\}.$$

tained, the instrument was designed to obtain each reflection twice. For this reason the counter has a 4° window. Even at that, not all the reflections of a given form will have the same intensity, but usually about a twofold axis, a form of reflections of moderate intensity will have two with the same intensity. About a threefold axis, perhaps eight of twelve reflections from a given $hk \cdot l$ form will have the same intensity or 12 out of 16 of a given $hk \cdot 0$ form. Unfortunately, the weaker reflections do not give as good results as the moderate to strong ones. In the case of the c-axis layers, if there was a variation in the height of peaks which appeared to have been fully in the window, the value taken for the integrated intensity was the average of the several peaks. In the case of the orthohexagonal A-axis layers, because there were fewer peaks contributing to a form and therefore a greater possibility that only one peak was squarely in the window, the value recorded in most cases was the measure of the highest peak.

In taking the averages of observed structure amplitudes, the weighting was in accord with the above. For example for a given $| F_{hk \cdot l} |$, $h,k,l \neq 0$, the value from the c-axis layer was weighted four times and a value from an orthohexagonal A-axis layer once. The standard deviation was calculated in accordance with the analysis given in Chapter 16 of the book by Dixon and Massey[11] and as suggested earlier by Ibers.[12] However, for the unobserved, the standard deviation was taken as equal to half the minimum observable. For $| F_{00 \cdot l} |$'s which would have unity weight since they appear only once, the σ was taken in accordance with a subjective estimate comparing the particular $| F_{00 \cdot l} |$ with others of similar value. The agreement between or among $| F_o |$'s from the same form but from different layers was quite good generally except for the weakest reflections.

In the Cu$K\alpha$ sphere, there is a total of 895 X-ray forms of guanidinium aluminum sulfate hexahydrate. The geometry of the Bond diffractometer allows us to observe only 842 of these. Of those possibly observable by the instrument, only 546 were actually observed.

V. ATTEMPT TO REFINE THE STRUCTURE

Because the major point of this paper is to demonstrate that the refined structure under discussion is effectively unattainable from the X-ray diffraction data, it seems worthwhile to give some of the details of the calculations. To make such a discussion simpler, the pertinent data are collected in tables. In Tables II and IV the values of parameters and some other important information are listed. In Table II two columns are assigned to each cycle; the left one lists the starting parameters,

Table II — Results of Least Squares Calculations (First Set of Weights)

Parameters		Cycle 1		Cycle 2		Cycle 3		Cycle 4		Cycle 5		Cycle 6		Cycle 7		Cycle 8		Cycle 9		Cycle 10		Cycle 12	
Scale Factors	1	0.6667	0.553		0.583	0.550	0.542		0.5395		0.5612		0.5552		0.5467	0.5552	0.5569		0.5557				
	2	0.6667	0.542		0.569	0.540	0.534		0.5552		0.5889		0.6162		0.6200	0.6162	0.6161		0.6186				
	3	0.6667	0.548		0.566	0.550	0.581		0.5695		0.5829		0.5879		0.5703	0.5879	0.5890		0.5933				
	4	0.6667	0.580		0.619	0.580	0.580		0.5879		0.6221		0.6327		0.6132	0.6327	0.6208		0.6161				
	5	0.6667	0.579		0.564	0.580	0.619		0.6187		0.6411		0.6492		0.6909	0.6492	0.6396		0.6449				
	6	0.6667	0.610		0.575	0.610	0.630		0.6494		0.6782		0.6887		0.6978	0.6887	0.7116		0.7327				
	7	0.6667	0.567		0.580	0.570	0.609		0.6533		0.6960		0.6887		0.6739	0.6887	0.6881		0.6854				
	8	0.6667	0.435		0.493	0.430	0.545		0.6096		0.6740		0.6826		0.6873	0.6826	0.6836		0.6856				
Atom 1 N (II)	x	0.553	0.551		0.554	0.551	0.558		0.5624		0.5632	0.5540	0.5600	0.5333	0.5626	-0.5533	-0.5531		-0.5522			-0.4630	-0.4614
	y	0.418	0.407		0.390	0.407	0.409	0.3333	0.3379		0.3344	0.3333	0.3386	0.3333	0.3312	-0.3333	-0.3349		-0.3338			0.2740	0.2791
	z	0.520	0.527		0.540	0.527	0.532		0.5544		0.5567		0.5492		0.5639	0.5492	0.5350		0.5275		0.5236	0.535	0.5321
	B or β_{11}	2.00	6.13	4.50	10.2	0.01086	0.01611	0.01000	0.01173		0.00887										0.02497	0.00890	
	β_{22}					0.01086	0.02353	0.01000	0.01478		0.00040	0.00700									0.01493	0.00700	
	β_{33}					0.01408	0.02927	0.02000	0.01223		0.01817										0.02670	0.01820	
	β_{12}					0.00543	0.00412		0.01032		0.00503										0.01499	0.00500	
	β_{13}					0.00000	0.00650		0.00188		0.00516										0.01027	0.00520	
	β_{23}					0.00000	0.00135		0.00084		0.00406										0.00615	0.00410	
Atom 2 O (V)	x	0.342	0.339		0.345	0.339	0.340		0.3422		0.3435		0.3437		0.3459	0.3437	0.3451		0.3449				0.3455
	y	0.120	0.113		0.116	0.113	0.114		0.1166		0.1252	0.1200	0.1191		0.1176	0.1191	0.1165		0.1166				0.1159
	z	-0.220	-0.221		-0.217	-0.221	-0.222		-0.2114		-0.2010		-0.2064		-0.2075	-0.2064	-0.2008		-0.1972		-0.1976		-0.2009
	B or β_{11}	2.00	1.75	1.50	0.74	0.00362	0.00614		0.00748		0.01087	0.00850									0.00843		
	β_{22}					0.00362	0.00314		0.00219		0.00215										0.00307		
	β_{33}					0.00469	0.01295		0.01947		0.01511										0.02193		
	β_{12}					0.00181	0.00463	0.00400	0.00510	0.00350	0.00611	0.00400									0.00375		
	β_{13}					0.00000	0.00221		-0.00016		-0.00338										-0.00333		
	β_{23}					0.00000	0.00029		-0.00022		-0.00222										-0.00206		
Atom 3 O (VI)	x	-0.314	-0.321		-0.333	-0.321	-0.324		-0.3222		-0.3235		-0.3214		-0.3208	-0.3214	-0.3212		-0.3211				-0.3218
	y	-0.120	-0.120		-0.126	-0.120	-0.120		-0.1171		-0.1130	-0.1170	-0.1122	-0.1150	-0.1119	-0.1150	-0.1139		-0.1137				-0.1173
	z	0.258	0.240		0.236	0.240	0.235		0.2349		0.2423		0.2362	0.2400	0.2328	0.2400	0.2448		0.2450				0.2439
	B or β_{11}	2.00	1.90	1.75	1.35	0.00423	0.00271		0.00080		-0.00106	0.00100									0.00287		
	β_{22}					0.00423	0.00033	0.00100	0.00084		0.00344										0.00149		
	β_{33}					0.00547	0.01067		0.01162		0.01523										0.01905		
	β_{12}					0.00211	0.00153		-0.00078		-0.00155										0.00066		
	β_{13}					0.00000	-0.00002	0.00000	0.00028		-0.00181										-0.00081		
	β_{23}					0.00000	-0.00030	0.00000	-0.00052		0.00012										-0.00057		
Atom 4 O (IX)	x	-0.477	-0.471		-0.475	-0.471	-0.471		-0.4687		-0.4673		-0.4654		-0.4702	-0.4654	-0.4649		-0.4654				-0.4677
	y	0.333	0.330		0.332	0.330	0.327		0.3258		0.3263		0.3288		0.3261	0.3288	0.3270		0.3272				0.3291
	z	-0.066	-0.058		-0.059	-0.058	-0.059		-0.0593		-0.0577		-0.0588		-0.0654	-0.0588	-0.0599		-0.0546.		-0.0520		-0.0504
	B or β_{11}	2.00	0.73	1.50	-0.22	0.00362	-0.00259	0.00000	-0.00104	0.00100	0.00063										0.00266		
	β_{22}					0.00362	0.00186		0.00003		0.00036										0.00009	0.00120	
	β_{33}					0.00469	0.01372		0.01363		0.01626										0.01613		
	β_{12}					0.00181	-0.00006	0.00000	-0.00223	0.00000	0.00017										0.00127		
	β_{13}					0.00000	-0.00052	0.00000	-0.00083		0.00083										-0.00122		
	β_{23}					0.00000	-0.00175	0.00000	-0.00028	0.00000	0.00147										0.00174		
Atom 5 O (X)	x	0.476	0.465		0.459	0.465	0.466		0.4614		0.4575		0.4635		0.4636	0.4635	0.4664		0.4647				0.4665
	y	-0.336	-0.343		-0.347	-0.343	-0.340		-0.3391		-0.3380		-0.3417		-0.3389	-0.3417	-0.3388		-0.3405		-0.3391		-0.3398
	z	0.174	0.172		0.176	0.172	0.174		0.1775		0.1816		0.1792		0.1762	0.1792	0.1717		0.1756		0.1758		0.1810
	B or β_{11}	2.00	2.47	2.00	2.16	0.00483	0.00924		0.00444		0.00459										0.00099	0.00200	
	β_{22}					0.00483	0.00118		0.00097		0.00385										0.00142		
	β_{33}					0.00626	0.00588		0.00905		0.01667										0.01636		
	β_{12}					0.00241	0.00112		-0.00097		0.00037										0.00135		
	β_{13}					0.00000	-0.00422	0.00420	-0.00243		-0.00322										0.00007		
	β_{23}					0.00000	0.00060		-0.00062		-0.00187										-0.00150		
Atom 6 N (I)	x	0.113	0.113		0.106	0.113	0.112		0.1213		0.1258	0.1130		-0.1130	-0.1138	0.1130	0.1135		0.1132				0.1198
	z	0.500	0.497		0.519	0.497	0.500		0.5037		0.5068		0.5038		0.4930	0.5038	0.5177		0.5065		0.5175		0.5137
	B or β_{11}	2.00	-2.50	3.00	-4.60	0.00724	-0.01095	0.00500	0.00378		0.01679	0.0038									0.00798	0.00380	
	β_{33}					0.00938	-0.00125	0.00500	0.01479		0.00380	0.01480									-0.00095	0.01480	
	β_{12}					0.00362	-0.00812	0.00000	0.00928	0.0032	0.02214	0.00320									0.01743	0.00320	
	β_{13}					0.00000	-0.00839	0.00000	-0.00028		-0.00593	0.00000									-0.01087	0.00000	
Atom 7 O (VIII)	x	-0.144	-0.135		-0.133	-0.135	-0.134		-0.1360		-0.1387		-0.1313		-0.1300	-0.1313	-0.1294		-0.1305				-0.1277
	z	-0.135	-0.124		-0.118	-0.124	-0.125		-0.1208		-0.1149		-0.1127		-0.1200	-0.1127	-0.1266		-0.1203		-0.1208		-0.1084
	B or β_{11}	2.00	1.53	1.50	-0.04	0.00362	0.00529		0.00732		0.00448										0.00649		
	β_{33}					0.00469	0.01168		0.00633		0.01204										0.01200		
	β_{12}					0.00181	0.00652	0.00400	0.00909	0.00650	0.00733	0.00400									0.00514		
	β_{13}					0.00000	0.00426		-0.00304	-0.00250	-0.00710	-0.00400									-0.01222	-0.00700	
Atom 8 O (VII)	x	0.140	0.130		0.124	0.130	0.131		0.1362		0.1432		0.1403		0.1376	0.1403	0.1441		0.1426				0.1383
	z	0.130	0.118		0.117	0.118	0.114		0.1153		0.1206		0.1181		0.1067	0.1181	0.1082		0.1124		0.1175		0.1222
	B or β_{11}	2.00	2.04	2.00	2.68	0.00429	0.00794		0.00856		0.00801										0.00728		
	β_{33}					0.00626	-0.00649	0.00500	0.00148		0.00469										0.00710		
	β_{12}					0.00241	0.00562		0.00903	0.00800	0.00841	0.00790									0.00749	0.00710	
	β_{13}					0.00000	-0.01109	0.00000	-0.01055	0.00000	-0.00842	-0.00400									-0.00545		
Atom 9 O (IV)	x	-0.328	-0.304		-0.292	-0.304	-0.297		-0.2929		-0.2881		-0.2900		-0.2937	-0.2900	-0.2928		-0.2912				-0.2938
	z	0.470	0.460		0.465	0.460	0.471		0.4708		0.4804		0.4741		0.4765	0.4741	0.4747		0.4756				0.4714
	B or β_{11}	2.00	5.04	2.00	0.71	0.00483	0.00767		0.00778		0.01081	0.01090									0.00823		
	β_{33}					0.00626	0.01094		0.00689		0.00549										0.00784		
	β_{12}					0.00241	0.00954	0.00550	0.00822	0.00700	0.00991										0.00627		
	β_{13}					0.00000	0.00105		-0.00071		-0.00066										0.00209		
Atom 10 O (III)	x	0.386	0.378		0.373	0.378	0.369		0.3726		0.3635	0.3730	0.3672		0.3647	0.3672	0.3708		0.3699				0.3682
	z	-0.409	-0.392		-0.400	-0.392	-0.398		-0.3939		-0.3925		-0.3955		-0.3937	-0.3955	-0.3913		-0.3894				-0.3965
	B or β_{11}	2.00	2.64	2.50	1.95	0.00604	0.00790		0.00280		0.00324										0.00651		
	β_{33}					0.00782	0.00884		0.01152		0.01338										0.01627		
	β_{12}					0.00302	0.00631		-0.00291	-0.00200	0.00107										0.00696	0.00640	
	β_{13}					0.0000	0.00389		0.00176		0.00171										0.00233		
Atom 11 O (II)	x	-0.433	-0.442		-0.445	-0.442	-0.444		-0.4418		-0.4324		-0.4410		-0.4418	-0.4410	-0.4403		-0.4412				-0.4440
	z	0.238	0.241		0.250	0.241	0.255		0.2789		0.2794		0.2743		0.2698	0.2743	0.2740		0.2764				0.2825
	B or β_{11}	2.00	1.80		2.71	0.00435	-0.00048	0.00200	-0.00165	0.00200	-0.00438	0.00200									-0.00066	0.00200	
	β_{33}					0.00563	0.02038		0.01860		0.02118										0.01777		
	β_{12}					0.00217	-0.00303	0.00000	-0.00452	-0.00100	-0.00646	0.00100									-0.00009		
	β_{13}					0.00000	-0.01026	0.00000	0.00062		0.00221										0.00540	0.00400	
Atom 12 O (I)	x	0.455	0.445		0.447	0.445	0.441		0.4491		0.4551		0.4536		0.4555	0.4536	0.4525		0.4538				0.4491
	z	-0.156	-0.154		-0.156	-0.154	-0.165		-0.1653		-0.1525		-0.1638		-0.1720	-0.1638	-0.1607		-0.1611				-0.1617
	B or β_{11}	2.00	-0.22	1.80	1.05	0.00435	0.00142	0.00300	0.00324		0.00268										-0.00227	0.00200	
	β_{33}					0.00563	0.01607		0.02608		0.00281	0.01500									0.00910		
	β_{12}					0.00217	0.00242	0.00200	0.00382	0.00250	0.00132										-0.00420	0.00200	
	β_{13}					0.00000	0.00270		0.00178		-0.01103	-0.00500									0.00376	0.00200	
Atom 13 S (I)	x	0.351	0.349		0.349	0.349	0.348		0.3486		0.3453		0.3475		0.3471	0.3475	0.3478		0.3479				0.3477
	z	-0.250	-0.251		-0.250	-0.251	-0.250		-0.2459		-0.2408		-0.2446		-0.2459	-0.2446	-0.2437		-0.2433				-0.2422
	B or β_{11}	0.80	1.21		1.19	0.00292	0.00360		0.00200.		0.00141										0.00163		
	β_{33}					0.00378	0.00604		0.00747		0.00601										0.00714		
	β_{12}					0.00146	0.00054		0.00308	0.00150	0.00161	0.00130									0.00067		
	β_{13}					0.00000	0.00049		-0.00087		-0.00203										-0.00041		
Atom 14 S (II)	x	-0.318	-0.322		-0.322	-0.322	-0.319		-0.3188		-0.3163		-0.3180		-0.3173	-0.3180	-0.3174		-0.3174				-0.3182
	z	0.306	0.307		0.309	0.307	0.308		0.3110		0.3179		0.3123		0.3125	0.3123	0.3138		0.3144				0.3152
	B or β_{11}	0.80	0.48	0.75	0.87	0.00181	0.00253		0.00375		0.00546										0.00194		
	β_{33}					0.00235	0.00574		0.00705		0.00724										0.00756		
	β_{12}					0.00091	0.00300	0.00200	0.00395	0.00320	0.00584	0.00500									0.00082		
	β_{13}					0.00000	0.00059		0.00034		-0.00079	-0.00078									0.00017		
Atom 15 C (II)	z	0.520	0.544		0.549	0.544	0.533		0.5459		0.5590		0.5530		0.5596	0.5530	0.5486		0.5420		0.5502		0.5507
	B or β_{11}	2.00	3.83	3.80	-0.95	0.00917	0.01477		-0.00094	0.00100	0.00263										0.00238		
	β_{33}					0.01189	-0.01835		0.00782		0.00694										0.01029		
Atom 16 Aℓ (II)	z	0.059	0.053		0.055	0.053	0.055		0.0574		0.0597		0.0566		0.0555	0.0566	0.0570		0.0572				0.0582
	B or β_{11}	1.50	0.49	0.50	1.42	0.00121	0.00191		0.00118		0.00134										0.00189		
	β_{33}					0.00156	0.00394		0.00517		0.00604										0.00770		
Atom 17 C (I)	z	0.500	0.501		0.488	0.501	0.482		0.5003		0.5141		0.5125		0.5262	0.5125	0.5031		0.4969		0.5072		0.4978
	B or β_{33}	2.00	3.36	3.40	11.14	0.01064	0.01083		0.01663		0.00569	0.01120									0.01033		
Atom 18 Aℓ (I)	B or β_{33}	1.50	1.01	1.00	-0.07	0.00313	0.00918		0.00845		0.00832										0.00891		
R		0.473		0.303				0.227		0.198		0.200		0.250		0.231		0.240		-0.236		0.238	
Weighted R		0.299		0.193				0.200		0.139		0.128		0.187		0.155		0.140		0.138		0.154	
s		12.97		8.36				2.95		3.54		2.82		4.10		3.40		3.08		3.03		3.38	
Estimated s			5.61		4.99		2.30		2.23		2.73		2.38		3.53		3.14		3.10		2.55		3.14
No. of data in L.S. calculation			695		695		790		809		847		755		755		755		755		755		755
Temperature factor not positive definite			6,12		4,6,7 15,18		2,3.4 6,7,8 9,11,12 13,14. 15		2,4,6 7,8,9 10,11,12 13,14 15		1,2,3 6,7,8 11,12,13 14										4,5,6 7,8,10 11,12		

the right, the calculated "corrected" parameters. A blank space in the left column indicates that the last previous calculated value was the starting value for the particular parameter. In the cases of cycles 9 and 10, all of the parameters had the last previous calculated values of cycles 8 and 9 respectively.

The order in which the atoms are listed in Tables II and IV is not the same as that of the paper[1] on the gallium isomorph, but the atom labeling is. In writing the special position symmetry patch for the Busing-Levy[13] IBM 704 least squares refinement program, it is most convenient to list the atoms in general positions first. Then to avoid mistakes in the listing of results, it is best to leave the order the same as that of the output of the program.

In the calculation of structure amplitudes the following atomic scattering factors were used: for O, Al^{3+}, N and C, those of Berghuis *et al*;[14] and for S, those of Viervoll and Ögrim.[15]

In cycles 1 and 2, 895 reflections, all those representing independent forms and observable in the Cu$K\alpha$ sphere, but perhaps not observable with the single-crystal diffractometer, were included. Eight of the parameters were scale factors, all of which were initially equal to 0.6667, one for each value of l from 0 to 6 and the eighth value for all the remaining l values. Also in the first two cycles, isotropic temperature factors were used despite the fact that it was obvious that the thermal motions of the atoms in this crystal must be highly anisotropic.

The starting structural parameters for the first cycle were those given for the gallium isomorph[1] except for changes in the S and Al temperature factors and the y-parameter of N(II), which was inadvertently taken as 0.418 instead of 0.333. Now it may be seen in Table II under cycle 1, that this y-parameter did not change as radically as one might have hoped, in fact as one might have expected, for an incorrect parameter. But the temperature factor of the atom did increase considerably, perhaps indicating that the atoms did not want to be at the positions indicated. On the other hand, the temperature factor of the N in the special position decreased considerably to a negative value as if to compensate for the other. This, in retrospect, was already indicative of strong interaction between the thermal parameters of these two atoms. Another important change was the large one, to -0.392, in the value of the O(III) z-parameter; this implies a very short S—O distance, 1.31 Å, in one set of the SO_4 groups.

The *estimated* error of fit[13,6] at the end of the least squares calculation of cycle 1 was very much lower than the first computed error of fit,[13] and it appeared that by readjustment of some of the temperature factors we could go a step further toward convergence before changing to anisotropic temperature factors. Initially cycle 2 showed that even with the readjustment of temperature factors, the R value† had dropped from 0.473 to 0.303, the weighted R from 0.299 to 0.193. But the error of fit was higher than that estimated in cycle 1 on the basis, of course, of the parameters computed in that cycle, some of which were physically impossible (i.e., negative temperature factors).

However, cycle 2 ended with an estimated error of fit somewhat lower than that of cycle 1. The N(II) y-parameter decreased toward the value which we believe to be the more nearly correct one, but the N(II) B value increased greatly and the N(I) B value became a large negative value. Also the x-parameter of N(I) decreased to imply an unlikely short C—N distance. Changes in the S and Al positional parameters were not large but several oscillations occurred. The O(III) (atom 10) z-parameter returned to -0.400, but even this value implied a rather short S—O distance, 1.37 Å.

At this point, it seemed necessary to change to anisotropic thermal parameters. The Busing-Levy program will compute these from the isotropic thermal parameters using the following relations: $\beta_{11} = Ba^{*2}/4$; $\beta_{12} = (Ba^*b^* \cos \gamma^*)/4$; etc.

The starting parameters were those computed in cycle 1 and adjusted for cycle 2 (see Table II). For cycle 3, a critical estimate of the reflections really observable by the single-crystal automatic diffractometer was made. This resulted in the removal from the calculation of 43 unobserved reflections, some of which had rather high calculated structure amplitudes when compared with the respective estimated threshold values. Included in cycle 3 was a rejection test: that is, when Δ/σ was >10.00, the reflection was not counted in the calculation of the R values or the standard error of fit, nor was it included in the least squares calculation. This reduced the number of $F_{hk \cdot l}$'s used in the least squares calculation to 790. (Unfortunately the R values and the calculated amplitudes computed in this cycle have been lost.)

The estimated error of fit resulting from the cycle 3 least squares calculation decreased from 4.99 to 2.30, an apparently tremendous improvement. However, the still incorrect N(II) y-parameter did not improve; also the values of the N(II) thermal parameters greatly increased. The O(III) values still implied a short S—O distance. The C(I) z-parameter indicated possible nonplanarity of the guanidinium ion in the special position, but this parameter also had an apparently

† Unless otherwise stated, the R value is that for the independent $F_{hk \cdot l}$'s, i.e., multiplicity is neglected. This is the R value calculated by the Busing-Levy program.

large estimated error, 0.0115, indicative of potential difficulty. Twelve of the atoms had calculated thermal parameters which did not satisfy all the criteria for physical reality (see Ref. 13). Therefore, for cycle 4 some of the thermal parameters had to be adjusted to satisfy these criteria. Also, the N(II) y-parameter was corrected. The R value and error of fit decreased considerably since cycle 2, but the weighted R value increased slightly. The same rejection test as used for cycle 3 allowed 809 reflections to be included in the cycle 4 calculation. The least squares calculation led to an estimated error of fit of 2.23, not too different from that estimated in the previous cycle.

In cycle 4, the values of the N(II) thermal parameters decreased, indicating that the high values had been caused by the wrong y-parameter. One would prefer to think, however, that the y-parameter should have tended to approach the correct value rather than to have the *thermal* parameters act as if the atom should be removed. This time the x-parameter of N(I) (atom 6) became rather large, implying too large a C—N distance. A number of the other positional parameters showed oscillation, and again there were twelve atoms which had thermal parameters not satisfying the criteria for physical reality (Table II). The O(III) z-parameter continued to imply a short S—O distance. The C(II) and N(II) atoms did not have the same values in z-parameter, nor did the C(I) and N(I) atoms have the same z-parameter. Also, in this cycle many of the scale factors, especially s_8, had almost reached their starting values after having at first decreased substantially.

The necessary adjustments were made on the thermal parameters before cycle 5 was carried out. Also, the rejection test was removed. Five reflections which appeared to have substantial contribution from the 54 hydrogen atoms or to have suffered from extinction were given zero weight. Thus, of the 852 reflections, 847 were used in the cycle 5 least squares calculation. Because some of the initially estimated $\sigma(F_o)$'s were really very small, a few of these also were readjusted. Initially the R value was 0.198, while the weighted R decreased to 0.139, this latter reduction resulting mostly perhaps from the few adjustments made on the $\sigma(F_o)$'s. The error of fit for the 847 reflections was larger than for the 809 of the previous cycle. The calculated estimated error of fit after the least squares calculation did decrease, however.

But in cycle 5 there was no improvement in the way the calculation was going. There were further oscillations, and, very important, the C(II)—N(II) distance continuing from cycle 3 was short, whereas the C(I)—N(I) distance continued to be long. Considering the guanidinium ion to be planar, the C—N distances were respectively 1.22 and 1.48 Å, the average is 1.35 Å in good agreement with the acceptable guanidinium C—N value 1.34 Å.[16] Again this indicated interaction between the N(II) x- and y-parameters and the N(I) x-parameter. Also, the parameter values of the S(I) and O(III) atoms still indicated an improbably short S—O distance. There were other indications of interaction: for example, the y-parameters of the O(V) and O(VI) atoms (2 and 3 respectively) behaved strangely, that of O(V) indicating an improbably large [SO_4] O—O distance, that of O(VI), too small an [SO_4] O—O distance.

It seemed at the time, however, that there might be other possibilities for explaining the course of events in the attempt to refine the structure. For example, there could be many reflections to which the hydrogen atoms would contribute, and, perhaps particularly because this is a noncentrosymmetric structure, the affected structure amplitudes were having a detrimental effect. Therefore, in cycle 6 all reflections for which $\sin^2/\lambda^2 < 0.0800$ were given zero weight. Necessary adjustments were made in thermal parameters (Table II); the N(I) and N(II) positional parameters were readjusted each to yield the C—N distance 1.34 Å; and the O(V) and O(VI) y-parameters were adjusted to yield more reasonable [SO_4] O—O distances. The R value for the 755 amplitudes (with nonzero weights) was 0.200, weighted $R = 0.128$ and error of fit, 2.82.

In the cycle 6 least squares calculation, only 43 parameters were varied: the scale factors and all positional parameters except the N(I) x-parameter. The estimated error of fit decreased to 2.38, but this cycle was also discouraging in that again there were oscillations and some rather large changes in parameter. The S(I)—O(III) distance continued to remain improbably short; the O(VI) y-parameter again implied too short an [SO_4] O—O distance; and the values of the N(II) x- and y-parameters implied a C(II)—N(II) distance of 1.25 Å.

In the paper on the gallium isomorph,[1] we had concluded that the arrangement of the guanidinium ions on the threefold axes were related to that at 3m to close approximation by $\frac{1}{3},\frac{2}{3},0$ and $\frac{2}{3},\frac{1}{3},0 - (u,0,w; 0,u,w; \bar{u},\bar{u},w)$. However, some doubt remained, and therefore it was decided to try some different orientations of the guanidinium groups.

For cycle 7, the N(II) parameters were readjusted, presumably back to the starting parameters of cycle 6. However, a card-punch error (0.5333 instead of 0.5533) was made in the x-parameter. The N(I) parameter was set to -0.1130. This we shall call the $(-,-)$† orienta-

† This symbolism is derived as follows: The ± orientations of N(I) are those for which in $(x,0,z)$ of positions $3c$, $x_{N(I)} = \pm u$ where u is very nearly $+0.113$. The ± orientations of N(II) are those for which in (x,y,z) of positions $6d$, $x_{N(II)} = \frac{1}{3} \pm u$, $y = \frac{2}{3}$. Thus $(-,-)$ here means that $x_{N(I)} = -0.113$, $x_{N(II)} = 0.220$, $y_{N(II)} = 0.667$. By symmetry the latter two are equivalent to 0.553 and 0.333 respectively.

tion. The positional parameters of O(VI) were also readjusted. The R value for the 755 reflections increased to 0.250, the weighted R to 0.187, and the error of fit to 4.10. In cycle 7 all scale and positional parameters were varied. At the end of the cycle, the estimated error of fit was 3.53. The C(II)—N(II) distance again was too short, ~1.21 A; again the O(VI) y-parameter decreased from the adjusted value; the difference in the C(I) and N(I) z-parameters increased. Also again there were oscillations. The results of cycle 7 did not look promising.

In cycle 8, the $(+,+)$ arrangement of the guanidinium ions was tried with the other starting parameters the same as those used in cycle 7. In this case the R value for the 755 amplitudes was 0.231, weighted R, 0.155, and error of fit, 3.40. Again only scale and positional parameters were varied. The estimated error of fit obtained at the end of the least squares calculation was 3.14. The results of this cycle looked promising. The C—N distances looked good; the O(V) and O(VI) parameters were not too bad. However, the S(I)—O(III) distance still looked improbably short. The agreement for individual amplitudes actually did not look as good as it did in cycle 6, but it was felt that perhaps some of this poorer agreement resulted from hydrogen contributions and/or from required changes in thermal parameters.

It was decided to continue to cycle 9 using the values of scale and positional parameters obtained in cycle 8. The R value increased to 0.240; the weighted R value decreased to 0.140; the error of fit was very close to that previously estimated. Despite this, the parameter results of this cycle (Table II) looked even better than those of the previous cycle, but the S(I)—O(III) distance continued to be improbably short.

The scale and positional parameters resulting from cycle 9 were used in cycle 10. There was not much change in R, weighted R or error of fit. In cycle 10, all scale and positional parameters which had changed less than 1σ in cycle 9 were held constant and all thermal parameters were allowed to vary. The estimated error of fit at the end of the cycle was 2.55. It appeared that the thermal parameters of the N(II) atom increased considerably as if trying to eliminate this atom, and as before this seemd to be an indication that the N(II) atom was not placed correctly. Also as if to compensate, the previously large β_{33} of N(I), 0.01480, decreased to -0.00095. Eight of the atoms had thermal parameter matrices which were not positive definite.

With this continued disappointment, another notion became more important. Was it possible that the structure given by Varfolomeeva *et al*[3] was correct? It seemed advisable to make the calculation with the model proposed by those authors. The results proved that the structure cannot possibly be correct. The initial R was 0.559, weighted R, 0.473 and error of fit, 10.38 for the 755 reflections. Examination of the calculated and observed amplitudes showed a great many very large discrepancies indicative of an improbable structure. Only the scale and positional parameters were varied in the least squares calculation. Thermal parameters for the N atoms were those initially used in cycle 6. All other thermal parameters were essentially those obtained in cycle 10 with necessary adjustments made. The initial and final positional parameters are shown separately in Table III. The estimated error of fit was 8.92, indicating no real possibility of convergence. The parameter changes were mostly drastic. The N(I) x-parameter, for example, would imply a C(I)—N(I) distance of 1.16 Å. Interestingly enough, the S(I)—O(III) distance continued to remain very short.

In cycle 12, the guanidinium ions on the two three-fold axes (i.e., at $\frac{2}{3}\,\frac{1}{3}$ and $\frac{1}{3}\,\frac{2}{3}$) were turned 30° from their original positions. The thermal parameters were the same as those used initially in cycle 11 and are shown in the next to the last columns of Table II. The R value was 0.238, weighted R, 0.154, and error of fit, 3.38 (the latter two being somewhat higher than for the starting parameters of cycle 10). The estimated error of fit obtained from the least squares calculation was 3.14. The results of this calculation did not look promising. The C(I)—N(I) distance was large; there was an extraordinarily large change in the z-parameter of O(VIII). Also, agreement of many individual amplitudes was poorer than for the very first orientation of the guanidiniums. In fact, from the calculations of cycles 7–10 and cycle 12, it had become apparent that the $(+,-)$ orientation was indeed the best. It also appeared that disorder or rotation† of the guanidinium ions was highly unlikely unless very subtle. In the case of complete disorder or the equivalent free rotation, there would be no contributions from the nitrogen atoms to the amplitudes $F_{hk\cdot l}$, $h - k \neq 3n$, exactly as in the case of the $(+,+)$ orientation. This alone makes it appear that the originally reported[1] $(+,-)$ orientation of the guanidinium ions was corroborated.

In cycle 12, the normal equations and inverse matrices were obtained.[13] Examination of the inverse matrix showed that there were large values of correlation coefficients, $\rho_{ij} = b_{ij}/\sqrt{b_{ii}b_{jj}}$, for many pairs of parameters. A few examples are:

† Two reports[17,18] based on nuclear magnetic resonance investigations of G.A.S.H. mention the possibility of rotation of the guanidinium groups. We have learned (by private communication) from, and have been permitted to quote, the author, D. W. McCall, of one of these,[17] that further investigation now indicates that this rotation is highly unlikely.

TABLE III — POSITIONAL PARAMETERS. CYCLE 11

Atom	Coordinates				Atom	Coordinates		
	x	y	z			x	y	z
1 - N(II)	0.2200	−0.3330	0.0000	Initial	10 - O(III)	0.3699	0	−0.0820
	0.2047	−0.3304	−0.0305	Final		0.3651	0	−0.1185
2 - O(V)	0.3449	0.1166	−0.3130	Initial	11 - O(II)	−0.4412	0	0.2764
	0.3512	0.1250	−0.3122	Final		−0.4246	0	0.2739
3 - O(VI)	−0.3211	−0.1137	0.2450	Initial	12 - O(I)	0.4538	0	−0.3260
	−0.3281	−0.1123	0.2025	Final		0.4688	0	−0.3351
4 - O(IX)	−0.4654	0.3272	0.3400	Initial	13 - S(I)	0.3479	0	−0.2433
	−0.4639	0.3327	0.3434	Final		0.3469	0	−0.2502
5 - O(X)	0.4647	−0.3391	0.5600	Initial	14 - S(II)	−0.3174	0	0.3144
	0.4647	−0.3422	0.5294	Final		−0.3200	0	0.3079
6 - N(I)	0.1132	0	0.4500	Initial	15 - C(II)	$\frac{1}{3}$	$\frac{2}{3}$	0.0000
	0.0987	0	0.4124	Final		$\frac{1}{3}$	$\frac{2}{3}$	0.0110
7 - O(VIII)	−0.1304	0	−0.1208	Initial	16 - Al^{3+}(II)	$\frac{1}{3}$	$\frac{2}{3}$	0.4430
	−0.1375	0	−0.1073	Final		$\frac{1}{3}$	$\frac{2}{3}$	0.4396
8 - O(VII)	0.1426	0	0.1175	Initial	17 - C(I)	0	0	0.4500
	0.1367	0	0.1260	Final		0	0	0.4694
9 - O(IV)	−0.2912	0	0.4756	Initial	18 - Al^{3+}(I)	0	0	0
	−0.3180	0	0.4471	Final		0	0	0

$z_{O(V)} - z_{O(II)}$,	0.81
$z_{O(V)} - z_{O(I)}$,	0.58
$z_{O(IX)} - z_{O(VII)}$,	0.84
$z_{O(IV)} - z_{O(III)}$,	0.65
$z_{S(I)} - z_{S(II)}$,	0.96.

It is noteworthy that the correlation coefficient for $x_{N(II)}—x_{N(I)}$ was very low, 0.10; it will be seen later that this low value resulted from the incorrect orientation of the guanidinium (II) ions.

It seemed unlikely that the weighting scheme could be the cause of the difficulties encountered. Nevertheless, it was decided to try a weighting scheme radically different from that used in the first twelve cycles.

In cycle 1′ (Table IV), all amplitudes with $\sin^2\theta/\lambda^2 < 0.0800$ were still weighted zero. Also all unobserved amplitudes were to be weighted zero and all observed, unity. However, a number of amplitudes which should have been weighted zero were weighted unity, and a few which should have been weighted unity were weighted zero. This left 534 reflections included in the least squares calculation. The initial parameters were those from cycles 9 and 10, except for the N's which were started at the exact $(+,-)$ orientation and the O(III) z-parameter which was started at -0.405 to give an S—O value closer to 1.48 Å. The R value was 0.204, weighted R, 0.149 and error of fit 2.19 for the 534 amplitudes and these parameter values. The least squares calculation gave an estimated error of fit of 1.90. Again the S(I)—O(III) distance decreased to 1.38 Å, the C(I)—N(I) distance increased again and the C(II)—N(II) distance decreased again. Some of the other distances are listed in Table V.

Starting with this calculation, the vector $v_i = \Sigma(\sqrt{w}D_i)(\sqrt{w}\Delta)$ was obtained as output† as well as the direct and inverse matrices,[13] the purpose being to see whether Δp_i's from the diagonal term approximations would be much different from those obtained by the exact solution of the normal equations. Not many of these were checked in this and subsequent cycles, but enough differences were found to indicate the importance of the off-diagonal terms.

It appeared that it would be most convenient to have the correlation or normalized inverse matrix to examine in each cycle. A program patch to enable us to do this was written by Misses D. C. Leagus and B. B. Cetlin.

† The program patch for this calculation was written by Miss D. C. Leagus.

TABLE IV — RESULTS OF LEAST SQUARES CALCULATIONS (SECOND SET OF WEIGHTS)

Parameters		Cycle 1'		Cycle 2'		Cycle 3'		Cycle 4'			Cycle 5'						
									a	b		a	b	c	d	σ	Par.N
Scale Factors	1	0.5557	0.5846		0.5627										0.5420	0.0079	1
	2	0.6186	0.6058		0.6069										0.5620	0.0086	2
	3	0.5933	0.5951		0.6032										0.5472	0.0072	3
	4	0.6161	0.5841		0.5903										0.5519	0.0092	4
	5	0.6449	0.6503		0.6539										0.5724	0.0136	5
	6	0.7327	0.6826		0.6849										0.6089	0.0208	6
	7	0.6854	0.6689		0.6760										0.5779	0.0279	7
	8	0.6856	0.6693		0.6751										0.5649	0.0427	8
Atom 1 N (II)	x	0.2200	0.2211	0.2184	0.2237	0.2200	0.2171		0.2148	0.2155		0.2169	0.2169	0.2146	0.2176	0.0027	9
	y	-0.3330	-0.3377		-0.3398	-0.3330	-0.3386		-0.3412	-0.3415		-0.3147		-0.3419	-0.3432	0.0026	10
	z	0.5350	0.5465		0.5521		0.5579		0.5552	0.5559		0.5591	0.5590	0.5580	0.5584	0.0044	11
	β_{11}	0.00890					0.00229		0.00431	0.00586		0.00424	0.00422		0.00456	0.00255	12
	β_{22}	0.00700					0.00970		0.01172	0.01386		0.01191	0.01222		0.01221	0.00330	13
	β_{33}	0.01820					0.02051		0.02269	0.02108					0.01509	0.00470	14
	β_{12}	0.00500					0.00227		0.00345	0.00427		0.00494	0.00508		0.00571	0.00201	15
	β_{13}	0.00520					0.00293		0.00185	0.00170		0.00228	0.00226		0.00191	0.00233	16
	β_{23}	0.00410					0.00039		0.00181	0.00217		0.00299	0.00309		0.00292	0.00269	17
Atom 2 O (V)	x	0.3449	0.3412		0.3429					0.3419		0.3430	0.3428	0.3423	0.3428	0.0025	18
	y	0.1166	0.1178		0.1167					0.1201		0.1174	0.1173	0.1140	0.1165	0.0022	19
	z	-0.1977	-0.2072		-0.2076					-0.2043		-0.2005	-0.2002	-0.2008	-0.2056	0.0043	20
	β_{11}	0.00843								0.00741		0.00689			0.00517	0.00248	21
	β_{22}	0.00307								0.00010	0.00300	0.00335			0.00244	0.00173	22
	β_{33}	0.02193								0.2388		0.02161			0.00999	0.00473	23
	β_{12}	0.00375								0.00267		0.00396	0.00402		0.00357	0.00153	24
	β_{13}	-0.00333								-0.00262		-0.00054	-0.00085		-0.00060	0.00230	25
	β_{23}	-0.00206								-0.00217					-0.00255	0.00207	26
Atom 3 O (VI)	x	-0.3211	-0.3225		-0.3243					-0.3215		-0.3233	-0.3231	-0.3230	-0.3232	0.0022	27
	y	-0.1137	-0.1155		-0.1162					-0.1143		-0.1149	-0.1148	-0.1166	-0.1155	0.0021	28
	z	0.2450	0.2437		0.2447					0.2385		0.2427	0.2436	0.2444	0.2365	0.0042	29
	β_{11}	0.00287								0.00027	0.00200	-0.00048	-0.00061		-0.00124	0.00163	30
	β_{22}	0.00149								0.00238		0.00261			0.00162	0.00180	31
	β_{33}	0.01905								0.02399		0.03084	0.03061		0.01944	0.00487	32
	β_{12}	0.00066								-0.00127		-0.00077	-0.00099		-0.00050	0.00113	33
	β_{13}	-0.00081								0.00014		0.00118	0.00117		0.00159	0.00176	34
	β_{23}	-0.00057								-0.00069					-0.00363	0.00216	35
Atom 4 O (IX)	x	-0.4654	-0.4682		-0.4701					-0.4702				-0.4688	-0.4717	0.0021	36
	y	0.3272	0.3259		0.3258					0.3255				0.3263	0.3245	0.0016	37
	z	-0.0520	-0.0573		-0.0571					-0.0629		-0.0549	-0.0552	-0.0581	-0.0515	0.0039	38
	β_{11}	0.00266								0.00357		0.00522	0.00522		0.00403	0.00166	39
	β_{22}	0.00120								0.00070		0.00379	0.00388		0.00323	0.00146	40
	β_{33}	0.01613								0.01697		0.00943	0.00926		0.00534	0.00401	41
	β_{12}	0.00127								0.00089		0.00271	0.00272		0.00218	0.00107	42
	β_{13}	-0.00122								0.00017		-0.00212	-0.00213		-0.00128	0.00172	43
	β_{23}	0.00174								0.00193					0.00194	0.00124	44
Atom 5 O (X)	x	0.4647	0.4669		0.4661					0.4626		0.4639	0.4637	0.4634	0.4630	0.0022	45
	y	-0.3391	-0.3377		-0.3373					-0.3369		-0.3377	-0.3377	-0.3374	-0.3372	0.0017	46
	z	0.1758	0.1748		0.1759					0.1735		0.1784	0.1785	0.1753	0.1798	0.0044	47
	β_{11}	0.00200								0.00383		0.00285	0.00278		0.00194	0.00187	48
	β_{22}	0.00142								0.00269		0.00048	0.00039		-0.00055	0.00180	49
	β_{33}	0.01636								0.01668					0.01066	0.00465	50
	β_{12}	0.00135								0.00189		0.00075	0.00071		0.00029	0.00131	51
	β_{13}	0.00007								-0.00176		0.00105	0.00097		0.00029	0.00190	52
	β_{23}	-0.00150								0.00049		-0.00002	0.00000		-0.00090	0.00143	53
Atom 6 N (I)	x	0.1132	0.1192	0.1132	0.1218	0.1130	0.1178	0.1130	0.1163	0.1172		0.1158	0.1157	0.1139	0.1159	0.0026	54
	z	0.5065	0.4959	0.5000	0.5043		0.5100		0.5055	0.5057		0.5110	0.5111	0.5100	0.5110	0.0044	55
	β_{11}	0.0038					0.00791		0.00634	0.00600		0.00658			0.00488	0.00295	56
	β_{33}	0.01480					0.01356		0.01021	0.01008		0.00970			0.00866	0.00430	57
	β_{12}	0.00320					0.00934	0.00600	0.01036	0.01189	0.00500	0.00792	0.00711		0.00646	0.00412	58
	β_{13}	0.00000					-0.01098	-0.00500	-0.00476	-0.00453					-0.00307	0.00432	59
Atom 7 O (VIII)	x	-0.1304	-0.1284		-0.1284					-0.1348		-0.1306	-0.1309	-0.1321	-0.1317	0.0025	60
	z	-0.1208	-0.1209		-0.1212					-0.1187		-0.1155	-0.1156	-0.1165	-0.1155	0.0053	61
	β_{11}	0.00649								0.00105	0.00400	0.00484	0.00501		0.00464	0.00265	62
	β_{33}	0.01200								0.01251					0.00971	0.00512	63
	β_{12}	0.00514								-0.00147	0.00300	0.00361	0.00376		0.00401	0.00354	64
	β_{13}	-0.00700								-0.00487		-0.01211	-0.01208		-0.01273	0.00380	65
Atom 8 O (VII)	x	0.1426	0.1366		0.1333					0.1333				0.1347	0.1330	0.0023	66
	z	0.1175	0.1184		0.1185					0.1123		0.1180	0.1176	0.1169	0.1198	0.0039	67
	β_{11}	0.00728								0.00459		0.00087	0.00071		0.00014	0.00228	68
	β_{33}	0.00710								0.00457		0.01211	0.01194		0.01199	0.00309	69
	β_{12}	0.00710								0.00414		0.00045	0.00021		-0.00011	0.00304	70
	β_{13}	-0.00545								-0.00608		-0.00294	-0.00276		-0.00440	0.00292	71
Atom 9 O (IV)	x	-0.2912	-0.2927		-0.2908					-0.2943		-0.2941		-0.2933	-0.2935	0.0020	72
	z	0.4756	0.4738		0.4762					0.4723		0.4755	0.4755	0.4725	0.4814	0.0039	73
	β_{11}	0.00823								0.00792					0.00342	0.00311	74
	β_{33}	0.00784								0.00919		0.00748	0.00731		0.00627	0.00324	75
	β_{12}	0.00627								0.00788		0.00688	0.00686		0.00192	0.00398	76
	β_{13}	0.00209								0.00155					-0.00473	0.00327	77
Atom 10 O (III)	x	0.3699	0.3707		0.3695		0.3692		0.3707	0.3741		0.3734	0.3736	0.3736	0.3755	0.0025	78
	z	-0.4050	-0.3958		-0.3940	-0.4050	-0.3928		-0.3946	-0.3974		-0.3919	-0.3921	-0.3954	-0.3878	0.0043	79
	β_{11}	0.00651					0.00743		0.00772	0.01052		0.01176	0.1206		0.00812	0.00357	80
	β_{33}	0.01627					0.01411		0.01235	0.01374					0.00369	0.00462	81
	β_{12}	0.00640					0.00495		0.00596	0.00862		0.01144	0.1182		0.00809	0.00467	82
	β_{13}	0.00233					0.00652		0.00401	0.00653					-0.00179	0.00437	83
Atom 11 O (II)	x	-0.4412	-0.4433		-0.4428					-0.4402		-0.4444	-0.4445	-0.4475	-0.4450	0.0023	84
	z	0.2764	0.2773		0.2752					0.2821		0.2822		0.2821	0.2803	0.0041	85
	β_{11}	0.00200								-0.00019	0.00200	0.00024	0.00009		0.00067	0.00180	86
	β_{33}	0.01777								0.00976		0.01433	0.01294		0.01353	0.00356	87
	β_{12}	-0.00009								-0.00144		0.00209	0.00165		0.00292	0.00315	88
	β_{13}	0.00400								-0.00003		0.00245	0.00216		0.00258	0.00319	89
Atom 12 O (1)	x	0.4538	0.4553		0.4533					0.4581		0.4567	0.4568	0.4551	0.4566	0.0021	90
	z	-0.1611	-0.1629		-0.1625					-0.1669		-0.1632	-0.1628	-0.1607	-0.1674	0.0040	91
	β_{11}	0.00200								0.00283		0.00427	0.00424		0.00473	0.00241	92
	β_{33}	0.00910								0.01021		0.00641	0.00666		0.00790	0.00369	93
	β_{12}	0.00000								0.00377		0.00603	0.00593		0.00677	0.00355	94
	β_{13}	0.00200								0.00134					0.00599	0.00404	95
Atom 13 S (I)	x	0.3479	0.3476		0.3477					0.3477				0.3474	0.3474	0.0008	96
	z	-0.2433	-0.2447		-0.2428					-0.2414		-0.2433	-0.2432	-0.2423	-0.2450	0.0030	97
	β_{11}	0.00163								0.00174					0.00118	0.00077	98
	β_{33}	0.00714								0.00684		0.00661	0.00655		0.00502	0.00126	99
	β_{12}	0.00067								0.00158		0.00155			0.00126	0.00105	100
	β_{13}	-0.00041								-0.00229		-0.00141	-0.00148		0.00003	0.00099	101
Atom 14 S (II)	x	-0.3174	-0.3189		-0.3187					-0.3187				-0.3183	-0.3189	0.0008	102
	z	0.3144	0.3123		0.3137					0.3156		0.3135	0.3136	0.3147	0.3112	0.0030	103
	β_{11}	0.00194								0.00223		0.00215			0.00172	0.00087	104
	β_{33}	0.00756								0.00871		0.00982	0.00987		0.00558	0.00181	105
	β_{12}	0.00082								0.00209		0.00179	0.00186		0.00198	0.00115	106
	β_{13}	0.00017								-0.00142		-0.00048	-0.00048		0.00069	0.00106	107
Atom 15 C (II)	z	0.5350	0.5529		0.5467					0.5338	0.5500	0.5535	0.5525	0.5432	0.5520	0.0091	108
	β_{11}	0.00238								0.00371		0.00343	0.00336		0.00208	0.00105	109
	β_{33}	0.01029								0.00672		0.00244	0.00236		-0.00152	0.00364	110
Atom 16 Aℓ (II)	z	0.0572	0.0571		0.0575					0.0573				0.0574	0.0587	0.0018	111
	β_{11}	0.00189								0.00089		0.00102			0.00022	0.00026	112
	β_{33}	0.00770								0.01072		0.00986	0.01009		0.00451	0.00235	113
Atom 17 C (I)	z	0.5072	0.5050		0.5060					0.4887	0.5000	0.5015		0.4948	0.4994	0.0088	114
	β_{33}	0.01033								0.01027		0.02050	0.02009		0.01523	0.00689	115
Atom 18 Aℓ (I)	β_{33}	0.00891								0.00630		0.00614			0.00511	0.00211	116
R		0.204		0.176		0.177		0.167			0.167						
Weighted		0.149		0.119		0.117		0.102			0.097						
s		2.19		1.85		1.90		1.69			1.60						
Estimated s			1.90		1.86		1.79		1.68	1.59		1.50	1.48	1.58	1.40		
No. of data in L.S. calculation			534		496		568		546	546		546	546	546	546		
Temperature factor not positive definite							6		6	2,3,6 7,8,11 12		3,6,7 8,11,12	3,6,7 8,11,12		2,3,5 6,7,8 9,11,12 13,14,15		

TABLE V — SOME INTERATOMIC DISTANCES OBTAINED FROM LEAST SQUARES CALCULATIONS (SECOND SET OF WEIGHTS)

Distance	Cycle 1′ Å	Cycle 2′ Å
C(I)—N(I)	1.40	1.43
C(II)—N(II)	1.29	1.25
S(I)—O(V)	1.46	1.44
S(I)—O(III)	1.38	1.38
S(I)—O(I)	1.46	1.44
S(II)—O(VI)	1.47	1.48
S(II)—O(IV)	1.48	1.50
S(II)—O(II)	1.50	1.49
Al(I)—O(VII)	1.92	1.89
Al(I)—O(VIII)	1.86	1.86
Al(II)—O(IX)	1.90	1.92
Al(II)—O(X)	1.91	1.91

In cycle 2′ the starting parameters were the same as those resulting from cycle 1′ (new weights) except for the x-parameters of N(I) and N(II) and the z-parameter of N(I). Also, it was found that under the conditions set for the weighting in cycle 1′, only 496 amplitudes should have been weighted unity. For these reflections and the starting parameters shown in Table III, the R value was 0.176, weighted R, 0.119 and error of fit, 1.85. Again only scale and positional parameters were allowed to vary. Changes were not large except for the N and C(II) parameters. Some distances calculated from these parameters are given in Table V. (C—N distances are always on assumption of planarity of the guanidinium group.) Note that again the C(I)—N(I) distance is short, the C(II)—N(II) long, but the average is the expected value for such a bond. Also noteworthy is the continued tendency of S(I)—O(III) to be short. In fact, there is a tendency throughout for the S(I)—O distances to be shorter on the average than the S(II)—O distances. Examining the correlation matrix for this cycle we may summarize the results as follows (Table VI). Only those pairs for which $|\rho| \geqq 0.40$ are listed. Thus of the 946 ρ_{ij} $(i \neq j)$ terms only 75 are $\geqq 0.40$. Important also is the fact that a large number, 677, of the terms are less than 0.10, many *much* less than 0.10; 194 of the $|\rho_{ij}|$ lie between 0.10 and 0.40. These could be important especially if one parameter has many interactions of moderate size with other parameters.

Earlier we gave some examples of $|\rho_{ij}|$ that were calculated from the inverse matrix of cycle 12 (old weights). It is seen from examination of Table VI that the values for the particular $|\rho_{ij}|$ obtained from cycle 2′ are essentially the same except for the value for the $x_{N(II)}$—$x_{N(I)}$ interaction. The value is much higher, 0.62, than the one, 0.10, obtained

TABLE VI — CORRELATION COEFFICIENTS FROM CYCLE 2′ (ONLY $|\rho_{ij}| > 0.4$ ARE LISTED)

$\lvert\rho\rvert$	$-j_1, j_2, j_3, \cdots$
0.40–0.50	11–17,20,23,27,29,37; 14–41; 17–20,23,27,29,31,35,43; 20–25,33,37; 23–31,33,37; 25–29,39; 27–31,37,43; 29–31,33,37; 31–37; 33–39; 35–37
0.50–0.60	9–10; 11–39,41,43; 12–15; 14–17,37; 17–39,41; 18–19; 20–31,43; 21–22; 23–43; 27–29,39,41; 29–39,41,43; 31–41,43; 33–41; 37–39,41,43
0.60–0.70	9–24; 11–25; 13–34; 16–36; 18–28; 20–23,27,35,39,41; 21–26; 23–29,41; 31–33,39; 38–40
0.70–0.80	23–39; 39–43; 41–43
0.80–0.90	14–35; 17–37; 20–29; 23–27; 42–44
0.90–1.00	39–41

Parameter Numbers

Parameter Number											
	s_1 1	s_2 2	s_3 3	s_4 4	s_5 5	s_6 6	s_7 7	s_8 8	N(II):x 9	y 10	z 11
	O(V):x 12	y 13	z 14	O(VI):x 15	y 16	z 17	O(IX):x 18	y 19	z 20		
	O(X):x 21	y 22	z 23	N(I):x 24	z 25	O(VIII):x 26	z 27	O(VII):x 28	z 29	O(IV):x 30	z 31
	O(III):x 32	z 33	O(II):x 34	z 35	O(I):x 36	z 37	S(I):x 38	z 39	S(II):x 40	z 41	C(II):z 42
	Al(I):z 43	C(I):z 44									

from the incorrect orientation of the guanidinium ions. Thus, *incorrect values for parameters can uncouple parameters.* Furthermore, this appears to be the reason that there was not much change in the incorrect y-parameter of N(II) in the first three cycles. That is, a parameter which is given a value which tends to make it independent may not change *rapidly* to a value which tends to make it dependent.

The purpose of the next cycle was to see the results of allowing the parameters, both positional and thermal, of only the N and O(III) atoms to vary. Before carrying out this calculation, however, the positions of hydrogen atoms were estimated. The guanidinium ions were considered to be essentially planar, and the z-parameters of the guanidinium H's taken as 0.55 for those about the threefold axes at $\frac{1}{3},\frac{2}{3}$ and $\frac{2}{3},\frac{1}{3}$, and 0.505 for those about the axis at 0,0. For the water molecules, the links with the SO_4 oxygen atoms were considered and the tilt of the water molecule estimated accordingly. In any given level of H_2O molecules about either of the nonequivalent axes, the z's were taken equal. The H—O—H angle was taken as 105° and the O—H distance, 0.96 Å. (The initial H-parameter values will not be listed; however, the last set used will be listed later.) First, H contributions to the $F_{hk\cdot l}$ for h,k,l positive were calculated for two different orientations of the guanidinium ions, namely: $(+,-)$ and $(+,+)$. (The program used for this calculation was written by R. G. Treuting; the atomic scattering factors for H were those of Viervoll and Ögrim.[15]) These calculations, together with consideration of previous calculations of the amplitudes, corroborated the conclusion that the $(+,-)$ orientation was the most probable one.

The N-parameters were readjusted to yield the most probable C—N distance, and the z-parameter of O(III) was started at -0.405. Those observed amplitudes with $\sin^2\theta/\lambda^2 < 0.0800$ which were not strongly affected by extinction were reweighted unity. The total number of reflections weighted unity was 568. The H atoms were put into the calculation as "fixed atoms" (see Ref. 13) with isotropic temperature factor $B = 3.00$ Å^2. The over-all R value was 0.177, weighted R, 0.117, and error of fit, 1.90.

The results of the least squares calculation are given in Table IV cycle 3′. It is seen that the O(III) z-parameter returned to that of the previous cycle. The N(I) x-parameter increased somewhat, implying a C(I)—N(I) distance of 1.37 Å. The parameters of N(II) imply a C(II)—N(II) distance of 1.33 Å.

In Table VII, we list those correlation coefficients greater than or equal to 0.40. If this table is compared with Table VI, one finds that the coupling of N(I) and N(II) positional parameters is still as strong as in the previous cycle. In both cycles 2′ and 3′, the correlation matrices showed no strong interaction between O(III) and nitrogen atom parameters. The correlation matrix of cycle 3′ indicated that there are some very strong interactions in pairs of thermal parameters. As expected, there was corroboration of a strong interaction between the β_{33}'s of the N atoms.

For this case, it might be worthwhile to show the Δp_i's obtained from the complete solution of the 21 normal equations compared with those obtained from the diagonal term approximation. These are given in Table VIII together with the σ's calculated by the Busing-Levy program. As expected, several of the Δp_i's for particular i are quite different, particularly for those which are highly correlated (see Table VII).

Before proceeding to the next cycle, the calculated and observed data were examined for any outstanding discrepancies and rechecks were made on the intensity data. It was found that 27 of the reflections which were listed as observed should have been listed as unobserved. It was also found that 5 reflections which were recorded as unobserved should have been observed by the instrument but were missed. These were obtained from film data.

Slight changes were made in the H-parameters; the x-parameter of N(I) was returned to 0.113 and necessary changes made in the β_{12} and β_{13} thermal parameters of N(I). Now the Busing-Levy program calculates and stores all derivatives, so that it is possible to allow different sets of parameters to remain constant and solve for sets of Δp_i for each initial set of parameters. In cycle 4′a, therefore, we first allowed only the N(I), N(II), and O(III) parameters to vary and then in 4′b,

TABLE VII — CORRELATION COEFFICIENTS FROM CYCLE 3′

$\lvert\rho\rvert$	i,j			
0.40–0.50	1,2	7,12	8,9	8,15
0.50–0.60	4,7			
0.60–0.70	1,10	5,7	5,14	
0.70–0.80	3,11	6,13		
0.80–0.90				
0.90–1.00	12,14	18,20		

Parameter numbers

	x	y	z	β_{11}	β_{22}	β_{33}	β_{12}	β_{13}	β_{23}
N(II)	1	2	3	4	5	6	7	8	9
N(I)	10		11	12		13	14	15	
O(III)	16		17	18		19	20	21	

TABLE VIII — PARAMETER CHANGES AND ERROR ESTIMATES FROM CYCLE 3′

Parameter number (see Table VII)	Busing-Levy	Diagonal term approx.	σ's from Busing-Levy
1	−0.0029	−0.0020	0.0027
2	−0.0056	−0.0027	0.0022
3	0.0058	0.0046	0.0037
4	−0.00661	−0.00105	0.00268
5	0.00270	0.00418	0.00245
6	0.00231	0.00477	0.00420
7	−0.00273	−0.00136	0.00202
8	−0.00227	−0.00333	0.00271
9	−0.00371	−0.00053	0.00225
10	0.0048	0.0051	0.0025
11	0.0057	−0.0030	0.0048
12	0.00411	0.00023	0.00296
13	−0.00124	0.00172	0.00527
14	0.00614	0.00113	0.00466
15	−0.01098	−0.00983	0.00552
16	−0.0003	−0.0009	0.0014
17	0.0121	0.0120	0.0023
18	0.00092	0.00160	0.00228
19	−0.00216	−0.00133	0.00311
20	−0.00145	−0.00230	0.00307
21	0.00419	0.00025	0.00351

varied all parameters except the scale factors. The results are shown in Table IV. Again in both cases, the N(I) x-parameter increased; there were changes in the N(II) parameters, but the implied C—N distance 1.35 Å was good. Also the z-parameter of O(III) seemed to improve, especially when all the parameters were allowed to vary. But in 4′a, the thermal parameter matrix of the N(I) atom was not positive definite, while in 4′b, seven atoms had thermal parameter matrices which were not positive definite. Also there were continued oscillations and large error estimates. It was evident that real convergence would not be attained.

However, because the N and O(III) parameters did look encouraging, it was decided to try one more cycle. This time the parameters of the water hydrogen atoms were recalculated in a somewhat different way. In a recent paper,[19] Aleksandrov, Lundin and Mikhailov report results of a study of the distribution of hydrogen atoms in guanidinium aluminum sulfate hexahydrate by means of proton magnetic resonance experiments. They report that the nearest neighbor p—p (proton-proton) vectors are perpendicular to the a_1, a_2 and a_3 axes.† They argue that on the basis of symmetry considerations all H atoms bonded to O's in a

† Previously, Spence and Muller[18] had reported this to be so for the p — p vectors of the water molecules, but had concluded that the p — p vectors of the guanidinium groups could be parallel to the c-axis with a separation of 2.05 Å.

single octahedron layer about a threefold axis must have the same z-parameter. Of course, this is true only for those hydrogen atoms bonded to N(I) atoms and to the water molecules about the threefold axis at 0,0. The trigonal axes and planes of symmetry are such that only three atoms about the axis at $\frac{1}{3},\frac{2}{3}$ and three about the axis at $\frac{2}{3},\frac{1}{3}$ must have the same value of z.

Thus contrary to the statements of Aleksandrov *et al,*[19] symmetry conditions do *not* require all the nearest neighbor H—H vectors to be parallel to the $(00\cdot1)$ plane, nor must they all be perpendicular to the a_1, a_2 and a_3 axes. Only for those about the threefold axis where the mirror planes intersect, namely at 0,0 must this be the case. However, it is possible that the nearest neighbor H—H vectors about the threefold axes at $\frac{1}{3}$ $\frac{2}{3}$, $\frac{2}{3}$ $\frac{1}{3}$ are *close* to parallelism with the $(00\cdot1)$ plane and perpendicularity to the a_1, a_2, a_3 axes.

Furthermore, Aleksandrov *et al*[19] refer to the trial structure reported by Varfolomeeva *et al.*[3] Although that structure is incorrect, it would have no noticeable effect on the conclusions of Aleksandrov *et al,* since they discuss only the nearest neighbor H—H vectors.

Thus, in calculating the H parameters, the tilting of the water H—H bonds out of the $(00\cdot1)$ plane and skewness to the a_1, a_2, a_3 axes was permitted in those water molecules about the threefold axes at $\frac{1}{3}$ $\frac{2}{3}$, $\frac{2}{3}$ $\frac{1}{3}$. (The guanidinium ions, however, were assumed to be planar.) In calculating the H positions, the water molecules were assumed to lie in the planes connecting the water oxygen atom with the two sulfate oxygen atoms involved in the hydrogen bonding. The bisector of the H—O—H angle of 105° was taken as the line passing through the center of the water oxygen atom and the center of the line connecting the two sulfate oxygen atoms involved. The parameters of the N and O atoms involved were those from cycle 4′b. The H-parameters thus deduced are listed in Table IX. The new parameters caused some differences in the

Table IX — H Parameters Used in Final Cycle

Description	x	y	z
on N(I)(atom 6)	0.205	0.086	0.51
on N(II)(atom 1)	0.465	0.256	0.56
	0.564	0.434	0.56
on O(VIII)(atom 7)	0.139	0.218	−0.148
on O(VII)(atom 8)	−0.072	0.134	0.156
on O(IX)(atom 4)	0.457	0.257	−0.124
	0.526	0.400	−0.111
on O(X) (atom 5)	−0.452	−0.260	0.205
	0.464	0.588	0.219

contributions to several amplitudes, but in general not very important ones.

Some necessary adjustments of thermal parameters resulting from cycle 4′b were made. In cycle 5′a,† only those positional parameters were varied in which changes greater than $\sigma/5$ occurred between previous cycles 2′ and 4′, all thermal parameters were varied in which there were changes greater than $\sigma/5$ between cycles 1′ and 4′; all scale factors were kept constant. In 5′b, only those parameters were varied in which changes in 5′a were greater than $\sigma/5$. In 5′c, only positional parameters were varied. In 5′d all parameters were varied. All results are listed in Table IV. Differences range from very small to very large and are indicative of the unattainability of convergence. We list also the σ's‡ in the Δp_i's for the last cycle 5′d in the last column of Table IV. These are especially large for most of the thermal parameters and for most of the z-parameters, and reflect the strong interdependence in pairs of parameters.

The correlation matrix‡ for cycle 5′d contains 6,670 $\rho_{ij}(i \neq j)$ terms. Thus we shall again only list the values of $|\rho_{ij}| \geqq 0.40$ (Table X). Of the 6,670 terms in the matrix, 176 have values greater than 0.40; 1,389 have values greater than 0.10.

On examining Table X, one finds that no interactions of scale factors with positional parameters are listed. In fact, the correlation coefficients for such combinations are all *very* low. However, there are all the other types of interactions, namely: scale factor-thermal parameter, thermal parameter-thermal parameter, positional parameter-positional parameter, and several (those with asterisk) positional parameter-thermal parameter. Most often, also, the interdependence is between analogous parameters; for example, a z-parameter of an atom interacts with z-parameters of other atoms. Even when a positional parameter interacts with a thermal parameter, an analogy exists, e.g., a z-parameter interacts with a β_{33}-parameter. This makes physical sense, of course, and gives us some confidence that the correlation coefficients reflect the structural interdependence of the parameters. Correlation may be caused partially by the experimental technique§ but it is unlikely to result mainly from the ill-conditioning of the normal equations by a

† It should be kept in mind that all cycles 5′ refer to the derivatives evaluated with the parameters of cycle 5′a.

‡ It is worth emphasizing that statistical theory precludes the use of the error estimates or normal equations matrix to determine the statistical significance of the parameters listed. Only if convergence is actually attained can these numbers be so used. Nevertheless, in a practical way, the error estimates and correlation coefficients do give us important information in the course of refinement or, as in the present case, relative to the unattainability of convergence.

§ X-ray vs neutron diffraction.

TABLE X — CORRELATION COEFFICIENTS FROM CYCLE 5′d.†
(ONLY $|\rho_{ij}| > 0.40$ ARE LISTED)

$\|\rho\|$	$i - j_1, j_2, j_3, \cdots$
0.40–0.50	3–4,5; 4–105; 6–75; 7–75; 8–75; 11–38,47,61,67,85; 15–56; 18–19,34; 20–47‡,61,97,103; 21–30; 24–30,86,88; 25–26,27; 29–47,61; 30–33; 36–37; 38–61,103; 40–51,62; 42–68; 43–66‡; 44–46‡; 45–46; 47–55, 67,73,85,91,103; 51–64; 52–53; 55–97,103; 57–59; 61–73,79,85,91; 63–69; 67–103; 75–81; 77–79‡,83; 79–83‡,97,103; 81–83; 97–108,114; 103–108,114; 105–113; 108–110‡; 113–116
0.50–0.60	4–5,6,7,8; 5–113; 8–99,113; 11–97,103; 12–15; 13–15,58; 16–59; 19–84; 20–91; 22–24; 26–89; 27–94‡; 29–85,97,103; 30–94; 37–53‡; 38–47,97; 39–42; 41–50,63,69; 43–71; 47–97; 48–51; 50–69; 51–62; 61–97,103; 67–97; 73–83‡,97,103; 85–91,97,103; 110–115
0.60–0.70	5–105; 13–56; 14–57; 20–29; 21–24,88; 23–86; 27–92‡; 28–90; 30–92; 37–46; 40–42,49,68,70; 49–64; 52–65; 72–78,80‡,82‡; 74–78‡; 76–78‡; 86–88; 96–104‡,106‡; 98–102‡; 100–102‡; 108–115‡; 111–116‡
0.70–0.80	5–6,7,8; 8–105; 9–54; 18–27; 21–86; 32–93; 35–95; 36–66; 45–60; 49–51,62; 50–63; 96–102; 108–114; 111–113‡
0.80–0.90	6–7,8; 11–55; 25–91; 38–67; 62–64; 73–79; 98–100
0.90–1.00	7–8; 47–61; 56–58; 68–70; 74–76; 80–82; 92–94; 97–103; 104–106

† See last column of Table IV for parameter numbers.
‡ Positional-thermal parameter correlation.

reasonable but not necessarily ideal weighting technique. It will be noticed also that the same pairs of parameters show very nearly the same measure of interdependence as indicated by earlier calculations, again corroborating the point that it is the structural model (including atomic form factors) which causes the interactions.

For the sake of completeness, we show in Table XI a list of observed amplitudes compared with those calculated from the parameters used initially in cycle 5′ and including the contributions of the H atoms with parameters shown in Table IX. Including consideration of multiplicity and the differences when calculated amplitudes are greater than the threshold values (with half the threshold value included in the denominator) for reflections not observed, the discrepancy factor is 0.11.†

The over-all agreement is quite good despite several discrepancies in which a calculated amplitude is above the threshold value for an unobserved reflection.‡ Table XI attests to the validity of the conclusion that the general features of the structure are correct.

† Six amplitudes, those of reflections $30\cdot 0$, $11\cdot 1$, $2\bar{1}\cdot 1$, $22\cdot 1$, $4\bar{2}\cdot 1$ and $2\bar{1}\cdot 2$, suffering from extinction were excluded in calculation of this discrepancy factor.

‡ These are a product of the instrument which sometimes missed reflections, which, according to visual estimates of photographic intensities, it should not have missed.

Table XI — Observed and Calculated Amplitudes Based on Parameters Used Initially in Cycle 5′

ℓ = 0

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
1	0	<5.8	0.8
2	0	<4.0	0.6
3	0	80.8	91.6
4	0	<11.4	3.9
5	0	6.2	8.2
6	0	114.1	124.3
7	0	<15.0	14.8
8	0	8.1	9.7
9	0	123.8	113.4
10	0	<15.2	14.8
11	0	<12.8	5.8
12	0	19.6	19.9
13	0	<5.6	9.1
1	1	27.4	23.7
2	1	18.1	15.4
3	1	28.9	27.8
4	1	57.9	55.3
5	1	22.1	19.6
6	1	31.8	28.5
7	1	38.5	38.7
8	1	35.5	31.9
9	1	22.7	21.1
10	1	<13.7	5.4
11	1	13.9	14.8
12	1	10.8	13.9
2	2	5.1	6.8
3	2	37.9	38.6
4	2	13.6	10.2
5	2	40.0	36.5
6	2	20.6	17.4
7	2	29.2	25.8
8	2	4.7	7.2
9	2	17.4	18.2
10	2	17.4	16.7
11	2	<10.4	10.0
3	3	16.9	22.5
4	3	10.6	10.3
5	3	19.5	19.0
6	3	61.9	58.4
7	3	<8.5	8.2
8	3	<15.4	8.5
9	3	24.2	22.8
10	3	<11.6	6.1
11	3	11.9	14.4
4	4	22.8	20.8
5	4	25.3	23.6
6	4	10.7	12.1
7	4	11.4	12.1
8	4	19.1	18.4
9	4	<12.3	6.1
10	4	<9.0	6.5
5	5	33.5	29.2
6	5	<15.8	2.1
7	5	24.3	21.2
8	5	<13.3	7.6
9	5	16.1	16.5
6	6	19.3	23.7
7	6	9.2	12.5
8	6	<10.3	5.2
7	7	32.4	31.2
8	7	13.7	18.2

ℓ = 1

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
0	0	23.5	20.5
1	0	<4.0	4.1
2	0	16.5	15.4
3	0	35.0	39.5
4	0	10.8	11.0
5	0	19.5	21.2
6	0	30.2	28.8
7	0	23.5	21.5
8	0	37.3	33.1
9	0	14.8	18.7
10	0	21.2	21.4
11	0	15.0	17.2
12	0	<6.7	6.6
13	0	<7.1	9.6
1	1	102.7	151.8
2	1	<6.1	4.0
3	1	<6.9	3.5
4	1	5.6	6.3
5	1	11.9	8.4
6	1	6.0	4.7
7	1	92.7	93.6
8	1	10.3	12.4
9	1	15.0	14.6
10	1	13.6	17.5
11	1	9.3	10.8
12	1	10.0	11.7
2	-1	57.8	71.9
3	-1	14.5	14.3
4	-1	27.0	29.8
5	-1	42.3	42.4
6	-1	28.7	28.1
7	-1	23.1	23.6
8	-1	27.0	23.4
9	-1	18.9	18.8
10	-1	12.5	17.3
11	-1	12.5	12.8
12	-1	13.9	15.5
13	-1	12.4	12.2
2	2	82.2	103.4
3	2	29.6	27.9
4	2	<8.3	2.6
5	2	22.7	21.8
6	2	34.6	30.5
7	2	13.8	15.2
8	2	32.1	28.5
9	2	15.7	18.4
10	2	<9.1	7.9
11	2	18.0	19.4
4	-2	112.2	152.5
5	-2	13.6	15.3
6	-2	9.8	12.0
7	-2	90.1	87.0
8	-2	<10.3	8.7
9	-2	<10.7	10.2
10	-2	48.3	49.1
11	-2	<10.0	10.5
12	-2	<9.1	5.8
13	-2	36.2	35.3
3	3	<9.6	3.1
4	3	<7.7	3.6
5	3	21.8	19.3

ℓ = 1

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
6	3	<11.4	11.7
7	3	14.5	17.9
8	3	13.1	17.0
9	3	<10.1	7.6
10	3	13.1	13.4
11	3	10.1	13.3
6	-3	42.2	42.4
7	-3	36.2	32.8
8	-3	13.2	15.9
9	-3	<11.2	2.6
10	-3	20.8	21.0
11	-3	24.2	22.7
12	-3	<10.1	10.1
13	-3	16.9	18.3
14	-3	5.9	10.9
4	4	28.1	30.4
5	4	<10.6	11.5
6	4	5.8	6.8
7	4	41.0	40.9
8	4	<10.5	8.9
9	4	<8.7	8.4
10	4	3.3	4.0
8	-4	10.6	11.9
9	-4	<10.6	7.6
10	-4	<11.7	10.3
11	-4	21.3	22.0
12	-4	<10.5	5.6
13	-4	12.5	12.1
14	-4	12.1	8.4
5	5	15.8	19.3
6	5	15.0	18.9
7	5	10.4	12.8
8	5	<9.8	3.6
9	5	10.7	12.4
10	-5	15.7	15.7
11	-5	16.0	13.8
12	-5	<10.7	8.5
13	-5	25.6	26.7
14	-5	11.9	12.0
6	6	<11.7	5.6
7	6	10.4	12.5
8	6	24.3	23.7
12	-6	21.3	16.9
13	-6	<9.7	8.9
14	-6	13.3	13.9
7	7	42.3	43.8
8	7	4.2	4.5
14	-7	<8.0	5.8
15	-7	2.9	5.9

ℓ = 2

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
0	0	43.8	51.8
1	0	19.2	19.6
2	0	18.5	16.6
3	0	3.4	4.0
4	0	32.6	31.3
5	0	<12.1	9.4
6	0	8.8	11.5
7	0	24.2	23.7
8	0	<7.1	1.6

ℓ = 2

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
9	0	30.7	25.2
10	0	12.3	13.1
11	0	<6.1	3.9
12	0	<4.4	3.3
1	1	70.4	74.2
2	1	10.4	11.8
3	1	25.2	22.3
4	1	19.2	16.3
5	1	12.4	12.0
6	1	16.7	18.3
7	1	61.6	64.2
8	1	12.0	13.0
9	1	21.1	19.1
10	1	21.0	19.8
11	1	9.2	11.2
12	1	<4.0	8.4
2	-1	131.0	171.6
3	-1	22.7	22.5
4	-1	32.2	32.5
5	-1	111.2	117.5
6	-1	12.5	13.9
7	-1	<6.4	9.0
8	-1	60.6	64.9
9	-1	21.3	21.5
10	-1	<7.0	6.6
11	-1	55.0	51.7
12	-1	19.8	19.5
13	-1	<4.0	5.7
2	2	43.8	43.4
3	2	32.4	31.3
4	2	12.4	11.9
5	2	24.8	28.4
6	2	<8:0	7.6
7	2	<7.5	8.3
8	2	48.1	44.5
9	2	10.8	10.4
10	2	15.3	15.1
11	2	4.0	2.6
4	-2	20.0	22.8
5	-2	18.7	18.3
6	-2	34.5	34.2
7	-2	9.0	14.4
8	-2	7.5	11.8
9	-2	30.4	27.1
10	-2	50.4	47.2
11	-2	<7.0	6.2
12	-2	27.4	24.8
13	-2	15.7	19.8
3	3	14.6	16.2
4	3	<7.1	3.2
5	3	<6.8	9.5
6	3	4.3	4.3
7	3	15.9	19.0
8	3	<9.6	2.8
9	3	20.1	21.0
10	3	<4.5	2.2
11	3	<3.0	2.1
6	-3	30.7	30.0
7	-3	35.4	33.7
8	-3	8.8	12.0
9	-3	23.7	23.7

ℓ = 2

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
10	-3	<7.5	3.2
11	-3	<8.0	6.1
12	-3	<6.7	7.8
13	-3	17.6	15.9
14	-3	<3.0	6.3
4	4	33.5	32.6
5	4	<8.0	9.2
6	4	36.4	33.8
7	4	<8.7	15.3
8	4	13.7	16.3
9	4	14.0	17.1
10	4	<4.0	10.9
8	-4	45.7	46.7
9	-4	<8.1	3.6
10	-4	<8.5	7.8
11	-4	10.8	14.7
12	-4	<7.0	9.5
13	-4	7.1	9.3
14	-4	42.8	40.8
5	5	76.4	81.8
6	5	<8.2	7.0
7	5	<7.2	0.6
8	5	28.5	27.7
9	5	3.4	6.3
10	-5	4.6	4.9
11	-5	10.8	12.1
12	-5	<7.2	10.5
13	-5	22.4	22.0
14	-5	12.0	10.2
6	6	<8.3	5.1
7	6	28.3	23.6
8	6	3.8	4.1
12	-6	<8.1	6.5
13	-6	<6.5	4.2
14	-6	9.4	7.3
7	7	19.4	20.5
14	-7	21.3	23.4

ℓ = 3

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
0	0	96.3	123.8
1	0	27.0	28.0
2	0	4.2	6.6
3	0	48.9	53.4
4	0	19.6	18.2
5	0	<8.4	9.0
6	0	12.3	15.5
7	0	<10.3	6.6
8	0	23.0	20.8
9	0	<10.3	5.6
10	0	19.0	18.4
11	0	6.3	9.9
12	0	25.7	23.3
1	1	82.3	85.0
2	1	<6.8	5.8
3	1	13.9	14.2
4	1	6.0	3.1
5	1	11.4	12.5
6	1	<9.1	3.8
7	1	55.5	52.6
8	1	<11.1	11.0

ℓ = 3

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
9	1	<9.6	1.0
10	1	27.2	25.0
11	1	<6.5	5.6
2	-1	37.9	39.6
3	-1	14.1	13.8
4	-1	8.3	6.3
5	-1	51.9	50.7
6	-1	6.6	11.2
7	-1	<9.1	10.5
8	-1	31.0	25.7
9	-1	<11.1	9.2
10	-1	8.6	12.9
11	-1	38.9	36.4
12	-1	<7.0	3.5
2	2	15.1	17.8
3	2	12.3	15.6
4	2	<8.9	7.7
5	2	10.9	15.5
6	2	<10.7	9.2
7	2	10.5	12.6
8	2	26.2	25.0
9	2	7.4	9.7
10	2	7.7	8.5
11	2	3.0	6.7
4	-2	25.2	24.7
5	-2	33.9	30.6
6	-2	<8.9	5.7
7	-2	31.3	27.4
8	-2	<10.7	1.7
9	-2	<10.5	1.9
10	-2	30.2	27.4
11	-2	6.8	7.1
12	-2	4.2	5.3
13	-2	<6.2	1.6
3	3	14.6	16.3
4	3	<12.2	5.9
5	3	10.8	15.3
6	3	45.4	43.0
7	3	10.5	11.3
8	3	<10.3	8.3
9	3	<10.0	11.9
10	3	8.5	10.5
6	-3	132.5	131.5
7	-3	14.8	16.8
8	-3	7.7	10.0
9	-3	<11.3	12.4
10	-3	<10.2	7.6
11	-3	<10.3	0.4
12	-3	45.8	41.4
13	-3	21.9	20.6
4	4	<11.2	8.2
5	4	25.3	22.2
6	4	<11.4	8.9
7	4	<11.7	13.7
8	4	<9.4	4.4
9	4	<7.4	5.3
8	-4	12.6	18.1
9	-4	<10.7	1.7
10	-4	17.3	17.4
11	-4	8.7	15.9
12	-4	<9.4	5.8

ℓ = 3

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
13	-4	30.8	27.7
5	5	51.1	49.1
6	5	<11.3	6.4
7	5	20.5	20.1
8	5	8.6	13.4
9	5	<6.2	9.8
10	-5	<12.6	13.5
11	-5	<11.3	3.4
12	-5	<9.7	3.6
13	-5	<8.6	5.8
14	-5	<5.4	4.7
6	6	59.6	52.3
7	6	<6.2	8.2
8	6	20.2	19.0
12	-6	<10.6	7.3
13	-6	10.9	12.7
14	-6	<5.8	3.6
7	7	21.0	21.9
14	-7	6.8	9.0

ℓ = 4

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
0	0	71.7	76.2
1	0	<4.2	5.4
2	0	9.4	10.2
3	0	59.7	64.2
4	0	9.7	12.2
5	0	<6.2	4.6
6	0	31.4	31.5
7	0	<7.6	10.0
8	0	<6.3	5.7
9	0	44.8	41.7
10	0	<6.2	6.7
11	0	9.2	13.8
1	1	28.6	24.7
2	1	10.2	12.9
3	1	26.9	26.1
4	1	31.3	32.3
5	1	16.7	18.6
6	1	7.2	7.6
7	1	11.5	12.9
8	1	12.0	15.4
9	1	<5.7	3.2
10	1	14.7	16.2
11	1	19.8	19.4
2	-1	13.1	17.2
3	-1	24.7	23.7
4	-1	27.7	28.2
5	-1	29.2	27.1
6	-1	<6.2	3.3
7	-1	10.4	12.7
8	-1	24.3	25.6
9	-1	<7.1	5.7
10	-1	15.1	15.2
11	-1	<5.0	6.5
12	-1	6.8	6.6
2	2	34.0	33.0
3	2	3.2	1.6
4	2	<6.2	5.1
5	2	<6.9	3.7
6	2	<7.8	7.9

ℓ = 4

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
7	2	10.0	13.2
8	2	<7.3	15.4
9	2	<5.5	5.4
10	2	12.7	13.8
4	-2	46.5	49.5
5	-2	41.7	40.6
6	-7	8.3	10.1
7	-2	19.1	23.1
8	-2	22.2	21.0
9	-2	17.2	17.3
10	-2	7.5	12.8
11	-2	23.4	23.2
12	-2	<4.7	3.1
3	3	22.4	27.3
4	3	<6.9	5.7
5	3	19.7	21.0
6	3	46.6	45.0
7	3	<6.3	3.2
8	3	18.3	19.0
9	3	24.4	25.9
10	3	<5.1	0.6
6	-3	54.0	61.5
7	-3	18.0	19.7
8	-3	15.7	15.7
9	-3	37.6	35.1
10	-3	<6.3	9.9
11	-3	18.1	20.0
12	-3	38.6	36.6
13	-3	<3.6	2.9
4	4	40.2	39.3
5	4	21.5	21.2
6	4	<7.3	6.4
7	4	32.4	29.3
8	4	20.6	20.1
9	4	<4.0	10.0
8	-4	20.3	20.4
9	-4	<7.3	2.5
10	-4	16.1	12.8
11	-4	<7.1	7.3
12	-4	<5.3	4.2
13	-4	<4.0	7.0
5	5	13.8	16.4
6	5	18.6	15.9
7	5	23.7	20.8
8	5	17.7	16.4
10	-5	33.4	31.3
11	-5	7.0	12.6
12	-5	4.7	4.2
13	-5	16.9	17.1
6	6	28.9	28.1
7	6	11.0	10.8
12	-6	8.4	15.6
13	-6	<4.7	3.2

ℓ = 5

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
0	0	36.6	34.3
1	0	20.7	21.4
2	0	<8.2	7.7
3	0	11.2	14.2
4	0	<9.8	2.8

ℓ = 5

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
5	0	16.4	17.4
6	0	21.5	24.0
7	0	34.7	30.1
8	0	<12.3	4.8
9	0	8.0	14.1
10	0	5.9	8.6
11	0	10.6	12.4
1	1	39.4	40.2
2	1	26.3	25.8
3	1	20.0	20.6
4	1	42.7	43.1
5	1	<11.9	4.5
6	1	<11.8	5.2
7	1	11.2	16.8
8	1	5.5	4.1
9	1	<10.5	5.1
10	1	25.7	24.3
11	1	<5.7	5.6
2	-1	54.7	58.7
3	-1	<9.9	4.0
4	-1	<10.6	10.5
5	-1	<11.9	6.8
6	-1	20.6	21.8
7	-1	22.8	21.8
8	-1	29.3	25.6
9	-1	6.4	10.3
10	-1	6.8	10.9
11	-1	<8.5	2.2
12	-1	19.3	19.0
2	2	24.2	22.7
3	2	45.6	44.0
4	2	19.6	19.9
5	2	<12.6	3.4
6	2	11.0	13.7
7	2	17.6	16.6
8	2	<11.3	6.1
9	2	19.9	18.8
10	2	5.1	8.2
4	-2	73.7	78.3
5	-2	<11.4	9.2
6	-2	<11.9	5.4
7	-2	47.6	47.1
8	-2	<13.1	8.7
9	-2	4.2	3.0
10	-2	21.7	22.6
11	-2	4.8	4.3
12	-2	13.7	11.7
3	3	<13.6	3.6
4	3	11.9	13.6
5	3	22.3	20.6
6	3	14.9	17.5
7	3	5.2	8.2
8	3	19.7	17.5
9	3	<9.0	4.4
6	-3	34.1	30.3
7	-3	13.5	13.9
8	-3	26.5	23.1
9	-3	<13.4	11.5
10	-3	<11.7	5.5
11	-3	16.3	16.0
12	-3	11.1	10.2

ℓ = 5

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
4	4	36.6	34.4
5	4	<12.5	6.3
6	4	<12.8	6.2
7	4	41.4	35.7
8	4	6.3	6.7
9	4	<7.3	6.7
8	-4	15.9	15.8
9	-4	22.7	22.8
10	-4	<12.8	5.8
11	-4	13.8	12.9
12	-4	22.9	18.3
13	-4	13.5	12.6
5	5	<14.5	8.9
6	5	22.1	18.2
7	5	15.5	13.6
8	5	<11.0	7.8
10	-5	28.2	29.0
11	-5	15.5	17.1
12	-5	26.5	21.8
13	-5	13.5	14.1
6	6	24.7	20.7
7	6	<7.6	7.3
12	-6	7.1	6.8
13	-6	6.3	9.0

ℓ = 6

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
0	0	27.1	22.8
1	0	14.2	15.6
2	0	14.1	16.5
3	0	12.0	14.6
4	0	22.8	24.0
5	0	<6.2	5.3
6	0	8.4	10.5
7	0	21.2	22.3
8	0	<4.9	5.3
9	0	4.1	5.8
10	0	19.6	18.3
11	0	<2.2	3.7
1	1	<6.0	8.8
2	1	16.1	15.0
3	1	20.0	19.4
4	1	<6.4	6.0
5	1	11.4	12.0
6	1	23.3	23.7
7	1	<5.6	3.0
8	1	<5.4	2.7
9	1	26.6	24.0
10	1	4.5	5.5
2	-1	57.1	56.0
3	-1	<5.4	3.3
4	-1	8.3	10.2
5	-1	45.6	46.8
6	-1	12.2	11.7
7	-1	<5.4	5.8
8	-1	29.5	27.8
9	-1	9.7	11.5
10	-1	5.2	8.3
11	-1	21.6	20.6
2	2	58.7	63.7
3	2	19.4	21.1

ℓ = 6

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
4	2	16.0	16.8
5	2	37.2	36.9
6	2	<6.5	7.7
7	2	<5.2	7.2
8	2	30.1	27.6
9	2	17.7	15.6
4	-2	31.0	30.3
5	-2	19.3	18.4
6	-2	9.6	13.5
7	-2	12.7	15.6
8	-2	<6.5	6.5
9	-2	25.9	23.3
10	-2	<4.4	4.0
11	-2	4.6	6.6
3	3	25.1	24.5
4	3	30.3	29.2
5	3	10.5	11.2
6	3	5.5	10.0
7	3	19.0	18.1
8	3	<4.4	4.6
9	3	<2.8	7.0
6	-3	33.9	31.8
7	-3	14.3	15.0
8	-3	<5.5	7.8
9	-3	<5.9	3.4
10	-3	26.6	24.9
11	-3	7.7	8.6
12	-3	<2.8	7.3
4	4	7.8	8.7
5	4	5.3	5.1
6	4	23.5	19.9
7	4	6.9	10.0
8	4	<3.1	4.0
8	-4	43.1	43.9
9	-4	25.7	24.4
10	-4	14.0	14.9
11	-4	28.5	27.5
12	-4	<3.1	7.6
5	5	30.8	27.2
6	5	<4.6	5.7
7	5	<3.3	6.0
10	-5	<5.0	0.9
11	-5	<4.6	6.2
12	-5	9.1	13.7
6	6	13.5	12.1
12	-6	12.7	11.5

ℓ = 7

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
0	0	26.1	29.0
1	0	<6.5	5.0
2	0	16.2	16.3
3	0	40.1	38.6

ℓ = 7

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
4	0	<4.7	10.1
5	0	<6.1	9.3
6	0	27.5	27.0
7	0	14.7	12.4
8	0	<3.0	2.5
9	0	8.6	11.8
10	0	<3.3	3.9
1	1	23.7	22.1
2	1	14.3	14.5
3	1	<5.6	7.6
4	1	9.4	14.0
5	1	19.9	19.3
6	1	5.2	7.9
7	1	12.5	11.7
8	1	13.3	13.4
9	1	<2.5	4.3
2	-1	<9.8	4.9
3	-1	<6.0	9.6
4	-1	<5.6	5.5
5	-1	<6.7	6.7
6	-1	18.4	17.2
7	-1	16.0	16.1
8	-1	<4.5	2.5
9	-1	<4.8	8.7
10	-1	8.6	10.6
2	2	15.6	15.3
3	2	9.7	11.8
4	2	30.8	28.0
5	2	7.2	8.4
6	2	<6.4	5.3
7	2	21.3	19.8
8	2	2.5	3.7
9	2	<2.1	5.5
4	-2	34.5	30.3
5	-2	<5.9	6.5
6	-2	8.8	11.3
7	-2	16.0	18.3
8	-2	12.9	14.0
9	-2	<3.8	5.7
10	-2	12.3	13.0
11	-2	<2.2	8.4
3	3	31.2	31.7
4	3	<6.1	3.9
5	3	<4.2	5.0
6	3	17.5	19.1
7	3	5.3	6.3
8	3	<3.6	2.7
6	-3	38.4	40.1
7	-3	6.0	10.2
8	-3	<4.2	4.6
9	-3	23.8	23.4
10	-3	5.9	7.7
11	-3	<3.6	6.5

ℓ = 7

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
4	4	17.8	16.4
5	4	12.5	12.4
6	4	14.6	14.4
7	4	13.7	14.3
8	-4	<6.4	3.1
9	-4	<5.6	6.2
10	-4	16.2	18.1
11	-4	<5.4	5.3
5	5	<7.3	5.4
6	5	12.4	11.5
10	-5	<7.3	8.0
11	-5	18.3	16.8

ℓ = 8

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
0	0	<10.2	6.0
1	0	24.9	22.4
2	0	15.6	15.0
3	0	<7.5	1.6
4	0	6.2	5.0
5	0	10.3	12.8
6	0	<4.1	3.4
7	0	11.6	11.2
8	0	29.0	26.1
9	0	<2.3	7.5
1	1	32.2	27.7
2	1	<5.8	5.7
3	1	9.9	10.7
4	1	31.0	28.4
5	1	14.7	14.7
6	1	4.1	6.8
7	1	17.6	15.1
8	1	18.2	15.7
2	-1	10.4	14.8
3	-1	9.7	12.2
4	-1	9.9	10.0
5	-1	28.2	24.2
6	-1	6.4	6.9
7	-1	13.6	13.0
8	-1	6.7	10.4
9	-1	18.5	17.9
2	2	10.1	13.8
3	2	<6.4	9.6
4	2	<5.0	2.7
5	2	7.3	11.3
6	2	16.8	15.7
7	2	7.0	5.9
4	-2	18.5	21.1
5	-2	<5.5	4.0
6	-2	6.1	3.1
7	-2	17.7	18.1
8	-2	20.4	18.2
9	-2	8.2	8.7

ℓ = 8

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
3	3	<7.0	3.4
4	3	<5.4	4.7
5	3	18.3	16.9
6	3	4.3	5.4
7	3	2.1	3.4
6	-3	<7.0	3.0
7	-3	<5.4	5.3
8	-3	9.4	10.7
9	-3	4.1	4.9
10	-3	<2.8	5.2
4	4	19.1	19.8
5	4	<4.6	3.7
6	4	7.2	9.6
8	-4	20.8	19.4
9	-4	8.8	8.3
10	-4	<3.9	3.4
5	5	21.0	19.1
10	-5	22.3	20.8

ℓ = 9

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
0	0	10.1	15.8
1	0	9.0	10.3
2	0	<5.2	3.5
3	0	18.5	17.8
4	0	<3.9	1.9
5	0	<4.8	2.7
6	0	16.9	16.2
7	0	9.0	5.7
1	1	<9.2	3.1
2	1	<5.5	2.9
3	1	<4.9	6.3
4	1	<5.1	4.6
5	1	<3.9	3.9
6	1	14.5	13.0
7	1	3.8	4.5
2	-1	<9.2	9.0
3	-1	<5.5	6.3
4	-1	<4.9	5.0
5	-1	7.6	10.5
6	-1	11.1	11.2
7	-1	6.8	6.6
8	-1	6.8	6.8
2	2	7.4	11.0
3	2	<4.7	1.7
4	2	<4.2	3.5
5	2	<7.8	9.3
6	2	7.0	8.0
4	-2	<6.9	10.3
5	-2	<4.7	1.3
6	-2	<4.2	6.5
7	-2	9.6	8.8
8	-2	<4.0	3.0

ℓ = 9

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
3	3	23.2	20.9
4	3	<4.4	4.1
5	3	4.7	3.2
6	3	15.2	14.3
6	-3	14.9	16.6
7	-3	<4.4	1.7
8	-3	<2.7	3.2
9	-3	15.2	15.8
4	4	4.0	5.4
5	4	<3.0	4.6
8	-4	8.4	9.5
9	-4	<3.0	3.9

ℓ = 10

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
0	0	<8.0	2.9
1	0	11.2	11.1
2	0	<4.3	5.8
3	0	<5.5	4.6
4	0	2.7	5.4
5	0	<3.5	3.6
6	0	<2.0	3.5
1	1	9.2	11.9
2	1	<4.4	4.9
3	1	<3.9	1.8
4	1	12.5	11.9
5	1	<2.7	0.7
2	-1	12.2	9.5
3	-1	<4.5	5.6
4	-1	10.8	10.3
5	-1	9.9	10.2
6	-1	12.8	10.1
2	2	17.2	15.2
3	2	<3.7	2.9
4	2	<3.1	2.7
4	-2	19.8	15.0
5	-2	4.7	4.9
6	-2	8.4	6.6
3	3	6.8	8.5
6	-3	8.9	6.0

ℓ = 11

h	k	$\lvert F_o \rvert$	$\lvert F_c \rvert$
0	0	14.8	13.2
1	0	<3.6	1.0
2	0	<3.0	4.1
1	1	6.0	5.7
2	1	<2.9	1.2
2	-1	10.6	9.9
3	-1	<2.9	1.9

VI. FURTHER COMMENTS ON THE INDETERMINACY OF THE EXACT STRUCTURE OF GUANIDINIUM ALUMINUM SULFATE HEXAHYDRATE

6.1 *Importance of the Weighting Procedure*

The use of two very different weighting procedures did not break down the high correlations existing between parameters. It is doubtful, especially in the case of so large a number of parameters, that any *reasonable* weighting procedure would succeed in uncoupling the parameters sufficiently to lead to greater determinacy.

6.2 *Effect of Keeping Some of the Parameters Constant while Allowing Others to Vary*

In the case that there is correlation between parameters, it would seem that, at least in the final stages of the refinement, holding of such parameters constant could lead to erroneous results. In a case involving a smaller number of parameters it might be possible to obtain a confidence region[6] for all the parameters by holding some of the parameters constant, but at several different values. For example, suppose the problem involves n almost independent parameters and two almost completely dependent parameters which appear to prevent convergence. Choosing several judicious values of one of the latter and making the calculation for each one will give sets of values for the other parameters which will allow the construction of the equiprobability ellipsoids.

However, in a problem involving many parameters, and many large and multiple correlations, such a technique would appear to be impractical. It should be mentioned that if the model were very nearly linear, only those correlations very near ±1 would be important in the unattainability of convergence. However, it is possible that the more nonlinear the model, the more important the other correlations become.

6.3 *Possible Effects of Increasing the Number of Observed Data*

There are two ways in which the number of data might be increased. One is to obtain more of the weak intensities by increasing the detector sensitivity. It does not seem that this would have the effect of decreasing the correlations. This was shown to some extent by the calculations based on the two different weighting schemes. In the first case the weighted evaluated derivatives for unobserved reflections were included; in the second, these were given zero weight and therefore excluded. Also, the exclusion of reflections for which $\sin^2\theta/\lambda^2 < 0.0800$ did not have an apparently significant effect on the correlations. (Compare, for example, analogous values in Tables VI and X.)

The other way in which to increase the number of data is to use shorter wavelength radiation. Now, it is not necessary actually to measure these data before determining the effect on the correlations because the correlation coefficients, as calculated, depend only on the model and the evaluated derivatives. It is unlikely that the situation would change very much if the additional terms were included because the relationship of the derivatives with respect to correlated parameters would probably not change very much.

In the case of tetragonal $BaTiO_3$,[5,6] higher index reflections would have almost no important contributions from the oxygen atoms. Thus the interactions among oxygen ion parameters will not be affected. Similarly, interactions among the metal ion parameters will probably not be much affected. But interactions between the two groups could be reduced. However, in the case of an all light atom structure, it would appear that the extra data would probably not reduce the correlations.

6.4 *Possible Effect of Greater Accuracy in Measurement of Observed Intensities*

The effect of greater accuracy in measurement of the observed intensities is not really predictable in this case. To be sure, in each iteration the reduction of $s = \sqrt{\Sigma(\sqrt{w}\Delta)^2}/\sqrt{m - n}$ would reduce the apparent *size* of the equiprobability surfaces. This we certainly know.

However, we must ask first whether there is a limit to the accuracy of the observed amplitude. One would suspect that there is such a limit. Furthermore as pointed out by Caticha-Ellis and Rimsky,[20] there will always be a discrepancy between the calculated and true values of the amplitudes. Thus s has a lower positive limit.

Reduction of s would not only decrease the size of the equiprobability surfaces (and therefore, of course, the standard estimates of error) but it would also decrease the components of the vector v, $v_j = \Sigma(\sqrt{w}D_j) \cdot (\sqrt{w}\Delta)$, where the D_i are the evaluated derivatives. Thus, for example, if cycle 5′d were repeated with *each* Δ decreased to $\frac{1}{2}$ of its value, each v_j, and therefore each $\Delta p_i = \sum_j b_{ij}v_j$ would be reduced to the same extent. Of course an *average* reduction of $\frac{1}{2}$ might not do the same thing. In fact, with a poor distribution of the reduction in Δ, the Δp_i in some cases could even be larger, depending on the algebraic values of the D_j.

Actually the nature of the shape of the equiprobability surfaces might give the best clue to what might happen if increased accuracy of measurement were attainable. The nonlinearity of the model would probably play an important part. The more nonlinear, the more important are apt to be those correlations which are not perfect. Of course, even one perfect correlation ±1, renders the whole problem indeterminate[6] if insistence is made on allowing all parameters to vary in an iteration. This is not necessary, however, and one could learn a great deal about the parameters of a structure which has only one perfect correlation and the rest very small ones (see Section 6.2). In the present case, there are many correlations having absolute values between 0.90 and 1.00 (Table X). These have the specific values: 0.917, 0.905, 0.913, 0.907, 0.975, 0.963, 0.901, 0.979, and 0.902, respectively. Perhaps the most important ones are the three closest to unity.

In the case of gross nonlinearity it seems possible that these and so many of the other high correlations of Table X could cause unattainability of convergence even if the lowest limit of s were attained. That is, the shape of the equiprobability surface may be such as to prevent the practical attainment of separate estimates of the parameters (see also Ref. 21) from the given data. This seems to be true of the $BaTiO_3$ case.[5,6]

Needless to say, a measure of doubt remains. Further work might aid in removing this doubt. This would involve trying to obtain more data and of greater accuracy, and further calculations. Our doing this is not presently contemplated.

6.5 *Fourier Synthesis vs Least Squares*

In the case of tetragonal barium titanate, Fourier synthesis produced no improvement on the least squares method.[22] It is likely that with the present data, the situation in the case of the G.A.S.H. would be the same. On the other hand, there is no requirement of linearity in the Fourier synthesis: the actual amplitudes are the Fourier coefficients. In the least squares technique, an approximation is used: i.e.,

$$\begin{aligned} F_{hkl}(p_1, p_2, \cdots, p_n) &= F_{hkl}(\bar{p}_1 + \Delta p_1, \bar{p}_2 + \Delta p_2, \cdots, \bar{p}_n + \Delta p_n) \\ &= F_{hkl}(\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_n) + \sum_{j=1}^{n} \left. \frac{\partial F_{hkl}}{\partial p_j} \right|_{\bar{p}_j} \Delta p_j \end{aligned}$$

where $\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_n$ are approximate but nearly true values of the parameters. It is possible that higher order terms could be important here, but it is *not* clear that inclusion of the next higher order terms would necessarily lead to improvement. Also, the calculation would increase in complexity.

Cochran has shown that a rather close relationship exists between the Fourier synthesis and least squares techniques. There are conditions on this relationship given by Cochran[24] and Hoard and Geller[24], and in addition in the actual least squares calculation, an approximation is made and nearness to linearity is assumed. Therefore, if the nonlinearity is not serious, convergence should be attainable in either case. If it is serious, the relationship could break down further and the Fourier synthesis could conceivably converge when the least squares calculation tends not to converge.

VII. COMMENTS ON THE SINGLE-CRYSTAL AUTOMATIC DIFFRACTOMETER

As mentioned earlier, the data used in this work were collected four years ago. Since that time only one or two attempts were made to use the instrument for other studies. These were unsuccessful because of difficulties which are probably surmountable, but require modification of the instrument.

The present instrument puts a lower limit on the sample size. To keep the time for recording a layer within reasonable bounds and to prevent the instrument from reacting to background scattering, only intensities above a certain preset count energize the circuitry which sets the crystal back and shifts speed. To obtain satisfactory counting rates the use of large crystals is required. (The intensity is proportional to the number of unit cells irradiated.) However, to obtain adequate or meaningful intensities from highly absorbing materials one must have small crystals. In short, the instrument presently is suited mainly to crystals with low absorption and from which sizable cylindrical specimens can be made.

The indexing of the reflections was a tedious process. The possibility of error, particularly at the high angles, was great, but the use of photographs and cross examination of data helped prevent errors. An improvement on the Bond-Benedict automatic single-crystal diffractometer would be provision for foolproof pre-indexing of the reflections.

VIII. SUMMARY

Extensive application of the least squares refinement technique (through the use of the Busing-Levy IBM 704 program) to three-dimensional X-ray data from crystals of guanidinium aluminum sulfate hexahydrate indicated that although the structure as originally reported for the isostructural guanidinium gallium sulfate is essentially correct, an *exact* structure is unattainable from the present data by means of the least squares method of refinement. The numerous high correlations of pairs of parameters, apparently linked with the nature of the structure, appear to be a primary cause of prevention of convergence.

The course of the calculations has been outlined with special emphasis on some of the more obvious parameter interactions, but tables are given to enable the more interested reader to examine the results in somewhat greater detail.

The work also further demonstrates the importance of the correlation matrix as a tool for establishing the existence or nonexistence of interdependence of structural parameters.

IX. ACKNOWLEDGMENTS

The authors wish to thank W. L. Bond and T. S. Benedict for making necessary changes on the single-crystal diffractometer; C. L. Mallows and W. C. Hamilton for helpful discussion; R. G. Treuting, W. R. Romanow and Misses D. C. Leagus and B. B. Cetlin for writing needed IBM 704 programs; Mrs. A. Werner for translating Refs. 3, 8 and 19 from the Russian; F. Barbieri, for preparing[25] the cylindrical crystals used to obtain the data; and the many people in the Bell Telephone Laboratories computation center for their cooperation.

REFERENCES

1. Geller, S., and Booth, D. P., Z. Kristallogr., **111,** 1959, p. 117.
2. Geller, S., Z. Kristallogr., **114,** 1960, p. 148.
3. Varfolomeeva, L. A., Zhdanov, G. S., and Umanskii, M. M., Kristallografiia, **3,** 1958, p. 368.
4. Bond, W. L., Acta Cryst., **8,** 1955, p. 741; Benedict, T. S., Acta Cryst., **8,** 1955, p. 747.
5. Evans, H. T., Jr., The Crystal Structure Analysis of Barium Titanate by X-ray Diffraction Methods, Research Laboratory Technical Report No. 54, Philips Laboratories, Inc., Jan., 1952. Also Acta Cryst., **14,** 1961, p. 1019.
6. Geller, S., Acta Cryst., **14,** 1961, p. 1026.
7. Wood, E. A., Acta Cryst., **9,** 1956, p. 618.
8. Ezhkova, Z. I., Zhdanov, G. S., and Umanskii, M. M., Kristallografiia, **3,** 1958, p. 231.
9. Tunell, G., Amer. Min., **24,** 1939, p. 448.
10. Bond, W. L., Acta Cryst., **12,** 1959, p. 375.
11. Dixon, W. J., and Massey, F. J., Jr., *Introduction to Statistical Analysis*, McGraw-Hill, New York, 1951, Chapter 16.
12. Ibers, J. A., Acta Cryst., **9,** 1956, p. 652.
13. Busing, W. R., and Levy, H. A., A Crystallographic Least Squares Refinement Program for the IBM 704, Oak Ridge National Laboratory Central Files Memorandum 59-4-37, April, 1959.
14. Berghuis, J., Haanappel, IJ. M., Potters, M., Loopstra, B. O., Mac Gillavry, C. H., and Veenendaal, A. L., Acta Cryst., **8,** 1955, p. 478.
15. Viervoll, H., and Ögrim, O., Acta Cryst., **2,** 1949, p. 277.
16. Drenth, J., Drenth, A., Vos, A., and Wiebenga, E. W., Acta Cryst., **6,** 1953, p. 424. Also Bryden, J. H., Burkardt, L. A., Hughes, E. W., and Donohue, J., Acta Cryst., **9,** 1956, p. 573.
17. McCall, D. W., J. Chem. Phys., **26,** 1957, p. 706.
18. Spence, R. D., and Muller, J., J. Chem. Phys., **26,** 1957, p. 706.
19. Aleksandrov, K. S., Lundin, A. G., and Mikhailov, G. M., Kristallografiia, **5,** 1960, p. 84.

20. Caticha-Ellis, S., and Rimsky, A., Acta Cryst., **11,** 1958, p. 481.
21. Box, G. E. P., Fitting Empirical Data, MRC Technical Summary Report No. 151, Mathematics Research Center, United States Army, University of Wisconsin, Madison, Wisc., May, 1960.
22. Evans, H. T., Jr., private communication.
23. Cochran, W., Acta Cryst., **1,** 1948, p. 138.
24. Hoard, J. L., and Geller, S., *Annual Reviews of Physical Chemistry*, Vol. 1, Annual Reviews, Inc., Stanford, Calif., 1950, p. 215.
25. Barbieri, F., and Durand, J. L., Rev. Sci. Instr., **27,** 1956, p. 871.

Discrimination Against Unwanted Orders in the Fabry-Perot Resonator

By D. A. KLEINMAN and P. P. KISLIUK

(Manuscript received September 20, 1961)

It is proposed here that the usual Fabry-Perot interferometer structure of the optical maser may be modified in a very simple way to provide discrimination against unwanted orders. The modification is an extra reflecting surface suitably positioned outside the maser which can greatly affect the losses of the various orders. A simple one-dimensional analysis is given for the effect, and numerical results are presented for a realistic case, showing that the effect can be large. It is concluded that this technique may be useful in preventing unwanted oscillations in the optical maser.

I. INTRODUCTION

The Fabry-Perot interferometer has recently become important as a resonant cavity for electromagnetic radiation at optical frequencies.[1,2,3,4] The nature of the modes of such a cavity has been discussed by Schawlow and Townes[1] and by Fox and Li.[5] The modes may be specified by three quantum numbers, one of which is the familiar *order number* giving the separation of the plates in units of the half-wavelength. The other two quantum numbers specify the possible field configurations across the plates, which are essentially identical in each order. Fox and Li have investigated these configurations and the corresponding frequencies and losses for interferometers consisting of perfectly reflecting plates in air. In the usual laboratory interferometer the Fox and Li modes cannot be resolved because of insufficient reflectivity of the plates. Therefore the role played by these modes in optical masers is not settled. On the other hand, fine structure which could be due to various *Fabry-Perot orders* has been seen in the output of both the gas[6] and the ruby[7] optical maser. It has been pointed out[1,8] that the optical maser is inherently a multimode device, and that the excitation of many modes can lead to undesirable effects in the noise, stability, and ultimate usefulness of the device. Therefore it is proposed here that it would be useful to discriminate against many of the Fabry-Perot orders which can occur in the output by increasing their losses relative to other "preferred" modes.

The Fabry-Perot orders present a problem only when the fluorescence emission of the maser covers a frequency band wider than $(2\mu d)^{-1}$ wave numbers, where μ is the refractive index and d the separation of the plates. This is the case in the gas maser of Javan, Bennett, and Herriott[3] where $(2\mu d)^{-1} \sim 0.005\ \text{cm}^{-1}$ and the doppler broadened Ne transition would be expected to have a width ~0.05. Also, in the ruby optical maser of Collins et al[2], $(2\mu d)^{-1} \sim 0.1$ while the fluoresence line width at room temperature ~10. The orders cannot be eliminated in these cases by shortening the maser and hence spreading the orders, because the gain would then be insufficient to produce oscillations.[9] In ruby, however, the gain could be increased[9] by more than an order of magnitude by cooling, so that the crystal could be shortened. At the same time, the cooling could decrease the line width by more than an order of magnitude,[10] so that elimination of orders appears possible in ruby optical masers by cooling. These examples show the interrelation of gain, line width, and the occurrence of Fabry-Perot orders in the optical maser output.

The idea of using a Fabry-Perot interferometer to discriminate against unwanted orders in the optical maser has occurred to a number of people.[11] Indeed, if the external beam contains several orders, a Fabry-Perot etalon could be constructed which would transmit only one of them. This, of course, would not necessarily have any effect on the losses of the various modes in the maser. If the etalon were put in the internal beam, elementary considerations do not tell us what to expect for the relative losses of the modes. The structure to be proposed in the next section is equivalent to making the etalon one of the reflecting ends of the maser. It is believed that a detailed discussion of how discrimination comes about in such structures is given here for the first time.

II. A MODIFIED INTERFEROMETER

It is proposed that another reflecting plate parallel to the maser plates be provided outside the maser with a means for adjusting the separation of the new plate from the maser. This would produce a modified interferometer having three essential optical surfaces with the active medium in the space between two of these surfaces. It is expected that the separation of the third surface from the maser will be much less than the length of the maser. The purpose of the extra surface is to provide discrimination between the Fabry-Perot orders of the original maser by making some orders very lossy compared to other orders. The losses may be due to scattering by inhomogeneities in the medium and irregularities on the reflecting surfaces, absorption by processes other than the fluorescence process of the active medium, and transmission through the outer reflecting surfaces. For convenience of discussion, all losses may be ascribed to the last mechanism by assigning suitable effective reflectivities to the outer surfaces. In any case, it is clear that the proposal has meaning only when losses are taken into account, since the only other effect of the extra surface would be to shift the frequencies of the already existing orders by amounts less than $(2\mu d)^{-1}$ and to introduce new frequencies corresponding to the increased over-all length of the modified interferometer. Therefore the performance of the device cannot be deduced in an elementary way by considering the two regions between the surfaces as two interferometers with the shorter preferentially selecting and rejecting certain orders of the longer. The truth is that the modified interferometer has more, not fewer, orders than the original maser, but unlike the latter the orders may have very different losses.

III. ANALYSIS

For analysis it is convenient to consider the one dimensional problem shown schematically in Fig. 1. A medium of real dielectric constant $\epsilon > 1$ and real conductivity σ occupies the region $-a \leqq z \leqq a$. For

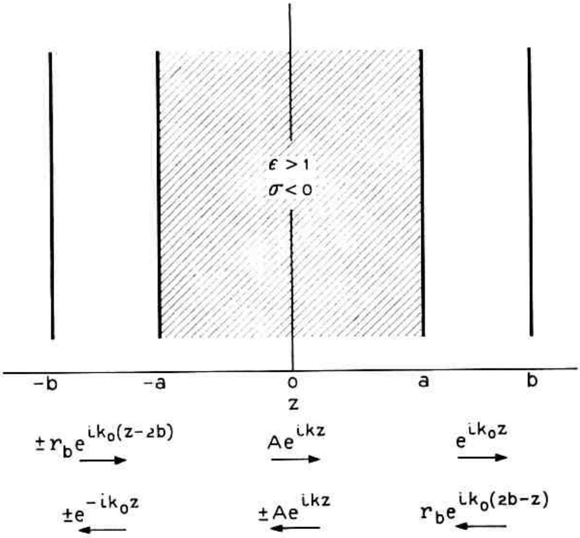


Fig. 1 — Schematic diagram of one-dimensional symmetric structure chosen for analysis of modified interferometer. The value of the constant A is not needed in the analysis.

$|z| > a$ it is assumed that $\epsilon = 1$ and $\sigma = 0$. At $z = \pm b$ are placed reflecting surfaces having the reflectivity for amplitude

$$r_b = e^{-2f}. \tag{1}$$

Since the phase angle of reflection is unimportant here, it has been assumed zero. For later use the quantity

$$T = \tanh f \tag{2}$$

will now be defined. The reflectivity of the surfaces at $z = \pm a$ is

$$r_a = (\sqrt{\epsilon} - 1)/(\sqrt{\epsilon} + 1). \tag{3}$$

From (1), (2) and (3) one can write

$$\begin{aligned} T &= (1 - r_b)/(1 + r_b) \\ 1/\sqrt{\epsilon} &= (1 - r_a)/(1 + r_a). \end{aligned} \tag{4}$$

It is therefore possible in this example to consider arbitrary reflectivities at $z = \pm a, \pm b$ by suitable choices for T and $1\sqrt{\epsilon}$ in the range 0 to 1.

The symmetry of Fig. 1 about a plane at $z = 0$ causes the field to be either even or odd with respect to reflection in this plane. The even solutions are shown by (+) and the odd solutions by (−) signs in Fig. 1. The propagation constants are given by

$$\begin{aligned} k_o &= \omega/c \\ k &= k_o\sqrt{\epsilon}[1 + i(4\pi\sigma/\epsilon\omega)]^{\frac{1}{2}} \\ &= k_o\sqrt{\epsilon} + i(2\pi\sigma/c\sqrt{\epsilon}) + \cdots. \end{aligned} \tag{5}$$

The continuity of the field and its derivative at $z = a$ gives the conditions

$$k \tan(ka) = -k_0 \tan(k_0 b - k_0 a + if) \tag{6}$$

for even modes, and

$$k_0 \tan(ka) = +k \cot(k_0 b - k_0 a + if) \tag{7}$$

for odd modes. These equations give, in general, complex eigenvalues for the angular frequency ω.

It is convenient to require that ω be real and allow σ to assume an appropriate negative value. Both ω and σ are determined by (6) or (7) for even or odd modes respectively. Physically this corresponds to supplying sufficient gain through the negative σ to maintain steady oscillations at frequency ω. The larger the value of $-\sigma$ the greater are the losses of the mode in question. Now let the dimensions be so chosen that

$$n(b - a) = ma\sqrt{\epsilon} \tag{8}$$

where m,n are positive integers. It is then possible to write

$$\begin{aligned} k_0(b-a) &= m\pi + (m/n)\Delta \\ \sqrt{\epsilon}k_0 a &= n\pi + \Delta. \end{aligned} \tag{9}$$

The conductivity to be determined is contained in the parameter

$$\chi = \tanh(2\pi\sigma a/c\sqrt{\epsilon}). \tag{10}$$

In any practical case the ratio k/k_0 occurring in (6) and (7) can be considered real, $k/k_0 \sim \sqrt{\epsilon}$. The equations for the real frequency and conductivity then reduce to

$$\tan\Delta = -\left(\tan\frac{m}{n}\Delta\right)\frac{1+\sqrt{\epsilon}T\chi}{\sqrt{\epsilon}+T\chi} \tag{11}$$

$$\chi = -T\,\frac{1-\sqrt{\epsilon}\tan\Delta\tan(m/n)\Delta}{\sqrt{\epsilon}-\tan\Delta\tan(m/n)\Delta} \tag{12}$$

for even modes, and to

$$\tan\Delta = \left(\cot\frac{m}{n}\Delta\right)\frac{\sqrt{\epsilon}+\chi T}{1+\sqrt{\epsilon}\chi T} \tag{13}$$

$$\chi = -T\,\frac{\sqrt{\epsilon}\tan(m/n)\Delta+\tan\Delta}{\tan(m/n)\Delta+\sqrt{\epsilon}\tan\Delta} \tag{14}$$

for odd modes. When $\tan\Delta$ is eliminated between (11) and (12) or between (13) and (14) the same quadratic equation for χ is obtained, namely

$$\chi^2 + 2p\chi + 1 = 0 \tag{15}$$

where

$$2p = \frac{\epsilon + T^2 + (1+\epsilon T^2)\tan^2(m/n)\Delta}{T\sqrt{\epsilon}(1+\tan^2(m/n)\Delta)}. \tag{16}$$

The solution of (15) which reduces properly as $r_b \to 0$ $(T \to 1)$ is

$$\chi = -p + |(p^2-1)^{\frac{1}{2}}|. \tag{17}$$

When $|\chi| \ll 1$, this reduces to

$$-\chi \sim \frac{T\sqrt{\epsilon}(1+\tan^2(m/n)\Delta)}{\epsilon+\tan^2(m/n)\Delta}. \tag{18}$$

The most practical method of solution is to find the frequencies by neglecting $T\chi$ in (11) and (13), which gives

$$\sqrt{\epsilon}\tan\Delta = -\tan(m/n)\Delta \qquad \text{(even)} \tag{19}$$

$$\tan\Delta = \sqrt{\epsilon}\cot(m/n)\Delta \quad \text{(odd)} \tag{20}$$

respectively. From these solutions, the values of $\tan(m/n)\Delta$ can be substituted into (17) or (18) to obtain χ.

From (15) and (16) it is seen that χ depends upon Δ through $\tan^2(m/n)\Delta$. As a result of the "tuning" condition (8), $\Delta = 0$ is a solution of (19); this is the "preferred" mode having the lowest loss

$$\chi_{\min} = -T/\sqrt{\epsilon}. \tag{21}$$

The largest losses belong to modes having $\tan^2(n/n)\Delta \gg 1$. The solut:on (17) gives two results in the limit $\tan^2(m/n)\Delta \rightarrow \infty$, depending on whether $\epsilon T^2 < 1$ or > 1

$$\begin{aligned} \chi_{\max} &= -T\sqrt{\epsilon} && (\epsilon T^2 < 1) \\ \chi_{\max} &= -1/(T\sqrt{\epsilon}) && (\epsilon T^2 > 1). \end{aligned} \tag{22}$$

Let the quantity

$$R = \chi/\chi_{\min} \tag{23}$$

be called the *discrimination ratio*; then $R_{\max} = \epsilon$ or $1/T^2$, whichever is smaller. Therefore the extra reflecting surface should satisfy

$$r_b \geqq r_a \tag{24}$$

to achieve the maximum discrimination, but there is no advantage in making r_b exceed r_a. It should be noted that the practical approximations (19) and (20) are not valid if $\epsilon T^2 > 1$.

IV. DISCUSSION

The properties of the solutions are best discussed with the aid of an example. For simplicity, a case is chosen in which (19), (20) are valid. Let

$$\begin{aligned} m/n &= 1/5 \\ \sqrt{\epsilon} &= 10 \\ T &= 0.02. \end{aligned} \tag{25}$$

The corresponding reflectivities for amplitude are $r_a = 0.82$, $r_b = 0.96$. According to (21) the loss of the preferred mode $\Delta = 0$ is measured by

$$\chi(0) = \chi_{\min} = -0.002. \tag{26}$$

From (19) it is seen that $\Delta = 5\pi/2$ is a solution with $\tan^2(m/n)\Delta = \infty$ so that according to (22)

$$\chi(5\pi/2) = \chi_{\max} = -0.2. \tag{27}$$

The graphical solution of (19) and (20) is sketched in Fig. 2 with circles representing even solutions and squares odd solutions. The results are summarized in Table I up to $\Delta = 5\pi/2 = 450°$; the remaining roots in the fundamental period of 5π may be obtained from the symmetry about $\Delta = 5\pi/2$. The roots are alternately even and odd as shown in the second column, and $\tan(m/n)\Delta$ in the fourth column rises monotonically from 0 to ∞ corresponding to increasing losses. The discrimination ratio R,

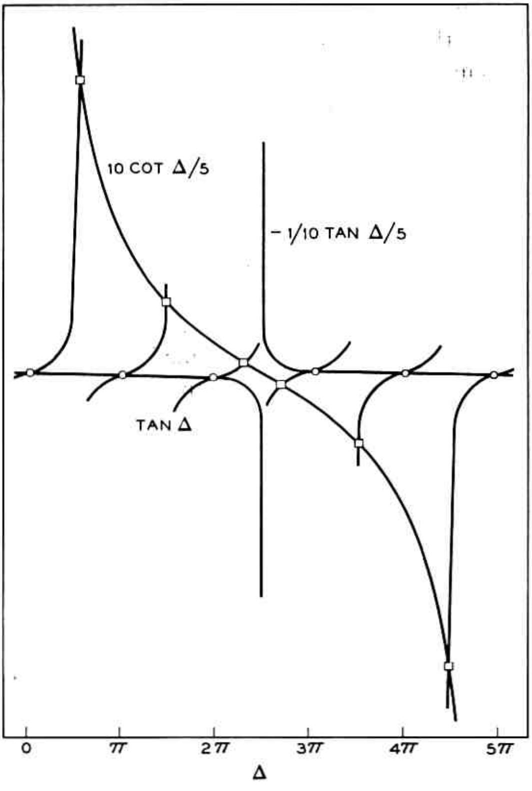


Fig. 2 — Graphical representations of (19) and (20) for $m/n = 1/5$, $\sqrt{\epsilon} = 10$. Odd solutions are indicated by squares and even solutions by circles except at $\Delta = 5\pi/2$, where the intersection is at $\pm\infty$.

TABLE I—SUMMARY OF RESULTS FOR NUMERICAL EXAMPLE WITH $m/n = 1/5$, $\sqrt{\epsilon} = 10$

Δ	type	$\tan \Delta$	$\tan \Delta/5$	R
0	even	0	0	1
88°10′	odd	+31.46	+0.318	1.1
175°58′	even	−0.0705	+0.705	1.49
262°34′	odd	+7.67	+1.304	2.66
345°21′	even	−0.261	+2.61	7.27
412°38′	odd	+1.31	+7.63	37.4
450°	even	∞	∞	100

given in the fifth column, increases from 1 to 100. These results are further summarized in Fig. 3, where the spectrum just calculated is compared with that of the "original" interferometer having no surfaces at $z = \pm b$. The loss in the original interferometer is $\chi = -1/\sqrt{\epsilon} = -0.1$ for all modes. The heights of the spectrum lines in Fig. 3 are proportional to $1/R$ to indicate the relative "Q" of each mode. The total number of frequencies in the fundamental period is twelve compared with ten in the original interferometer for the same period. This is exactly what one would expect, corresponding to the 20 per cent increase in optical length of the modified interferometer. Also as one would expect, the spacing of the preferred modes corresponds to the orders of an ordinary Fabry-Perot interferometer of spacing $d = b - a$.

It will be seen in Fig. 3 that the three modes on either side of a preferred mode have frequencies very close to modes of the original interferometer at $\Delta = \pm\pi/2, \pm\pi, \pm 3\pi/2$. The losses of these modes can be

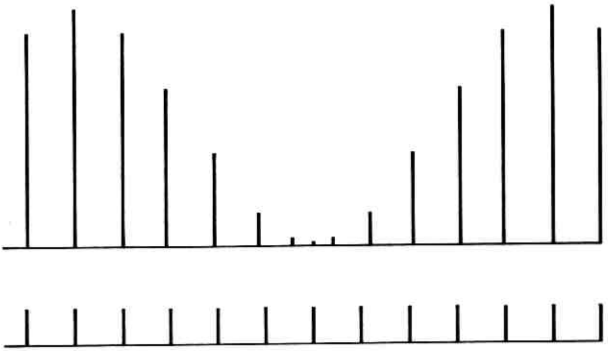


Fig. 3 — The calculated spectrum with the "Q" of each mode indicated by the height of the lines. Shown below for comparison is the spectrum of the "original" interferometer.

calculated to a good approximation from these values of Δ. In general the approximation is

$$\tan(m/n)\Delta \sim \tan[(m/n)N\pi/2] \qquad (28)$$

where $N = 0,1,2, \cdots$ but $N \ll n/m$. Using (28) the evaluation of (17) or (18) can then be carried out immediately without solving for all of the frequencies. This is very convenient since only the modes near the preferred mode are expected to be of interest. It will be noted that the extra modes introduced by the extra surface are among the lossy modes. The periodicity in the above example is a result of choosing n/m an integer. If n/m is chosen not an integer, the periodicity is destroyed, but $\Delta = 0$ remains a preferred mode with minimum loss. Except for extra modes in the regions of high loss, the general effect of the extra surface is to impose a modulation of period $(n/m)\pi$ on the original modes. It is of course not essential for the desired effect that this modulation have a period commensurate with the period of the orders of the original interferometer. Greatest advantage in discrimination against unwanted Fabry-Perot orders is obtained by setting

$$b - a \sim (2\Delta\nu)^{-1} \qquad (29)$$

where $\Delta\nu$ is the half-width at half-maximum of the fluorescence emission.

V. SUMMARY

The theory of the orders of the modified interferometer has been treated in one dimension by considering the symmetrical structure of Fig. 1. The analysis clearly shows the nature and magnitude of the effects to be expected. These effects do not depend in any essential way upon the symmetry assumed for convenience in the analysis, and similar results would be expected for an unsymmetrical modified interferometer with only one extra reflecting surface. It is clear that details in the analysis could be generalized in various ways without changing the substance of the conclusions. The most important of these would be to allow arbitrary reflection and absorption at the interfaces at $\pm a$ to represent the properties of deposited metal layers. On the basis of what has been presented, however, it can be asserted that a third surface of suitable reflectivity and properly positioned can provide considerable discrimination between the orders of a Fabry-Perot interferometer.

REFERENCES

1. Schawlow, A. L., and Townes, C. H., Phys. Rev., **112,** 1958, p. 1940.
2. Collins, R. J., Nelson, D. F., Schawlow, A. L., Bond, W. L., Garrett, C. G. B., and Kaiser, W. K., Phys. Rev. Letters, **5,** 1960, p. 303.

3. Javan, A., Bennett, W. R., Jr., and Herriott, D. R., Phys. Rev. Letters, **6,** 1961, p. 106.
4. Sorokin, P. P., and Stevenson, M. J., Phys. Rev. Letters, **5,** 1960, p. 557.
5. Fox, A. G., and Li, T., B.S.T.J., **40,** March, 1961, p. 453.
6. Bennett, W. R., to be published.
7. Garrett, C. G. B., private communication.
8. Garrett, C. G. B., and Kaiser, W. K., private communication.
9. Kisliuk, P. P., and Boyle, W. S., to be published.
10. Schawlow, A. L., article in *Advances in Quantum Electronics*, to be published.
11. The authors are aware of unpublished notes on this problem by C. G. B. Garrett, W. R. Bennett, Jr., and D. R. Herriott.

The One-Sided Barrier Problem for Gaussian Noise

By DAVID SLEPIAN

(Manuscript received September 21, 1961)

This paper is concerned with the probability, $P[T,r(\tau)]$, that a stationary Gaussian process with mean zero and covariance function $r(\tau)$ be nonnegative throughout a given interval of duration T. Several strict upper and lower bounds for P are given, along with some comparison theorems that relate P's for different covariance functions. Similar results are given for $F[T,r(\tau)]$, the probability distribution for the interval between two successive zeros of the process.

INTRODUCTION

Let $X(t)$ be a real continuous parameter Gaussian process, stationary and continuous in the mean. We shall assume throughout that $EX(t) = 0$ and shall write $r(\tau) = EX(t)X(t + \tau)$. We further assume throughout that we are dealing with a separable, measurable version of the process.

Our main concern in this paper is the probability $P[T,r(\tau)]$ that $X(t)$ be nonnegative for $0 \leqq t \leqq T$. This quantity is of interest as a means of describing the duration of the excursions taken by the process from its mean. From $P[T,r(\tau)]$, the distribution function $F[\lambda,r(\tau)]$ of the interval between successive zeros of the process can be determined by differentiation [see (19)]. This latter quantity is of importance in a variety of engineering applications of noise theory.

Considerable effort has been directed in the past toward the numerical determination of $F[\lambda,r(\tau)]$ both theoretically[1-25] and empirically.[26-32] These researches have resulted in various approximations for $F[\lambda,r(\tau)]$, but many of these are neither upper nor lower bounds for F, and exact circumstances under which they are good approximations are not clear. Generally speaking, they are good for small values of λ and become nugatory for sufficiently large λ. There appears to be nothing rigorous in the literature concerning the asymptotic behavior of F for large λ. (An approximation method is given in Ref. 21.)

In this paper we summarize some known results and present a number of new strict bounds and comparison theorems for $P[T,r(\tau)]$ and $F[\lambda,r(\tau)]$. The most important of these are: Theorem 1, Section 1.3; Theorem 3, Section 1.4; and Theorem 10, Section 1.8. Theorems 12 and 13 (Section 2.7) dealing with class 2 covariances (defined in Section 1.1), though of less importance for our goal, are perhaps of more than passing interest. These and other results presented shed some light on theoretical questions regarding P and F. Their utility in numerically determining these quantities will be discussed elsewhere.

The paper is divided into two parts: Part I presents definitions, results, and discussions; Part II contains the more technical aspects of proofs and other supportive material for Part I.

PART I — DEFINITIONS, RESULTS AND DISCUSSIONS

1.1 *Preliminaries*

From its definition, it is clear that $P[T,r(\tau)]$ is a nonincreasing function of T. It assumes the value $\frac{1}{2}$ for $T = 0$. It obeys the scaling laws

$$P[T,\lambda r(\tau)] = P[T,r(\tau)] \tag{1}$$

$$P[T,r(\lambda\tau)] = P[\lambda T,r(\tau)] \tag{2}$$

$$\lambda > 0.$$

In asserting (2) for all $\lambda > 0$ we have assumed $r(\tau)$ given for all τ. This is a convention that will be adhered to throughout this paper. It is to be noted, however, that $P[T,r(\tau)]$ for $0 \leqq T \leqq T_o$ depends only on the "piece" of the covariance function $r(\tau)$, $0 \leqq \tau \leqq T_o$.

The scaling law (1) suggests normalizing the covariances to be considered so that

$$r(0) = 1. \tag{3}$$

We adopt this convention hereafter.

The scaling law (2) suggests that a normalization of the time scale is in order. There does not appear to be a convenient way to do this for the class of all covariances. For processes continuous in the mean, such as are being considered here, all one can say in general about covariances is that they are even continuous nonnegative-definite functions. This is a rather large class of functions containing a great variety of pathologies such as nowhere differentiable continuous functions. In what follows we shall have occasion to consider covariances $r(\tau)$, strictly monotone in some right neighborhood of $\tau = 0$ and such that $r(\tau) - 1$ behaves like a nonnegative power of $|\tau|$ for sufficiently small $|\tau|$. We normalize and define this class as follows: *The continuous covariance $r(\tau)$ is said to be of class α if, as τ approaches zero,*

$$r(\tau) = 1 - \frac{|\tau|^\alpha}{\Gamma(\alpha + 1)} + o(|\tau|^\alpha),$$

and if $r(\tau)$ is strictly monotone in some right neighborhood $0 < \tau < \tau_o$ of the origin. Here necessarily $0 \leqq \alpha \leqq 2$ and $\Gamma(\alpha + 1)$ is the usual gamma function. The normalization is contained in the specific choice of the coefficient of $|\tau|^\alpha$.

To the author's knowledge, when the scaling laws (1) and (2) are taken into account, there are only three distinct covariances for which $P[T,r(\tau)]$ is known explicitly. These are:

$$(i) \quad r_1(\tau) = e^{-|\tau|}, \qquad 0 \leqq \tau \leqq \infty,$$

$$P[T,r_1(\tau)] = \frac{2}{\pi} \arcsin e^{-T}, \qquad 0 \leqq T < \infty;$$

$$(ii) \quad r_2(\beta,\tau) = 1 - \beta^2 + \beta^2 \cos\left(\frac{\tau}{\beta}\right), \qquad 0 \leqq \tau < \infty, \qquad 0 \leqq \beta \leqq 1,$$

$$P[T,r_2(\beta,\tau)] = \begin{cases} \dfrac{1}{2} - \dfrac{T}{4\pi} - \dfrac{1}{2\pi} \arcsin\left[\beta \sin\left(\dfrac{T}{2\beta}\right)\right], & 0 \leqq \dfrac{T}{\beta} \leqq 2\pi, \\ \tfrac{1}{2}[1 - \beta], \quad 2\pi \leqq \dfrac{T}{\beta} < \infty; \end{cases}$$

$$(iii) \quad r_3(\tau) = \begin{cases} 1 - |\tau|, & |\tau| \leqq 1 \\ 0, & |\tau| \geqq 1, \end{cases}$$

$$P[T,r_3(\tau)] = \frac{1}{4} + \frac{1}{2\pi}[\arcsin(1 - T) - \sqrt{T(2 - T)}], \qquad 0 \leqq T \leqq 1.$$

The process with covariance $r_1(\tau)$ is Markovian, and it is this special property that permits determination of $P[T,r_1(\tau)]$ in this case (see Ref. 22 or Ref. 21, Section IX).

Case (*ii*) corresponds to the stochastic process

$$X(t) = A + B \cos\left[\frac{t}{\beta} + \Phi\right],$$

with A, B and Φ independent random variables, the two former being normal with mean zero and variances $1 - \beta^2$ and β^2 respectively, and the latter being distributed uniformly in $(0,2\pi)$. The determination of P in this case is an exercise in integration and elementary probability theory that will be omitted here. For the obvious generalization of this case, namely,

$$X(t) = A + \sum_1^N B_i \cos[t/\beta_i + \Phi_i],$$

$P[T,r(\tau)]$ can be expressed in principle as a $(2N + 1)$-fold integral. Except in the case $N = 1$ presented, the integrals appear untractable.

The form for $P[T,r_3(\tau)]$ given follows from results found in Ref. 23. Note that it is valid only for $T \leqq 1$. We have been unable to extend P beyond this point.

These examples shed little light on the many questions that naturally arise concerning the behavior of $P[T,r(\tau)]$, both as a function of T and as a functional of $r(\tau)$. What are possible asymptotic behaviors of P for large T? What features of $r(\tau)$ determine this behavior? To what extent is P determined by the behavior of $r(\tau)$ in the neighborhood of $\tau = 0$? (For example, if $r(\tau)$ is analytic in the neighborhood of $\tau = 0$, then it can be extended as a covariance in only one way, namely, by its analytic continuation. In this case, then, $P[T,r(\tau)]$ is completely determined by the behavior of $r(\tau)$ near $\tau = 0$.) If $q(\tau)$ is another covariance, in some sense close to $r(\tau)$ for $0 \leqq \tau \leqq T$, is $P[T,r(\tau)]$ close in some sense to $P[T,q(\tau)]$? How can $P[T,r(\tau)]$ be determined numerically for a given covariance $r(\tau)$?

These and many other basic questions await to be answered in full.

1.2 *$P[T,r(\tau)]$ as a Limit*

Let $0 = t_1 < t_2 < \cdots < t_n = T$ be a partition of the interval $(0,T)$ into $n - 1$ parts. The n random variables $X(t_1)$, $X(t_2)$, $\cdots$, $X(t_n)$ are jointly Gaussian with covariance matrix $\mathbf{r} = (r_{ij})$, where $r_{ij} = r(t_i - t_j)$. Denote by $P_n(\mathbf{r})$ the probability that these n random variables be nonnegative. Because of the assumed separability of the process,

$$P[T,r(\tau)] = \lim_{n\to\infty} P_n(\mathbf{r}), \tag{4}$$

where it is understood that the limit is taken as the partition is refined with mesh tending to zero. If $r(\tau)$ is positive definite, then $|\mathbf{r}| \neq 0$ for any choice of partition, and one can write explicitly

$$P_n(\mathbf{r}) = (2\pi)^{-n/2} \,|\mathbf{r}|^{-\frac{1}{2}} \int_0^\infty dx_1 \cdots \int_0^\infty dx_n e^{-\frac{1}{2}\Sigma r_{ij}^{-1} x_i x_j}. \tag{5}$$

It is somewhat surprising that information about $P[T,r(\tau)]$ is so difficult to obtain when it can be expressed as the limit of the apparently not too unwieldy expression on the right of (5). This integral is deceptive. For $n > 3$ it cannot be expressed in terms of elementary functions of the covariance elements r_{ij}. Series expansions and upper and lower bounds can be easily written for this integral, but most of the obvious ones yield vacuous results in the limit as the partition is refined.

The integral (5) admits of a simple geometric interpretation obtained by reducing the quadratic form in the exponent to a sum of squares by a linear transformation and performing a radial integration. $P_n(\mathbf{r})$ is the fraction of the unit sphere in Euclidean n-space cut out by n-hyperplanes through the center of the sphere. The angle θ_{ij} between the normals to the ith and jth hyperplanes directed into the cutout region is given by $\cos\theta_{ij} = r_{ij}, i,j = 1,2, \cdots, n$. This geometric interpretation of $P_n(\mathbf{r})$ holds even when $|\mathbf{r}| = 0$. For $n = 2$ and 3, this picture gives at once

$$P_2 = \frac{1}{2\pi}[\pi - \theta_{12}] = \frac{1}{4} + \frac{1}{2\pi}\arcsin r_{12} \tag{6}$$

$$\begin{aligned} P_3 &= \frac{1}{4\pi}[2\pi - \theta_{12} - \theta_{13} - \theta_{23}] \\ &= \frac{1}{8} + \frac{1}{4\pi}[\arcsin r_{12} + \arcsin r_{13} + \arcsin r_{23}]. \end{aligned} \tag{7}$$

Seen on the surface of the sphere, the region described above is the generalization of the spherical triangle in three-space and is known as an n-dimensional spherical simplex. Geometers have studied the problem of expressing the content of the spherical simplex in terms of the angles between its bounding surfaces.[33-38] Many of their results can be readily derived from known results in probability theory using the connection with $P_n(\mathbf{r})$ just mentioned (see Section 2.1).

It is clear that $P_n(\mathbf{r})$ is an upper bound for $P[T,r(\tau)]$. The result (7) then is a simple upper bound for $P[T,r(\tau)]$, where $r_{12} = r(t_2 - t_1)$, $r_{13} = r(t_3 - t_1)$, $r_{23} = r(t_3 - t_2)$ and t_1, t_2, t_3 are any three points in the interval $(0,T)$. For very small values of T, this upper bound can be made close to the true value of $P[T,r(\tau)]$. For large values of T, this is generally not the case. If, for example, $r(\tau)$ is never negative, P_3 is always greater than $\frac{1}{8}$. If $r(\tau)$ oscillates in sign, there is a minimum value for P_3 different from zero (unless $r(\tau)$ achieves the value -1) obtainable for any choice of $t_1 \leqq t_2 \leqq t_3$, and hence this bound for $P[T,r(\tau)]$ does not approach zero for large T.

1.3 *A Comparison Theorem for $P[T,r(\tau)]$*

Recall that in the geometric picture of $P_n(\mathbf{r})$, $r_{ij} = \cos\theta_{ij}$ where θ_{ij} is the angle between the inward normals to hyperplanes i and j. Intuitively, it is clear that if this angle is decreased, i.e., if r_{ij} is increased, $P_n(\mathbf{r})$ should also increase. This is borne out by the following

Lemma 1† — *Let $P_n(\mathbf{r})$ be the probability that n jointly Gaussian variates with mean zero and normalized covariance matrix $\mathbf{r}(r_{ii} = 1)$ be nonnegative. Let q be another normalized $n \times n$ covariance matrix. If $r_{ij} \geqq q_{ij}$ for $i,j = 1, 2, \cdots, n$, then $P_n(\mathbf{r}) \geqq P_n(\mathbf{q})$.*

Note that the matrices $\mathbf{r}$ and $\mathbf{q}$ need only be nonnegative definite (as distinguished from positive definite).

By regarding $P[T,r(\tau)]$ as a limit of $P_n(\mathbf{r})$, as explained in the preceding section, Lemma 1 can be used to deduce the following comparison theorem.

Theorem 1 — If $r(\tau) \geqq q(\tau)$ for $0 \leqq \tau \leqq T_o$, then $P[T,r(\tau)] \geqq P[T,q(\tau)]$ for $0 \leqq T \leqq T_o$.

The covariance function of a process is generally regarded as a rough measure of how much the process "hangs together." This view is supported by the theorem. A process with a larger covariance function hangs together more and is more likely to maintain the same sign than one with a smaller covariance.

The comparison theorem can be used with the three covariances (Section 1.1) for which $P[T,r(\tau)]$ is known exactly to bound this quantity for other covariances. The theorem is particularly useful for comparing covariances of the same class. Let $r(\tau)$ and $q(\tau)$ both be of class α, and suppose that $r(\tau) \geqq q(\tau)$ in some neighborhood of the origin. Then $P[T,r(\tau)] \geqq P[T,q(\tau)]$ in this neighborhood. But, for any $\lambda > 1, q(\tau) \geqq r(\lambda\tau)$ in some sufficiently small neighborhood of the origin, so that also $P[T,q(\tau)] \geqq P[T,r(\lambda\tau)] = P[\lambda T,r(\tau)]$ by the scaling law (2). Choosing λ appropriately leads to the following

Theorem 2 — Let $r(\tau)$ and $q(\tau)$ be of class α with $r(\tau) \geqq q(\tau)$ in some neighborhood of $\tau = 0$. Then for some $T^* > 0$,

$$P[T,r(\tau)] \geqq P[T,q(\tau)] \geqq P[r^{-1}(q(T)),r(\tau)], \quad 0 \leqq T \leqq T^*.$$

The theorem is proved in Section 2.3 where the determination of T^* and the choice of proper branch for $r^{-1}(q)$ are also discussed. Knowledge of $P[T,r(\tau)]$ thus provides both upper and lower bounds for $P[T,q(\tau)]$ near $\tau = 0$.

† Proved in Section 2.2. A special case of this lemma was proved by J. Chover[4] by a completely different method. He applied his result to obtain a weak version of our Theorem 1. Chover's result inspired much of the present paper.

1.4 *Some Related Results Useful for Large T*

From Lemma 1, it is easy to deduce (see Section 2.4)

Theorem 3 — Let $T_1 \geqq 0, T_2 \geqq 0, T_3 \geqq 0$ be such that $T_1 + T_2 = T_3$. If $r(\tau) \geqq 0$ for $0 \leqq \tau \leqq T_3$, then

$$P[T_3, r(\tau)] \geqq P[T_1, r(\tau)]P[T_2, r(\tau)]. \tag{8}$$

This theorem provides some asymptotic information on $P[T,r(\tau)]$ for covariances that are never negative. It implies for these covariances that $-(1/T) \log P[T,r(\tau)]$ approaches a nonnegative limit as T becomes infinite. In this sense, then, for *nonnegative covariances, $P[T,r(\tau)]$ cannot decrease asymptotically more rapidly than exponentially.* An exponential lower bound for these covariances is found by iterating (8). Thus, if $T = NT_o$, $P[T,r(\tau)] = P[NT_o, r(\tau)] \geqq P[T_o, r(\tau)]^N$. One obtains in this manner the exponential bound

$$P[T,r(\tau)] \geqq P_o P_o^{T/T_o} \qquad T \geqq T_o \tag{9}$$

which holds for nonnegative $r(\tau)$ with $P_o = P[T_o, r(\tau)]$, $T_o > 0$.

For covariances for which $P[T,r(\tau)]$ is not known, (9) still gives useful information by replacing P_o by a lower bound. For example, from the lower bounds presented below Theorem 6 in Section 1.6, it follows that for nonnegative $r(\tau)$ of class 2, $P[T,r(\tau)] \geqq f(T)$ where

$$f(T) = \begin{cases} \dfrac{1}{2}\left[1 - \dfrac{T}{\pi}\right], & 0 \leqq T \leqq \dfrac{\pi}{2} \\[2ex] \dfrac{1}{4}\left[\dfrac{3}{2} - \dfrac{T}{\pi}\right], & \dfrac{\pi}{2} \leqq T \leqq \dfrac{3\pi}{2}. \end{cases} \tag{10}$$

By choosing T_o to maximize $f(T_o)^{1/T_o}$ and using this maximum value for P_o in (9), one obtains the following

Lower Bound — If $r(\tau)$ is of class 2 and nonnegative, then

$$P[T,r(\tau)] \geqq 0.121\, e^{-2.078(T/\pi)}, \qquad T \geqq (1.016)\pi.$$

For a specific nonnegative covariance of class 2, a somewhat smaller exponent can often be obtained by using for f the lower bound of Theorem 6, or a lower bound obtained from the comparison theorem and example (*ii*) of Section 1.1.

For covariances (such as $r_3(\tau)$ of Section 1.1) that are identically zero for $\tau \geqq T_1$ for some $T_1 > 0$, an exponential upper bound can readily be written for $P[T,r(\tau)]$. For example, if $T = (2N - 1)T_1$, then $P[(2N - 1)T_1, r(\tau)]$ is certainly not greater than the probability that the process be nonnegative in the intervals $(0,T_1)$, $(2T_1, 3T_1)$, $\cdots$,

$((2N - 2)T_1, (2N - 1)T_1)$. But the process in any one of these intervals is independent of the process in the other intervals because of the vanishing of $r(\tau)$ for $\tau \geqq T_1$. Thus, $P[T,r(\tau)] \leqq \{P[T_1, r(\tau)]\}^N$. Arguing in this manner, one arrives at the

Upper Bound — If $r(\tau)$ vanishes for $\tau \geqq T_1$, then

$$P[T,r(\tau)] \leqq \frac{1}{\sqrt{P_1}} P_1^{T/2T_1}, \qquad T \geqq T_1,$$

where $P_1 = P[T_1, r(\tau)]$.

1.5 *Bounds from Rice's Series*

Let $0 = t_1 < t_2 < \cdots < t_n = T$ be a partition of the interval $(0,T)$ into $n - 1$ parts. Let A_i denote the event: "$X(t)$ changes sign at least once in the interval $t_i \leqq t < t_{i+1}$," $i = 1,2,\cdots,n - 1$. Then, by the method of inclusion and exclusion,

$$\begin{aligned} 2P[T,r(\tau)] = 1 - \sum_i \Pr\{A_i\} + \sum_{i<j} \Pr\{A_i \cap A_j\} \\ - \sum_{i<j<k} \Pr\{A_i \cap A_j \cap A_k\} \\ + \cdots + (-1)^{n-1}\Pr\{A_1 \cap A_2 \cap \cdots \cap A_{n-1}\}, \end{aligned}$$

is the probability that none of the events A_i occur. If $r''(0)$ exists, the above series approaches as a limit as the partition is refined with mesh tending to zero

$$2P[T,r(\tau)] = 1 - \int_0^T q_1(t_1)\,dt_1 + \frac{1}{2!}\int_0^T dt_1 \int_0^T dt_2 q_2(t_1, t_2) - \cdots ;$$

(compare Rice,[19] Equation 3.4–11) which we write as

$$\begin{aligned} 2P[T,r(\tau)] &= 1 + \sum_1^\infty \frac{(-1)^n B_n}{n!}, \\ B_n &= \int_0^T dt_1 \cdots \int_0^T dt_n q_n(t_1, \cdots, t_n). \end{aligned} \tag{11}$$

Here $q_n(t_1, \cdots, t_n)dt_1 \cdots dt_n$ is the probability that $X(t)$ has one or more zeros in each of the intervals $(t_1, t_1 + dt_1), \cdots, (t_n + dt_n)$. The existence of $r''(0)$ assures us that $X(t)$ has a derivative almost everywhere in $(0,T)$ for almost all sample functions. One then has

$$\begin{aligned} q_n(t_1, \cdots, t_n) = \int_{-\infty}^\infty d\xi_1 \cdots \int_{-\infty}^\infty d\xi_n \mid \xi_1 \cdots \xi_n \mid \\ \cdot [p(\xi_1, \cdots, \xi_n, x_1, \cdots, x_n)]_{x's=0}. \end{aligned} \tag{12}$$

Here $p(\xi_1, \cdots, \xi_n, x_1, \cdots, x_n)$ is the joint density for the random variables $X'(t_1), \cdots, X'(t_n), X(t_1), \cdots, X(t_n)$ with ξ_i associated with $X'(t_i)$ and x_i associated with $X(t_i), i = 1,2,\cdots,n$. $X'(t)$ is the derivative of $X(t)$ with respect to t.

From the derivation of the method of inclusion and exclusion, successive partial sums of (11) alternately overestimate and underestimate $2P(T)$. We therefore have the sequence of bounds

$$\begin{gathered} 0 \leqq 2P[T, r(\tau)] \leqq 1, \\ 1 - \frac{B_1}{1!} \leqq 2P[T, r(\tau)] \leqq 1 - \frac{B_1}{1!} + \frac{B_2}{2!}, \\ 1 - \frac{B_1}{1!} + \frac{B_2}{2!} - \frac{B_3}{3!} \leqq 2P[T, r(\tau)] \leqq 1 - \frac{B_1}{1!} + \frac{B_2}{2!} - \frac{B_3}{3!} + \frac{B_4}{4!}, \end{gathered} \tag{13}$$

etc. Unfortunately, except for $n = 1,2,3$, the integrand $q_n(t_1, \cdots, t_n)$ occurring in the definition of B_n cannot be expressed in terms of elementary functions. For covariances $r(\tau)$ of class 2, one has

$$q_1(t_1) = \frac{1}{\pi},$$

$$q_2(t_1, t_2) = \frac{1}{\pi^2} \frac{\mu^{3/2}[\sqrt{1 - \alpha^2} + \alpha \arcsin \alpha]}{(1 - \alpha^2)^{3/2}},$$

where

$$\mu = (1 - r^2)(1 - r''^2) - r'^2(2 + 2rr'' - r'^2),$$

$$\alpha = [(1 - r^2)r'' + rr'^2]/[1 - r^2 - r'^2],$$

and

$$r = r(t_2 - t_1), \qquad r' = r'(t_2 - t_1), \qquad r'' = r''(t_2 - t_1).$$

The expression for q_3 is too complicated to warrant display here.

Bounds given by partial sums such as (13) cannot be expected to yield useful results for large T. Typically, for large T, B_n behaves like T^n: the upper bounds exceed unity for large T and the lower bounds become negative.

For small T, however, (13) yields useful information. One has

$$B_1 = \frac{T}{\pi}.$$

If $r(\tau) = 1 - \tau^2/2 + c\tau^4/4! + O(\tau^6)$, a very tedious computation shows that for small T,

$$B_2 = \frac{c-1}{24}\frac{T^3}{\pi} + o(T^3),$$

$$B_3 = O(T^6).$$

From this and the inequalities (13) follows

Theorem 4 — If for small τ

$$r(\tau) = 1 - \frac{\tau^2}{2} + \frac{c\tau^4}{4!} + O(\tau^6),$$

then the first three right-hand derivatives of $P[T,r(\tau)]$ *with respect to* T *exist at* $T = 0$ *and are given by*

$$P[0,r(\tau)] = \tfrac{1}{2},$$

$$P'[0+,r(\tau)] = -\frac{1}{2\pi},$$

$$P''[0+,r(\tau)] = 0,$$

$$P'''[0+,r(\tau)] = \frac{3}{2}\frac{c-1}{48\pi}.$$

The assumed form for $r(\tau)$ in Theorem 4 is important. It has been shown by Longuet-Higgins[14] that if $r(\tau) = 1 - \tau^2/2 + b\,|\,\tau\,|^3 + O(\tau^4)$, $b \neq 0$, then for small T, $B_n = O(T^2)$ for $n = 2,3,4,\cdots$. One can only conclude in this case that $P'[0+,r(\tau)] = -1/2\pi$.

The power series $1 + \sum_1^\infty B_n\lambda^n/n!$ can be written formally as

$$\exp \sum_1 c_n\lambda^n/n.$$

Expand the latter in a power series, equate coefficients of like powers of λ and set $\lambda = -1$. There results the formal identity using (11)

$$2P[T,r(\tau)] = e^{-c_1+c_2/2!-c_3/3!+\cdots}, \tag{14}$$

where

$$\begin{aligned} c_1 &= B_1 = \frac{T}{\pi} \\ c_2 &= B_2 - B_1^2 \\ c_3 &= B_3 - 3B_1B_2 + 2B_1^3 \\ c_4 &= B_4 - 4B_1B_3 - 3B_2^2 + 12B_1^2B_2 - 6B_1^4, \end{aligned} \tag{15}$$

etc., with the B's given by (11) and (12). Relations (15) are the usual ones connecting semi-invariants with central moments (see Ref. 39, p. 37 or Ref. 40, p. 186). Kuznetsov, Stratonovich and Tikhonov[12] have suggested the use of (14) keeping a finite number of c's as a better approximation to P than series (11). For large T, (14) will perhaps yield a better approximation than (11), but it is difficult to see under just what circumstances this will be true. A knowledge of the asymptotic behavior of the c's for large T is needed, but this appears to be a difficult point.

A truncated form of (14) will not in general yield the correct asymptotic behavior of $P[T,r(\tau)]$. For example, retaining only c_1, (14) gives $2\,P[T,r(\tau)] \sim e^{-T/\pi}$ for all class 2 covariances. That this is not in general correct can be seen from a family of simple counterexamples. If $q(\tau)$ is of class 2, then so is

$$r^*(\tau) = q(\alpha\tau)\,\frac{\sin\beta\tau}{\beta\tau}, \tag{16}$$

where $\alpha = \sqrt{1-\beta^2/3}$ and $0 < \beta \leqq \sqrt{3}$. If $X(t)$ has covariance $r^*(\tau)$, then since $r^*(n\pi/\beta) = 0, n = \pm1,\pm2,\cdots$, the random variables

$$X(\pi/\beta),\, X(2\pi/\beta),\, X(3\pi/\beta),\cdots$$

are independent. Set $N = [\beta T/\pi]$. Then

$$\begin{aligned} P[T,r^*(\tau)] &\leqq \Pr\{X(j\pi/\beta) \geqq 0, j = 1,\cdots,N\} = (\tfrac{1}{2})^N \\ &\leqq 2(\tfrac{1}{2})^{\beta T/\pi} = 2e^{-(\beta \log 2)T/\pi}. \end{aligned}$$

Thus if

$$\sqrt{3} = 1.732 \geqq \beta > \frac{1}{\log 2} = 1.442, \tag{17}$$

$e^{T/\pi}P[T,r^*(\tau)]$ approaches zero exponentially for large T, and the first term in the exponent of (14) yields an incorrect asymptotic behavior.

It is interesting to note that the form $e^{-T/\pi}$ obtained from (14) by retaining only c_1 would be correct for a process in which the axis crossings were independent. One would then have $q_n(t_1,\cdots,t_n) = \prod q_1(t_j)$, $B_n = (B_1)^n$ and $c_n = 0, n > 1$. For processes with the covariance (16) with β given by (17), $P[T,r^*(\tau)]$ decays even more rapidly. This has nothing to do with the asymptotic behavior of r^*: by proper choice of $q(\tau)$, this can be altered at will. One must suppose this rapid decay of $P[T,r^*(\tau)]$ is due to the fact that typically $r^*(\tau)$ takes negative values so that at certain time separations the process is anticorrelated. Indeed, it is tempting to conjecture that for nonnegative class 2 covariances, $e^{T/\pi}P[T,r(\tau)]$ increases without limit for large T.

1.6 *Some Other Bounds for $P[T,r(\tau)]$*

In this section we list a few miscellaneous bounds on $P[T,r(\tau)]$.

Theorem 5 —

$$P[T,r(\tau)] \leqq \frac{2}{\pi} \int_0^1 (1 - u) \arcsin r(Tu)\, du.$$

The theorem is proved in Section 2.5. If $\tau \arcsin r(\tau)$ is integrable, the bound in Theorem 5 approaches zero like $1/T$.

Lower bounds for $P[T,r(\tau)]$ are difficult to obtain. One is given by (see Section 2.6)

Theorem 6 — If $r(\tau)$ is of class 2,

$$P[T,r(\tau)] \geqq \frac{3}{8} - \frac{T}{4\pi} + \frac{1}{4\pi} \arcsin r(T).$$

This bound goes negative for relatively small values of T (at least before $T = 2\pi$). It gives somewhat more information than the bound

$$P[T,r(\tau)] \geqq \frac{1}{2}\left[1 - \frac{T}{\pi}\right], \tag{18}$$

obtained from Rice's series (Section 1.5) by retaining only B_1. The bound obtained by retaining B_1, B_2 and B_3 is of course generally much better than that of Theorem 6 but is so complicated that it can be used only with difficulty even with a modern computer. For nonnegative covariances of class 2, Theorem 6 gives $P[T,r(\tau)] \geqq 3/8 - T/4\pi$. This, together with (18), gives (10).

Theorem 7 — If in the neighborhood of $\tau = 0$,

$$r(\tau) = 1 - \frac{\tau^2}{2} + \frac{1}{\beta^2}\frac{\tau^4}{4!} + o(\tau^4),$$

then

$$P[T,r(\tau)] \leqq \frac{1}{2} - \frac{T}{4\pi} - \frac{1}{2\pi} \arcsin\left[\beta \sin\left(\frac{T}{2\beta}\right)\right], \qquad 0 \leqq T \leqq T_1,$$

where $T_1 = \min(\beta\pi,\tau_o)$ and τ_o is the smallest positive value of τ for which $r(\tau) = 1 - 2\sqrt{\beta}$. This theorem follows from the comparison Theorem 1, the result (*ii*) of Section 1.1 and the fact (see Theorem 14, p. 494), that for $0 \leqq \tau \leqq T_1$, the covariance of Theorem 7 is dominated by $r_2(\beta,\tau)$.

Theorem 8 — If $r(\tau)$ is nonnegative and of class 2, then

$$P[T,r(\tau)] \geqq \frac{1}{2} - \frac{T}{4\pi} - \frac{1}{2\pi} \arcsin\left[\frac{1}{\sqrt{2}} \sin \frac{T}{\sqrt{2}}\right], \qquad 0 \leqq T \leqq \frac{\pi}{\sqrt{2}}.$$

This theorem follows from the comparison Theorem 1, the result (*ii*) of Section 1 and the fact (see Theorem 13 in Section 2.7) that for $0 \leqq \tau \leqq \pi/\sqrt{2}$, every nonnegative covariance of class 2 is greater than $r_2(1/\sqrt{2},\tau)$.

We conclude this section with a rather weak, but sometimes useful, result proved in Section 2.8.

Theorem 9 — Let $h(\xi)$ be nonnegative for $0 \leqq \xi \leqq \theta$ and let $h(\xi) = 0$ for $\xi < 0$ and $\xi > \theta$. Define

$$G_\theta(x) = \int_{-\infty}^{\infty} h(x + \xi)h(\xi)\, d\xi$$

and set

$$r_\theta(\tau) = \int_{-\infty}^{\infty} r(\tau - x)G_\theta(x)\, dx.$$

Then

$$P[T,r_\theta(\tau)] \geqq P[T + \theta,r(\tau)].$$

1.7 *Relationship Between $P[T,r(\tau)]$ and $F[\lambda,r(\tau)]$*

If $r''(0)$ exists, then almost all sample functions $X(t)$ possess a derivative almost everywhere. If $r''(0)$ does not exist, then almost all sample functions are nowhere differentiable. In this latter case, if a realization $X(t)$ has a zero at $t = 0$, it almost certainly has infinitely many zeros in every right neighborhood of $t = 0$. In discussing $F[\lambda,r(\tau)]$, the distribution of the interval, l, between successive zeros of $X(t)$, *we accordingly restrict our attention to covariances for which $r''(0)$ exists.*

The quantity $P[T,r(\tau)] - P[T + \Delta,r(\tau)]$ is the measure of those sample functions which are nonnegative in $(0,T)$ but are not nonnegative in $(-\Delta,0)$, i.e., the measure of those sample functions that are nonnegative in $(0,T)$ and have at least one axis crossing in $(-\Delta,0)$. Divide this quantity by the probability $\nu\Delta + o(\Delta)$ that $X(t)$ have one or more upward axis crossings in $(-\Delta,0)$ and allow Δ to approach zero. There results

$$Q[T,r(\tau)] = -\frac{1}{\nu}\frac{d}{dT} P[T,r(\tau)] = 1 - F[T,r(\tau)]. \tag{19}$$

Here $Q[T,r(\tau)]$ is the conditional probability that $X(t)$ be nonnegative in $(0,T)$ given an upcrossing of the axis at $t = 0$; $F[\lambda,r(\tau)] = \Pr(l \leqq \lambda)$ is the distribution function for the interval l between zeros. One should note carefully that the condition in the definition of Q is in the "horizontal window sense" (see Ref. 10, Section 2 for a more complete discussion of this term). We shall find $Q[T,r(\tau)]$ more convenient to deal with than $F[T,r(\tau)]$.

From its definition, $Q[T,r(\tau)]$ is nonincreasing as a function of T. It assumes the value 1 for $T = 0$. Like $P[T,r(\tau)]$, it satisfies the scaling laws

$$\begin{gathered} Q[T,\lambda r(\tau)] = Q[T,r(\tau)] \\ Q[T,r(\lambda\tau)] = Q[\lambda T,r(\tau)] \\ \lambda > 0. \end{gathered} \tag{20}$$

For most purposes, then, it suffices to consider only class 2 covariances. In this case (see Ref. 19, Equation (3.3–10)) $\nu = \dfrac{1}{2\pi}$ and (19) becomes

$$Q[T,r(\tau)] = -2\pi \frac{d}{dT} P[T,r(\tau)]. \tag{21}$$

Clearly upper and lower bounds on $Q[T,r(\tau)]$, say

$$\begin{aligned} Q_U[T,r(\tau)] &\geqq Q[T,r(\tau)], \qquad 0 \leqq T \leqq T_o \\ Q_L[T,r(\tau)] &\leqq Q[T,r(\tau)], \qquad 0 \leqq T \leqq T_o\,, \end{aligned}$$

furnish bounds on $P[T,r(\tau)]$ by integration:

$$\frac{1}{2} - \frac{1}{2\pi}\int_0^T Q_U[x,r(\tau)]\,dx \leqq P[T,r(\tau)] \leqq \frac{1}{2} - \frac{1}{2\pi}\int_0^T Q_L[x,r(\tau)]\,dx,$$
$$0 \leqq T \leqq T_0\,.$$

However, since Q is nonincreasing, it is also possible to obtain weak bounds on Q from known bounds on P. For example, since Q is nonincreasing, if $b > a \geqq 0$,

$$(b - a)Q[a,r(\tau)] \geqq \int_a^b Q[\gamma,r(\tau)]\,d\gamma \geqq (b - a)Q[b,r(\tau)],$$

or from (21)

$$Q[a,r(\tau)] \geqq 2\pi\,\frac{P[a,r(\tau)] - P[b,r(\tau)]}{b - a} \geqq Q[b,r(\tau)]. \tag{22}$$

Thus if $P_U(T)$ and $P_L(T)$ are respectively upper and lower bounds for $P[T,r(\tau)]$ valid for all T,

$$\max_{x \geqq T} 2\pi \frac{P_L(T) - P_U(x)}{x - T} \leqq Q[T,r(\tau)] \leqq \min_{0 \leqq x \leqq T} 2\pi \frac{P_U(x) - P_L(T)}{T - x}. \tag{23}$$

Note that the left inequality of (22) for $a = 0$, $b = T$ again gives (18). Also from (21) and the fact that Q is nonincreasing, it follows that $P[T,r(\tau)]$ for class 2 covariances must be convex downward.

To the author's knowledge, when the scaling laws (20) are taken into account, the only covariance for which $Q[T,r(\tau)]$ is known explicitly is $r_2(\beta,\tau)$ of (*ii*), Section 1.1. One has

$$r_2(\beta,\tau) = 1 - \beta^2 + \beta^2 \cos\left(\frac{\tau}{\beta}\right), \qquad 0 < \beta \leqq 1,$$

$$Q[T,r_2(\tau)] = \begin{cases} \dfrac{1}{2}\left[1 + \dfrac{\cos\left(\dfrac{T}{2\beta}\right)}{\sqrt{1 - \beta^2 \sin^2\left(\dfrac{T}{2\beta}\right)}}\right], & 0 \leqq \dfrac{T}{\beta} \leqq 2\pi, \\ 0, & 2\pi \leqq \dfrac{T}{\beta} \leqq \infty. \end{cases}$$

1.8 *A Comparison Theorem for* $Q[T,r(\tau)]$

Imposing the condition that $X(t)$ have an upcrossing at $t = 0$ in the horizontal window sense greatly complicates computation of probabilities associated with the process. For instance, when $X(t)$ is conditioned in this manner, the random variables $X(t_1),X(t_2),\cdots,X(t_n)$ are no longer jointly Gaussian. If $r(\tau)$ is of class 2, their joint density is

$$2\pi \int_0^\infty d\xi\, \xi p(\xi, x_0, x_1, \cdots, x_n)_{x_0=0},$$

where $p(\xi, x_o, x_1, \cdots, x_n)$ is the Gaussian density of the unconditioned variables $X'(0)$, $X(0)$, $X(t_1),\cdots,X(t_n)$.

It is possible, nevertheless, to derive a comparison theorem for $Q[T,r(\tau)]$ and $Q[T,q(\tau)]$ for class 2 covariances somewhat in the spirit of Theorem 1. (See Section 2.9 for proof.) The function $g(t) = q^{-1}[r(t)]$ plays a role here. Writing $\tau = g(t)$, then $q(\tau) = r(t)$. For a given value of t, we choose $g(t)$ as the smallest positive value of τ for which $q(\tau) = r(t)$. At $t = 0$, we have $\tau = 0$. As t increases from 0, so does τ. One of two difficulties can occur as t increases: $r(t)$ may reach a local minimum $r(t_o)$ at $t = t_o$ before $q(\tau)$ has reached its first local minimum, say $q(\tau_1)$; τ may assume the value τ_1 when t assumes the value $t_1 \leqq t_o$. In the former case we define $g(t)$ only for $0 \leqq t < t_o$; in the latter case, we define $g(t)$ only for $0 \leqq t \leqq t_1$. The comparison theorem can now be stated as follows:

Theorem 10 — Let $r(\tau)$ and $q(\tau)$ be of class 2 and let $g(t) = q^{-1}[r(t)]$ be defined as above. If for all nonnegative x and y with $x + y \leqq T_o$,

$$g(x) + g(y) \geqq g(x + y), \tag{24}$$

then for $0 \leqq T \leqq T_o$

$$Q[T,r(\tau)] \leqq Q[g(T),q(\tau)]. \tag{25}$$

It is easy to show that if $r(\tau) \geqq q(\tau)$ in some neighborhood of the origin, then $g(t)$ has the subadditive property (24) in some sufficiently small neighborhood of the origin so that the theorem is not vacuous.

The steps which led from Theorem 1 to Theorems 2 and 3 are no longer valid when $X(t)$ is conditioned to have an upcrossing at $t = 0$. We have found no analogue of these theorems for $Q[T,r(\tau)]$.

By using (21), one can integrate the inequality (25) to obtain a more complicated comparison theorem for $P[T,r(\tau)]$, namely

$$P[T,r(\tau)] \geq \tfrac{1}{2} + \int_0^{g(T)} h'(\xi) \frac{d}{d\xi} P[\xi,q(\tau)]\, d\xi = P[g(T),q(\tau)]/g'(T) - \int_0^{g(T)} P[\xi,q(\tau)]h''(\xi)\, d\xi,$$

valid for $0 \leqq T \leqq T_o$. Here $h(\xi) = g^{-1}(\xi) = r^{-1}[q(\xi)]$.

PART II — PROOFS AND SUPPORTIVE MATERIAL

2.1† *The Geometric Approach to P_n*

We wish to consider the probability $P_n(\mathbf{r})$ that n jointly normal variates, each with mean zero and normalized covariance matrix $\mathbf{r}$, be nonnegative. Throughout this section we assume that $\mathbf{r}$ is nonsingular. Then $P_n(\mathbf{r})$ can be written as in (5). Denote the eigenvalues and nor-

† The material in this section was developed in 1952. Many of the results have been obtained independently by other workers and have been reported in the literature. Cf. Plackett[41] in particular.

malized eigenvectors of $\mathbf{r}$ by λ_i and $\psi^i = (\psi_1^{\ i},\psi_2^{\ i},\cdots,\psi_n^{\ i}), i = 1,2,\cdots,n$. One has

$$\begin{aligned} \sum_k r_{ik}\psi_k^{\ j} &= \lambda_j\psi_i^{\ j}, \\ \sum_k \psi_k^{\ i}\psi_k^{\ j} &= \sum_k \psi_i^{\ k}\psi_j^{\ k} = \delta_{ij}, \\ r_{ij} &= \sum_k \lambda_k\psi_i^{\ k}\psi_j^{\ k}, \\ i,j &= 1,2,\cdots,n. \end{aligned} \tag{26}$$

In (5) make the substitution $x_i = \sum_k \psi_i^{\ k}\sqrt{\lambda_k}y_k$. There results

$$P_n(\mathbf{r}) = (2\pi)^{-n/2}\int_R \cdots \int dy_1 \cdots dy_n\, e^{-\frac{1}{2}\Sigma y_k^2},$$

where the region R is defined by

$$H_i \equiv \sum_k \psi_i^{\ k}\sqrt{\lambda_k}y_k \geqq 0, \qquad i = 1,2,\cdots,n.$$

Denote by A_n the $(n - 1)$-dimensional content of the intersection of this region with the surface of the unit sphere having center at the origin. Then, by changing to a spherical coordinate system,

$$P_n = (2\pi)^{-n/2}A_n\int_0^\infty dr\, r^{n-1}e^{-r^2/2} = \frac{A_n}{S_n},$$

where $S_n = 2\pi^{n/2}/\Gamma(n/2)$ is the area of the unit sphere. Thus, P_n is the fraction of the unit sphere on the positive side of the n hyperplanes $H_i = 0$. The unit normal $\mathbf{a}^i$ to H_i directed into R has components $a_k^{\ i} = \psi_i^{\ k}\sqrt{\lambda_k}$. From the last of (26), we find for the angle θ_{ij} between $\mathbf{a}^i$ and $\mathbf{a}^j$, $\cos\theta_{ij} = \mathbf{a}^i\cdot\mathbf{a}^j = r_{ij}$.

As mentioned in Section 1.2, expressions for the content A_n of the spherical simplex in terms of the angles between its bounding surfaces are not known for $n > 3$. However, for the determination of $P[T,r(\tau)]$ one is concerned with the limit as $n \to \infty$ of P_n where the angles $\theta_{ij}^{(n)}$ are given, for example, by $\cos\theta_{ij}^{(n)} = r[(i - j)T/n]$ with $r(\tau)$ a given positive definite function. Thus, sufficiently tight bounds for P_n might in the limit yield useful results concerning $P[T,r(\tau)]$. The geometric picture suggests a large number of such bounds. Unfortunately, none has been found which yields useful limits. Since, however, approximations for the n-variable normal integral P_n are of interest in their own right, we digress to mention several such bounds which may be useful. (See Ref. 42 for a bibliography on the multivariate normal integral.)

Circular cones with vertices at the origin can be inscribed and circumscribed about the region R. The half-angle of the inscribed cone is found to be given by

$$\sin \theta_i = \frac{1}{\sqrt{\sum_{ij} r_{ij}^{-1}}}, \tag{27}$$

and the half-angle of the circumscribed cone is given by

$$\cos \theta_c = \frac{1}{\sqrt{\sum_{ij} r_{ij} \sqrt{r_{ii}^{-1}} \sqrt{r_{jj}^{-1}}}}. \tag{28}$$

The fraction of the unit sphere cut out by a circular cone of half-angle θ is

$$F_n(\theta) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \int_0^\theta d\varphi \sin^{n-2} \varphi = \tfrac{1}{2} I_{\sin^2\theta} \left(\frac{n-1}{2}, \frac{1}{2}\right) \tag{29}$$

where I is Pearson's incomplete beta function.[43] One has

$$F_n(\theta_i) \leqq P_n \leqq F_n(\theta_c). \tag{30}$$

Bounds for P_n can also be written in terms of inscribed and circumscribed Euclidean simplexes. The planes $H_i = 0$ intersect $n - 1$ at a time in lines which pass through the origin and a vertex of the spherical simplex. Let $\mathbf{b}^i$ denote the unit vector from the origin to the vertex not contained in $H_i = 0$. One finds for the components $b_k^i = \psi_i^k (\lambda_k r_{ii}^{-1})^{-1/2}$ and for the content of the Euclidean simplex determined by the origin and the end points of the $\mathbf{b}^i$,

$$G_n = \frac{1}{n! \sqrt{|\mathbf{r}|} \sqrt{\Pi r_{ii}^{-1}}}. \tag{31}$$

This simplex lies within the region of interest. The hyperplane through the end points of the vectors $\mathbf{b}^i \sec \theta_c$ is tangent to the unit sphere. The Euclidean simplex determined by the origin and the ends of these vectors therefore contains the region of interest. Thus,

$$\frac{G_n}{V_n} \leqq P_n \leqq \frac{\sec^n \theta_c G_n}{V_n}, \tag{32}$$

where $V_n = \pi^{n/2}/\Gamma(n/2 + 1)$ is the content of the unit sphere, θ_c is given by (28) and G_n by (31). Incidentally, for the cosines of the angles between the **b**'s one finds the interesting reciprocal relations

$$s_{ij} \equiv \mathbf{b}^i \cdot \mathbf{b}^j = \frac{{r_{ij}}^{-1}}{\sqrt{{r_{ii}}^{-1}{r_{jj}}^{-1}}}, \qquad r_{ij} = \frac{{s_{ij}}^{-1}}{\sqrt{{s_{ii}}^{-1}{s_{jj}}^{-1}}},$$

which is the natural generalization of the usual relationship between the sides and angles of a spherical triangle in three-space.

One can expect the bounds in (30) to be close to each other when the $\mathbf{b}^i$ are nearly coplanar, e.g., when all the entries of $\mathbf{r}$ are near unity. One can expect the bounds in (32) to be close to each other when the $\mathbf{b}^i$ are nearly codirectional, e.g., when all the entries of $\mathbf{r}^{-1}$ are nearly equal.

An important differential recursion relation first derived by Schläfli[33] for the content of the spherical simplex can be obtained in an analytic manner from the expression (5) for P_n. We write

$$P_n(\mathbf{r}) = \int_0^\infty dx_1 \cdots \int_0^\infty dx_n g_n(x_1, \cdots, x_n;\mathbf{r}) \tag{33}$$

where the n-variate Gaussian density is given in terms of its characteristic function by

$$g_n(x_1, \cdots, x_n;\mathbf{r}) = \int_{-\infty}^\infty d\xi_1 \cdots \int_{-\infty}^\infty d\xi_n\, e^{i\Sigma x_j \xi_{j_e} - \frac{1}{2}\Sigma r_{jk}\xi_j\xi_k}.$$

From this latter expression it follows that

$$\frac{\partial g_n}{\partial r_{jk}} = \frac{\partial^2 g_n}{\partial x_j \partial x_k}, \qquad k > j. \tag{34}$$

Here we regard g_n as a function of the $n(n-1)/2$ variables r_{jk}, $k > j$, and recall that $r_{ii} = 1$, $r_{jk} = r_{kj}$. Regarding P_n as a function of this same set of variables, we find from (33) and (34)

$$\frac{\partial P_n(\mathbf{r})}{\partial r_{12}} = \int_0^\infty dx_1 \cdots \int_0^\infty dx_n \frac{\partial^2}{\partial x_1 \partial x_2} g_n(x_1, \cdots, x_n;\mathbf{r}).$$

Perform the integrations indicated on x_1 and x_2. There results

$$\frac{\partial P_n(\mathbf{r})}{\partial r_{12}} = \int_0^\infty dx_3 \cdots \int_0^\infty dx_n g_n(0, 0, x_3, \cdots, x_n;\mathbf{r}) \geqq 0. \tag{35}$$

Now if g_n is the density for the random variables $X_1, \cdots, X_n$,

$$g_n(x_1, \cdots, x_n;\mathbf{r}) = p(x_1, x_2)p(x_3, \cdots, x_n \mid x_1, x_2),$$

where $p(x_1, ,x_2)$ is the joint density for X_1 and X_2 and

$$p(x_3, \cdots, x_n \mid x_1, x_2)$$

is the conditional density of $X_3, \cdots, X_n$ given that $X_1 = x_1$ and $X_2 = x_2$. In the case of Gaussian variates, these densities are well known Evaluating this expression at $x_1 = x_2 = 0$, one finds

$$g_n(0,0,x_3, \cdots, x_n;\mathbf{r}) = \frac{1}{2\pi\sqrt{1 - r_{12}^2}}\, g_{n-2}(x_3, \cdots, x_n;\mathbf{r}_{\cdot 12}).$$

When combined with (35) and generalized for arbitrary indices, this yields

$$\frac{\partial P_n(\mathbf{r})}{\partial r_{jk}} = \frac{1}{2\pi\sqrt{1 - r_{jk}^2}}\, P_{n-2}(\mathbf{r}_{\cdot jk}) \geqq 0. \tag{36}$$

Here $\mathbf{r}_{\cdot jk}$ is the customary notation of the statistician for partial correlation coefficients (see Ref. 40, Section 23.4 and pp. 318–319), so that, for example with $\mu \neq j,k$, $\nu \neq j,k$

$$r_{\mu\nu\cdot jk} = \frac{\begin{vmatrix} r_{\mu\nu} & r_{\mu j} & r_{\mu k} \\ r_{j\nu} & 1 & r_{jk} \\ r_{k\nu} & r_{kj} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{\mu j} & r_{\mu k} \\ r_{j\mu} & 1 & r_{jk} \\ r_{k\mu} & r_{kj} & 1 \end{vmatrix}^{\frac{1}{2}} \begin{vmatrix} 1 & r_{\nu j} & r_{\nu k} \\ r_{j\nu} & 1 & r_{jk} \\ r_{k\nu} & r_{kj} & 1 \end{vmatrix}^{\frac{1}{2}}}.$$

Equation (36) is Schläfli's celebrated differential recursion formula. His many relations connecting the angles of the boundary simplexes are familiar to the statistician as identities among partial correlation coefficients.

We close this section with a simple demonstration that for odd n, P_n can be expressed in terms of the content of lower dimensional simplexes. Let p_i denote the probability that X_i be nonnegative, p_{ij} denote the probability that X_i and X_j be nonnegative, etc. Then $P_n = p_{12\cdots n}$. Set $M_1 = \Sigma p_i$, $M_2 = \sum_{i<j} p_{ij}$, etc. Then from the well-known inclusion and exclusion formula, the probability Q_n that none of the variates be nonnegative is

$$Q_n = 1 - M_1 + M_2 - \cdots + (-1)^n M_n .$$

But from symmetry, $P_n = Q_n = M_n$ so that

$$[1 - (-1)^n]P_n = 1 - M_1 + M_2 - \cdots + (-1)^{n-1} M_{n-1} .$$

(Cf. Sommerville,[35] Chapter IX, Section 1.9.) No recursion is known for even n.

2.2 *Proof of Lemma 1*

Lemma 1 follows directly from (35). Note that in the derivation of this result, it was not necessary to normalize the covariance matrix. This result thus states that if $\boldsymbol{\varrho}$ is a position definite symmetric matrix, then

$$\frac{\partial P_n(\boldsymbol{\varrho})}{\partial \rho_{ij}} \geqq 0, \qquad j > i, \tag{37}$$

with $P_n(\boldsymbol{\varrho})$ defined by (5).

Now let $\mathbf{r}$ and $\mathbf{q}$ be nonnegative definite $n \times n$ symmetric matrices with $r_{ii} = q_{ii} = 1$. Then $\boldsymbol{\varrho} = \lambda\mathbf{r} + (1 - \lambda)\mathbf{q} + \epsilon\mathbf{I}$, where $\mathbf{I}$ is the $n \times n$ unit matrix, is positive definite for each $\epsilon > 0$ and each λ satisfying $0 \leqq \lambda \leqq 1$. Consider $P_n(\boldsymbol{\varrho})$ as a function of λ. It is readily established that $P_n(\boldsymbol{\varrho})$ possesses a continuous derivative and indeed that

$$\begin{aligned} \frac{dP_n(\boldsymbol{\varrho})}{d\lambda} &= \sum_{j>i} \frac{\partial P_n(\boldsymbol{\varrho})}{\partial \rho_{ij}} \frac{d\rho_{ij}}{d\lambda} \\ &= \sum_{j>i} \frac{\partial P_n(\boldsymbol{\varrho})}{\partial \rho_{ij}} (r_{ij} - q_{ij}). \end{aligned}$$

If now $r_{ij} \geqq q_{ij}, j > i$, (37) then gives

$$\frac{dP_n(\boldsymbol{\varrho})}{d\lambda} \geqq 0.$$

Integration on λ from 0 to 1 yields $P_n(\mathbf{r} + \epsilon\mathbf{I}) \geqq P_n\ (\mathbf{q} + \epsilon\ \mathbf{I})$. From well-known continuity theorems (see Cramer,[40] Section 24.3 and 10.7), Lemma 1 follows by letting ϵ tend to zero.

2.3 *Proof of Theorem 2*

Let $r(\tau)$ and $q(\tau)$ both be of class $\alpha > 0$ and suppose that $r(\tau) \geqq q(\tau)$ for $0 \leqq \tau \leqq T_o$. Then for any $\lambda > 1$,

$$\begin{gathered} r(\tau) \geqq q(\tau) \geqq r(\lambda\tau) \\ 0 < \tau \leqq \tau_1(\lambda), \end{gathered} \tag{38}$$

for some suitable $\tau_1(\lambda)$. By Theorem 1, then, and the scaling law (2)

$$\begin{gathered} P[T,r(\tau)] \geqq P[T,q(\tau)] \geqq P[\lambda T,r(\tau)] \\ 0 \leqq T \leqq \tau_1(\lambda). \end{gathered} \tag{39}$$

To see how best to choose λ to obtain a good lower bound for $P[T,q(\tau)]$, it is convenient to define a version of $h(\tau) = r^{-1}[q(\tau)]$. Let τ_q be the smallest value of $\tau > 0$ for which $q(\tau)$ is not decreasing. (Strictly speaking, τ_q = inf. of those T for which $q(\tau)$ is not strictly monotone for $0 < \tau \leqq T$. If this T set is empty, define $\tau_q = \infty$.) Define τ_r in an analogous manner. The function $r^{-1}(q)$ is defined for $1 \geqq q \geqq r(\tau_r)$ by the branch having values between 0 and τ_r. Similarly we define $q^{-1}(r)$ for $1 \geqq r \geqq q(\tau_q)$ by the branch having values between 0 and τ_q. If $q(\tau_q) \leqq r(\tau_r)$, we define $h(\tau) = r^{-1}[q(\tau)]$ only for $0 \leqq \tau < q^{-1}[r(\tau_r)]$. If $q(\tau_q) \geqq r(\tau_r)$, we define $h(\tau)$ for $0 \leqq \tau < \tau_q$. Clearly $h(0) = 0$. As τ increases from zero, $h(\tau)$ is at first at least as large as τ, since $r(\tau) \geqq q(\tau)$ near $\tau = 0$. For small τ, $r(h) = q(\tau)$, so that $h'(\tau)r'(h) = q'(\tau)$ or

$$h'(0+) = \lim_{t \to 0+} \frac{q'(\tau)}{r'(h)} = \lim_{t \to 0+} \frac{\tau^{\alpha-1}}{h^{\alpha-1}} = \lim_{t \to 0+} \left(\frac{h(\tau)}{\tau}\right)^{1-\alpha} = h'(0+)^{1-\alpha},$$

so that $h'(0+) = 1$. Three typical curves for $y = h(\tau)$ are shown in Fig. 1. Note that $h(\tau)$ is strictly monotone in its domain of definition.

Consider now the plots of $y = h(\tau)$ and $y = \lambda\tau$ as shown on Fig. 1. For all values of λ, these curves have the origin as a point in common. When $\lambda = 1$, the straight line $y = \lambda\tau$ is tangent to $y = h(\tau)$ at the origin. As λ is increased from 1, a second point of intersection moves out from the origin. It may happen, as in Fig. 1(a), that the line $y = \lambda\tau$ becomes tangent to $y = h(\tau)$. If so, we denote by T^* the abscissa of the first such point of tangency as λ increases from unity and we denote the corresponding value of λ by λ^*. If no such tangency occurs, we denote by T^* the largest value of τ in the domain $h(\tau)$. In this case we set $\lambda^* = h(T^*)/T^*$. (Note that λ^* may be infinite.) Observe that for a given $\lambda < \lambda^*$, the abscissa of the first point of intersection of $y = \lambda\tau$ with $y = h(\tau)$ to the right of the origin, say τ_1, satisfies $h(\tau_1) = \lambda\tau_1$ or $q(\tau_1) = r(\lambda\tau_1)$. For $\tau \leqq \tau_1$, the right inequality of (38) maintains; for $\tau = \tau_1 + \epsilon, r(\lambda\tau) > q(\tau)$ for small positive ϵ.

The lower bound $P[\lambda T, r(\tau)]$ on the right of (38) is a nonincreasing function of λ for a fixed T. For a given $T \leqq T^*$, then, this bound is made as large as possible by choosing λ as the smallest value greater than unity for which $q(T) = r(\lambda T)$. With this choice, λT has the value $h(T)$ and Theorem 2 is proved. The largest T^* for which the theorem as stated in Section 1.3 is true is the value T^* defined in the previous paragraph.

Note that if $r(\tau)$ and $q(\tau)$ cross at $\tau_o > 0$, i.e., $r(\tau_o) = q(\tau_o)$, T^* is necessarily less than τ_o, for in this case, $y = h(\tau)$ crosses $y = \tau$ at τ_o as in Fig. 1(a) and a tangency occurs as indicated.

2.4 *Proof of Theorem 3*

Let $T_1 > 0$ and $T_2 > 0$ be given and set $T_3 = T_1 + T_2$. Consider the approximation to $P[T_3, r(\tau)]$ given by the probability $P_n(\mathbf{r})$ that

$$X(t_1),\cdots,X(t_{n_1}),\ X(\tau_1),\cdots,X(\tau_{n_2})$$

all be nonnegative. Here $0 = t_1 < t_2 < \cdots < t_{n_1} = T_1$ is a partition of $(0,T_1)$ and $T_1 < \tau_1 < \tau_2 < \cdots < \tau_{n_2} = T_3$ is a partition of $(T_1, T_1 + T_2)$ and $n_1 + n_2 = n$. The covariance matrix **r** can be written in block form

$$\mathbf{r} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{C} \end{pmatrix},$$

where **A** is an $n_1 \times n_1$ normalized covariance matrix with elements $r(t_i - t_j)$ **C** is an $n_2 \times n_2$ normalized covariance matrix with elements

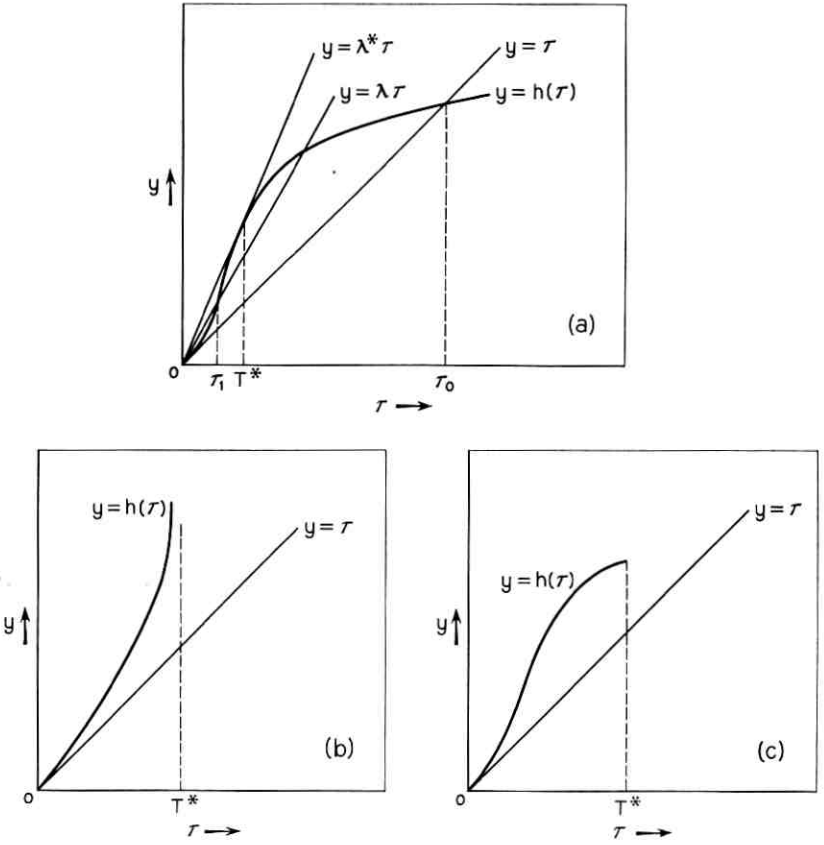


Fig. 1 — The curve $y = h(\tau)$.

$r(\tau_i - \tau_j)$, and $\mathbf{B}$ has n_1 rows and n_2 columns and elements $r(t_i - \tau_j)$. Now

$$\hat{\mathbf{r}} = \begin{pmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{C} \end{pmatrix},$$

is also a covariance matrix, and if $r(\tau) \geqq 0$ for $0 \leqq \tau \leqq T_3$, the elements of $\mathbf{r}$ are not less than the corresponding elements of $\hat{\mathbf{r}}$. From Lemma 1, it follows that $P_n(\mathbf{r}) \geqq P_n(\hat{\mathbf{r}})$. But $\hat{\mathbf{r}}$ is the covariance matrix for two independent sets of random variables so that

$$P_n(\mathbf{r}) \geqq P_n(\hat{\mathbf{r}}) = P_{n_1}(\mathbf{A})P_{n_2}(\mathbf{C}).$$

By refining the partition with mesh tending to zero, one has $P[T_3, r(\tau)] \geqq P[T_1, r(\tau)]P[T_2, r(\tau)]$ and the theorem is established. (It is trivially true if T_1 or T_2 or both are zero.)

2.5 *Proof of Theorem 5*

Theorem 5 is a consequence of the following more general

Theorem 11 — Let the random variables $X_1, X_2, \cdots, X_n$, $n > 2$ have a joint density $p(x_1, \cdots, x_n)$ with the property $p(-x_1, \cdots, -x_n) = p(x_1, \cdots, x_n)$. Then

$$\Pr\{X_i \geqq 0, i = 1,2,\cdots,n\} \leqq -\frac{1}{2} + \frac{4}{n(n-2)} \sum_{i<j} \Pr\{X_i \geqq 0, X_j \geqq 0\}.$$

The proof of this theorem follows that of a theorem by Gaddum[44] concerning spherical simplexes and their angle sums. We introduce the following notations: $P_{ij} = \Pr(X_i \geqq 0, X_j \geqq 0)$, $P = \Pr\{X_i \geqq 0, i = 1,2,\cdots,n\}$, $R(a_1, a_2, \cdots, a_n) = \Pr\{a_1X_1 \geqq 0, a_2X_2 \geqq 0, \cdots, a_nX_n \geqq 0\}$, $a_i = \pm 1$, $i = 1,\cdots,n$. Thus $P = R(1,1,\cdots,1)$ and

$$\sum_{a_1,\cdots,a_n} R(a_1, a_2, \cdots, a_n) = 1,$$

where in the sum each a takes values $+1$ and -1. The 2^n symbols R are equal in pairs;

$$R(a_1, a_2, \cdots, a_n) = R(-a_1, -a_2, \cdots, -a_n).$$

We call $R(-a_1, -a_2, \cdots, -a_n)$ the complement of $R(a_1, a_n, \cdots, a_n)$.

One has

$$\begin{aligned} P_{12} &= P + \Sigma' R(1,1,a_3, a_4, \cdots, a_n) \\ P_{13} &= P + \Sigma' R(1,a_2, 1,a_4, \cdots, a_n) \\ &\vdots \\ P_{n(n-1)} &= P + \Sigma' R(a_1, a_2, \cdots, a_{n-2}, 1,1). \end{aligned} \tag{40}$$

Here the R symbol on the right of the equation having P_{ij} as left member has a 1 in the i^{th} and j^{th} places and a's elsewhere. In each equation, the sum is over all combinations of plus and minus 1 for the a's except for the combination all a's plus 1.

Now consider adding the $n(n-1)/2$ equations (40). One has

$$\sum_{i<j} P_{ij} = [n(n-1)/2]P + S,$$

where S is the sum of all the sums of R symbols on the right of (40). A given R symbol with precisely j of its arguments $+1$ will occur $j(j-1)/2$ times in S, $j = 2,3,\cdots,n-1$. Denote by T_j the sum of all R symbols that have precisely j of their arguments $+1$. Then

$$\sum_{i<j} P_{ij} = \frac{n(n-1)}{2} P + \frac{(n-1)(n-2)}{2} T_{n-1} + \sum_{j=2}^{n-2} \frac{j(j-1)}{2} T_j. \tag{41}$$

Now

$$\sum_{j=2}^{n-2} \frac{j(j-1)}{2} T_j = \sum_{j=2}^{n-2} \frac{(n-j)(n-j-1)}{2} T_{n-j},$$

so that

$$\sum_{j=2}^{n-2} \frac{j(j-1)}{2} T_j = \frac{1}{2} \sum_{j=2}^{n-2} \left[\frac{j(j-1)}{2} T_j + \frac{(n-j)(n-j-1)}{2} T_{n-j}\right].$$

But since an R symbol and its complement are numerically equal, $T_j = T_{n-j}$, so that (41) becomes

$$\sum_{i<j} P_{ij} = \frac{n(n-1)}{2} P + \frac{(n-1)(n-1)}{2} T_{n-1} + \frac{1}{2} \sum_{j=2}^{n-2} \left[\frac{j(j-1)}{2} + \frac{(n-j)(n-j-1)}{2}\right] T_j.$$

Now, for $j = 2,3,\cdots,n-2$,

$$\frac{j(j-1)}{2} + \frac{(n-j)(n-j-1)}{2} \geqq \frac{n(n-2)}{4},$$

so that

$$\sum_{i<j} P_{ij} \geqq \frac{n(n-1)}{2} P + \frac{(n-1)(n-2)}{2} T_{n-1} + \frac{n(n-2)}{8} \sum_{j=2}^{n-2} T_j \geqq \frac{n(n-2)}{2} P + \frac{n(n-2)}{8} \sum_{j=1}^{n-1} T_j.$$

However, the last appearing sum is $1 - 2P$ and Theorem 11 follows directly.

In the case of a Gaussian process $X(t)$ with normalized covariance function $r(\tau)$, we consider the application of Theorem 11 to the random variables $X_i = X(iT/n), i = 1,2,\cdots,n$. Then from (6), $P_{ij} = \frac{1}{4} + 1/2\pi \arcsin r[(i - j)T/n]$. By taking limits as n becomes infinite, Theorem 11 then yields

$$P[T,r(\tau)] \leqq \frac{2}{\pi} \frac{1}{T^2} \int_0^T dy \int_0^y dx \arcsin r(y - x).$$

Elementary manipulations then lead to the result stated as Theorem 5.

2.6 *Proof of Theorem 6*

Consider n random variables, $X_1, X_2, \cdots, X_n$, and the following mutually exclusive events: (A) the variables are all nonnegative; (B_j) the first j variables are nonnegative and the $(j + 1)^{\text{st}}$ is negative, $j = 1,2,3,\cdots,n - 1$. The union C of these events is the event $X_1 \geqq 0$. We suppose $\Pr\{C\} = \frac{1}{2}$ and write $P_n = \Pr\{A\}, V_j = \Pr\{B_j\}, j = 1,2,\cdots,n - 1$ so that

$$P_n = \tfrac{1}{2} - \sum_{j=1}^{n-1} V_j .$$

But $V_j \leqq \Pr\{X_1 \geqq 0, X_j \geqq 0, X_{j+1} < 0\}, j = 2,\cdots,n - 1$ so that

$$P_n \geqq \tfrac{1}{2} - \Pr\{X_1 \geqq 0, X_2 \leqq 0\} - \sum_{j=2}^{n-1} \Pr\{X_1 \geqq 0, X_j \geqq 0, X_{j+1} < 0\}. \quad (42)$$

Consider a stationary Gaussian process $X(t)$ with a class 2 covariance $r(\tau)$. In (42) set $X_j = X(jT/n)$. From (7), one obtains

$$\Pr\{X_1 \geqq 0, X_j \geqq 0, X_{j+1} < 0\}$$
$$= \frac{1}{8} + \frac{1}{4\pi}\left[\arcsin r\left[(j - 1)\frac{T}{n}\right] - \arcsin r\left[j\frac{T}{n}\right] - \arcsin r\left[\frac{T}{n}\right]\right],$$

and from (6)

$$\Pr\{X_1 \geqq 0, X_2 \leqq 0\} = \frac{1}{4} - \frac{1}{2\pi} \arcsin r\left(\frac{T}{n}\right).$$

Insert these values in (42) and pass to the limit as n becomes infinite. Theorem 6 results.

2.7 *On Class 2 Covariances*

Let $r(\tau)$ be a class 2 covariance. From the Bochner representation

$$r(\tau) = \int_0^\infty \cos \lambda\tau \, dF(\lambda),$$

where we now have

$$1 = \int_0^\infty dF(\lambda) = \int_0^\infty \lambda^2 dF(\lambda),$$

it is not hard to show that r is continuous, that $r'(\tau)$ exists everywhere and is continuous, and that $r''(\tau)$ exists and is continuous everywhere except perhaps at $\tau = 0$.

If the process $X(t)$ with mean zero has $r(\tau)$ as its covariance function, then the four random variables $X(0), X'(0), X(t), X'(t)$ have covariance matrix

$$\begin{pmatrix} 1 & 0 & r & r' \\ 0 & 1 & -r' & -r'' \\ r & -r' & 1 & 0 \\ r' & -r'' & 0 & 1 \end{pmatrix}$$

where we write $r = r(t)$, $r' = d/dt \, r(t)$, $r'' = d^2/dt^2 \, r(t)$. For this to be a nonnegative definite matrix it is necessary that the determinant of all major diagonal submatrices be nonnegative. Evaluating these determinants, one finds the system of differential inequalities

$$(1 - r^2 - r'^2)(1 - r'^2 - r''^2) - (rr' + r'r'')^2 \geqq 0, \qquad (43)$$

$$1 - r^2 - r'^2 \geqq 0, \qquad (44)$$

$$1 - r^2 - r''^2 \geqq 0, \qquad 1 - r'^2 - r''^2 \geqq 0,$$

$$1 - r^2 \geqq 0, \qquad 1 - r'^2 \geqq 0, \qquad 1 - r''^2 \geqq 0.$$

These inequalities can also be derived without raising the question of existence of the derivative process by demanding that the covariance matrix of the four random variables $X(0), X(\epsilon) - X(0), X(t), X(t+\epsilon) - X(t)$ be nonnegative definite for arbitrarily small values of ϵ.

Consider now the family of covariances

$$r_2(\beta,\tau) = 1 - \beta^2 + \beta^2 \cos\left(\frac{\tau}{\beta}\right), \qquad 0 \leqq \beta \leqq 1, \qquad (45)$$

introduced in Section 1.1. In what follows, we shall be concerned with the family, F, of curves $r = r_2(\beta,\tau)$, where for each β with $0 < \beta \leqq 1$ we restrict our attention to the interval $0 \leqq \tau \leqq \pi\beta$. Several members of the family are shown in Fig. 2. The following statements, evident from the figure, are easy to prove analytically. (1) The curves of the family do not intersect each other except at $\tau = 0$. (2) A horizontal line $r = r_o$ with $| r_o | < 1$ intersects exactly once each member of F with parameter value in the range $1 \geqq \beta \geqq \sqrt{(1 - r_o)/2}$. For each value of α satisfying $-\sqrt{1 - r_o^2} \leqq \alpha \leqq 0$, there is a unique member of the famly that intersects the line $r = r_o$ with slope α. If $\beta(\alpha)$ denotes the parameter value of this member of F, $\beta(\alpha)$ is a continuous strictly monotone decreasing function of α, $-\sqrt{1 - r_o^2} \leqq \alpha \leqq 0$.

We shall say that the curve $r = r(\tau)$ intersects the curve $r = g(\tau)$ from below if at the point of intersection $r' > g'$.

Lemma 2 — Let $r(\tau)$ be of class 2.

a. *If the first local minimum of $r(\tau)$ is at τ_1, then $r = r(\tau)$ cannot intersect from below any member of the family F,*

$$r = r_2(\beta,\tau) = 1 - \beta^2 + \beta^2 \cos\left(\frac{\tau}{\beta}\right), \qquad 0 \leqq \tau \leqq \pi\beta, \qquad 0 \leqq \beta \leqq 1,$$

in the interval $0 \leqq \tau \leqq \tau_1$.

b. *If $r = r(\tau)$ passes down through the point (τ_o, r_o) with slope r_o' satisfying $-\sqrt{1 - r_o^2} \leqq r_o' \leqq 0$, then there is a unique translated member of F, say $r = r_2(\beta_o, \tau - \mu)$ which passes through (τ_o, r_o) with slope r_o'. If $r_2(\beta_o, \tau - \mu)$ and $r(\tau)$ are nonincreasing for $\bar{\tau} \leqq \tau \leqq \tau_o$, then $r(\tau) \leqq r_2(\beta_o, \tau - \mu)$ for $\bar{\tau} \leqq \tau \leqq \tau_o$.*

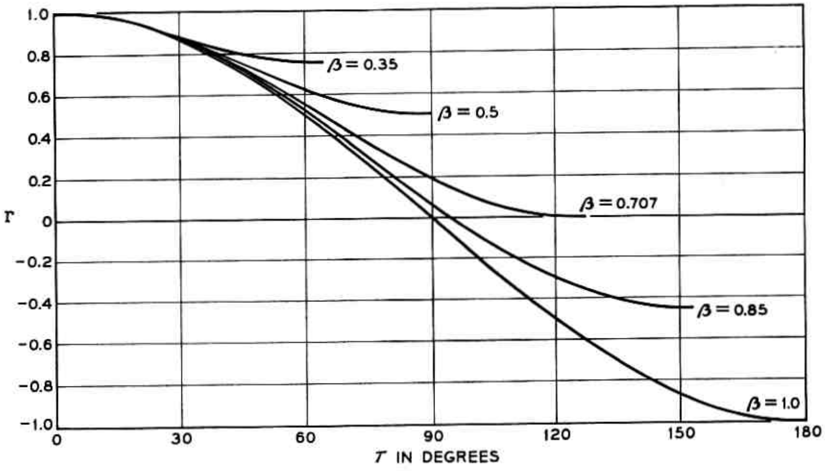


Fig. 2 — The family F.

Proof — Part a of the lemma will be deduced from part b. The first conclusion of part b is the remark (2) above. The second conclusion of part b follows from the inequality (43). If $| r | \neq 1$, this latter can be written by elementary algebraic manipulations as

$$-\frac{1 - r^2 - r'^2}{1 - r^2} \leqq r'' + \frac{rr'^2}{1 - r^2} \leqq \frac{1 - r^2 - r'^2}{1 - r^2}.$$

The right-hand inequality can be rewritten as

$$\frac{r''}{(1 - r)^2} + \frac{r'^2}{(1 - r)^3} \leqq \frac{1}{(1 - r)^2},$$

or, if $r' \leqq 0$, as

$$\frac{2r'r''}{(1 - r)^2} + \frac{2r'^3}{(1 - r)^3} \geqq \frac{2r'}{(1 - r)^2},$$

or

$$\frac{d}{d\tau} \frac{r'^2}{(1 - r)^2} \geqq 2 \frac{d}{d\tau} \frac{1}{1 - r}.$$

Integrate this expression from τ to τ_o with $\tau < \tau_o$ to obtain

$$\frac{r'^2}{(1 - r)^2} - \frac{2}{1 - r} \leqq \frac{r_o'^2}{(1 - r_o)^2} - \frac{2}{1 - r_o}, \tag{46}$$

where the subscript o refers to quantities evaluated at τ_o. Denote the right member of this inequality by $- 1/h^2$, and note that, as is indicated by the notation,

$$\frac{1}{h^2} = \frac{2(1 - r_o) - r_o'^2}{(1 - r_o)^2} \geqq \frac{(1 + r_o)(1 - r_o) - r_o'^2}{(1 - r_o)^2} = \frac{1 - r_o^2 - r_o'^2}{(1 - r_o)^2} \geqq 0,$$

by (44). Inequality (46) now becomes

$$r'^2 - 2(1 - r) \leqq -\frac{1}{h^2}(1 - r)^2,$$

or what is the same

$$r'^2 \leqq \frac{1}{h^2}(1 - r)(r - \lambda),$$

where

$$\lambda = 1 - 2h^2. \tag{47}$$

It follows then that

$$\frac{r'}{\sqrt{(1 - r)(r - \lambda)}} \geqq \frac{1}{h},$$

with h a nonnegative quantity. Integrate this again from τ to τ_o to obtain

$$\arcsin \frac{r_o - (1 + \lambda)/2}{(1 - \lambda)/2} - \arcsin \frac{r - (1 + \lambda)/2}{(1 - \lambda)/2} \geqq -\frac{\tau_o - \tau}{h}.$$

Thus one finds

$$r(\tau) \leqq \frac{1 + \lambda}{2} + \frac{1 - \lambda}{2} \sin \left[\frac{\tau_o - \tau}{h} + \arcsin \frac{r_o - (1 + \lambda)/2}{(1 - \lambda)/2} \right] \qquad (48)$$
$$\equiv q(\tau).$$

This inequality is valid in a τ-range to the left of τ_o until either $q(\tau)$ or $r(\tau)$ has a local maximum.

Now by (47), $q(\tau)$ can be written

$$q(\tau) = 1 - h^2 + h^2 \cos \left(\frac{\tau - \mu}{h} \right),$$

for suitably defined μ, and one finds by using the various definitions

$$q(\tau_o) = r_o$$
$$q'(\tau_o) = r_o'.$$

Thus $q(\tau)$ is the member of the family F which, when translated in the τ-direction, passes through the point (τ_o, r_o) with slope r_o'. To the left of τ_o, the curve $r = r(\tau)$ remains below this translated member of F. Part b is thus proved.

Now suppose that $r = r(\tau)$ intersects a member of the family F from below, say at (τ_o, r_o) with $\tau_o \leqq \tau_1$. Let the parameter value of this member of F be β_o. Since $0 \geqq r'(\tau_o) > r_2'(\beta_o, \tau_o)$, the translated member of F passing through (τ_o, r_o) with slope $r'(\tau_o)$ has a parameter value $\beta = \beta_1 < \beta_o$. This translated version of $r = r_2(\beta_1, \tau)$ has no local maximum in the interval $(0, \tau_o)$, and its value at $\tau = 0$ is less than unity. One thus has the contradiction $r(0) < 1$ and the lemma is proved.

Theorem 12 — Let $r(\tau)$ be a class 2 covariance. Then

$$r(\tau) \geqq \cos \tau, \qquad 0 \leqq \tau \leqq \pi.$$

Proof: In a region where $r'(\tau) \leqq 0$, inequality (44) implies

$$-1 \leqq -\frac{r'}{\sqrt{1 - r^2}} \leqq 1.$$

Integrating from τ_o to $\tau > \tau_o$ assuming that $r'(\tau) \leqq 0$ throughout (τ_o, τ), one finds

$$-(\tau - \tau_o) + \arccos r_o \leqq \arccos r \leqq (\tau - \tau_o) + \arccos r_o,$$

where $r_o = r(\tau_o)$. This in turn implies $\cos[\tau - \tau_o - \arccos r_o] \geqq r(\tau)$ and $r(\tau) \geqq \cos[\tau - \tau_o + \arccos r_o]$, where the former inequality holds from $\tau = \tau_o$ until the cosine assumes the value unity, and the latter inequality holds from $\tau = \tau_o$ until the cosine assumes the value minus unity. The result may be stated as follows: Let the class 2 covariance $r(\tau)$ pass downward (= not upward) through the point (τ_o, r_o) in the τ-r plane. The curve $r = \cos \tau$ can be translated in the τ-direction to pass downward through (τ_o, r_o). Then to the right of τ_o, $r = r(\tau)$ lies above this translated cosine curve until either the cosine curve or $r(\tau)$ has its next local minimum. Similarly, a cosine curve can be translated to pass up through (τ_o, r_o). To the right of τ_o, $r = r(\tau)$ lies below this translated cosine curve until either $r(\tau)$ has its next local minimum or the cosine curve has its next maximum.

A similar result holds if $r(\tau)$ increases through (τ_o, r_o).

Now let $\tau_o = 0$, $r_o = 1$. Then $r = r(\tau)$ lies above $r = \cos \tau$ until the first minimum of either. If the first minimum of $r(\tau)$ occurs at $\tau_1 \geqq \pi$, the theorem is proved. Suppose now $\tau_1 < \pi$ and that $r = r(\tau)$ crosses $r = \cos \tau$ in $(0,\pi)$. The first such crossing must be downward, since $r(\tau) \geqq \cos \tau$ from 0 to τ_1. If the crossing is at $\bar{\tau}$, then $r(\bar{\tau}) = \cos \bar{\tau}$, and $r'(\bar{\tau}) \leqq -\sin \bar{\tau}$. If indeed $r'(\bar{\tau}) < -\sin \bar{\tau}$, one obtains from (43) the contradiction $1 \geqq r^2(\bar{\tau}) + r'^2(\bar{\tau}) > \cos^2\bar{\tau} + \sin^2\bar{\tau} = 1$. On the other hand, if the crossing takes place with $r'(\bar{\tau}) = -\sin \bar{\tau}$, then *b* of Lemma 2 shows that $r(\tau) \leqq \cos \tau$ for $\tau < \bar{\tau}$ which contradicts the assumption that the crossing was downward. Thus, the theorem is proved.

Theorem 13 — If $r(\tau)$ is of class 2 and

$$r(\tau) \geqq 0, \qquad 0 \leqq \tau \leqq \frac{\pi}{\sqrt{2}},$$

then

$$r(\tau) \geqq r_2\left(\frac{1}{\sqrt{2}}, \tau\right) = \tfrac{1}{2} + \tfrac{1}{2}\cos\sqrt{2}\,\tau = \cos^2\left(\frac{\tau}{\sqrt{2}}\right)$$

for

$$0 \leqq \tau \leqq \frac{\pi}{\sqrt{2}}.$$

The theorem is a consequence of repeated applications of Lemma 2. We prove the theorem by supposing it false and then arrive at a contradiction. We refer to the curve $r = r_2(1/\sqrt{2},\tau)$, $0 \leqq \tau \leqq \pi/\sqrt{2}$ as C.

Suppose now that $r(\tau) \geqq 0$ for $0 \leqq \tau \leqq \pi/\sqrt{2}$ and that some point P_o on $r = r(\tau)$, say (τ_o, r_o), lies below C. Denote $r'(\tau_o)$ by r_o'. We can suppose P_o chosen so that $r_o' < 0$, since $r = r(\tau)$ cannot be nondecreasing at all points where it lies below C. Let the horizontal line $r = r_o$ through P_o intersect C at P_1 and denote the slope of C at P_1 by $C'(r_o)$. The point P_1 has larger abscissa than the point P_o. The curve $r = r(\tau)$ possesses a continuous derivative. As the height r_o of the horizontal line $r = r_o$ is continuously decreased to zero from its initial value, a value must be found with P_o to the left of P_1 and $r_o' \geqq C'(r_o)$. By b of Lemma 2, a curve of the family F with parameter value $\beta \leqq 1/\sqrt{2}$ can be translated to the left to pass through P_o with slope r_o'. In the interval $0 \leqq \tau \leqq \tau_o$, this translated member of F lies strictly below C and is monotone. The first local maximum of $r = r(\tau)$ to the left of P_o therefore lies below C as must also the local minimum just preceding this maximum. A curve of F can then be translated to pass through this local minimum with slope zero, and repetition of the argument shows that all local maxima of $r = r(\tau)$ for $0 \leqq \tau \leqq \tau_o$ lie below C. In particular $r(0) < 1$, which contradicts the initial assumption concerning $r(\tau)$. Q.E.D.

Theorem 14 — Let the covariance $r(\tau)$ have the behavior

$$r(\tau) = 1 - \frac{\tau^2}{2} + m\frac{\tau^4}{4!} + o(\tau^4),$$

near $\tau = 0$. Then

$$r(\tau) \leqq r_2\left(\frac{1}{\sqrt{m}}, \tau\right), \qquad 0 \leqq \tau \leqq T_1,$$

with $r_2(\beta,\tau)$ given by (45). Here $T_1 = \min(\beta\pi,\tau_o)$ and τ_o is the smallest positive value of τ for which $r(\tau) = 1 - 2/m$.

Proof — The first four derivatives of $r(\tau)$ exist at $\tau = 0$. From the Bochner representation for $r(\tau)$, it is easy to show using Schwarz's inequality that

$$v^2 \equiv m - 1 \geqq 0. \tag{49}$$

It also follows that $r''(\tau)$ exists everywhere and is continuous.

The Gaussian process $X(t)$ having covariance $r(\tau)$ has first and second derivates $X'(t)$ and $X''(t)$ almost everywhere with probability 1. The covariance matrix of the random variables $X(0), X(t), X'(t), X''(t)$ is

$$\begin{vmatrix} 1 & r & r & r'' \\ r & 1 & 0 & -1 \\ r' & 0 & 1 & 0 \\ r'' & -1 & 0 & m \end{vmatrix}.$$

The determinant of this matrix cannot be negative. This is equivalent to the inequalities

$$-v \leqq \frac{r + r''}{\sqrt{1 - r^2 - r'^2}} \leqq v.$$

In any region where $r' \leqq 0$, the right-hand inequality gives

$$\frac{r'(r + r'')}{\sqrt{1 - r^2 - r'^2}} = -\frac{d}{d\tau}\sqrt{1 - r^2 - r'^2} \geqq vr'.$$

Integrate this from 0 to τ to obtain

$$\sqrt{1 - r^2 - r'^2} \leqq v(1 - r). \tag{50}$$

Note that if τ_1 is the first positive value of τ for which $r'(\tau) = 0$, (50) gives

$$r(\tau_1) \leqq \frac{v^2 - 1}{v^2 + 1}.$$

Thus we have the interesting side result that if $r(\tau)$ is everywhere non-negative $v^2 \geqq 1$ or $m \geqq 2$.

Squaring the inequality (50) and rearranging the terms, one finds

$$r'^2 \geqq (1 + v^2)(1 - r)(r - \alpha),$$

where

$$\alpha = \frac{v^2 - 1}{v^2 + 1} < 1. \tag{51}$$

Since $r' \leqq 0$, this implies

$$\frac{r'}{\sqrt{(1 - r)(r - \alpha)}} \leqq -\sqrt{1 + v^2},$$

if $r > \alpha$. Integration from 0 to τ yields

$$\arcsin\frac{r - (1 - \alpha)/2}{(1 - \alpha)/2} - \frac{\pi}{2} \leqq -\sqrt{1 + v^2}\,\tau,$$

where it is assumed that $\tau \leqq \tau_o$. If then

$$\left| \frac{\pi}{2} - \sqrt{1 + v^2}\,\tau \right| \leqq \frac{\pi}{2},$$

$$\frac{r - (1 + \alpha)/2}{(1 - \alpha)/2} \leqq \sin\left(\frac{\pi}{2} - \sqrt{1 + v^2}\tau\right),$$

or, what is the same thing in virtue of the definitions (49) and (51),

$$r(\tau) \leqq 1 - \frac{1}{m^2} + \frac{1}{m^2} \cos (m\tau).$$

The theorem is thus proved.

2.8 *Proof of Theorem 9*

Let $h(\xi)$ be nonnegative for $0 \leqq \xi \leqq \theta$ and zero elsewhere. Then

$$Y(t) = \int_{t-\theta}^{t} h(t - t')X(t')\,dt' = \int_{-\infty}^{\infty} du\, h(u)X(t - u)\,du,$$

will certainly be nonnegative for $0 \leqq t \leqq T$ whenever $X(t)$ is nonnegative for $-\theta \leqq t \leqq T$. The probability that the Y process be nonnegative in $(0,T)$ is therefore not less than the probability that the X process be nonnegative in $(-\theta,T)$. If X is Gaussian with mean zero and covariance $r(\tau)$, then Y is Gaussian with mean zero and covariance

$$\begin{aligned} r_\theta(\tau) &= EY(t)Y(t + \tau) = \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\; h(u)h(v)EX(t - u)X(t + \tau - v) \\ &= \int_{-\infty}^{\infty} du \int_{-\infty}^{\infty} dv\; h(u)h(v)r(\tau - u + v) \\ &= \int_{-\infty}^{\infty} dx\; r(\tau - x) \int_{-\infty}^{\infty} d\xi\; h(x + \xi)h(\xi). \end{aligned}$$

One has then $P[T,r_\theta(\tau)] \geqq P[T + \theta,r(\tau)]$, which is Theorem 9.

2.9 *Proof of Theorem 10*

Let $0 = t_1 < t_2 < \cdots < t_n = T$ be a partition of $(0,T)$. Define $Q_n(\mathbf{r})$ by

$$Q_n(\mathbf{r}) = \frac{\Pr\,(X(t_1) < 0, X(t_i) \geqq 0, i = 2, 3, \cdots, n)}{\Pr\,(X(t_1) < 0, X(t_2) \geqq 0)}, \tag{52}$$

where $X(t)$ is a Gaussian process with zero mean and class 2 covariance

$r(\tau)$. As the partition is refined with mesh tending to zero, $Q_n(\mathbf{r})$ approaches $Q[T,r(\tau)]$ as a limit. The numerator on the right of (52) is $P_n(\hat{\mathbf{r}})$ where

$$\hat{\mathbf{r}} = \begin{vmatrix} 1 & -r(t_2) & -r(t_3) & \cdots & -r(t_n) \\ -r(t_2) & 1 & r(t_3 - t_2) & \cdots & r(t_n - t_2) \\ -r(t_3) & r(t_3 - t_2) & 1 & \cdots & r(t_n - t_3) \\ \vdots & & & & \vdots \\ -r(t_n) & r(t_n - t_2) & r(t_n - t_3) & \cdots & 1 \end{vmatrix}, \quad (53)$$

and as usual $P_n(\mathbf{r})$ denotes the probability that n normal variates of mean zero and covariance matrix $\mathbf{r}$ be nonnegative. Note that the denominator of the right of (52) depends only on $r(t_2)$.

Let another Gaussian process, $Y(t)$, have class 2 covariance $q(\tau)$. We define $r^{-1}(\tau), q^{-1}(\tau), h(\tau) = r^{-1}[q(\tau)]$ as in Section 2.3 and set $g(t) = q^{-1}[r(t)] = h^{-1}(t)$. Note that $g(t)$ is strictly monotone within its domain of definition. Assume that T is within the domain of definition of g. With the points t_i given as in (52), set $\tau_i = g(t_i), i = 1,2,\cdots,n$. The points $0 = \tau_1 < \tau_2 < \cdots < \tau_n = g(T)$ form a partition of the interval $(0, g(T))$. The mesh of this partition tends to zero with the mesh of the t_i partition.

Consider now the approximation to $Q[g(T),q(\tau)]$ given by

$$Q_n(\mathbf{q}) = \frac{\Pr\{Y(\tau_1) < 0, Y(\tau_i) \geqq 0, i = 1, 2, 3, \cdots, n\}}{\Pr\{Y(\tau_1) < 0, Y(\tau_2) \geqq 0\}}. \quad (54)$$

The numerator here is $P_n(\hat{\mathbf{q}})$ where $\hat{\mathbf{q}}$ is given by (53) with r replaced by q and t replaced by τ. Since $\tau_i = g(t_i), q(\tau_i) = r(t_i), i = 1,2,\cdots,n$, so that the first row and column of $\hat{\mathbf{r}}$ are the same as the first row and column of $\hat{\mathbf{q}}$. For any other entry of $\hat{\mathbf{r}}$ with $t_i \geqq t_j$, one has

$$\begin{aligned} r(t_i - t_j) &= q[g(t_i - t_j)] \\ &= q[\tau_i - \tau_j + \{g(t_i - t_j) - g(t_i) + g(t_j)\}]. \end{aligned}$$

Since $q(\tau)$ is nonincreasing

$$r(t_i - t_j) \leqq q(\tau_i - \tau_j)$$

and hence by Lemma 1

$$P_n(\hat{\mathbf{r}}) \leqq P_n(\hat{\mathbf{q}}),$$

provided

$$g(t_i - t_j) - g(t_i) + g(t_j) \geqq 0.$$

or what is the same thing, provided

$$g(x) + g(y) \geqq g(x + y), \tag{55}$$

where $0 \leqq x = t_j < t_i = x + y$.

When (55) is satisfied, the numerator of (54) is not less than the numerator of (52). The denominators of these expressions are equal since they are the same function of $r(t_2) = q(\tau_2)$. Therefore, $Q_n(\mathbf{q}) \geqq Q_n(\mathbf{r})$. The conclusion of Theorem 10 results by passing to the limit as the t partition is refined.

2.10 *Generalizations*

A number of the results presented in this paper can be generalized in a direct manner. We only mention here an obvious extension of Theorem 1.

In the derivation of Lemma 1, the lower limit of integration for x_i in (33) can be replaced by a_i. Now choose $a_i = a(t_i)$ with $a(t)$ a given function defined for $0 \leqq t \leqq T$, and where the points t_i form a partition of $(0,T)$. Proceeding as in the derivation of Theorem 1, one arrives at the following more general result. Let $X(t)$ be a Gaussian process with $EX(t) = 0$, $EX(t)X(s) = r(s,t)$. Let $Y(t)$ be a Gaussian process with $EY(t) = 0$, $EY(t)Y(s) = q(s,t)$. Then if

$$r(s,s) = q(s,s), \qquad 0 \leqq s \leqq T$$

and

$$r(s,t) \geqq q(s,t), \qquad 0 \leqq s,t \leqq T$$

$$\Pr\{X(t) \geqq a(t), 0 \leqq t \leqq T\} \geqq \Pr\{Y(t) \geqq a(t), 0 \leqq t \leqq T\}.$$

2.11 *Asymptotics*

As already remarked in the introduction of this paper, there appears to be little in the literature concerning the asymptotic behavior of $P[T,r(\tau)]$ for large T. Intuition would indicate exponential falloff for a wide class of covariances. Example (*ii*) of Section 1.1, though special in nature since $r_2(\beta,\tau)$ is periodic, provides a counterexample to exponential behavior, and so the class must be carefully defined. Here, by the two bounds presented in Section 1.4, we have shown exponential behavior for nonnegative covariances that vanish identically for τ greater than some $\tau_o > 0$. Recently, by using Theorem 1, M. Rosenblatt has established an asymptotic exponential upper bound for $P[T,r(\tau)]$ for all covariances which are ultimately majorized by a decaying exponential. This, together with the lower bound of Section 1.4, establishes the asymptotic exponential behavior of $P[T,r(\tau)]$ for all nonnegative covariances that themselves decay exponentially. Professor Rosenblatt has also established that if $r(\tau) \rightarrow 0$ with increasing τ, then $T^n P[T,r(\tau)] \rightarrow 0$ with increasing T for every $n > 0$.

We conclude with the remark that from (23) of Section 1.7, one can show that asymptotic exponential behavior of $P[T,r(\tau)]$ implies asymptotic exponential behavior for $Q[T,r(\tau)]$.

III. ACKNOWLEDGMENTS

This paper is the outgrowth of many years of interest in the one-sided barrier problem. Some of the results reported were obtained as early as 1952. During these years, I have discussed the problem in detail with many of my friends. Among them are J. Chover, E. N. Gilbert, M. Kac, S. P. Lloyd, B. McMillan, H. O. Pollak, S. O. Rice, and M. Rosenblatt. Through their many suggestions and their counsel, all have contributed to my understanding of the problem and to this paper. To them my thanks. I am indebted to J. A. McFadden for a number of the references that follow.

REFERENCES

1. Bartlett, M. S., *An Introduction to Stochastic Processes*, Cambridge University Press, 1955, Chapter 3.
2. Bendat, J. S., *Principles and Applications of Random Noise Theory*, John Wiley and Sons, New York, 1958, Chapter 10.
3. Brown, W. M., Some Results on Noise Through Circuits, I.R.E. Trans. on Info. Theory, **IT-5,** Special Supplement, May, 1959, pp. 217–227.
4. Chover, J., Certain Convexity Conditions on Matrices with Applications to Gaussian Processes, to appear in Duke Math. J.
5. Darling, D. A., and Siegert, A. J. F., The First Passage Problem for a Continuous Markov Process, Ann. Math. Stat., **24,** 1953, pp. 624–639.
6. Debart, H. P., Zeros d'un Signal Aléatoire Stationnaire, Câbles et Transmission, **14,** 1960, pp. 191–199.
7. Ehrenfeld, S., et al, *Theoretical and Observed Results for the Zero and Ordinate Crossing Problems of Stationary Gaussian Noise with Application to Pressure Records of Ocean Waves*, Tech. Report No. 1, Research Division, College of Engineering, New York University, New York, Dec., 1958.
8. Helstrom, C. W., The Problem of First-passage Times and Number of Crossings for Gaussian Stochastic Processes, I.R.E. Trans. on Info. Theory, **IT-3,** 1957, pp. 232–237.
9. Kac, M., Some Remarks on Oscillators Driven by a Random Force, I.R.E. Trans. on Circuit Theory, **CT-7,** 1960, pp. 476–479.
10. Kac, M., and Slepian, D., Large Excursions of Gaussian Processes, Ann. Math. Stat., **30,** 1959, pp. 1215–1228.
11. Kohlenberg, A., Notes on the Zero Distribution of Gaussian Noise, Technical Memorandum No. 44, M.I.T. Lincoln Lab., Lexington, Mass., Oct., 1953.
12. Kuznetsov, P. I., Stratonovich, P. L., and Tikhonov, V. I., On the Duration of Exceedances of a Random Function, Zhurnal Tekhnicheskoi Fiziki, **24,** 1954, pp. 103–112 (in Russian). English translation by N. R. Goodman available as Sci. Paper No. 5, Eng. Stat. Group, College of Eng., New York University, New York, March, 1956.

13. Longuet-Higgins, M. S., On the Intervals between Successive Zeros of a Random Function, Pro. Roy. Soc. (A), **246,** 1958, pp. 99–118.
14. Longuet-Higgins, M. S., The Distribution of Intervals between Zeros of a Stationary Random Function, Phil. Trans. Royal Society, to appear.
15. McFadden, J. A., The Axis-Crossing Intervals of Random Functions, I.R.E. Trans. on Info. Theory, **IT-2,** 1956, pp. 146–150; The Axis-Crossing Intervals of Random Functions II, I.R.E. Trans. on Info. Theory, **IT-4,** 1958, pp. 14–24; The Fourth Product Moment of Infinitely Clipped Noise, I.R.E. Trans. on Info. Theory, **IT-4,** 1958, pp. 159–162; The Axis Crossings of a Stationary Gaussian Markov Process, I.R.E. Trans. on Info. Theory, **IT-7,** 1961, pp. 150–153.
16. Middleton, David, *Statistical Communication Theory*, McGraw-Hill Book Co., New York, 1960, Section 9.4, pp. 426–434.
17. Miller, I., and Freund, J. E., Some Results on the Analysis of Random Signals by Means of a Cut-counting Process, J. Appl. Phys., **27,** 1956, pp. 1290–1293; Some Distribution Theory Connected with Gaussian Processes, Va. Poly. Inst. Dept. of Stat. Report, 1956.
18. Palmer, D. S., Properties of Random Functions, Proc. Cambridge Phil. Soc., **52,** 1956, pp. 672–686.
19. Rice, S. O., Mathematical Analysis of Random Noise, B.S.T.J., **23,** 1944, pp. 282–332; **24,** 1945, pp. 46–156. See esp. Sections 3.3, 3.4.
20. Rice, S. O., Statistical Properties of a Sine Wave Plus Random Noise, B.S.T.J., **27,** 1948, pp. 109–157.
21. Rice, S. O., Distribution of the Duration of Fades in Radio Transmission, B.S.T.J., **37,** 1958, pp. 581–635.
22. Siegert, A. J. F., On the Roots of Markoffian Random Functions, Report RM-447, Rand Corporation, Santa Monica, Calif., Sept., 1950; On the First Passage Time Probability Problem, Phys. Rev., **81,** 1951, pp. 617–623.
23. Slepian, D., First Passage Time for a Particular Gaussian Process, Ann. Math. Stat., **32,** 1961, pp. 610–612.
24. Tikhonov, V. I., and Amiantov, I. N., Probability Density for the Duration of Fluctuations, Radiotekhnika, **15,** 1960, pp. 10–20 (in Russian). A number of other Russian references are given here.
25. Wang, M. C., and Uhlenbeck, G. E., On the Theory of Brownian Motion II, Rev. Mod. Phys., **17,** 1945, pp. 323–342. See Sections 12b, 12c.
26. Bialyi, L. I., Density of Distribution of Intervals between Zeros of a Narrow-Band Normal Stationary Random Process, Radiotekhnika i Elektronika, **4,** 1959, p. 266 (in Russian).
27. Blötekjaer, K., An Experimental Investigation of Some Properties of Band-pass Limited Gaussian Noise, I.R.E. Trans. on Info. Theory, **IT-4,** 1958, pp. 100–102.
28. Favreau, R. R., Low, H. and Pfeffer, I. Evaluation of Complex Statistical Functions by an Analog Computer, I.R.E. Convention Record, 1956, Part 4, pp. 31–37.
29. Rainal, A. J., Digital Measurement of Axis-Crossing Intervals, Electronics, **33,** No. 23, June 3, 1960, pp. 88–91.
30. Steinberg, H., Schultheiss, P. M., Wogrin, C. A., and Zweig, F., Short-Time Frequency Measurements of Narrow-Band Random Signals by Means of a Zero Counting Process, J. Appl. Phys., **26,** 1955, pp. 195–201.
31. Velichkin, A. I. and Ponomareva, V. D., Experimental Investigation of the Duration of Noise Peaks, Radiotekhnika, **15,** 1960, pp. 21–26 (in Russian).
32. White, G. M., An Experimental System for Studying the Zeros of Noise, Cruft Lab. Tech. Report 261, Harvard University, May, 1957; Experimental Determination of the Zero-crossing Distribution W(N; T), Cruft Lab. Tech. Report 265, Sept. 1957; Zeros of Gaussian Noise, J. Appl. Phys., **29,** 1958, pp. 722–729.
33. Schläfli, L., On the Multiple Integral $\int^n dxdy \cdots dz$ whose limits are $p_1 = a_1x + b_1y + \cdots + h_1z > 0$, $p_2 > 0, \cdots, p_n > 0$, $x^2 + y^2 \cdots + z^2 < 1$, Quart. J. Pure and Appl. Math., **2,** 1858, pp. 261–301, and **3,** 1860, pp. 54–68, pp. 97–107.

34. Coxeter, H. S. M., *Regular Polytopes*, Pitman Co., New York, 1947, pp. 126–144; The Functions of Schläfli and Lobatschefsky, Quart. J. of Math., **6,** 1935, pp. 13–29.

35. Sommerville, D. M. Y., *The Geometry of N Dimensions*, E, P. Dutton and Co., New York, 1929.

36. Van der Vaart, H. R., The Content of Some Classes of Non-Euclidean Polyhedra for Any Number of Dimensions, Pro. Akademie van Wetenschappen, Amsterdam, A., **18,** 1955, pp. 199–221.

37. Rogers, C. A., An Asymptotic Expansion for Certain Schläfli Functions, J. London Math. Soc., **36,** 1961, pp. 78–80.

38. Ruben, H., A Power Series Expansion for a Class of Schläfli Functions, J. London Math. Soc., **36,** 1961, pp. 69–77.

39. Riordan, J., *Combinatorial Analysis*, John Wiley, New York, 1958.

40. Cramér, H., *Mathematical Methods of Statistics*, Princeton University Press, Princeton, N. J., 1946.

41. Plackett, R. L., A Reduction Formula for Normal Multivariate Integrals, Biometrika, **41,** 1954, pp. 351–360.

42. Gupta, S. S., Bibliography on the Multivariate Normal Integrals and Related Topics, Ann. Math. Stat., to appear.

43. Pearson, K., *Tables of the Incomplete Beta-Function*, Cambridge University Press, 1934.

44. Gaddum, J. W., Distance Sums on a Sphere and Angle Sums in a Simplex, Am. Math. Monthly, **63,** 1956, pp. 91–96.

Probability Distributions for the Phase Jitter in Self-Timed Reconstructive Repeaters for PCM

By M. R. AARON and J. R. GRAY

(Manuscript received August 25, 1961)

Probability distributions for the timing jitter in the output of an idealized self-timed repeater for reconstructing a PCM signal are approximated. Primary emphasis is focused on self-timed repeaters employing complete retiming. In this case the probability distribution for the timing jitter reduces to the computation of the phase error in the zero crossings at the output of the tuned circuit excited by a jitter-free binary pulse train. It is assumed that the tuned circuit is mistuned from the pulse repetition frequency, and the individual pulses are either impulses or raised cosine pulses. Both random pulse trains and random plus periodic trains are considered. In general, the probability distributions are skewed in the direction of increasing phase error. The approach to the normal law in the neighborhood of the mean when the circuit Q becomes arbitrarily large is demonstrated. Results obtained from the analytical approach are compared with two computer methods for the case of random impulse excitation of a tuned circuit characterized by a Q of 125 and mistuning of 0.1 per cent. Excellent agreement between the three techniques is displayed. For no mistuning and raised cosine excitation two methods for computing the phase error are given and numerical results obtained from both techniques agree closely.

Some attention is given to an idealized version of a reconstructive repeater employing partial retiming and it is shown that the timing performance of such a repeater for random signals is very much inferior to the completely retimed repeater.

I. INTRODUCTION

Over the past several years the problem of maintaining pulse spacing within very close bounds in PCM transmission has received considerable attention both theoretically and experimentally. The effects of timing jitter in degrading repeater performance, in introducing distortion in the decoded analog signal, and in enhancing the difficulty of dropping or adding several pulse trains in time have been documented.[1-8] Sources of mistiming in a self-timed reconstructive repeater are well catalogued and include: noise, crosstalk, mistuning, finite pulse width effects, and amplitude to phase conversion in nonlinear devices. The first four of these sources have been considered in various analyses of timing jitter in self-timed and separately-timed PCM repeaters. Amplitude to phase conversion in nonlinear circuits has received attention primarily from the experimental viewpoint.

The majority of the theoretical work to date has been concerned with timing errors in self-timed repeaters when the timing-wave extractor is a simple tuned circuit. For a random pulse train exciting the tuned circuit in the presence of noise and mistuning, results have been obtained for the mean displacement and the standard deviation of the zero crossings from their ideal location. This analysis is appropriate to repeaters employing complete retiming. These time displacements can also be considered as phase errors and we will use this terminology in what follows. If the probability density function for the phase error is normal, the mean and standard deviation are sufficient for a complete statistical description. In this paper we will show that in general the probability density function is not normal, and is inherently unsymmetrical about the mean.

An approximation to the probability density and the cumulative distribution for the phase error at the output of a mistuned resonant circuit will be derived for both random and random plus periodic pulse trains. A completely random pulse train is defined to be one in which pulses and spaces are equally likely. The individual pulses of the binary pulse train are assumed to be jitter free and are either impulses or raised cosine pulses. The approach to the normal law when the circuit Q is large is demonstrated. For a value of Q of 125, and a mistuning of 0.1 per cent from the pulse repetition frequency a comparison of numerical results obtained from the analytical approach and two computer methods is made. Agreement among the three approaches is excellent.

Our plan of attack is to place all of the manipulations required to specify the tuned circuit response to the most general pulse trains in the Appendix and concentrate on most of the probabilistic notions in the main body of the paper. Appendix A covers the response of the tuned circuit to a random or random plus periodic binary pulse train of arbitrary pulse shape, and Appendix B is concerned with the specialization to raised cosine pulses. Section II of the text deals with the terminology required, covers the tuned circuit response to impulses, and briefly summarizes the results of Appendices A and B. In Section III, the probability density function for the phase error is derived. Section IV is devoted to the cumulative distribution function and Section V alludes to the semi-invariants that are required in the evaluation of the density and cumulative distribution functions. These semi-invariants are derived in Appendix C. The approach of the probability density function for the phase error to the normal law as the circuit Q becomes arbitrarily large is displayed in Section VI with the algebraic support relegated to Appendix D. The comparison of numerical results mentioned previously with other computer approaches is made in Section VII. For zero mistuning, but finite pulse width excitation, it can be shown that the probability distributions for the phase error can be related directly to the probability distribution for the timing wave amplitude. This is demonstrated in Section VIII. A discussion of further numerical results is given in Section IX. We consider an idealized model of a partially retimed repeater in Section X for purposes of comparison with the results of Section IX. A wrap-up of the procedures, results, and future work concludes the paper.

II. RESPONSE OF THE TIMING CIRCUIT

Before we go on to the general equation for the phase error due to finite pulse width and mistuning, we will specialize to impulse excitation of a simple tuned circuit characterized by its Q and mistuning from the pulse repetition frequency. This should provide the casual reader with some feel for how the more general equation for the phase error arises without going through the detailed manipulations of Appendices A and B. The procedure adopted in the analysis to follow is equivalent to that of H. E. Rowe.[2]

Assuming the input to the timing circuit to be a train of jitter-free unit impulses occurring at random with spacing T, the excitation may be represented as

$$f(t) = \sum_{n=-\infty}^{n=\infty} a_n \delta(t - nT), \tag{1}$$

where a_n is a random variable taking the values 0 or 1 with probability $\frac{1}{2}$,* $\delta(t - nT)$ is a unit impulse whose time of arrival is nT, and the spacing T is the reciprocal of the pulse repetition frequency f_r. For a parallel resonant circuit the impulse response is given by

* Unless otherwise specified, the case of equal likelihood will be considered in all calculations.

$$h(t) = A\,e^{-(\pi/Q)f_o t}\cos(2\pi f_o t + \varphi), \tag{2}$$

where

$$f_o = \frac{1}{2\pi}\sqrt{\frac{1}{LC} - \left(\frac{1}{2RC}\right)^2}, \qquad A = \frac{1}{2QC}\sqrt{4Q^2 + 1},$$

$$Q = 2\pi f_o RC, \qquad \text{and} \qquad \varphi = \tan^{-1}\frac{1}{2Q}.$$

Here f_o is the natural resonant frequency as distinguished from the steady-state resonant frequency $f_s = (1/2\pi)\sqrt{1/LC}$. Combining (1) and (2), the total response to all impulses occurring in time slots up to and including the one at $t = 0$ may be written as

$$F(t) = A\sum_{n=-\infty}^{n=0} a_n\,e^{-(\pi/Q)f_o(t-nT)}\cos[2\pi f_o(t - nT) + \varphi]. \tag{3}$$

This expression gives the output of the timing circuit for values of t in the interval between $t = 0$ and the arrival time of the next impulse. Rewriting (3) in the form of a carrier with both amplitude and phase modulation we get

$$F(t) = A\sqrt{x^2 + y^2}\,e^{-(\pi/Q)f_o t}\cos[2\pi f_o t + \varphi + \theta], \tag{4}$$

where

$$\theta = \tan^{-1}\frac{y}{x},$$

$$x = \sum_{n=0}^{\infty} a_n\,e^{-(\pi/Q)f_o nT}\cos 2\pi f_o nT, \quad \text{and}$$

$$y = \sum_{n=0}^{\infty} a_n\,e^{-(\pi/Q)f_o nT}\sin 2\pi f_o nT.$$

In the above x and y represent the in-phase and quadrature components of the response. If the tank could be tuned exactly to the pulse repetition frequency ($f_o \equiv f_r \equiv 1/T$), then the phase modulation would disappear and the amplitude modulation would be dependent on x alone. In practical applications this is not possible and the phase shift θ does occur. If we denote the fractional mistuning $\Delta f/f_r$ by k, we may write f_o in terms of f_r as follows

$$f_o = f_r(1 + k).$$

In this case (4) becomes, neglecting k with respect to unity in the exponential term

$$F(t) = A\sqrt{x^2 + y^2}\, e^{-(\pi/Q)f_r t} \cos\,[2\pi f_r(1 + k)t + \varphi + \theta], \tag{5}$$

with

$$x = \sum_{n=0}^{\infty} a_n\, e^{-(\pi/Q)\,n} \cos 2\pi kn,$$

$$y = \sum_{n=0}^{\infty} a_n\, e^{-(\pi/Q)\,n} \sin 2\pi kn,$$

and

$$\theta = \tan^{-1} y/x.$$

To illustrate the relationship between the timing deviation t_d and the phase error θ, it is assumed that repeater delays have been adjusted so that the timing wave supplied to the regenerator in the absence of mistuning is properly aligned with the signal impulses in the information-bearing channel. In this case, the negative-going zero crossing occurring ideally at $t_o = T/4$ determines the instant of regeneration. When mistuning is present this zero crossing is displaced such that it occurs at the instant $t_o' = T(\frac{1}{4} - \theta/2\pi)$. The difference $t_o - t_o'$ will then give the timing deviation which, expressed as a fractional part of the pulse spacing, is

$$\frac{t_d}{T} = \frac{\theta}{2\pi}\,. \tag{6}$$

From (6) and the definition of θ, the phase error corresponding to the timing deviation is related to the random variables x and y by

$$\theta = \tan^{-1}\frac{y}{x}. \tag{7}$$

In deriving (7) it should be recalled that only the incidental approximation $k \ll 1$ has been made. When we consider a binary pulse train in which the pulses representing the binary "one" are of arbitrary pulse shape, it is necessary to make other approximations to arrive at a tractable expression for the phase error. Furthermore, the excitation encompasses the infinite past as well as the tails of succeeding pulses to accommodate driving pulses that may overlap or are not time limited. The most general result given by (59) is an extension along two lines of Rowe's relationship for the timing jitter in the output of the tuned circuit due to mistuning and finite pulse width. First, the results are applicable to arbitrary pulse shape. Secondly, our relationship for the phase error is based on a different approximation in the case of finite width pulses.

In appendix B we specialize to the case of raised cosine pulses in order to make use of some of Rowe's results. For this case the phase error is given by (73) and takes the form

$$\theta = \frac{y + a}{x + b} + c, \tag{8}$$

where a, b, and c are constants that depend upon Q, k, and the pulse width T/s of the raised cosine pulse. x and y are correlated random variables that depend upon Q, k, and the pulse pattern. They are defined below (5) with the additional constraint that $a_o = 1$ when we consider finite width pulse; i.e., a pulse definitely occurs at the origin. In our notation, a positive phase error corresponds to the zero crossing of interest occurring prior to the reference. The largest pulse width we consider is $1.5T$. This avoids the necessity of considering the effect of the presence or absence of a following pulse on the negative-going zero crossing of interest. Similarly, for positive-going zero crossings we do not have to use special methods for considering the occurrence or nonoccurrence of a preceding pulse. This is not a serious analytical restriction, since larger pulse widths can be handled by the machinery provided in Section A-4. As a practical matter in the design of a self-timed reconstructive repeater for operation in a long repeater chain, wider pulses would introduce intolerable phase jitter. In the following, we will also neglect the constant c in (8), since it is independent of pulse pattern and can in principle be compensated for in either the timing path or information-bearing path in a self-timed reconstructive repeater.

III. PROBABILITY DENSITY FOR THE PHASE ERROR

3.1 *Preliminaries*

From the above, the random variable of interest is

$$\theta = \frac{y + a}{x + b} = \frac{y_1}{x_1}. \tag{9}$$

To determine the probability density $p(\theta)$ or the cumulative distribution $F(\theta)$, we consider the joint probability density of the correlated random variables x_1 and y_1, $p(x_1, y_1)$. $F(\theta) = \Pr(y_1/x_1) \leqq \theta)$, which may be written

$$F(\theta) = \int_0^\infty dx_1 \int_{-\infty}^{\theta x_1} dy_1 p(x_1, y_1) + \int_{-\infty}^0 dx_1 \int_{\theta x_1}^\infty dy_1 p(x_1, y_1).$$

Differentiation of $F(\theta)$ with respect to θ plus rearrangement yields

$$p(\theta) = \int_0^\infty x_1 p(x_1, \theta x_1)\, dx_1 + \int_0^\infty x_1 p(-x_1, -\theta x_1)\, dx_1. \qquad (10)$$

Therefore if $p(x_1, y_1)$ is known, $p(\theta)$ can be determined by integration. As is typical of this class of problems when x_1 and y_1 are not correlated normal variables, the exact determination of $p(x_1, y_1)$ is rarely obtainable. Therefore, we find it essential to proceed along approximate lines.

We can write the characteristic function $\varphi(u,v)$ for $p(x_1, y_1)$ as

$$\varphi(u,v) = \int_{-\infty}^\infty dx_1 \int_{-\infty}^\infty dy_1 e^{i(ux_1 + vy_1)} p(x_1, y_1). \qquad (11)$$

If we take the partial derivative of (11) with respect to u, evaluate it at $u = -\theta v$, divide both sides by $2\pi i$, and integrate over v from $-\infty$ to ∞, we get

$$\frac{1}{2\pi i} \int_{-\infty}^\infty \left.\frac{\partial \varphi(u,v)}{\partial u}\right|_{u=-\theta v} dv = \frac{1}{2\pi} \int_{-\infty}^\infty dv \int_{-\infty}^\infty dx_1 \int_{-\infty}^\infty dy_1 x_1 e^{iv(y_1 - \theta x_1)} p(x_1, y_1).$$

When we interchange the order of integration to integrate over v first,

$$\frac{1}{2\pi i} \int_{-\infty}^\infty \left.\frac{\partial \varphi(u,v)}{\partial u}\right|_{u=-\theta v} dv = \int_{-\infty}^\infty dx_1 \int_{-\infty}^\infty dy_1 x_1 \delta(y_1 - \theta x_1) p(x_1 y_1),$$

where $\delta(y_1 - \theta x_1)$ is the Dirac delta function. Integration over y_1 then results in

$$\begin{aligned} \frac{1}{2\pi i} \int_{-\infty}^\infty \left.\frac{\partial \varphi(u,v)}{\partial u}\right|_{u=-\theta v} dv &= \int_{-\infty}^\infty x_1 p(x_1, \theta x_1)\, dx_1 \\ &= \int_0^\infty x_1 p(x_1, \theta x_1)\, dx_1 - \int_0^\infty x_1 p(-x_1, -\theta x_1)\, dx_1. \end{aligned} \qquad (12)$$

A comparison of (10) with (12) reveals that they are equal provided that x_1 is always positive, in which case $p(-x_1, -\theta x_1)$ is zero. Under this condition[9]

$$p(\theta) = \frac{1}{2\pi i} \int_{-\infty}^\infty \left.\frac{\partial \varphi(u,v)}{\partial u}\right|_{u=-\theta v} dv. \qquad (13)^*$$

In the following we will use (13) to approximate $p(\theta)$; before doing so we make a few remarks about the range of the random variables x_1 and θ.

* The result in (13) is given as an exercise for the reader on p. 317 of Ref. 9.

3.2 *Minimum Values of x_1 and y_1*

Our comments in this section will largely be confined to the case of impulse excitation in which case $x_1 = x$ and $y_1 = y$, where x and y are defined following (5). From the definition of x it can be seen that it attains its minimum value for the set of $a_n = 1$ in which the argument of $\cos 2\pi kn$ is in the second and third quadrants (modulo 2π). With this pulse pattern it is easily shown that

$$x_{\min} = -\frac{\beta \sin 2\pi k e^{-(\pi/4kQ)}(1 + e^{-(\pi/2kQ)})}{(1 - e^{-(\pi/2kQ)})(1 - 2\beta \cos 2\pi k + \beta^2)} = -2\bar{y}\,\frac{e^{-(\pi/4kQ)}}{(1 - e^{-(\pi/2kQ)})}$$

where $\beta = e^{-(\pi/Q)}$ and $\bar{y}$ = average value of y (from Appendix D). For the values of k and Q that we consider, namely kQ less than about 0.1 and $Q \geqq 100$, an excellent approximation for $x_{\min}$ is

$$x_{\min} = -2\bar{y}e^{-(\pi/4kQ)}.$$

When kQ is fixed at 0.1,

$$x_{\min} \doteq \frac{4kQ^2}{\pi}\, e^{-2.5\pi}$$

and for $Q = 100$, $x_{\min} = -0.005$. The ratio $x_{\min}/\bar{x}$, where $\bar{x}$ = average value of x, can be shown to be

$$\frac{x_{\min}}{\bar{x}} \doteq -4kQ\, e^{-(\pi/4kQ)},$$

which for $kQ = 0.1$ is -0.00016, or very close to zero. Based on unpublished work of one of the authors, the probability of $x/\bar{x}$ of even going negative is so remote as to be completely unimportant and decreases with increasing Q for kQ fixed.

Another interesting way of looking at the probability of x becoming negative is to consider the probability of pulses occurring in the first quadrant of the argument of $\cos 2\pi kn$ to constrain the minimum value of x to zero. This can occur in any of several ways. One possibility is to choose a single pulse (a single $a_n = 1$) in the sector of the first quadrant bounded by $n = 0$ and the largest integral value of n that satisfies

$$\beta^n \cos 2\pi kn > |\, x_{\min} \,|.$$

For $Q = 100$ and $kQ = 0.1$, the above is satisfied for a value of n that is less than about 148. The probability of at least one pulse in this range of n is $1 - (1 - p)^{148}$ which is about $1 - 10^{-18}$ for equally likely pulses and spaces. Therefore, x is positive with probability very close to unity.

For increasing values of Q, with kQ fixed at 0.1, the probability that x is >0 approaches unity even more closely.

By an argument that parallels the above, the probability that $y < 0$ for $k > 0$ and impulse excitation is very small. Similarly, probability $y > 0$ for $k < 0$ is extremely small.

For raised cosine excitation, $x_{\min}$ is increased by $1 + b$, which for the pulse widths considered herein is always >0.25, thereby making $x_{\min}$ positive for the Q's of interest to us. We also note that long strings of zeros as required in attaining $x_{\min}$ cannot be tolerated in a PCM repeater with a simple tuned circuit timing extractor, since the timing wave amplitude would fall well below the point at which it would be useful in the repeater. A higher minimum on the timing wave amplitude can be assured by constraining the transmitted pulse train to avoid such long strings of spaces.[7] In this paper we simulate this constraint by the introduction of a forced periodic pattern of pulses in the otherwise random train. This serves to increase $x_{\min}$ and decrease the range of θ as we shall see below and in Sections VII and VIII.

3.3 *Range of θ*

For random impulse excitation, it is apparent from (5) that θ is unbounded when we choose a single $a_n = 1$ for n large and all the rest zero. However, with $a_o = 1$ and the values of Q we consider, x is always positive, and from the results of Section 3.2 θ is essentially confined to $(0, \pi/2)$ for $k > 0$ and $[0, -(\pi/2)]$ for $k < 0$. In the following we seek tighter bounds under the practically important case $a_o = 1$. Experimentally, $a_o = 1$ means that we examine only those time slots containing pulses.

For the general form of θ, D. Slepian and E. N. Gilbert of Bell Telephone Laboratories* have developed an algorithm for determining the pattern that yields the maximum value of θ. Their result is particularly simple when $kQ \ll 1$; then we can approximate x by

$$1 + \sum_{1}^{\infty} a_n \, e^{-(\pi/Q) n}$$

and y by

$$2\pi k \sum_{1}^{\infty} a_n \, n e^{-(\pi/Q) n}.$$

Under this condition Gilbert and Slepian have shown that the pulse

* Private communication.

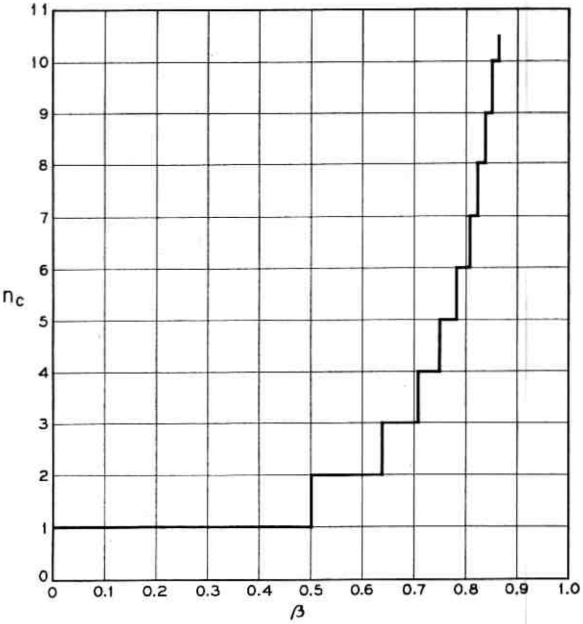


Fig. 1 — n_c vs β for random impulse excitation.

pattern giving the largest value of θ is specified by all pulses present for $n \geqq n_c$ and pulses absent for $n < n_c$. The value of n_c is obtained from*

$$\frac{\beta^{n+1}}{(1-\beta)^2} = n_c(1+b) - \frac{a}{2\pi k}. \tag{14}$$

where $\beta = e^{-(\pi/Q)}$. For random impulse excitation $a = 0 = b$. For this case, n_c versus β obtained from (14) is shown in Fig. 1. For $\beta < \frac{1}{2}$, all pulses present ($n_c = 1$) yields the maximum value for θ. In the range $\frac{1}{2} < \beta < 0.639$ the pulse immediately adjacent to the origin is dropped out to obtain $\theta_{\max}$ and so on.

The maximum value attained in a specified interval is achieved for the largest β in the interval and the maximum value is given simply by $2\pi k$ times the n_c defined by the β interval. The β intervals corresponding to constant n_c get smaller and smaller as β approaches one. This is illustrated in Fig. 2, where we have plotted n_c against Q rather than β, showing a continuous approximation to the actual staircase characteristic. We note that for $Q = 100$, $n_c = 80$ and $\theta_{\max} = 2\pi k n_c = 160\pi k$. With $k = 10^{-3}$, $\theta_{\max} = 0.16\pi$ radians.

* See Appendix E for the proof.

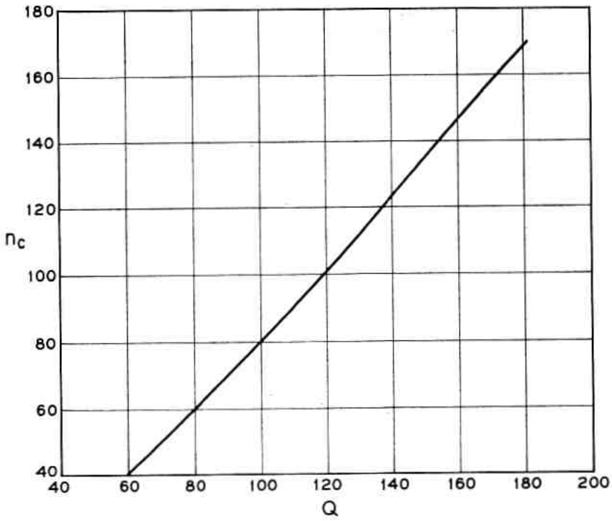


Fig. 2 — n_c vs Q.

For finite width pulses, a and b are non-zero. With raised cosine pulses of pulse width less than 1.5 time slots $a < 0.65$ and $b > -0.75$ with the largest negative value of b corresponding to the consideration of positive going time slots. When the mistuning, k, is positive, the effect of finite pulse width then is to raise the maximum value of n_c over the impulse case and consequently to raise $\theta_{\max}$. On the other hand, when $k < 0$, $\theta_{\max}$ can be reduced over the impulse case. We will demonstrate this effect in connection with the cumulative distribution in Section IX of the paper.

As noted previously, the long string of spaces implied by large n_c make the timing wave amplitude so small as to be useless in a real repeater. The timing wave amplitude can be increased by forcing a periodic pulse pattern. With the constraint that every Mth pulse must occur, the pattern that yields the maximum value for θ is as before where n_c is now given by

$$\frac{\beta^{n_c+1}}{(1-\beta)^2} = -\frac{a}{2\pi k} + n_c\left[1 + b + \beta^M \frac{(1-\beta^{rM})}{(1-\beta^M)}\right] + \frac{M\beta^M(1-\beta^{rM})}{2\pi k(1-\beta^M)^2} - \frac{rM\beta^{(r+1)M}}{2\pi k(1-\beta^M)}, \tag{15}$$

where r is the largest integer less than n_c/M. It can be seen that (15) reduces to (14) as $M \to \infty$ as expected. Furthermore, since the difference in the last two terms of (15) is positive and the term added to $1 + b$ is also positive, it is apparent that the effect of the periodic pattern is to reduce n_c and consequently $\theta_{\max}$ as expected.

3.4 *Probability Density Function, $p(\theta)$*

With the above preliminaries disposed of, we will proceed to use (13) to develop an approximate expression for $p(\theta)$. To do this we assume that the logarithm of the characteristic function possesses a power series expansion in the neighborhood of $u = 0 = v$. The general form of this series is[10]

$$\log \varphi(u,v) = \sum_{\substack{r=0 \\ r+s \neq 0}}^{\infty} \sum_{s=0}^{\infty} \frac{\lambda_{rs}}{r!s!} (iu)^r (iv)^s \tag{16}$$

where the λ_{rs} are the semi-invariants of the distribution for x_1 and y_1. Since

$$\frac{\partial \varphi}{\partial u} = \varphi \frac{\partial}{\partial u} [\log \varphi],$$

we may write

$$p(\theta) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{\partial}{\partial u} [\log \varphi] \bigg|_{u=-\theta v} \exp [\log \varphi] \bigg|_{u=-\theta v} dv. \tag{17}$$

Using (17) and performing the differentiation indicated in the integrand, we get

$$p(\theta) = \frac{d}{d\theta} \left[\frac{i}{2\pi} \int_{-\infty}^{\infty} \exp\left[\sum_{\substack{r=0 \\ r+s\neq 0}}^{\infty} \sum_{s=0}^{\infty} \frac{\lambda_{rs}}{r!s!} (-1)^r \theta^r (iv)^{r+s} \right] \frac{dv}{v} \right]. \tag{18}$$

We now remove terms from the double summation for which $r + s \leqq 2$. The remaining terms we treat as u, and expand e^u in a power series retaining only the first two terms ($e^u \sim 1 + u$). In this case $p(\theta)$ becomes approximately

$$p(\theta) \backsim p_o(\theta) + \sum_{\substack{r \\ r+s>2}}^{r+s=6} \sum_{s} \frac{\lambda_{rs}}{r!s!} (-1)^r p_{rs}(\theta), \tag{19}$$

where

$$p_o(\theta) = \frac{d}{d\theta} \left[\frac{i}{2\pi} \int_{-\infty}^{\infty} \frac{dv}{v} \exp -iv(\lambda_{10}\theta - \lambda_{01}) - \frac{v^2}{2} (\lambda_{20}\theta^2 - 2\lambda_{11}\theta + \lambda_{02}) \right],$$

or $p_o(\theta) = (d/d\theta)f_o(\theta)$, where $f_o(\theta)$ is defined by comparison with the above.

Similarly,

$$p_{rs}(\theta) = \frac{d}{d\theta}\left[\theta^r \frac{i}{2\pi}\int_{-\infty}^{\infty}\frac{dv}{v}\,(iv)^{r+s} \cdot \exp -iv(\lambda_{10}\theta - \lambda_{01}) - \frac{v^2}{2}\,(\lambda_{20}\theta^2 - 2\lambda_{11}\theta + \lambda_{02})\right],$$

or

$$p_{rs}(\theta) = \frac{d}{d\theta} f_{rs}(\theta)\,.$$

An upper limit for the double summation in (19) is set in order to make the approximation for $p(\theta)$ consistent with the number of terms used in the power series expansion for e^u. The reason for 6 as an upper limit will become apparent when we discuss the semi-invariants, λ_{rs}, in detail in Section V. Performing the differentiations and integrations indicated in (19) we finally arrive at

$$p(\theta) \sim \frac{1}{\sqrt{2\pi}}\frac{A_2(\theta)}{A_1(\theta)^{\frac{3}{2}}}\exp\left[-\frac{A_o(\theta)^2}{2A_1(\theta)}\right] \cdot\left\{1 + \sum_{\substack{r \\ r+s>2}}^{r+s=6}\sum_{s}(-1)^r\frac{\lambda_{rs}}{r!s!}\,\theta^r\left[\frac{H_{r+s}\left(\dfrac{A_o(\theta)}{\sqrt{2A_1(\theta)}}\right)}{(\sqrt{2A_1(\theta)})^{r+s}} + \frac{H_{(r+s)-1}\left(\dfrac{A_o(\theta)}{\sqrt{2A_1(\theta)}}\right)}{(\sqrt{2A_1(\theta)})^{(r+s)-1}}\cdot\frac{A_{rs}(\theta)}{\theta A_2(\theta)}\right]\right\}, \tag{20}$$

where

$$\begin{aligned} A_o(\theta) &= \lambda_{10}(\theta - \theta_o),\\ A_1(\theta) &= \lambda_{20}\theta^2 - 2\lambda_{11}\theta + \lambda_{02}\,,\\ A_2(\theta) &= \lambda_{10}[\theta(\lambda_{20}\theta_o - \lambda_{11}) - (\lambda_{11}\theta_o - \lambda_{02})],\\ A_{rs}(\theta) &= s\lambda_{20}\theta^2 + (r - s)\lambda_{11}\theta - r\lambda_{02}\,, \end{aligned}$$

and

$$\theta_o = \frac{\lambda_{01}}{\lambda_{10}}\,.$$

The H's are Hermite polynomials defined by

$$H_n(Z) = (-1)^n e^{Z^2} \frac{d^n}{dZ^n} (e^{-Z^2}).$$

The result in (20) gives a general expression for $p(\theta)$ as a function of the semi-invariants of the distribution of x_1 and y_1. The solution obtained is approximate in that it depends upon an asymptotic expansion analogous to the Edgeworth Series. As noted by Cramer,[9] one is not particularly interested in whether series of this type converge or not, but whether a small number of terms suffice to give a good approximation to the probability density function over a specified range of its argument. In our case, the statistical properties of the input pulse pattern, and the parameters of the timing circuit are controlling in this regard. With this in mind, the determination of the range in θ over which a valid approximation may be obtained in various cases is deferred for the present.

IV. CUMULATIVE DISTRIBUTION FUNCTION

The cumulative distribution function $F(\theta)$ may be determined using the results derived in the preceding section. Beginning with (19) we may write

$$p(\theta) \sim f_o'(\theta) + \sum_{\substack{r \\ r+k>2}}^{r+k=6} \sum_{k} \frac{\lambda_{rk}}{r!k!} (-1)^r f_{rk}'(\theta). \tag{21}$$

By definition*

$$F(\theta) = \int_{-\infty}^{\theta} p(u)\, du.$$

Integrating (21) between the limits indicated, $F(\theta)$ becomes

$$F(\theta) \doteq f_o(\theta) + \sum_{\substack{r \\ r+s>2}}^{r+s=6} \sum_{s} \frac{\lambda_{rs}}{r!s!} (-1)^r f_{rs}(\theta) + \frac{1}{2}. \tag{22}$$

Referring back to (19) and performing the integration over v necessary to determine $f_o(\theta)$ and $f_{rk}(\theta)$, we get

* The significance of the lower limit of integration in the definition of $F(\theta)$ will be discussed in connection with the numerical results.

$$F(\theta) \sim \frac{1}{2} + \frac{erf}{2}\left[\frac{A_o(\theta)}{\sqrt{2A_1(\theta)}}\right] - \frac{1}{\sqrt{2\pi A_1(\theta)}}$$

$$\cdot \exp\left[-\frac{A_o(\theta)^2}{2A_1(\theta)}\right] \sum_{\substack{r \\ r+s>2}}^{r+s=6} \sum_{s} \frac{\lambda_{rs}}{r!s!}(-1)^r\theta^r \cdot \frac{H_{(r+s)-1}\left(\dfrac{A_o(\theta)}{\sqrt{2A_1(\theta)}}\right)}{(\sqrt{2A_1(\theta)})^{(r+s)-1}}, \tag{23}$$

where $A_o(\theta)$, $A_1(\theta)$ and H_{r+s-1} have been previously defined.

V. SEMI-INVARIANTS FOR THE DISTRIBUTION OF x AND y

In this section we consider the coefficients of the power series expansion for the logarithm of the characteristic function $\varphi(u,v)$. These are determined as functions of the parameters of the timing circuit, and the excitation and provide the necessary information for an explicit solution for $p(\theta)$ and $F(\theta)$. A closed form for the λ_{rs} is obtainable for all excitations of interest under the condition $p = \frac{1}{2}$ (pulses and spaces equally likely). [The semi-invariants for any p can be obtained by appropriate differentiations of $\log \varphi(u,v)$. We have not expended the energy for this exercise.] The semi-invariants are shown below for random impulse excitation under the condition $kQ \ll \pi$ and are derived for all excitations we consider in Appendix D.*

$$\lambda_{10} = \frac{1}{2(1-\beta)} \qquad \lambda_{01} = \frac{\pi k \beta}{(1-\beta)^2} \tag{24}$$

$$\lambda_{rs}\,|_{r+s>1} = \frac{(-1)^s B_{r+s}(2^{r+s}-1)}{} \cdot (2\pi k)^s \frac{d^s}{dg^s}\left(\frac{1}{1-e^{-g}}\right) \tag{25}$$

where $\beta = e^{-(\pi/Q)}$, $g = \pi/Q\,(r+s)$, and the B_{r+s} are Bernoulli numbers. Since $B_{r+s} = 0$ for $r + s$ odd and >1, we note that the odd order semi-invariants given in (24) and (25) vanish beyond order 1. Therefore since the λ_{rs} for $r + s = 3$ are zero, one can extend the upper limit in the double summation in (19) to 6, and still maintain consistency with the fact that only 2 terms in the power series expansion for the exponential, e^u, were used in the approximation for $p(\theta)$. This conclusion is valid for all excitations of interest.

VI. BEHAVIOR OF $p(\theta)$ FOR LARGE Q

When the Q of the resonant circuit becomes large, the past history of the input signal becomes increasingly important in determining the

* The more general semi-invariants without the restriction $kQ \ll \pi$ are given in Appendix D; however, they are too long to be repeated here.

statistical properties of x and y. This follows from the form of the exponential term in the expressions for x and y given in (5). Invoking the Central Limit Theorem under this condition, one would expect the values of x and y to begin heaping up about their respective means with the probability density function $p(x,y)$ approaching a two dimensional normal distribution. Analogous behavior is expected of θ and we will now consider $p(\theta)$ as given by (20) in the neighborhood of its mean for large Q. The discussion is restricted to the case of random impulse excitation, but the results for other excitations parallel those of this section.

To determine $p(\theta)$ near its mean, we write, using the previous condition $kQ \ll \pi$,

$$\theta \doteq \frac{y}{x} \doteq \frac{2\pi k \sum_{n=0}^{\infty} a_n n\, e^{-\alpha n}}{\sum_{n=0}^{\infty} a_n\, e^{-\alpha n}}, \tag{26}$$

where

$$\alpha = \frac{\pi}{Q}.$$

For this to hold as Q becomes arbitrarily large, we require the kQ product to be constant. Since

$$x \sim \sum_{n=0}^{\infty} a_n\, e^{-\alpha n},$$

θ can also be written as

$$\theta \sim -2\pi k \frac{d}{d\alpha}[\log x] = -2\pi k \frac{d}{d\alpha}\left[\log \frac{x}{\bar{x}} + \log \bar{x}\right], \tag{27}$$

where $\bar{x}$ is the average value of x. Expanding $\log x/\bar{x}$ in a power series in the neighborhood of 1 (x near $\bar{x}$), and keeping only the first term, θ becomes

$$\theta \sim -2\pi k \frac{d}{d\alpha}[\log \bar{x}] - 2\pi k \frac{d}{d\alpha}\left[\frac{x - \bar{x}}{\bar{x}}\right]. \tag{28}$$

Differentiating the above with respect to α we get for θ in the neighborhood of its mean

$$\theta \sim \frac{\bar{y}}{\bar{x}} + \frac{\bar{x}y - x\bar{y}}{\bar{x}^2}. \tag{29}$$

In determining this result we make use of the fact that

$$\bar{y} = -2\pi k \frac{d}{d\alpha} [\bar{x}]. \tag{30}$$

Using (29) one can determine the logarithm of the characteristic function of θ, and the associated semi-invariants of the θ distribution. When this is done, the mean of θ is

$$\bar{\theta} \sim \frac{\bar{y}}{\bar{x}} = \theta_o = \frac{2\pi k\beta}{1 - \beta} \tag{31}$$

which also can be derived directly from (29). The standard deviation and the 4th semi-invariant are given by

$$\sigma = \sqrt{\frac{2(2\pi k)^2\beta^2}{(1 - \beta)(1 + \beta)^3}}$$

$$\begin{aligned}\lambda_4 = \frac{-2(2\pi k)^4\beta^4}{(1 - \beta^4)} \Bigg[1 &- \frac{4\beta^3(1 - \beta)}{(1 - \beta^4)} + \frac{6\beta^2(1 + \beta^4)(1 - \beta)^2}{(1 - \beta^4)^2} \\ &- \frac{4\beta(1 - \beta)^3(1 + 4\beta^4 + \beta^8)}{(1 - \beta^4)^3} \\ &+ \frac{(1 - \beta)^4(1 + 11\beta^4 + 11\beta^3 + \beta^{12})}{(1 - \beta^4)^4} \Bigg]\end{aligned} \tag{32}$$

with $\beta = e^{-\alpha}$. These same results can be derived using (20) and including only the first correction term from the double sum (i.e., only those λ_{rs} for which $r + s = 4$). The details of the calculation along with the λ_{rs} of interest are given in Appendix D. The final result for $p(\theta)$ is

$$p(\theta) \sim \frac{1}{\sqrt{2\pi}\,\sigma} \exp - \frac{(\theta - \theta_o)^2}{2\sigma^2} \left(1 + \frac{\lambda_4}{4!} \frac{H_4 \dfrac{\theta - \theta_o}{\sqrt{2}\,\sigma}}{4\sigma^4} \right). \tag{33}$$

The above equation for $p(\theta)$ is in the form of the standard Edgeworth approximation. In the limit as Q becomes large ($\beta \to 1$), and with kQ constant, $p(\theta)$ reduces to

$$p(\theta) \sim \frac{1}{\sqrt{2\pi\sigma}} \exp - \frac{(\theta - \theta_o)^2}{2\sigma^2} \left[1 - \frac{5\pi}{128Q} H_4 \left(\frac{\theta - \theta_o}{\sqrt{2}\,\sigma} \right) \right] \tag{34}$$

with $\theta_o \doteq 2kQ$ and $\sigma \doteq k\sqrt{\pi Q}$. Equation (26) indicates the approach to the normal law as Q becomes large with the first correction term going as $1/Q$. The above results for θ_o and σ correspond to those derived earlier by Bennett[1] by another method. If we rewrite σ as $kQ\sqrt{\pi/Q}$ we notice that $p(\theta)$ becomes more peaked with increasing Q, and falls off quite rapidly as θ departs from the mean. In the high Q case the concentration about θ_o becomes more pronounced as expected.

It is to be emphasized that the general properties of $p(\theta)$ for large Q demonstrated here will be true for the other inputs also. For example, with random impulse excitation plus 1 out of M pulses forced, the average value will remain the same as above but σ will be a function of M;

$$\sigma \doteq kQ\sqrt{\frac{\pi}{Q}\frac{M(M-1)}{(M+1)^2}} \qquad \text{for } \frac{Q}{M\pi} \gg 1.$$

The effect of M is to reduce σ and therefore increase the concentration about the mean. As M becomes large (fewer pulses required to occur), the effect of M becomes insignificant for this large Q case.

VII. NUMERICAL RESULTS FOR $p(\theta)$ AND $1 - F(\theta)$: IMPULSE EXCITATION

7.1 $p(\theta)$

To determine the behavior of the probability density function for finite Q, we must use the general form of the approximation to $p(\theta)$ given by (20), since most of the approximations made in the previous section for Q arbitrarily large are no longer valid. By way of illustration we consider the case $Q = 100$, $k = 10^{-3}$ with impulse excitation and all pulses random ($p = \frac{1}{2}$). For negative mistuning, $k = -10^{-3}$, the curve for $p(\theta)$ will be identical with that for k positive except that θ is replaced with $-\theta$. The result for the probability density function is shown in Fig. 3. The calculations* upon which this curve is based include the first and second correction terms of (20); i.e., terms for which $r + s = 4$ and $r + s = 6$. Points beyond $\theta = 0.13$ radians on the lower end and $\theta = 0.35$ radians on the upper end are not included, since the approximation begins to fail at these extremes. More specifically, the probability density obtained from (20) goes negative somewhere between $\theta = 0.13$ radians and $\theta = 0.12$ radians and $\theta = 0.35$ and $\theta = 0.36$ radians. However, as we shall see later, up to these points the results for the cumulative distribution are in good agreement with computer simulation. The cumulative distribution is also shown on Fig. 3 to point out the fact that the median occurs slightly below the approximate mean given by $2kQ$. In addition, it is apparent from the shape of $p(\theta)$ and

* Equation (20) and all subsequent calculations for $p(\theta)$ and $F(\theta)$ were programmed for the IBM 7090 computer by Miss E. G. Cheatham.

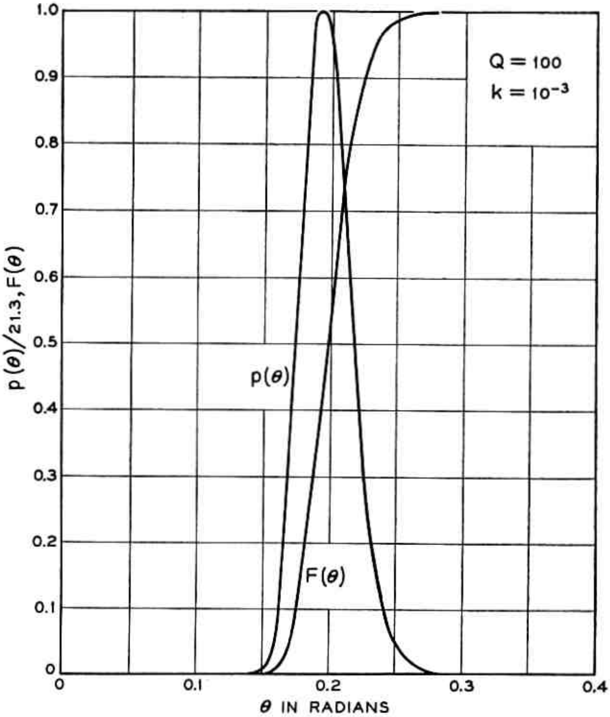


Fig. 3 — $p(\theta)$ and $F(\theta)$ as a function of θ for $k = 10^{-3}$ and $Q = 100$. Random impulse excitation.

$F(\theta)$ that the probability density is skewed in the direction of increasing phase error. This is more easily visualized from Fig. 4 where we have shown $p(\theta)$ as in Fig. 3 plotted on log paper. The normal probability density with the same mean and variance as our computed curve is also shown to further illustrate the skewness.

On Fig. 5 we have plotted $p(\theta)$, as defined in (20), to illustrate the contribution of its constituent terms. From this figure we see that the principal term (always positive) predominates over most of the range. At the tails, the terms involving λ_{rs} for $r + s = 4$ pulls $p(\theta)$ in and forces the density to become negative. The last term in the approximation, for which $r + s = 6$, serves to extend the region over which $p(\theta)$ remains positive.

When $1/M$ pulses are forced, the skewness is reduced, as is the variance. There are several ways of explaining this effect. First, as discussed in Section 3, the denominator of θ in (8) or (9) is raised, thereby reducing

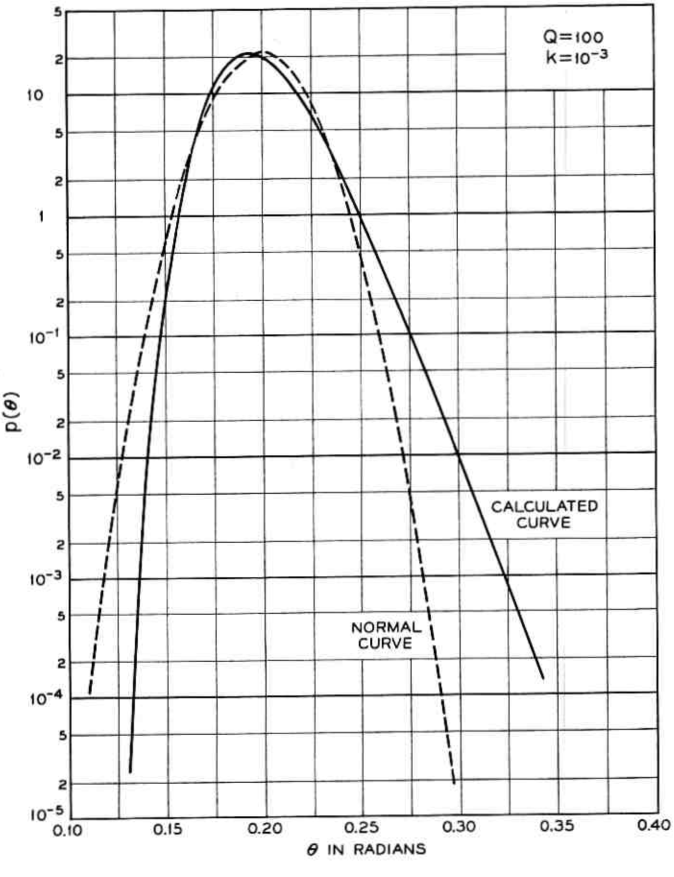


Fig. 4 — $p(\theta)$ for $k = 10^{-3}$ and $Q = 100$. The normal curve with the same mean and variance is also shown for comparison. Random impulse excitation.

the range of variation of the timing wave amplitude and confining θ to a narrower range. This is expected from the physical standpoint, since forcing a periodic pattern with the remaining pulses and spaces equally likely is similar to increasing the probability of occurrence of a pulse in an all-random sequence. Since the pulses, when they occur, have the proper spacing, they will tend to correct for the departure of the zero crossings from the mean that has occurred during the free response of the tuned circuit in the absence of a pulse. Indeed, in the limit when $M = 1$ (all pulses definitely occur), all the probability is concentrated at the mean, $2kQ$, which is identical to the steady state phase shift of the tuned circuit in response to a sine wave at the pulse repetition frequency. This behavior is also predicted mathematically from (20) and the fact that λ_{rs} goes to zero for $r + s > 1$ when $M = 1$. The same effect occurs when Q approaches infinity with kQ constant and it can be shown from the results of the previous section that $p(\theta)$ goes to $\delta(\theta)$ when the limit is taken. In this light, we can view the introduction of forced pulses as effectively increasing the Q of the tuned circuit while maintaining kQ fixed.

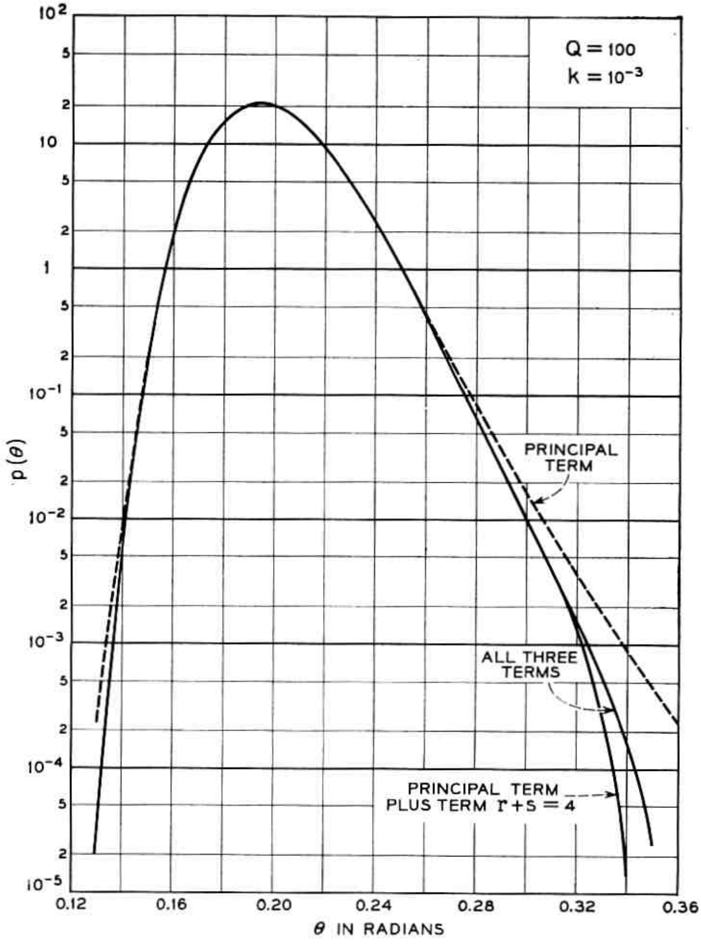


Fig. 5 — Contributions of various terms involved in the $p(\theta)$ approximation given by (20). Random impulse excitation is assumed, with $k = 10^{-3}$ and $Q = 100$.

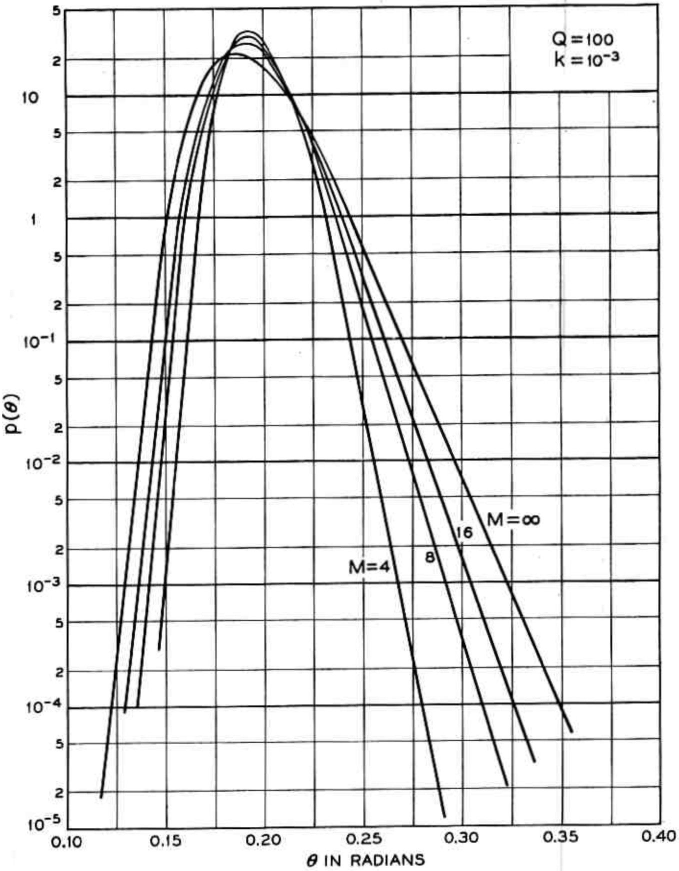


Fig. 6 — The effect on $p(\theta)$ of requiring $1/M$ impulses to occur. $k = 10^{-3}$, $Q = 100$.

In practical applications, the effect of a pulse at the origin is of particular interest. Mathematically, this corresponds to $M = \infty$. Physically this means we examine and record phase error only for those time slots containing a pulse. Fig. 6 illustrates the narrowing of the density function for $M = \infty$ (pulse at the origin), and $M = 16$, 8, and 4. It is interesting to note that, for these cases, the probability density function remains positive over the range of θ we have used in the computations from 0.1 to 0.4 radian. This encompasses values of $p(\theta) < 10^{-7}$ on the left of the mean and $p(\theta) < 10^{-5}$ to the right of the mean. This is to be expected since λ_{rs} decrease with decreasing M for $r + s \geqq 2$, thereby reducing the importance of the terms involving the Hermite polynomials in (20) and improving the approximation.

Fig. 7 depicts the behavior of $p(\theta)$ as Q grows with kQ fixed at 0.1. The results are consistent with the predictions of the previous section.

7.2 $1 - F(\theta)$

For a closer inspection of the behavior of the distribution at its tails, $1 - F(\theta)$ will be examined. This function as evaluated from (23) for

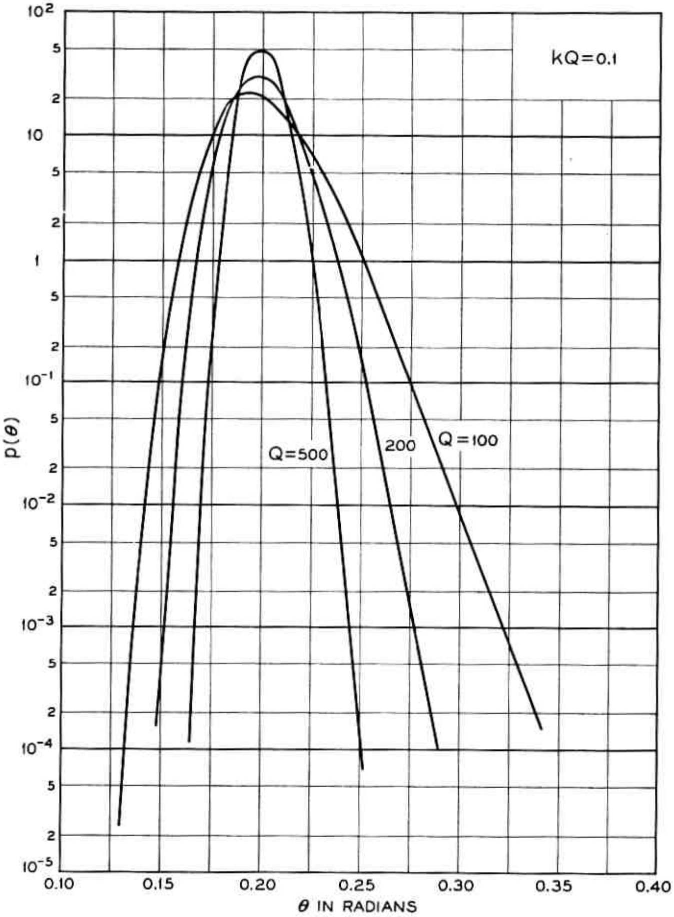


Fig. 7 — The effect on $p(\theta)$ of increasing Q with $kQ = 0.1$ and random impulse excitation.

$Q = 100$, $k = 10^{-3}$, and purely random excitation ($p = \frac{1}{2}$) is shown in Fig. 8. The plot shown gives the probability that θ deviates from its mean by more than some constant C times σ. In the same figure a comparison of the calculated approximation with the normal curve of identical mean and standard deviation indicates a substantial departure from the normal law as the phase error increases. When periodic patterns are interspersed with the random train, the departure from the mean is further reduced, as can be seen from Fig. 9. Similar behavior is exhibited in Fig. 10, where Q is increased from 100 to 500 and kQ maintained constant at 0.1.

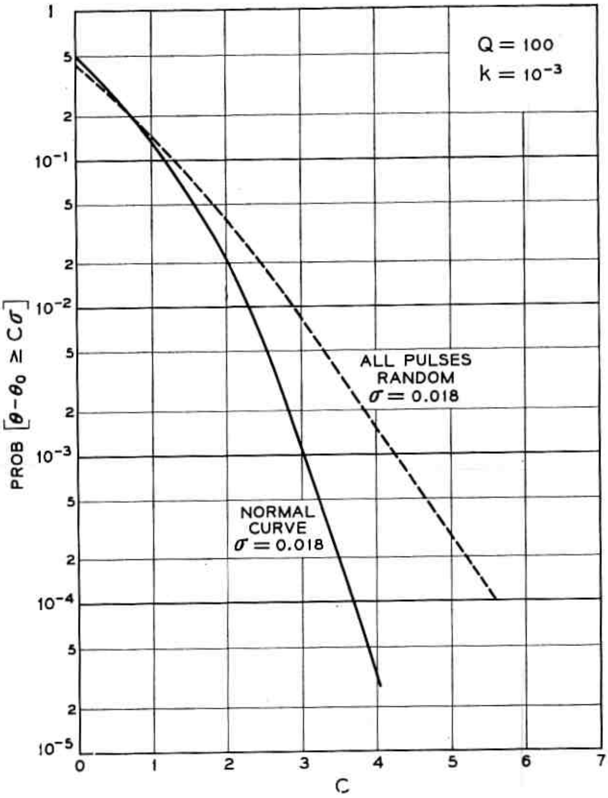


Fig. 8 — Comparison of $1 - F(\theta)$ with the normal curve in the vicinity of the tails. The normal curve is computed assuming the same mean and variance used in determining $1 - F(\theta)$. Random impulse excitation with $Q = 100$ and $k = 10^{-3}$ is assumed for computing $1 - F(\theta)$.

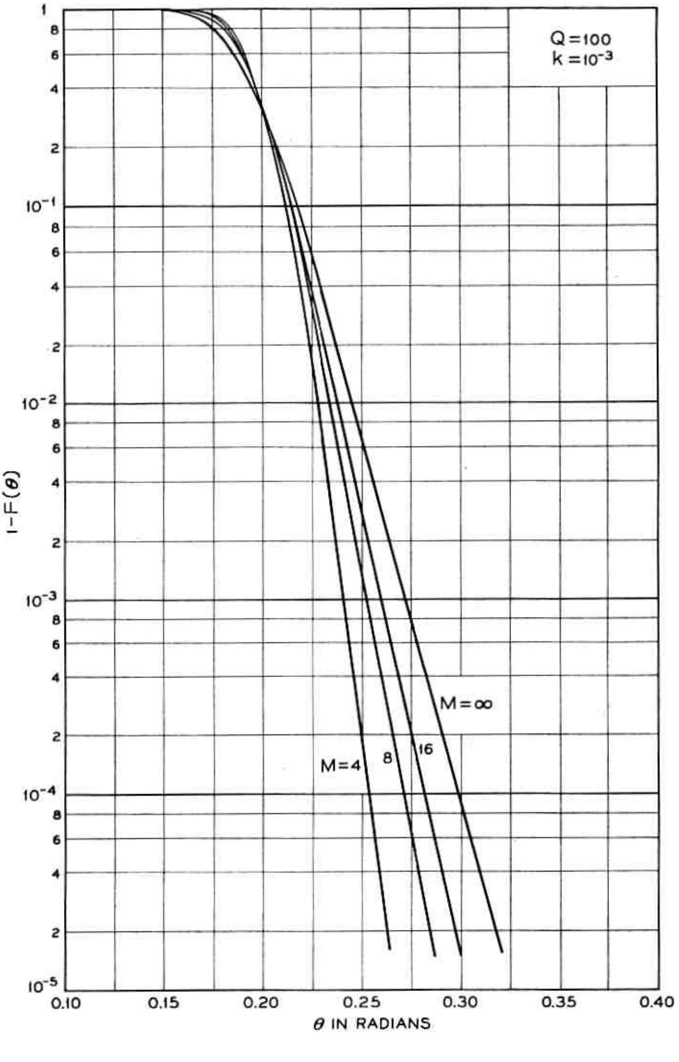


Fig. 9 — The effect on $1 - F(\theta)$ of requiring $1/M$ impulses to occur. $k = 10^{-3}$, $Q = 100$.

7.3 *Comparison with other approaches*

Since we have made approximations in arriving at our expression for the phase error, it is natural to ask how these approximations affect our computed results. A comparison of our results with two other approaches

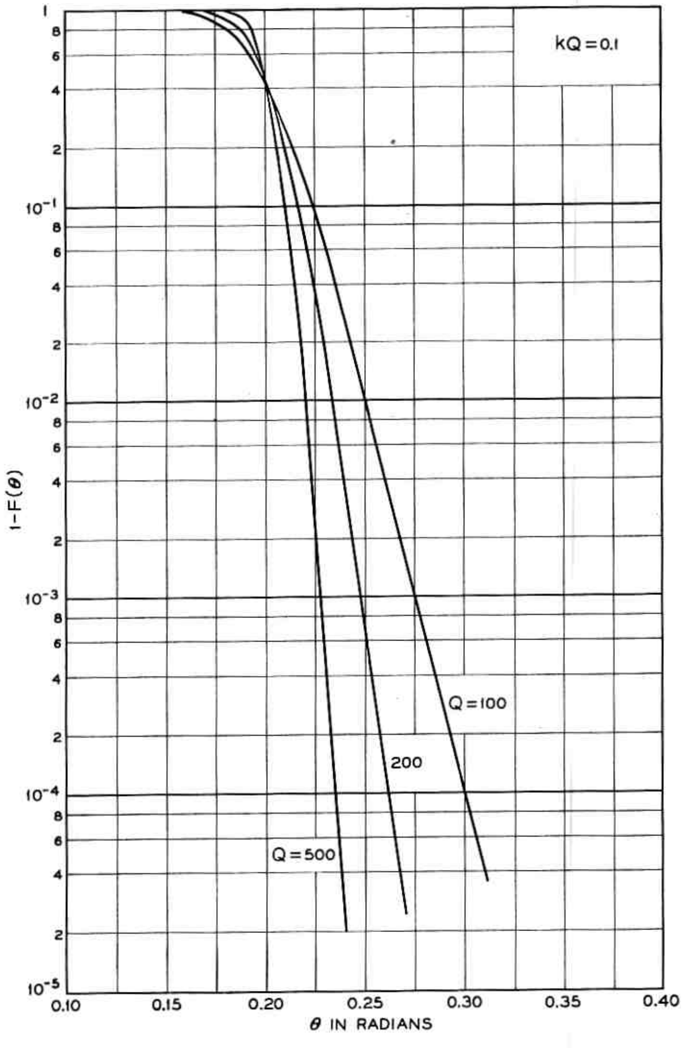


Fig. 10 — The effect on $1 - F(\theta)$ of increasing Q with $kQ = 0.1$ and random impulse excitation.

will be made for the case of impulse excitation. We recall from Section 2 that the phase error under impulse excitation is given by

$$\tan \theta = \frac{y}{x}.$$

For kQ sufficiently small we can write

$$\frac{\theta}{2\pi k} \doteq \frac{\sum_{n=0}^{\infty} na_n\beta^n}{\sum_{n=0}^{\infty} a_n\beta^n}. \tag{35}$$

The approximation of tan θ by its argument is not crucial in this case, since a straightforward transformation can be made on the probability distribution to correct for this approximation [i.e., $p(\theta) = \sec^2 \theta p(\tan \theta)$].

H. Martens* shows that (35) can be manipulated to yield a recursion relationship for the phase error that is in a convenient form for digital computer evaluation. T. V. Crater and S. O. Rice used this approach in some of their work, and a probability distribution so determined is shown by the dots in Fig. 11 for $Q = 125$. For the same value of Q, we have computed the probability distribution from the series in (23), and it is displayed as the solid curve of Fig. 11. It can be seen that the agreement between the two approaches is excellent. The scattering of the "experimental" points at the 10^{-3} level and below is due to the limited number of pulse positions considered by Crater and Rice. Specifically, 10^4 pulse positions were processed after an initial transient of some 5×10^3 pulse positions had elapsed.

In addition, S. O. Rice in unpublished work has shown that the tail of the distribution should behave as $A(\frac{1}{2})^{\theta/2\pi k}$, where A is an unknown constant. When we take the values of θ at the 10^{-3} and 10^{-4} levels and substitute these in Rice's asymptotic form and form a ratio, the constant A cancels out and we should obtain 10. The actual value for the ratio is 10.9, which tends to indicate that the asymptotic behavior has virtually been reached. This suggests that an extrapolation of the distribution to larger values of θ by merely continuing with the same slope should be valid.

We also note that we can write

$$(\tfrac{1}{2})^{\theta/2\pi k} = (\tfrac{1}{2})^{\theta Q/\pi\theta_o}$$

where we have made use of $\theta_o = 2kQ$. With kQ constant, one would expect the cumulative probability to fall off faster for larger Q, as is indeed the case. The slopes of the curves of Fig. 10 follow Rice's predictions quite closely.

While the above comparisons are comforting, they only indicate that our final expressions for $p(\theta)$ and $F(\theta)$ are accurate for computing these quantities from the initial defining equation for θ. Approximations have been made in arriving at the starting relationship. A check on these initial approximations may be obtained from a simulation of the problem.

* Unpublished memorandum.

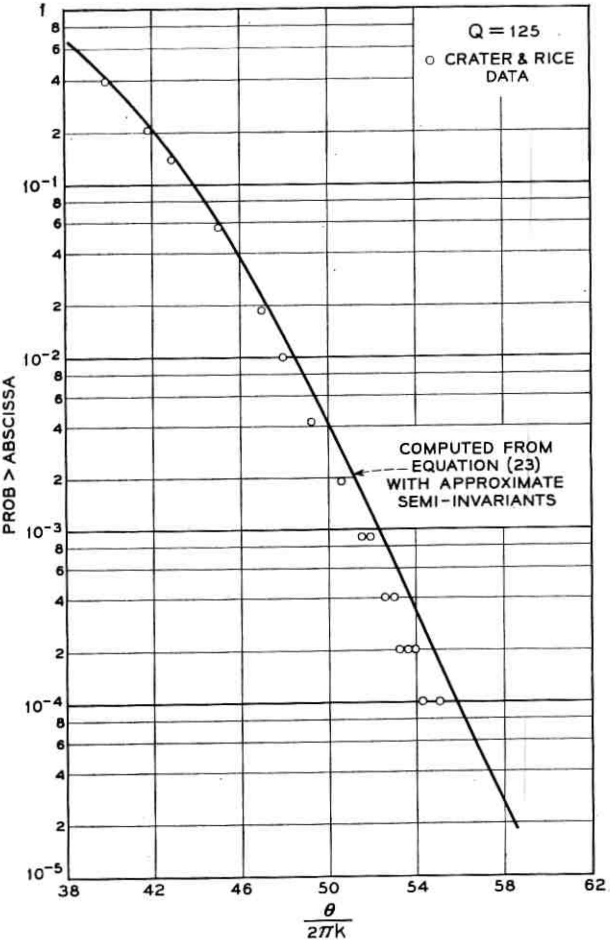


Fig. 11 — Comparison of $1 - F(\theta)$ computed by (23) with the results of the Crater-Rice simulation for $Q = 125$. Random impulse excitation is assumed.

One such simulation has been accomplished by Miss M. R. Branower using a combination of analogue and digital computers. The principal errors introduced in this process involve the stability of the analogue computer with time and the number of pulses processed. For a tuned circuit characterized by a Q of 125 and mistuning $k = +10^{-3}$, the computer simulation yields the results of Fig. 12. Results obtained using (23), the exact semi-invariants of Appendix C, and the tan θ transformation mentioned previously yield the "computed curve" of Fig. 12.

Again the results are in very close agreement. To indicate the effect of the approximation $kQ \ll \pi$, we have repeated the computed curve of Fig. 11 on Fig. 12.

VIII. RAISED COSINE EXCITATION

8.1 *Results for 1 − F(θ)*

With raised cosine excitation, the computations are performed as before and only the semi-invariants λ_{rs} for $r + s = 1$ are changed from the

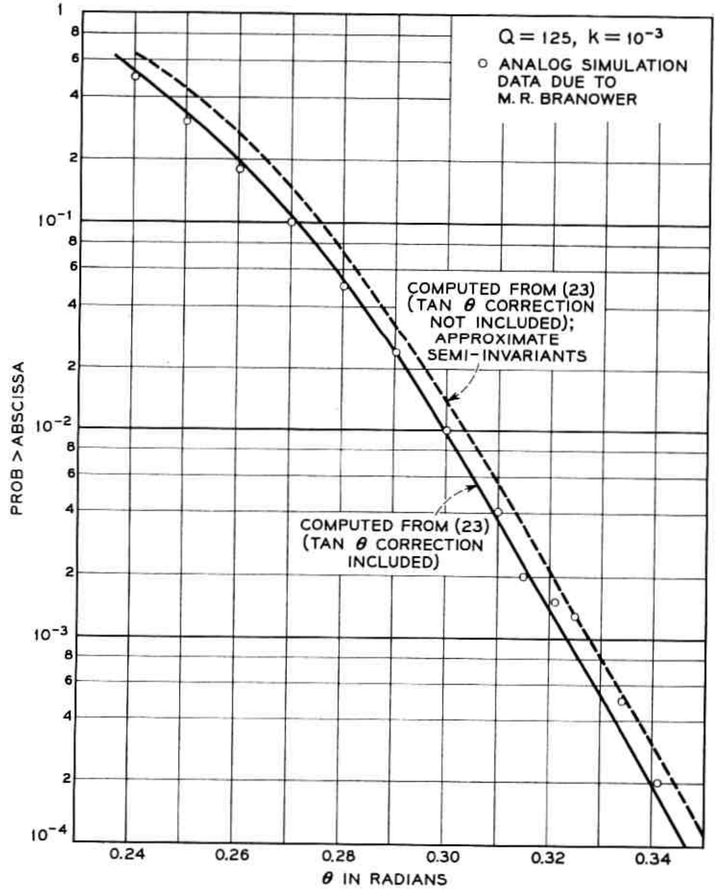


Fig. 12 — Comparison of $1 - F(\theta)$ computed by (23) with the results of an analog simulation due to M. R. Branower. Random impulse excitation with $Q = 125$ and $k = 10^{-3}$ is assumed. The effect of the tan θ approximation is shown together with results for both approximate and exact semi-variants.

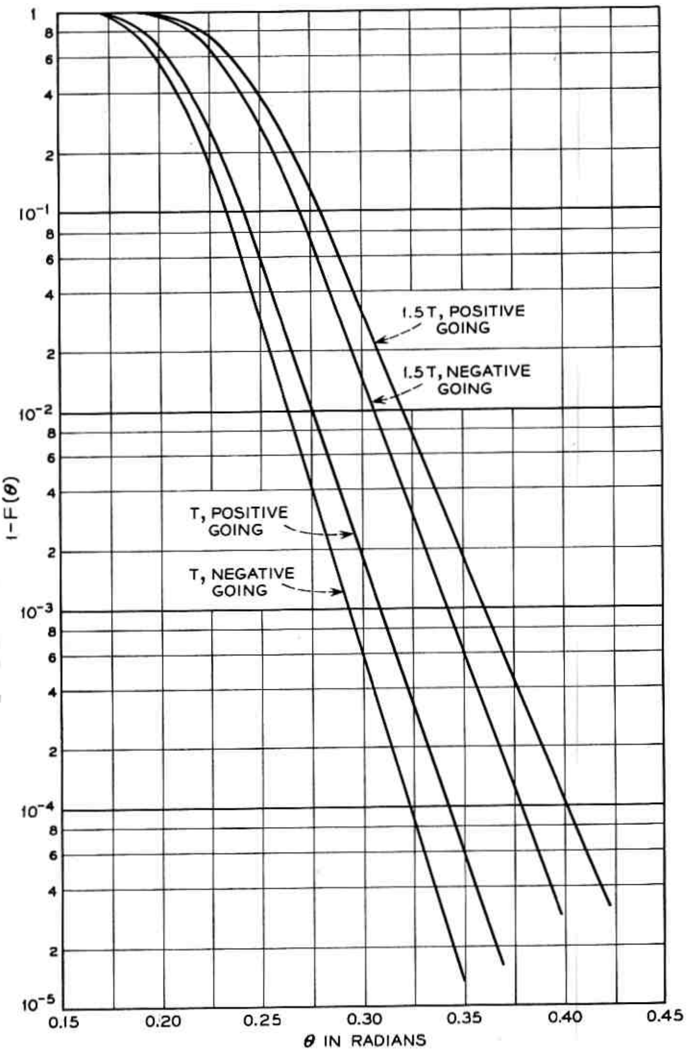


Fig. 13 — Plot of $1 - F(\theta)$ for raised cosine excitation. Pulses of width T and $1.5T$ are assumed in the calculation. The distribution of the phase error for both positive and negative-going zero crossings is shown. $Q = 100$, $k = 10^{-3}$.

previous case. Results obtained for this excitation are shown on Fig. 13, where it is apparent that the use of widest pulses and positive-going zero crossings yields the largest phase error. The effect of Q and M with this type of input is the same as with impulses.

8.2 *Comparison with another approach when $k = 0$*

In the absence of mistuning, the phase error becomes

$$\theta = \frac{a}{x + b}, \tag{36}$$

and the probability distribution for θ may be obtained by methods given previously, or by the following relationship:

$$\begin{aligned} \text{Prob } (\theta \geqq \lambda) &= \text{Prob } \left(\frac{a}{x + b} \geqq \lambda \right) \\ &= \text{Prob } \left(x \leqq \frac{a - b\lambda}{\lambda} \right). \end{aligned} \tag{37}$$

Therefore, if the distribution for x is known, the distribution for θ may be determined from it. The random variable x is the normalized timing wave amplitude defined by Rowe. This random variable has been considered by S. O. Rice in unpublished work and he has developed a procedure for closely approximating its probability distribution. Using the method of moments, one of the authors also computed this distribution. The results were in excellent agreement with Rice's results and the cumulative distribution obtained by the moment method is shown in Fig. 14. It can be shown that the probability density for x is unimodal and symmetric about its mean; therefore, the data on Fig. 14 suffices to specify the complete distribution. With this data and (37) we can determine the distribution for θ. Alternately, we can use (23) to make this computation. A comparison of the distribution obtained by the two approaches is shown in Fig. 15 and it can be seen that the agreement is very close. Thus we have found another check on our series approximation for $p(\theta)$. Conversely, we can use the distribution for θ to compute the distribution for x. In this regard it is interesting to note that when the Edgeworth expansion including semi-invariants through order 6 is used to approximate the distribution for x, the density function begins to turn negative in the neighborhood of 3σ from the mean indicating failure of the approximation. On the other hand, using the same number of semi-invariants in the expansion for $p(\theta)$, where θ in this case is essentially the reciprocal of x, we obtain a good approximation to the cumulative distribution for x. This is believed to be due to the narrowness of the range of θ as compared with x; i.e., x varies from 1 to $1/(1 - \beta) = Q/\pi$, while $1/x$ goes from $1 - \beta \doteq \pi/Q$ to 1.

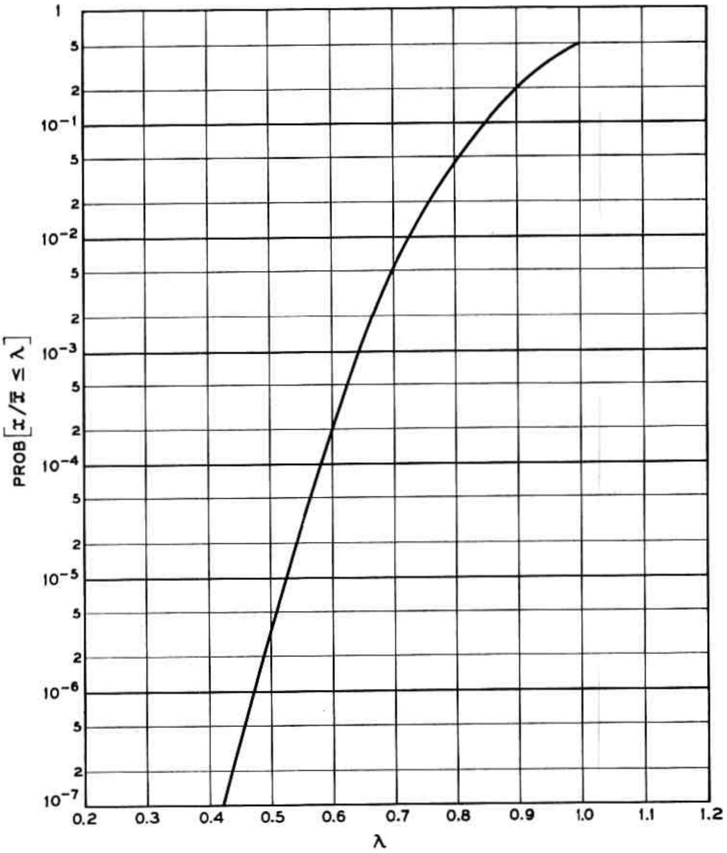


Fig. 14 — Probability distribution of the timing wave amplitude. $Q = 100$.

IX. OPTIMUM TUNING — FINITE PULSE WIDTH

In the case of impulse excitation it should be apparent that zero mistuning, $k = 0$, is the desired objective for no phase error. On the other hand, with finite width pulses zero mistuning does not yield zero phase error. Mistuning can be purposely introduced in the finite pulse width case to make the mean value of θ zero, to minimize the variance of θ, or to optimize some other parameter of the θ distribution.

An approximation to making the mean of θ zero may be obtained by choosing k such that the average value of the numerator of θ is zero. This means that

$$\bar{y}_1 = a + \bar{y} \doteq a + \frac{\pi k \beta}{(1 - \beta)^2} = 0, \tag{38}$$

or

$$k = -\frac{a(1 - \beta)^2}{\pi \beta}. \tag{39}$$

For example, when $Q = 100$ and $a = 0.65$, as for raised cosine pulses of width $1.5T$, then $k = -2.05 \times 10^{-4}$ to satisfy (39). In the high Q case (39) becomes $k \doteq -(a\pi/Q^2)$.

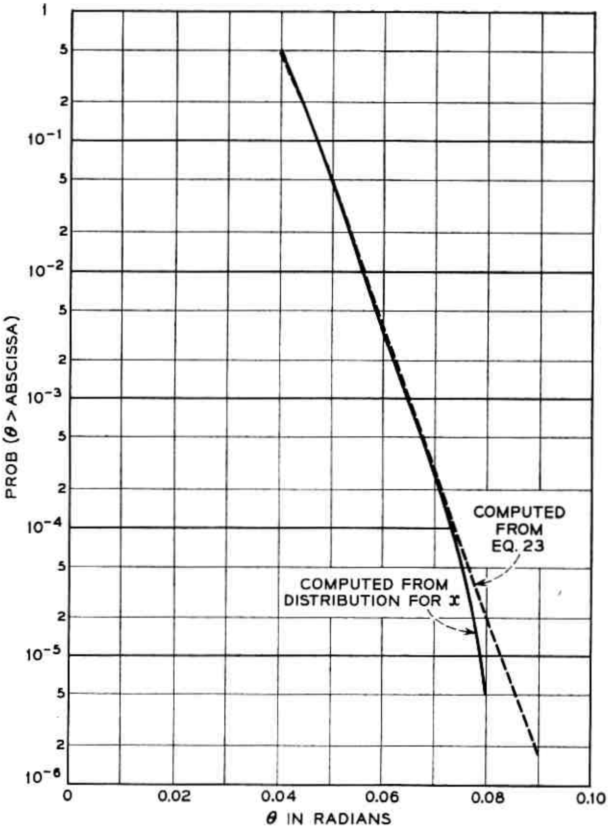


Fig. 15 — Comparison of the distribution of θ as computed by (23) and that determined from the distribution of the timing wave amplitude of Fig. 14. Raised cosine pulses of width 1.5T drive a tuned circuit with a $Q = 100$ and zero mistuning. Timing deviations in the neighborhood of negative-going zero crossings are considered.

When the objective is to minimize the variance of θ, we consider σ as defined in Appendix D; i.e.

$$\sigma = \left| \frac{(\lambda_{20}\theta_o^2 - 2\lambda_{11}\theta_o + \lambda_{02})^{1/2}}{\lambda_{10}} \right|. \tag{40}$$

A plot of σ versus k is shown in Fig. 16, where it is seen that the minimum σ occurs close to the "zero mean" value of k. Probability distributions for values of k that encompass the optimum are shown on Fig. 17. The narrowing of the density function for the optimum value of k is evident.

The results of this section suggest that when the tuned circuit in a self-timed repeater is adjusted, it should be excited with a random pulse train and the tuning adjusted to minimize the jitter on the leading edge

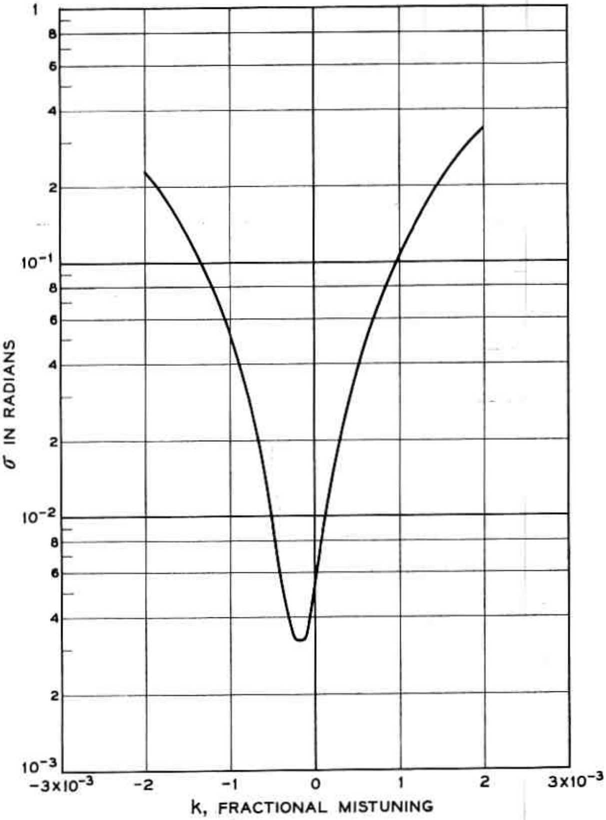


Fig. 16 — Standard deviation of phase error as a function of mistuning with raised cosine pulses 1.5T wide. Negative-going zero crossings are considered. Q = 100.

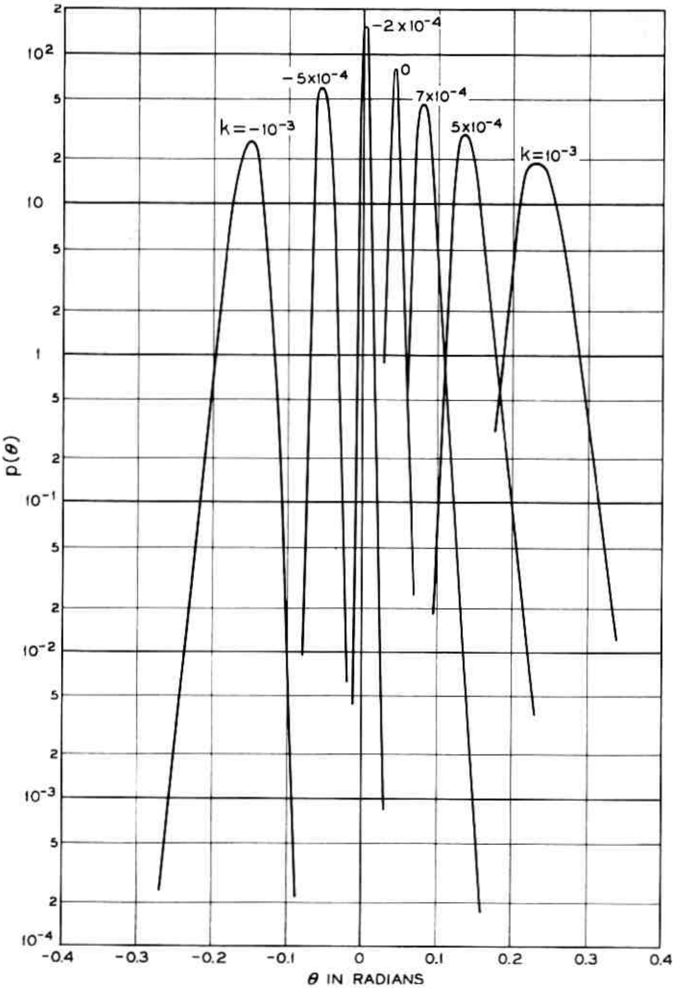


Fig. 17 — $p(\theta)$ for raised cosine excitation with various mistunings in the neighborhood of the optimum mistuning. Negative-going zero crossings and pulses $1.5T$ wide are assumed in making the calculations. $Q = 100$.

of the output pulse train as viewed, for example, on an oscilloscope. This is the method used for the adjustment of the repeater of Ref. 8.

X. PARTIAL RETIMING

In Section VIII we have shown that, in the absence of mistuning, the variable θ can be related to the normalized timing wave amplitude x

and the distribution for θ determined from the distribution for x. Here we will also make use of the distribution for x in order to analyze an idealized version of a forward-acting partial retiming scheme. The scheme we consider has been described by E. D. Sunde[5] and analyzed for periodic pulse patterns in Ref. 7. We make the same assumptions here as in the later reference, namely

1. The pulses exciting the tuned circuit are so narrow that they can be considered impulses. They are obtained by processing incoming pulses to the repeater and they excite a simple tuned circuit.
2. The timing wave is so clamped that its maximum excursion is at ground.
3. Reconstruction of the raised cosine pulse takes place when the algebraic sum of the timing wave and the raised cosine pulse crosses a threshold assumed to be at half the peak pulse amplitude.

For random impulse excitation of the tuned circuit prior to $t = 0$ and the definite occurrence of a pulse at $t = 0$, we have, according to the above assumptions (with no pulse overlap)

$$\frac{1}{2}\left(1 + \cos\frac{2\pi ts}{T}\right) - \frac{x}{2\bar{x}}\left(1 - \cos\frac{2\pi t}{T}\right) = \frac{1}{2} \tag{41}$$

for $|t| \leqq T/2s$
where

$$x = \sum_{n=0}^{\infty} a_n\beta^n,$$

$$a_o = 1 \text{ (the pulse at the origin definitely occurs),}$$

and

$$\bar{x} \equiv \text{average value of } x.$$

Equation (41) is based on the assumption that the average timing wave has a peak-to-peak amplitude equal to the peak pulse height (i.e., when $x = \bar{x}$, the timing wave amplitude varies between -1 and 0). If we define t_p as the time at which regeneration takes place and $\theta_p = 2\pi t_p/T$ as the corresponding phase angle, then it can be seen from (41) that this phase is a random variable dependent upon the random variable x. We will solve for θ_p under the condition $s = 1$, which means that the information-bearing pulses are resolved.* Under this condition $-(\pi/2) < \theta_p < 0$. Consistent with our previous definition of phase error, we will consider the negative of θ_p, since this makes the phase error positive

* Other pulse widths and different ratios of average timing wave amplitude to pulse peak can be handled, but we will not consider them here.

when we take our reference as the phase corresponding to the time at which the pulse peak occurs (at $t = 0$). In this way a positive phase error corresponds to regeneration prior to the pulse peak and permits direct comparison with the results of section 8 for the complete retiming approach. Solving (41) for $\cos \theta_p$ gives

$$\cos \theta_p = \frac{\dfrac{x}{\bar{x}}}{1 + \dfrac{x}{\bar{x}}} \tag{42}$$

and

$$\text{Prob}\,(\cos\theta_p \leqq \lambda) = \text{Prob}\,(\theta_p \geqq \cos^{-1}\lambda)$$

$$= \text{Prob}\left(\frac{\dfrac{x}{\bar{x}}}{1 + \dfrac{x}{\bar{x}}} \leqq \lambda\right) = \text{Prob}\left(x \leqq \frac{\lambda\bar{x}}{(1 - \lambda)}\right). \tag{43}$$

It is apparent from the above that we can use the distribution for x to determine the distribution for θ_p. For $Q = 100$, the distribution for x is shown in Fig. 14 and with (43) enables us to obtain the distribution for θ_p as shown in Fig. 18. When we compare this result with that of Fig. 15, which shows $1 - F(\theta)$ for the case of complete retiming, it is apparent that partial retiming results in a considerably larger variation of phase error. This supports the contention made in Ref. 7.

XI. CONCLUSIONS AND FUTURE WORK

We have derived an approximate relationship for the probability density and cumulative distribution for the phase error at the output of a tuned circuit when it is excited by a random or random plus periodic pulse train. The effects of mistuning of the tuned circuit and the finite widths of the driving pulses have been considered. Three independent checks of our results indicate that the expressions given are excellent approximations to the true state of affairs for $kQ < 0.1$ and $Q > 100$. Regions defined by these limits encompass values of k and Q of interest in PCM systems under consideration.

More specifically, we have shown that the distributions are not normal and are skewed in the direction of increasing phase error. When we consider pulse positions in which a pulse definitely occurs, it has been shown that the maximum phase error is bounded. In addition with raised cosine excitation we have demonstrated that the mistuning can be adjusted to minimize the mean or variance of the distribution for the

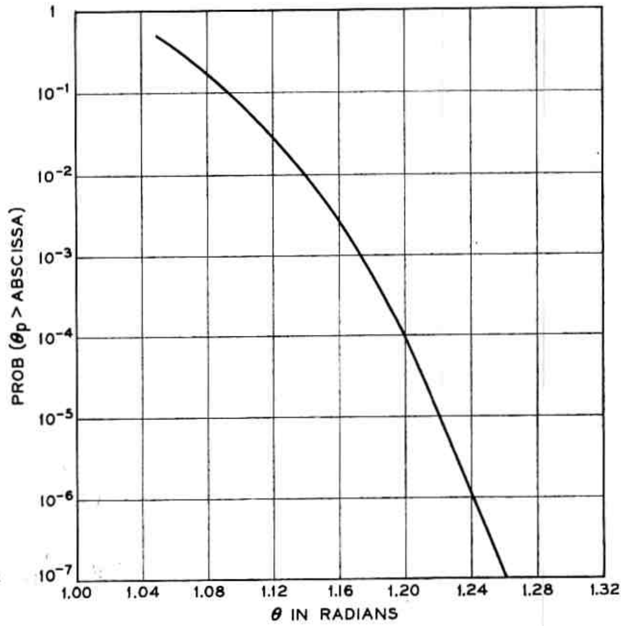


Fig. 18 — Distribution of the phase error with partial retiming. $Q = 100$ and $k = 0$. Raised cosine excitation pulse width $= T$.

phase error. The performance of an idealized version of a forward-acting partial retiming scheme has been analyzed and shown to be considerably inferior to a completely retimed repeater.

There are several desirable directions to proceed from our present position. First, it appears to be possible, in the case where we examine pulses only, to start from the maximum value of θ and work back toward the mean to better approximate the distribution near the tails. S. O. Rice has used this approach in related problems with success. Second, it is of interest to determine the pattern to give the maximum phase error at the output of a string of repeaters. This is not necessarily the pattern that creates $\theta_{\max}$ in a single repeater. In this regard, we have concentrated on only a single repeater. Obviously it is of interest to extend our results to a repeater string. This extension remains elusive.

XII. ACKNOWLEDGMENTS

The authors are indebted to several people for contributions to this effort. We would like to acknowledge the work of Miss E. G. Cheatham in computer programming. T. V. Crater kindly made the results of his digital computations available prior to publication. Miss M. R. Branower was most cooperative in providing us with distributions obtained from an analogue simulation of the problem. Our thanks go to E. N. Gilbert and D. Slepian for deriving the conditions for the maximum value of θ. We are grateful to S. O. Rice for helpful discussions, for results on the asymptotic behavior of the probability distribution, and for prior work on a related problem which provided us with helpful clues on how to proceed in our problem.

APPENDIX A. DERIVATION OF EQUATION FOR NORMALIZED TIMING ERROR

A-1. *Response of tuned circuit to random pulse train*

The impulse response of a parallel resonant circuit is well known to be

$$h(t) = \text{Real part of} \left[\frac{1}{C} \left(1 + \frac{j}{2Q} \right) e^{-(\pi/Q) f_o t} \, e^{+j2\pi f_o t} \right]. \tag{44}$$

Following Rowe,[2] we will imply the real part in all subsequent calculations involving complex quantities. The pulse train applied to the tuned circuit is given by

$$r(t) = \sum_{-\infty}^{\infty} a_n g(t - nT), \tag{45}$$

where:

$$a_n = 1 \text{ with probability } p,$$

$$a_n = 0 \text{ with probability } 1 - p, \text{ and}$$

$$g(t) = \text{pulse shape representing the binary 1.}$$

The response of the tuned circuit to $r(t)$ is

$$z(t) = \int_{-\infty}^{t} r(\tau) h(t - \tau) \, d\tau. \tag{46}$$

In view of (45), this can be written

$$z(t) = T \sum_{-\infty}^{+\infty} a_n h(t - nT) \cdot \int_{-\infty}^{(t/T)-n} g(xT) \exp \left[\left(\frac{f_o T \pi}{Q} - j2\pi f_o T \right) x \right] dx. \tag{47}$$

Define

$$f_o \equiv \frac{1+k}{T} = f_r(1+k), \tag{48}$$

with $k \equiv$ fractional mistuning from the pulse repetition frequency. Equation (47) can be manipulated to yield

$$z(t) = |\, A(t) \,|\, e^{j[2\pi f_r t + \Phi(t)]}, \tag{49}$$

where

$$\begin{aligned}\Phi(t) = {} & \tan^{-1}\frac{1}{2Q} + 2\pi f_r k t \\ & + \tan^{-1}\frac{\displaystyle\sum_{-\infty}^{\infty} a_n e^{\pi/Q(1+k)n}\left[-I_1\left(\frac{t}{T}-n\right)\sin 2\pi kn + I_2\left(\frac{t}{T}-n\right)\cos 2\pi kn\right]}{\displaystyle\sum_{-\infty}^{\infty} a_n e^{\pi/Q(1+k)n}\left[I_1\left(\frac{t}{T}-n\right)\cos 2\pi kn + I_2\left(\frac{t}{T}-n\right)\sin 2\pi kn\right]}\end{aligned} \tag{50}$$

and

$$\begin{aligned} I_1\left(\frac{t}{T}-n\right) &= \mathrm{Re}\int_{-\infty}^{(t/T)-n} g(xT)\exp\left[\left(\frac{\pi}{Q}f_oT - j2\pi f_oT\right)x\right]dx, \quad \text{and} \\ I_2\left(\frac{t}{T}-n\right) &= \mathrm{Im}\int_{-\infty}^{(t/T)-n} g(xT)\exp\left[\left(\frac{\pi}{Q}f_oT - j2\pi f_oT\right)x\right]dx. \end{aligned} \tag{51}$$

In (49), $|\, A(t) \,|$ represents the amplitude modulation on the carrier, while $\Phi(t)$ represents the phase modulation, the quantity of primary interest here.

A-2. *Equation for normalized timing error*

There is no loss in generality and it is convenient if the timing error is evaluated in the neighborhood of the pulse that occurs for $n = 0$.

In this neighborhood, negative-going zero crossings occur where

$$2\pi f_r t + \Phi(t) = \frac{\pi}{2}$$

or

$$\frac{t}{T} = \frac{1}{4} - \frac{\Phi(t)}{2\pi}. \tag{52}$$

Similarly, positive-going zero crossings occur for

$$\frac{t}{T} = -\frac{1}{4} - \frac{\Phi(t)}{2\pi}. \tag{53}$$

In the absence of tuning error, and with impulse excitation, $\Phi = 0$ and the negative and positive-going zero crossings occur close to $\pm T/4$ respectively.* Using these zero crossings as a reference, it is easily seen that the equations for normalized timing error become

$$\frac{e_1}{T} = -\frac{\Phi\left(\frac{1}{4} + \frac{e_1}{T}\right)}{2\pi} \tag{54}$$

for negative-going zero crossings and

$$\frac{e_2}{T} = -\Phi\left(-\frac{1}{4} + \frac{e_2}{T}\right) \tag{55}$$

for positive-going zero crossings.

With the exception of the minor generalization to arbitrary pulse shape, the method employed thus far is identical with that used by Rowe.[2] At this point in the evaluation of the timing error, we depart from his approximate solutions of (54) and (55) and attempt other approaches. Before proceeding in this direction, an indication of the approximation used by Rowe will be given. For the high Q case, Φ will be small and will change only a small amount for small changes in $2\pi f_r t$. Based on this assumption,

$$\begin{aligned} \frac{e_1}{T} &\doteq -\frac{\Phi(\frac{1}{4})}{2\pi}, \\ \frac{e_2}{T} &\doteq -\frac{\Phi(-\frac{1}{4})}{2\pi}. \end{aligned} \tag{56}$$

* Neglecting $\tan^{-1} \dfrac{1}{2Q}$ in (50)

It should be pointed out that these initial approximations are good for Rowe's purposes (steady-state error for $1/M$ patterns). However, for our purposes they need to be improved.

A-3. *Approximate solution of equation for normalized timing error*

One method for improving the accuracy of the initial approximation is to expand Φ in a power series about $T/4$ for negative-going zero crossings and retain two terms in the expansion to get

$$\frac{e_1}{T} = -\frac{\Phi(\frac{1}{4})}{2\pi + \Phi'(\frac{1}{4})}. \tag{57}$$

The form of Φ makes this approach messy and makes the determination of the probability distribution more difficult.

Another approach that is more tractable involves the separate Taylor expansion of I_1 and I_2 (51) in Φ about the reference time. If we retain only the first two terms in the Taylor expansion, replace the arctangent by its argument, and neglect k with respect to unity, we obtain for negative-going zero crossings

$$\begin{aligned}\frac{e_1}{T} &= -\frac{1}{4\pi Q} - \frac{k}{4} \\ &- \frac{1}{2\pi}\,\frac{\begin{aligned}\sum_{-\infty}^{\infty} a_n e^{(\pi/Q)n}\,[-\sin 2\pi kn\,(I_1(\tfrac{1}{4}-n) + e_1 I_1'(\tfrac{1}{4}-n) \\ + \cos 2\pi kn(I_2(\tfrac{1}{4}-n) + e_1 I_2'(\tfrac{1}{4}-n))]\end{aligned}}{\begin{aligned}\sum_{-\infty}^{\infty} a_n e^{(\pi/Q)n}[\cos 2\pi kn\,(I_1(\tfrac{1}{4}-n) + e_1 I_1'(\tfrac{1}{4}-n)) \\ + \sin 2\pi kn\,(I_2(\tfrac{1}{4}-n) + e_1 I_2'(\tfrac{1}{4}-n))]\end{aligned}}.\end{aligned} \tag{58}$$

If terms in $(e_1/T)^2$ are neglected, multiplication of both sides of (58) by the long denominator on the right results in a linear equation for e_1/T. This equation is applicable to arbitrary pulse shape, time-limited or not, and has been applied by one of the authors to periodic patterns of both Gaussian and raised cosine pulses in unpublished work. The results were compared with digital computer simulation and were in excellent agreement, thereby giving us confidence in using this approach for random pulse patterns. In this paper, we will concentrate on raised cosine pulses. This enables us to make use of some of the results given by H. E. Rowe in Section 2.5 of his paper.[2] For these time-limited pulses, the limits of integration on the I's of (51) are modified in an obvious way, and the upper limit on the sum over n is limited to the pulse immediately succeeding the time slot of interest at $n = 0$ for negative-going zero crossings. The evaluation of the various I's required is discussed in Appendix B.

Subject to the above conditions, the normalized timing error, as derived in Appendix B, can be written in the following form:

$$\frac{e_1}{T} = \frac{Ay + Bx + C}{Dy + Ex + F}, \tag{59}$$

where

$$\begin{aligned} y &\equiv \sum_{n=0}^{\infty} a_n e^{-(\pi/Q)n} \sin 2\pi kn, \\ x &\equiv \sum_{n=0}^{\infty} a_n e^{-(\pi/Q)n} \cos 2\pi kn, \end{aligned} \tag{60}$$

and $a_o \equiv 1$ (a pulse definitely occurs for $n = 0$). A through F are defined in Appendix B and are functions of the pulse width and Q and mistuning of the tuned circuit. In addition, C and F are functions of the presence or absence of a pulse in the succeeding time slot for negative-going zero crossings if sufficient pulse overlap exists. For positive-going zero crossings the form of the equation for the normalized timing error is the same and the new C and F are dependent upon the presence or absence of a pulse in the preceding time slot. This assumes that the pulse width is less than $2.5T$.

A-4. *Modification of probability distributions for pulse overlaps*

With the dependence on the occurrence of a succeeding pulse, as is the case for negative-going zero crossings with sufficient pulse overlap, we must modify the determination of the probability distribution as given in the main body of the paper. If we denote e_{11}/T and $C = C_1$, $F = F_1$ for $\underline{a}_1 = 1$ (a succeeding pulse definitely occurs), and denote e_{12}/T and $C = C_2$, $F = F_2$ for $\underline{a}_1 = 0$, then the average probability distribution for the timing deviation will be given by

$$\text{Prob}\left(\frac{e_1}{T} \leqq \lambda\right) = p \text{ Prob}\left(\frac{e_{11}}{T} \leqq \lambda\right) + (1 - p) \text{ Prob}\left(\frac{e_{12}}{T} \leqq \lambda\right). \tag{61}$$

When the pulse width is less than $1.5T$, $C_1 = C_2$, $F_1 = F_2$, and therefore $e_{11} = e_{12}$ and the above modification is not required. A similar procedure is applicable for positive-going zero crossings.

APPENDIX B. RAISED COSINE PULSES

B-1. *Determination of I's*

For a raised cosine pulse centered at the origin and of width T/s, I of equation (51) becomes

$$
\begin{aligned}
I(x) &= 0 && x < -\frac{1}{2s} \\
I(x) &= \int_{(1/2s)}^{x} (1 + \cos 2\pi s x_1) e^{[(\pi/Q) - j2\pi]Kx_1} dx_1 && |x| \leqq \frac{1}{2s} \qquad (62) \\
I(x) &= I\left(\frac{1}{2s}\right) && x > \frac{1}{2s}
\end{aligned}
$$

where

$$K \equiv (1 + k)$$

The integral in (62) is readily evaluated to give

$$
\begin{aligned}
-jI = \frac{1}{2\pi K} & \left[\frac{e^{[(\pi/Q) - j2\pi]Kx} - e^{-[(\pi/Q) - j2\pi]K/2s}}{\left(1 + \frac{j}{2Q}\right)} \right] \\
& + \frac{1}{4\pi} \left[\frac{e^{[(\pi/Q) - j2\pi]Kx} e^{+j2\pi s x} - e^{-[(\pi/Q) - j2\pi]K/2s}}{(K - s) + j\frac{K}{2Q}} \right] \qquad (63) \\
& + \frac{1}{4\pi} \left[\frac{e^{[(\pi/Q) - j2\pi]Kx} e^{-j2\pi s x} - e^{-[(\pi/Q) - j2\pi]K/2s}}{(K + s) + j\frac{K}{2Q}} \right]
\end{aligned}
$$

The derivatives required in the evaluation of (58) may be obtained from

$$\frac{dI}{dx} = e^{(\pi/Q)Kx} [e^{-j2\pi Kx} + \tfrac{1}{2} e^{-j2\pi(K-s)x} + \tfrac{1}{2} e^{-j2\pi(K+s)x}]. \qquad (64)$$

In the evaluation of I and dI/dx, mistuning makes very little difference for the allowable values in practical systems. Therefore, with $K = 1$

$$I']_{x=1/4} = -j\, e^{\pi/4Q} \left[1 + \cos\frac{\pi s}{2}\right] \qquad (65)$$

$$I']_{x=-1/4} = j\, e^{\pi/4Q} \left[1 + \cos\frac{\pi s}{2}\right] \qquad (66)$$

$$I']_{x=3/4} = j\, e^{\pi/4Q}\left[1 + \cos\frac{3\pi s}{2}\right] \tag{67}$$

$$I']_{x=-3/4} = -j\, e^{\pi/4Q}\left[1 + \cos\frac{3\pi s}{2}\right]. \tag{68}$$

Equations (65) and (68) above are required for negative-going zero crossings, while (66) and (67) are needed for positive-going zero crossings.

B-2. *Equation for Normalized Timing Error with Raised Cosine Pulses*

From (58) we can write the equation for normalized timing error as

$$\frac{e_1}{T} \doteq -\frac{1}{4\pi Q} - \frac{k}{4} - \frac{1}{2\pi}\frac{N}{P}, \tag{69}$$

where N and P are defined by comparison with (58). Cross multiplication by P, neglecting terms in e_1^2 and collecting terms, yields

$$\frac{e_1}{T} = \frac{Ay + Bx + C}{Dy + Ex + F}, \tag{70}$$

where x and y are defined by (60), and A through F are as follows:

$$A = -\frac{1}{2\pi} I_1\left(\frac{1}{2s}\right) + \left[\frac{1}{4\pi Q} + \frac{k}{4}\right] I_2\left(\frac{1}{2s}\right)$$

$$B = -\frac{1}{2\pi} I_2\left(\frac{1}{2s}\right) - \left[\frac{1}{4\pi Q} + \frac{k}{4}\right] I_1\left(\frac{1}{2s}\right)$$

$$\begin{aligned} C = {} & -\frac{1}{2\pi}\left[I_2\left(\frac{1}{4}\right) - I_2\left(\frac{1}{2s}\right)\right] - \left[\frac{1}{4\pi Q} + \frac{k}{4}\right]\left[I_1\left(\frac{1}{4}\right) - I_1\left(\frac{1}{2s}\right)\right] \\ & + \underline{a}_1\, e^{(\pi/Q)}\left\{\frac{1}{2\pi}\left[\sin 2\pi k I_1\left(-\frac{3}{4}\right) - \cos 2\pi k I_2\left(-\frac{3}{4}\right)\right]\right. \\ & \left. - \left[\frac{1}{4\pi Q} + \frac{k}{4}\right]\left[\sin 2\pi k I_2\left(-\frac{3}{4}\right) + \cos 2\pi k I_1\left(-\frac{3}{4}\right)\right]\right\} \end{aligned}$$

$$D = -I_2\left(\frac{1}{2s}\right)$$

$$E = I_1\left(\frac{1}{2s}\right)$$

$$F = I_1\left(\frac{1}{4}\right) - I_1\left(\frac{1}{2s}\right) + \left[\frac{1}{4\pi Q} + \frac{k}{4}\right] I_1'\left(\frac{1}{4}\right) + \frac{1}{2\pi} I_2'\left(\frac{1}{4}\right)$$

$$+ \underline{a}_1 e^{(\pi/Q)} \left\{ \cos 2\pi k \left[\frac{1}{2\pi} I_2' \left(-\frac{3}{4}\right) + I_1 \left(-\frac{3}{4}\right) \right.\right.$$
$$\left. + \left(\frac{1}{4\pi Q} + \frac{k}{4}\right) I_1' \left(-\frac{3}{4}\right) \right] + \sin 2\pi k \left[-\frac{1}{2\pi} I_1' \left(-\frac{3}{4}\right) \right.$$
$$\left.\left. + I_2 \left(-\frac{3}{4}\right) + \left(\frac{1}{4\pi Q} + \frac{k}{4}\right) I_2' \left(-\frac{3}{4}\right) \right] \right\}.$$

For positive-going zero crossings, only the constants C and F are changed.

B-3. *Numerical Evaluation of Constants*

In order to make use of some of Rowe's results, we will choose the same two cases for pulse width that he used.

Case 1. $s = 1$, *Pulses Resolved*

a. *Negative-Going Zero Crossings.* Since mistuning has a small effect on the evaluation of the I's, we neglect it in this regard. Neglecting terms in $1/Q^2$ and k/Q, after some arithmetic one arrives at

$$\frac{e_1}{T} = -\frac{\dfrac{1}{4\pi} y - \left(\dfrac{1}{16\pi Q} - \dfrac{k}{8}\right) x + \dfrac{0.0795}{2\pi} + \dfrac{0.0316}{Q} + 0.0085k}{\dfrac{3}{16\pi Q} y + \dfrac{1}{2}(x - 1) + 0.375 + \dfrac{0.06}{Q}}. \quad (71)$$

$Q > 50$ and $kQ < 0.2$ encompass values of practical interest. In this region the term in y in the denominator of (71) can be neglected and the numerator term $0.0085k$ is also negligible. It is also convenient to deal with phase error rather than timing error. Therefore, we rewrite (71) as

$$\theta_1 = -\frac{2\pi e_1}{T} = \frac{y - \dfrac{1}{Q}\left[\dfrac{1}{4} - \dfrac{\pi}{2} kQ\right] x + 0.159 + \dfrac{0.397}{Q}}{x - 0.25 + \dfrac{0.12}{Q}}. \quad (72)$$

The multiplication by -2π is used to avoid any questions later on as to which way certain inequalities are to be taken. This means that θ is the negative of the phase error as previously defined. A positive value of θ signifies that the zero crossing occurs prior to $\pm T/4$ for negative going and positive going zero crossings respectively. The general form of θ for all the cases to be considered herein then can be written as

$$\theta = \frac{y + a}{x + b} + c. \tag{73}$$

For the situation under consideration in this section,

$$a = 0.159 + \frac{0.334}{Q} + \frac{\pi}{8} k$$

$$b = -0.25 + \frac{0.12}{Q}$$

$$c = -\frac{1}{Q}\left[\frac{1}{4} - \frac{\pi}{2} kQ\right].$$

b. *Positive-Going Zero Crossings.* Proceeding in the same way as in Sections B-2 and B-3 above, the phase error for positive-going zero crossings is as in (73) with

$$a = 0.159 - \frac{0.2}{Q} + \frac{3\pi}{8} k$$

$$b = -0.75 + \frac{0.62}{Q}$$

$$c = -\frac{1}{Q}\left[\frac{1}{4} - \frac{\pi}{2} kQ\right].$$

In this case it should be noted that with zero mistuning ($y = 0$) and with a pulse for $n = 0$ and nowhere else, a positive-going zero crossing does not occur in the neighborhood of $-T/4$. Under this special condition, $x = 1$ and (73) with the constants of this section would predict an incorrect error in the positive-going zero crossing. Of course such a sparse pattern occurs with probability zero. Fortunately, for all other more reasonable periodic patterns, results obtained from (73) are in good agreement with computer simulation.

Case 2. $s = \frac{2}{3}$, Pulses Overlapping, Base Width = 1.5T

a. *Negative-Going Zero Crossings.* In this section we will dispense with all of the algebra and arithmetic and simply write down the final results. For the case at hand

$$\frac{e_1}{T} \doteq -\frac{\frac{0.255}{2\pi} y - \frac{1}{Q}(0.073 - 0.064kQ)x + 0.0264 + \frac{1}{Q}\cdot(0.034 - 0.02kQ)}{0.255x - 0.062 + 0.048/Q}. \tag{74}$$

When this is converted to the form of (73), we have

$$a = 0.65 + \frac{0.4}{Q} - 0.21k$$

$$b = -0.243 + \frac{0.188}{Q}$$

$$c = -\frac{1}{Q}[1.8 - 1.58kQ].$$

b. *Positive-Going Zero Crossings*

$$a = 0.65 - \frac{0.4}{Q} + 0.94k$$

$$b = -0.753 + \frac{1.66}{Q}$$

$$c = -\frac{1}{Q}[1.8 - 1.58kQ].$$

The remarks made in connection with positive-going zero crossings for Case 1 are equally applicable here.

APPENDIX C. SEMI-INVARIANTS FOR THE JOINT DENSITY FUNCTION OF x_1 AND y_1

C-1. *One out of M pulses definitely occur; the remaining pulses are independent and occur with probability ½; raised cosine pulses.*

The characteristic function is defined as

$$\varphi(u,v) = E \exp i(ux_1 + vy_1), \tag{75}$$

where E is the expectation operator, and from Appendix B

$$\begin{aligned} x_1 &\equiv \sum_{m=0}^{\infty} e^{-\alpha Mm} \cos 2\pi kMm + b + \sum_{n \neq mM}^{\infty} a_n e^{-\alpha n} \cos 2\pi kn, \\ y_1 &\equiv \sum_{m=0}^{\infty} e^{-\alpha Mm} \sin 2\pi kMm + a + \sum_{n \neq mM}^{\infty} a_n e^{-\alpha n} \sin 2\pi kn, \end{aligned} \tag{76}$$

with $\alpha \equiv \pi/Q$. Substituting (76) in (75) and performing the expectation operation gives

$$\begin{aligned}\varphi(u,v) &= \exp i \sum_{m=0}^{\infty} e^{-(\pi/Q)Mm} \, (u \cos 2\pi kMm + v \sin 2\pi kMm) \\ &\cdot \exp i(ub + va) \times \prod_{n \neq mM}^{\infty} \exp\left\{\frac{i}{2} e^{-(\pi/Q)n}(u \cos 2\pi kn + v \sin 2\pi kn)\right\} \quad (77) \\ &\times \prod_{n \neq mM}^{\infty} \cos\left\{\frac{e^{-(\pi/Q)n}}{2} (u \cos 2\pi kn + v \sin 2\pi kn)\right\}\end{aligned}$$

which may be rearranged to

$$\begin{aligned}\varphi(u,v) &= \exp\left[\frac{i}{2} \sum_{n=0}^{\infty} \left\{e^{-(\pi/Q)Mn} \, (u \cos 2\pi kMn + v \sin 2\pi kMn)\right.\right. \\ &\quad \left.\left. + e^{-(\pi/Q)n} \, (u \cos 2\pi kn + v \sin 2\pi kn)\right\}\right] \exp i(ub + va) \\ &\quad \times \frac{\displaystyle\prod_{n=0}^{\infty} \cos\left\{\frac{e^{-(\pi/Q)n}}{2} \quad (u \cos 2\pi kn + \sin 2\pi kn)\right\}}{\displaystyle\prod_{n=0}^{\infty} \cos\left\{\frac{e^{-(\pi/Q)Mn}}{2} \, (u \cos 2\pi kMn + v \sin 2\pi kMn)\right\}}. \quad (78)\end{aligned}$$

When we take the logarithm of (78), we obtain

$$\begin{aligned}\log \varphi(u,v) &= \frac{i}{2} \sum_{n=0}^{\infty} [\beta^{Mn}(u \cos 2\pi kMn + v \sin 2\pi kMn) \\ &\qquad + \beta^{n}(u \cos 2\pi kn + v \sin 2\pi kn)] \\ &\quad + i(ua + vb) + \sum_{n=0}^{\infty} \log \cos\left[\frac{\beta^{n}}{2} \, (u \cos 2\pi kn + v \sin 2\pi kn)\right] \quad (79) \\ &\quad - \sum_{n=0}^{\infty} \log \cos\left[\frac{\beta^{Mn}}{2} \, (u \cos 2\pi kMn + v \sin 2\pi kMn)\right],\end{aligned}$$

where $\beta = e^{-(\pi/Q)}$.

The first sum in (79) may be carried out, and when combined with $i(ua + vb)$ yields the semi-invariants λ_{10} and λ_{01} which are of course the mean values for x_1 and y_1 respectively. Since the last two terms of (79) are similar in form, we will confine our manipulations to the next to the last term. We denote this term by

$$F(u,v) = \sum_{n=0}^{\infty} \log \cos\left[\frac{\beta^{n}}{2} \, (u \cos 2\pi kn + v \sin 2\pi kn)\right]. \qquad (80)$$

Using the infinite product expansion for the cosine and the power series expansion for the log; i.e.,

$$\cos z = \prod_{m=0}^{\infty} \left[1 - \left(\frac{2z}{(2m+1)\pi} \right)^2 \right] \qquad (z^2 < \infty)$$

and

$$\log (1 - x) = - \sum_{j=1}^{\infty} \frac{x^j}{j} \qquad (x^2 < 1).$$

$F(u,v)$ becomes

$$F(u,v) = - \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \sum_{j=1}^{\infty} \frac{(uC_n + vS_n)^{2j}}{i(2m+1)^{2j}\pi^{2j}}, \tag{81}$$

where $C_n \equiv e^{-\alpha n} \cos 2\pi kn$ and $S_n \equiv e^{-\alpha n} \sin 2\pi kn$. The sum over j may be obtained by virtue of

$$\sum_{m=0}^{\infty} \frac{1}{(2m+1)^{2j}} = \frac{(2^{2j} - 1)(-1)^{j-1}(2\pi)^{2j} B_{2j}}{2^{2j+1}(2j)!}$$

where the B_{2j} are the Bernoulli numbers. With the above sum over m and the expansion of $(uC_n + vS_n)^{2j}$ in a binomial series, we arrive at

$$F(u,v) = \sum_{j=1}^{\infty} \frac{(-1)^j B_{2j}(2^{2j} - 1)}{2j(2j)!} \sum_{r=0}^{2j} \binom{2j}{r} \sum_{n=0}^{\infty} C_n^{\,r} u^r (S_n v)^{2j-r} \tag{82}$$

Proceeding in the same manner that took us from (80) to (82), it can be verified that the last term of (79) takes the same form as the right-hand side of (82) with n replaced by nM. These results and comparison with the definition of the semi-invariants for a two dimensional distribution[10] lead to the following for the semi-invariants for the process under consideration:

$$\lambda_{10} = \frac{1}{2} \left[\frac{1 - \beta \cos 2\pi k}{1 - 2\beta \cos 2\pi k + \beta^2} + \frac{1 - \beta^M \cos 2\pi kM}{1 - 2\beta^M \cos 2\pi kM + \beta^{2M}} \right] + b,$$

$$\lambda_{01} = \frac{1}{2} \left[\frac{\beta \sin 2\pi b}{1 - 2\beta \cos 2\pi k + \beta^2} + \frac{\beta^M \sin 2\pi kM}{1 - 2\beta^M \cos 2\pi kM + \beta^{2M}} \right] + a,$$

and

$$\lambda_{rs}]_{r+s>1} = \frac{B^{r+s}(2^{r+s} - 1)}{r + s} \sum_{n=0}^{\infty} [C_n^{\,r} S_n^{\,s} - C_{nM}^{\,r} S_{nM}^{\,s}]. \tag{83}$$

The sum over n can be shown to be a geometric series multiplied by two finite series if the sines and cosines in S and C respectively are represented in exponential form and use is made of the binomial expansion. After some algebra, an alternate form for (83) can be shown to be

$$\lambda_{rs}]_{r+s>1} = \frac{B_{r+s}(2^{r+s} - 1)r!\,s!}{(r + s)2^r(2i)^s}\, G(r,s,\beta,k,M), \tag{84}$$

where $G(r,s,\beta,k,M)$ is (shortened to G)

$$\begin{aligned} G = \sum_{p=0}^{r} \sum_{q=0}^{s} & \frac{(-1)^q}{q!(r - p)!q!(s - q)!} \\ & \cdot\left[\frac{1}{1 - \beta^{(r+s)} \exp\left[i2\pi k(r + s - 2p - 2q)\right]}\right. \\ & \quad \left. - \frac{1}{1 - \beta^{M(r+s)} \exp\left[i2\pi kM(r + s - 2p - 2q)\right]}\right]. \end{aligned} \tag{85}$$

For u and v in the neighborhood of zero, the contributions to the series in (79) become smaller as n becomes larger. The importance of successive terms is judged by the exponential decay factor $e^{-(n\pi/Q)}$. If we consider all terms up to some $n_{\max}$ where $n_{\max} \gg Q/\pi$ and $kn_{\max} \ll 1$, then we arrive at the following inequality

$$\frac{kQ}{\pi} \ll 1. \tag{86}$$

Under the above condition $\cos 2\pi kn$ can be replaced by unity and $\sin 2\pi kn$ by $2\pi kn$ for all terms of importance in the series and (79) becomes approximately

$$\begin{aligned} \log \varphi(u,v) \sim i & \left\{\frac{u}{2}\left[\frac{1}{1 - \beta} + \frac{1}{1 - \beta^M}\right] + v\pi k\left[\frac{\beta}{(1 - \beta)^2} + \frac{M\beta^M}{(1 - \beta^M)^2}\right]\right. \\ & \left. + (ub + va)\right\} + \sum_{n=0}^{\infty} \log \cos\left\{\frac{\beta^n}{2}(u + 2\pi knv)\right\} \\ & - \sum_{n=0}^{\infty} \log \cos\left\{\frac{\beta^{Mn}}{2}(u + 2\pi kMnv)\right\}. \end{aligned} \tag{87}$$

Paralleling the operations performed on (80) to obtain (82) it can be shown that the semi-invariants obtained from (87) under the condition (86) are

$$\begin{aligned} \lambda_{10} &= \frac{1}{2}\left(\frac{1}{1 - \beta} + \frac{1}{1 - \beta^M}\right) + b, \\ \lambda_{01} &= \pi k\left(\frac{\beta}{(1 - \beta)^2} + \frac{M\beta^M}{(1 - \beta^M)^2}\right) + a, \quad \text{and} \\ \lambda_{rs}]_{r+s>1} &= (-1)^s \frac{B_{r+s}(2^{r+s} - 1)}{(r + s)}(2\pi k)^s \frac{d^s}{dg^s}\left(\frac{1}{1 - e^{-g}} - \frac{1}{1 - e^{-Mg}}\right), \end{aligned} \tag{88}$$

with $g = (r + s)\pi/Q$.

C-2. *Same as I Above Except That Pulses are Impulses*

For this case the semi-invariants are as above with $a = 0 = b$.

C-3. *Impulse Excitation, All Pulses Random*

With this type of excitation, we have

$$\lambda_{10} = \frac{1}{2}\left[\frac{1 - \beta \cos 2\pi k}{1 - 2\beta \cos 2\pi k + \beta^2}\right],$$

$$\lambda_{01} = \frac{1}{2}\left[\frac{\beta \sin 2\pi k}{1 - 2\beta \cos 2\pi k + \beta^2}\right],$$

and

$$\lambda_{rs}]_{r+s>1} = \frac{B_{r+s}(2^{r+s} - 1)_{r!s!}}{(r + s)2^r(2i)^s}\left[\sum_{p=0}^{r}\sum_{q=0}^{s}\frac{(-1)^q}{p!(r - p)!q!(s - q)!} \cdot \frac{1}{1 - \beta^{r+s}\exp\left[i2\pi k(r + s - 2p - 2q)\right]}\right]. \tag{89}$$

It is readily shown in this case that the approximate semi-invariants [subject to (86)] are

$$\begin{aligned}\lambda_{10} &= \frac{1}{2}\left(\frac{1}{1 - \beta}\right),\\ \lambda_{01} &= \frac{\pi k\beta}{(1 - \beta)^2}\\ \lambda_{rs}]_{r+s>1} &= (-1)^s\,\frac{B_{r+s}(2^{r+s} - 1)}{r + s}\,(2\pi k)^s\,\frac{d^s}{dg^s}\left(\frac{1}{1 - e^{-g}}\right),\end{aligned} \tag{90}$$

with $g = (r + s)\pi/Q$.

APPENDIX D

High Q Behavior of $p(\theta)$

To illustrate the behavior of the probability density function when the Q of the resonator becomes large, we consider $p(\theta)$ in the neighborhood of the mean, θ_o. We include terms of the double summation in (19) for which $r + s = 4$. Since the Bernoulli numbers $B_{r+s} = 0$ for $r + s$ odd and >1, the terms λ_{rs} for $r + s = 3$ are zero. For $\theta \sim \theta_o$, therefore, $p(\theta)$ becomes

$$p(\theta) \doteq \frac{1}{\sqrt{2\pi}} \frac{\lambda_{10}}{(\lambda_{20}\theta_o^2 - 2\lambda_{11}\theta_o + \lambda_{02})^{\frac{1}{2}}} \cdot \exp -\frac{{\lambda_{10}}^2}{2} \frac{(\theta - \theta_o)^2}{(\lambda_{20}\theta_o^2 - 2\lambda_{11}\theta_o + \lambda_{02})} \cdot \left[1 + \frac{H_4\left(\dfrac{\lambda_{10}(\theta - \theta_o)}{\sqrt{2}(\lambda_{20}{\theta_o}^2 - 2\lambda_{11}\theta_o + \lambda_{02})^{\frac{1}{2}}}\right)}{\dfrac{\sqrt{2}(\lambda_{20}{\theta_o}^2 - 2\lambda_{11}\theta_o + \lambda_{02})^{\frac{1}{2}}}{\lambda_{10}}} \sum_{\substack{r \\ r+s>2}}^{r+s=4}\sum_{s} (-1)^r \frac{\lambda_{rs}}{r!s!} \frac{{\theta_o}^r}{{\lambda_{10}}^4} \right]. \tag{91}$$

The semi-invariants of interest in the above equation are given below and were determined using the results of the previous section for the case "all impulses random," subject to $kQ \ll \pi$.

$$\lambda_{10} = \frac{1}{2(1 - \beta)} \qquad \theta_o = \frac{\lambda_{01}}{\lambda_{10}} = \frac{2\pi k\beta}{1 - \beta}$$

$$\lambda_{20} = \frac{1}{4}\frac{1}{(1 - \beta^2)} \qquad \lambda_{11} = \frac{\pi k}{2}\frac{\beta^2}{(1 - \beta^2)^2} \qquad \lambda_{02} = \frac{(\pi k)^2\beta^2(1 + \beta^2)}{(1 - \beta^2)^3}$$

$$\lambda_{40} = -\frac{1}{8}\frac{1}{(1 - \beta^4)} \quad \lambda_{31} = -\frac{\pi k}{4}\frac{\beta^4}{(1 - \beta^4)^2} \quad \lambda_{22} = -\frac{(\pi k)^2}{2}\frac{\beta^4(1 + \beta^4)}{(1 - \beta^3)^3}$$

$$\lambda_{13} = -(\pi k)^3 \frac{\beta^4}{(1 - \beta^4)^4} (1 + 4\beta^4 + \beta^8)$$

$$\lambda_{04} = -2(\pi k)^4 \frac{\beta^4}{(1 - \beta^4)^5} (1 + 11\beta^4 + 11\beta^8 + \beta^{12}).$$

Using the above expressions for the λ's, the following quantities in (91) may be reduced to

$$\frac{\lambda_{10}}{(\lambda_{20}\theta_o^2 - 2\lambda_{11}\theta_o + \lambda_{02})^{\frac{1}{2}}} = \frac{1}{\dfrac{\sqrt{2}(2\pi k)\beta}{(1 - \beta)^{\frac{1}{2}}(1 - \beta)^{\frac{3}{2}}}} = \frac{1}{\sigma},$$

$$\sum_{\substack{r \\ r+s>2}}^{r+s=4}\sum_{s} (-1)^r \frac{\lambda_{rs}}{r!\,s!}\frac{{\theta_o}^r}{{\lambda_{10}}^4} = -\frac{1}{4!}\frac{2(2\pi k)^4\beta^4}{1 - \beta^4} \cdot \left[1 - \frac{4\beta^3(1 - \beta)}{(1 - \beta^4)} + \frac{6\beta^2(1 + \beta^4)(1 - \beta)^2}{(1 - \beta^4)^2} - \frac{4\beta(1 - \beta)^3(1 + 4\beta^4 + \beta^8)}{(1 - \beta^4)^3} + \frac{(1 - \beta)^4(1 + 11\beta^4 + 11\beta^8 + \beta^{12})}{(1 - \beta^4)^4}\right],$$

or

$$\sum_{r}\sum_{s}^{r+s=4}{}_{r+s>2} (-1)^r \frac{\lambda_{rs}}{r!\,s!}\frac{\theta_o^{\,r}}{\lambda_{10}^{\,4}} \equiv \frac{\lambda_4}{4!}.$$

The probability density therefore takes the form

$$p(\theta) \doteq \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{(\theta-\theta_o)^2}{2\sigma^2}\left[1+\frac{\lambda_4}{4!}\frac{H_4\left(\dfrac{\theta-\theta_o}{\sqrt{2}\sigma}\right)}{4\sigma^4}\right]. \qquad (92)$$

This result is in the form of the standard Edgeworth approximation with θ_o, σ, and λ_4 the mean, the standard deviation and the 4th semi-invariant of the θ distribution, respectively. In the limit as Q becomes large ($\beta \to 1$) we approximate $1 - \beta$ by π/Q and

$$\sigma \to k\sqrt{\pi Q} \quad \theta_o \to 2kQ$$

The coefficient of the 4th Hermite polynomial approaches $-(5\pi/128Q)$. Equation (92) then indicates the approach to the normal law with the first correction term going as $1/Q$. The results for θ_o and σ correspond to those derived earlier by Bennett, Rice and others.

APPENDIX E

Determination of $\theta_{\max}$

For $kQ \ll \pi$, a good approximation for θ is (from Appendix B)

$$\theta = \frac{a + 2\pi k \sum_{n=0}^{\infty} a_n n\beta^n}{b + \sum_{n=0}^{\infty} a_n\beta^n}. \qquad (93)$$

When $a_o = 1$, we have

$$\frac{\theta}{2\pi k} = \frac{\dfrac{a}{2\pi k} + \sum_{n=1}^{\infty} a_n n\beta}{1 + b + \sum_{n=1}^{\infty} a_n\beta^n}. \qquad (94)$$

It is of interest to determine the pulse pattern that yields the maximum value of $\theta/2\pi k$. This is equivalent to the determination of a one-zero sequence of a_n's such that (94) is a maximum.

Assume that an initial pattern has been chosen such that $\theta/2\pi k = A_o/B_o$. If a single a_n is changed from zero to one (pulse added), then $\theta/2\pi k$ is changed to $(A_o + n\beta^n)/(B_o + \beta^n)$. Clearly, we should effect this conversion if

$$\frac{A_o + n\beta^n}{B_o + B^n} \geqq \frac{A_o}{B_o}$$

or

$$n \geqq \frac{A_o}{B_o}. \tag{95}$$

On the other hand if a one is changed to a zero (pulse removed), then $\theta/2\pi k$ will be increased if

$$\frac{A_o - n\beta^n}{B_o - \beta^n} > \frac{A_o}{B_o}$$

or

$$n < \frac{A_o}{B_o}. \tag{96}$$

The process is continued in this manner until all $a_n = 1$ for $n \geqq n_c$ and all $a_n = 0$ for $n < n_c$ (except a_o, which is constrained to be unity). n_c may be determined from the above process, since

$$n_c = \frac{\theta_{\max}}{2\pi k} = \frac{\dfrac{a}{2\pi k} + \sum\limits_{n=n_c}^{\infty} n\beta^n}{1 + b + \sum\limits_{n=n_c}^{\infty} \beta^n}, \tag{97}$$

which can be rearranged to

$$\frac{\beta^{n_c+1}}{(1 - \beta)^2} = -\frac{a}{2\pi k} + n_c(1 + b). \tag{98}$$

When a periodic pulse pattern of 1 out of every M pulses is forced, $\theta_{\max}$ is found in the same manner as above and the relationship between the various parameters to achieve this maximum is given by (15) of the main body of the paper.

REFERENCES

1. Bennett, W. R., B.S.T.J., **37,** Nov., 1958, p. 1501.
2. Rowe, H. E., B.S.T.J., **37,** Nov., 1958, p. 1543.

3. DeLange, O. E., B.S.T.J., **37,** Nov., 1958, p. 1455.
4. DeLange, O. E., and Pustelnyk, M., B.S.T.J., **37,** Nov., 1958, p. 1487.
5. Sunde, E. D., B.S.T.J., **36,** July, 1957, p. 891.
6. DeLange, B.S.T.J., **35,** Jan., 1956, p. 67.
7. Aaron, M. R., B.S.T.J., **41,** Jan., 1962, p. 99.
8. Mayo, J. S., B.S.T.J., **41,** Jan., 1962, p. 25.
9. Cramér, H., *Mathematical Methods of Statistics*, Princeton University Press, Princeton, N. J., 1946, p. 317 and p. 224.
10. Laning and Battin, *Random Processes in Automatic Control*, McGraw-Hill Book Co., New York, 1956, p. 61.

Properties and Design of the Phase-Controlled Oscillator with a Sawtooth Comparator

By C. J. BYRNE

(Manuscript received September 1, 1961)

A sawtooth phase comparator has advantages over the more common sinusoidal comparator in a phase-controlled oscillator because its output is linear for larger values of phase error. For some applications, it is no more complex or expensive than the sinusoidal comparator.

This paper analyzes properties of the phase-controlled oscillator with a sawtooth comparator that have been mentioned in the literature for sinusoidal comparators. In addition, there is new theoretical material on the effect of fast jitter and noise.

The properties of the circuit are presented in a manner which is convenient for design.

Since it is easier to analyze the circuit with a sawtooth comparator, many applications of the device have been considered. Because of this wide viewpoint, the paper may be helpful in understanding the phase-controlled oscillator in general.

Table of Contents

I. INTRODUCTION

The phase-controlled oscillator (see Fig. 1), otherwise known as the phase-locked oscillator, is often used to produce a signal whose frequency and phase are controlled by an input signal. The literature[1,2,3] on the subject assumes that the phase comparator, which is the error detector of the loop, produces an output which is proportional to the sine of the phase difference.

This paper considers the case of the sawtooth comparator, whose output is a linear function of the phase difference over a periodic range (see Fig. 2a). Because of this linearity, the sawtooth comparator is superior in operation to the sinusoidal comparator for some applications. In general, the sinusoidal comparator is simpler and cheaper, but in applications involving digital signals, the two are comparable in cost and complexity.

The purpose of this paper is to present a comprehensive survey of many properties of the phase-controlled oscillator, relating to many different applications. We have drawn heavily on the literature, modifying the analysis to make it apply to the sawtooth comparator. In addition, there is new theoretical material on the effect of fast jitter and noise. New results derived by A. J. Goldstein in a companion paper[4] are presented in an abbreviated form, more suitable for design.

Most of the properties are presented in a graphical form which facilitates design.

II. DESCRIPTION

2.1 *General*

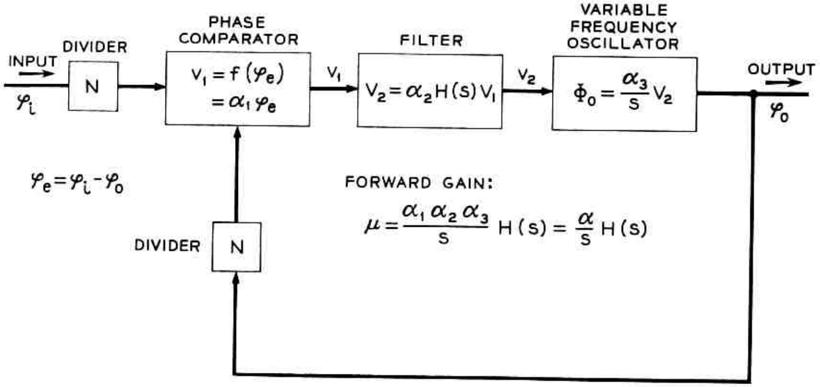
The block diagram of a phase-controlled oscillator is shown in Fig. 1. Notice the resemblance to a negative feedback amplifier or a servo loop.[2] There is a forward gain path, a feedback path, and a subtracting or error detecting device.

The input and output signals are not the voltages themselves, but are the phases of the nearly periodic voltages. If the input and output voltages are at different frequencies, dividers or multipliers must be used to bring them to a common frequency at the phase comparator. In this paper, we will assume that the output and the input are at the same frequency. We will however, consider the use of dividers to allow the comparator to operate on the Nth submultiple of the input and output frequency. We will measure phase of the submultiple signals in radians of the original frequency.

2.2 *Phase Comparator*

The phase comparator is the error detector of the servo loop. It produces a voltage which depends on the phase difference between the input submultiple and the output submultiple.

Of course, the comparator cannot distinguish between different cycles



INPUT
φ_i
DIVIDER
N
PHASE COMPARATOR
$V_1 = f(\varphi_e)$
$= \alpha_1 \varphi_e$
V_1
FILTER
$V_2 = \alpha_2 H(s) V_1$
V_2
VARIABLE FREQUENCY OSCILLATOR
$\Phi_0 = \frac{\alpha_3}{s} V_2$
OUTPUT
φ_0
$\varphi_e = \varphi_i - \varphi_0$
FORWARD GAIN:
$\mu = \frac{\alpha_1 \alpha_2 \alpha_3}{s} H(s) = \frac{\alpha}{s} H(s)$
DIVIDER
N

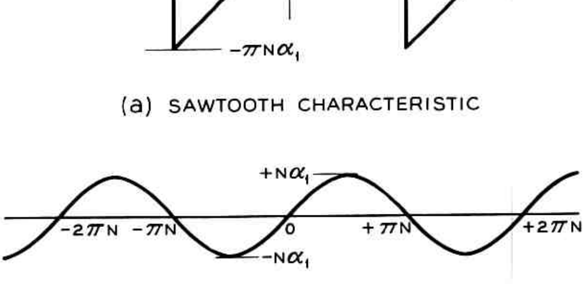
Fig. 1 — Block diagram of the phase-controlled oscillator.

of the input and output submultiples. Therefore, its output must be a periodic function of the phase difference between input and output, with a period equal to one cycle of the submultiple frequency or N cycles of the input and output frequency. We see that the greater the divider ratio, the greater the range of the phase comparator, in cycles of the input and output frequency.

The sawtooth and sinusoidal comparator functions are shown in Fig. 2. The phase error is measured in radians of the input and output frequency. The gains have been adjusted so that the slopes at zero are identical. This means that the functions have the same small-signal performance at zero quiescent phase error. Note that the peak output of the sawtooth comparator is π times the peak of the sinusoidal comparator.

The sampler and mixer types of sinusoidal comparator are described in the literature.[5]

Since the sawtooth characteristic is not common, we will describe one method of building such a comparator. We assume that the input and output signals are available as short pulses. If the signals are originally sinusoids, the pulses can be obtained from zero crossings. As shown in Fig. 3, these pulses control a flip-flop. The input is sent into the set terminal of the flip-flop and the output is sent into a comple-



(a) SAWTOOTH CHARACTERISTIC

(b) SINUSOIDAL CHARACTERISTIC

Fig. 2 — Characteristics of the sawtooth and sinusoidal phase comparators.

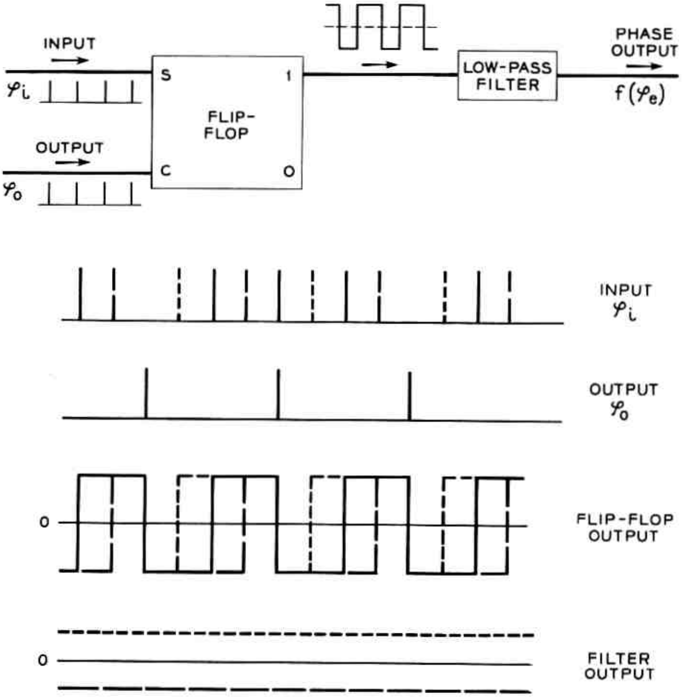


Fig. 3 — Flip-flop sawtooth phase comparator.

ment (or count) terminal. Therefore the time spent in the set state will be the time between the input pulse and the output pulse.

If the flip-flop puts out a positive voltage in the set state and an equal negative voltage in the reset state, the average output voltage will be a linear function of the phase error. The average output will be zero when the pulses are 180° out of phase. Therefore one of the signals should be inverted before pulse forming if the output is desired to be in phase with the input.

If the phase error exceeds $\pm\pi$, the pulses will pass each other. There will be a sudden discontinuity, and the voltage will change quickly from one extreme to the other.

If the input signal is turned off, the flip-flop acts like a binary counter, driven by the output signal. The average output voltage will be zero.

The average voltage will be extracted from the flip-flop output by the low-pass filter. It should have a cutoff frequency low enough to remove signal components near the submultiple frequency.

Since this type of comparator works on zero crossings, its conversion gain is independent of signal amplitude.

A sampler comparator can also have a sawtooth characteristic, if the input has a sawtooth waveform.

Because of the operation of the phase comparator, the phase-controlled oscillator is really a sampled system. E. G. Kimme has shown[6] that the phase-controlled oscillator can be treated as a continuous system if the sampling frequency is so high that its effects are strongly attenuated by the closed loop. We will assume this to be the case throughout this paper.

2.3 *Filter*

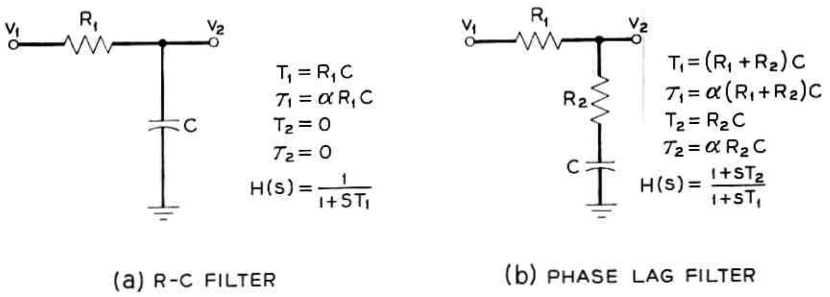
The filter has a low pass characteristic to attenuate fast changes in the phase error due to noise in the input signal. It also helps to smooth out the high frequency component of the phase comparator output. Usually a simple RC filter or a phase lag filter is used, as shown in Fig. 4.

2.4 *Oscillator*

The variable oscillator produces the output signal. When its input voltage v_2 is zero, the output frequency is the design center frequency ω_c. If v_2 is not zero, the output frequency varies in proportion to v_2. Since the important property of the output is the phase, which is the integral of the frequency, the variable oscillator acts like a perfect integrator.

III. OPERATION

Readers who have a background in servo systems may find it helpful to think of the phase-controlled oscillator as a type 1 servo system, such



(a) R-C FILTER

(b) PHASE LAG FILTER

Fig. 4 — Filters.

as a velocity motor with position feedback.[2] The analogy is clear from Fig. 1.

3.1 *Aligned Operation*

Let the frequency of the input signal be identical to the center frequency of the oscillator and let the phase error be zero. Then the input to the oscillator is zero and its frequency will be identical to that of the input.

Now let us quickly advance the input phase by a small amount and continue at the center frequency. There will be a positive error voltage which will increase the output frequency. The output phase will advance until it catches up to the input. The circuit cannot settle down until the output phase is identical to the input phase, because of the integrating action of the oscillator.

3.2 *Mistuning*

Assume that the input frequency increases a little, causing the input phase to continually advance. As before, a positive error signal will result, increasing the output frequency. Therefore the output phase will continually advance. When the circuit settles down to a steady state, the phase error will be constant, and just sufficient to detune the oscillator so that its frequency will be *identical* to the input frequency. The greater the phase-to-frequency gain of the forward path, the less phase error will result from a given input frequency deviation.

3.3 *Jitter*

Now let the average input frequency be constant, but assume that the phase is jittering back and forth. Suppose the jitter is very rapid. Even if there were no filter, the integrating action of the oscillator would smooth out the jitter so that the output would be more stable than the input. The low-pass filter, of course, smooths the error signal before it gets to the oscillator and attenuates the jitter even more.

If the amplitude of the jitter is too great, the phase comparator will go through a discontinuity, and when the circuit settles down again, it will have slipped N cycles of the input, either ahead or behind.

As the rate of jitter decreases, the operation of the loop becomes more complex. Because of the integration, jitter in the oscillator phase lags the fluctuations in its input voltage by 90°. If the low pass filter also has about 90° phase lag at some frequency of jitter, we see that we have positive feedback instead of negative feedback. The open loop phase gain is the ratio of a change in output phase to a change in phase error. If this is large enough at a frequency where we have positive feedback, we can actually have an increase in jitter, or even a jitter oscillation[1] which would destroy the usefulness of the device for most purposes.

If the jitter is very slow compared to the loop time constants, the servo loop will track it, and the jitter will be passed on to the output.

If the jitter is distributed in a wide band, such as that caused by the addition of white noise to a coherent signal, the circuit will respond only to that jitter resulting from noise components near the frequency of the coherent signal. Therefore the circuit can be used to enhance the signal-to-noise ratio of a phase-modulated carrier. This property also allows the circuit to lock on a coherent signal of approximately known frequency although it is surrounded by strong wide-band noise.

3.4 *Phase Modulation*

The error signal v_1 (see Fig. 1) is essentially proportional to the phase modulation of the input at frequencies higher than the circuit can track, and to the frequency modulation at frequencies that can be tracked. The signal at v_2 is filtered to reduce noise. Therefore the circuit can be used as a demodulator of phase or frequency modulated signals in noise.

The circuit can also be used as a phase modulator. The carrier is connected to the input. The modulating voltage is added to the output of the comparator. The feedback tends to keep the oscillator input voltage small. Therefore the comparator output must be nearly equal to the negative of the modulating voltage. This means that the output phase is nearly proportional to the modulating voltage. At high frequencies, the loop gain drops, and these relations are no longer valid.

3.5 *Quieting*

If the input signal is smooth, but the oscillator itself is jittery because of internal noise, the oscillator will be quieted by the feedback, especially at low frequencies where the problem is likely to be most serious.

3.6 *Discontinuities*

We have looked at the small-signal linear performance of the phase-controlled oscillator; now let us examine its operation when it is passing through discontinuities. Suppose we increase the input frequency until the phase error is nearly equal to $+N\pi$, where N is the divider ratio. A small further increase will cause the phase comparator to go through a discontinuity, making the error $-N\pi$. This will start to decrease the oscillator frequency and the error will rapidly return to $+N\pi$, and then jump to $-N\pi$ again. After a short time, the error will settle down to a periodic behavior, with discontinuities at regular intervals. Since the average error must be somewhat less than $+N\pi$, the average output frequency will be somewhat less than the input frequency, and the frequency of the phase error will be the beat frequency between input and output, divided by N.

3.7 *Pull-in*

As the input frequency is reduced in this "flickering" state, the beat frequency decreases. Finally, the phase error does not quite hit a discontinuity at its highest excursion, and the error settles down to a static value. We say the loop has pulled into lock with the input.

Depending on the nature of the filter, there may or may not be hysteresis in the pull-in action. If there is hysteresis the pull-in frequency deviation will be less than the deviation which can be held in lock, once lock has been established.

IV. APPLICATIONS

The phase-locked oscillator has many interesting capabilities, and consequently has found many diverse applications.[1] Some of the functions and examples of use are:

a. Locking a high frequency signal to a submultiple; television sync signals are locked to the power frequency.

b. Locking a strong steady signal to a weak, intermittent signal; television color carrier recovery.

c. Locating and locking on a weak coherent signal in wide-band noise; space communication.

d. Detecting phase or frequency shifts in a signal; space communication.

e. Smoothing a jittery signal; smoothing jitter in a digital signal.

f. Locking a high-power oscillator to a more stable low-power oscillator; microwave generation.

g. Phase modulation of a reference carrier.

h. Frequency synthesis.

Each of these applications requires a different viewpoint in analyzing the circuit. An optimization process for one application may be useless in another. Even an expression such as noise bandwidth may not have the same meaning with a jitter reducing circuit as with a microwave source.

The application we have foremost in mind is that of capturing and smoothing a jittering timing signal for a digital channel. Most of the properties we analyze are chosen for this application. However, we present additional material which is needed for other applications. We have attempted to be explicit in revealing our viewpoint when we define noise bandwidth, figure of merit, etc.

V. QUIESCENT OPERATION

5.1 *Steady-State Error*

If a phase-locked oscillator is synchronized with a signal whose frequency is not identical with the oscillator's center frequency, there must be a steady phase error. The comparator converts this phase error into the voltage required to tune the oscillator so that its output frequency will be identical to the input frequency.

The gain α is the low frequency conversion gain from phase error to frequency (see Fig. 1). It is the change in output frequency (in radians per second) that results from a change in phase error of one radian. The mistuning frequency ω_m is the difference between the input frequency and the oscillator center frequency. Then the steady phase error is

$$\varphi_e = \frac{\omega_m}{\alpha}. \tag{1}$$

The phase error is directly proportional to the mistuning. With a given mistuning, the error may be made as small as desired by increasing the gain, α. However, we shall see that high gain has undesirable effects also.

5.2 *Lock Frequency*

The greatest frequency mistuning that can be locked in synchronism is determined by the maximum output of the phase comparator. At the limit,

$$|\omega_m| = \omega_L = N\pi\alpha. \tag{2}$$

We call ω_L the *lock frequency*.

5.3 *Phase Error Margin*

One of the advantages of the sawtooth comparator over a sinusoidal comparator is that the small-signal performance is independent of the steady mistuning, since the gain does not depend on the phase error. However, mistuning reduces the margin between the steady phase error and the error which will cause a discontinuity. This limits the permissible peak jitter amplitude, if no discontinuities are allowed.

The phase error margin is

$$\varphi_{er} = N\pi - \frac{|\omega_m|}{\alpha}. \tag{3}$$

VI. RESPONSE IN THE LINEAR REGION

As long as the circuit is in synchronism and the phase error does not exceed the bounds of $\pm N\pi$, the phase controlled oscillator acts like a linear feedback system.

6.1 *Phase Response*

From Fig. 1, we see that the forward gain of the loop is the product of the gains of the comparator, filter, and oscillator:

$$\begin{aligned} \mu &= [\alpha_1][\alpha_2 H(s)]\left[\frac{\alpha_3}{s}\right] \\ &= \alpha\frac{H(s)}{s}, \end{aligned} \tag{4}$$

where $\alpha = \alpha_1\alpha_2\alpha_3$.

The feedback is

$$\beta = 1. \tag{5}$$

The response of the output phase to changes in the input phase is given by the familiar negative feedback equation:

$$Y = \frac{\Phi_o}{\Phi_i} = \frac{\mu}{1+\mu\beta} = \frac{\alpha H(s)}{s+\alpha H(s)}. \tag{6}$$

The signals Φ_i and Φ_o are phases of the input and output voltages.

The phase error, as a function of the input phase, is

$$\Phi_e = \Phi_i - \Phi_o = \frac{s}{s+\alpha H(s)}\Phi_i. \tag{7}$$

Notice that we measure phase of the submultiple signals in radians of the original signals.

The filter is usually either an *RC* filter or a phase lag filter, as shown in Fig. 4. For the phase lag, the more general case,

$$H(s) = \frac{1+sT_2}{1+sT_1}, \tag{8}$$

where

$$T_1 \geqq T_2 .$$

In the RC case, $T_2 = 0$.

When we substitute (8) into (6), the transfer ratio becomes

$$Y = \frac{1 + s\,\frac{\tau_2}{\alpha}}{1 + s\,\frac{1 + \tau_2}{\alpha} + s^2\,\frac{\tau_1}{\alpha^2}}, \tag{9}$$

where

$$\tau_1 = \alpha T_1 , \qquad \tau_2 = \alpha T_2 .$$

The phase error response is found from (7):

$$\Phi_e = \frac{\frac{s}{\alpha}\left(1 + s\,\frac{\tau_1}{\alpha}\right)}{1 + s\,\frac{1 + \tau_2}{\alpha} + s^2\,\frac{\tau_1}{\alpha^2}}\,\Phi_i . \tag{10}$$

Note that the denominator of transfer functions (9) and (10) is a second order polynomial, of the form

$$1 + s\,\frac{2\xi}{\omega_n} + s^2\left(\frac{1}{\omega_n}\right)^2,$$

where:

$$\omega_n = \frac{\alpha}{\sqrt{\tau_1}}, \qquad \xi = \frac{1}{2}\,\frac{\tau_2 + 1}{\sqrt{\tau_1}} . \tag{11}$$

Equations (9) and (10) appear frequently in the literature, but have been included here for completeness. Some of the literature[1,2] uses the natural frequency ω_n and the damping ratio ξ as defining parameters of the system. We shall use α, τ_1 and τ_2 more often, because they are more closely related to physical quantities.

Most of the important properties of the phase-controlled oscillator can be expressed as normalized ratios which are independent of α. Therefore we shall present these properties as functions of the two remaining design parameters, τ_1 and τ_2. As an example of our method of presentation, contours of constant damping ratio ξ are shown on a plot of τ_2 vs τ_1 in Fig. 5. We will call this the filter plot. Properties of the filter plot are discussed in Section IX.

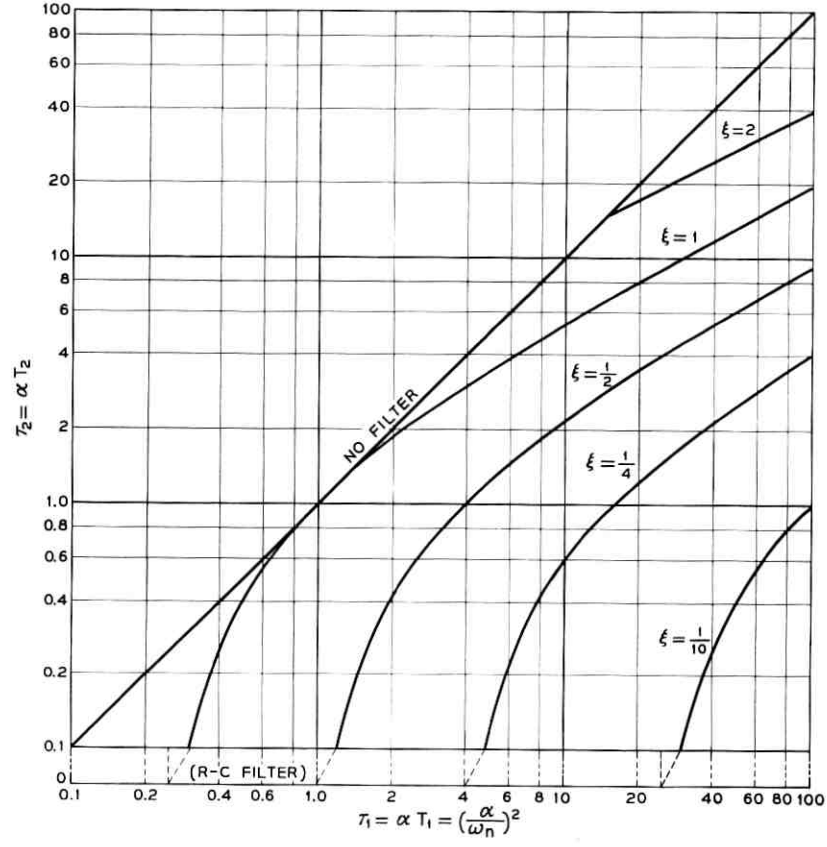


Fig. 5 — Contours of constant damping ratio on the filter plot.

6.2 *Voltage Response*

As we have mentioned, the phase-locked oscillator can be used as a phase modulator by adding a modulating voltage v_M to the error voltage v_1. The response of the output phase is:

$$\Phi_o = \frac{1}{\alpha_1} Y V_M . \tag{12}$$

where Y is given by (6) and (9). Note that we have used V_M for the transform of v_M. Examination of (9) shows that the output phase will follow the input voltage as long as the modulating frequency is low enough, since Y approaches unity as s approaches zero.

If the phase-locked oscillator is used as a demodulator the output can be taken before or after the filter. Therefore we present the response equations for the voltages at each point (see Fig. 1).

$$V_1 = \alpha_1(1 - Y)\Phi_i \tag{13}$$

$$V_2 = \frac{1}{\alpha_3} sY\Phi_i. \tag{14}$$

VII. SMALL-SIGNAL PROPERTIES

The small-signal properties we shall analyze are the response of the output phase to sinusoidal jitter of the input phase, the noise bandwidth, the peak jitter gain, the response to a step change in phase, and the response to a step change in frequency. All of these effects are not pertinent to every system, but each is useful in some of the applications.

7.1 *Sine Wave Jitter Response*

The small signal transfer ratio Y between input phase jitter and output phase jitter was given in (9). For sinusoidal jitter, the squared magnitude (power gain) of $Y(\omega)$ is

$$|\,Y(\omega)\,|^2 = \frac{1 + \left(\frac{\omega}{\alpha}\right)^2 \tau_2^{\,2}}{1 + \left(\frac{\omega}{\alpha}\right)^2 [1 - 2(\tau_1 - \tau_2) + \tau_2^{\,2}] + \left(\frac{\omega}{\alpha}\right)^4 \tau_1^{\,2}}. \tag{15}$$

The phase of $Y(\omega)$ is

$$\theta(\omega) = \tan^{-1}\left(\frac{\omega}{\alpha}\right)\tau_2 - \tan^{-1}\frac{\left(\frac{\omega}{\alpha}\right)(1 + \tau_2)}{1 - \left(\frac{\omega}{\alpha}\right)^2 \tau_1}. \tag{16}$$

The jitter attenuation curves for several sets of filter parameters are plotted in Fig. 6.

In Case I, no filter, we have simply an integrator with unity feedback around it. At low frequencies the jitter is not attenuated; at high frequencies there is a 6 db per octave roll-off. When an RC filter is added, the additional high frequency attenuation produces a 12 db per octave roll-off. When the filter time constant is very large, the phase shift in

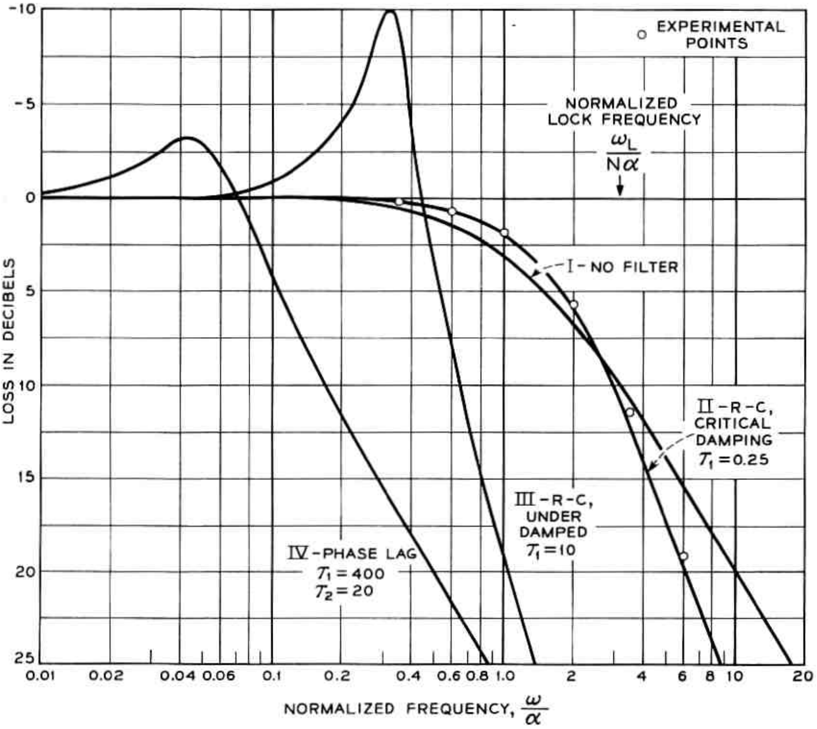


Fig. 6 — Jitter attenuation with various filter parameters.

the forward loop results in positive feedback, and causes a region where jitter is amplified.

When a phase lag filter is used, the second break point caused by the resistor in series with the capacitor can be used to stabilize the feedback loop and reduce the peak jitter gain. Since the attenuation of the phase lag filter is constant at high frequencies, the final slope is 6 db per octave.

7.2 *Noise Bandwidth*

One of the functions of a phase-controlled oscillator is to reduce noise. In the absence of better information, it is usual to assume that somewhere in the system the noise is white and Gaussian. Since most of the noise at the output is usually restricted to a narrow band by the filtering action of the circuit, it is convenient to express the amount of noise that remains as the bandwidth of an ideal filter (i.e., rectangular filter) that would pass the same mean square noise. The familiar formula for computing noise bandwidth is

$$B = \int_0^\infty | G(\omega) |^2 \, d\omega, \tag{17}$$

where $G(\omega)$ is the normalized transfer function between noise input and noise output, and B is in radians per second. The transfer function which is used for $G(\omega)$ will depend on where the noise input and output are, and this in turn will depend on the application.

When the phase-controlled oscillator is used to clear up jitter in digital signals, the appropriate transfer function is Y, the ratio of output phase shift to input phase shift, as given in (9). We will call the noise bandwidth of Y the jitter bandwidth B_j. When we substitute (9) into (17), we have

$$\frac{B_j}{\pi\alpha} = \frac{1}{2} \frac{1 + \frac{\tau_2^2}{\tau_1}}{1 + \tau_2}. \tag{18}$$

We recall that $N\pi\alpha$ is the lock frequency. An increase in N increases the lock frequency without changing B_j.

For no filter, or for any RC filter, the normalized jitter bandwidth is $\frac{1}{2}$. For τ_2^2/τ_1 much greater than 1, the normalized jitter bandwidth approaches $\frac{1}{2}(\tau_2/\tau_1)$. The jitter bandwidth is shown on the filter plot in Fig. 7.

With a sawtooth phase comparator, the jitter bandwidth is independent of the mistuning. This is not true of the sinusoidal comparator. The jitter bandwidth for the sinusoidal case is

$$\frac{B_j}{\pi\alpha} = \frac{1}{2} \cos \varphi_e \frac{1 + \frac{\tau_2^2}{\tau_1} \cos \varphi_e}{1 + \tau_2 \cos \varphi_e} \tag{19}$$

where α is the gain at zero error.

Equation (19) can be obtained from (18) by replacing α by (α cos φ_e), the small signal gain at a quiescent phase error φ_e. Note that α is a factor in τ_1 and τ_2. Notice that the jitter bandwidth for the sinusoidal comparator decreases as the mistuning (and therefore φ_e) increases.

Now let us consider the effect of interference due to broad-band noise added to the input signal. To justify a small signal analysis, we must assume that filtering limits the total energy of the interference, to keep it well below the signal level. However, we assume that the filtered noise is essentially flat in a band around the signal which is much wider than the interference noise bandwidth which we shall derive.

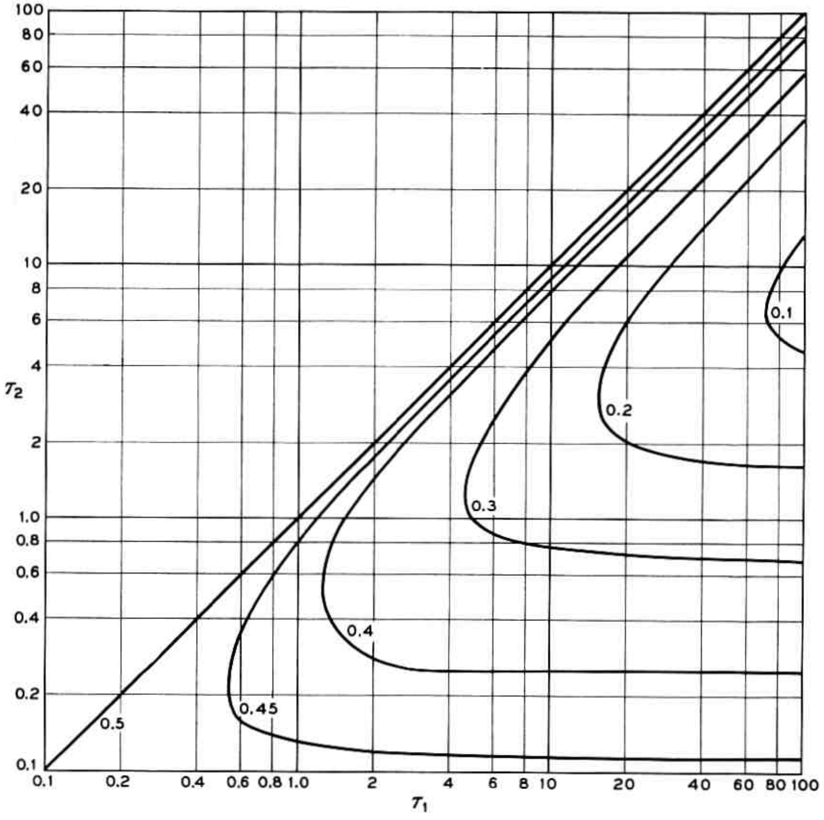


Fig. 7 — Contours of normalized jitter noise bandwidth on the filter plot.

The effect of interference depends strongly on the type of phase comparator in the system. We shall analyze the linear zero-crossing case and the sinusoidal mixer or sampler case.

Interference noise disturbs both the phase and the amplitude of the input signal. When a zero-crossing comparator is used, only the phase disturbance is detected. If the noise power density is $\overline{v_n^2}$ (volts^2 per radians per second) and the input sinusoid has a peak v_i, the jitter "power" density for phase in radians is $\overline{(v_n/v_i)^2}$ (radians^2 per radians per second).

The output jitter will be

$$\overline{\varphi_o^2} = B_j \frac{\overline{v_n^2}}{v_i^2} \tag{20}$$

The effect of broad band input noise is quite different when a sinusoidal sampler or mixer phase comparator is used. The following discussion assumes that the reader is familiar with the literature of the sinusoidal comparator. With this type of comparator, noise at the input produces a voltage at the output of the comparator which is independent of the amplitude and phase of the input signal. The noise density of the comparator output voltage is $(\alpha_1/v_i)^2\overline{v_n^2}$ where v_i is the *expected* peak signal amplitude used in computing the expected α_1 at zero error (if no limiting is used with this type of comparator, the gain depends on the signal amplitude). When the comparator is connected in a feedback loop, the appropriate method of analysis is to consider the interference noise injected at the *output* of the phase comparator. The appropriate transfer ratio is that previously used for modulation in (12).

The interference bandwidth B_i can be found by substituting (12) into (17);

$$\frac{B_i}{\pi \propto} = \frac{1}{2}(\cos \varphi_e)^{-1} \frac{1 + \left(\frac{\tau_2{}^2}{\tau_1} \cos \varphi_e\right)}{1 + (\tau_2 \cos \varphi_e)} \tag{21}$$

This is the noise bandwidth given by Rey.[1]

Notice that the interference bandwidth B_i increases as the phase error increases, while the jitter bandwidth B_j decreases. The reason for the difference is that the sampler and mixer comparators are sensitive to the amplitude of the input signal as well as the phase.

Now we can compare the output phase noise performance of the linear zero-crossing comparator with the sinusoidal sampler or multiplier type. If they have the same gain at zero error, they will have the same response to jitter and interference at zero error. In the presence of mistuning, however, the sinusoidal comparator will be more sensitive to interference and less sensitive to jitter while the linear comparator will not change.

When the phase-controlled oscillator is used as a demodulator, still another definition of noise bandwidth is required. If we take the output signal after the filter, which cuts off some of the noise, and assume that interference noise is added to the input signal, we have for the zero-crossing detector,

$$v_2 = \left[\frac{s}{\alpha}\, Y\right] \frac{\alpha_1\alpha_2}{v_1}\, V_n\,. \tag{22}$$

By substituting the expression in brackets into (17), we can find the demodulator noise bandwidth, B_D. This bandwidth is not finite for the phase lag filter, because the transfer ratio does not approach zero at high frequencies. Therefore higher order filters are desirable for this application.

7.3 *Peak Jitter Gain*

We have shown in Fig. 6 that it is possible for the jitter transfer ratio to be greater than unity. In most systems, this is not very harmful. However, where phase-controlled oscillators are connected in cascade, gain can be very troublesome.

We can find the peak gain $|\hat{Y}|$ by examining (15) for its maximum. The frequency at the peak is

$$\left(\frac{\hat{\omega}}{\alpha}\right)^2 = \frac{1}{{\tau_2}^2}\left[\left(1 + \left(\frac{\tau_2}{\tau_1}\right)^2 [2(\tau_1 - \tau_2) - 1]\right)^{\frac{1}{2}} - 1\right]. \tag{23}$$

A. J. Goldstein[4] has shown that the square of the peak magnitude can be written

$$|\hat{Y}|^2 = \frac{1}{1 - {\tau_1}^2 \left(\frac{\hat{\omega}}{\alpha}\right)^4}. \tag{24}$$

An examination of (23) shows that there is no peak, and the gain is never greater than unity if

$$\tau_1 - \tau_2 < \tfrac{1}{2}. \tag{25}$$

The peak gain is shown on the filter plot in Fig. 8.

7.4 *Response to a Step Change in Phase*

Fast phase changes can occur because of quick changes in the transmission path or because the signal has been deliberately modulated. When a step in phase occurs there is a sudden change in the phase error, since the phase of the oscillator cannot change instantaneously. The error signal controls the oscillator so that the error returns eventually to its quiescent value.

To act like a step change, the phase shift does not have to be instantaneous, as long as the rise time is much less than the shortest time constant of the phase-controlled loop. Therefore if the phase comparator works from a subharmonic of the input frequency the amplitude of the phase change can be several input periods, as long as the change is slow enough for the subharmonic generator (counter, etc.) to follow, but faster than the loop time constants.

If a counter is used as a subharmonic generator, an error in the counter,

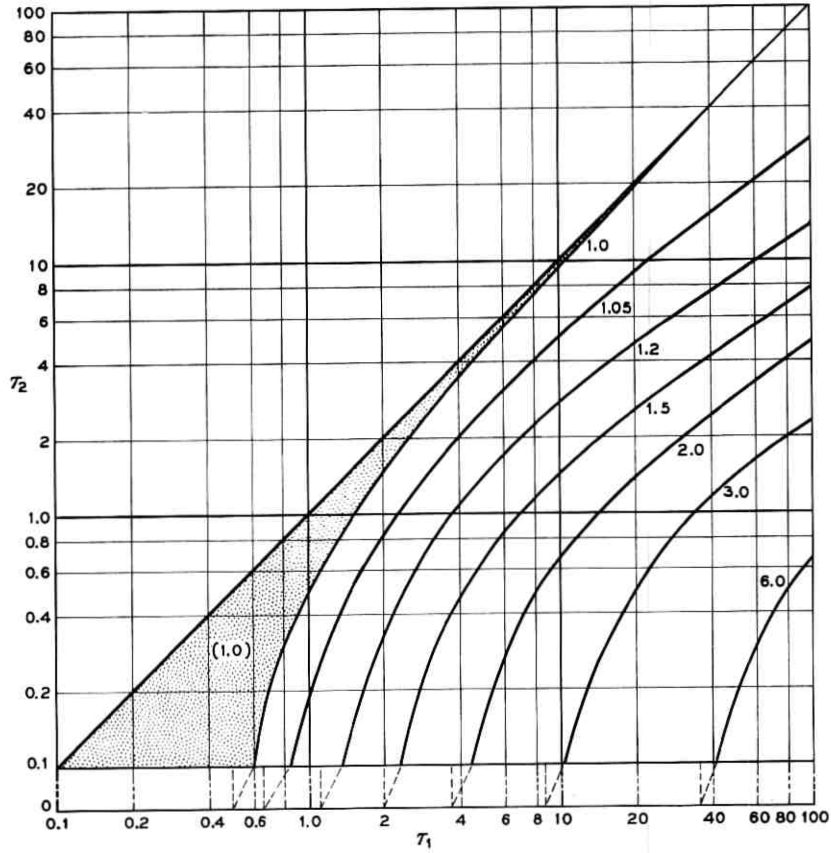


Fig. 8 — Contours of peak jitter gain on the filter plot.

or an extraneous pulse introduced into the counter, will act like a step change in input phase.

The response of the phase error to the phase input is given in (10). When the input phase is a step of amplitude $\Delta\varphi_i$, the time response of the phase error can be shown to be

$$\varphi_e = \Delta\varphi_i e^{-\xi\omega_n t} \left[\cosh\left(\sqrt{\xi^2 - 1}\,\omega_n t\right) - \frac{\xi - \dfrac{\omega_n}{\alpha}}{\sqrt{\xi^2 - 1}} \cdot \sinh\left(\sqrt{\xi^2 - 1}\,\omega_n t\right) \right]. \qquad (26)$$

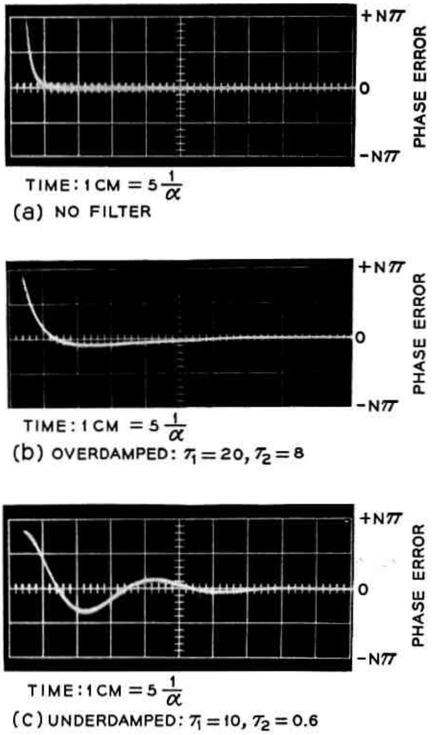
For the underdamped case ($\xi < 1$) the hyperbolic functions in (26) become trigonometric functions. The damping ratio ξ and the natural frequency ω_n have been defined in (11).

At $t = 0$, just after the step, we see that the phase error equals the change in input phase. If we examine the initial derivative of (26), we find that it is never positive. This means that the phase error will never exceed its initial value.

Some examples of the phase error response to a step change in phase are shown in Fig. 9.

7.5 *Response to a Step Change in Frequency*

A sudden change in frequency can occur because of a change from one source to another, because of malfunction, or because the signal has been modulated. When a frequency step occurs, the error signal builds up until the oscillator frequency catches up to the input frequency,



TIME: 1 CM = $5\frac{1}{\alpha}$

(a) NO FILTER

TIME: 1 CM = $5\frac{1}{\alpha}$

(b) OVERDAMPED: $\tau_1 = 20, \tau_2 = 8$

TIME: 1 CM = $5\frac{1}{\alpha}$

(C) UNDERDAMPED: $\tau_1 = 10, \tau_2 = 0.6$

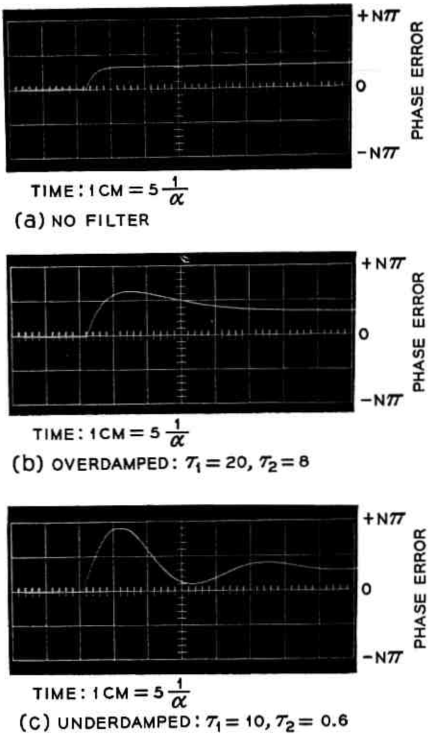
Fig. 9 — Scope traces of the response of the phase error to a step change in phase of the input.

leaving a static change in phase error. If a low-pass filter is used between the phase comparator and the oscillator, the transient phase error can have a peak value much greater than the quiescent phase change.

Let us assume a frequency change $\Delta\omega_i$. This is equivalent to a ramp phase input, $\Delta\omega_i t$. We can use (10) to find the response of the phase error:

$$\varphi_e = \frac{\Delta\omega_i}{\alpha}\left\{1 - e^{-\xi\omega_n t}\left[\cosh\left(\sqrt{\xi^2 - 1}\,\omega_n t\right) - \frac{\dfrac{\alpha}{\omega_n} - \xi}{\sqrt{\xi^2 - 1}}\sinh\left(\sqrt{\xi^2 - 1}\,\omega_n t\right)\right]\right\}. \tag{27}$$

Some examples of the phase error response to a step change in input frequency are shown in Fig. 10.



TIME: 1 CM = $5\frac{1}{\alpha}$

(a) NO FILTER

TIME: 1 CM = $5\frac{1}{\alpha}$

(b) OVERDAMPED: $\tau_1 = 20$, $\tau_2 = 8$

TIME: 1 CM = $5\frac{1}{\alpha}$

(c) UNDERDAMPED: $\tau_1 = 10$, $\tau_2 = 0.6$

Fig. 10 — Scope traces of the response of the phase error to a step change in frequency of the input.

The peak phase error is of particular interest. For the overdamped case it is

$$\hat{\varphi}_e = \frac{\Delta\omega_i}{\alpha}\left[1 + (\tau_1 - \tau_2)^{\frac{1}{2}} \exp\left(-\frac{\xi}{\sqrt{\xi^2 - 1}} \tanh^{-1} \frac{\sqrt{\xi^2 - 1}}{\xi - \frac{\omega_n}{\alpha}}\right)\right]. \quad (28)$$

For the underdamped case, the inverse hyperbolic tangent is replaced by the inverse trigonometric tangent. The value of this angle is between zero and π. An expression closely related to (28) has been derived by R. D. Barnard,[7] as a capture condition.

A large value of τ_1 can result in overshoot which is many times the quiescent phase error. This means that the response of such a system to a sudden frequency shift looks like a pulse. This characteristic is useful in demodulation of a frequency shift signal.

The large overshoot can throw the loop out of synchronism if it exceeds the capacity of the phase comparator. This effect will be discussed more fully in Section 8.5.

The normalized peak phase error is shown on the filter plot in Fig. 11.

VIII. LARGE-SIGNAL PROPERTIES

We have examined the operation of the synchronized phase-controlled oscillator when the error is within the range of the phase comparator. For this "small-signal" case, the problem was completely linear. When the circuit is not in synchronism, or when disturbances of the input signal are large enough to produce a phase error which exceeds the range of the comparator, discontinuities are present in the output and the problem becomes nonlinear. Despite this difficulty, we have been able to analyze certain large-signal properties of the phase-controlled loop with a sawtooth comparator. These are the pull-in frequency, the seize frequency, the settling time, the maximum allowed frequency shift, and the effect of certain types of jitter on large-signal performance.

8.1 *Pull-in Frequency*

A very important property of the phase-locked loop is the range of frequencies that can pull the oscillator into synchronism. In general, this range is smaller than the range of frequencies which can be held in lock. When the system is not synchronized, the phase comparator goes through periodic discontinuities, which prevent the loop from synchronizing. Whether or not a loop will pull a given frequency into lock depends on the past history of the loop and the jitter of the signal.

We define the *pull-in frequency* as the maximum steady mistuning

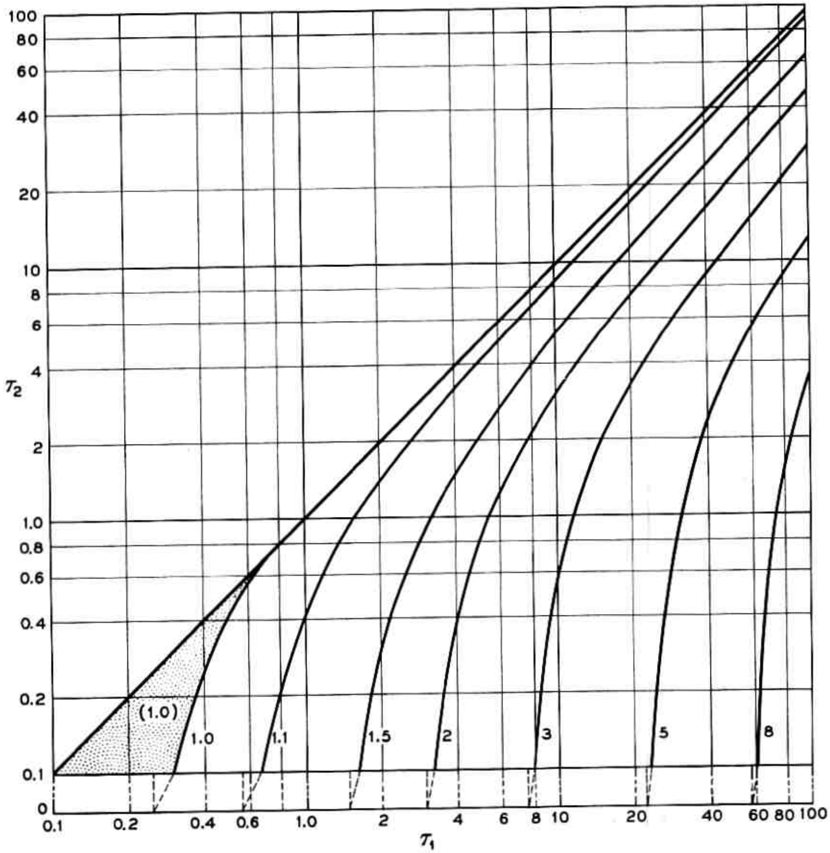


Fig. 11 — Contours of normalized peak phase error caused by a step change in frequency.

of the input frequency that will always pull the circuit into synchronism. Frequencies outside of the pull-in range but inside the lock range may or may not be pulled in, depending on the initial conditions.

We can determine the pull-in frequency experimentally by mistuning the input beyond the lock frequency and then slowly reducing the mistuning until the circuit locks. When the mistuning exceeds the lock range, there are frequent discontinuities in the phase error; it appears to "flicker." As the mistuning is slowly decreased, the flicker rate decreases.

When the mistuning is brought down to the pull-in frequency, the flicker mode becomes unstable. With the mistuning then held constant, just under the pull-in frequency, the phase error trajectory from discontinuity to discontinuity slowly changes as shown in Fig. 12. Finally, the error misses a discontinuity and synchronism is achieved.

The pull-in frequency, then, is the mistuning for which the stable asynchronous mode disappears. For lower values of mistuning, the circuit must eventually reach a synchronous condition since there is no asynchronous solution.

A. J. Goldstein[4] has found an exact answer for the pull-in frequency ω_p;

$$\frac{\omega_p}{N\pi\alpha} = \frac{1 - D}{\tanh \frac{1}{2}\xi\omega_n T_0} + (D) \tanh \tfrac{1}{2}\xi\omega_n T_0 . \tag{29}$$

where T_0, the critical flicker period, is the smallest positive solution of

$$\sqrt{\tau_1}\,\sqrt{\xi^2 - 1}\,\frac{\tanh \frac{1}{2}\xi\omega_n T_0}{\tanh \frac{1}{2}\sqrt{\xi^2 - 1}\;\omega_n T_0} = \sqrt{\tau_1}\,\xi - \frac{\tau_1}{\tau_2}\left(1 - \sqrt{1 - \frac{\tau_2}{\tau_1}}\right) = c_1 ,$$

and D is given by

$$D = \frac{c_1(\sqrt{\tau_1}\,\xi - 1) - \tau_1(\xi^2 - 1)}{c_1^2 - \tau_1(\xi^2 - 1)} .$$

For the underdamped case (damping ratio $\xi < 1$) the hyperbolic tangent is replaced by the trigonometric tangent.

A. J. Goldstein[4] has used a digital computer to evaluate (29). The data is presented on the filter plot in Fig. 13.

We can see from (29) that the pull-in frequency is directly proportional to the lock frequency $N\pi\alpha$, for a given set of parameters τ_1 and τ_2. We will call $\omega_p/N\pi\alpha$ the pull-in to lock ratio, or the relative pull-in.

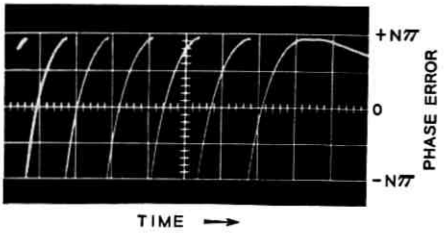


Fig. 12 — Scope trace of the phase error after the mistuning is brought just below the pull-in frequency. The flicker mode becomes unstable.

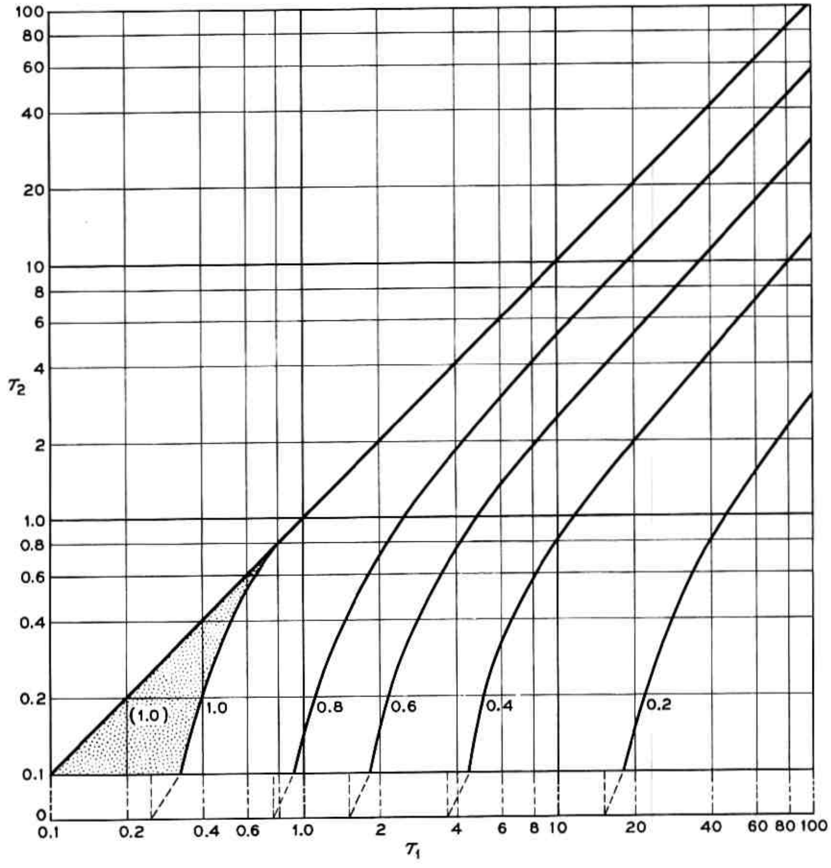


Fig. 13 — Contours of the pull-in to lock ratio on the filter plot.

We have shown that the small-signal properties of the phase-controlled oscillator are completely specified by the parameters τ_1, τ_2 and α. Therefore, for constant small-signal performance (such as noise bandwidth), the pull-in range is proportional to the count ratio N. We can get any pull-in frequency we wish by using a large enough count ratio.

There are two limitations on increasing the count ratio. The first is economy; high counts require more equipment. The second is theoretical. The comparator supplies data only once every period of the submultiple frequency. For our analysis to be valid, the submultiple frequency should be much higher than the cutoff frequency of the forward path, which is of the order of ω_n. This limits the maximum count.

For $\tau_2 \gg 1$, and $\tau_2/\tau_1 < 0.5$, the pull-in frequency approaches

$$\frac{\omega_p}{N\pi\alpha} \cong \frac{2}{\sqrt{3}} \sqrt{\frac{\tau_2}{\tau_1}}. \tag{30}$$

It is interesting to compare the pull-in frequency of a sawtooth comparator to that of a sinusoidal comparator[1] with the same gain at zero error. The normalized pull-in frequencies for both types of comparator are shown in Fig. 14, for a damping ratio ξ of $\frac{1}{2}$.

Fig. 14 shows that the sawtooth phase detector has a pull-in frequency at least twice that of a sinusoidal detector which has the same small-signal performance.

8.2 *Figure of Merit*

In most applications, a large pull-in frequency and a small noise bandwidth are desired. Unfortunately, these requirements are antagonistic, since a small noise bandwidth means that the loop cannot react to a rapidly flickering phase error. Examination of the formulas for pull-in (29) and jitter noise bandwidth (18) shows that both are proportional to the gain, α. Therefore a natural figure of merit is the ratio of pull-in frequency to the jitter noise bandwidth:

$$M = \frac{\omega_p}{B_j}. \tag{31}$$

Since the pull-in frequency is proportional to the count ratio N while

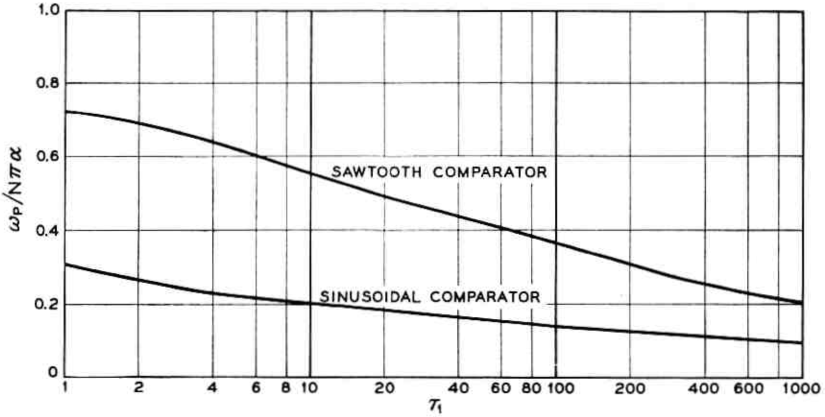


Fig. 14 — Normalized pull-in of the sinusoidal and sawtooth comparators for a damping ratio of 1/2.

the noise bandwidth is independent of N, the figure of merit is proportional to N. This means that we can get as large a value of pull-in as we wish for a given noise bandwidth, if we are willing to use a large count ratio.

The normalized figure of merit M/N is shown on the filter plot in Fig. 15.

D. Richman[8] has defined a different figure of merit, since he wished to compromise between noise bandwidth and gain. His figure of merit is equivalent to our normalized noise bandwidth (18), plotted in Fig. 7.

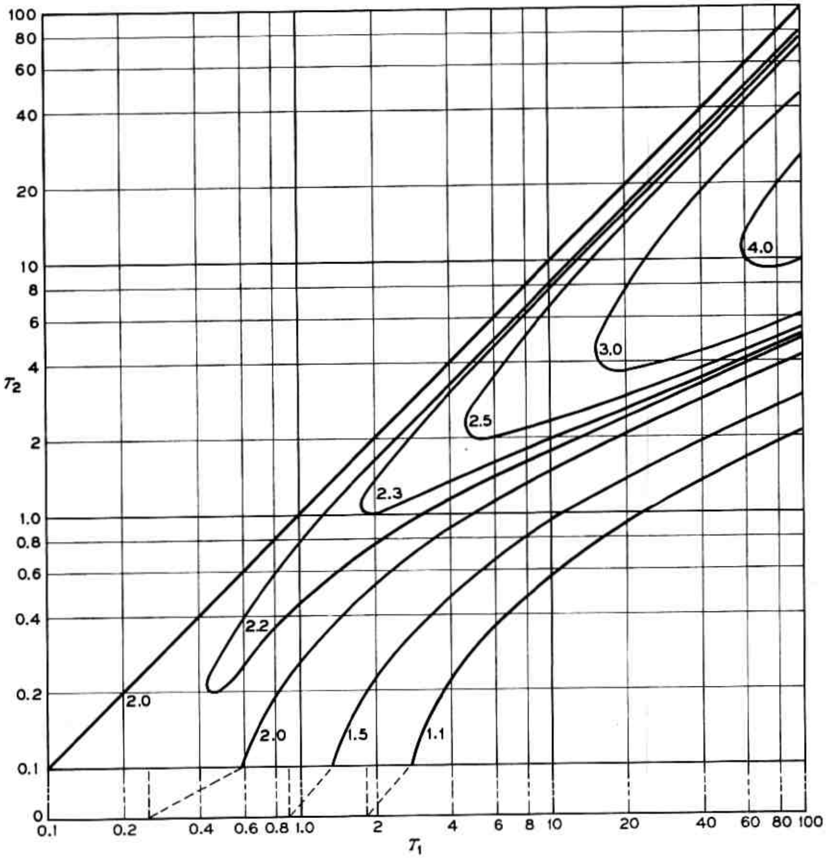


Fig. 15 — Contours of normalized figure of merit on the contour plot. The normalized figure of merit is the ratio of the pull-in to the noise bandwidth, divided by the count ratio N.

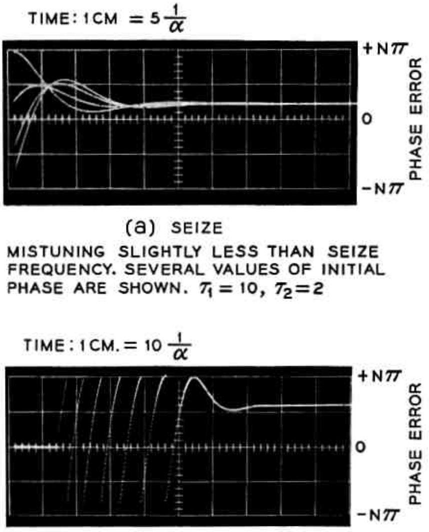
8.3 *Seize Frequency*

As long as the mistuning of a signal is less than the pull-in frequency, we can be sure the circuit will lock; but it may flicker for a long while before it does.

For some applications, it is important that the circuit synchronize immediately on a signal that has just started, without flickering through discontinuities. We define the *seize frequency* ω_s as the maximum mistuning of a suddenly connected signal that cannot cause a discontinuity after the initial phase jump (see Fig. 16).

We have described a phase comparator which produces a zero error signal when there is no input signal. With such a device, the effect of suddenly connecting a signal is equivalent to a step phase shift of an arbitrary value between $-N\pi$ and $+N\pi$ and a step change in frequency equal to the mistuning of the signal ω_m .

In the marginal case, the phase error between the oscillator and the signal at the instant of connection is nearly $N\pi$. The seize frequency is the value of mistuning for which the initial derivative of the phase error



(a) SEIZE
MISTUNING SLIGHTLY LESS THAN SEIZE FREQUENCY. SEVERAL VALUES OF INITIAL PHASE ARE SHOWN. $\tau_1 = 10$, $\tau_2 = 2$

(b) PULL-IN
MISTUNING GREATER THAN SEIZE BUT LESS THAN PULL-IN FREQUENCY. NOTE THE DISCONTINUITIES. $\tau_1 = 10$, $\tau_2 = 2$

Fig. 16 — Scope traces of the phase error during capture.

is zero, so that no discontinuity results. It is easily shown that

$$\frac{\omega_s}{N\pi\alpha} = \frac{\tau_2}{\tau_1}. \tag{32}$$

Note that a circuit with an RC filter ($\tau_2 = 0$) may go through a discontinuity for any nonzero mistuning, if the initial phase shift is large enough.

According to Richman[8] the seize frequency for the sinusoidal comparator is $\alpha(\tau_2/\tau_1)$. As indicated by a comparison with (32), the seize frequency in general is simply τ_2/τ_1 times the lock frequency.

8.4 *Settling Time*

The settling time is the time required for the phase error to settle nearly to its steady state value after a change in input conditions. If no discontinuity occurs, the settling time t_s may be estimated to be the time at which the damping term $e^{-\xi\omega_n t_s}$ [in (26) and (27)] decays to 0.1. Then, substituting for ξ according to (11),

$$t_s = \frac{4.6}{\tau_2 + 1} T_1 . \tag{33}$$

If a discontinuity is crossed, an additional time will be required to allow the flickering to die out. During each flicker period a small charge is added to the filter capacitor, bringing the average output frequency of the oscillator closer to the input frequency. Finally, the circuit locks.

The flicker time for a given mistuning depends on the initial conditions, especially on the initial capacitor voltage. For the special case of a suddenly connected signal (initial capacitor voltage zero), D. Richman has derived[8] an approximation for t_F, the time in the flicker state, for the sinusoidal comparator.

He assumes that the capacitor voltage does not change appreciably during a single flicker period; in effect, he replaces the capacitor with a variable battery. Further, Richman neglects the effect of the initial phase. By applying his methods to the sawtooth comparator, we obtain:

$$\frac{t_F}{T_2} = \int_{\omega_m/\omega_L}^{\tau_2/\tau_1} \frac{\dfrac{\tau_1}{\tau_2}\, d\left(\dfrac{\omega_I}{\omega_L}\right)}{\dfrac{\omega_m}{\omega_L} - \left(\dfrac{\tau_1}{\tau_2}\dfrac{\omega_I}{\omega_L}\right) + \left(1 - \dfrac{\tau_2}{\tau_1}\right)\left[\coth^{-1}\left(\dfrac{\tau_1}{\tau_2}\dfrac{\omega_I}{\omega_L}\right)\right]^{-1}}, \tag{34}$$

where ω_I is an "instantaneous mistuning" parameter introduced by Richman.

Equation (34) is a good approximation for $\tau_2 \gg 1$ and $t_F \gg t_s$.

To carry out the integration, we must use numerical methods. We have plotted t_F/T_2 against ω_m/ω_L for various values of τ_2/τ_1 in Fig. 17. Experimental results are also shown in Fig. 17.

Note that t_F goes to zero as ω_m approaches the seize frequency and to infinity as ω_m approaches the pull-in frequency. If a short pull-in time is important, the mistuning frequency should not be allowed to approach the pull-in frequency.

8.5 *Maximum Frequency Shift*

Consider a phase-controlled oscillator which is locked on an input which is frequency modulated by a digital signal (frequency-shift key-

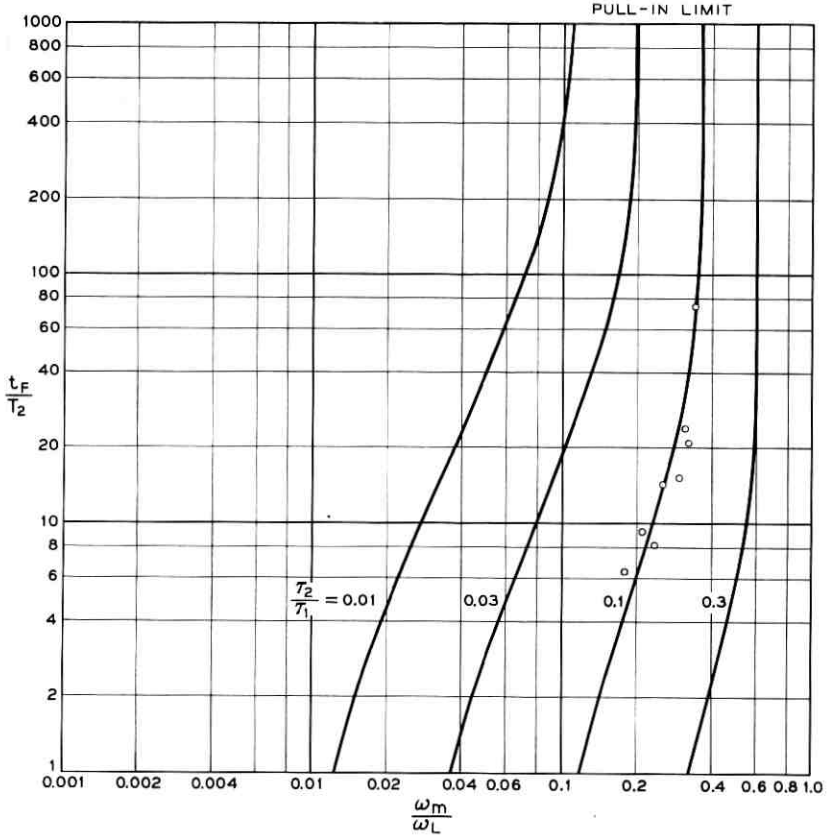


Fig. 17 — Flicker time during pull-in. The time is zero for mistuning less than the seize frequency and infinity for mistuning greater than the pull-in frequency.

ing). Let us find the maximum frequency shift that will not cause the phase error to cross a comparator discontinuity. We assume that the center frequency of the oscillator is set midway between the two signal frequencies. We further assume that the time constants of the phase-controlled oscillator are much smaller than the maximum time between shifts, so that the circuit may be in steady state before the next shift occurs.

We will consider the case of a sudden increase of frequency $\Delta\omega_i$. The initial phase error is $-(\Delta\omega_i/2\alpha)$. The maximum allowed phase error is $+N\pi$. Thus the peak change in phase error $\hat{\varphi}_e$, caused by the maximum allowable change of input frequency $\Delta\hat{\omega}_i$, is

$$\hat{\varphi}_e = N\pi + \frac{\Delta\hat{\omega}_i}{2\alpha}. \tag{35}$$

The error $\hat{\varphi}_e$ has been given in (28). Solving (35) for $\Delta\hat{\omega}_i$ in terms of $\hat{\varphi}_e/\Delta\hat{\omega}_i$, we have:

$$\frac{\Delta\hat{\omega}_i}{N\pi\alpha} = \frac{1}{\dfrac{\alpha\hat{\varphi}_e}{\Delta\hat{\omega}_i} - \dfrac{1}{2}}. \tag{36}$$

In the presence of mistuning, $N\pi$ in (35) and (36) is replaced by the margin φ_{er}, given in (3). Values of $\alpha\hat{\varphi}_e/\Delta\omega_i$ have been plotted in Fig. 11.

8.6 *Effective Comparator Characteristic in the Presence of Fast Jitter*

One of the functions of a phase-locked oscillator is to produce a steady output despite jitter and noise in the input signal. Therefore, we can expect that a major part of the phase comparator output will have frequencies much higher than the oscillator can follow. In such a situation only the low frequency component of the comparator output is significant in controlling the circuit.

The low-frequency component of the comparator output is the time average taken over a time interval which is longer than the period of the predominant jitter, but shorter than the response time of the circuit. The following analysis assumes that such an intermediate time range exists; i.e., that there is very little jitter whose frequency is low enough to cause the circuit to respond.

Let us write the phase error as the sum of a low-frequency component φ_{e0} and a fast jitter component φ_{ej}. Then the instantaneous output of the phase comparator is $f(\varphi_{e0} + \varphi_{ej})$. The average output of the comparator is

$$\overline{v_1} = \frac{1}{T_a} \int_0^{T_a} f(\varphi_{e0} + \varphi_{ej})\, dt, \tag{37}$$

where T_a is the averaging time.

Let us define an effective comparator characteristic in the presence of jitter:

$$\overline{v_1} = f_j(\varphi_{e0}). \tag{38}$$

This new characteristic governs the response of the circuit to slow phase changes in the presence of fast jitter.

Now we assume that the time of integration is such that the time spent at each value of φ_{ej} is proportional to the probability density of φ_{ej} at the value. For random processes, this requires that T_a be much greater than the correlation time of the process. If φ_{ej} is periodic, it is sufficient that T_a be equal to one period.

If this assumption is valid, we can replace the time integral (37) by an ensemble integral:

$$f_j(\varphi_{e0}) = \int_{-\infty}^{+\infty} f(\varphi_{e0} + \varphi_{ej}) p(\varphi_{ej})\, d\varphi_{ej}\,, \tag{39}$$

where $p(\varphi_{ej})$ is the probability density of the jitter.

Equation (39) represents a smoothing operation by the jitter probability function upon the comparator characteristic. If the density function has even symmetry, (39) is analogous to the general filter equation

$$v_{\text{out}}(t) = \int_{-\infty}^{+\infty} v_{\text{in}}(t - \tau) i(\tau)\, d\tau \tag{40}$$

where $i(\tau)$ is the impulse response of a hypothetical filter.

The effective sawtooth comparator characteristic for Gaussian, sinusoidal, and square wave jitter is shown in Fig. 18. These photographs were obtained by opening the phase-controlled oscillator loop and allowing the oscillator to free run. This means that the average phase error φ_{e0} increases linearly with time. The phase comparator output was passed through a low-pass filter to obtain $f_j(\varphi_{e0})$.

Note that jitter always decreases the peak comparator output voltage.

For Gaussian noise, we can evaluate (39) by neglecting the possibility of jitter crossing two or more discontinuities. Then the effective comparator characteristic for $(-N\pi < \varphi_{e0} < +N\pi)$ is

$$f_j(\varphi_{e0}) = \varphi_{e0} + N2\pi\left[\int_{-\infty}^{-x_1} \frac{1}{\sqrt{2\pi}} e^{-(x^2/2)}\, dx - \int_{x_2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x^2/2)} dx\right] \tag{41}$$

where

$$x_1 = \frac{N\pi + \varphi_{e0}}{(\varphi_e)\text{rms}},$$

$$x_2 = \frac{N\pi - \varphi_{e0}}{(\varphi_e)\text{rms}}, \quad \text{and}$$

$(\varphi_e)rms$ is the root mean square phase error due to fast jitter.

The peak effective comparator output for Gaussian jitter is plotted in Fig. 19.

The effective comparator function for the sinusoidal comparator is very easy to find, using the filter analogy:

$$\begin{aligned} f(\varphi_e) &= \sin \varphi_e, \\ f_j(\varphi_{e0}) &= e^{-\frac{1}{2}(\varphi_e)\text{rms}^2} \sin \varphi_{e0}. \end{aligned} \tag{42}$$

Therefore the effect of high-frequency Gaussian jitter for the sinusoidal comparator is simply to reduce the loop gain.

In general, the presence of fast jitter causes a deterioration of large

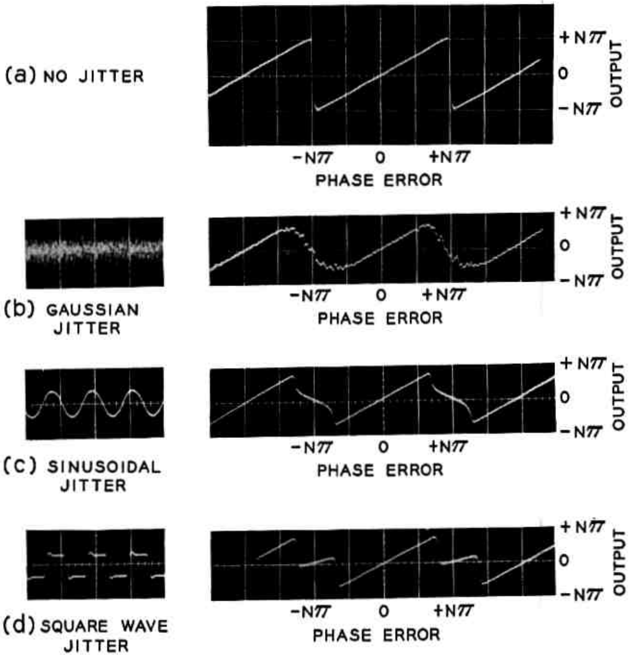


Fig. 18 — The phase comparator characteristic in the presence of fast jitter.

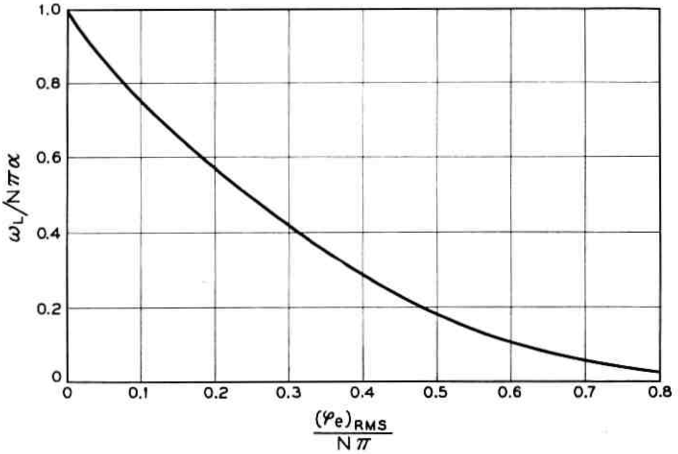


Fig. 19 — Normalized lock frequency (peak comparator output) in the presence of fast Gaussian jitter.

signal performance. For example the lock frequency depends directly upon the peak comparator output, which decreases as jitter increases.

8.7 *False Synchronization Mode*

As shown in Fig. 18(d), the sawtooth comparator characteristic can have a region with positive slope centered on an average phase error $N\pi$. This means that the circuit can synchronize in this region instead of the region near zero error. In this false mode the jitter continually crosses and recrosses the discontinuity.

Fortunately, this undesirable mode is possible only for certain types and amplitudes of jitter. We can test for the possibility of the false mode by examining the slope of $f_j(\varphi_{e0})$ at $N\pi$. We can write $f(\varphi_e)$ in the vicinity of $N\pi$ as $\varphi_e - 2N\pi U(\varphi_e - N\pi)$, where U is the unit step function. Substituting this in (39) and taking the derivative, we have

$$\left.\frac{df_j(\varphi_{e0})}{d(\varphi_{e0})}\right|_{N\pi} = 1 - 2N\pi p(0), \qquad (43)$$

where $p(0)$ is the probability density of φ_{ej} at 0. Therefore the false mode is possible when $p(0) < 1/2N\pi$.

For square wave jitter, $p(0) = 0$. Therefore the false mode is always possible.

For sine wave jitter with an amplitude $A, p(0) = 1/A\pi$. Therefore the false mode is possible only when $A > 2N$. Since the comparator can accommodate only jitter error amplitudes less than πN in the normal mode, we are not likely to encounter sinusoidal jitter large enough to support the false mode.

It can be shown by using the filter analogy that Gaussian jitter cannot produce the false mode; the slope of $f_j(N\pi)$ is always negative. The $p(0)$ criterion is not applicable in the case of Gaussian jitter because more than one discontinuity is involved.

We see that the false mode need be considered only for signals with jitter such that $p(0)$ is very small.

Even if the false mode has been established, a lull in the jitter will cause the circuit to jump to the normal mode. It will stay in the normal mode even if jitter returns, as long as no discontinuities occur.

IX. DESIGN METHODS

We have analyzed many properties of the phase-controlled oscillator with a sawtooth comparator. Some of these properties, notably the lock range, pull-in range, and noise bandwidth are significant in nearly all applications of the device. Others, such as peak jitter gain, seize frequency, and settling time are important only for certain specific applications.

Usually, in a particular design problem, two or three of the properties will be of prime importance and the rest can be neglected. Then the problem is to find the values of the design parameters (α, τ_1, τ_2) which yield the best combination of the important properties. If the properties are simple, like the lock frequency ($N\pi\alpha$), it is easy to find the best design.

9.1 *Filter Plot*

Unfortunately, most of the properties of the phase-controlled oscillator turn out to be complicated transcendental functions of the design parameters τ_1 and τ_2. Therefore we have presented many of the properties as contour curves on a plot of τ_1 vs. τ_2, which we call the filter plot (Figs. 5, 7, 8, 11, 13, and 15). Most of the properties are normalized through division by the gain constant α. In some cases, the count ratio N is also used as a normalizing factor.

τ_1 and τ_2 are the time constants of the phase lag filter (Fig. 4b), multiplied by the gain α. We have plotted τ_1 and τ_2 on logarithmic scales, to allow the presentation of large ranges. A useful property of these scales is that a given percentage change in τ_1 or τ_2 appears as a constant displacement on the plot. This facilitates estimating the effect of variations of the parameters.

τ_1 is always larger than τ_2; therefore the possible designs are restricted to the region below the 45° line on the plot. Points along this 45° line are identical, and correspond to the case of no filter. When τ_2 is zero, the phase lag filter degenerates to the RC filter. Since this case is of some interest, we have provided a zero τ_2 axis below the plot and indicated the intersection of the various contours with this line.

9.2 *Approximate Relations*

An examination of the filter plots shows that there are large regions where the contours approach straight lines. It is possible to derive simplified formulas for these regions. A summary of these approximations and the conditions for their validity is given below.

Pull-in frequency: $$\frac{\omega_p}{N\pi\alpha} \cong \frac{2}{\sqrt{3}}\sqrt{\frac{\tau_2}{\tau_1}} \quad (\tau_2 \gg 1) \tag{44}$$

Noise bandwidth: $$\frac{B_j}{\pi\alpha} \cong \tfrac{1}{2}\frac{\tau_2}{\tau_1} \quad (\tau_2^2/\tau_1 \gg 1) \tag{45}$$

Figure of merit: $$\frac{M}{N} \cong \frac{4}{\sqrt{3}}\sqrt{\frac{\tau_1}{\tau_2}} \quad (\tau_2^2/\tau_1 \gg 1) \tag{46}$$

Peak error, frequency step: $$\frac{\alpha\hat{\varphi}_e}{\Delta\omega_i} = \frac{\tau_1}{\tau_2} \quad (\tau_2^2/\tau_1 \gg 1) \tag{47}$$

Equation (44) has been derived from (29) by A. J. Goldstein.[4] Equation (45) can easily be found from (18). Equation (46) is found by dividing (44) by (45), according to the definition (31). Equation (47) can be derived from (28).

These approximations sometimes allow analytic methods to be used to find an approximate optimum solution. This requires justification of operating in the region where the approximations are valid.

9.3 *Optimization Techniques*

There are several types of optimization methods, which we shall discuss in order of increasing difficulty.

The simplest method optimizes one property by varying one parameter, all other parameters being fixed. This yields a class of designs which has one less parameter than the general case. The remaining parameters can be assigned to satisfy requirements on other properties, in confidence that the final design will have high performance for the optimized property.

An example of this approach has appeared in the literature.[1,9] The gain α and the time constant τ_1 (which together specify the resonant frequency) are held constant and the time constant τ_2 is varied to minimize the noise bandwidth B_j. This process restricts the design to

$$\tau_2 + 1 = \sqrt{\tau_1 + 1}. \tag{48}$$

For large values of τ_1, the damping ratio ξ approaches 0.5. Equation (48) is plotted in Fig. 20, against the figure of merit contours.

Let us describe one procedure for designing a circuit using (48). The gain α can be set to give the proper lock frequency. Then τ_1 and τ_2 can be set to give the required pull-in frequency, while satisfying (48).

This approach is good, and yields rather useful designs. However, it does not necessarily produce the best possible design for a given set of requirements.

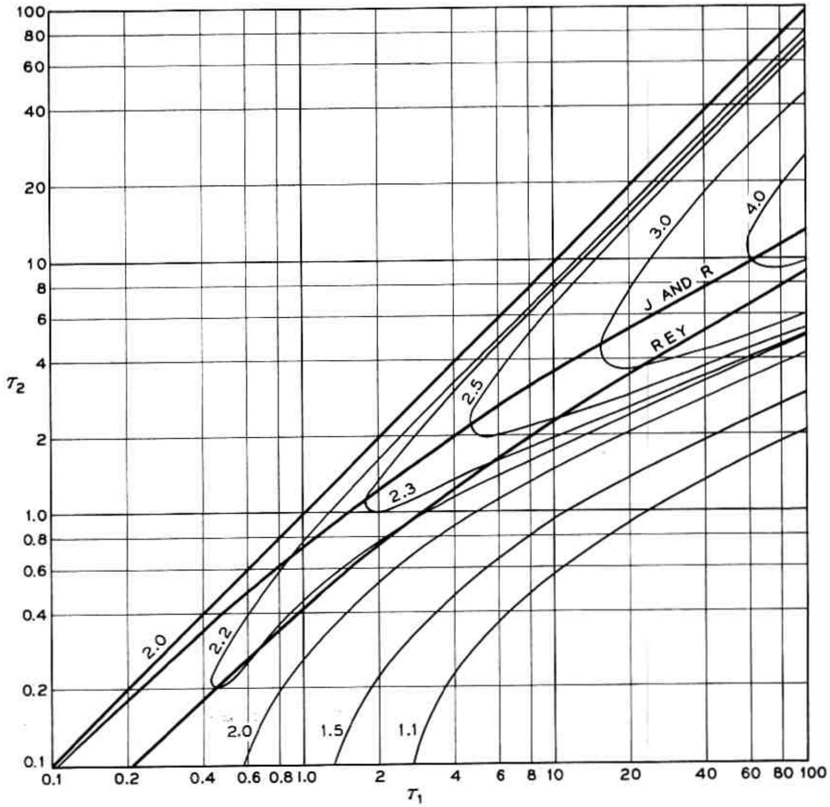


Fig. 20 — "Optimized" designs of Jaffe and Rechtin[10] and T. Rey,[1] with the figure of merit contours on the filter plot.

For example suppose the lock and pull-in frequencies have been specified, with a pull-in to lock ratio of 0.5. Following the above procedure, we compare Figs. 13 and 20 to find that τ_1 and τ_2 should be 18 and 3.2 to satisfy (48) and have $\omega_p/\omega_L = 0.5$. From Fig. 7 we find that the normalized noise bandwidth is 0.19.

To see that a better design than this is possible, suppose that τ_1 and τ_2 were 100 and 20. Then the noise bandwidth would be 0.12, a large improvement.

A more powerful technique is possible when some properties are specified by system requirements and another property should be optimized. The specified properties are used to restrict the range of the design parameters. Then the remaining range is examined to seek the optimum design.

For example, suppose that the lock range has been specified, and the normalized noise bandwidth is required to be less than 0.2. It is desired to maximize the pull-in frequency. A comparison of Fig. 7 and Fig. 13 shows that the design should lie on the upper part of the 0.2 noise bandwidth contour, and τ_1 and τ_2 should be as large as possible.

The most common problems require a compromise design which yields good results for two or more properties. Sometimes it is possible to express the relative importance of the properties mathematically. Then the optimum design can be derived analytically. A good example of this is given by Jaffe and Rechtin,[10] where the desirable properties are low-noise bandwidth and a high peak phase error due to a frequency step. Their design curve is shown in Fig. 20.

More often the relative importance of the properties is indistinctly known, and the engineer must use his judgment in striking a compromise. The filter plots are intended to aid this process by giving the engineer a "feel" for the circuit properties over the entire range of the parameters.

9.4 *Numerical Example*

To show how the design aids we have presented can be used in practice, we will do a realistic problem.

A phase-controlled oscillator is to be designed to smooth jitter in a 1.5 megacycle signal. In the worst case of mistuning, the circuit must pull itself into synchronism. We wish to design a circuit with low jitter noise bandwidth.

The uncertainty of the input signal is $\pm 10^{-5}$ or ± 15 cycles per second.

The oscillator center frequency is controlled by a frequency determining element, which we shall assume to be a crystal, and by the surrounding circuit. We take the range of the crystal as $\pm 10^{-5}$ or ± 15 cps. The effect of variations in the circuit will depend on the control the circuit has on the crystal, which is in turn related to the gain α. We assume that the range of center frequency due to circuit variations is ± 0.2 $N\pi\alpha$.

The count ratio N is 4.

Let us make the following definitions:

δ_s — maximum deviation of the signal frequency (rad per sec)

δ_0 — maximum deviation of the crystal tuning (rad per sec)

ϵ — maximum deviation of the oscillator center frequency due to circuit variations, divided by $N\pi\alpha$.

Then the maximum mistuning (which determines the pull-in frequency) is

$$\hat{\omega}_m = \omega_p = \delta_s + \delta_0 + \epsilon N\pi\alpha. \tag{49}$$

If we assume that the final design will be in a region where the approximate relations hold, we can use (44) and (45) for the pull-in frequency ω_p and the jitter noise bandwidth B_j.

When (44) and (49) are combined, we find

$$\frac{\tau_2}{\tau_1} = \frac{3}{4}\left(\frac{\delta_s + \delta_0}{N\pi\alpha} + \epsilon\right)^2. \tag{50}$$

For this value of τ_2/τ_1, the jitter bandwidth is

$$B_j = \frac{3}{8}\pi\alpha\left(\frac{\delta_s + \delta_0}{N\pi\alpha} + \epsilon\right)^2. \tag{51}$$

Note that the only variable is α. When we minimize B_j by varying α, we obtain

$$\begin{aligned} \alpha &= \frac{\delta_s + \delta_0}{N\pi\epsilon}, \\ \frac{\tau_2}{\tau_1} &= 3\epsilon^2, \\ \omega_p &= 2(\delta_s + \delta_0), \quad \text{and} \\ B_j &= \frac{3}{2}\frac{(\delta_s + \delta_0)\epsilon}{N}. \end{aligned} \tag{52}$$

When the numerical substitutions are made, we have

$$\begin{aligned} \alpha &= 75 \text{ rad/sec per radian,} \\ \frac{\tau_2}{\tau_1} &= 0.12, \\ \omega_p &= 377 \text{ rad/sec, or 60 cps,} \quad \text{and} \\ B_j &= 14 \text{ rad/sec, or 2.25 cps.} \end{aligned} \tag{53}$$

Now we have not yet completely specified the design, because we only have the ratio of τ_2 and τ_1. We can be confident of the numbers above for any value of τ_1, as long as we have the proper value of τ_2/τ_1 and as long as we stay in the region where the approximate relations are valid.

If we make τ_1 very large, we will require a very long time constant in the filter. Therefore we will make τ_1 just large enough to satisfy the condition for the approximate noise bandwidth relation, $\tau_2^2/\tau_1 \gg 1$. Let us set $\tau_2^2/\tau_1 = 4$. Then, from (53)

$$\begin{aligned} \tau_2 &= 33, \\ \tau_1 &= 275, \\ T_2 &= \frac{\tau_2}{\alpha} = 0.44 \text{ sec,} \quad \text{and} \\ T_1 &= \frac{\tau_1}{\alpha} = 3.67 \text{ sec.} \end{aligned} \tag{54}$$

If high accuracy is required, the values of τ_2 and τ_1 given in (54) can be used to find the exact values of ω_p and B_j, instead of using the approximate values given in (53).

X. CIRCUIT MODIFICATIONS

A two mode system has often been used[11] to increase the pull-in frequency. In this system, a frequency detector as well as a phase detector is used; the output of the frequency detector adjusts the oscillator tuning until the phase-controlled loop can synchronize. This scheme greatly extends the pull-in range, but requires additional hardware.

Another means of extending the pull-in frequency has been published by R. Ley.[9] Back-to-back diodes are placed across the series filter resistor R_1. When the circuit is in synchronism and the jitter is small, the diodes do not conduct. The small signal properties are just as we have analyzed them. However, if the circuit is not synchronized, the flickering error voltage will cause the diodes to conduct, shorting out R_1. This will bring the pull-in frequency up near the lock frequency.

The major drawback of this method is that large jitter error voltages will make the diodes conduct, and be passed on to the oscillator.

Either or both of these methods may be used to greatly extend the pull-in range if the other system requirements permit their use.

XI. SUMMARY

Nearly all the properties of the phase-controlled oscillator which have appeared in the literature have been analyzed for the case of the sawtooth comparator and the phase lag filter.

New theoretical material has been introduced on the effects of fast noise and jitter.

The sawtooth comparator has advantages over the sinusoidal comparator for many applications. The reason for this is that the gain of the sawtooth comparator remains constant over a broader range of operation.

The properties of the phase-controlled oscillator are presented in a manner which facilitates design without unnecessary restrictions. Various methods of design are discussed, and numerical examples are provided to illustrate the methods.

XII. ACKNOWLEDGMENTS

A. J. Goldstein did the analysis and programming for many of the properties presented here. D. E. Rowlinson processed the computer data to produce most of the contour plots. The author is indebted to H. A. Schneider for many suggestions for improving the manuscript. The construction of the experimental equipment was started by J. Beyer and completed by the author.

This work was suggested and encouraged by M. Karnaugh.

GLOSSARY OF IMPORTANT SYMBOLS

A Laplace transform is denoted by capitalizing the symbol.

B_j	jitter noise bandwidth
B_i	interference noise bandwidth
B_D	demodulator noise bandwidth
$f(\varphi_e)$	comparator function
$f_j(\varphi_{e0})$	effective comparator function
$G(\omega)$	any normalized noise transfer function

$H(s)$	filter transfer function
α_1	dc gain, comparator
α_2	dc gain, filter
α_3	frequency to voltage ratio, oscillator
$\alpha = \alpha_1\alpha_2\alpha_3$	open loop dc gain
$M = \dfrac{\omega_p}{B_j}$	figure of merit
N	count ratio
T_1	large filter time constant
T_2	small filter time constant
t_s	settling time
t_F	flicker time
v_1	comparator output voltage
v_2	oscillator input voltage
v_n	interference noise density
v_i	signal voltage amplitude
v_M	modulating voltage
Y	jitter transfer function
$\hat{Y}$	peak jitter gain
$\xi = \dfrac{1}{2}\dfrac{\tau_2 + 1}{\sqrt{\tau_1}}$	damping ratio
φ_i	input phase
$\Delta\rho_i$	change in input phase
φ_o	output phase
$\varphi_e = \varphi_i - \varphi_o$	phase error
$\hat{\varphi}_e$	peak phase error
φ_{er}	phase error margin
φ_{e0}	short-time average phase error
φ_{ej}	phase error due to fast jitter
$(\varphi_e)_{rms}$	root mean square of φ_{ej}
$\tau_1 = \alpha T_1$	large filter time constant (normalized)
$\tau_2 = \alpha T_2$	small filter time constant (normalized)
ω_i	input frequency
$\Delta\omega_i$	change in input frequency
$\hat{\Delta}\omega_i$	maximum frequency shift
ω_m	mistuning frequency
$\omega_n = \dfrac{\alpha}{\sqrt{\tau_1}}$	natural frequency
ω_L	lock frequency
ω_p	pull-in frequency
ω_s	seize frequency

References

1. Rey, T. J., Proc. I.R.E., **48,** Oct., 1960, p. 1760.
2. McAleer, H. T., Proc. I.R.E., **47,** June, 1959, p. 1137.
3. Hazeltine Staff, *Principles of Color Television,* John Wiley and Sons, Inc., New York, 1956, p. 180.
4. Goldstein, A. J., B.S.T.J., this issue, p. 603.
5. Rideout, V. C., *Active Networks,* Prentice-Hall, Inc., New York, 1954, p. 368.
6. Kimme, E. G., unpublished work.
7. Barnard, R. D., B.S.T.J., **37,** Jan., 1962, p. 227.
8. Richman, D., Proc. I.R.E., **34,** Jan., 1954, p. 106.
9. Ley, R., Annales de Radioélectricité, **13,** July, 1958, p. 212.
10. Jaffe, R., and Rechtin, R., I.R.E. Trans. on Info. Theory, **IT-1,** No. 1, 1955, p. 66.
11. Richman, D., Proc. I.R.E., **34,** Jan., 1954, p. 288.

Analysis of the Phase-Controlled Loop with a Sawtooth Comparator

By A. JAY GOLDSTEIN

(Manuscript received October 18, 1961)

Because of the recent interest in phase-controlled oscillators, a discussion of the phase-controlled loop with a sawtooth comparator is presented. The main emphasis is on finding the pull-in range of the loop. A companion paper in this issue (Ref. 4) deals with applications and shows how design parameters can be obtained from results developed here.

I. INTRODUCTION

The phase-controlled oscillator has evoked much interest in recent years. Some of its applications are to synchronism in television,[1,2] synchronization to a harmonic of a crystal oscillator,[3] elimination of jitter in pulse code modulation,[4] tracking filters, etc.

The general phase-controlled oscillator loop is given in Fig. 1. The incoming signal and the variable oscillator have the same free-running frequency ω_c. The phase comparator has as its output some function f of the phase difference $\varphi_e = \varphi_i - \varphi_0$. As examples of $f(\varphi_e)$ we have

the linear case: $f(\varphi_e) = \varphi_e$

the sinusoidal case: $f(\varphi_e) = \sin \varphi_e$

the sawtoothed case $f(\varphi_e) = \varphi_e$ for $-\frac{d}{2} < \varphi_e < \frac{d}{2}$

(see Fig. 2): $f(\varphi_e + nd) = f(\varphi_e)$ for $n = \cdots -1,0,1, \cdots$.

The output of the phase comparator passes through a filter whose impulse response is $h(t)$. The output of the filter $v(t)$ controls the variable oscillator according to the equation

$$\frac{d\varphi_0}{dt} = \alpha v(t). \tag{1}$$

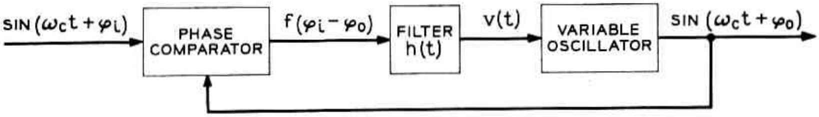

SIN $(\omega_c t + \varphi_i)$
PHASE COMPARATOR
$f(\varphi_i - \varphi_0)$
FILTER h(t)
v(t)
VARIABLE OSCILLATOR
SIN $(\omega_c t + \varphi_0)$

Fig. 1 — The general phase-controlled loop.

Thus, the frequency of the controlled oscillator is

$$\omega_c + \frac{d\varphi_0}{dt} = \omega_c + \alpha v(t).$$

In a companion paper in this issue, C. J. Byrne[4] discusses the engineering origins and applications of the sawtoothed comparator and shows how design parameters can be obtained from the results of this article.

This article is primarily concerned with finding the pull-in range of the loop. This is defined precisely in Section III. Briefly it is *the maximum asymptotic (in time) value of the mistuning $d\varphi_i/dt$ for which the slave oscillator eventually synchronizes or locks to the input frequency*. All of the literature cited in the references deals with this problem for the case of a sinusoidal or linear phase comparator. The linear case is easily solved since the resulting differential equation is linear. (See in particular Labin[5] for a detailed discussion.) In the sinusoidal case the differential equation of the system is nonlinear. Only in the cases of no filter and an ideal integrator has the equation, up to the present, been solved in closed form. See Labin[5] for an excellent discussion of the no-filter case. In order to handle the nontrivial filter, many authors have used methods of phase plane analysis.[6,7,8] Phase plane analysis is restricted to the problem of *capture range* in which the mistuning and phase error are zero for negative time, and the mistuning is constant for positive time. This

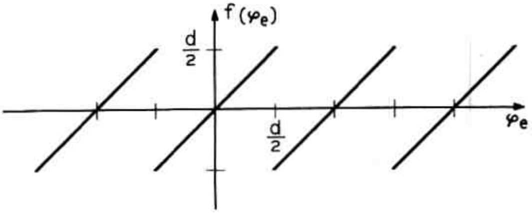


Fig. 2 — The sawtoothed phase comparator characteristic. The phase error φ_e is difference between the input and output phases of the loop.

kind of analysis gives only upper and lower bounds for the capture range and is restricted to a lag filter (Fig. 3). For an RC filter ($R_2 = 0$), Barnard[8] shows how phase plane analysis can give exact results.

To obviate the mathematical complexities, people have resorted to making various hypotheses about the nature of the solution of the nonlinear differential equation. These assumptions are based upon physical intuition and gross behavior observed in the laboratory. Different assumptions have led to different approximate solutions for the capture range. Moreover, they deal primarily with the lag filter, since it leads to a second-order differential equation while a more general filter gives a higher-order differential equation.

The loop equation when expressed as an integral equation is

$$\frac{d\varphi_e}{dt} = -\alpha \int_0^t f[\varphi_e(t')]h(t - t')\, dt' + \frac{d\varphi_i}{dt} - \alpha v_0(t).$$

It is surprisingly tractable for the sawtooth comparator, and the pull-in range can be computed for any filter. Fig. 4 shows the excellent agreement between theory and experiment for the lag filter. These experimental results were obtained by C. J. Byrne.

To obtain our results, we too must make an assumption. While the assumptions other authors have made deal with the behavior (in steady state) when far outside the pull-in range, ours deals with the behavior just outside of the pull-in range (see Section 4.4). This hypothesis is easily verified experimentally and has been so verified by C. J. Byrne for a representative selection of RC filters.

A brief description of each section follows.

Section II gives the basic integro-differential equation of the loop.

Section III defines the lock and pull-in range. The former is called by some the pull-out range. The lock range is the maximum frequency difference that the loop can lock to. It is given by

$$\omega_L = \alpha f_{\max} H(0)$$

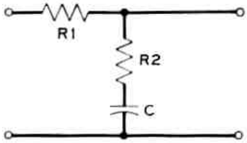


Fig. 3 — The integral compensating or lag filter. The normalized time constants are $\tau_1 = \alpha(R_1 + R_2)C$ and $\tau_2 = \alpha R_2 C$. For an RC filter $\tau_2 = 0$. α = (V. F. O. output frequency shift)/(V. F. O. input voltage).

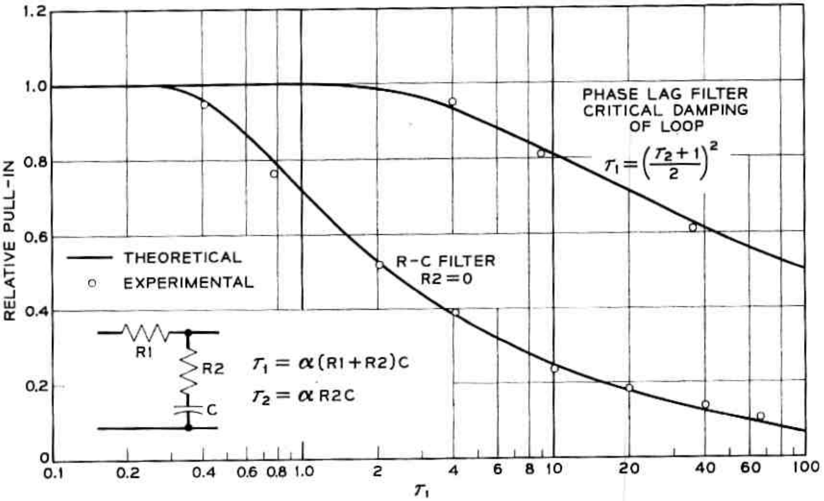


Fig. 4 — The relative pull-in range. For critical damping $(\tau_2 + 1)^2/4 = \tau_1$ and for the RC filter $R_2 = 0$.

where $H(0)$ is the dc gain of the filter and $f_{\max}$ is $d/2$ for the sawtooth comparator.

Section 4.1 gives the solution of the basic loop equation. This solution is the sum of (1) the solution of the linear phase comparator problem, (2) a series of step functions, and (3) a series of damped exponentials. The solution is obtained by representing the phase comparator function as the sum of the phase difference [giving (1)] and a series of translated unit step functions [giving (2) and (3)].

Section 4.2 gives the steady-state solution when *not captured*. In this case the output of the phase comparator is a periodic function whose period for a fixed filter depends on the asymptotic relative mistuning (Fig. 5). By examining this non-capture situation we obtain the pull-in range. We observe that in non-capture state the period and relative mistuning *must* correspond to a point on a curve typified in Fig. 5. Hence a relative mistuning lying below the minimum point of the curve corresponds to a capture or synchronized situation, and the height of the minimum gives the ratio of pull-in to lock range (the relative pull-in γ_p).

Section V gives all the explicit design formulae for the lag filter. For the special case of the RC filter ($R_2 = 0$ in Fig. 3) an explicit formula for relative pull-in can be given, namely

$$\gamma_p = \begin{cases} \tanh \frac{\pi}{4} (\alpha R_1C - \frac{1}{4})^{-\frac{1}{2}} & (\alpha R_1C \geqq \frac{1}{4}) \\ 1 & (\alpha R_1C \leqq \frac{1}{4}). \end{cases}$$

In all other cases we must find the roots of a transcendental equation by numerical approximation methods.

Byrne[4] gives graphs of the results of Section V for the lag filter. These are graphs of relative pull-in (Fig. 13), noise bandwidth (small signal) (Fig. 7), figure of merit (relative pull-in/noise bandwidth) (Fig. 15), and maximum loop gain (small signal) (Fig. 8).

The noise bandwidth is a measure of the ability of the loop to reject small phase noise. More explicitly, the noise bandwidth N of a network is defined to be the bandwidth of that ideal low-pass filter which passes the same white noise power as the given network.

There are many possible ways of defining a single measure of the performance of the system, depending on the particular application in mind. We have chosen the figure of merit γ_p/N, i.e., a large figure of merit implies high noise rejection and large relative pull-in.

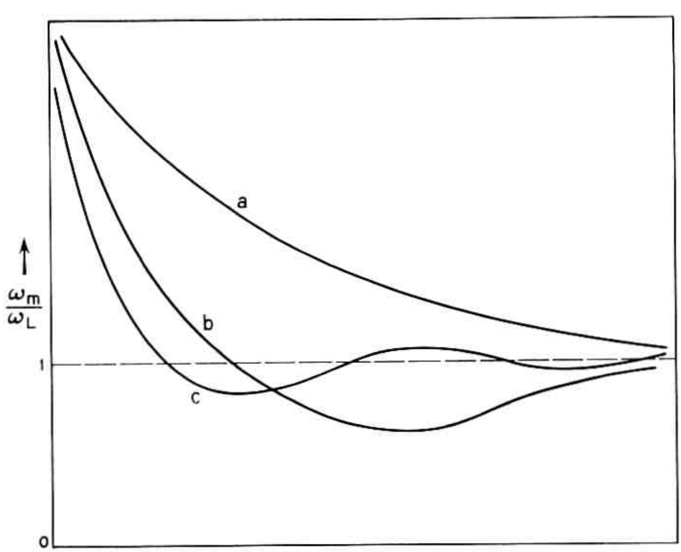


Fig. 5 — Relative mistuning ω_m/ω_L in a non-synchronized steady state vs the period T of the comparator output. (a) no filter, (a) and (b) overdamped loop and (c) underdamped loop.

For small phase deviations of the input, the comparator can be considered linear. We can then discuss the gain of the loop as a function of the frequency of the phase deviation. The maximum of the loop gain is denoted by $\hat{Y}$. In some applications $\hat{Y}$ is restricted by stability considerations to be less than unity.

Section VI is devoted to the derivation of several interesting asymptotic results for the lag filter. A simple formula is obtained for the relative pull-in for large values of the filter time constants. It is also shown that if the maximum loop gain is allowed to have a fixed value greater than unity, then, by appropriate choice of the time constants, arbitrarily large values of the figure of merit can be obtained.

This work could not have been completed without the aid of M. Karnaugh who suggested the problem, E. G. Kimme who proved that the sawtooth comparator is a continuous approximation to the original discrete sample data system, C. J. Byrne whose experimental work confirmed the formulae derived here, D. E. Rowlinson who constructed the contour curves from the computer data, and R. D. Barnard with whom many fruitful discussions were held.

II. THE BASIC LOOP EQUATION

We obtain an integro-differential equation for the loop by noting that the output of the filter can be written as a convolution plus initial conditions

$$v(t) = \int_0^t f[\varphi_e(t')]h(t - t')\, dt' + v_0(t)$$

where $v_0(t)$ is the filter output due to residual charges and fluxes in the filter at time zero. $v_0(t)$ damps out exponentially in all filters of interest. Substituting this into (1) and replacing φ_0 by $\varphi_i - \varphi_e$ we obtain

$$\frac{d\varphi_e}{dt} = -\alpha \int_0^t f[\varphi_e(t')]h(t - t')\, dt' + \frac{d\varphi_i}{dt} - \alpha v_0(t). \qquad (2)$$

In order that the derivations which follow not be unduly complicated by inessential parameters, we make the following normalizations

$$\begin{aligned} x(t) &= \varphi_e(t)/f_{\max} \quad (f_{\max} = d/2) \\ C(x(t)) &= f(\varphi_e(t))/f_{\max} \\ \varphi(t) &= \varphi_i(t)/f_{\max}\,. \end{aligned}$$

The normalized form of (2) becomes

$$\frac{dx}{dt} = -\alpha \int_0^t C[x(t')]h(t - t')\, dt' + \frac{d\varphi}{dt} - \alpha v_0(t)/f_{\max}\,. \qquad (3)$$

III. DEFINITIONS OF LOCK RANGE AND RELATIVE PULL-IN

If the input frequency $\omega_c + d\varphi_i/dt$ is increased "very slowly" to a value which is not too large, the output frequency $\omega_c + d\varphi_0/dt$ will follow it (i.e., be always equal to, or locked to, the input frequency). The maximum value of $d\varphi_i/dt$ for which lock-in will occur is called the *lock range* and is denoted by ω_L. More precisely, ω_L will be determined from (1) when the maximum dc voltage v is obtained. This maximum value is clearly the product of α, $f_{\max}$ the maximum value of the comparator function f and $H(0)$ the dc gain of the filter.*

$$\omega_L = \alpha f_{\max} H(0). \tag{4}$$

Suppose that the input frequency is not increased slowly, but in some sudden or erratic manner. Suppose moreover that the input frequency approaches a limiting value, ω_m, the *mistuning;* i.e.

$$\lim_{t \to \infty} \frac{d\varphi_i}{dt} = \omega_m .$$

In general, even if $0 < \omega_m < \omega_L$ (that is, we are in the lock range), the output frequency will not asymptotically lock to the input frequency (that is, be captured), but will be a modulated frequency. We define the relative *pull-in range* γ_p to be that normalized maximum frequency difference such that

$$-\gamma_p \omega_L < \lim_{t \to \infty} \frac{d\varphi_i}{dt} = \omega_m < \gamma_p \omega_L \tag{5}$$

implies

$$\lim_{t \to \infty} \frac{d\varphi_0}{dt} = \omega_m . \tag{6}$$

Notice that we make no restriction on how $d\varphi/dt$ approaches ω_m, as long as $| \omega_m | < \gamma_p \omega_L$.

IV. DERIVATION OF RESULTS†

4.1 *Basic Equation*

Let

$$0 < t_0 < t_1 \cdots < t_n < \cdots$$

* We shall use capital letters to denote the Laplace transform of the function denoted by corresponding lower-case letters.

† From here on we are dealing with the sawtooth comparator.

be all the instants (called discontinuity points) at which the phase difference $x(t)$ crosses the discontinuity of C, i.e.,

$$\lim_{\substack{\Delta \to 0 \\ \Delta > 0}} x(t_n - \Delta) = x(t_n -) = 1 + 2n'$$

where the first equality is a definition of $x(t_n -)$ and where n' is an integer dependent on n. Let

$$\begin{aligned} a_n &= 1 \text{ if } x \text{ is increasing at } t_n \\ a_n &= -1 \text{ if } x \text{ is decreasing at } t_n \\ a_n &= 0 \text{ if } x \text{ is stationary at } t_n \,. \end{aligned}$$

Using the unit step function

$$u(t) = \begin{cases} 0 & \text{for} \quad t < 0 \\ 1 & \text{for} \quad t \geqq 0 \end{cases}$$

we can express $C(x(t))$ in the analytically useful form

$$\frac{C(x(t))}{2} = \frac{x(t)}{2} - n_0 - \sum_{j=0}^{\infty} a_j u(t - t_j) \tag{7}$$

where n_0 is an integer so chosen that this equation holds at $t = 0$.

We note here for future reference that

$$x(t_n -) = n_0 \pm \tfrac{1}{2} + \sum_{j=0}^{n} a_j \,. \tag{8}$$

Substituting (7) into the loop equation (3), we obtain

$$\begin{aligned} \frac{1}{2}\frac{dx}{dt} = -\frac{\alpha}{2}\int_0^t x(t')h(t - t')\,dt' + \alpha n_0 \int_0^t h(t')\,dt' \\ + \alpha \sum_{j=0}^{\infty} a_j \int_0^t u(t' - t_j)h(t - t')\,dt' \\ + \frac{1}{2}\frac{d\varphi}{dt} - \alpha v_0(t)/2f_{\max} \,. \end{aligned} \tag{9}$$

Solving this by Laplace transform methods we obtain

$$\begin{aligned} \tfrac{1}{2}X(s) = \frac{s\Phi(s) - [\varphi_0(0) + \alpha V_0(s)]/f_{\max}}{2(s + \alpha H(s))} + \frac{n_0}{s}\,\frac{\alpha H(s)}{s + \alpha H(s)} \\ + \sum_{j=0}^{\infty} \frac{a_j e^{-st_j}}{s}\,\frac{\alpha H(s)}{s + \alpha H(s)} \,. \end{aligned}$$

Letting

$$R(s) = \frac{1}{s + \alpha H(s)} \tag{10}$$

we have

$$1 - sR(s) = \frac{\alpha H(s)}{s + \alpha H(s)}. \tag{11}$$

Note that $sR(s)$ is the transfer function between input phase φ_i and comparator output phase φ_e for the linear comparator case. $r(t)$ is then the phase response at the linear comparator output due to a step in input phase. Since applications will require the system to synchronize to a step in phase, we will assume that $r(t) \to 0$ as $t \to \infty$.

Using this equation and taking inverse transforms in the equation for $X(s)$ we obtain

$$\frac{x(t)}{2} = \frac{x_L(t)}{2} + n_0(1 + r(t)) + \sum_{j=0}^{\infty} a_j u(t - t_j) - \sum_{j=0}^{\infty} a_j r(t - t_j) \tag{12}$$

where

$$X_L(s) = \frac{s\Phi(s) - [\varphi_0(0) + \alpha V_0(s)]/f_{\max}}{s + \alpha H(s)}.$$

$x_L(t)$ is the solution of the loop equation in the case of a linear comparator function $f(x) = x$.

Using the final value theorem[10] we have

$$\begin{aligned} x_L(\infty) &= \lim_{t\to\infty} x_L(t) = \lim_{s\to 0} sX_L(s) = \lim_{t\to\infty} \frac{\varphi'(t)}{\alpha H(0)} \\ x_L(\infty) &= \frac{2\omega_m/d}{\alpha H(0)} = \frac{\omega_m}{\omega_L} \end{aligned} \tag{13}$$

From (12) and (7) we have for the comparator output

$$\frac{C(x(t))}{2} = \frac{x_L(t)}{2} - \sum_{j=0}^{\infty} a_j r(t - t_j) + n_0 r(t). \tag{14}$$

In a steady-state condition this reduces to

$$C(x(t)) = \frac{\omega_m}{\omega_L} - 2 \sum_{-\infty}^{+\infty} a_j r(t - t_j) \tag{15}$$

where the $n_0 r(t)$ term vanishes because of the remarks following the definition of $R(s)$.

4.2 *Steady-State Solution When Not Captured*

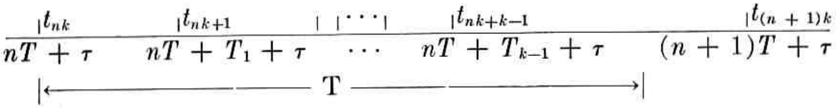
When we are not locked and in steady state, the output of the phase comparator will be a periodic function.* We give here a simplified heuristic derivation of the steady-state periodic solution. A rigorous derivation is easily obtained using the heuristics as a guide. In steady state, the normalized comparator output $y(t) = C(x(t))$ will be periodic with a period which we will call T. In a given period there may be many discontinuity points t_j ; let us suppose there are k. Then assuming we are in steady state, we can write

$$t_{nk+i} = nT + T_i + \tau, \quad \begin{cases} n = \cdots -1,0,1, \cdots \\ i = 0,1, \cdots ,k - 1 \end{cases} \tag{16}$$

where

$$0 = T_0 < T_1 < \cdots < T_{k-1} < T.$$

These relations are illustrated below.



The a_n's will be periodic in steady state and we let

$$a_{nk+i} = A_i \quad \begin{cases} n = \cdots -1,0,1, \cdots \\ i = 0,1, \cdots ,k - 1 \end{cases}. \tag{17}$$

It is no restriction to assume a time shift so that $\tau = 0$. Then, let

$$t = mT + u \; (0 < u \leqq T) \tag{18}$$

and combine the above three equations with (14). We obtain

$$\begin{aligned} y(t) &= C[x(t)] = C[x(mT + u)] \\ &= \omega_m/\omega_L - 2 \sum a_{nk+i} r(mT + u - t_{nk+i}) \\ &= \omega_m/\omega_L - 2 \sum_{i=0}^{k-1} A_i \sum_{n=-\infty}^{m} r[(m - n)T + u - T_i]. \end{aligned}$$

(The second summation has the upper limit m because $r(t) = 0$ for $t \leqq 0$.) Letting $j = m - n$, we obtain

$$y(t) = \frac{\omega_m}{\omega_L} - 2 \sum_{i=0}^{k-1} A_i \sum_{j=0}^{\infty} r(jT + u - T_i). \tag{19}$$

* A mathematical proof is not at hand. Indications of its truth are given in Benes[9] and experimental observations confirm this.

Let us define a periodic function

$$p(t,T) = \begin{cases} \sum_{j=0}^{\infty} r(t + jT) & (0 < t \leqq T) \\ p(t - nT,T) & (nT < t \leqq (n + 1)T) \end{cases} \tag{20}$$

or

$$p(t,T) = \sum_{-\infty}^{+\infty} r(t + jT).$$

$(r(t)) = 0$ for $t \leqq 0$ makes $p(t,T)$ a well defined function.) With this definition, the normalized steady-state comparator output, when not locked, can be written

$$y(t) = \frac{\omega_m}{\omega_L} - 2 \sum_{i=0}^{k-1} A_i p(t - T_i ,T). \tag{21}$$

The expression for $p(t,T)$ is familiar to those in the field of sample data systems.* Though superficially formidable, it can be expressed in closed form quite easily for the only important class of the filter transfer functions $H(s)$, namely rational functions. In that case $R(s)$ is a rational function too. Hence $r(t)$ is a linear combination of exponentials of the form $t^m e^{\beta t}$ (real part of β negative). Then $p(t,T)$ for $0 < t \leqq T$ is a linear combination of geometric series, each of the form

$$\begin{aligned} z(t) &= \sum_{j=0}^{\infty} (t + jT)^m e^{\beta(t+jT)} \\ &= \frac{d^m}{d\beta^m} \sum_{j=0}^{\infty} e^{\beta(t+jT)} \\ &= \frac{d^m}{d\beta^m} \frac{e^{\beta t}}{1 - e^{\beta T}} . \end{aligned} \tag{22}$$

This steady-state solution consists of a constant term $2\omega_m/d\alpha H(0)$, which is the normalized steady-state output for a linear phase comparator plus a linear combination (with coefficients ± 1) of time translates of the function $p(t,T)$, which is periodic of period T. The derivation shows that every steady-state periodic solution of the loop equation has the form of (21).

Equation (21) hides several pitfalls. These are:

1. We must have $| y(t) | \leqq 1$. Hence only certain T and T_i are admissible.

* It is the response of a filter $R(s)$ to an input $\sum_{j=-\infty}^{+\infty} \delta(t + jT)$.

2. Are the solutions represented by (21) physically realizable?
3. Are the solutions represented by (21) stable with respect to small noise perturbations?

These three topics are grouped under the title Boundary Conditions and will be discussed following a discussion of the pull-in range.

4.3 *Relative Pull-in*

From the definition of T, $y(T-) = \pm 1$ and by an appropriate choice of τ in (16) (if $\omega_m > 0$) we may assume $y(T) = 1$. Then from (21)

$$\frac{\omega_m}{\omega_L} = 1 + 2 \sum_{i=0}^{k-1} A_i p(T - T_i ,T). \tag{23}$$

Now the minimum value of $\omega_m > 0$ for which we have a non-constant periodic steady-state stable solution is by definition $\gamma_p \omega_L$, hence

$$\gamma_p = 1 + 2 \min \sum_{i=0}^{k-1} A_i p(T - T_i , T). \tag{24}$$

where the minimum is taken over all T and over all steady-state solutions satisfying conditions 1, 2 and 3 above.

4.4 *Boundary Conditions*

4.4.1 *Discontinuity Point Condition*

$y(t)$, being the normalized phase comparator output, satisfies $-1 \leqq y(t) \leqq 1$. Also $y(t'-) = \pm 1$ if and only if for some n and i, $t' = T_i + nT$, or $y(t)$ is stationary at t' (i.e., $y'(t') = 0$ and y at t' is increasing if $y(t') = -1$ or decreasing if $y(t') = 1$). These are equivalent to

$$\sum A_i(p(t' - T_i , T) - p(T - T_i , T)) = 0$$

if and only if $t' = nT + T_i$ or $y(t') - y(T)$ is stationary at t'. This restriction will be called the *discontinuity point condition.*

To analytically determine whether this condition is satisfied, in a general case, is clearly very difficult. For the case of the lag filter we can solve the problem analytically but must rely on an experimental fact. C. J. Byrne has found experimentally, in a large class of RC filters, that there is just one discontinuity per period T, i.e., the k in (21) is one. We will call this the *Experimental Hypothesis.* Thus

$$y(t) = \frac{\omega_m}{\omega_L} - 2p(t,T) \tag{25}$$

and

$$\gamma_p = 1 + 2 \min_T p(T,T). \tag{26}$$

In the section on the lag filter we show that if

$$p(T',T') = \min_T p(T,T) \tag{27}$$

then $p(t,T')$ satisfies the discontinuity point condition. Thus if $p(t,T')$ is realizable (it is — see below) and is stable under noise (we do not know, but have some evidence — see below) then

$$\gamma_p = 1 + 2p(T',T') \tag{28}$$

for the lag filter.

4.4.2 *Realizability Condition*

Does there exist, for each of the steady-state functions represented in (21) satisfying the discontinuity point condition, a corresponding input function $\varphi(t)$? That is, are the $y(t)$ in (21) physically realizable?

In Appendix A we prove realizability for any filter but not in quite the form stated above. We do the following:

(a) A particular input $\varphi(t) = 2\omega_m/d$ is injected.

(b) The loop is broken at the output of the phase comparator.

(c) Into the filter, at this point, is injected a voltage which asymptotically has the form (21).

(d) One shows that the output of the phase comparator has asymptotically the same form.

(e) In steady state the loop is closed.

4.4.3 *Non-Synchronous Stability*

Are the solutions stable? By this we mean: Will a steady-state solution be thrown into synchronism by a "small" noise? In formal terms, we suppose that a solution $y(t)$ has a discontinuity point, say t_0 shifted by noise to $t_0 + \Delta_0$. Each of the following discontinuity points $t_1, t_2, \cdots, t_n, \cdots$ is shifted to $t_1 + \Delta_1, t_2 + \Delta_2, \cdots, t_n + \Delta_n, \cdots$. It suffices for our purposes that the $(t_n + \Delta_n)$'s be asymptotically periodic (i.e., the noise sends us into another periodic solution and not into synchronism). The best we have been able to prove is that

$$\lim_{n\to\infty} \left[\frac{d\Delta_n}{d\Delta_0} \bigg|_{\Delta_0=0} \right] = c < \infty.$$

This has been done for the lag filter using the experimental assumption that $k = 1$ and that $T' - \epsilon < T < T'$, for ϵ sufficiently small, where

T' is given in (28). Now it would suffice for stability to show that Δ_n is bounded for Δ_0 sufficiently small, but the above does not imply this, for all it says is that

$$\Delta_n = c\Delta_0 + \epsilon_n \Delta_0^2$$

and we do not know that ϵ_n is bounded.

V. LAG (INTEGRAL COMPENSATING) FILTER

5.1 *General Results*

This section gives all the explicit formulae for design procedures in the case of the lag filter (Fig. 3). We assume the experimental hypothesis (see Section 4.4) throughout this section.

The transfer function of the filter is

$$H(s) = \frac{t_2 s + 1}{t_1 s + 1}$$

where

$$t_1 = (R_1 + R_2)C.$$
$$t_2 = R_2 C$$

Hence

$$R(s) = \frac{1}{s + \alpha H(s)} = \frac{t_1 s + 1}{t_1 s^2 + (\alpha t_2 + 1)s + \alpha}$$
$$= \frac{p_1 + \frac{1}{t_1}}{p_1 - p_2} \frac{1}{s - p_1} - \frac{p_2 + \frac{1}{t_2}}{p_1 - p_2} \frac{1}{s - p_2}$$

where p_1 and p_2 are the roots of denominator of $R(s)$. In particular, introducing the normalized dimensionless time constants

$$\tau_i = \alpha t_i, \qquad i = 1,2$$

we have for the roots

$$p_i = \frac{1}{t_1}(a + (-1)^i b)$$

where

$$a = (\tau_2 + 1)/2 \geqq \tfrac{1}{2}$$
$$b^2 = a^2 - \tau_1 .\text{*}$$

* The real or imaginary part of b is non-negative.

The denominator of $R(s)$ can be written in the form

$$s^2 + 2\omega_n \xi s + \omega_n^2$$

where

$$\omega_n^2 = (\alpha/t_1)$$

and ξ, the damping factor, is

$$\xi = (\tau_2 + 1)/2\sqrt{\tau_1} = a/(a^2 - b^2)^{\frac{1}{2}}.$$

In this notation we obtain

$$r(t) = \frac{1}{2b}[-(a - b - 1)\exp(-(a - b)t/t_1) \\ + (a + b - 1)\exp(-(a + b)t/t_1)].$$

Because $r(t)$ is a linear combination of exponentials, we can easily sum the infinite series for $p(t,T)$, obtaining

$$p(t,T) = \tilde{p}(\eta',\eta) = \frac{1}{2b}\left[-(a - b - 1)\frac{\exp[-(a - b)\eta']}{1 - \exp[-(a - b)\eta]} \right. \\ \left. + (a + b - 1)\frac{\exp[-(a + b)\eta']}{1 - \exp[-(a + b)\eta]}\right] \tag{29}$$

where $\eta' = t/t_1$ and $\eta = T/t_1$ are dimensionless time variables.

To obtain γ_p using the results of (27) we must find

$$\min_T p(T,T)$$

or the roots of

$$0 = \frac{d\tilde{p}(\eta,\eta)}{d\eta}.$$

Differentiating the expression for $\tilde{p}(\eta,\eta)$ we obtain $\eta \neq 0$ and

$$\frac{\sinh^2 (a - b)\eta/2}{\sinh^2 (a + b)\eta/2} = \frac{(a - b)(a - b - 1)}{(a + b)(a + b - 1)} \tag{30}$$

or $\eta = \infty$. And upon using the addition formula for the hyperbolic sine, we have

$$\frac{\tanh a\eta/2}{\frac{1}{b}\tanh b\eta/2} + \frac{b\tanh b\eta/2}{\tanh a\eta/2} = 2\frac{a^2 + b^2 - a}{2a - 1} = 2c \tag{31}$$

which defines c, or $\eta = \infty$.

Use of the quadratic formula gives

$$\frac{\tanh a\eta/2}{\frac{1}{b}\tanh b\eta/2} = c + \sqrt{c^2 - b^2} = c_1(a,b) \tag{32}$$

or $\eta = \infty$.* In special cases considered it was found that the minimum of $\tilde{p}(\eta,\eta)$ occurs at the first positive zero of its derivative (or at $\eta = \infty$).

5.2 *Critical Damping*

From (31) we see that as b approaches zero (damping factor equals one),

$$\begin{gathered}\frac{\tanh a\eta/2}{\eta/2} = 2\,\frac{a(a-1)}{2a-1} = c_1(a,0), \quad \text{if} \quad a > 1 \\ \eta = \infty, \quad \text{if} \quad \tfrac{1}{2} < a < 1.\end{gathered} \tag{33}$$

Thus $\gamma_p = 1$ for $b = 0$ and $\frac{1}{2} < a < 1$.

5.3 *No Filter and RC Filter*

The filter parameters satisfy

$$0 \leqq \tau_2 \leqq \tau_1$$

which upon conversion to the a and b parameters become

$$(a - 1)^2 \geqq b^2$$

and

$$a \geqq \tfrac{1}{2}.$$

Equality holds in the first case, when $R_1 = 0$ or $C = 0$ (i.e., there is no filter) and in the second case, when $R_2 = 0$, (i.e., a simple RC filter.)

For no filter, $a + b - 1 = 0$ or $a - b - 1 = 0$, and referring to (30) we have only $\eta = \infty$. Thus min $p(T,T) = 0$ and $\gamma_p = 1$.

For the RC filter $R_2 = 0$, $a = \frac{1}{2}$, we obtain from (30) $\eta \neq 0$ and $\sinh b\eta = 0$ or $\eta = \infty$. If b is real, $\eta = \infty$ and $\gamma_p = 1$. If b is imaginary

$$\eta = m\pi/b \qquad m = 1,2, \cdots$$

* If the negative sign were used in the quadratic formula then η would be negative (complex) when b was imaginary (real).

and we easily find that $\tilde{p}(m\pi/b, m\pi/b)$ is minimum at $m = 1$, giving finally

$$\gamma_p = \begin{cases} \tanh \dfrac{\pi}{4} \left(\tau_1 - \dfrac{1}{4}\right)^{-\frac{1}{2}} & \text{if} \quad \tau_1 > \dfrac{1}{4} \\ \\ 1 & \text{if} \quad \tau_1 \leqq \dfrac{1}{4} \end{cases} . \tag{34}$$

The results of these special cases are graphically summarized in Fig. 6. (Also see Fig. 13 of Byrne, Ref. 4.) In the shaded area of Fig. 6 the

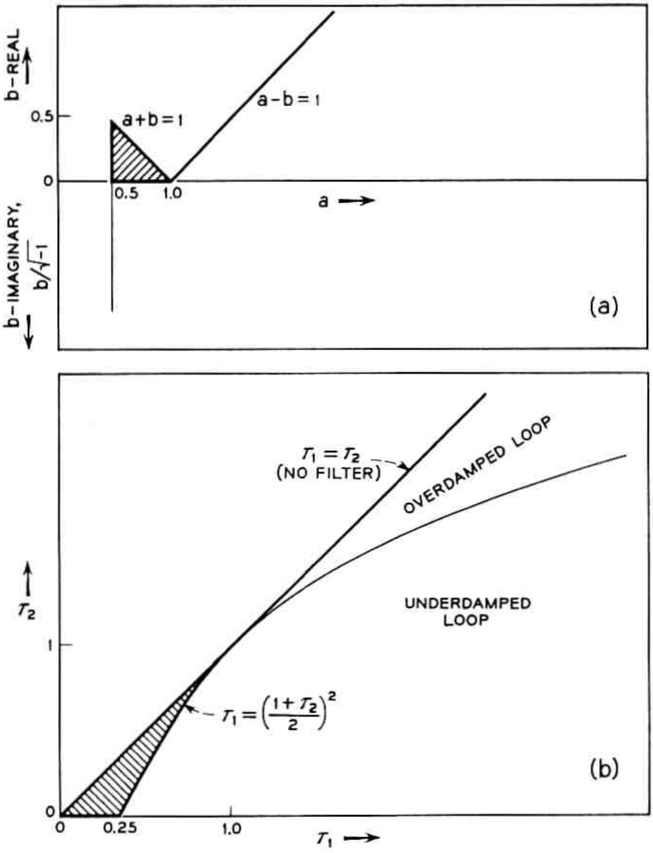


Fig. 6 — In part (a) the parameters a and b are restricted to lie below and/or to the right of the polygonal curve. The heavy lines and the shaded area give values of a and b for which the relative pull-in is unity. In part (b) the same information is given for the normalized time constants τ_1 and τ_2.

relative pull-in is unity. This follows from the fact that the left-hand side of (31) is bounded below by $2b$* while

$$2c - 2b = 2(a - b - 1)(a - b)/(2a - 1)$$

is negative in that region. Hence in (31) we must have $\eta = \infty$.

5.4 *Computational Procedures*

Except in the special cases of no filter ($R_1 = 0$) and the RC filter ($R_2 = 0$), there is no simple way of computing the relative pull-in. We must solve (32) by an iterative procedure and substitute the result into the equation for $\tilde{p}(\eta,\eta)$. If η is the solution of (32) or (33) we have a simpler equation for γ_p, namely

$$\gamma_p = [1 - D \operatorname{sech}^2 a\eta/2]/\tanh a\eta/2$$

where

$$D = \frac{(a - 1)c_1 - b^2}{c_1^2 - b^2} \quad (b \neq 0)$$

and

$$D = (a - \tfrac{1}{2})/a \quad (b = 0).$$

An upper bound for η is obtained from (32) and (33). Using the fact that $\tanh x < 1$, we obtain

$$\eta < \begin{cases} 2(\tanh^{-1} b/c_1)/b & (b \neq 0) \\ 2/c_1 & (b = 0) \end{cases}. \tag{35}$$

A lower bound for η in the case b is real is obtained by using the inequalities

$$z - z^3/3 \leqq \tanh z \leqq z.$$

Using this in the equation for η we have

$$a\eta/2 - (a\eta/2)^3/3 \leqq \tanh a\eta/2 = \begin{Bmatrix} c_1(a,b) \dfrac{\tanh b\eta/2}{b} \\ c_1(a,0)\eta/2 \end{Bmatrix} \leqq c_1\eta/2$$

giving the lower bound

$$2\left(3\,\frac{a - c_1}{a^3}\right)^{\frac{1}{2}} \leqq \eta.$$

* The left-hand side of (31) is of the form $b(x + 1/x)$. For x positive this is bounded below by $2b$.

We note here for future reference that if b is imaginary, $b = ib'$ then (35) implies the inequality

$$0 \leqslant \eta b' < \pi. \tag{36}$$

5.5 *Discontinuity Point Condition for the Lag Filter*

To prove that this condition is satisfied, it suffices to show that

$$\frac{dy}{dt} \neq 0 \quad \text{for} \quad 0 < t < T. \tag{37}$$

For, since we may suppose

$$y(T) = 1,$$

it follows that if

$$y(t') = 1 \quad (0 < t' < t)$$

then Rolle's theorem tells us that there exists a t'' with $t' < t'' < T$ such that $y'(t'') = 0$. This contradicts (37). It suffices also to prove (37) for that T which minimizes $p(T,T)$.

Recall that we are assuming we have a lag filter and that $k = 1$ in (21) (experimental hypothesis). Assuming (37) false, we obtain from (29) after some calculation

$$\frac{e^{-(a+b)\eta'/2}}{e^{-(a-b)\eta'/2}} = \frac{1 - e^{-(a+b)\eta/2}}{1 - e^{-(a-b)\eta/2}} \tag{38}$$

where η minimizes $\tilde{p}(u,u)$. Note that $0 < \eta' < \eta$.

Case 1, b real. Then $a > b$ and

$$\begin{aligned} \frac{e^{-(a+b)\eta'/2}}{e^{-(a-b)\eta'/2}} &= e^{-b\eta'} \\ &> e^{-b\eta} \\ &= \frac{e^{-(a+b)\eta/2}}{e^{-(a-b)\eta/2}} \\ &> \frac{1 - e^{-(a+b)\eta/2}}{1 - e^{-(a-b)\eta/2}}.^{*} \end{aligned}$$

Hence (38) is false.

[*] If $0 < x < y < 1$, then $x/y > x - 1/y - 1$, for $-x > -y$ implies $xy - x > xy - y$; hence in factoring and dividing we obtain the desired inequality.

Case 2, b imaginary. Let $b = ib'$, then (38) becomes

$$-b'\eta'/2 + m\pi = \arg(e^{-(a+ib')\eta/2} - 1). \tag{39}$$

Now the real part of $e^{-(a+ib')\eta/2} - 1$ is negative and the imaginary part is negative (since by (36), $0 < b'\eta/2 < \pi/2$). Hence the right-hand side is an angle in the third quadrant. But the left-hand side is an angle which can only be in the second or fourth quadrant, since

$$0 < b'\eta' < b'\eta < \pi.$$

Hence (39) is false, proving the discontinuity point condition.

5.6 *Small-Signal Properties of the Loop*

In this section we give formulae for design parameters of the loop when we are operating on the linear portion of the phase comparator. Then the closed loop transfer function Y is

$$Y(s) = \frac{\alpha(t_2 s + 1)}{s^2 t_1 + (\alpha t_2 + 1)s + \alpha}.$$

Restricting our attention to real frequencies and normalizing the frequency ω by

$$\Omega = \omega/\alpha$$

and recalling that

$$\tau_1 = \alpha t_1, \qquad \tau_2 = \alpha t_2$$

we obtain

$$|\, Y(\Omega)\,|^2 = \frac{\tau_2^2\Omega^2 + 1}{(\tau_2 + 1)^2\Omega^2 + (1 - \Omega^2\tau_1)^2}.$$

With the phase shift

$$\theta = -\arctan \frac{(1 + \tau_1\tau_2\Omega^2)\Omega}{(1 - \tau_1\Omega^2) + \tau_2(1 + \tau_2)^2\Omega^2} + \begin{cases} 0 \text{ if denominator positive} & * \\ \pi \text{ if denominator negative} \end{cases}.$$

Important parameters for design are the maximum gain and the frequency and phase shift at which it occurs and the range of frequencies for which the gain exceeds one. Differentiating $|\, Y(\Omega)\,|^2$ and solving for its zero gives

* The arctan is an angle in the first or fourth quadrant.

$$\Omega_{\max}^2 = \begin{cases} (2\tau_1 - 1)/2\tau_1^2 \quad \text{if} \quad \tau_2 = 0, \quad \tau_1 \geqq \frac{1}{2} \\ \{[1 + (2(\tau_1 - \tau_2) - 1)\tau_2^2/\tau_1^2]^{\frac{1}{2}} - 1\}/\tau_2^2 \quad \text{if} \quad \tau_1 - \tau_2 \geqq \frac{1}{2} \\ 0 \quad \text{if} \quad \tau_1 - \tau_2 \leqq \frac{1}{2}. \end{cases}$$

Solving $| Y(\Omega) |^2 \geqq 1$ gives

$$\Omega^2 \leqq \Omega_1^2$$

where

$$\Omega_1^2 = \begin{cases} \dfrac{2(\tau_1 - \tau_2) - 1}{\tau_1^2} & \text{if} \quad \tau_1 - \tau_2 \geqq \frac{1}{2} \\ 0 & \text{if} \quad \tau_1 - \tau_2 \leqq \frac{1}{2} \end{cases}.$$

We also have the interesting inequality

$$\sqrt{2}\,\Omega_{\max} \leqq \Omega_1$$

with equality when $\tau_2 = 0$. The cases $\tau_2 = 0$ and $\tau_1 - \tau_2 \leqq \frac{1}{2}$ are immediate. The case $\tau_1 - \tau_2 \geqq \frac{1}{2}$ gives

$$\begin{aligned} \Omega_{\max}^2 &= \{[1 + \tau_2^2\Omega_1^2]^{\frac{1}{2}} - 1\}/\tau_2^2 \\ &= \frac{\Omega_1^2}{[1 + \tau_2^2\Omega_1^2]^{\frac{1}{2}} + 1} \\ &\leqq \frac{\Omega_1^2}{2} \end{aligned}$$

proving the result in this case.

We wish to emphasize that the maximum gain is unity if and only if $\tau_1 - \tau_2 \leqq \frac{1}{2}$. Peak gain = constant contours are given in Fig. 8 of Ref. 4.

The 3 *db* point occurs at $\Omega = \Omega_{\frac{1}{2}}$ where

$$| Y(\Omega_{\frac{1}{2}}) |^2 = \tfrac{1}{2}$$

from which we obtain

$$\Omega_{\frac{1}{2}}^2 = B + (B^2 + \tau_1^{-2})^{\frac{1}{2}}$$

where

$$B = (\tau_2^2 + 2(\tau_1 - \tau_2) - 1)/2\tau_2^2.$$

The noise bandwidth N is defined by[4]

$$N = \int_0^\infty | Y(\omega) |^2 \, d\omega.$$

It can be evaluated in various ways, for example see Ref. 10. One obtains

$$N = \pi\alpha(1 + \tau_2^2/\tau_1)/2(\tau_2 + 1).$$

In the no-filter case ($\tau_2 = \tau_1 = 0$) and RC case ($\tau_2 = R_2 = 0$) we have $N = \pi\alpha/2$. $N =$ constant contours are given in Ref. 4, Fig. 7.

As discussed in the introduction, the figure of merit was chosen to be the ratio N/γ_p . $N/\gamma_p =$ constant contours are given in Ref. 4, Fig. 15.

VI. ASYMPTOTIC RESULTS

In this section we obtain the asymptotic results stated in the introduction. Since the derivations are tedious, the results are first summarized.

From computer data, the contour curves of relative pull-in $\gamma_p =$ constant with ordinate and abscissa the normalized time constants

$$\tau_1 = \alpha(R_1 + R_2)C$$

$$\tau_2 = \alpha R_2 C$$

seem to be asymptotic to straight lines for large values of the normalized parameters. (See Fig. 13 in Ref. 4.) This observation led to the conjecture that for fixed γ_p and large τ_2

$$\tau_1 = K(\tau_2 + 1).$$

In Appendix C we prove this and show that

$$1/K = 1 - (1/\gamma_p - \gamma_p)^2(\tanh^{-1}\gamma_p)^2.$$

With respect to the figure of merit (see Fig. 15 in Ref. 4), the following very important results are derived in Appendix B for the lag filter. Suppose the peak small-signal phase gain $\hat{Y}$ of the loop is restricted to be unity (it is always unity at dc). Then the maximum merit obtainable for filters giving the unity peak loop gain is 2.27. If, however, we permit a fixed peak gain greater than unity, we can have an arbitrarily large merit figure. This usually results in very poor transient response. More precisely, the following results are derived in Appendix B. Let us consider those lag filters for which the peak small-signal (phase) gain is fixed at $\hat{Y}$. Define M by

$$M^2 = 1 - \hat{Y}^{-2}$$

Then for a filter with normalized time constants τ_1 and τ_2 and normalized frequency $\Omega = \omega/\alpha$, for which the loop has peak gain $\hat{Y}$ occurring at frequency $\Omega_{\max}$, we have

$$\Omega_{\max}{}^2 = M/\tau_1$$

and

$$\tau_1 = (M^2\tau_2{}^2 + 2\tau_2 + 1)/2(1 + M).$$

Asymptotically for τ_2 large we obtain for the noise bandwidth (with $a = (\tau_2 + 1)/2$)

$$N/\pi\alpha = \left(1 + \frac{2(1 - M)}{M^2}\right) \Big/ 4a + 0(a^{-2})*$$

and for the relative pull-in range

$$\gamma_p = a^{-\frac{1}{2}} \frac{2}{\sqrt{3}} \frac{\sqrt{M - 1}}{M} + 0(a^{-\frac{3}{2}})$$

$$= \frac{2}{\sqrt{3}} \left(\frac{\tau_2 + 1}{\tau_1}\right)^{\frac{1}{2}} + 0((\tau_2/\tau_1)^{\frac{3}{2}}).$$

Thus the noise bandwidth decreases as a^{-1} while the relative pull-in decreases as $a^{-\frac{1}{2}}$. Hence the figure of merit increase as $a^{\frac{1}{2}}$.

The derivations of the preceding results are given in Appendices B and C.

APPENDIX A

Realizability of Steady-State Solutions

Recall that (assuming $d = 2$)

$$y(t) = \frac{\omega_m}{\alpha H(0)} - 2 \sum_{i=1}^{k-1} A_i \sum_{n=0}^{\infty} r(t - T_i - nT) \tag{40}$$

where [see (13)]

$$x_L = x_L(\infty) = \frac{\omega_m}{\alpha H(0)}.$$

Since we assume $y(t)$ satisfies the discontinuity point condition

$$\frac{\omega_m}{\alpha H(0)} = 1 + 2 \sum_{i=0}^{k-1} A_i p(T - T_i, T).$$

Break the loop at the output of the phase comparator, inject $y(t)$

* Two functions $f(x)$ and $g(x)$ satisfy $f(x) = 0(g(x))$ if and only if $| f(x)/g(x) | \leqq$ constant $< \infty$ for x sufficiently large.

into the filter, and let the input phase be $\omega_m t + x_L - c$ (where c is defined below). The phase output of the oscillator is given by

$$\frac{d\varphi_0(t)}{dt} = \alpha \int_0^\infty y(t')h(t - t')\,dt'$$

and upon integrating once and substituting (40),

$$\varphi_0(t) = \frac{\omega_m}{H(0)} \int_0^t \int_0^{t'} h(t'')\,dt''\,dt' - 2 \sum_{i=0}^{k-1} A_i \sum_{n=0}^{\infty} \int_0^t \int_0^{t'} \alpha r(t'' - T_i - nT)h(t' - t'')\,dt''\,dt'.$$

By taking the Laplace transform of the double integral in the summation and by using the relations in (10) and (11), we find

$$\varphi_0(t) = \frac{\omega_m}{H(0)} \int_0^t \int_0^{t'} h(t'')\,dt''\,dt' - 2 \sum_{i=0}^{k-1} A_i \left\{ \sum_{n=0}^{\infty} u(t - T_i - nT) - \sum_{n=0}^{\infty} r(t - T_i - nT) \right\}.$$

Now the remaining double integral is the integral of the step response of the filter and for large t is of the form $H(0)t + c$. Using this and the definition of $y(t)$, we obtain for large t

$$\varphi_0(t) \sim \omega_m t + c - 2 \sum_{i=0}^{k-1} A_i \sum_{n=0}^{\infty} u(t - T_i - nT) - y(t) + x_L\,.$$

Now using the discontinuity point condition and the representation of the comparator in (7) we find the comparator output is asymptotically $y(t)$. Hence in steady state we may close the circuit without any disturbance.

APPENDIX B

Figure of Merit for Constant Peak Gain and Large Time Constants (Lag Filter)

From Section 5.6 we have for the closed loop small-signal (phase) gain

$$|\,Y(\Omega)\,|^2 = \frac{\tau_2^2\Omega^2 + 1}{(\tau_1 + 1)^2\Omega^2 + (1 - \Omega^2\tau_1)^2}. \tag{41}$$

Differentiating with respect to Ω and equating the result to zero gives

$$\tau_1^2\tau_2^2\Omega_{\max}^4 + 2\tau_1^2\Omega_{\max}^2 - [2(\tau_1 - \tau_2) - 1] = 0. \tag{42}$$

We can also represent the square of peak gain $\hat{Y}^2$ as the ratio of the derivatives of the numerator and denominator of (41) evaluated at $\Omega_{\max}$.*

$$\hat{Y}^2 = \frac{\tau_2^2}{(\tau_2 + 1)^2 - 2\tau_1(1 - \tau_1\Omega_{\max}^2)}$$

$$= \frac{\tau_2^2}{\tau_2^2 - [2(\tau_1 - \tau_2) - 1] + 2\tau_1^2\Omega_{\max}^2}.$$

This, after using (42), gives

$$\hat{Y}^2 = \frac{1}{1 - \tau_1^2\Omega_{\max}^4}. \tag{43}$$

Defining $M \geqq 0$ by

$$M^2 = 1 - \hat{Y}^{-2},$$

we have

$$0 \leqq M < 1, \quad \text{since} \quad 1 \leqq \hat{Y} < \infty.$$

Also (43) gives

$$\tau_1\Omega_{\max}^2 = M. \tag{44}$$

Substitute (44) into (42) and solve for τ_1. Then

$$\tau_1 = (M^2\tau_2^2 + 2\tau_2 + 1)/2(1 - M).$$

Using this result in the formula for the noise bandwidth (Section 5.6), we have for $\hat{Y}$ constant and τ_2 large (and hence $a = (\tau_2 + 1)/2$ is large)

$$N = \frac{\pi\alpha}{4a}\left(1 + \frac{2(1 - M)}{M^2}\right) + 0(a^{-2}). \tag{45}$$

We now turn to the problem of obtaining asymptotic expressions for the relative pull-in range for $\hat{Y}$ fixed and greater than unity.

We can rewrite the expression for τ_1 as

$$\tau_1 = \frac{2M^2}{1 - M}a^2 + 2(1 + M)a - \frac{M + 1}{2}. \tag{46}$$

* If $f(x) = p(x)/q(x)$, then $f'(x_0) = 0$ implies $f(x_0) = p'(x_0)/q'(x_0)$. One obtains this result by logarithmic differentiation of $f(x)$.

Using the definition $b^2 = a^2 - \tau_1$, (46) and the binomial expansion, we have for large a

$$\frac{b}{a} = \left(1 - \frac{2M^2}{1 - M}\right)^{\frac{1}{2}} \left(1 - \frac{1 - M}{1 - 2M}\frac{1}{a} + 0(a^{-2})\right) \tag{47}$$

if $M \neq \frac{1}{2}$ and

$$b^2 = -3(a - \tfrac{1}{4}) \tag{48}$$

if $M = \frac{1}{2}$. *In the following we suppose* $M \neq \frac{1}{2}$. Recall (31) that to find the relative pull-in we need the root of

$$\tanh a\eta/2 = c_1 \frac{\tanh b\eta/2}{b}$$

where

$$c_1 = c + (c^2 - b^2)^{\frac{1}{2}}$$

and

$$c = (a^2 + b^2 - a)/(2a - 1).$$

Hence

$$\begin{aligned} c &= (2a^2 - a - \tau_1)/(2a - 1) \\ &= a[1 - \tau_1/a(2a - 1)] = a - \frac{\tau_1}{\tau_2} \qquad (49) \\ c &= a\left[1 - \frac{M^2}{1 - M} + 0(a^{-1})\right]. \end{aligned}$$

Also

$$\begin{aligned} c^2 - b^2 &= \left(a - \frac{\tau_1}{\tau_2}\right)^2 - (a^2 - \tau_1) \\ &= \left(\frac{\tau_1}{\tau_2}\right)^2\left(1 - \frac{\tau_2}{\tau_1}\right) \end{aligned} \tag{50}$$

giving

$$(c^2 - b^2)^{\frac{1}{2}} = \frac{\tau_1}{\tau_2}\left[1 - \frac{1}{2}\left(\frac{\tau_2}{\tau_1}\right) - \frac{1}{8}\left(\frac{\tau_2}{\tau_1}\right)^2 + 0\left(\frac{\tau_2^{\,3}}{\tau_1^{\,3}}\right)\right].$$

Finally

$$c_1 = \left(a - \frac{\tau_1}{\tau_2}\right) + (c^2 - b^2)^{\frac{1}{2}}$$
$$c_1 = a\left[1 - \frac{1}{2a} + 0(a^{-2})\right]. \tag{51}$$

Setting $z = a\eta/2$, we have

$$\tanh z = \frac{c_1}{b} \tanh \frac{b}{a} z. \tag{52}$$

We will show that for large a, z is small and then obtain an approximation to z by using a power series expansion for $\tanh z$. First note that the derivative at zero of the right-hand side of (52) is

$$\frac{c_1}{a} = 1 - \frac{1}{2a} + 0(a^{-2})$$

which approaches 1 from below for large a. Also

$$\frac{c_1}{b} = \frac{a}{b} + 0(b^{-1})$$
$$= \left(1 - \frac{2M^2}{1 - M}\right)^{-\frac{1}{2}} + 0(b^{-1}).$$

Hence $| c_1/b |$ is bounded away from 1 (and greater than 1).

A sketch of the curves of the two sides of (52) with the above two facts shows that

$$\lim_{a\to\infty} z = 0.$$

Using power series expansions in (52) we obtain for large a

$$z - \frac{z^3}{3} = \frac{c_1}{b}\left(z\left(\frac{b}{a}\right) - \frac{1}{3} z^3 \left(\frac{b}{a}\right)^3\right)$$

or

$$z^2 = 3\,\frac{1 - c_1/a}{1 - c_1b^2/a_3}$$
$$= 3\,\frac{\dfrac{1}{2a} + 0(a^{-2})}{\dfrac{2M^2}{1 - M} + 0(a^{-1})} \tag{53}$$
$$z = \frac{(3(1 - M))^{\frac{1}{2}}}{2M}\, a^{-\frac{1}{2}} + 0(a^{-1}).$$

From Section 5.4 the relative pull-in is

$$\gamma_p = \frac{1 - D}{\tanh\left(\frac{a\eta}{2}\right)} + D \tanh\left(\frac{a\eta}{2}\right)$$

where

$$D = ((a - 1)c_1 - b^2)/(c_1^2 - b^2).$$

Then

$$1 - D = \frac{(c_1 - a + 1)}{c_1 - b^2/c_1}$$

and

$$1 - D = \frac{c_1 - a + 1}{2c} \tag{54}$$

since

$$c_1 + b^2/c_1 = 2c.$$

Using (49) and (51) we have

$$1 - D = \frac{1 - M}{4aM^2} + 0(a^{-2}). \tag{55}$$

We now obtain the asymptotic formula for γ_p by substitution into the formula for γ_p the approximations for D, $1 - D$ and the approximation $\tanh(a\eta/2) \sim a\eta/2 = z$ with z approximated as in (53).

$$\begin{aligned} \gamma_p &= \frac{2(1 - M)^{\frac{1}{2}}}{3^{1/2}M} a^{-\frac{1}{2}} + 0(a^{-\frac{3}{2}}) \\ &= \frac{2}{3^{\frac{1}{2}}} ((\tau_2 + 1)/\tau_1)^{\frac{1}{2}} + 0(\tau_2^{-\frac{3}{2}}). \end{aligned} \tag{56}$$

APPENDIX C

Relative Pull-in (Lag Filter, Large τ_1 and τ_2)

Assuming that for a large a

$$\tau_1 = 2Ka + L + 0(a^{-1}) \tag{57}$$

we obtain from the definition

$$b^2 = a^2 - \tau_1$$

that

$$b = (a - K)\left[1 - \frac{L + K^2}{(a - K)^2} + 0(a^{-3})\right]^{\frac{1}{2}}.$$

Expanding the square root we obtain

$$b = (a - K)\left[1 - \frac{1}{2}\,\frac{L + K^2}{(a - K)^2} + 0(a^{-3})\right] \qquad (58)$$

and

$$\frac{b}{a} = 1 - \frac{K}{a} + 0(a^{-2}). \qquad (59)$$

From (57) since

$$a = (\tau_2 + 1)/2$$

we have

$$\frac{\tau_1}{\tau_2} = K + \frac{L + K}{\tau_2} + 0(\tau_2^{-2}) \qquad (60)$$

From (50) we have

$$c^2 - b^2 = \left(\frac{\tau_1}{\tau_2}\right)^2 - \left(\frac{\tau_1}{\tau_2}\right)$$

and by using (60) we have

$$c^2 - b^2 = (K^2 - K)\left[1 + \frac{(L + K)(2K - 1)}{K(K - 1)}\,\frac{1}{\tau_2} + 0(\tau_2^{-2})\right].$$

Using the binomial expansion

$$(c^2 - b^2)^{\frac{1}{2}} = (K^2 - K)^{\frac{1}{2}}\left[1 + \frac{1}{2}\,\frac{(L + K)(2K - 1)}{K^2 - K}\,\frac{1}{\tau_2} + 0(\tau_2^{-2})\right]. \qquad (61)$$

From (49)

$$c = a - \frac{\tau_1}{\tau_2}$$

and using (60), we obtain

$$c = a - K - \frac{L + K}{\tau_2} + 0(\tau_2^{-2}). \qquad (62)$$

Then

$$\begin{aligned} c_1 &= c + (c^2 - b^2)^{\frac{1}{2}} \\ &= (a - K) + (K^2 - K)^{\frac{1}{2}} + 0(\tau_2^{-1}). \end{aligned} \tag{63}$$

Finally from (58) and (63)

$$\begin{aligned} \frac{c_1}{b} &= \frac{c_1/(a-K)}{b/(a-K)} \\ &= 1 + \frac{(K^2 - K)^{\frac{1}{2}}}{a} + 0(a^{-2}). \end{aligned} \tag{64}$$

Letting $z = a\eta/2$, (31) becomes

$$\tanh z = [1 + (K^2 - K)^{\frac{1}{2}}/a + 0(a^{-2})] \tanh (1 - \mathrm{K}/a + 0(a^{-2}))z. \tag{65}$$

Using the addition formula for $\tanh (A + B)z$ and simplifying, we have

$$\begin{aligned} &\tanh^2 z \tanh (K/a + 0(a^{-2}))z - [(K^2 - K)^{\frac{1}{2}}/a + 0(a^{-2})] \tanh z \\ &\quad + [1 + (K^2 - K)^{\frac{1}{2}}/a + 0(a^{-2})] \tanh (K/a + 0(a^{-2}))z = 0. \end{aligned} \tag{66}$$

We show that z/a approaches zero with a and use this to simplify (66). From (35)

$$z < \frac{2 \tanh^{-1} [1 + (K^2 - K)^{\frac{1}{2}}/a + 0(a^{-2})]}{1 - K/a + 0(a^{-2})}$$

$$= \frac{\ln\left(\dfrac{2a}{(K^2 - K)^{\frac{1}{2}}} + 1 + 0(a^{-1})\right)}{1 - K/a + 0(a^{-2})}.$$

Since

$$\lim_{u \to \infty} \ln u/u = 0$$

we have

$$\lim_{a \to \infty} z/a = 0.$$

Returning to (66), we now have asymptotically

$$\tanh^2 z + \left(\frac{K-1}{K}\right)^{\frac{1}{2}} \frac{\tanh z}{z} - 1 = 0.$$

Solving for K we obtain

$$1/K = 1 - \left[\frac{1}{\tanh z} - \tanh z\right]^2 z^2. \tag{67}$$

Now the relative pull-in given in Section 5.4 is

$$\gamma_p = \frac{1 - D}{\tanh z} + D \tanh z$$

and we easily show that [using (54)]

$$\begin{aligned} 1 - D &= \frac{c_1 - a + 1}{2c} \\ &= \frac{(K^2 - K)^{\frac{1}{2}} - K + 1 + 0(a^{-1})}{a - K + 0(a^{-1})} \\ &= 0(a^{-1}). \end{aligned}$$

Hence asymptotically for fixed z,

$$\gamma_p = \tanh z + 0(a^{-1}). \tag{68}$$

Thus for given relative pull-in, the above gives us z and $\tanh z$, and then (67) gives K from which (57) gives for large τ_1

$$\tau_1 = K(\tau_2 + 1). \tag{69}$$

REFERENCES

1. Richman, D., Proc. I.R.E., **42,** 1954, p. 106.
2. Richman, D., Proc. I.R.E., **42,** 1954, p. 288.
3. Ley, R., Annales de Radioelectricté, **13,** No. 53, 1958, p. 212.
4. Byrne, C. J., This issue, p. 559.
5. Labin, E., Phillips Res. Rep., 1949, p. 291.
6. Gruen, W. J., Proc. I.R.E., **41,** 1953, p. 1043.
7. Jaffe, R., and Rechtin, R., I.R.E., P.G.I.T., **1,** 1955, p. 66.
8. Barnard, R. D., B.S.T.J., **41,** January, 1962, p. 227.
9. Benes, V. E., B.S.T.J., **41,** January, 1962, p. 257.
10. James, H. M., Nichols, N. B., and Phillips, R. S., *Theory of Servomechanisms*, M.I.T. Radiation Laboratories Series, McGraw-Hill Book Co., New York, 1947.
11. Preston, G. W., and Tellier, J. C., Proc. I.R.E., **41,** 1953, p. 249.

Reliability of Components for Communication Satellites

By I. M. ROSS

(Manuscript received August 15, 1961)

This article considers the reliability of components such as transistors, diodes, and solar cells in relation to the design of a communication satellite with adequate reliability. Consideration is given to methods for determining the reliability of high-quality components and of techniques for selecting the most stable components for this application. It is concluded that, at least for a simple communication satellite, components can now be obtained that will lead to a satisfactory life.

I. INTRODUCTION

All the necessary components and circuit techniques are available to fabricate a simple communication system using low-orbit satellites.[1] Such a system would use many satellites at an altitude of a few thousand miles and be capable of global communications with a few megacycles baseband. The ground receiver portion of the system could achieve adequate signal-to-noise for very low received power by use of high-gain receiving antennas, low-noise maser receivers and FM modulation with feedback. The satisfactory performance of this type of receiver was demonstrated in the Echo I experiment.[2] In conjunction with such sensitive ground receiver equipment, it is possible to use a satellite repeater putting out only a few watts of power from an isotropic antenna, and hence avoiding the additional complexity of attitude stabilization. The components needed for such a satellite, including the traveling-wave tubes, transistors, diodes and solar cells, are all either available or achievable within the capability of existing technology. Thus a communication satellite system is feasible in principle. Whether or not it is economical and therefore practical, depends upon the life expectancy of the system, and specifically on the life of the satellite itself. It will be assumed here that a satellite life of at least five years is a reasonable target in the design of a practical communication system. By the very nature of the system, repair of the satellite is presently impossible (and if ever possible, would be exorbitantly expensive), and because of the cost penalty of additional weight in orbit, extensive redundancy is most undesirable. Thus, the practicality of the system depends critically on the reliability of the components that make up the satellite itself. This paper is devoted to a discussion of the reliability of components in relation to the design of a satellite with adequate reliability. Although the discussion is directed specifically to low-orbit (several thousand miles altitude) satellites, many of the ideas could apply equally well to higher orbits.

In Section II, below, consideration is given to the order of component reliability needed in a simple communication satellite. Section III deals with the reliability of components in general with emphasis on means for attaining highly reliable components and for determining quantitatively their degree of reliability. Section IV discusses the level of reliability that can be achieved in three critical classes of components, namely transistors and diodes, traveling-wave tubes and solar cells. Finally, it is concluded that, with careful manufacture and selection, components can be obtained for a practical communication satellite system.

II. COMPONENT RELIABILITY REQUIRED FOR COMMUNICATION SATELLITES

For the consideration of reliability it is convenient to divide the life of a satellite into three periods, namely pre-launch, launch, and orbit. It is usual practice to assume that any failure that occurs a reasonable time prior to lift-off can be corrected by replacement and that, at the worst, this could result in some delay in the launch time. For such an assumption to be valid, it is necessary that components or batches of components be accessible and removable so that failed portions of the satellite can be replaced. The design for such flexibility does necessitate some weight increase. Although the launch period is short, it is accompanied by large mechanical stresses liable to cause failure. As will be discussed later, in the section on traveling-wave tubes, experience with many launches has shown that with well designed components and equipment, failure during launch of the electronic equipment in a satellite is not a significant factor in the over-all reliability of the satellite. It is the third period, life in orbit, which dominates the reliability design of a satellite. In this section we consider the relationship between the reliability of components and the anticipated life in orbit.

In calculating the probability of survival of a system containing a large number of components, it is frequently assumed that the failure distribution of any type of component is exponential. On such an assumption, the performance of a given type of component can be characterized by a mean time to failure or a failure rate. One of the more convenient ways to represent the failure rate is in terms of a number of failures for a given number of component operating hours. A method which is in increasing use defines failure rate as the number of failures per 10^9 component hours (1 failure per 10^9 hours corresponds to a failure rate of 0.0001 per cent per 1000 hours). By way of calibration, a good resistor or capacitor has a failure rate in the range 5 to 10 per 10^9 component hours, while an entertainment receiver tube will have a rate in the neighborhood of 100,000 per 10^9 hours.

If we assume that a given system contains n_1 components of a given type, and that the failure rate for that type is f_1 per 10^9 hours, we expect statistically that there will be $n_1 f_1$ failures per 10^9 hours. Hence in a time t hours we expect $tn_1 f_1/10^9$ failures. Assuming that failure probability is random and that the failure of any one of these components leads to failure of the system — that is, assuming no redundancy — the probability P_1 that the system will not fail in t hours due to failure of one of the n_1 components, is given by:

$$P_1 = \exp\left[-\frac{tn_1 f_1}{10^9}\right]. \tag{1}$$

Similarly, if we have a system composed of n_1, n_2, etc., components of types having failure rates f_1, f_2, etc., and we again assume no redundancy, the probability P_m of survival for time t is given by:

$$P_m = \exp\left[-\frac{t}{10^9}\sum_1^m (n_m f_m)\right]. \tag{2}$$

This simple equation can be used to estimate probability of system's survival, provided that the following conditions are met:

a) The failure mode of the components is assumed random with recognized exceptions being treated separately.

b) The system contains no redundancy.

Assumption b) is unrealistic since a certain degree of redundancy will be featured in any good design. However, because of weight limitations in a satellite, redundancy cannot be used to correct for poor reliability performance of a majority of the devices. Hence the equation is useful in determining desired objectives.

Table I shows the results of reliability calculations for a hypothetical communication satellite. At the left of the table are listed the types and numbers of critical components used. These types and numbers, which are representative of a very simple repeater of a few megacycles baseband, do not include any allowance for redundancy, nor do they include allowance for the telemetry invariably associated with such a system.

TABLE I — RELIABILITY CALCULATION FOR SIMPLE COMMUNICATION SATELLITE

Type of Component	Number (n)	Case I		Case II		Case III	
		Failure Rate (f) (Failures/10^9 hrs)	P Product (nf)	Failure Rate (f) (Failures/10^9 hrs)	Product (nf)	Failure Rate (f) (Failures/10^9 hrs)	Product (nf)
Transistor	140	20	2800	10	1400	5	700
Diodes	161	15	2415	10	1600	5	805
Resistor	400	5	2000	5	2000	2	800
Capacitor	250	10	2500	5	1250	2	500
Inductor and Transformer	40	20	800	15	600	5	200
Relays	6	50	300	25	150	6	120
Ni-Cd Cells	20	50	1000	25	500	15	300
Totals	1017		11,815		7510		3425
Average Failure Rate		11.6		7.4		3.4	
Probability of success — 1 year		0.901		0.94		0.97	
Probability of success — 5 years		0.60		0.72		0.86	

Excluded from the list is the traveling-wave tube. The unique life properties of the single traveling-wave tube in a nonredundant satellite warrant special treatment. Also excluded are the solar cells which, as will be discussed later, will probably fail through wear-out resulting from radiation damage and thus cannot be treated with the statistics of equation (2).

The table shows three cases, each assuming somewhat different failure rates for the components. For each case the table gives the failure rate f assumed for the component, the product of the failure rate times the number n of each component, the total sum $\sum_1^m (n_m f_m)$ and the average failure rate. Also shown in the table is the probability of success of the satellite, i.e., no failure of any component as calculated using (2), for one-year operation and for five-year operation. It is seen that case 1 represents satisfactory performance for one year and poor performance for five, while case 3 represents satisfactory performance for five years. Case 2 is an intermediate case. Using some judgment as to the relative values of failure rates for various components, the failure rates were chosen in the three cases to give the above results. Thus the table shows what level of component reliability is needed to meet a given systems performance.

It must be emphasized that considerable caution is needed in the interpretation of the results shown in Table I. Implicit in the calculations are many assumptions, the validity of which could be questioned. The results should therefore be used as a guide to the order of magnitudes of reliability required and should not be considered to be precise predictions of systems performance. There are, nevertheless, a number of general conclusions to be drawn from the table. The first is that although this is a fairly simple system — 1000 components — average failure rates in the neighborhood of 10 per 10^9 component hours are required to give anything approaching economical life. As seen from (2), the life expectancy for a given probability of success varies inversely with the average failure rate. Thus, an average failure rate in the neighborhood of 100 would be intolerable, while an average failure rate in the order of 1 would permit increased design life and/or complexity. A second conclusion is that all the components that are numerous, i.e., all the transistors, diodes, resistors and capacitors, require an equally high order of reliability. This conclusion results directly from forbidding redundancy for the high-runner components. A final conclusion is that, at least for the more reliable designs, the reliability of connections between components cannot be ignored. For the 1000 components of Table I there would be several thousand connections and hence, in order that there be an insignificant probability of a connection failure, they must have failure rates substantially less than 1 per 10^9 hours. Although there is little quantitative information regarding reliability of connections, it is believed that those liable to fail are eliminated during the vibration, temperature cycle, and vacuum tests normally carried out as part of the acceptance test of a complete satellite.

III. RELIABILITY OF COMPONENTS

Fig. 1 shows a possible failure pattern for a batch of components. Such a curve could be obtained by taking a large number of new components of a specific type, operating them under typical conditions, and plotting the failure rate versus time for the batch. The distribution has two regions of relatively high failure rate, one early in life and attributable to "manufacturing freaks," one later in life attributable to "wear-out," separated by a region of low failure rate labeled "random failure." These three regions will be discussed separately.

3.1 *Wear-Out Failure*

In some manufactured products there is a mechanism or a collection of mechanisms which systematically reduces the useful performance of the product until a point is reached at which it has no further utility and is "worn out." Typical examples of wear-out mechanisms are friction of

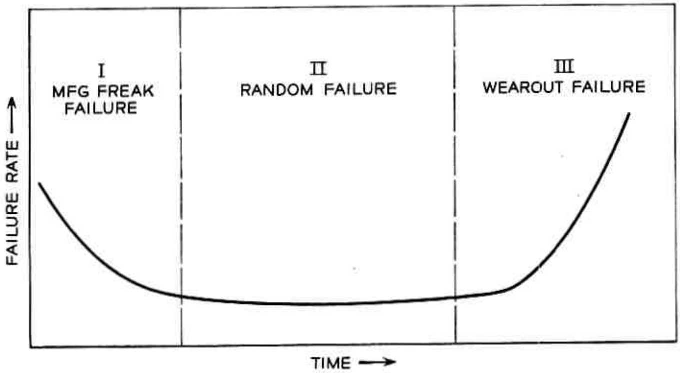


Fig. 1 — Possible failure distribution for a large number of new components.

bearings, corrosion of relay contacts, and deactivation of electron tube cathodes. If, for a given batch of components, conditions were identical during fabrication and use, then all components would fail in response to wear-out simultaneously. However, because conditions are not identical, simultaneous failure does not occur, and the failure distribution is characterized by a peak of finite width. Region III in Fig. 1 shows the onset of wear-out. Once wear-out failure commences, the failure rate of the batch of components increases vary rapidly, and effectively all components of that type must be replaced. In systems such as satellites, where replacement is not possible, the time at which wear-out becomes significant should be greater than the designed life of the satellite. Lengthening of the time to wear-out can only be achieved by understanding the wear-out mechanisms and by designing the components either to minimize or eliminate these effects.

3.2 *Manufacturing Freak Failure*

There is a certain percentage, preferably small, of any product that fails unusually early in life because of some defect in manufacture. These are, in a sense, objects that were not made according to the design. For example, such early failures can occur both in tubes and semiconductor devices as a result of defective seals or of the presence of particles inside the encapsulations. The prevalence of manufacturing freaks can be reduced drastically by quality control in manufacture. Remaining freaks can usually be detected and rejected by rigorous pre-aging tests, such as leak tests, vibration and shock tests. In addition, the product can be aged for a period longer than that corresponding to Region I, so that the remaining freaks will fail during this "pre-age period."

3.3 *Random Failure*

Even in a well designed and well manufactured product there may be a substantial period, after that exhibiting high failure rate due to manufacturing freaks and before wear-out occurs, of a continuing failure rate. These failures include components which, through presumably detectable causes, fail in response to manufacturing weaknesses much later than the majority of freaks, and others which fail through similar causes to, but earlier than, the wear-out failures. The failures that occur during this period may generally be attributable to a large number of different causes, each of which occurs so rarely that it would be exorbitantly expensive to identify all of them. This period is in essence the useful life of the product. If the frequency of such failures is sufficiently low, as indicated, these may be essentially below the noise level of identification of mechanisms, and a random failure mechanism, and hence a constant failure rate, may be assumed. Although there may be considerable doubt as to the validity of this assumption for some components, it has proved useful in the estimation of over-all systems reliability.

Fig. 2 summarizes the steps that can be taken to cope with the various modes of failure shown in Fig. 1. The region of high failure rate corresponding to wear-out can be moved further out in time by design based upon knowledge of the failure mechanisms. The number of devices subject to early failure through manufacturing freaks can be reduced by quality control, rejected after testing and annihilated by pre-aging. Hence, provided sufficient care is taken, it is possible to obtain a product which, during the intended life of the system, will exhibit substan-

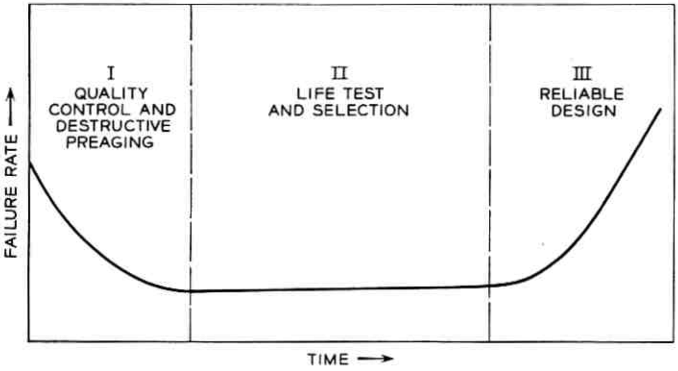


Fig. 2 — Summary of steps that can be taken to reduce failures of various types.

tially only a low failure rate corresponding to Region II. This failure rate can be determined from the results of extensive life tests involving, for the most reliable components, thousands of devices for thousands of hours.

The low failure rate of Region II is that characteristic of the product. Where reliability is of supreme importance, it is desirable to select from the product as a whole those components that exhibit the greatest degree of stability. This can be achieved by putting on life test a number of components many times that needed in the system, and after a given length of time selecting from the batch only those components which have shown the minimum change in their parameters. The duration of the life test prior to selection will depend upon a number of factors, including the life required in the system and the system's schedule which, itself, frequently limits the life-test period. In the selection of submarine cable tubes, a period of seven months is used. Although it is expected that the selected product will have a lower failure rate than the batch from which it was selected, it is difficult, if not impossible, to estimate the degree of this improvement. The consensus, however, is that a factor of 10–100 improvement could be achieved.

In order to achieve the reliability potential of a carefully designed and manufactured component, it is essential that the same care go into the design and assembly of circuits and subsystems. Circuits must be designed with adequate margins, and power dissipations must be determined so that temperatures do not reach values at which reliability of the components is no longer adequate. Assembly procedures should be arranged to avoid excessive mechanical or thermal shock. The conservative use of a component is thus an important part of the achievement of reliability.

IV. RELIABILITY OF SPECIFIC COMPONENTS

The components that appear in large number in a typical satellite and require reliabilities corresponding to 10 failures per 10^9 hours, include transistors, diodes, resistors and capacitors. Passive components, resistors and capacitors, have for many years been available with reliability in this range. However, until recently such low failure rates had not been achieved in the active components. For this reason the discussion in Section 4.1 below is restricted to transistors and diodes.

The traveling-wave tube used to generate the output power in most communication satellite designs does not require the high degree of statistical reliability called for in transistors and diodes. However, it is required to operate without failure for a period much longer than the life of ordinary tubes and also to withstand severe mechanical stress during launch. The expected performance of satellite tubes is discussed in Section 4.2 below.

The solar cells, although as numerous as the transistors and diodes, are expected to fail due to "wear-out" from radiation damage. The expected life of these components is discussed in Section 4.3.

4.1 *Transistors and Diodes*

As indicated previously, the reliability of a component in the final analysis is limited both by the design of the component and the care with which it is manufactured. The attention to design and manufacture is particularly important in the case of transistors and diodes which are both delicate and particularly sensitive to contamination, yet are required to exhibit failure rates comparable to those of the more rugged, passive components. Mechanical techniques have been developed whereby small semiconductor wafers can be bonded to headers and even smaller leads connected between the wafers and the headers, such that the resulting structure will easily withstand the mechanical shock and vibration experienced during the launch of a satellite and the temperature cycling that may be experienced while in orbit. Final cleaning and sealing techniques have also been developed which insure a degree of initial cleanliness and subsequent protection from outside contamination, such that adequate reliability for satellite applications can be achieved.

Table II outlines the complete reliability testing program proposed by Bell Laboratories for providing transistors and diodes for satellite applications. The first step is to insure that the design itself has adequate reliability potential. In order for a design to qualify for satellite use, it must pass mechanical tests which represent conditions more rugged than will be experienced during launch. The devices are further subjected to electron and proton bombardment simulating many years exposure to Van Allen radiation. Finally, devices are subjected to reliability evaluation to determine the reliability potential of the design.

The second step, that of screening and pre-aging, is designed to eliminate those few remaining freaks that were not eliminated by quality control. These tests include mechanical shock and vibration tests to eliminate weak components. In the reliability portion of these tests, a sample from the particular manufacturing lot is tested at increasing temperatures until all devices in the sample have failed. The median temperature for failure and the distribution of failures with temperature, when compared with similar figures for previous manufactured lots,

TABLE II—RELIABILITY PROGRAM FOR SATELLITE TRANSISTORS AND DIODES

1. Design Qualification Tests	
Mechanical	
Temperature cycling	−65C to +85C (−120C to +40C for blocking diodes)
Temperature-humidity cycling	
Shock	2,000 g
Centrifuge	5,000–10,000 g
Vibration	60g, 100–2,000 cycles
Radiation	
Reliability	
Accelerated aging	
Life testing	
Field experience	
2. Screening and Pre-aging	
Mechanical	
Centrifuge	2,000 g
Temperature-humidity cycle	
Tap or shock	
Reliability	
Accelerated temperature sample	
High-temperature aging	
3. Life Test and Selection	
Reliability	
System simulation and selection	

indicate whether or not there are major differences from previous lots. In addition, all the devices,that may be used in satellites are subjected to a short period of high temperature aging. Since, as discussed later, aging is accelerated by raising temperature, this pre-age eliminates many devices that otherwise would have exhibited unusually early failure.

The third step consists of choosing from the components that have passed step two, a number many times greater than the number that are finally to be used, and putting them on life test for six months under power and temperature conditions simulating those anticipated in operation. The duration of this test, which ideally should be a substantial fraction of the design life of the system, is frequently limited by economic factors or by the time available prior to the system's operation. During the life-test period, the characteristics of the components are measured at frequent intervals. The components needed for the system are chosen on the basis of their performance during the life-test period. If proper choices have been made, the components used should be ones which have shown no change in characteristics.

Steps 2 and 3 in this program are intended to insure that the components selected are truly representative of the design and do not include any freaks. Assuming these steps to be successful, the most significant portion of the program in determining system performance is the evaluation in step 1 of the reliability potential of the product. Since the reliability required is in the neighborhood of a few failures per 10^9 hours, this reliability evaluation can involve tens of thousands of components for tens of thousands of hours. It is with the object of reducing the numbers and times involved that considerable emphasis has been put on the development of accelerated aging techniques.[3,4,5] The results of a typical accelerated aging experiment are shown in Fig. 3. Plotted in the figure is the median life of a germanium transistor as a function of the temperature at which the transistor is operated. The data shown as solid points were obtained for some germanium transistors manufactured by the Western Electric Company in 1958. The temperatures at which the transistors were tested range from 100°C to as high as as 350°C, while the range in time to median failure is from about 20 minutes to just over 1 year, nearly 5 decades. The fact that the points fit a straight line on a $1/T$ versus log time plot suggests that raising the temperature is accelerating some failure mode which can be characterized by an activation energy. It has been found that within experimental error, the apparent activation energy is the same for all germanium transistors and, in addition, that there is a single but slightly different activation energy for all silicon transistors and diodes. The

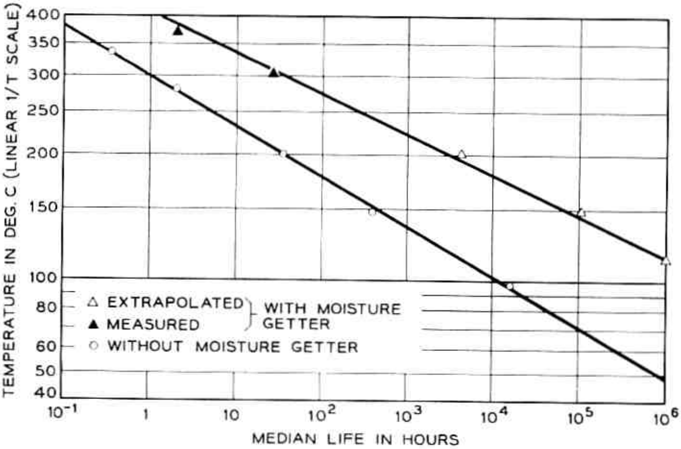


Fig. 3 — Results of a typical accelerated aging experiment on germanium transistors.

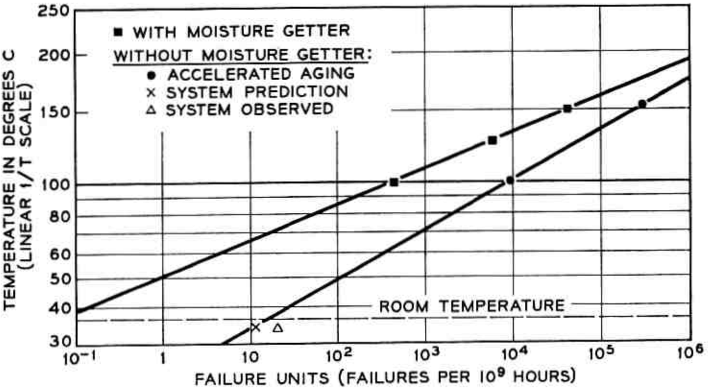


Fig. 4 — Failure rate vs temperature for germanium transistors.

triangles in Fig. 3 are for transistors manufactured by the Western Electric Company more recently. It is apparent that substantial improvements have been made at least in the high-temperature performance of the product. The data in Fig. 3 are for the median life. In performing the accelerated aging experiments, one also obtains the distribution of failures in time for a given temperature or, alternatively, in temperature for a given time. It is found that these distributions have the same shape, i.e., log normal in time* and normal in temperature, for all transistors and diodes. The widths of the distributions do not change with temperature for a given device type, that is, for fixed design and manufacturing procedure. This uniformity of failure distribution gives further confidence that raising temperature is accelerating a failure mode characteristic of the product.

Knowing the variation of median life with temperature and the distribution of failures in time for a fixed temperature, it is possible to derive a more useful plot for the systems designer, that of failure rate against temperature as shown in Fig. 4. The points are for the older transistors from the previous figure. A straight line is observed in the plot of $1/T$ against log failure rate. Extrapolating the line to room temperature, one would predict a failure rate of 10 per 10^9 hours for these transistors. The prediction of a failure rate of 10 per 10^9 hours from the acceleration curve of Fig. 4 is, however, liable to be optimistic because there is no guarantee that the curve does not dip below the straight line for times greater than the longest at which a measurement was made. There is no guarantee that in raising the temperature we are

* This is an example of a component that in the region of low failure rate does not exhibit the exponential failure distribution usually assumed.

accelerating all the failure mechanisms or even a guarantee that we are accelerating the most important failure mechanism at operating temperatures. For example, although one might expect that raising the temperature would increase the rate of reaction between the germanium surface and any water vapor inside the transistor can, one has no reason to suspect that elevated temperature would affect the occurrence of a short-circuit caused by a metal chip falling between emitter and base contact.

The accelerated aging curve, when extrapolated to room temperature, indicates the potential reliability of the design, and in the final analysis one must depend upon laboratory tests or field experience under operating conditions. The triangle on Fig. 4 shows the failure rate observed in the field trial of a new system using about 40,000 of these same transistors for about 10,000 hours. It is encouraging that the failure rate is only a factor of about 2 higher than that predicted from accelerated aging, and particularly so since the system failure rate includes failures due to mishandling and is for devices which were subjected to no special selection. It is therefore reasonable to estimate that the failure rate for these older germanium transistors, when properly handled and selected in a manner proposed for satellite use, would lie somewhere in the neighborhood of 10 to 20 per 10^9 hours.

The line through the squares in Fig. 4 is the accelerated aging curve for the more recent Western Electric product. Note again that there is a substantial improvement. The accelerated aging curve for recent silicon transistors and silicon diodes does not differ significantly from that for germanium transistors. With such an improvement in the reliability potential of the product, and with careful pre-aging and selection, one is confident that failure rates substantially lower than 10 per 10^9 hours are now achievable and that they may well be lower than 1 per 10^9 hours. However, complete confirmation of this prediction will have to await results of field trials.

The acceleration curves serve to emphasize the importance of conservative circuit design in the achievement of high reliability. It is seen frɔm the slope of the curves that failure rate increases very rapidly with temperature. It is therefore important that power dissipation in the device be maintained sufficiently low that temperature rise above ambient does not impair reliability. It is equally important that the ambient temperature be maintained at a suitably low value.

4.2 *Traveling-Wave Tubes*

Fig. 5 is a photograph of the traveling-wave amplifier under development at Bell Telephone Laboratories for use in experimental communi-

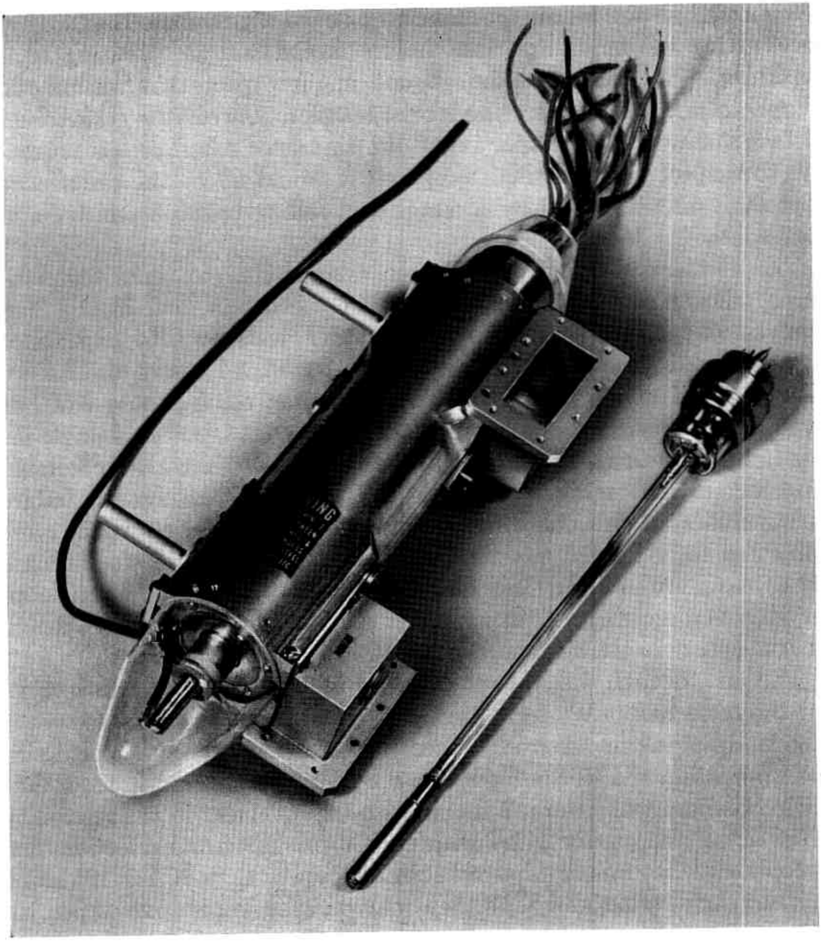


Fig. 5 — Traveling-wave amplifier under development for satellite use.

cation satellites. Table III lists the more important characteristics of this tube. Before discussing the performance and reliability of the M4041 satellite traveling-wave tube, a few words are in order on the reasons for selecting traveling-wave tubes to provide the output power in the satellite. It would appear that if a solid-state device could produce several watts at a few thousand megacycles, it would be, because of its small weight and potential reliability, an obvious choice over the traveling-wave tube. To date, however, schemes for generating power at several thousand megacycles using solid-state devices — harmonic

TABLE III—SATELLITE TUBE CHARACTERISTICS M4041 (7/7/61)

Operating point	0 dbm input saturated output
Output power (minimum)	3.5 w
Gain (at saturation)	35.5 db
Gain (low level)	41 db
Anode voltage	1770 volts
Helix voltage	1540 volts
Collector voltage	740 volts
Cathode current	17.0 ma
Cathode current density	85 ma/cm^2
Collector power (including helix and anode)	12.5 w
Heater power	1.5 w
Weight	7.1 lbs.

generators, for example — operate at efficiencies very much lower than that of a traveling-wave tube, even when heater power is included. The weight of the additional solar cells needed to provide power for the solid-state device would more than offset the decrease in weight from that of a traveling-wave tube. The weight penalty for extra power is particularly severe for satellites subject to Van Allen radiation, where account must be taken not only of the weight of the solar cells and their mounting but also of the necessary protective covers. The higher gain of the traveling-wave tube gives it a distinct advantage over other tubes such as triodes, which would require at least two stages and, through consequent loss of efficiency, lead again to greater over-all weight. The high efficiency of the traveling-wave tube results from the distinct separation between the microwave interaction region and the beam formation and collection regions. After the microwave interaction takes place, the beam is allowed to enter a region of retarding field, where the beam is slowed before collection. This is usually done by depressing the collector voltage below that of the helix, as shown in Fig. 6. Since very little current is intercepted on the helix and the anode, the input power is very nearly proportional to the collector voltage. By depressing the collector voltage, efficiencies as high as 39 per cent have been achieved and 36 per cent is typical. When the power required by the cathode heater is included, this value falls to typical value of 31 per cent. A second effect of collector depression is that ions generated between the anode and the collector will flow to the collector and not to the cathode. This results in a substantial decrease in the possible ion current bombarding and consequently damaging the cathode.

The traveling-wave amplifier for a satellite must be a new design in order optimally to meet the specific needs of the system. With any

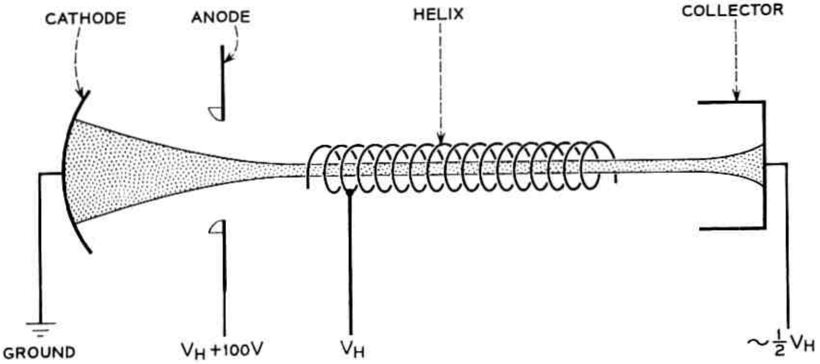


Fig. 6 — Traveling-wave tube circuit with depressed collector.

reasonable time scale, it is not possible to carry out a long-term evaluation of tube life, nor is it possible to do shorter experiments on very large numbers of models as is done with semiconductor devices. It is therefore necessary from the viewpoint of reliability to employ a design closely derived from experience gained with previous tubes and to utilize a "pedigree" approach in the assembly process. These earlier tubes include the pentodes used in telephone submarine cables,[6] the traveling-wave tubes used for microwave transmission at 6 kmc[7] and the rocket-borne traveling-wave tube used in a Bell Telephone Laboratories missile guidance system.[8] The salient features of these tubes are discussed in the next few paragraphs.

The submarine cable tube, the 175HQ, was the first tube designed to meet long-life reliability requirements somewhat similar to those encountered in satellite work. The failure pattern for this tube was found to agree with that shown in Fig. 1. The dominant wear-out mechanism in this case was determined to be the deactivation of the cathode, an effect which increases rapidly with increasing cathode temperature. Design information was developed which permitted the choice of a cathode temperature low enough to insure the desired life of the tube. The techniques of quality control to eliminate manufacturing freaks, and of life test and selection to insure the minimum random failure rate, were used extensively on this tube. As a result, the tubes that have been manufactured and put into operation in submarine cables easily meet the systems requirements. For example, Fig. 7 shows the accumulated tube life of the tubes in operation to date in submarine cables. There are now over 1600 tubes in such operation, some for as long as five years, with an accumulated life of 49 million tube-hours and no failures. It is on the basis of this evidence that it is believed possible to make long-life tubes and, in particular, to eliminate failure due to cathode deactivation.

The second tube of interest is a 6 kmc traveling-wave tube used as a ground-based microwave repeater, the M1789, now the WECo 444A. This traveling-wave tube was the first designed by Bell Telephone Laboratories specifically for long life, and it used many of the design principles and many of the selection techniques developed for submarine cable tubes. This tube also was designed to operate with a depressed collector. A little over four years ago, twelve of these tubes were placed on life test at their normal operating power of 5 watts. Table IV shows the accumulated hours on each of these tubes as of May, 1961, at which time there had been no tube failures. On the basis of this experience and the fact that the satellite traveling-wave tube has been designed to have a substantially lower cathode loading and cathode temperature than the 6 kmc tube, the satellite tube has an expectation of a life considerably in excess of four years.

The third tube is a traveling-wave tube designed to operate in the Bell Telephone Laboratories Command Guidance System, the M1958, now the 7116. In this system, the rocket to be guided contains a receiver, decoder and transmitter. There is a component count approximating 1000, including one traveling-wave tube. This system has been

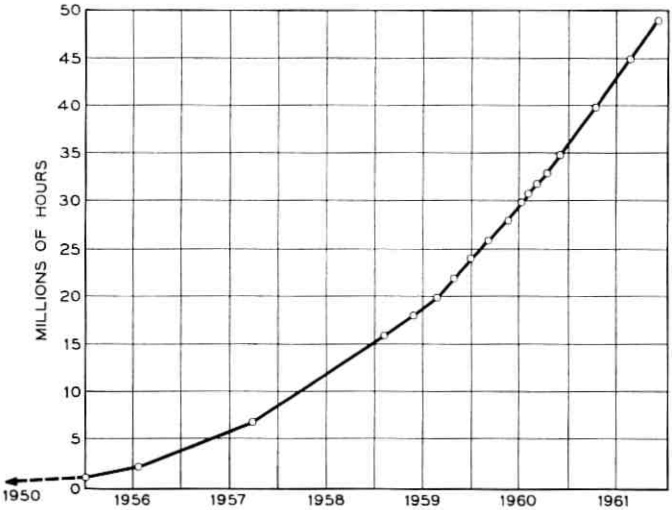


Fig. 7 — Operational life of electron tubes in undersea cable system repeaters.

TABLE IV—M1789 TRAVELING-WAVE TUBE LIFE TEST

Tube Number	Accumulated Hours 5/1/61
BC-856	39502
BC-1342	39630
BC-1363	39319
BD-14	39256
BD-660	39401
BH-69	39256
BH-208	33994
BH-413	37813
BH-464	36840
BH-559	35394
BS-41	36615
BS-102	34352

used in the guidance for about one-third of the U.S. satellites now in orbit. It was used, for example, with Echo I and with the three Tiros satellites. There have been to date over fifty successive firings using this guidance package with no failure. Since the guidance system needs only to operate for a few minutes, it gives us little information on long term reliability. However, since it not only must survive launch but must also operate during launch, this performance is a very potent demonstration that traveling-wave tubes can be made rugged enough to withstand the strains of launch. It further demonstrates that an electronic system containing roughly the number and kind of components needed in an active satellite can also survive launch.

To summarize, then, it is known from experience with the submarine cable tube and with the microwave relay tube that traveling-wave tubes can be designed with a life expectancy considerably in excess of four years. The performance of the guidance tube demonstrates that techniques are available for making a traveling-wave tube sufficiently rugged to withstand launch.

4.3 *Solar Cells*

Communication satellite designs for the immediate future rely on silicon solar cells as the prime source of power. These cells will be subject to radiation in the Van Allen belt,[9] which consists of electrons with substantial densities at energies up to 1 mev and protons at energies as high as 100 mev. Fig. 8 is a map of the Van Allen belt on a plane containing the earth's magnetic axis. There is a peak in the electron intensity at an altitude of about 2000 miles, and a second peak at about 10,000 miles with a substantial density of electrons at intermediate altitudes.

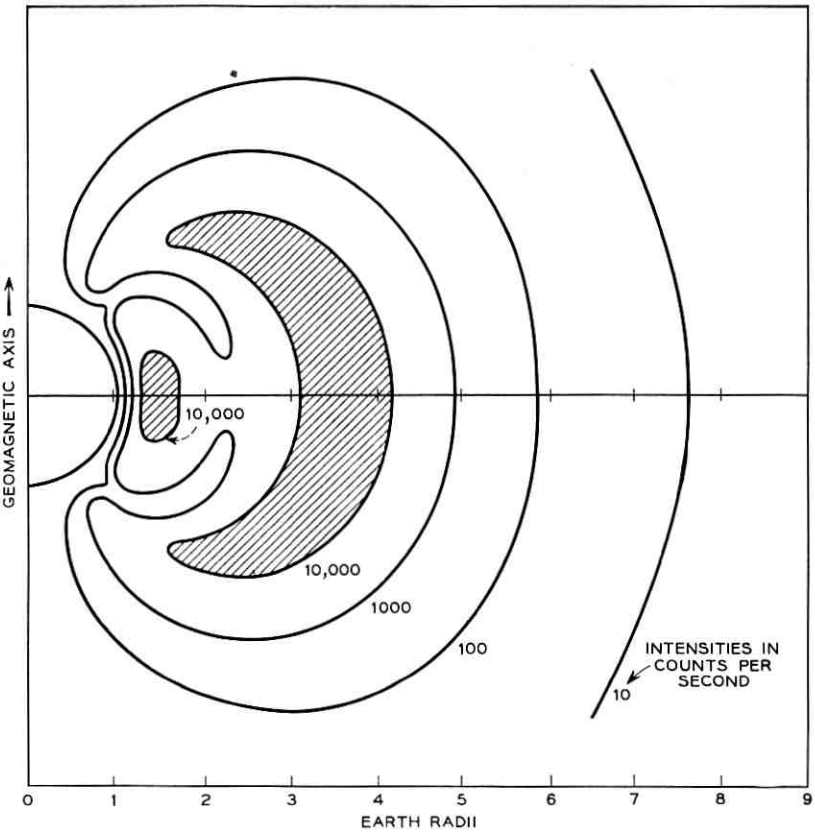


Fig. 8 — Map of Van Allen radiation belt in plane through earth's axis.

The protons, which are much less numerous, have a distribution which also peaks at around 2000 miles and falls off in some undetermined manner to negligible values beyond 10,000 miles. Bombardment of solar cells with particles of such energy results in a continual decrease of power output with time, at such a rate that this degradation could result in the failure of the power supply within the desired life of the satellite. Here then is an example of probable failure due to wear-out, in which case it is particularly important both to understand the mechanism of wear-out and to design the devices to minimize the effect. In this section, we discuss the effects of Van Allen belt radiation on solar cells, the means of designing cells to minimize the effects, and the predicted performance of such specially designed cells.

As shown in Fig. 9, a solar cell typically consists of a slice of n-type silicon with a thin p-type layer on one surface and contacts made to both surfaces. When light falls on the p-type surface, the photons penetrate the silicon to depths dependent upon their wavelengths and are absorbed with the creation of free carriers, hole-electron pairs, in the silicon. The free carriers created in response to the longer wavelength light are created deeper in the material. Some of the carriers move to the junction, and in crossing the junction create a current flow in the external circuit. Thus an illuminated solar cell is a source of electric power and has a voltage-current characteristic typically as shown in the figure.

In discussing the optimum design of solar cells, it is convenient to divide the generated carriers into two classes, namely those that are generated in the body of the material beneath the pn junction, and those that are generated in the surface layer above the pn junction. Those generated beneath the junction will reach it only if they are generated within a distance called the diffusion length, that is, the distance that generated carriers may move in the material before being annihilated by recombination. The diffusion length is a property of a

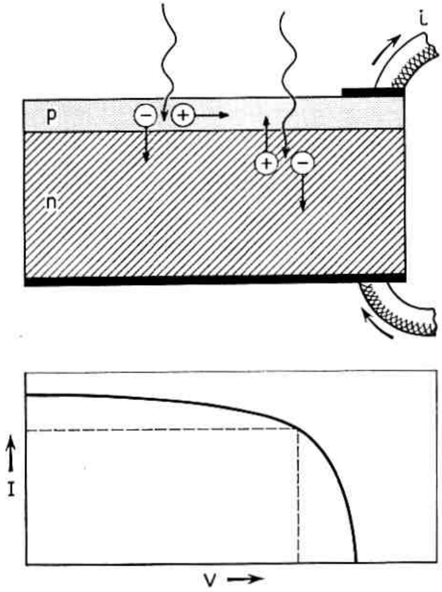


Fig. 9 — Solar cell construction and typical voltage-current characteristic.

particular material and depends critically upon its perfection and purity. For a solar cell to have the maximum efficiency, this diffusion length should be as long as possible in order that effectively all carriers generated beneath the junction may reach the junction and contribute to the output current. A somewhat different situation exists for the carriers generated in the surface layer. This layer is usually quite thin compared to a diffusion length. However, the surface of the semiconductor acts as a sink for carriers and thus competes with the junction for carrier collection. The net result is that the efficiency for collection of carriers generated above the junction is less than that for carriers generated below the junction. It is therefore desirable to minimize the thickness of the surface layer.

The perfect solar cell therefore would have a zero thickness of surface layer and an infinite diffusion length. A zero thickness surface layer, however, would lead to infinite series resistance. Obviously a compromise is necessary. Fig. 10 shows the distribution of carriers generated in silicon in response to sunlight. The plot gives the percentage of carriers generated beyond the value of the abscissa. It is seen that about 75 per cent of the carriers are generated below 1 micron depth, and that for a junction depth about $\frac{1}{4}$ micron, essentially all the carriers are generated below the junction.

When high-energy electrons or high-energy protons are incident on a silicon solar cell, they create local disorder in the crystal which results in a steady decrease of diffusion length with time. A simple theory for

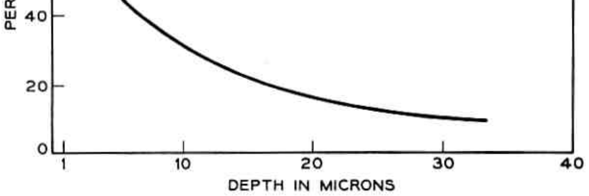


Fig. 10 — Distribution of free carriers generated in silicon in response to sunlight.

the degradation of diffusion length predicts that the diffusion length L should depend on the total flux Φ of electrons or protons according to the equation:

$$\frac{1}{L^2} = \frac{1}{L_0^2} + K\Phi \tag{3}$$

where L_0 is the value of the diffusion length before irradiation and K is a constant for a given energy of particle and for a given semiconductor. Hence, for large enough radiation fluxes, the diffusion length is inversely proportional to the square root of the flux. Fig. 11 shows a plot of diffusion length versus flux of 1 mev electrons. The experimental points were obtained by measuring the diffusion length in silicon after successive exposure to 1 mev electrons from a Van de Graaff generator. The line on Fig. 11 is a two-parameter fit of (3) to the experimental data. Similar results are obtained for proton bombardment.

As the diffusion length in a solar cell decreases with exposure to radiation, fewer and fewer of the carriers generated deep in the silicon are collected at the junction. Thus, the power output of the solar cell decreases. Since, as pointed out earlier, the depth of generation increases with the wavelength of light, the solar cell degrades initially by loss of response to the longer wavelength, i.e., the red light. This fact has a number of implications for the design of solar cells for use in the Van Allen belt. Firstly, since it is the blue response that is likely to be maintained, and this response involves the carriers generated closest to the surface, it is most important for satellite solar cells that the junction depth be minimized. Secondly, it is important that any antireflective

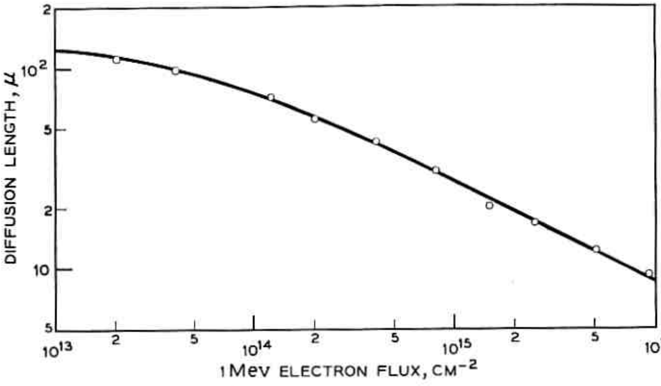


Fig. 11 — Diffusion length vs flux of 1 mev electrons.

coating be optimized for blue light, not for red. Initial good response to red light, which calls for long diffusion length, becomes of lesser importance.

It has been found by several investigators that the decrease of diffusion length in response to electron and proton bombardment is less rapid in p-type silicon than it is in n-type silicon.[10] For this reason, cells for satellite use are preferably made with a thin n-skin on a p-type body rather than the other way around. Fig. 12 is a schematic diagram of a solar cell designed at Bell Telephone Laboratories and incorporating the features just discussed.[11] It is made on a p-type silicon body with an n-layer $\frac{1}{4}$ micron thick. In order to produce such a thin layer with good properties, it is necessary to minimize surface damage. For this reason the surface used is given an optical polish. Such a thin layer tends to have high sheet resistance and calls for many contact fingers to minimize the effect of series resistance. Finally, the cell is given an antireflection coating of thickness designed to optimize the response to blue light.

Having designed a cell to minimize the effects of radiation damage, it is then necessary to consider what, if anything, can be done to shield the cells from the radiation. In the case of electrons, substantially all of which have energies of less than 1 mev, such shielding is practical using materials like quartz or sapphire. Fig. 13 shows the measured degradation of the short-circuit current of variously shielded solar cells after electron bombardment corresponding to increasing time in the Van Allen belt. The shield thicknesses are represented as g/cm². It is seen that over the range for which the measurements were made — which

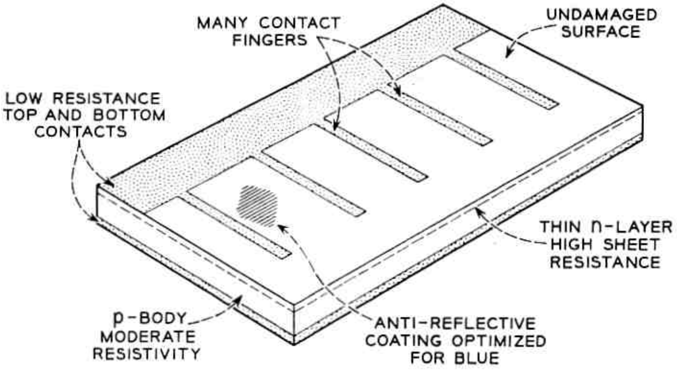

MANY CONTACT FINGERS
UNDAMAGED SURFACE
LOW RESISTANCE TOP AND BOTTOM CONTACTS
THIN n-LAYER HIGH SHEET RESISTANCE
p-BODY MODERATE RESISTIVITY
ANTI-REFLECTIVE COATING OPTIMIZED FOR BLUE

Fig. 12 — Structure of Bell Laboratories solar cell for satellite use.

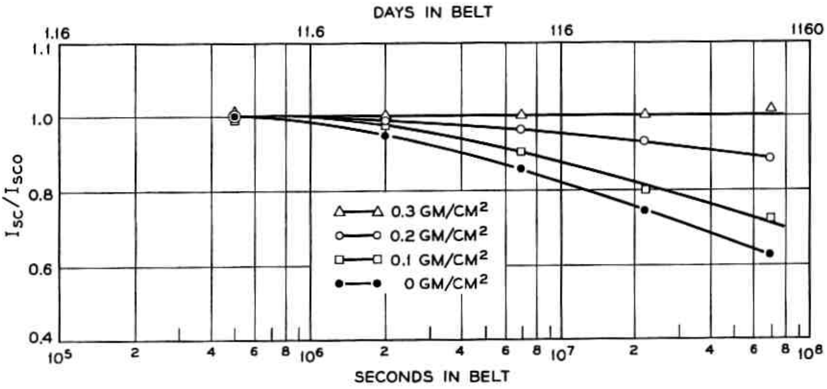


Fig. 13 — Solar cells with various shielding: measured degradation of short-circuit current after electron bombardment.

was equivalent to two years in the Van Allen belt — the effect of electrons was eliminated by the use of 0.3 g/cm² of shielding. Shielding of protons, which are much more energetic, would require intolerable weights of material. However, the 0.3 g/cm², which eliminates the electron damage, does provide some reduction in the proton damage.

Fig. 14 is a plot of the anticipated power output of the solar cells

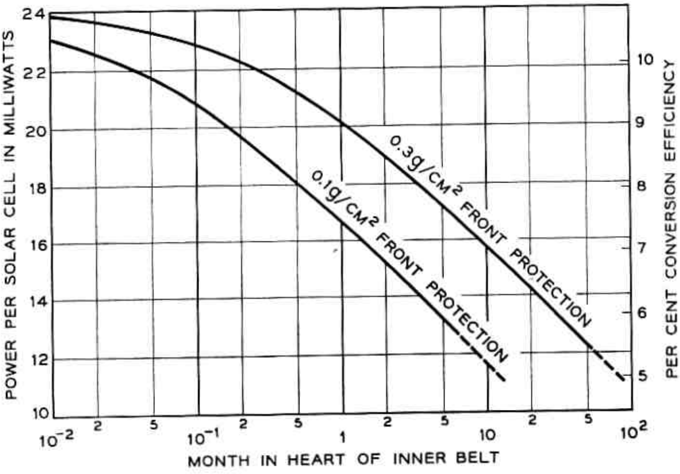


Fig. 14 — Anticipated power output of solar cells as function of time in Van Allen belt; with present data, error factor may be as great as 3 in time.

shown in Fig. 12 as a function of months in the heart of the Van Allen belt. The curves were obtained by estimating the densities and energy distributions of electrons and protons in the Van Allen belt and subjecting the cells to electron and proton bombardments simulating Van Allen conditions. There may be considerable errors in the estimation of Van Allen radiation and, as a result, the time to a given degradation may well be in error by a factor as great as 3. It should further be noted that the curves have been calculated for the case of a satellite that spends all its time in the Van Allen belt, and this is certainly pessimistic. A satellite in a circular polar orbit, for example, would spend approximately $\frac{1}{6}$ of the time in the Van Allen belt.

The most significant feature of the curves in Fig. 14 is that the plot of power output per solar cell versus log time is approximately linear after initial degradation. This dependence is consistent with the anticipated variation of diffusion length with flux, Fig. 11, and the distribution of carriers generated in the silicon, Fig. 10. The degradation with time becomes progressively less severe at longer times. Thus, for the case of 0.3 g/cm^2 protection, the output after 10 months has dropped from an initial value of 24 mw to about 16 mw while at the end of 100 months it has dropped further only to 11 mw. This additional decrease in power output for a factor of 10 increase in time could be compensated for by a 50 per cent increase in the number of solar cells. It appears then that provided there has been no gross underestimate of the nature and effect of the Van Allen belt radiation, solar cell power can be provided for a design life of five years and that the design life could be increased without excessive penalty. The curves also illustrate the design choices that can be made in selecting the mass of front protection. It is seen that for a given power output per cell, a factor of 3 increase in weight of protection yields about a factor of 5 improvement in life. However, the same improved life for a given power output could be achieved by retaining the lighter front protection but increasing the number of cells by 30 per cent. Just which is the best design of front protection thickness will depend on the particular satellite under consideration. For the case of the experimental satellite being designed at Bell Telephone Laboratories, a front protection consisting of 0.3 g/cm^2 of sapphire was found to be the best choice. Fig. 15 is a photograph of some solar cell modules with and without the sapphire protection.

The solar cell is yet another example of a component which can give adequate life performance only if the component is properly designed and used conservatively. In this case, conservative use involves paying

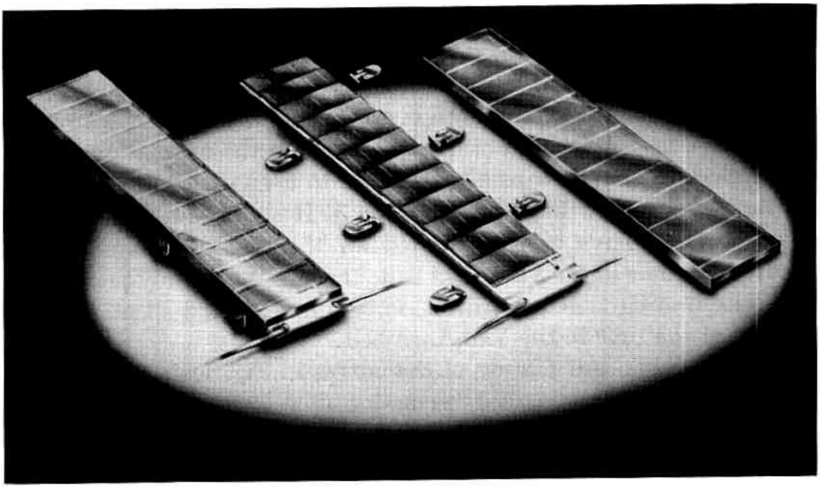


Fig. 15 — Photograph of solar cells, without protection (center) and with sapphire shields.

the weight penalty of sufficient radiation protection and increasing the number of solar cells to allow for some inevitable loss of power output per cell in response to radiation.

V. CONCLUSIONS

Returning to Table I, it is seen that the failure rate of 20 per 10^9 hours chosen for transistors in case I is probably a conservative figure. This degree of reliability has already been observed in the field on older devices that did not have the benefit of more recent design improvements and that were not life tested, selected and carefully handled as devices would be for satellite use. With proper selection and handling care, these older devices would almost certainly meet the requirements for case II and possibly for case III. The results of accelerated aging of the newer product lead to predictions of at least one order of magnitude improvement in transistor reliability. Assuming that at least some of this improvement will be realized under operating conditions, one expects that transistor performance is adequate for case III. The reliability of diodes, which approximates that for transistors, is similarly adequate for case III. Should transistor and diode failure rates indeed turn out to be in the region of one per 10^9 hours, then more complex satellites could be designed with life expectancy much longer than five years.

It further appears that traveling-wave tubes can be made that will survive launch and should not limit the life in orbit. Finally, even under the most pessimistic assumptions as to the nature of the Van Allen belt, solar cell power plants can be provided, at a weight penalty, to meet the required life. More precise design of solar cell power supplies will only be possible when more precise and extensive data are available on the nature of the Van Allen belt.

Adequately reliable communication satellites can therefore be made, provided they incorporate components of proven integrity which are used in a conservative design. The use of components of proven integrity involves expense for high-quality design, careful manufacture and painstaking selection. The use of such components does not permit the performance advantages that might be gained with use of developmental components. In the final analysis, conservative design leads to more weight per given function. Typical examples are the increased weight of a rugged traveling-wave tube, the weight of solar cell protective covers, the weight of additional solar cells to allow for the inevitable degradation in the Van Allen belt, and the additional weight of circuitry designed with ample margins.

Hence, limitations of weight in orbit and requirements of long life in orbit both result in a limit on the complexity of the satellite. Communication satellites in the immediate future must be simple. As higher component reliability is demonstrated and as improved vehicles permit greater payloads, so can the complexity of the satellites increase.

VI. ACKNOWLEDGMENTS

In preparing this paper, the author has drawn heavily on the work and advice of many colleagues at Bell Telephone Laboratories. He is particularly indebted to B. T. Howard and to D. S. Peck.

REFERENCES

1. Glaser, J. L., The Design of Medium Height Random Orbit Satellite Systems, NEREM Record — 1961, p. 39.
2. The Bell System Technical Journal, **40,** July, 1961, complete issue.
3. Peck, D. S., A Mesa Transistor Reliability Program, Solid State J., Nov./Dec., 1960.
4. Peck, D. S., Semiconductor Reliability Predictions from Life Distribution Data, in *Semiconductor Reliability*, edited by Shwop, J. E. and Sullivan, H. J., Engineering Publishers, Elizabeth, N.J., 1961.
5. Dodson, G. A. and Howard, B. T., High Stress Aging to Failure of Semiconductor Devices, Proc. Seventh National Symposium on Reliability and Quality Control, Philadelphia, Penn., Jan., 1961.
6. McNally, J. O., Metson, G. H., Veazie E. A., and Holmes, M. F., Electron Tubes for the Transatlantic Cable System, BSTJ, **36,** 1957, p. 163.

7. Laico, J. P., McDowell, H. L., and Mosler, C. R., A Medium Power Traveling-Wave Tube for 6000 mc Radio Relay, B.S.T.J., **35,** 1956, pp. 1285.
8. Bradford, C. E., and Laico, J. P., Ruggedized Traveling-Wave Tubes for Missile Use, Proc. of 1960 Electronic Components Conference, Washington, D.C., pp. 91–95.
9. Van Allen, J. A., J. Geophys. Res., **64,** 1959, p. 1683.
   Holly, F. E., and Johnson, R. G., J. Geophys. Res., **65,** 1960, p. 771.
   Freden, F. C., and White, R. S., Phys. Rev. Lett., **3,** 1959, p. 9 and J. Geophys. Res., **65,** 1960, p. 1399.
   Naugle, J. E., and Fichtel, C. E., Am. Phys. Soc. Bull. II, **6,** 1961, p. 53.
   Singer, S. F., *Space Research,* North Holland Publishing Co., Amsterdam, 1960, p. 797.
   Naugle, J. E., Nucleonics, **19,** April, 1961, p. 89.
10. Madelkorn, J., et al, A New Radiation-Resistant High-Efficiency Solar Cell, delivered at a meeting on "Radiation Damage to Semiconductors by High Energy Protons," sponsored by N.A.S.A., Wash., D.C., Oct. 20, 1960. Available from U.S. Army Signal Res. and Dev. Lab., Fort Monmouth, N.J.
11. Smits, F. M., Smith, K. D., and Brown, W. L., Solar Cells for Communication Satellites in the Van Allen Belt, presented at 1961 Convention, The British I.R.E., Oxford, to be published in Proc. of the British I.R.E.

Automatic Stereoscopic Presentation of Functions of Two Variables

By BELA JULESZ and JOAN E. MILLER

(Manuscript received September 21, 1961)

Spatial models of functions of two variables are often a valuable research tool. Nomograms and artistic relief drawings in two dimensions are difficult to prepare and still lack the direct impact of a spatial object. It has been demonstrated (see Ref. 2) that objects with a randomly dotted surface permit the determination of binocular parallax and, thus, can be seen in depth even though they are devoid of all other depth cues. This random surface presentation has the advantage that the random brightness points can be evenly and densely placed, whereas the classical contour-line projection at equally spaced heights may leave empty spaces between adjacent contour-lines. A digital computer is used to generate the three-dimensional image of a given $z = f(x, y)$ function and to wrap its surface with points of random brightness. The stereo projections of the function are obtained and, when viewed stereoscopically, give the impression of the three-dimensional object as being viewed along the z-axis. The random surface prevents the accumulation of clusters of uniform regions or periodic patterns which yield ambiguities when fused. Two stereo demonstrations are given of surfaces obtained by this method.

I. INTRODUCTION

Pictorial representations and visual displays are invaluable aids in conveying scientific or technical information. In particular, the problem of presenting three-dimensional data is of interest both from the standpoint of its wide range of applicability and the difficulty involved in the production of such representations.

The methods usually employed to present functions of two variables in the fields of applied mathematics, engineering, cartography, etc., fall into two categories: 1) two-dimensional and 2) three-dimensional displays. The first has the obvious advantage of being suitable for the printed page, thus permitting a wide circulation for the information so presented. The techniques of nomography, orthography, isarithmic (contour line) representations (see Fig. 1) and relief drawings (see Fig. 2) belong to this category and are widely used despite the expense and difficulty in their preparation. However, the greatest objection is perhaps the failure of such displays to match the capabilities of human observers, who are equipped to perceive a three-dimensional object in depth. The second category — that of spatial models or sculpture — answers this objection, but these models are usually much too difficult to execute and much too limited in their applicability.

There is, therefore, a need for a technique which a) eliminates the tedious effort required of draftsmen in producing such displays, b) presents displays complete with the spatial effects inherently belonging to three-dimensional objects and appreciated by human observers, and c) generates displays suitable for the printed page. This first requirement has already been met for two-dimensional representations by the de-

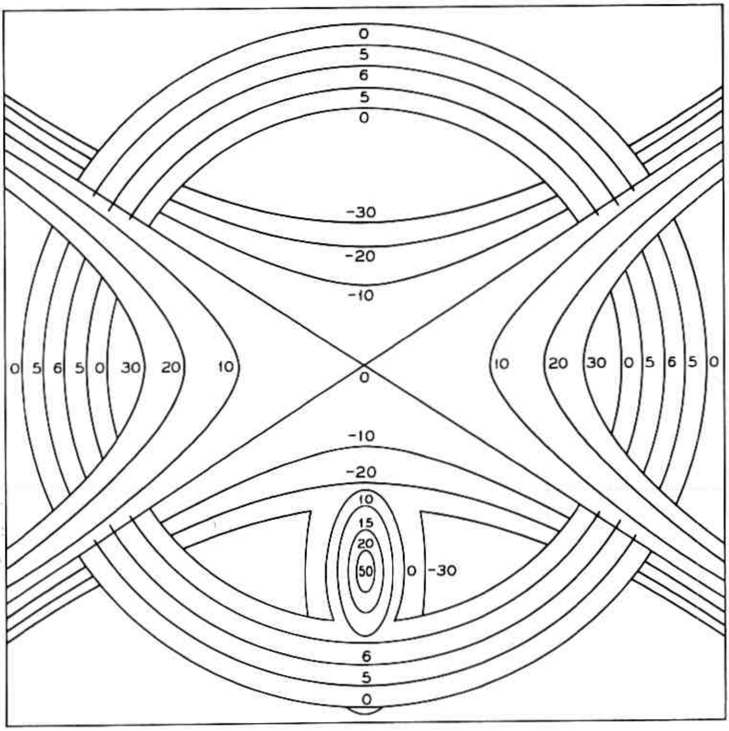


Fig. 1 — Isarithmic (contour-line) drawing (Example 1).

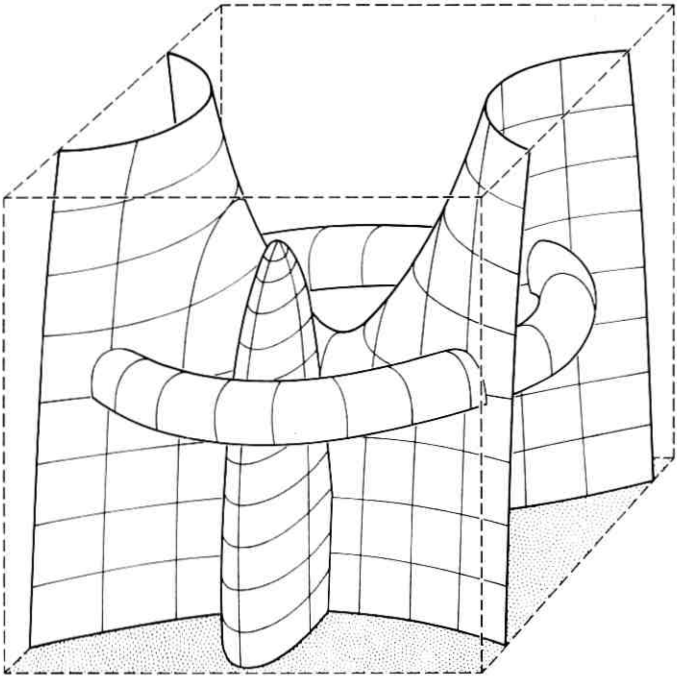


Fig. 2 — Relief drawing (Example 1).

velopment of oscilloscopic displays which automatically project onto the screen of the oscilloscope the object surface defined by one dependent voltage and two independent voltages.[1] The second and third requirement, however, seem of particular interest, and therefore this paper discusses a method employing a computer to make stereoscopic presentations of functions of two variables.

II. METHOD

The technique to be described here may be outlined as follows: the three-dimensional image of a given function $z = f(x,y)$, which is supplied as a table of corresponding x, y and z values, is stored in a digital computer. The computer is programmed to generate a stereo picture pair which, when fused, gives the subjective impression of the three-dimensional object as being viewed along the z-axis perpendicular to the base plane of x and y. This procedure for obtaining the stereo projections of an object can be considered in three parts: 1) defining the function to be presented as a three-dimensional object, 2) "wrapping" the object with a textured surface, and 3) generating a stereo pair by taking proper projections of the object.

In practice, the variables x, y and z must be evenly sampled with a given resolution. Therefore, the object can be defined only by approximation, and the various approximations differ in their fine structure. The classical method is the contour-line approach shown in Fig. 3(a)

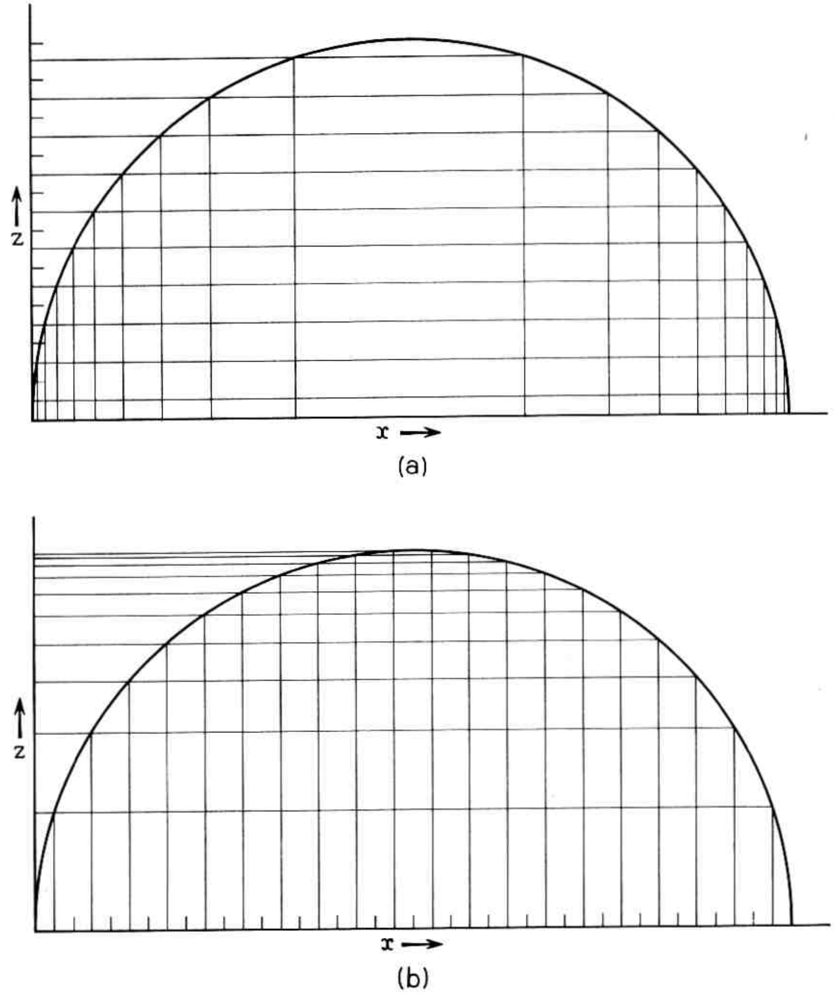


Fig. 3 — (a) Surface definition with even z-axis quantization (contour lines); (b) surface definition with even x, y plane quantization.

Here the z values are quantized into equal levels, and to any such z level the corresponding x and y values are taken (rounded to their nearest sample). This approximation yields uniformly distributed z values but uneven coverage of the x and y values. If the surface rises sharply toward the observer, the (x,y) points become densely packed, whereas if the surface becomes flat, these points become farther apart, resulting in gaps. Another possible approximation is shown in Fig. 3(b). Here the evenly sampled x and y values are taken and the corresponding z values are determined and rounded to their nearest sample. Hence, the surface to be displayed is defined by a dense covering of points obtained by projection up from the base plane. This second method of approximation is chosen since the object is to be viewed from above, and since the dense covering will result in efficient use of the available stereo picture area. In the case of multiple-valued functions or several functions considered in one display, the projection is made onto the maximum z value, which is the point closest to the observer.

The surface of the object is thus defined but in a rather abstract sense. In order that the object be visible, brightness values must be assigned to every surface point. The use of identical brightness values for all points would yield a surface of homogeneous texture when viewed perpendicular to the base plane. Such a surface would have no patterns, shadows, or brightness changes due to different angles of reflection; that is, it would have neither monocular nor binocular depth cues and thus would be inappropriate for the purpose. Therefore, to obtain depth cues each point must be printed at varying brightness levels. It has been shown[2] that stereo picture pairs comprised of points of random brightness and thus devoid of all cues except binocular parallax can be perceived in depth when fused. Therefore, it is sufficient to assign randomly to each point (x,y,z) a brightness level. The brightness selection on a random basis is a simple procedure, eliminating any consideration for appropriate monocular cues, and has the further advantage of avoiding periodicities and regions of ambiguities. That is, a point domain seen by the left eye may be fused with any periodically repeating domain seen by the right, if such exists, thus producing confusion as to the correct binocular parallax. Therefore, random brightness patterns are used to produce unique point domains which can be fused unambiguously. The question of how many brightness levels to use in the random selection is answered by the requirements of the system of output to be used. However, the use of few levels increases the probability of occurrence of any one level, and clusters of points of equal brightness can produce areas of indeterminate depth on the surface to be viewed. For a photographic output procedure requiring a small number of levels, it would be desirable, therefore, to apply rules to the random selection which would prevent these clusters. Also, a useful monocular cue could be provided by regulating the occurrence of certain brightness levels in a manner dependent upon the z-level of the point domain. Thus, there are many possible refinements to the basic procedure of giving texture to the surface by randomly assigning brightness levels.

The object has been defined and its surface has been invested with brightness levels, albeit random and uninformative when viewed monocularly. It remains now to produce a stereo pair in which the point domains are given the proper parallax shift. The calculations for two such pictures follow the simple formulas for projection, which are shown in Fig. 4. The center of projection is considered at a distance H from the base plane of the object. The centers of projection for the stereo picture pair are separated by a base distance B and are positioned symmetrically about the z-axis. The plane of the pair is at a distance F from the centers of projection. The projections for each point (x,y,z) of the surface where $z = f(x,y)$ onto the left and right members of the pair are then given by the relations

$$x_L = \left(x + \frac{B}{2}\right) \cdot \frac{F}{H - z} \qquad \text{and} \qquad x_R = \left(x - \frac{B}{2}\right) \cdot \frac{F}{H - z}$$

$$y_L = y \cdot \frac{F}{H - z} \qquad\qquad y_R = y_L$$

The total parallax for the point (x,y,z) is seen to be $\Delta = BF/(H - z)$ and is shared equally by the two pictures.

It should be pointed out that binocular parallax alone constitutes only the perception of relative depth. Without other depth cues it is not possible to determine absolute depth when fusing the pair obtained by the above projections. That is to say, the perceived z-scaling, which is some monotonic function of $\Delta = \text{const.}/(H - z)$, is obtained by some arbitrary selection for the value H. (In stereoscopic viewing the supplementary depth cues determine the absolute distance of the plane of the stereo pictures from the observer, which is subjectively substituted for H.) If the function $\Delta = \text{const.}/(H - z)$ is used, the parallax shifts will be similar to those experienced by the human optical system and thus will give rise to familiar percepts of z-scaling. Inasmuch as the perceived depth is some monotonic function of the binocular parallax, which is in turn a monotonic function of the height of the surface, it suffices to choose any monotonic function $\Delta = f(z)$. For example, if the range of z is limited, the function $\Delta = z$ gives a good approximation to

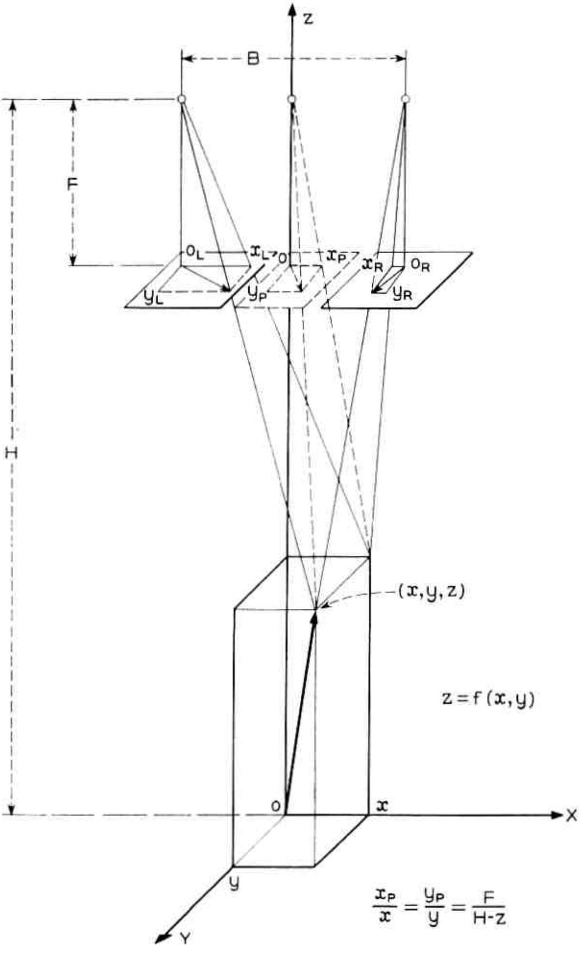


Fig 4 — Projection of an object onto a stereo pair.

the projection of Fig. 4. This function merely provides a different subective z-scaling. If a numerical z-scale is provided, which can be perceived in depth together with the surface to be presented, and if both are generated according to the same projection rules, then the problem of a correctly labelled stereoscopic projection is solved.

Consequently, the parallax shift can be computed, having selected a function $\Delta = f(z)$, which gives a new position to each point, and an identical brightness level can be assigned at random to the corresponding points in the left and right fields. The stereo pair then results by use of a suitable output medium, that is, a video transducer in which the x,y positions correspond to the deflection, and in which the brightness values correspond to the intensity of the beam in a cathode-ray tube display. Inasmuch as a digital system is used, the video transducer generates a sampled display. Since the projection can produce expansions and contractions of point domains on the surface or parallax shifts which are not integral multiples of the sampling intervals, over-sampling is required. This, however, results in great strain on the storage capacity of available computers and on the resolution requirements of video transducers. Therefore, it is necessary to make compromises by trading resolution in object definition for resolution in depth. In the present state of technology, however, there are devices available which will satisfy this requirement.

III. INSTRUMENTATION

The above steps were carried out by quantizing the base plane into 10,000 points with the scale on the x and y axes running from 1 to 100. An IBM 7090 computer was used to generate an array of corresponding values $z = f(x,y)$ and to assign a random number designating the brightness for each of the points. For simplicity, the function $\Delta = z$ was chosen for the parallax shift instead of the geometric projection and was applied in the x-direction only. That is, the coordinates of the point x,y in the left and right pictures were

$$\begin{aligned} x_L &= x + z/2 & \qquad & & x_R &= x - z/2 \\ & & \text{and} & & & \\ y_L &= y & & & y_R &= y. \end{aligned}$$

This corresponds to projecting the stereo pair with a cylindrical lens, the axis of which runs parallel to the y-direction. This position and brightness information was then written on digital magnetic tape and put into a General Dynamics S-C 4020 microfilm printer, which served as the output device for the stereo pair. Different brightness levels were achieved by randomly employing each of the sixty-four type characters available on the microfilm printer. The variation in density of each of the characters gave sufficient variation in brightness level and provided an efficient means for plotting brightness information. The grid size of the microfilm output was 1024 x 1024 and provided, therefore, an oversampling of ten to one for the chosen picture size. This oversampling gave enough stereo resolution for most applications. In order that the type characters did not overlap and totally obscure each other, the maximum shift permitted between two points was taken to be six micropositions. This limited the total parallax shift to twelve units. That is, the maximum angle of rise between two points on the surface which could be displayed was approximately 85°. The total range of the surface was further restricted by the locations of the peaks and valleys relative to the side boundaries of the grid. For the examples to be shown, a scaling on the z-values of about 60 levels running from -30 to $+30$ was chosen.

An alternative semi-automatic method of output using an optical system was also investigated. This technique resulted in the production of a solid model of the surface to be displayed, which was then photographed by a stereo camera to obtain the desired picture pair. The model was prepared in layers by printing the points belonging to each of the quantized z-levels on transparent glass slides as black and white dots. The slides were then stacked together in register to form a solid cube, where the width of the glass plates determined the scale factor for the z-axis. The prints for each level were obtained by writing the picture information on magnetic tape in digital form as computer output and by using a digital-to-analog converter and a slow-speed television monitor to produce oscilloscope displays, which were then photographed.[3,4,5] A secure mounting of the stack of glass slides in which the entire stack remained transparent was achieved by making an air-tight seal between each plate with a polyester resin having an index of refraction sufficiently near that of the glass.* This technique results, therefore, in a stereo pair in which the parallax shift corresponds to the geometric projection, and furthermore, gives rise to a solid model which is a desirable by-product.

IV. RESULTS

Example 1, generated and displayed automatically, is shown in Fig. 5, and can be perceived in depth when viewed stereoscopically. Viewing may be facilitated by use of Fresnel lenses accompanying the article cited in Ref. 2. The following three surfaces are presented:

1) the hyperbolic paraboloid,

$$\left(\frac{x-50}{30}\right)^2 - \left(\frac{y-50}{20}\right)^2 = \frac{z}{30},$$

with saddle point at (50, 50, 0),

* The slide mounting techniques were developed by R. A. Payne of Bell Telephone Laboratories.

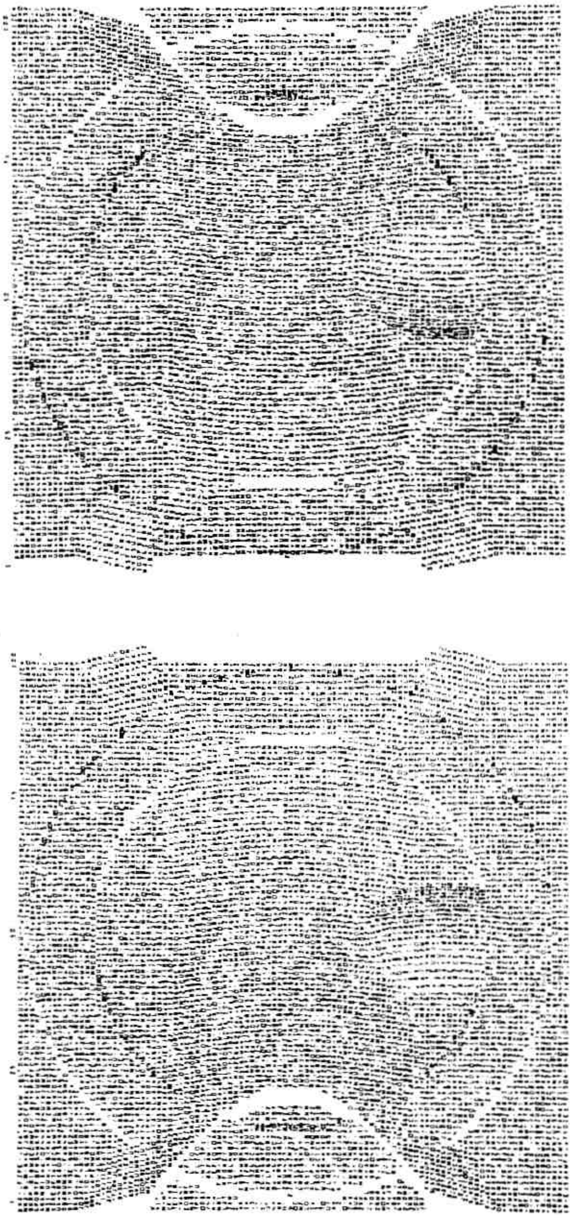


Fig. 5 — Automatic stereoscopic presentation of a function (Example 1).

2) the elliptical paraboloid,

$$\left(\frac{x-50}{10}\right)^2 + \left(\frac{y-79}{20}\right)^2 = \frac{-(z-20)}{50},$$

with vertex at (50, 79, 20), and

3) the torus,

$$(\sqrt{(x-50)^2 + (y-50)^2} - 42)^2 + z^2 = 6^2,$$

centered at (50, 50, 0) and having a radius of 6. (Conventional two-dimensional displays of Example 1 were given in Figs. 1 and 2.)

The reduction required for reproducing the stereo pictures here has made the resolution of individual type characters very difficult. For this reason, a presentation in depth of the numerical z-scale was omitted. However, a very effective display can be achieved, including the z-scale, by using a larger picture size.

In Fig. 6 the same surface is displayed by the optical method. Two

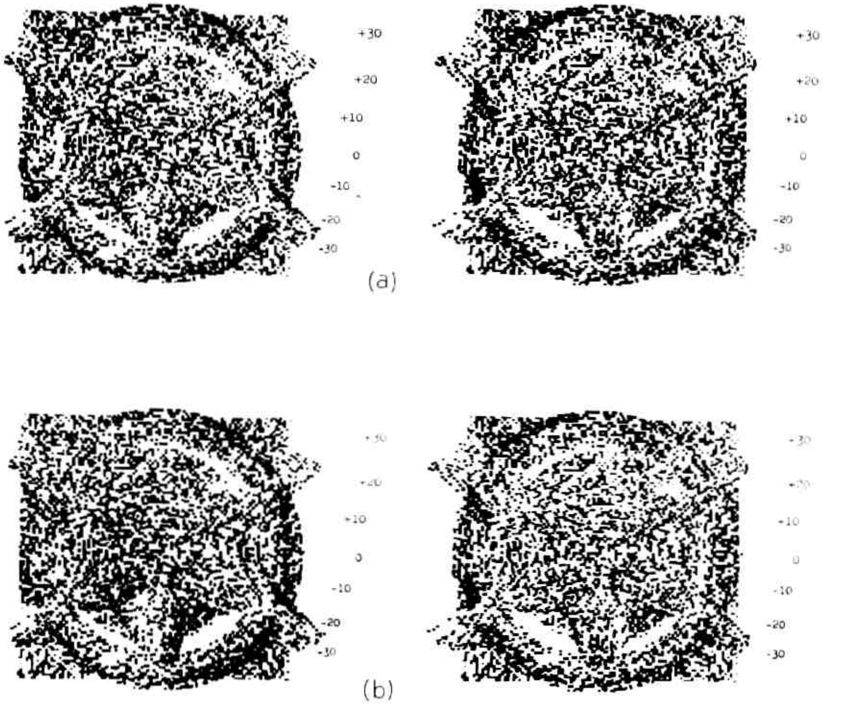


Fig. 6—(a) Semi-automatic stereoscopic presentation of a function (Example 1); (b) same as (a), but with increased z-axis scaling.

views are presented to show that variable scaling on the z-axis is possible. The amount of depth is determined by both the width of the glass plates and the base distance between the two lenses of the stereo camera (or distance between two positions of a single-lens camera). Fig. 6 (b) was produced with greater base distance, and consequently the surface has greater stretching in the z-direction. The numerals on the right edge of the displays indicating the z-levels were applied to the appropriate slides by hand and were not generated as picture material. It will also be pointed out that in this method, the points of the two pictures are more clustered and less uniform in distribution. This demonstrates the expanding and contracting of point domains produced by the transformation of projection and provides a helpful monocular cue. All displays are far more evenly filled with brightness elements, however, than if the contour-line method had been used.

Example 2 shown in Fig. 7 is that of a spiral given by the parametric equations

$$x = \rho \cos \theta + 50.5$$

$$y = \rho \sin \theta + 50.5$$

$$z = \frac{10}{\pi} \theta - 30$$

with

$$0 \leqq \theta \leqq 6\pi$$

$$\rho = 50 - \frac{15}{2\pi} \theta.$$

This presentation is another display from the microfilm printer, illustrating the procedure in its completely automatic form. Approximately one minute of time is required to generate and display the stereo information.

V. SUMMARY

A method for automatically presenting three-dimensional information in depth has been described. The advantages are threefold in that (*i*) such presentations make possible displays which are very difficult if not impossible to obtain by other means, (*ii*) they carry the spatial impact enjoyed by human observers, (*iii*) and they are suitable for the printed page. The technique has been outlined in three steps: definition of surface, texturing of the surface with brightness elements, and generation

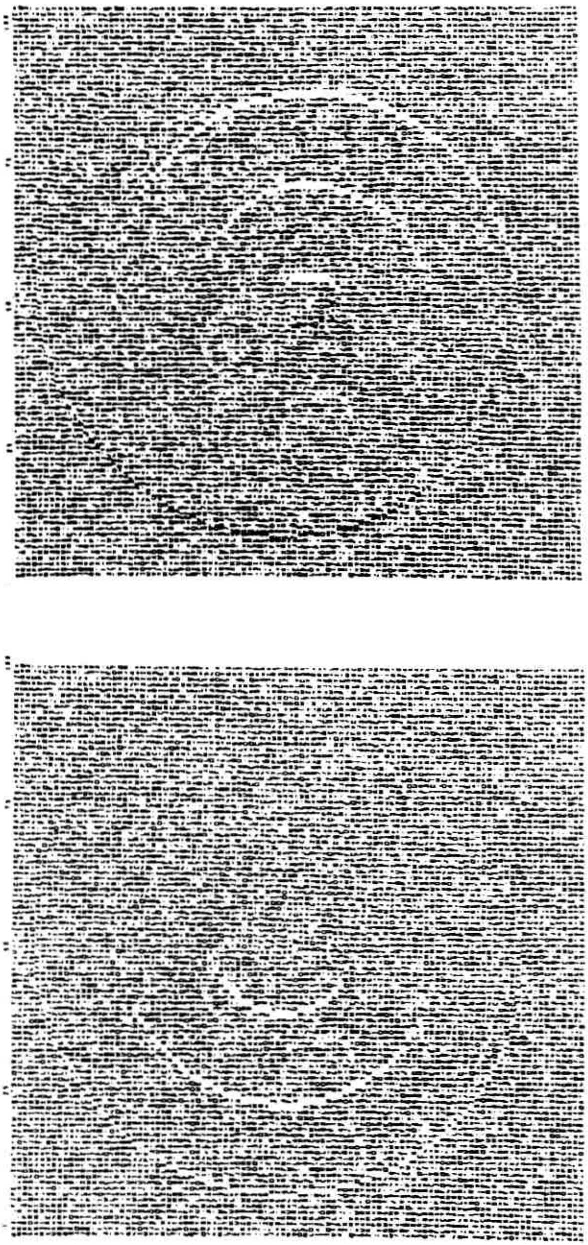


Fig. 7 — Automatic stereoscopic presentation of a function (Example 2).

of stereo projections of the surface. By use of digital computers and the special-purpose output devices now available, this procedure can be carried out in a completely automatic fashion, thus making possible a simple and effective demonstration of three-dimensional data.

REFERENCES

1. MacKay, D. M., A Simple Multi-Dimensional CRT Display Unit, Electronic Engineering, **32,** London, June, 1960, pp. 344–347.
2. Julesz, B., Binocular Depth Perception of Computer-Generated Patterns, B.S.T.J., **39,** Sept., 1960, pp. 1125–1161.
3. David, E. E., Jr., Mathews, M. V., McDonald, H. S., Experiments with Speech Using Digital Computer Simulation, I.R.E. Wescon Convention Record-Audio, Aug., 1958, p. 3.
4. Graham, R. E., Kelly, J. L., Jr. A Computer Simulation Chain for Research on Picture Coding, I.R.E. Wescon Convention Record — Computer Applications, Aug., 1958, pp. 41–46.
5. Julesz, B., A Method of Coding Television Signals Based on Edge Detection, B.S.T.J., **38,** July, 1959, pp. 1001–1020.

Maximization of the Fundamental Power in Nonlinear Capacitance Diodes

By J. A. MORRISON

(Manuscript received October 6, 1961)

In this paper we consider the problem of determining the maximum fundamental power in a nonlinear capacitance diode, when the charge waveform has a given periodicity and (i) varies between prescribed maximum and minimum values, (ii) has a prescribed maximum and a prescribed maximum slope. Under (i) the maximum obtainable fundamental power is first determined. The charge waveform is then further restricted to contain no higher than second harmonics, so that the diode is being used as a frequency doubler, and the maximum power transfer is determined. The maximum power transfer is also determined under (ii). Particular diodes considered are the abrupt-junction and the graded-junction ones, with operation in the forward conduction region being permitted.

I. ENGINEER'S SUMMARY

This section of the paper is a summary which stresses some of the contents of the introduction and summary that follow. It is hoped that this will make it easier for the engineer who is involved in parametric amplifier and varactor design to deduce the relevant applications of the results contained in this paper.

In the first instance it should be emphasized that an idealized problem, based on a mathematical model, is considered. The nonlinear capacitor is assumed to be isolated from any external circuits, and we do not discuss how the power is fed into or taken from the device. Clearly there will be some power lost in the external circuit, and the maximum obtainable fundamental power determined in this paper is only a theoretical maximum, but it would seem to be worthwhile to understand this theoretical maximum. When the maximum power transfer from the first to the second harmonic is considered, the charge waveform, and hence the current, giving this maximum is determined. Clearly there is some relative phase between the first and second harmonics in the current, and the reactance of the output circuit must be adjusted so as to obtain this relative phase.

It is also important to stress that some of the results obtained hold for a general, i.e., arbitrary single-valued, voltage-charge relationship, and are accordingly applicable to any particular such voltage-charge relationship in which the engineer may be interested. We have, for simplicity, considered just the abrupt-junction and the graded-junction diodes as special cases, and have idealized the voltage-charge relationship in the forward conduction region, but other particular diodes can be considered as special cases of the general results. We discuss below the results which are pertinent to the general voltage-charge relationship.

Firstly, we have derived the functional form of the charge waveform (of given periodicity and varying between prescribed values) which gives the maximum power in the fundamental. The charge waveform is composed (see (33) below) of intervals in which it takes on either the maximum or minimum prescribed value, or else follows a certain curve. The form of the curve depends on the voltage-charge relationship and involves parameters which are functionals of the charge waveform throughout the entire period, and hence are not known a priori. These parameters have to be determined for each particular voltage-charge relationship, by solving simultaneous transcendental equations. It is also necessary to allow for finite jumps in the charge waveform, and (36) below must hold at such a jump. Of course, a jump is not physically realizable, since it would correspond to an infinite current, and this makes it evident that the maximum is a theoretical one, quite apart from losses in the external circuit. It does, however, provide an upper bound on the maximum realizable fundamental power.

In view of the fact that the maximum fundamental power has to be determined separately for each specific diode, we derive upper and lower bounds for the maximum fundamental power, (11) to (13), which apply to a general voltage-charge relationship. For a wide class, the ratio of the upper to the lower bound is 1.54. It turns out that, for the particular diodes considered, the lower bound is quite close to the actual value. Further use is made of the charge waveform giving this lower bound, when the power transfer from the fundamental to the second harmonic is considered, subject to the charge waveform containing no higher than second harmonics. A good approximation to the maximum power transfer is obtained by taking the Fourier approximation, up to second harmonics, and suitably normalizing so that the approximating charge waveform has the prescribed maximum and minimum values.

In connection with maximizing the power transfer from the fundamental to the second harmonic, we consider the diode to be a harmonic generator, there being input power in the fundamental only. In order to make the mathematical problem more tractable, it is supposed that the entire output is in the second harmonic. Equations (18) and (19) simply state that the maximum power output in the second harmonic, when there is input power in the fundamental only, is not greater than the maximum obtainable fundamental power without such restrictions, and is not less than the maximum fundamental power when there is no output or input power in the third and higher harmonics. It is assumed here that the charge waveform is continuous. We have already discussed the maximum obtainable fundamental power.

The problem of determining the maximum fundamental power when there is no output or input power in the third and higher harmonics is still not very tractable, without additional restrictions on the charge waveform, and it is thus further supposed that the charge waveform contains no higher than second harmonics. The maximum subject to this additional restriction is obviously not greater than the maximum without it. The significant point about this restriction is that there is then no power output or input in the third and higher harmonics, whatever the voltage-charge relationship. We thus determine a canonical representation of the charge waveform which contains no higher than second harmonics and has prescribed maximum and minimum values. By suitable choice of the time origin, this representation contains just two parameters which lie in a bounded region.

Now, it is a straightforward matter to compute numerically the fundamental power for any given voltage-charge relationship and a given charge waveform. The numerical maximization of this power with respect to the two parameters in the above canonical representation is also a straightforward process. Thus it is clear that the above procedure has general applicability. We add that in the numerical maximization process, the two parameters which give the approximating charge waveform (obtained from the charge waveform giving the good lower bound to the maximum obtainable fundamental power) are used for starting values.

Consideration is also given to the current-limited diode, in which the charge waveform has a prescribed maximum value and a prescribed maximum slope (corresponding to maximum current magnitude). Again, we determine a two-parameter canonical representation for the charge waveform containing no higher than second harmonics, and the numerical maximization of the fundamental power, for any given voltage-charge relationship, proceeds along the same lines as in the previous case, except that we no longer have predetermined starting values for the two parameters. Lack of space has prevented inclusion of the determination of the functional form of the charge waveform which gives the maximum obtainable fundamental power (without restriction on the harmonic content of the charge waveform) in the current-limited case.

II. INTRODUCTION AND SUMMARY

2.1 *Introduction*

We will be concerned with various nonlinear capacitance diodes, these being characterized by a nonlinear voltage-charge relationship. Specific examples are the abrupt-junction diode and the graded-junction diode, which are composed of diffused p-n junctions. In the former case the voltage difference, v, across the diode is proportional to the square of the stored charge (per unit area), q, i.e., $v \propto q^2$, while in the latter case $v \propto q^{\frac{3}{2}}$, provided, in both cases, that $q \geqq 0$, which implies that operation of the diode does not take place in the forward conduction region. Now as electric field strength and barrier width increase, creation of electron-hole pairs through secondary impact ionization by both holes and electrons leads to avalanche multiplication, resulting finally in an effectively infinite increase of current with added applied voltage, and this is termed reverse breakdown. There is thus a maximum voltage $v_{\max}$, and a corresponding maximum value $q_{\max}$ of the charge density (which may be related to $v_{\max}$ through the actual voltage-charge relationship), above which it is not desirable to operate the diode.

We define the normalized voltage V and the normalized charge Q by

$$V = \frac{v}{v_{\max}}; \qquad Q = \frac{q}{q_{\max}}. \tag{1}$$

Hence the normalized voltage-charge relationships for the abrupt-junction and graded-junction diodes, operated in the region between forward conduction and reverse breakdown, are

$$V = \begin{cases} Q^2, & \text{(abrupt)} \\ Q^{\frac{3}{2}}, & \text{(graded)} \end{cases}, \quad 0 \leqq Q \leqq 1. \tag{2}$$

It is also possible to operate the diodes partially in the forward conduction region, corresponding to $Q < 0$. The voltage is not very dependent on the charge in this region and as an idealization we may assume that it is zero throughout. A physical restriction is placed on the maximum possible current magnitude, in that the electron velocity is limited by lattice scattering. Throughout most of our analysis we replace this condition by a limitation on the minimum charge, so that

$$q \geqq q_{\min} = -m(q_{\max}). \tag{3}$$

Thus, in the forward conduction region,

$$V = 0, \qquad -m \leqq Q \leqq 0. \tag{4}$$

We do, however, give some consideration to the current-limited diode in which, instead of (3),

$$| i | \leqq i_{\max} . \tag{5}$$

We will consider charge waveforms that are periodic in time, t, with angular frequency ω. We define the normalized time x and the normalized current I by

$$x = \omega t; \qquad I = \frac{i}{\omega q_{\max}}. \tag{6}$$

Thus $Q(x)$ is periodic in x with period 2π and, since $i = dq/dt$,

$$I = \frac{dQ}{dx} = Q'(x). \tag{7}$$

The average real and reactive powers (per unit area) in the nth harmonic, p_n and r_n, are given by

$$(p_n + jr_n) = \frac{1}{2}\left(\frac{\omega}{\pi}\right)^2 \left(\int_0^{2\pi/\omega} i\, e^{-j\omega nt}\, dt\right)\left(\int_0^{2\pi/\omega} v\, e^{j\omega nt}\, dt\right). \tag{8}$$

We define the normalized real and reactive powers in the nth harmonic, P_n and R_n, by

$$(P_n + jR_n) = \frac{2\pi^2(p_n + jr_n)}{\omega q_{\max} v_{\max}}. \tag{9}$$

We will be concerned with the maximization of the real fundamental power, under various conditions, and summarize the results below. We note that P_n is not affected by a time shift in the charge waveform, but it is reversed in sign by a time reversal of the waveform.

2.2 *The Maximum Obtainable Fundamental Power, When the Charge Waveform is Subject to Bounded Variation*

The functional form of the charge waveform which, subject to the restriction $-m \leqq Q(x) \leqq 1$, maximizes the fundamental power, P_1, is found for the general voltage-charge relationship, $V = V(Q)$. The specific form is determined for diodes of interest and the corresponding value of max P_1, the maximum obtainable fundamental power, calculated. Thus, for the abrupt-junction diode operated in the region between forward conduction and reverse breakdown, (2), max $P_1 = 0.687$ and the charge waveform $Q(x)$ giving rise to this value is depicted in Fig. 1. The corresponding value of the reactive fundamental power is $R_1 = 2.43$. For the graded-junction diode, operated in the region between forward conduction and reverse breakdown, it is found that max $P_1 = 0.408$, with $R_1 = 2.48$. The charge waveform giving rise to these values is depicted in Fig. 2. The abrupt-junction diode is also considered when the region of operation includes forward conduction. Thus, from (2) and (4), $V(Q) = [\max(0,Q)]^2$, $-m \leqq Q(x) \leqq 1$. Fig. 4 depicts max P_1 and the corresponding R_1 as functions of m. The charge waveform $Q(x)$ which gives these values when $m = 1$ is shown in Fig. 5. The somewhat idealized voltage-charge relationship given by $V(Q) = \max(0,Q)$, $-m \leqq Q(x) \leqq 1$, $m > 0$, may be treated analytically. It is found in this case that

$$\max P_1 = \frac{3\sqrt{3}}{2}\, m; \qquad R_1 = \frac{3}{2}\,(m + 2). \tag{10}$$

The charge waveform giving these values is composed of $Q(x) = 1$, 0 and $-m$ in consecutive intervals of x of length $2\pi/3$.

It is observed that the charge waveform which gives rise to max P_1, for the various diodes, contains at least one discontinuity (or jump) in a period. A jump, of course, is not physically realizable, since it would correspond to an infinite current, so max P_1 cannot actually be attained.

Finally, upper and lower bounds are obtained on the maximum obtainable fundamental power, max P_1, for the general voltage-charge relationship $V = V(Q)$, with $-m \leqq Q(x) \leqq 1$. Thus, it is shown that

$$\frac{3\sqrt{3}}{2}\, L \leqq \max P_1 \leqq 4(1 + m)U, \tag{11}$$

where

$$L = \max_{-m \leqq (\sigma,\rho,\tau) \leqq 1} [(\rho - \tau)V(\sigma) + (\tau - \sigma)V(\rho) + (\sigma - \rho)V(\tau)]; \tag{12}$$

and

$$U = \min_{\lambda} \{ \max_{-m \leqq \sigma \leqq 1} [\lambda\sigma - V(\sigma)] - \min_{-m \leqq \sigma \leqq 1} [\lambda\sigma - V(\sigma)]\}. \tag{13}$$

Moreover, it is shown that

$$L \leqq (1 + m)U \leqq 2L. \tag{14}$$

The bounds in (14) cannot be improved without restriction on $V(Q)$, but if $[(\rho + m)V(1) - (m + 1)V(\rho) + (1 - \rho)V(-m)]$ does not change sign in $-m \leqq \rho \leqq 1$, then $L = (1 + m)U$ and the ratio of the upper to the lower bound in (11) becomes 1.54. The class of voltage-charge relationships

$$V(Q) = [\max\ (0,Q)]^{\nu}, \quad -m \leqq Q \leqq 1; \qquad m \geqq 0, \quad \nu \geqq 1, \quad (15)$$

which includes the particular diodes considered, satisfies the above condition, and in this case

$$L = m + \left(1 - \frac{1}{\nu}\right)[(1 + m)\nu]^{-1/(\nu-1)}. \qquad (16)$$

For the particular cases considered, the lower bound in (11) is fairly close to $\max P_1$.

A lower bound is also obtained, for a general voltage-charge relationship $V = V(Q)$, with $-m \leqq Q(x) \leqq 1$, for P_1 such that $P_1 + P_2 = 0$. It is shown that

$$\max\ [P_1 \mid P_1 + P_2 = 0] \geqq (1.87)L. \qquad (17)$$

2.3 *The Maximization of the Power Transfer in a Frequency Doubler, With Bounded Charge Waveform*

Here we are interested in maximizing the power transfer from the fundamental to the second harmonic, when the diode is being used as a harmonic generator. Thus there must be input power at the fundamental frequency only, i.e., $P_1 > 0$ and $P_n \leqq 0$, $n \geqq 2$. In order to make the problem more tractable we suppose that the entire power output is put is in the second harmonic, so that $P_n = 0$, $n \geqq 3$. It follows that $P_1 + P_2 = 0$, provided that the charge waveform is continuous, since then $\sum_{n=1}^{\infty} P_n = 0$. We observe that

$$\begin{aligned} &\max\ [-P_2 \mid P_n \leqq 0, \quad n \geqq 3] \\ &\qquad \leqq \max\ [P_1 \mid P_n \leqq 0, \quad n \geqq 3] \leqq \max P_1, \end{aligned} \qquad (18)$$

and

$$\begin{aligned} \max\ [-P_2 \mid P_n \leqq 0, \quad n \geqq 3] &\geqq \max\ [-P_2 \mid P_n = 0, \quad n \geqq 3] \\ &= \max\ [P_1 \mid P_n = 0, \quad n \geqq 3]. \end{aligned} \qquad (19)$$

Even the problem of determining $\max\ [P_1 \mid P_n = 0,\ n \geqq 3]$, that is, $\max P_1$ subject to $P_n = 0$, $n \geqq 3$, is not very tractable, without additional restrictions on the charge waveform. Thus, it is supposed that the charge, and hence the current, contains no higher than second harmonics. The conditions $P_1 + P_2 = 0$ and $P_n = 0$, $n \geqq 3$, are then identically satisfied, independently of the voltage-charge relationship $V = V(Q)$.

Now, a change in the time origin does not affect the power transfer. Hence, the canonical representation of a charge waveform which contains no higher than second harmonics and is such that $Q(\pi) = Q_{\min} = -m$ and $Q(2 \tan^{-1} s) = Q_{\max} = 1$, is constructed. In addition to the parameter s there is the parameter y which is subject to the restriction $0 \leqq y \leqq (1 - s^2)$, which of course also implies that $s^2 \leqq 1$. It is found that $P_1 = 0$ on $y = 0$ and on $y = (1 - s^2)$, independently of the voltage-charge relationship. Moreover, $P_1(s,y) = -P_1(-s,y)$ and in particular $P_1 = 0$ on $s = 0$ also, so that it is sufficient to consider only the region $-1 \leqq s \leqq 0$, $0 \leqq y \leqq (1 - s^2)$ and to maximize $| P_1 |$. The abrupt-junction diode, operated in the region between forward conduction and reverse breakdown, may be treated analytically, and it is found that the maximum power transfer is 0.281, as compared with the maximum obtainable fundamental power of 0.687. The corresponding reactive fundamental powers are 1.46 and 2.43, and the charge waveform giving the maximum power transfer is depicted in Fig. 6, which should be compared with Fig. 1.

In order to determine the maximum power transfer for a general voltage-charge relationship, recourse must be made to numerical computation. However, a prior step is the determination of a charge waveform which provides a reasonable approximation to the maximum power transfer, and hence provides starting values for s and y in the numerical maximization process. A good lower bound was obtained for the maximum obtainable fundamental power. Furthermore, for a wide class of voltage-charge relationships $V = V(Q)$, the charge waveform $Q(x)$ giving this lower bound satisfies $Q_{\max} = 1$ and $Q_{\min} = -m$. The class of voltage-charge relationships (15) falls within this class. Thus it would seem feasible that a reasonable approximation to the maximum power transfer will be obtained by taking the Fourier approximation, up to the second harmonics, of the charge waveform giving the good lower bound for the maximum obtainable fundamental power, and suitably shifting and expanding (or contracting) the Fourier approximation so that the resulting charge waveform $\tilde{Q}(x)$ satisfies $\tilde{Q}_{\max} = 1$ and $\tilde{Q}_{\min} = -m$. This is the procedure adopted and, for the abrupt-junction diode, operated in the region between forward conduction and reverse breakdown, it actually yields the charge waveform that gives the maximum power transfer.

The results of the numerical maximization process are tabulated in Section 5.4. Tables II and III, for the cases $\nu = 2$ and $\nu = \frac{3}{2}$, in (15), show the values of *max* P_1, the maximum power transfer, and the corresponding values of R_1 and R_2, the reactive powers in the fundamental and second harmonic, $I_{\max}$, the maximum normalized current magnitude, and $(b^2 + c^2)$ and $(d^2 + e^2)$, the squares of the amplitudes of the first and second harmonics in the charge waveform, for several values of m. Tables IV and V show the values of $-s$ and y which give *max* P_1 and also $y^{(1)}$ and $P_1^{(1)}$, the value of P_1 corresponding to the starting values $y^{(1)}$ and $-s^{(1)} = 1/\sqrt{3}$. It is interesting to observe how close $P_1^{(1)}$ is to *max* P_1, particularly for the smaller values of m. Table VI compares *max* P_1 with the maximum obtainable fundamental power, max P_1, in the case $\nu = 2$, for several values of m. It is also worth noting that in the case $\nu = \frac{3}{2}$, $m = 0$ we have *max* $P_1 = 0.162$, whereas max $P_1 = 0.408$.

2.4 *The Maximization of the Power Transfer in a Frequency Doubler, for the Current-Limited Diode*

We finally turn our attention to the current-limited diode in which (5), instead of (3), holds. Thus, from (5) to (7),

$$| Q'(x) | \leqq \frac{i_{\max}}{(\omega q_{\max})} = \frac{\kappa}{\omega}. \tag{20}$$

For the P^+N abrupt-junction diode of germanium[1,2]

$$\begin{aligned} v_{\max} &\simeq 1.03 \times 10^{13}(N)^{-0.725} \text{ volts,} \\ i_{\max} &\simeq 1.6 \times 10^{-12}N \text{ amps/cm}^2, \end{aligned} \tag{21}$$

where N is the donor concentration in cm^{-3}. But, from the voltage-charge relationship,

$$q_{\max}{}^2 = 2e\,\epsilon\,Nv_{\max}, \tag{22}$$

where e denotes electron charge. Hence,

$$q_{\max} \simeq 2.16 \times 10^{-9}(N)^{0.1375} \text{ coulombs/cm}^2, \tag{23}$$

and

$$\kappa = \frac{i_{\max}}{q_{\max}} \simeq 0.74 \times 10^{-3}(N)^{0.8625} \text{ sec}^{-1}. \tag{24}$$

For $N = 2 \times 10^{16}$, a reasonable value, $\kappa \approx 10^{11}$ sec^{-1}, which is in the range of angular frequencies of interest.

We consider the problem of maximizing the power transfer from the fundamental to the second harmonic, when the diode is being used as a frequency doubler, and, as previously, the additional assumption is made that the charge waveform $Q(x)$, and hence the current, contains no higher than second harmonics. The first step is the construction of the canonical representation of $Q(x)$ such that $Q_{\max} = 1$, $Q'(\pi) = Q'_{\min} = -k$ and $Q'(2 \tan^{-1} s) = Q'_{\max} \leqq k$. In addition to the parameter s there is the parameter y which is subject to the restriction $0 \leqq y \leqq \frac{1}{2}(1 - s^2)$, which of course also implies $s^2 \leqq 1$. It is found that $P_1 = 0$ on $y = \frac{1}{2}(1 - s^2)$, independently of the voltage-charge relationship. Since, if $\tilde{Q}(x) = Q(\pi - x)$, then $\tilde{Q}_{\max} = 1$, $\tilde{Q}'_{\max} = k$ and $\tilde{Q}'_{\min} \geqq -k$, it is sufficient to consider the above canonical representation and to maximize $| P_1 |$, in order to maximize P_1 subject to $Q_{\max} = 1$, $| Q' |_{\max} = k$. We denote this maximum by $\Pi(k)$. For the abrupt-junction diode operated in the region between forward conduction and reverse breakdown, the determination of $\Pi(k)$ is carried out analytically for k sufficiently small that $Q_{\min} \geqq 0$. It is found that $\Pi(k) = 0.731k^3$, for $0 \leqq k \leqq 0.681$. Combining this result with that obtained when the charge waveform is subject just to bounded variation, $0 \leqq Q(x) \leqq 1$, it is shown that, from the viewpoint of maximizing the actual fundamental real power p_1, the optimum operating frequency lies in the range

$$1.299 \leqq \frac{(\omega q_{\max})}{i_{\max}} \leqq 1.468, \tag{25}$$

and that

$$1 \leqq \frac{54(\max p_1)}{(i_{\max} v_{\max})} \leqq \frac{2(4)^{\frac{1}{3}}}{3} < 1.06. \tag{26}$$

For the abrupt-junction diode which is allowed to operate partly in the forward conduction region, the maximization of the power transfer is determined by numerical computation. For the values of s and y which give $\max | P_1 |$, i.e., $\Pi(k)$, the reactive powers R_1 and R_2, and $Q_{\min}$, i.e., $-M(k)$, were calculated, the results being given in Table VII (Section 6.4). It is shown that $\max P_1$ subject to $Q_{\max} \leqq 1$ and $| Q' |_{\max} \leqq k$ is attained with $Q_{\max} = 1$ and $| Q' |_{\max} = k$. For $k < 0.681$ it can also be attained with $1.468k \leqq Q_{\max} < 1$ and $| Q' |_{\max} = k$. Optimizing with respect to the frequency it appears that $20(\max p_1) \sim i_{\max} v_{\max}$. Thus a considerable improvement is obtained by permitting operation in the forward conduction region. The optimum frequency in this case is roughly one-fifth that in the case when operation is not allowed in the forward conduction region, although close to max p_1 may be obtained at one-third the frequency.

In conclusion, we add that lack of space has necessitated the omission of several aspects of this problem, and in particular of the determination of the maximum obtainable fundamental power when the periodic charge waveform is restricted only to have bounded slope.

III. THE CHARGE WAVEFORM WHICH, SUBJECT TO BOUNDED VARIATION, MAXIMIZES THE POWER IN THE FUNDAMENTAL HARMONIC

3.1 *The Functional Form of the Charge Waveform*

From (1), (6), (7), (8) and (9),

$$P_n + jR_n = \left(\int_0^{2\pi} Q'(x)\, e^{-jnx}\, dx\right)\left(\int_0^{2\pi} V[Q(x)]\, e^{jnx}\, dx\right). \quad (27)$$

It is noted that P_n is not affected by a time shift in the charge waveform $Q(x)$, but it is reversed in sign by a time reversal of the waveform. On the other hand, R_n is not affected by either a time shift or a time reversal in the charge waveform. Integrating by parts the first integral in (27), and remembering that $Q(x)$ is periodic with period 2π, and then separating real and imaginary parts,

$$P_n = n(\alpha_n\delta_n - \beta_n\gamma_n); \qquad R_n = n(\alpha_n\gamma_n + \beta_n\delta_n), \quad (28)$$

where

$$\begin{aligned} \alpha_n &= \int_0^{2\pi} Q(x)\sin nx\, dx; & \beta_n &= \int_0^{2\pi} Q(x)\cos nx\, dx; \\ \gamma_n &= \int_0^{2\pi} V[Q(x)]\sin nx\, dx; & \delta_n &= \int_0^{2\pi} V[Q(x)]\cos nx\, dx. \end{aligned} \quad (29)$$

From (28) and (29) we may express P_n as a double integral,

$$\frac{1}{n}P_n = \int_0^{2\pi}\int_0^{2\pi} Q(x)V[Q(y)]\sin n(x-y)\, dx\, dy. \quad (30)$$

To find the functional form of $Q(x)$ which, subject to the restriction

$$-m \leqq Q(x) \leqq 1, \quad (31)$$

maximizes P_1, we set

$$Q(x) = [(1+m)\operatorname{sech} R(x) - m], \quad (32)$$

so that the inequalities in (31) are satisfied. A variational procedure applied to (30) then shows that for stationary values of P_1, we have, for each x,

$$\begin{gathered} Q(x) = -m, \qquad or \qquad Q(x) = 1, \qquad or \\ V'[Q(x)] = \frac{(\gamma_1 \cos x - \delta_1 \sin x)}{(\alpha_1 \cos x - \beta_1 \sin x)}, \end{gathered} \tag{33}$$

where α_1, β_1, γ_1, and δ_1 are as defined in (29). This, then, is the functional form of $Q(x)$ which maximizes P_1. Evaluation of the integrals in (29) will lead to four equations for the four unknowns α_1, β_1, γ_1, and δ_1. Note, however, that

$$\begin{aligned} \frac{d}{dx}\left[\frac{(\gamma_1 \cos x - \delta_1 \sin x)}{(\alpha_1 \cos x - \beta_1 \sin x)}\right] &= -\frac{(\alpha_1\delta_1 - \beta_1\gamma_1)}{(\alpha_1 \cos x - \beta_1 \sin x)^2} \\ &= \frac{-P_1}{(\alpha_1 \cos x - \beta_1 \sin x)^2}, \end{aligned} \tag{34}$$

from (28), is of one sign. Since we are not interested in $P_1 = 0$, which case arises in particular if $Q(x) =$ const, it follows that allowance must be made for discontinuities in $Q(x)$, since we require that $Q(x)$ be periodic. Supposing that $Q(x)$ is discontinuous at $x = \varphi$, we obtain a condition by integrating the equation

$$V'[Q(x)]Q'(x) = \frac{(\gamma_1 \cos x - \delta_1 \sin x)}{(\alpha_1 \cos x - \beta_1 \sin x)} Q'(x), \tag{35}$$

from $x = \varphi - 0$ to $x = \varphi + 0$. This gives

$$[V[Q(x)]]_{\varphi-0}^{\varphi+0} = \frac{(\gamma_1 \cos \varphi - \delta_1 \sin \varphi)}{(\alpha_1 \cos \varphi - \beta_1 \sin \varphi)} [Q(x)]_{\varphi-0}^{\varphi+0}. \tag{36}$$

3.2 *The Charge Waveform for the Abrupt-Junction Diode*

In normalized form the voltage-charge relationship for the abrupt-junction diode operated in the region between forward conduction and reverse breakdown is

$$V(Q) = Q^2, \qquad 0 \leqq Q(x) \leqq 1, \tag{37}$$

so that $m = 0$ in (31). We make use of the fact that P_1 is invariant under the transformation $Q(x) \rightarrow Q(x - \theta)$, and choose θ so that $\beta_1 = 0$, since this leads to a simplification of the analysis. Let us define a and b by the equations

$$\gamma_1 = 2a\alpha_1, \quad \delta_1 = 2b\alpha_1; \qquad \beta_1 = 0. \tag{38}$$

Then, from (28),

$$P_1 = 2b\alpha_1^2. \tag{39}$$

It is clear that max $P_1 > 0$, and hence that $b > 0$. The functional form of $Q(x)$ for max P_1 is, from (33), (37) and (38),

$$Q(x) = 0, \qquad \textit{or} \qquad Q(x) = 1, \qquad \textit{or} \qquad Q(x) = (a - b \tan x). \tag{40}$$

Rejecting combinations which lead to $P_1 = 0$, we are led to the conclusion that, within a cycle, $Q(x) = 1$ for an interval, it then follows the curve $Q(x) = (a - b \tan x)$ and then $Q(x) = 0$ for an interval, after which it jumps from 0 to 1 and the cycle is repeated.

Let φ be a value of x at which a jump in $Q(x)$ from 0 to 1 occurs. Then (36), (37), and (38) give

$$\tan \varphi = \frac{(2a - 1)}{2b}. \tag{41}$$

Thus we obtain max P_1 by taking

$$Q(x) = \begin{cases} 1, & \text{for } \varphi < x \leqq \pi + \tan^{-1}[(a-1)/b]; \\ (a - b \tan x), & \text{for } \pi + \tan^{-1}[(a-1)/b] \leqq x \\ & \qquad \leqq \pi + \tan^{-1}(a/b); \\ 0, & \text{for } \pi + \tan^{-1}(a/b) \leqq x < 2\pi + \varphi, \end{cases} \tag{42}$$

where

$$-\frac{\pi}{2} < \tan^{-1}[(a-1)/b] < \varphi < \tan^{-1}(a/b) < \frac{\pi}{2}, \tag{43}$$

and

$$Q(x + 2\pi) = Q(x), \qquad \text{all } x. \tag{44}$$

Now α_1, β_1, γ_1, and δ_1 may be calculated from (29), (37), (42) and (44). Substitution into (38) then leads to

$$\begin{aligned} (2a - 1)\cos\varphi &= 2b\{[(a-1)^2 + b^2]^{\frac{1}{2}} - (a^2 + b^2)^{\frac{1}{2}}\}; \\ 2b\cos\varphi + \sin\varphi + 3b\tau &= \{(a+1)[(a-1)^2 + b^2]^{\frac{1}{2}} \\ &\qquad - a(a^2 + b^2)^{\frac{1}{2}}\}; \\ \sin\varphi &= \{[(a-1)^2 + b^2]^{\frac{1}{2}} - (a^2 + b^2)^{\frac{1}{2}}\}, \end{aligned} \tag{45}$$

where

$$\tau = b[\tanh^{-1}\{a(a^2 + b^2)^{-\frac{1}{2}}\} - \tanh^{-1}\{(a-1)[(a-1)^2 + b^2]^{-\frac{1}{2}}\}]. \tag{46}$$

It would appear that we now have one too many conditions on a, b and φ because of the relationship in (41), which was obtained from the jump condition at $x = \varphi$, but it is observed that the first and last equations in (45) are consistent with (41). Since $|\varphi| < \pi/2$ and $b > 0$, (41) gives

$$\begin{aligned} \cos\varphi &= 2b[(2a-1)^2 + 4b^2]^{-\frac{1}{2}}; \\ \sin\varphi &= (2a-1)[(2a-1)^2 + 4b^2]^{-\frac{1}{2}}. \end{aligned} \tag{47}$$

Substituting into the first equation in (45), we obtain

$$(2a-1)[(2a-1)^2 + 4b^2]^{-\frac{1}{2}} = \{[(a-1)^2 + b^2]^{\frac{1}{2}} - (a^2 + b^2)^{\frac{1}{2}}\}. \tag{48}$$

A solution to (48) is $a = \frac{1}{2}$ and, moreover, this is the only solution since if a $> \frac{1}{2}$ the L.H.S. > 0 and the R.H.S. < 0, and vice versa. Thus,

$$a = \tfrac{1}{2}, \qquad \varphi = 0. \tag{49}$$

The second equation in (45), using the definition of τ given in (46), now leads to an equation for b, namely

$$3b^2 \tanh^{-1}[(1 + 4b^2)^{-\frac{1}{2}}] = [\tfrac{1}{4}(1 + 4b^2)^{\frac{1}{2}} - b], \tag{50}$$

and (39) and the expression for α_1 give

$$P_1 = 2b\{1 + 2b\tanh^{-1}[(1 + 4b^2)^{-\frac{1}{2}}]\}^2 = \frac{1}{18b}[2b + (1 + 4b^2)^{\frac{1}{2}}]^2, \tag{51}$$

using (50). Equation (50) was solved numerically and it was found that

$$b = 0.14136; \qquad \max P_1 = 0.6868. \tag{52}$$

The shape of $Q(x)$ which gives this maximum value of P_1 is shown in Fig. 1. From (28) and (38) the corresponding reactive fundamental

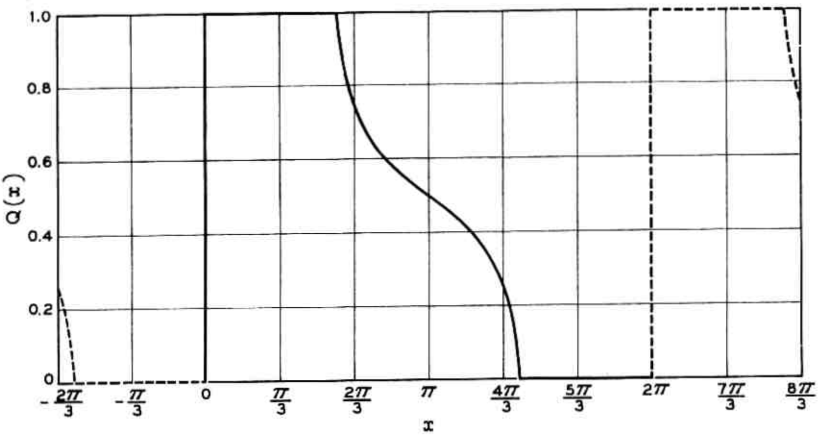


Fig. 1 — Charge waveform for maximum obtainable fundamental power in abrupt-junction diode operated in the region between forward conduction and reverse breakdown.

power is given by

$$R_1 = 2a\alpha_1^{\;2} = \frac{P_1}{(2b)} = 2.429. \tag{53}$$

Note that the reactive power is about three and a half times as large as the real power.

3.3 *The Charge Waveform for the Graded-Junction Diode*

We now turn our attention to the second diode of interest, namely the graded-junction diode, and suppose that it is operated in the region between forward conduction and reverse breakdown. In normalized form the voltage-charge relationship is

$$V(Q) = Q^{\frac{3}{2}}, \qquad 0 \leqq Q(x) \leqq 1. \tag{54}$$

The determination of the maximum obtainable fundamental power, max P_1, is carried out along the same lines as for the abrupt-junction diode, although the details are more involved. The analytical form of the charge waveform $Q(x)$ which gives max P_1 is

$$Q(x) = \begin{cases} 1, & \text{for} \quad \psi < x \leqq \pi + \tan^{-1}[(a-1)/b]; \\ (\mathrm{a} - b\tan x)^2, & \text{for} \quad \pi + \tan^{-1}[(a-1)/b] \leqq x \\ & \qquad \leqq \pi + \tan^{-1}(a/b); \\ 0, & \text{for} \quad \pi + \tan^{-1}(a/b) \leqq x < 2\pi + \psi, \end{cases} \tag{55}$$

where

$$-\frac{\pi}{2} < \tan^{-1}[(a-1)/b] < \psi < \tan^{-1}(a/b) < \frac{\pi}{2}, \tag{56}$$

and (44) holds. Here

$$\gamma_1 = \frac{3}{2}a\alpha_1; \qquad \delta_1 = \frac{3}{2}b\alpha_1; \qquad \beta_1 = 0, \tag{57}$$

which leads to three equations for a, b and ψ. These equations are consistent with the jump condition (36) which gives

$$\tan\psi = (3a - 2)/(3b). \tag{58}$$

Elimination of ψ leads to two equations for a and b which were solved numerically, giving

$$b = 0.11098; \qquad a = 0.67375. \tag{59}$$

These values lead to

$$\max P_1 = 0.4084; \qquad R_1 = 2.479. \tag{60}$$

The corresponding charge waveform $Q(x)$ is depicted in Fig. 2.

3.4 *The Abrupt-Junction Diode When the Region of Operation Includes Forward Conduction*

In this case the normalized voltage-charge relationship is, from (2) and (4),

$$V(Q) = [\max (0,Q)]^2, \qquad -m \leqq Q(x) \leqq 1, \qquad m > 0. \tag{61}$$

As previously, we translate $Q(x)$ so that $\beta_1 = 0$ and again define a and b by (38), so that (39) for P_1 also holds. From (33), (38), and (61), the functional form of $Q(x)$ for max P_1 is

$$\begin{aligned} Q(x) &= -m, \qquad \textit{or} \qquad Q(x) = 1, \\ \textit{or} \qquad \max [0,Q(x)] &= (a - b \tan x). \end{aligned} \tag{62}$$

Thus we are led to the conclusion that within a cycle $Q(x) = 1$ for an interval, it then follows the curve $Q(x) = (a - b \tan x)$ until the point at which $Q(x) = 0$ where it jumps to the value $-m$, and after $Q(x) = -m$ for an interval it jumps to the value 1 and the cycle is repeated. Thus in this idealized case there are two discontinuities in $Q(x)$ in one cycle. Note that according to (36), together with (38), the jump of

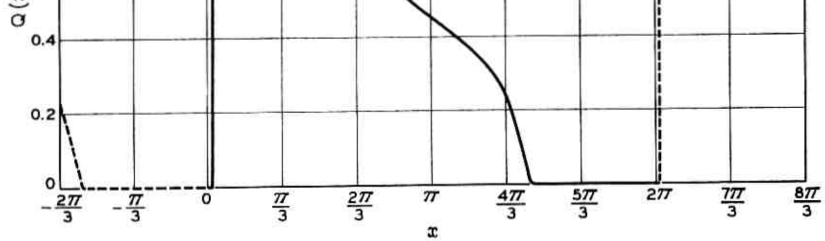


Fig. 2 — Charge waveform for maximum obtainable fundamental power in graded-junction diode operated in the region between forward conduction and reverse breakdown.

$Q(x)$ from 0 to $-m$ occurs at $x = \psi$ where $(a - b \tan \psi) = 0$, since $V(0) = 0$ and $V(-m) = 0$. Hence we obtain max P_1 by taking

$$Q(x) = \begin{cases} 1, & \text{for} \quad \varphi < x \leqq \pi + \tan^{-1}[(a-1)/b]; \\ (a - b \tan x), & \text{for} \quad \pi + \tan^{-1}[(a-1)/b] \leqq x \\ & \qquad < \pi + \tan^{-1}(a/b); \\ -m, & \text{for} \quad \pi + \tan^{-1}(a/b) < x < 2\pi + \varphi, \end{cases} \tag{63}$$

where (43) and (44) hold. In this case the jump condition at $x = \varphi$, gives

$$\tan \varphi = \frac{[2a(1+m) - 1]}{2b(1+m)}. \tag{64}$$

The calculation of α_1, β_1, γ_1, and δ_1, and substitution into (38), leads to three equations for a, b, and φ, which are consistent with (64). The elimination of φ leads to two equations for a and b, which quantities of course are functions of m. It was found to be possible to eliminate m analytically from these two equations, so that instead of solving the two simultaneous equations for a and b for given values of m, the single relation between a and b which did not involve m was solved for b for given values of a. Thus a parametric solution was obtained in the form $b = b(a)$, $m = m(a)$. From this a and b were plotted graphically against m and the results are shown in Fig. 3. It was shown analytically that

$$1 < 4a(1+m) \leqq 2, \tag{65}$$

the upper bound being attained for $m = 0$ and the lower bound being approached for $m \to \infty$. Also, as $m \to \infty$ it is found that

$$b \sim \sqrt{3}a; \qquad \max P_1 \sim \frac{3\sqrt{3}}{2} m; \qquad R_1 \sim \frac{3}{2} m, \tag{66}$$

where R_1 is the reactive power in the fundamental. Fig. (4) shows max P_1 and the corresponding R_1 as functions of m. It is interesting to note that the ratio (max P_1)/R_1 increases with increasing m from its initial value of 0.28, its asymptotic value being $\sqrt{3}$, from (66). The charge waveform $Q(x)$ giving rise to max P_1 is shown, for $m = 1$, in Fig. 5.

3.5 *The Charge Waveform for an Idealized Voltage-Charge Relationship*

We now consider a special voltage-charge relationship which may be handled analytically. Thus we suppose that the capacitance has a finite

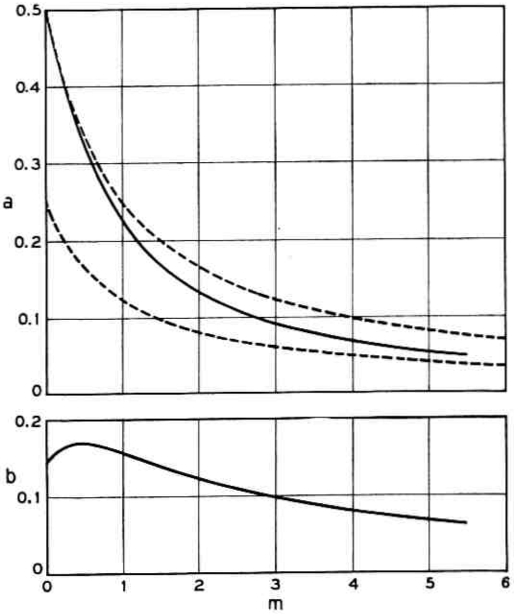


Fig. 3 — Parameters in charge waveform for maximum obtainable fundamental power in abrupt-junction diode operated partly in forward conduction region, vs. minimum charge.

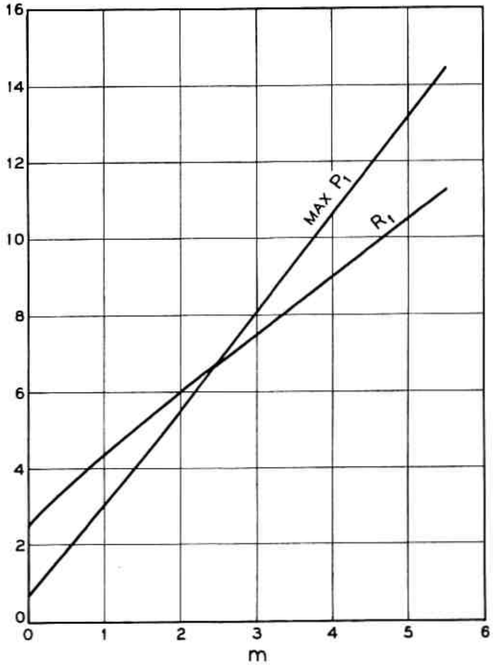


Fig. 4 — Maximum obtainable fundamental power, and corresponding fundamental reactive power, in abrupt-junction diode operated partly in forward conduction region, vs. minimum charge.

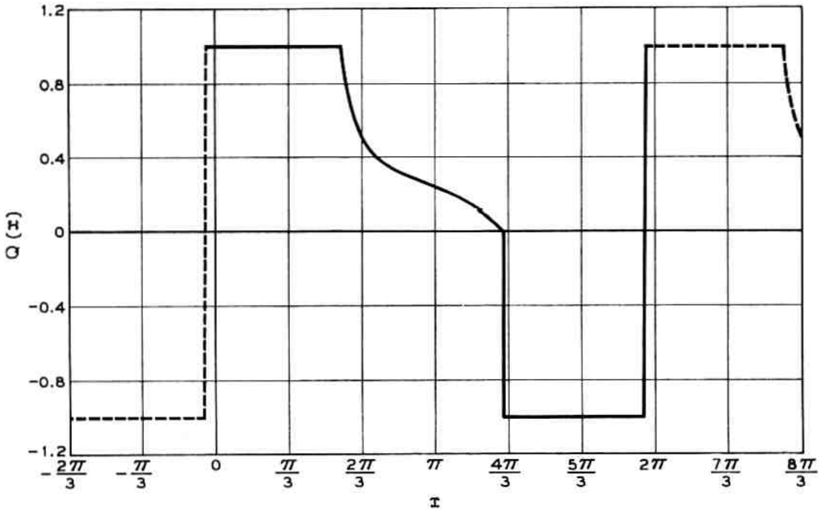


Fig. 5 — Charge waveform for maximum obtainable fundamental power in abrupt-junction diode operated partly in forward conduction region.

constant value for reverse bias and is infinite for forward bias, and hence in normalized form

$$V(Q) = \max\,(0,Q), \qquad -m \leqq Q(x) \leqq 1; \qquad m > 0. \tag{67}$$

Since $V'(Q)$ is constant except possibly at $Q = 0$, where it is indeterminate, we deduce from (33) that $Q(x)$ has one of the values 1, 0, and $-m$ at each point. Omitting further details, it is found that max P_1 is given by

$$Q(x) = \begin{cases} 1, & 0 < x < 2\pi/3; \\ 0, & 2\pi/3 < x < 4\pi/3; \\ -m, & 4\pi/3 < x < 2\pi. \end{cases} \tag{68}$$

Also,

$$\max P_1 = \frac{3\sqrt{3}}{2}\,m; \qquad R_1 = \frac{3}{2}\,(m + 2). \tag{69}$$

Note that, as might be expected, these values are asymptotically, as $m \to \infty$, the same as for the voltage-charge relationship in (61), as is seen from (66).

IV. BOUNDS ON THE MAXIMUM OBTAINABLE FUNDAMENTAL POWER

4.1 *Lower Bounds*

We now derive some lower bounds for the maximum obtainable power in the fundamental, for a general voltage-charge relationship, by the simple expedient of choosing specific charge waveforms. Any P_1 which we obtain is, of course, a lower bound for max P_1. Thus we consider the charge waveforms

$$Q(x) = \begin{cases} \sigma, & \text{on } \Gamma_1 \, ; \\ \rho, & \text{on } \Gamma_2 \, ; \\ \tau, & \text{on } \Gamma_3 \, , \end{cases} \tag{70}$$

where each $\Gamma_j (j = 1,2,3)$ is a finite collection of nonintersecting intervals, open at the left and closed at the right, and furthermore

$$\Gamma_j \cap \Gamma_k = 0, \quad j \neq k; \qquad \bigcup_{j=1}^{3} \Gamma_j = (0,2\pi]. \tag{71}$$

From (28), (29), (70) and (71),

$$\begin{aligned} P_n = nL(\sigma,\rho,\tau) \bigg[\left(\int_{\Gamma_1} \cos nx \, dx \right) \left(\int_{\Gamma_2} \sin nx \, dx \right) \\ - \left(\int_{\Gamma_1} \sin nx \, dx \right) \left(\int_{\Gamma_2} \cos nx \, dx \right) \bigg], \end{aligned} \tag{72}$$

where

$$L(\sigma,\rho,\tau) = [(\rho - \tau)V(\sigma) + (\tau - \sigma)V(\rho) + (\sigma - \rho)V(\tau)]. \tag{73}$$

The significant point here is that we can choose the intervals Γ_1 and Γ_2 to make P_1 as large as possible, for the waveform class of (70), independently of the functional form of the voltage-charge relationship, $V = V(Q)$. This is still true if we wish to make P_1 as large as possible subject to the condition $P_1 + P_2 = 0$, say, since the factor containing V, namely $L(\sigma,\rho,\tau)$, occurs in each P_n. Note, from (73), that $L(\sigma,\rho,\tau)$ vanishes unless σ, ρ, and τ are unequal. Also, if (σ,ρ,τ) undergo a cyclic permutation then $L(\sigma,\rho,\tau)$ is unaltered, but if (σ,ρ,τ) undergo an anticyclic permutation then $L(\sigma,\rho,\tau)$ is reversed in sign. We suppose that the charge waveform has bounded variation as in (31) and define

$$L = \max_{-m \leqq (\sigma,\rho \ \tau) \leqq 1} [L(\sigma,\rho,\tau)] \geqq 0. \tag{74}$$

Also from (72) it is seen that P_1 changes sign if Γ_1 and Γ_2 are interchanged, which is equivalent to an anticyclic permutation of (σ,ρ,τ).

Thus we are interested in making the modulus (or magnitude) of the bracketed expression following the factor $L(\sigma,\rho,\tau)$ in (72) as large as possible, in order to obtain as large as possible a lower bound for max P_1. We will restrict ourselves to special Γ_j and find the maximum modulus of the bracketed expression in (72) for these subclasses. In particular, we consider

$$\Gamma_1 = (0,\lambda]; \qquad \Gamma_2 = (\mu,\nu], \quad 0 < \lambda \leqq \mu < \nu < 2\pi. \tag{75}$$

Then, from (72),

$$P_n = \frac{1}{n} L(\sigma,\rho,\tau) F(n\lambda,n\mu,n\nu), \tag{76}$$

where

$$\begin{aligned} \mathrm{F}(\lambda,\mu,\nu) &= [\sin(\nu-\lambda) - \sin(\mu-\lambda) + \sin\mu - \sin\nu] \\ &= 4\sin[(\nu-\mu)/2]\sin(\lambda/2)\sin[(\nu+\mu-\lambda)/2]. \end{aligned} \tag{77}$$

We first set $\lambda = \mu$ and determine μ and ν to maximize $F(\mu,\mu,\nu)$ which from (75) and (77) is seen to be positive. The stationary values of $[\sin(\nu-\mu) + \sin\mu - \sin\nu]$ are given by

$$\cos\mu = \cos(\mu-\nu) = \cos\nu. \tag{78}$$

Hence $F(\mu,\mu,\nu)$ is a maximum for $\mu = 2\pi/3$, $\nu = 4\pi/3$ and from (74), (76), and (77) the corresponding maximum of P_1 is

$$P_1 = \frac{3\sqrt{3}}{2} L. \tag{79}$$

Now for the voltage-charge relationship (37) it is readily verified that $L(\sigma,\rho,\tau)$, as defined in (73), has a maximum value of $\frac{1}{4}$ which is attained for $\sigma = 1$, $\rho = \frac{1}{2}$, $\tau = 0$, and hence in this case we obtain the value $P_1 = 0.650$, which is quite close to the value of max P_1 given in (52).

We now consider the maximization of $F(\lambda,\mu,\nu)$ subject to the condition

$$F(\lambda,\mu,\nu) + \tfrac{1}{2}F(2\lambda,2\mu,2\nu) = 0, \tag{80}$$

corresponding to $P_1 + P_2 = 0$. Using the second part of (77), (80) becomes

$$1 + 4\cos[(\nu-\mu)/2]\cos(\lambda/2)\cos[(\nu+\mu-\lambda)/2] = 0, \tag{81}$$

supposing that $F(\lambda,\mu,\nu) \neq 0$. It is interesting to note that (81) cannot be satisfied with $\lambda = \mu$. It is found that $F(\lambda,\mu,\nu)$ is maximized, subject to (75) and (81), by

$$\lambda = 2(\pi - \theta); \qquad \mu = \theta; \qquad \nu = (2\pi - \theta), \tag{82}$$

where

$$\cos\theta = -(\tfrac{1}{2})^{\frac{2}{3}}, \qquad (2\pi/3 < \theta < \pi), \tag{83}$$

and the corresponding value of P_1, with $P_1 + P_2 = 0$, is

$$P_1/(4L) = [1 - (\tfrac{1}{2})^{\frac{4}{3}}]^{\frac{3}{2}} = 0.468. \tag{84}$$

4.2 *An Upper Bound, and its Relationship to a Lower Bound*

In Appendix A we give the derivation of an upper bound, for a general voltage-charge relationship, on the maximum obtainable fundamental power, using the fact that the charge waveform is of bounded variation, (31). It is shown that

$$\max P_1 \leqq 4(1 + m)U, \tag{85}$$

where

$$U = \min_{\lambda}\{\max_{-m \leqq \sigma \leqq 1} [\lambda\sigma - V(\sigma)] - \min_{-m \leqq \sigma \leqq 1} [\lambda\tau - V(\sigma)]\}. \tag{86}$$

In the previous section we showed, by example, that

$$\max P_1 \geqq \frac{3\sqrt{3}}{2} L, \tag{87}$$

where L is defined by (73) and (74). From Appendix A, we have

$$1 \leqq (1 + m)U/L \leqq 2, \tag{88}$$

and these bounds cannot be improved without restriction on the voltage-charge relationship. However, there is a large class of voltage-charge relationships for which the lower bound is attained, namely those for which $[(\rho + m)V(1) - (m + 1)V(\rho) + (1 - \rho)V(-m)]$ does not change sign in $-m \leqq \rho \leqq 1$. From (85) and (87) it follows that

$$\frac{3\sqrt{3}}{2} \leqq \frac{\max P_1}{L} \leqq 4, \qquad \text{if} \qquad L = (1 + m)U. \tag{89}$$

Also, for the above class, L in (74) is given with $\sigma = 1$, $\tau = -m$, or vice versa, and U in (86) is given with $\lambda = [V(1) - V(-m)]/(1 + m)$. A class of voltage-charge relationships of interest is

$$\mathrm{V}(Q) = [\max\ (0,Q)]^{\nu}, \quad -m \leqq Q \leqq 1; \qquad m \geqq 0, \quad \nu \geqq 1, \tag{90}$$

of which we have already considered the cases $\nu = 2$, $\nu = 1$, and $\nu = \frac{3}{2}$ (with $m = 0$). It is readily seen that this class satisfies the above condition, and hence

$$
\begin{aligned}
L &= \max_{-m \leqq \rho \leqq 1} \{(\rho + m) - (1 + \ m)\,[\max\,(0,\rho)]^{\nu}\}\,; \\
U &= \max_{-m \leqq \rho \leqq 1} \left\{\frac{\rho}{(1 + m)} - [\max\,(0,\rho)]^{\nu}\right\} \\
&\qquad\qquad - \min_{-m \leqq \rho \leqq 1} \left\{\frac{\rho}{(1 + m)} - [\max\,(0,\rho)]^{\nu}\right\}.
\end{aligned}
\tag{91}
$$

Thus,

$$L = m + \left(1 - \frac{1}{\nu}\right) [(1 + m)\nu]^{-1/(\nu-1)} = (1 + m)U, \tag{92}$$

and the bounds on $(\max P_1)/L$ in (89) hold. From (69) and (92) it is seen that the lower bound is exact for the case $\nu = 1$, $m > 0$. For $\nu = \frac{3}{2}$ and $m = 0$, $3\sqrt{3}\ L/2 = 0.385$ as compared with $\max P_1 = 0.408$. For $\nu = 2$, $L = (2m + 1)^2/[4(1 + m)]$, and Table I shows the ratio $2(\max P_1)/(3\sqrt{3}L)$ for several values of m, and it is noted that the lower bound improves with increasing m.

TABLE I — ($\nu = 2$)

m	0	0.589	1.20	1.89	3.07	5.50
$\dfrac{2(\max P_1)}{3\sqrt{3}L}$	1.058	1.042	1.027	1.018	1.012	1.006

V. THE MAXIMIZATION OF THE POWER TRANSFER FROM THE FUNDAMENTAL TO SECOND HARMONIC, WITH BOUNDED CHARGE WAVEFORM

5.1 *The Canonical Representation of the Charge Waveform*

We wish to consider the problem of maximizing the power transfer from the fundamental to the second harmonic, when the charge waveform contains no higher than second harmonics, so that

$$Q(x) = a + b \sin x + c \cos x + d \sin 2x + e \cos 2x. \tag{93}$$

We also impose the conditions

$$Q_{\max} = 1; \qquad Q_{\min} = -m. \tag{94}$$

Note that it does not follow a priori that the maximum power transfer subject to (94) is equal to the maximum subject to $Q_{\max} \leqq 1$, $Q_{\min} \geqq$

$-m$. We observe, however, that for the voltage-charge relationship (90),

$$\begin{aligned}&\max\,[P_1 \mid Q_{\max} = r,\, Q_{\min} = -q] \\ &\quad = r^{(\nu+1)} \max\,[P_1 \mid Q_{\max} = 1,\, Q_{\min} = -q/r], \qquad 0 < r \leqq 1,\end{aligned} \tag{95}$$

as may be deduced from (28) and (29). Thus it is sufficient to determine $\max\,[P_1 \mid Q_{\max} = 1, Q_{\min} = -m]$, that is, $\max P_1$ subject to the conditions of (94), for a range of values of m.

A canonical representation of $Q(x)$ is found in Appendix B. In addition to the two conditions in (94) it is supposed, by a suitable choice of time origin, that

$$Q(\pi) = Q_{\min} = -m. \tag{96}$$

Thus the five coefficients in (93) are given in terms of two parameters and it is found that

$$\begin{aligned}a &= [(c - e) - m]; \qquad\qquad b = 2d = (1 + m)sy; \\ c &= (1 + m)[\tfrac{1}{2}(1 - s^4) - s^2y]; \\ e &= \tfrac{1}{4}(1 + m)[y(1 - s^2) - \tfrac{1}{2}(1 + s^2)^2].\end{aligned} \tag{97}$$

The parameter s arises from the equation

$$Q(2 \tan^{-1} s) = Q_{\max} = 1. \tag{98}$$

The parameter y is subject to the condition

$$0 \leqq y \leqq (1 - s^2), \tag{99}$$

which of course also implies that $s^2 \leqq 1$. Thus we have a two-parameter canonical representation of $Q(x)$, and these two parameters lie in a bounded region. Moreover, it is shown in Appendix B that, independently of the voltage-charge relationship $V = V(Q)$,

$$P_1\,|_{y=0} = 0; \qquad P_1\,|_{y=(1-s^2)} = 0, \tag{100}$$

so that P_1 vanishes on the boundary of this region. Also it is seen, from (93) and (97), that changing the sign of s is equivalent to the transformation $Q(x) \rightarrow Q(2\pi - x)$, and hence

$$P_1(-s,y) = -P_1(s,y); \qquad P_1\,|_{s=0} = 0. \tag{101}$$

5.2 *The Abrupt-Junction Diode*

We now consider the abrupt-junction diode operated in the region between forward conduction and reverse breakdown. From (28), (29), (37), (93) and (97), with $m = 0$, it follows that

$$P_1 = -\frac{\pi^2}{4}\, s(1 + s^2)^2 y[(1 - s^2) - y]. \tag{102}$$

The maximum of (102) subject to (99) is

$$\max P_1 = \frac{4\pi^2}{81\sqrt{3}} = 0.2814, \tag{103}$$

being given by $s = -(1/\sqrt{3})$, $y = \frac{1}{3}$. Thus a charge waveform giving $\max P_1$ is

$$\bar{Q}(x) = \frac{1}{2} + \frac{1}{3\sqrt{3}}\left[2\sin\left(x + \frac{2\pi}{3}\right) + \sin 2\left(x + \frac{2\pi}{3}\right)\right], \tag{104}$$

and the corresponding fundamental reactive power is found to be

$$R_1 = \frac{4\pi^2}{27} = 1.462. \tag{105}$$

$Q(x) = \bar{Q}(x - (2\pi/3))$ is depicted in Fig. 6. It is interesting to compare (52) and (103), and Figs. 1 and 6. We comment that the above results may be obtained quite elegantly, without using the canonical representation of the charge waveform.

5.3 *A Charge Waveform Which Provides an Approximation to the Maximum Power Transfer*

In Section IV we obtained a lower bound to the maximum obtainable fundamental power, (87), and it was seen to be a close bound in the particular cases considered. The charge waveform giving this lower bound is one which has values σ, ρ, and τ on consecutive intervals of

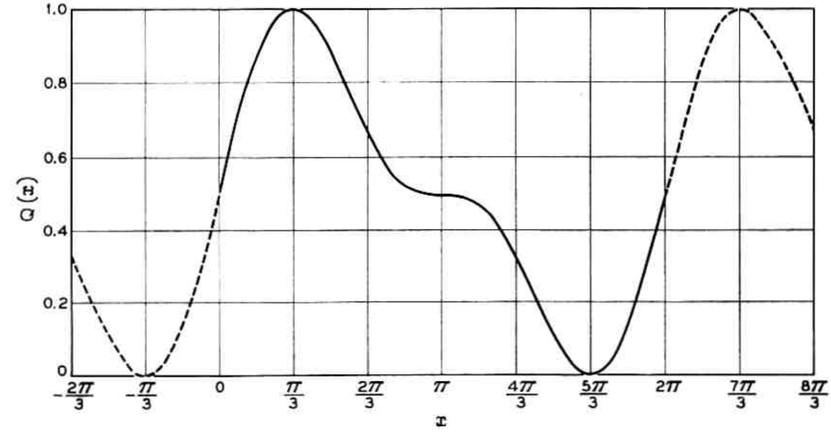


Fig. 6 — Charge waveform for maximum power transfer from fundamental to second harmonic in abrupt-junction diode operated in the region between forward conduction and reverse breakdown.

x of length $2\pi/3$. Here σ, ρ, and τ are those values which, subject to $-m \leqq (\sigma,\rho,\tau) \leqq 1$, maximize $L(\sigma,\rho,\tau)$, as defined in (73). It was also pointed out that if $[(\rho + m)V(1) - (1 + m)V(\rho) + (1 - \rho)V(-m)]$, i.e., $L(1,\rho,-m)$, does not change sign in $-m \leqq \rho \leqq 1$, then $L(\sigma,\rho,\tau)$ is maximized with $\sigma = 1$, $\tau = -m$ (or vice versa) and a suitable value of ρ. The class of voltage-charge relationships given in (90) satisfies this condition and then

$$\rho = [(1 + m)\nu]^{-1/(\nu-1)}. \tag{106}$$

Now the Fourier coefficients, up to the second harmonic as in(93), of the charge waveform giving the close lower bound to the maximum obtainable fundamental power, are

$$\begin{aligned} a &= \frac{1}{3}(\sigma + \rho + \tau); \qquad b = 2d = \frac{3}{2\pi}(\sigma - \tau); \\ c &= -2e = \frac{\sqrt{3}}{2\pi}(\sigma + \tau - 2\rho). \end{aligned} \tag{107}$$

We will restrict ourselves to that class of voltage-charge relationships, $V = V(Q)$, for which $L(\sigma,\rho,\tau)$ in (73) attains its maximum, subject to $-m \leqq (\sigma,\rho,\tau) \leqq 1$, when

$$\sigma = 1, \qquad \tau = -m, \qquad -m < \rho < 1. \tag{108}$$

It would seem feasible that we might obtain a reasonable approximation to the maximum power transfer from the fundamental to the second harmonic, by suitably shifting and expanding (or contracting) the above Fourier approximation, so that (94) is satisfied. Setting $\sigma = 1$, $\tau = -m$ in (107) and carrying out this procedure, we obtain the approximating charge waveform

$$\begin{aligned} \tilde{Q}(x) = \frac{1}{3}(1 + \rho - m) &+ \frac{\sqrt{3}}{9}(1 + m)(2\sin x + \sin 2x) \\ &+ \frac{1}{9}(1 - m - 2\rho)(2\cos x - \cos 2x). \end{aligned} \tag{109}$$

If we define $\bar{Q}(x) = \tilde{Q}[x + (2\pi/3)]$, then (96) is satisfied and in the canonical representation of $\bar{Q}(x)$, (97), we have

$$s = -\frac{1}{\sqrt{3}}, \qquad y = \frac{2(1 - \rho)}{3(1 + m)}. \tag{110}$$

For the abrupt-junction diode operated in the region between forward conduction and reverse breakdown, $\rho = \frac{1}{2}$, setting $m = 0$, $\nu = 2$ in

(106). Hence, from (110), $s = -(1/\sqrt{3})$ and $y = 1/3$, so that, from the previous section, the approximating charge waveform is actually the one which gives the maximum power transfer.

5.4 *The Numerical Computation of the Maximum Power Transfer, for Particular Diodes*

We have already obtained a two-parameter canonical representation of the charge waveform containing no higher than second harmonics and satisfying (94). The two parameters s and y lie in the bounded region given by (99), and P_1 vanishes, independently of the voltage-charge relationship, on the boundary of this region. Also, since P_1 is antisymmetric in s, it is sufficient to consider only half the region and to maximize $| P_1 |$. The maximization was carried out numerically for particular diodes, by means of the iterative process of fitting a quadric surface. As a starting point $s^{(1)},y^{(1)}$ in the process, that point corresponding to the approximating charge waveform, derived in the previous section, was used.

The results of the numerical computations for the voltage-charge relationship of (90), with $\nu = 2$ and $\nu = \frac{3}{2}$, and several values of m, are tabulated below. Tables II and III give the values of the maximum power transfer, *max* P_1, together with the corresponding values of the

TABLE II—($\nu = 2$)

m	*max* P_1	R_1	R_2	$I_{\max}$	$(b^2 + c^2)$	$(d^2 + e^2)$
0	0.2814	1.462	0.7310	0.7698	0.1482	0.0370
$\frac{1}{2}$	0.7773	1.966	1.060	1.160	0.3289	0.0865
1	1.284	2.300	1.300	1.549	0.5947	0.1573
2	2.198	2.921	1.561	2.310	1.451	0.3484
$\frac{7}{2}$	3.366	3.854	1.679	3.422	3.616	0.7414
5	4.371	4.788	1.642	4.515	6.869	1.250
7	5.544	6.020	1.474	5.951	12.92	2.097
9	6.586	7.228	1.230	7.372	20.95	3.096

TABLE III—($\nu = \frac{3}{2}$)

m	*max* P_1	R_1	R_2	$I_{\max}$	$(b^2 + c^2)$	$(d^2 + e^2)$
0	0.1623	1.514	0.7389	0.7684	0.1499	0.0366
$\frac{1}{2}$	0.6782	2.137	1.182	1.162	0.3257	0.0878
1	1.246	2.428	1.529	1.560	0.5635	0.1652
2	2.271	3.023	1.882	2.330	1.367	0.3682
$\frac{7}{2}$	3.575	4.034	2.006	3.445	3.479	0.7703
5	4.691	5.115	1.907	4.533	6.710	1.277

reactive powers in the fundamental and second harmonic, R_1 and R_2, and the maximum current $I_{\max}$ associated with the charge waveform $Q(x)$, that is the maximum value of $|Q'(x)|$. It is worth noting that R_2 does not continue to increase with m. Also included are the squares of the amplitudes of the first and second harmonics in the charge waveform, $(b^2 + c^2)$ and $(d^2 + e^2)$. These, together with the real and reactive powers, determine the normalized impedances. Tables IV and V give the values of $-s$ and y which given *max* P_1, and also $y^{(1)}$ and $P_1^{(1)}$, the value of P_1 corresponding to $y^{(1)}$ and $-s^{(1)} = 1/\sqrt{3} = 0.5774$. It is interesting to note how close $P_1^{(1)}$ is to *max* P_1, except for the larger values of m. Table VI compares *max* P_1 with the maximum obtainable fundamental power, max P_1, as obtained in Section III, for the case $\nu = 2$ and several values of m. It is also worth comparing the value of *max* $P_1 = 0.162$ for the case $\nu = \frac{3}{2}$, $m = 0$ with the corresponding value of max $P_1 = 0.408$.

Table IV—($\nu = 2$)

m	$-s$	$y^{(1)}$	y	$P_1^{(1)}$	*max* P_1
0	0.5774	0.3333	0.3333	0.2814	0.2814
$\frac{1}{2}$	0.5839	0.2963	0.2942	0.7770	0.7773
1	0.5848	0.2500	0.2426	1.283	1.284
2	0.5716	0.1852	0.1782	2.192	2.198
$\frac{7}{2}$	0.5465	0.1317	0.1301	3.307	3.366
5	0.5246	0.1019	0.1046	4.199	4.371
7	0.5008	0.0781	0.0844	5.197	5.544
9	0.4816	0.0633	0.0717	6.047	6.586

Table V—($\nu = \frac{3}{2}$)

m	$-s$	$y^{(1)}$	y	$P_1^{(1)}$	*max* P_1
0	0.5742	0.3704	0.3704	0.1622	0.1623
$\frac{1}{2}$	0.5871	0.3566	0.3562	0.6775	0.6782
1	0.5977	0.2963	0.2829	1.241	1.246
2	0.5875	0.2112	0.1989	2.262	2.271
$\frac{7}{2}$	0.5591	0.1449	0.1419	3.518	3.575
5	0.5331	0.1097	0.1130	4.536	4.691

Table VI—($\nu = 2$)

m	0	$\frac{1}{2}$	1	2	$\frac{7}{2}$	5
max P_1	0.281	0.777	1.28	2.20	3.37	4.37
max P_1	0.687	1.83	3.02	5.50	9.33	13.15

5.5 *On the Power Transfer From the Fundamental to the Third Harmonic*

Breitzer et al[3] were also concerned with the abrupt-junction diode operated in the region between forward conduction and reverse breakdown and considered charge waveforms containing no higher than third harmonics. They treated in detail the power transfer from the fundamental to the third harmonic, subject to $P_2 = 0$, and obtained a maximum value of

$$P_1 = (0.0242)\pi^2 = 0.238 = -P_3, \tag{111}$$

making allowance for the difference in notation. This value of P_1 arose from two distinct charge waveforms. One was

$$Q(x) = (0.5) + (0.310)\sin x + (0.168)\sin 2x + (0.155)\sin 3x, \tag{112}$$

and the other was quite close to this. We saw previously how by taking the Fourier approximation, containing up to second harmonics, of a charge waveform which gives a good lower bound for max P_1 subject only to restrictions on $Q_{\max}$ and $Q_{\min}$, and suitably shifting and expanding (or contracting) so that the restrictions on $Q_{\max}$ and $Q_{\min}$ are satisfied by the approximating charge waveform, we could obtain a good approximation to the maximum power transfer from the first to second harmonic, when no higher than second harmonics are allowed. In the case of the abrupt-junction diode operated in the region between forward conduction and reverse breakdown, which is the diode that we will consider in this section, it was found that the charge waveform so derived was precisely one that gives the maximum power transfer.

Now, it is found that the best mean square approximation containing up to third harmonics, and subject to $P_2 = 0$, to the charge waveform which gives the good lower bound to the maximum obtainable fundamental power is

$$\bar{Q}(x) = \left[\frac{1}{2} + \frac{3}{\pi}f(x)\right], \tag{113}$$

where

$$\mathrm{f}(x) = [(0.4)\sin x + (0.25)\sin 2x + (0.2)\sin 3x]. \tag{114}$$

We shift and contract $\bar{Q}(x)$ by setting

$$Q(x) = \frac{1}{2}\left[1 + \frac{f(x)}{M}\right]; \qquad M = \max\,[f(x)], \tag{115}$$

so that $Q_{\max} = 1$ and $Q_{\min} = 0$. For this charge waveform,

$$P_1 = (0.0075)\pi^2/M^3 = -P_3; \qquad P_2 = 0. \tag{116}$$

It is found that

$$M = 0.680; \qquad P_1 = 0.235, \tag{117}$$

and $Q(x)$, as given by (114) and (115) is plotted in Fig. 7(a). The value of P_1 in (117) is very close to the maximum value obtained by Breitzer et al, (111), and it is interesting to compare Fig. 7(a) with Fig. 7(b) which depicts $Q(x)$ as given by (112).

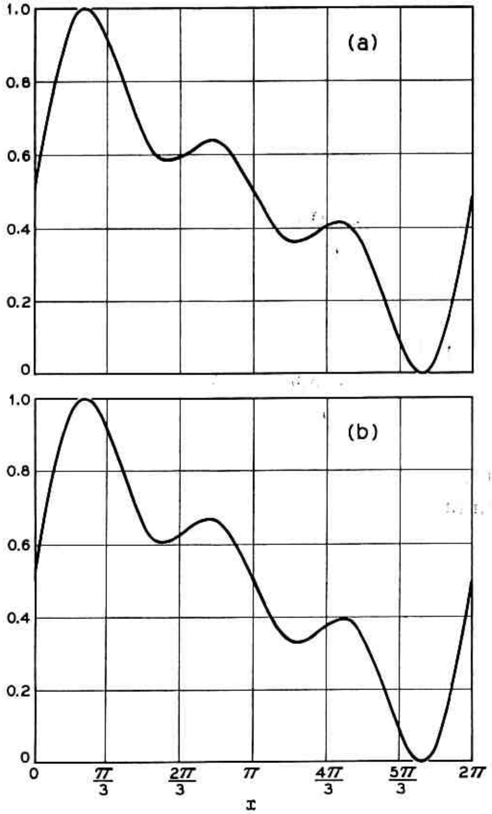


Fig. 7 — Charge waveforms giving (a) approximately, and (b) exactly, the maximum power transfer from fundamental to third harmonic in abrupt-junction diode operated in the region between forward conduction and reverse breakdown.

VI. THE MAXIMIZATION OF THE POWER TRANSFER FROM THE FUNDAMENTAL TO THE SECOND HARMONIC, FOR THE CURRENT-LIMITED DIODE

6.1 *The Canonical Representation of the Charge Waveform*

We are concerned with charge waveforms as in (93) and impose the restrictions

$$Q_{\max} = 1, \qquad | Q' |_{\max} = k. \tag{118}$$

We observe that, for voltage-charge relationships of the form given by (90),

$$\begin{aligned} &\max \left[P_1 \mid Q_{\max} = p, | Q' |_{\max} = l\right] \\ &= p^{(\nu+1)} \max \left[P_1 \mid Q_{\max} = 1 , | Q' |_{\max} = \frac{l}{p} \right], \qquad 0 < p \leqq 1 . \end{aligned} \tag{119}$$

In Appendix C we determine a canonical representation of $Q(x)$, subject to the conditions

$$Q_{\max} = 1; \qquad Q'_{\max} \leqq k; \qquad Q'_{\min} = -k, \tag{120}$$

by making use of the canonical representation obtained in Section 5.1, when the charge waveform has prescribed maximum and minimum values. Note that if $\tilde{Q}(x) = Q(\pi - x)$, where $Q(x)$ satisfies the conditions of (120), then

$$\tilde{Q}_{\max} = 1; \qquad \tilde{Q}'_{\max} = k; \qquad \tilde{Q}'_{\min} \geqq -k. \tag{121}$$

From Appendix C, the five coefficients in (93) are given in terms of two parameters s and y. It is found that

$$b = \frac{kw}{(w - z)}; \qquad 2d = \frac{kz}{(w - z)}; \qquad c = -\frac{ksy}{(w - z)} = 4e, \tag{122}$$

where

$$w = [\tfrac{1}{2}(1 - s^4) - s^2y]; \qquad z = \tfrac{1}{4}[y(1 - s^2) - \tfrac{1}{2}(1 + s^2)^2], \tag{123}$$

and that

$$a = [1 - \max (b \sin x + c \cos x + d \sin 2x + e \cos 2x)], \tag{124}$$

which in general has to be determined numerically. The waveform is translated so that

$$Q'(\pi) = -k = Q'_{\min}. \tag{125}$$

The parameter s arises from the equation

$$Q'(2 \tan^{-1} s) = Q'_{\max} \leqq k, \tag{126}$$

and the parameter y is subject to the condition

$$0 \leqq y \leqq \tfrac{1}{2}(1 - s^2), \tag{127}$$

which of course also implies that $s^2 \leqq 1$. It is shown in Appendix C that

$$P_1 \equiv 0 \quad \text{for} \quad y = \tfrac{1}{2}(1 - s^2). \tag{128}$$

In order to maximize P_1 subject to (118), it is sufficient, in view of the correspondence between $Q(x)$ and $\tilde{Q}(x) = Q(\pi - x)$ given by (120) and (121), to use the above canonical representation and to maximize $| P_1 |$.

6.2 *The Abrupt-Junction Diode*

We now consider the abrupt-junction diode operated in the region between forward conduction and reverse breakdown, for which the voltage-charge relationship is $V(Q) = Q^2$, $0 \leqq Q \leqq 1$. We first maximize $| P_1 |$ subject to the conditions of (120), and suppose that k is sufficiently small that $Q_{\min} \geqq 0$. Using the canonical representation obtained in the previous section, P_1 may be expressed in terms of s and y. Omitting the details, it is found that $| P_1 |$ is maximized, subject to the restriction (127), for $s^2 = \frac{1}{3}$, $y = 0$. The charge waveform giving this maximum is

$$Q(x) = 1 + k[S(x) - S_{\max}], \tag{129}$$

where

$$S(x) = \frac{(4 \sin x - \sin 2x)}{6}. \tag{130}$$

It is readily verified that $S_{\max} = g = -S_{\min}$, where

$$g = \frac{(1 + \sqrt{3})}{2\sqrt{2}(3)^{\frac{1}{4}}} = 0.734. \tag{131}$$

Thus $Q_{\min} = (1 - 2gk)$, so that $Q_{\min} \geqq 0$ for $2gk \leqq 1$. This $Q(x)$ actually gives a negative value of P_1, so that $\tilde{Q}(x) = Q(\pi - x)$ maximizes P_1, and it is found that

$$\max P_1 = \frac{2\pi^2}{27} k^3 = 0.7311k^3, \qquad \text{for} \quad k \leqq \frac{1}{2g} = 0.681. \tag{132}$$

Fig. 8 depicts $S(\pi - x)$.

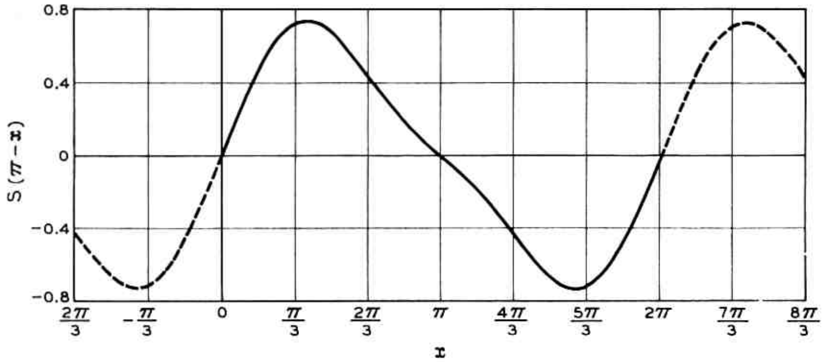


Fig. 8 — Shifted and normalized charge waveform for maximum power transfer from fundamental to second harmonic in current-limited abrupt-junction diode, with maximum current less than a critical value.

The fundamental reactive power corresponding to max P_1 is

$$R_1 = \frac{8\pi^2}{9}(1 - gk)k^2. \tag{133}$$

But, for the voltage-charge relationship $V(Q) = Q^2$, the addition of a constant to the charge waveform does not affect P_1. Hence, if instead of requiring $Q_{\max} = 1$ we just require $0 \leqq Q(x) \leqq 1$, we have

$$R_1 = \frac{8\pi^2}{9} ak^2 = 8.78ak^2; \qquad gk \leqq a \leqq (1 - gk). \tag{134}$$

6.3 *The Optimum Operating Frequency*

So far, no discussion has been made of the angular frequency ω of the actual periodicity of the charge waveform. We here consider this factor in the case of the abrupt-junction diode operated in the region between forward conduction and reverse breakdown. Now the physical limitation placed on the maximum current magnitude takes the form

$$| Q'(x) | \leqq \frac{\kappa}{\omega}, \tag{135}$$

from (20). Also, the actual fundamental power p_1 is, from (9), proportional to ωP_1. We thus consider the maximization of ωP_1 as ω varies, where the charge waveform $Q(x)$, containing no higher than second harmonics, is subject to $0 \leqq Q(x) \leqq 1$ and the condition in (135). We make use of results from Section V, as well as from the previous section.

Thus, we define

$$\max [P_1 \mid Q_{\max} = 1, Q_{\min} = -m] = P(m), \tag{136}$$

and let the value of $| Q' |_{\max}$ for the charge waveform which gives $P(m)$ be denoted by $K(m)$. Then, remembering that the addition of a constant to the charge waveform does not affect P_1, since $V(Q) = Q^2$, we obtain from (95) and (103),

$$P(m) = \frac{4\pi^2}{81\sqrt{3}} (1 + m)^3, \tag{137}$$

and also, from (104),

$$K(m) = (1 + m)K(0) = \frac{4}{3\sqrt{3}} (1 + m). \tag{138}$$

Similarly, we define

$$\max [P_1 \mid Q_{\max} = 1, | Q' |_{\max} = k] = \Pi(k), \tag{139}$$

and let the value of $Q_{\min}$ for charge waveforms which give $\Pi(k)$ be denoted by $-M(k)$. Then, from the previous section,

$$\Pi(k) = \frac{2\pi^2}{27} k^3; \qquad M(k) = -(1 - 2gk), \tag{140}$$

where g is given by (131).

Now if $Q(x)$ is subject to just the restriction $0 \leqq Q(x) \leqq 1$, then $\max P_1 = P(0)$, from (137). But, from (138), if $(\omega/\kappa) \leqq (3\sqrt{3})/4$ then the $Q(x)$ which give this value of $\max P_1$ satisfy (135). Hence,

$$\left(\frac{\omega}{\kappa}\right) \max P_1 = 0.2814 \left(\frac{\omega}{\kappa}\right), \qquad 0 \leqq \left(\frac{\omega}{\kappa}\right) \leqq 1.299. \tag{141}$$

Note that if $(\omega/\kappa) > 1.299$, then this gives an upper bound on $(\omega/\kappa) \max P_1$. Also, if $(\omega/\kappa) > 1.299$, then $\max P_1 \geqq P(m)$ if $K(m) = (\kappa/\omega)$, and hence, from (137) and (138),

$$\left(\frac{\omega}{\kappa}\right) \max P_1 \geqq 0.617 \left(\frac{\kappa}{\omega}\right)^2, \qquad \left(\frac{\omega}{\kappa}\right) \geqq 1.299. \tag{142}$$

From (140), setting $k = (\kappa/\omega)$, we have

$$\left(\frac{\omega}{\kappa}\right) \max P_1 = 0.731 \left(\frac{\kappa}{\omega}\right)^2, \qquad \left(\frac{\omega}{\kappa}\right) \geqq 1.468, \tag{143}$$

and if $0 \leqq (\omega/\kappa) < (1.468) = 2g$, then this provides an upper bound on $(\omega/\kappa) \max P_1$. Also, if $0 \leqq (\omega/\kappa) < 2g$, then $\max P_1 \geqq \Pi[1/(2g)]$, from (140). Hence,

$$\left(\frac{\omega}{\kappa}\right) \max P_1 \geqq 0.231 \left(\frac{\omega}{\kappa}\right), \qquad 0 \leqq \left(\frac{\omega}{\kappa}\right) \leqq 1.468. \tag{144}$$

Fig. 9 shows (ω/κ) max P_1 as a function of (ω/κ). For

$$1.299 \leqq \left(\frac{\omega}{\kappa}\right) \leqq 1.468, \tag{145}$$

the curve lies between the dashed lines. Thus, from the viewpoint of maximizing the actual fundamental real power, the optimum operating frequency, when the diode is not allowed to operate in the forward conduction region, lies in the range given by (145). Also, we can assert that

$$0.3655 = \frac{\pi^2}{27} \leqq \max\left[\left(\frac{\omega}{\kappa}\right) P_1\right] \leqq \frac{2(4)^{\frac{1}{3}}\pi^2}{81} = 0.387. \tag{146}$$

6.4 *Maximization of the Power Transfer, When the Region of Operation Includes Forward Conduction*

In a previous section we obtained a canonical representation of a charge waveform $Q(x)$, containing no higher than second harmonics, for which $Q_{\max} = 1$, $Q'_{\max} \leqq k$ and $Q'_{\min} = -k$. This canonical representation is given by (93), (122), (123) and (124), and involves two parameters s and y which lie in a bounded domain given by $0 \leqq y \leqq \frac{1}{2}(1 - s^2)$. It was shown that, independently of the voltage-charge relationship, $P_1 = 0$ on $y = \frac{1}{2}(1 - s^2)$. Moreover, it was seen that in order to maximize P_1 subject to $Q_{\max} = 1$ and $|\, Q' \,|_{\max} = k$, it is sufficient to consider this canonical representation and to maximize $|\, P_1 \,|$.

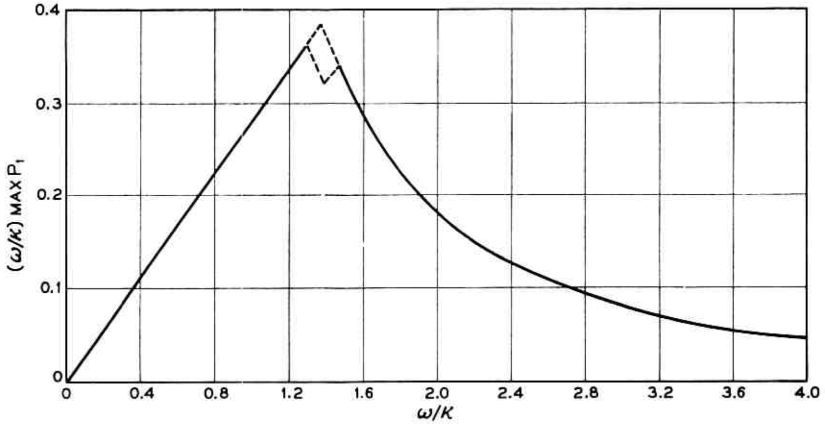


Fig. 9 — Maximum power transfer from fundamental to second harmonic in current-limited abrupt-junction diode operated in the region between forward conduction and reverse breakdown, vs. frequency.

The maximization was carried out analytically for the abrupt-junction diode when k is sufficiently small that the diode does not operate in the forward conduction region. We treat here, by means of numerical computation, the abrupt-junction diode when partial operation in the forward conduction region takes place, the normalized voltage-charge relationship being given by (61).

Again, the maximization process was that of fitting a quadric surface, and this time it was also necessary to calculate a in (124) numerically. Further, it was desirable to first compute the value of P_1 over a rough grid, and then to pick appropriate values $s^{(1)}$ and $y^{(1)}$, as a starting point in the maximization process. Thus, for several values of k, $\max |P_1|$, i.e., $\Pi(k)$ in the notation of (139), was computed in the manner described above. For the values of s and y which gave $\max |P_1|$, the corresponding values of R_1 and R_2, the reactive powers in the fundamental and second harmonic, and of $Q_{\min}$, i.e., $-M(k)$ in the notation of the previous section, were calculated, together with $(b^2 + c^2)$ and $(d^2 + e^2)$, the squares of the amplitudes of the first and second harmonics in the charge waveform. The results of the numerical computations are tabulated in Table VII. We note that the values of P_1 corresponding to the given values of s and y are negative. If $Q(x)$ is the charge waveform corresponding to s and y, (93), (122), (123), and (124), then the positive value of P_1, that is $\Pi(k)$, is obtained from the charge waveform $\tilde{Q}(x) = Q(\pi - x)$, or any translation thereof.

Now, from (119) with $\nu = 2$, and from (139),

$$\max [P_1 \mid Q_{\max} = p, |Q'|_{\max} = l] = p^3\Pi\left(\frac{l}{p}\right), \qquad 0 < p \leqq 1. \quad (147)$$

For $Q_{\max} \leqq 0$ we have $P_1 \equiv 0$, from (61). We may write

$$\frac{p^3\Pi\left(\frac{l}{p}\right)}{\Pi(k)} = \frac{\left(\frac{l}{p}\right)^{-3}\Pi\left(\frac{l}{p}\right)}{l^{-3}\Pi(l)} \cdot \frac{\Pi(l)}{\Pi(k)}. \quad (148)$$

The quantity $k^{-3}\Pi(k)$ is depicted in Fig. 10(a), and it is seen to be a nonincreasing function of k. It follows, from (147) and (148), since $\Pi(k)$ is a strictly increasing function of k, that $\max P_1$ subject to $Q_{\max} \leqq 1$ and $|Q'|_{\max} \leqq k$ is attained with $Q_{\max} = 1$ and $|Q'|_{\max} = k$. For $k < 1/(2g) = 0.681$, it can also be attained with $2gk \leqq Q_{\max} < 1$ and $|Q'|_{\max} = k$. We comment that for the voltage-charge relationship $V(Q) = \max(0,Q)$, $\max P_1$ subject to $Q_{\max} \leqq 1$ and $|Q'|_{\max} \leqq k$ is not attained with $Q_{\max} = 1$, for sufficiently small k, since in this case $P_1 \equiv 0$ if $Q_{\min} \geqq 0$.

Let us now consider the frequency factor, as we did at the end of the previous section, so that (135) holds. Hence, setting $k = (\kappa/\omega)$.

$$\max\left[\frac{\omega}{\kappa} P_1\right] = \max\left[\frac{\Pi(k)}{k}\right]. \tag{149}$$

The curve in Fig. 10(b) depicts $\Pi(k)/k$ and it is seen to be an increasing function of k in the range shown, although it is to be expected that it tends to zero as $k \rightarrow \infty$. It appears that max $[\Pi(k)/k] \sim 1$, so that, from (146), a considerable improvement is obtained if the diode is permitted to operate in the forward conduction region. We must bear in mind, however, that we have idealized the voltage-charge relationship in the forward conduction region.

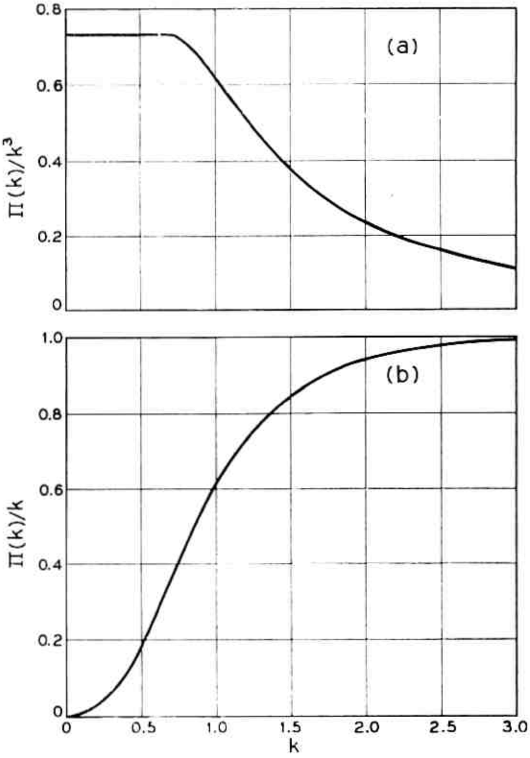


Fig. 10 — Maximum power transfer from fundamental to second harmonic divided by (a) the cube of the maximum current, and (b) the maximum current, for current-limited abrupt-junction diode with operation in forward conduction region permitted, vs. the maximum current.

Table VII

k	$\Pi(k)$	R_1	R_2	$M(k)$
0.75	0.3058	2.167	0.3033	0.0896
1.0	0.6159	2.440	0.5075	0.3980
1.5	1.265	2.840	0.8274	1.027
2.25	2.169	3.483	1.058	2.013
3.0	2.979	4.153	1.135	3.024
k	$-s$	y	$(b^2 + c^2)$	$(d^2 + e^2)$
0.75	0.5988	0.0018	0.2404	0.0169
1.0	0.6647	0.0288	0.3653	0.0393
1.5	0.7035	0.0834	0.7159	0.1124
2.25	0.7058	0.1354	1.613	0.2775
3.0	0.6996	0.1663	3.006	0.5039

VII. ACKNOWLEDGMENTS

The writer takes pleasure in thanking J. M. Early, who posed this problem, for his several discussions and helpful comments. He is also grateful to C. F. Pease for programming the numerical maximization processes, and to Mrs. J. D. Root and Miss M. A. Lounsberry for programming the numerical computations in connection with the maximum obtainable fundamental power.

APPENDIX A

From (28) and (29), for any λ (which we take to be real),

$$\begin{aligned} P_1 &= \left(\int_0^{2\pi} \{\lambda Q(x) - V[Q(x)]\} \sin x \, dx\right)\left(\int_0^{2\pi} Q(x) \cos x \, dx\right) \\ &\quad - \left(\int_0^{2\pi} \{\lambda Q(x) - V[Q(x)]\} \cos x \, dx\right)\left(\int_0^{2\pi} Q(x) \sin x \, dx\right) \qquad (150) \\ &= \Delta \int_0^{2\pi} \{\lambda Q(x) - V[Q(x)]\} \sin (x - \theta) \, dx, \end{aligned}$$

where

$$\Delta \sin \theta = \int_0^{2\pi} Q(x) \sin x \, dx; \qquad \Delta \cos \theta = \int_0^{2\pi} Q(x) \cos x \, dx. \qquad (151)$$

Hence,

$$\Delta = \int_0^{2\pi} Q(x) \cos (x - \theta) \, dx. \qquad (152)$$

Now, $\max P_1 = \max | P_1 |$. Since

$$\left| \int_0^{2\pi} f(x) \sin (x - \varphi)\, dx \right| \leqq 2(\max f - \min f), \tag{153}$$

(31), (150) and (152) lead to (85) and (86) in Section 4.2. We next derive the inequalities (88), where L is defined by (73) and (74). Now,

$$\begin{aligned} L &\geqq \max_{-m \leqq \sigma \leqq 1} L(\sigma, 1, -m) \\ &= \max_{-m \leqq \sigma \leqq 1} [(1 + m)V(\sigma) - (m + \sigma)V(1) + (\sigma - 1)V(-m)] \\ &= -(1 + m) \min_{-m \leqq \sigma \leqq 1} \left\{ \frac{\sigma[V(1) - V(-m)]}{(1 + m)} - V(\sigma) \right\} \\ &\qquad - [mV(1) + V(-m)]. \end{aligned} \tag{154}$$

Also,

$$\begin{aligned} L &\geqq \max_{-m \leqq \rho \leqq 1} L(1, \rho, -m) \\ &\qquad = (1 + m) \max_{-m \leqq \rho \leqq 1} \left\{ \frac{\rho[V(1) - V(-m)]}{(1 + m)} - V(\rho) \right\} \\ &\qquad\qquad + [mV(1) + V(-m)]. \end{aligned} \tag{155}$$

Hence, from (86), (154) and (155),

$$2L \geqq (1 + m)U. \tag{156}$$

Also, from (73) and (74),

$$\begin{aligned} L = \max_{-m \leqq (\sigma, \rho, \tau) \leqq 1} \{ &(\tau - \rho)[\lambda\sigma - V(\sigma)] + (\sigma - \tau)[\lambda\rho - V(\rho)] \\ &+ (\rho - \sigma)[\lambda\tau - V(\tau)]\}, \end{aligned} \tag{157}$$

for any (real) λ. In view of the remarks preceding (74) we may assume either that $-m \leqq \sigma \leqq \rho \leqq \tau \leqq 1$, or that $-m \leqq \tau \leqq \rho \leqq \sigma \leqq 1$, without loss of generality. In the former case

$$\begin{aligned} &(\tau - \rho)[\lambda\sigma - V(\sigma)] + (\sigma - \tau)[\lambda\rho - V(\rho)] + (\rho - \sigma)[\lambda\tau - V(\tau)] \\ &\leqq (\tau - \rho) \max_{-m \leqq \kappa \leqq 1} [\lambda\kappa - V(\kappa)] + (\sigma - \tau) \min_{-m \leqq \kappa \leqq 1} [\lambda\kappa - V(\kappa)] \\ &\qquad + (\rho - \sigma) \max_{-m \leqq \kappa \leqq 1} [\lambda\kappa - V(\kappa)] \\ &= (\tau - \sigma)\{ \max_{-m \leqq \kappa \leqq 1} [\lambda\kappa - V(\kappa)] - \min_{-m \leqq \kappa \leqq 1} [\lambda\kappa - \mathrm{V}(\kappa)]\}. \end{aligned} \tag{158}$$

Hence

$$\begin{aligned}(\tau - \rho)[\lambda\sigma - V(\sigma)] + (\sigma - \tau)[\lambda\rho - V(\rho)] + (\rho - \sigma)[\lambda\tau - V(\tau)]& \\ \leqq (1 + m)\{\max_{-m \leqq \kappa \leqq 1} [\lambda\kappa - V(\kappa)] - \min_{-m \leqq \kappa \leqq 1} [\lambda\kappa - V(\kappa)]\}.&\end{aligned} \tag{159}$$

Equation (159) may be derived, in a similar manner, when $-m \leqq \tau \leqq \rho \leqq \sigma \leqq 1$. Thus, from (157),

$$L \leqq (1 + m)\{\max_{-m \leqq \kappa \leqq 1} [\lambda\kappa - V(\kappa)] - \min_{-m \leqq \kappa \leqq 1} [\lambda\kappa - V(\kappa)]\}. \tag{160}$$

But this is true for all (real) λ. Hence, from (86)

$$L \leqq (1 + m)U. \tag{161}$$

If we do not restrict the voltage-charge relationship then the bounds given by (156) and (161) cannot be improved. This is demonstrated by considering the (somewhat artificial) relationship

$$V(Q) = \begin{cases} 1, & Q = [(1 + m)\alpha - m]; \\ -1, & Q = [1 - (1 + m)\alpha]; \\ 0, & \text{otherwise}; \quad m > -1, \qquad 0 < \alpha < \tfrac{1}{2}. \end{cases} \tag{162}$$

It may be verified that in this case

$$L = (1 + m) = (1 - \alpha)(1 + m)U. \tag{163}$$

We now find a class of voltage-charge relationships for which the bound in (161) is attained. If $\sigma \geqq \tau$, then, by the definition of U in (86),

$$\begin{aligned}(\sigma - \tau)U = \leqq \max_{-m \leqq \rho \leqq 1} \{[V(\sigma) - V(\tau)]\rho - (\sigma - \tau)V(\rho)\}& \\ - \min_{-m \leqq \rho \leqq 1} \{[V(\sigma) - V(\tau)]\rho - (\sigma - \tau)V(\rho)\}.&\end{aligned} \tag{164}$$

Let $-m \leqq \tau \leqq \sigma \leqq 1$. Then,

$$\begin{aligned} L &\geqq \max_{-m \leqq \rho \leqq 1} [(\rho - \tau)V(\sigma) + (\tau - \sigma)V(\rho) + (\sigma - \rho)V(\tau)] \\ &= \max_{-m \leqq \rho \leqq 1} \{[V(\sigma) - V(\tau)]\rho - (\sigma - \tau)V(\rho)\} \\ &\qquad\qquad + [\sigma V(\tau) - \tau V(\sigma)] \\ &\geqq (\sigma - \tau)U + [\sigma V(\tau) - \tau V(\sigma)] \\ &\qquad + \min_{-m \leqq \rho \leqq 1} \{[V(\sigma) - V(\tau)]\rho - (\sigma - \tau)V(\rho)\} \\ &= (\sigma - \tau)U \\ &\qquad + \min_{-m \leqq \rho \leqq 1} [(\rho - \tau)V(\sigma) + (\tau - \sigma)V(\rho) + (\sigma - \rho)V(\tau)]. \end{aligned} \tag{165}$$

$$Q(\pi/2) = [\tfrac{1}{4}(3 - 2y)(1 + m) - m] = [1 - \tfrac{1}{4}(1 + 2y)(1 + m)], \quad (181)$$

so that $Q_{\max} > 1$ for $s = 0, y < -\frac{1}{2}$ and $Q_{\min} < -m$ for $s = 0, y > \frac{3}{2}$. Hence $Q_{\max} > 1$ for y < 0, and $Q_{\min} < -m$ for $y > (1 - s^2)$. The region of interest, i.e., $Q_{\max} = 1$ and $Q_{\min} = -m$, is given by

$$0 \leqq y \leqq (1 - s^2). \quad (182)$$

We next consider the fundamental power when the charge waveform $Q(x)$ contains no higher than second harmonics. From (28), (29) and (170),

$$P_1 = \pi \int_0^{2\pi} V[Q(x)](b \cos x - c \sin x)\, dx. \quad (183)$$

We determine conditions under which $P_1 \equiv 0$, independently of the voltage-charge relationship $V = V(Q)$. This is clearly the case if $b = 0 = c$, or if $Q(x)$, as given by (170), is a single-valued function of $(b \sin x + c \cos x)$, for then the integrand in (183) is the derivative of a periodic function. Noting that

$$2(b \sin x + c \cos x)^2 = (b^2 + c^2) + 2bc \sin 2x + (c^2 - b^2) \cos 2x, \quad (184)$$

it follows from (170) that the latter condition holds if

$$d = 2\lambda bc; \qquad \mathrm{e} = \lambda(c^2 - b^2), \quad (185)$$

for some λ. Combining this condition with $b = 0 = c$,

$$2bce + d(b^2 - c^2) = 0 \Rightarrow P_1 \equiv 0. \quad (186)$$

We now consider the canonical representation of $Q(x)$, with $Q_{\max} = 1$ and $Q_{\min} = -m$, wherein the coefficients in (170) are given by (97). Then condition (186) becomes, upon reduction,

$$y = 0, \qquad \text{or} \qquad y = (1 - s^2), \qquad \text{or} \qquad s = 0 \Rightarrow P_1 \equiv 0. \quad (187)$$

APPENDIX C

We here determine the canonical form of $Q(x)$, as given by (170), such that

$$Q_{\max} = 1; \qquad Q'_{\max} \leqq k; \qquad Q'_{\min} = Q'(\pi) = -k. \quad (188)$$

Now, when the charge waveform $\bar{Q}(x)$ is subject to $\bar{Q}_{\max} = 1$ and $\bar{Q}_{\min} = -m$, the five coefficients corresponding to those in (170) are given in terms of two parameters s and y, from (97), by

$$\begin{aligned} \bar{a} &= [(\bar{c} - \bar{e}) - m]; & \bar{b} &= 2\bar{d} = (1 + m)sy; \\ \bar{c} &= (1 + m)w; & \bar{e} &= (1 + m)z, \end{aligned} \tag{189}$$

where

$$w = [\tfrac{1}{2}(1 - s^4) - s^2 y]; \quad z = \tfrac{1}{4}[y(1 - s^2) - \tfrac{1}{2}(1 + s^2)^2]. \tag{190}$$

The charge waveform is translated so that $\bar{Q}(\pi) = -m$, which may be done without loss of generality. The parameter s arises from the condition $\bar{Q}(2 \tan^{-1} s) = 1$, and the parameter y is subject to the condition $0 \leqq y \leqq (1 - s^2)$, which of course also implies $s^2 \leqq 1$. If, in addition, $\bar{a} = 0$, then

$$(1 + m)(w - z) = m, \tag{191}$$

and hence, from (190),

$$2y(1 + 3s^2) = [(1 + s^2)(5 - 3s^2) - 8m/(1 + m)]. \tag{192}$$

Now $0 \leqq y \leqq (1 - s^2)$, but if we require $m \geqq 1$ then

$$0 \leqq y \leqq \tfrac{1}{2}(1 - s^2), \qquad (m \geqq 1). \tag{193}$$

Turning to a charge waveform $Q(x)$, as given by (170), which satisfies the conditions of (188), we may write

$$Q'(x) = \frac{k}{m} \bar{Q}(x), \qquad m \geqq 1, \tag{194}$$

where (191) and (193) hold. Hence,

$$Q'(x) = \frac{k[sy(\sin x + \frac{1}{2} \sin 2x) + w \cos x + z \cos 2x]}{(w - z)}. \tag{195}$$

Integrating, and remembering that $Q_{\max} = 1$,

$$Q(x) = \{1 + k[S(x) - S_{\max}]\}, \tag{196}$$

where

$$S(x) = \frac{\left[w \sin x + \dfrac{z}{2} \sin 2x - sy(\cos x + \frac{1}{4} \cos 2x)\right]}{(w - z)}. \tag{197}$$

In general, $S_{\max} = \max[S(x)]$ is determined numerically.

We now turn to the fundamental power, P_1, when $Q(x)$ has the above canonical representation. From (170), (186), (190), (196) and (197), we find that $P_1 \equiv 0$, independently of the voltage-charge relationship, if

$$[(1 - s^2) - 2y]\{(1 - s^2)(1 + 3s^2) - s^2[(1 - s^2) - 2y]^2\} = 0. \quad (198)$$

In view of (193), the second factor vanishes only if $s^2 = 1$, $y =$ 0 Hence we conclude that

$$P_1 \equiv 0 \qquad \text{for} \qquad y = \tfrac{1}{2}(1 - s^2). \quad (199)$$

REFERENCES

1. Early, J. M., Maximum Rapidly-Switchable Power Density in Junction Triodes, IRE Trans. on Electron Devices, **ED-6,** July, 1959, pp. 322–325.
2. Early, J. M., private communication.
3. Breitzer, D. I., Gardner, R., Greene, J. C., Lombardo, P. P., Salzberg, B., and Seigel, K., Third Quarterly Progress Report, Application of Semiconductor Diodes to Low-Noise Amplifiers, Harmonic Generators, and Fast-Acting TR Switches, Airborne Instruments Laboratory. Report No. 4589-I-3 (March, 1959).

The Design and Analysis of Pattern Recognition Experiments

By W. H. HIGHLEYMAN

(Manuscript received March 2, 1961)

A popular procedure for testing a pattern recognition machine is to present the machine with a set of patterns taken from the real world. The proportion of these patterns which are misrecognized or rejected is taken as the estimate of the error probability or rejection probability for the machine. In Part I, this testing procedure is discussed for the cases of unknown and known a priori probabilities of occurrence of the pattern classes. The differences between the tests that should be made in the two cases are noted, and confidence intervals for the test results are indicated. These concepts are applied to various published pattern recognition results by determining the appropriate confidence interval for each result.

In Part II, the problem of the optimum partitioning of a sample of fixed size between the design and test phases of a pattern recognition machine is discussed. One important nonparametric result is that the proportion of the total sample used for testing the machine should never be less than that proportion used for designing the machine, and in some cases should be a good deal more.

PART I — ON ANALYSIS

INTRODUCTION

There are two distinct and consecutive processes usually involved in the feasibility study of a pattern recognition method or machine. The first process is the actual design of the machine. This might be based upon a set of sample patterns which the experimenter has gathered, from which he estimates the parameters of the machine. Alternatively, the experimenter may base his design on some *a priori* knowledge concerning the pertinent characteristics of the pattern classes under study. The second process is then the testing of this machine either in its hardware form or by its simulation on a general purpose computer. A different set of sample patterns from that used in the design is used in this stage.

The popular procedure for interpreting the test results is to take the proportion of patterns in the test data which have been misrecognized or rejected by the machine as the estimates of the error probability and rejection probability, respectively, for the machine. There are several questions which might be raised concerning this testing procedure, such as:

1. Are these estimates the best estimates?
2. If so, how good are these estimates?
3. How does the estimate improve as the sample size is increased?

Questions such as these are discussed in Part I of this paper. Two cases are considered; one is the case in which the *a priori* probabilities of class occurrence are unknown, and the other case assumes full knowledge of the *a priori* probabilities.

Case 1. Unknown a priori Probabilities — Random Sampling

Let the number of allowable pattern classes be c. It will be assumed that, for each allowable class i, there exists an *a priori* probability of occurrence ω_i, a probability of error e_i, and a probability of rejection r_i. (For the rest of this paper, the term "error" will refer to an undetected error; all detected errors will be assumed to be rejected.) These probabilities are unknown to the experimenter, who is interested in estimating the overall probability of error for the machine.

$$e = \sum_{i=1}^{c} \omega_i e_i , \tag{1}$$

and the over-all probability of rejection,

$$r = \sum_{i=1}^{c} \omega_i r_i . \tag{2}$$

Let him perform the following experiment, which will be called random sampling. Consider the patterns to be randomly generated by a "pattern source" according to the *a priori* probabilities of occurrence. He takes a pattern from the source, identifies it, and then lets his pattern recognition machine attempt identification. He notes which of the three possible outcomes occurs: correct recognition, misrecognition, or rejection. This experiment is repeated n times, resulting in m_e samples which have been misrecognized and m_r samples which have been rejected.

Since these outcomes are mutually exclusive, and each experiment independent, then the resulting random variables, m_e and m_r, clearly are distributed according to the multinomial probability distribution. That is, the joint probability distribution of m_e and m_r, $P(m_e, m_r)$, is given by

$$P(m_e, m_r) = \binom{n}{m_e \; m_r} e^{m_e} r^{m_r} (1 - e - r)^{n - m_e - m_r} . \tag{3}$$

The maximum-likelihood estimates for e and r, denoted by $\hat{e}$ and $\hat{r}$, are then[1]

$$\begin{aligned} \hat{e} &= \frac{m_e}{n}, \\ \hat{r} &= \frac{m_r}{n}, \end{aligned} \tag{4}$$

which are the estimates in common use. Further, each of these estimates is proportional to a single random variable having a binomial distribution; therefore, $n\hat{e}$ and $n\hat{r}$ are themselves binomially distributed. The mean value of each estimate is the parameter for which it is an estimate; the variance of each is[1]

$$\sigma_{\hat{e}}^2 = \frac{1}{n^2} \sigma_{m_e}^2 = \frac{e(1 - e)}{n} \tag{6}$$

$$\sigma_{\hat{r}}^2 = \frac{r(1 - r)}{n}. \tag{7}$$

Because it is known that $n\hat{e}$ and $n\hat{r}$ are binomially distributed, confidence intervals can be applied to these estimates.* These confidence intervals require rather involved computations, but fortunately have been plotted for several values of n by various people.[3,4] In Fig. 1 is shown such a plot of intervals for a 95 per cent confidence level computed by C. S. Clopper and E. S. Pearson. The use of this graph is fairly simple. A vertical line extended upward from the observed value of the estimate given on the abscissa will intersect the pair of curves pertaining to the particular sample size used. Projecting these two intersections horizontally onto the ordinate axis gives an interval for the parameter being estimated. The probability is 0.95 that the interval drawn in this manner includes the parameter. For instance, if a sample size of $n = 250$ yielded 50 errors, then the estimate of the probability of error is 0.20. Using Fig. 1 it can be stated that, with probability 0.95, the true probability of error is included in the interval from 0.15 to 0.27.

* Mattson[2] has used a similar argument for determining convergence of an adaptive system. However, he used Tchebycheff's inequality to obtain confidence intervals which are necessarily larger than if he had used such intervals pertaining to the binomial distribution.

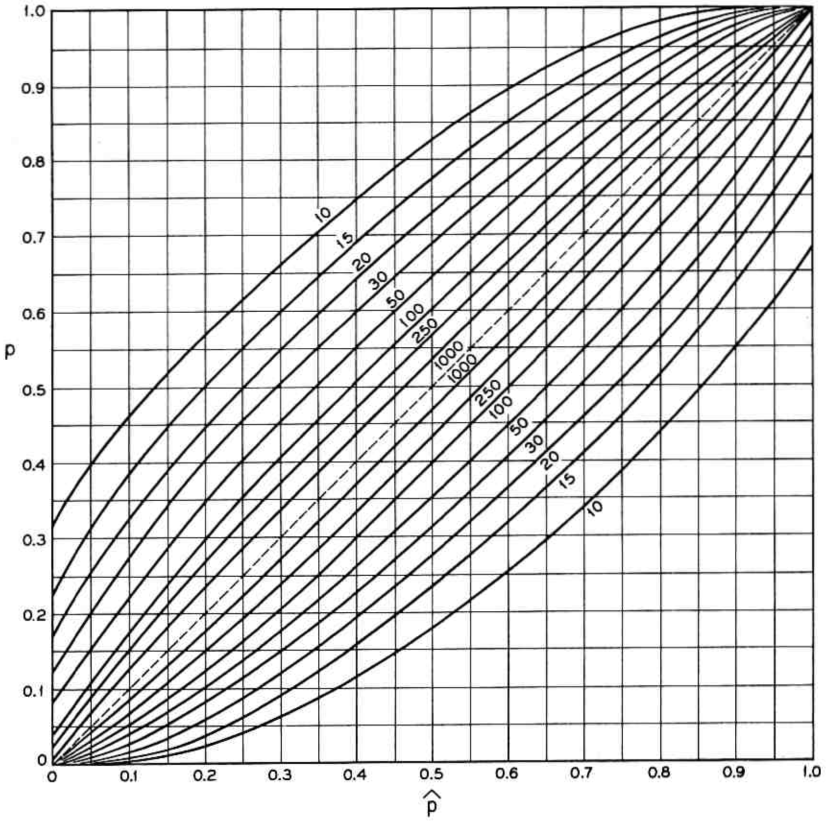


Fig. 1 — 95 per cent confidence intervals for a binomially distributed variable.

Case 2. Known a priori Probabilities — Selective Sampling

It is now assumed that the *a priori* probability of occurrence for each class, ω_i, is known. To take advantage of this knowledge, the experimenter takes n_i samples from each class i such that

$$\frac{n_i}{n} = \omega_i , \tag{8}$$

where n is the total number of samples. This process will be referred to as selective sampling.* (It will be assumed that the ω_i are such that (8) can be fulfilled with the desired sample size, n.)

* This sort of sampling dichotomy has been previously noted by others. For instance, Bowley[5] and Neyman[6] have referred to these two methods as "unrestricted" and "stratified" sampling.

The machine is again allowed to attempt recognition of these patterns, resulting in m_{e_i} samples from class i being misrecognized, and m_{r_i} samples from class i being rejected.

For any class i, the joint probability distribution for m_{e_i} and m_{r_i} again is multinomial:

$$P(m_{e_i}, m_{r_i}) = \binom{n_i}{m_{e_i}\; m_{r_i}} e_i^{m_{e_i}} r_i^{m_{r_i}} (1 - e_i - r_i)^{n_i - m_{e_i} - m_{r_i}}. \quad (9)$$

Since each of these distributions is independent of the others in this experiment, then the joint probability of the outcome for all c classes is the product of the individual probabilities (9):

$$P(m_{e_1}, \cdots, m_{e_c}, m_{r_1}, \cdots, m_{r_c}) \\ = \prod_{i=1}^{c} \binom{n_i}{m_{e_i}\; m_{r_i}} e_i^{m_{e_i}} r_i^{m_{r_i}} (1 - e_i - r_i)^{n_i - m_{e_i} - m_{r_i}}. \quad (10)$$

This is no longer a multinomial probability distribution. However, since the maximum-likelihood estimate of a sum of independent variables is the sum of the maximum-likelihood estimates, then these estimates for e and r are

$$\hat{e} = \frac{\sum_{i=1}^{c} m_{e_i}}{n}, \quad (11)$$

$$\hat{r} = \frac{\sum_{i=1}^{c} m_{r_i}}{n}, \quad (12)$$

which again agree with the popular practice of using the proportions as estimates. The random variables of which $n\hat{e}$ and $n\hat{r}$ are values are not now binomially distributed, since a sum of binomially distributed variables is not itself a binomial distribution in general.

The mean of each estimate is again the particular parameter being estimated. The variance of each of these estimates can be computed:

$$\sigma_{\hat{e}}'^2 = \frac{1}{n^2} \sum_{i=1}^{c} \sigma_{m_{e_i}}'^2 = \frac{1}{n^2} \sum_{i=1}^{c} n_i e_i (1 - e_i) = \frac{1}{n} \sum_{i=1}^{c} \omega_i e_i (1 - e_i), \quad (13)$$

in which use of (8) is made, and the prime distinguishes this variance from that for random sampling. Similarly,

$$\sigma_{\hat{r}}'^2 = \frac{1}{n} \sum_{i=1}^{c} \omega_i r_i (1 - r_i). \quad (14)$$

It is of interest to compare these variances for selective sampling with those obtained for the case of random sampling. Since the variance for $\hat{r}$ has the same form as $\hat{e}$ in both cases, it is necessary to consider only one of them, say $\hat{e}$. First note that $\sigma_{\hat{e}}^2$ can be written, using (1) and (6), as

$$\sigma_{\hat{e}}^2 = \frac{1}{n}\left(\sum_{i=1}^{c} \omega_i e_i\right)\left(1 - \sum_{k=1}^{c} \omega_k e_k\right). \tag{15}$$

From (13),

$$\sigma_{\hat{e}}^2 - \sigma_{\hat{e}}'^2 = \frac{1}{n}\sum_{i=1}^{c} \omega_i e_i^2 - \frac{1}{n}\left(\sum_{i=1}^{c} \omega_i e_i\right)^2. \tag{16}$$

Noting that $\sum_{i=1}^{c} \omega_i = 1$, (16) can be written as

$$\sigma_{\hat{e}}^2 - \sigma_{\hat{e}}'^2 = \frac{1}{n}\sum_{i=1}^{c} \omega_i\left(e_i - \sum_{k=1}^{c} \omega_k e_k\right)^2 = \frac{1}{n}\sum_{i=1}^{c} \omega_i (e_i - e)^2 = \sigma_e^2 \geqq 0. \tag{17}$$

Hence, the variance in the case of random sampling is greater than the variance in the case of selective sampling, the difference being what might be interpreted as the variance of the class errors. That is, if e_i is treated as a random variable with probability distribution ω_i, then σ_e^2 is the variance of e_i. (A similar derivation holds for the variance of the rejection probability estimates.) That the selective sampling variance should be smaller than the random sampling variance might be expected, since in selective sampling more information is used, namely the *a priori* probabilities.

Although statements have been made concerning the mean and variance of the estimates in the selective sampling case, nothing has been said yet concerning confidence intervals. This is a much more complicated problem than that in the case of random sampling, since the estimates do not have a simple distribution function. In fact, the confidence intervals will in general depend on the particular set of e_i's (or r_i's) pertaining to the machine, and not simply on e (or r).

However, for small probabilities, the binomial distribution is quite closely approximated by the Poisson distribution, the fit becoming perfect as the probability approaches zero. For any reasonable recognition machine, one would expect the probabilities of error and rejection to be small; consequently, the marginal form of (9) for m_{e_i} or m_{r_i} may be approximated by a Poisson distribution. The estimates given by (11) and (12) are now sums of random variables with Poisson distributions (approximately) which are then themselves Poisson distributed. If the over-all error is also small, as is usually the case, the binomial-Poisson approximation can now be used in reverse, and one may state that, for small error rates, the error and rejection estimates

(11) and (12) are approximately binomially distributed. Consequently, one can use Fig. 1 to obtain 95 per cent confidence intervals for the error and rejection probabilities. Further, from (17), we would expect this confidence interval to be on the safe side, that is, the actual 95 per cent confidence interval should be slightly smaller than this.

APPLICATION TO PUBLISHED RESULTS

To illustrate the ease of determining these confidence intervals, some published results in pattern recognition are listed in Table 1 along with the 95 per cent confidence intervals as determined from Fig. 1. It should be emphasized that Table I is not meant to compare one method against another, since the methods obviously treat problems of various complexities. Rather, the table is meant to compare the accuracies of the various evaluating experiments.

Three points of caution should be noted concerning the validity of the confidence intervals in this table. First, the author is not positive that the test data is different from the design data in every case. Second, to the best of the author's knowledge, in every case the number of samples taken from each allowable pattern class was predetermined. This is selective sampling; therefore, it is assumed that the proportion of samples taken from each class represents its *a priori* probability of occurrence. The third assumption is that the patterns used to test the machine are a reasonable sampling from the real-life world of patterns, and are not biased toward either well-formed or poorly-formed (noisy) patterns.

CONCLUSION

Two important cases concerning the testing of pattern recognition methods or machines have been considered: Random sampling for the case of unknown *a priori* probabilities of class occurrence, and selective sampling for the case of known *a priori* probabilities. The most predominant form of testing in the present day art is to assume that the pattern classes have equal *a priori* probabilities of occurrence, and consequently to use equal sample sizes for each class; this is a special case of selective sampling.

It has been shown that, for both cases, the maximum-likelihood estimate for the error probability or rejection probability is simply the proportion of samples misrecognized or rejected. In the case of random sampling, the estimates are binomially distributed, and accurate confidence intervals can be obtained. In the case of selective sampling, tighter estimates are obtained which are approximately binomially distributed

TABLE I — 95 PER CENT CONFIDENCE INTERVALS FOR SOME PUBLISHED RESULTS

Author	Pattern Classes	Measured Characteristics	Recognition Criteria	Sample Size	Error	95% Confidence Interva
Baran, Estrin[7]	Machine Printed Numbers	Presence of ink in elements of 30 x 32 matrix	Maximize *a posteriori* probability (Bayes' Equation)	480	9%	7%–12%
Bledsoe, Browning[8]	Hand-Printed Alpha-Numerics	Presence of mark in elements of 10 x 15 matrix	Matching 2-tuples of matrix elements against table	180	21.6%	13%–29%
Bomba[9]	Hand-Printed Alpha-Numerics	Topological features (orientation of straight lines, intersections, etc.)	Decision tree	112	0%	0%–4%
Doyle[10]	Hand-Printed AEILMNORST	Simply measured topological features	Maximize *a posteriori* probability (Bayes' Equation)	~450	12.5%	10%–16%
Frishkopf[11]	Handwritten words	Extremes, and interconnections between extremes	Cross-correlation against dictionary	160	68%	57%–77%
Harmon[11]	Unsegmented Handwritten Letters	Topological features (cusps, closures, special marks, etc.)	Decision tree	412	41.1%	37%–46%
Mathews, Denes[12]	Spoken digits	Frequency vs time spectra	Cross-correlation against previous averaged spectra from same speaker	99	6%	2%–12%
Marill, Green[13]	Handwritten A,B,C (done as example only)	Distance of character from field edge along eight different line segments	Likelihood function assuming normal distribution of measures	90	3%	1%–10%
Sebestyen[14]	Spoken digits	Frequency vs time spectra	Minimization of non-Euclidean distance measure to average spectra	20	0%	0%–18%

for small error rates. Conservative confidence limits may then be obtained for these estimates.

Using these notions, the experimenter can now determine the sample size required to obtain results which he deems significant. Alternatively, if he has a limited sample size, he can determine the significance of his results. Note that in both cases considered, the variance is inversely proportional to the sample size. This does not mean that the confidence interval is inversely proportional to the square root of the sample size, however, since a binomial rather than a normal distribution pertains. However, perusal of Fig. 1 seems to indicate that this is a good rule of thumb. Note also that the total number of samples required to obtain a certain confidence in the results seems to be independent of the number of allowable pattern classes. This is an interesting philosophical point to ponder.

PART II — ON DESIGN

INTRODUCTION

Part I of this paper was concerned with the estimation of the performance of a given pattern recognition machine. There it was shown how confidence intervals could be found for these estimates. These results are nonparametric in that they hold for any categorization machine (or procedure) regardless of its structure.

We now consider the following problem. An experimenter desires to solve a particular pattern recognition problem. He has at his disposal a set of different methods for solving this problem, but it is not clear to him which is the best to use. Consequently, he desires to estimate the performance of each method when applied to this problem, and choose the best. Let us assume that each method is characterized by certain key parameters which, when known, completely determine the recognition machine. To evaluate any particular recognition method, the experimenter plans to design the corresponding machine by estimating its parameters on the basis of one sampling from the real world of patterns, and then to test this machine based on another sampling (either by constructing the machine or by simulating it).

However, in many practical applications, the total sample size available to the experimenter for design and test purposes is limited. For instance, he may be interested in building a machine to read hand-printed numbers, but he may not have an automatic scanner available to him. Since simulating a scanner by hand is very tedious, he may not be willing to scan more than a certain number of samples.

Or, he may be interested in distinguishing between radar returns caused by missiles and those caused by decoys. Since it is expensive to actually run the sort of experiment required to gather data for this problem, budget limitations will certainly place a limit on the number of available samples.

Another example is in the field of automatic diagnosis of diseases. The experimenter may, for instance, be interested in building a machine which would determine the presence of cancer based on a list of symptoms. However, records have been maintained for only a certain number of people who have contracted this disease, and the sample size is thus definitely limited.

The following problem then arises. If the total sample size is fixed, what is the optimum partitioning of this sample between the design and test phases? This is a rather loose, but concise, statement of the problem. A more accurate one follows.

Assume that the experimenter is concerned with the study of a particular pattern recognition method as applied to some particular problem. The optimum pattern recognition machine based upon this method would have an error probability e_o. The experimenter is interested in estimating e_o so that he can decide whether the particular method under study is adequate for the solution of his problem, or alternately whether it is better than another method. To do this, he takes a sample of a certain size t from the real-life world of patterns. He desires to use part of this sample to design a machine according to the particular method under study. The machine which he thus designs will have an actual error probability $e \geqq e_o$ (both quantities are unknown to the experimenter). He then uses the remaining part of his original sample to test the machine (according to the procedures of Part I). He thus obtains an estimate of e, which will be denoted by $\hat{e}$. It will be shown that $\hat{e}$ is a biased estimate of e_o, and that the bias can be computed. Consequently $\hat{e}$ can be adjusted so that it gives an unbiased estimate, $\hat{e}_o$, of e_o. The optimum partitioning of the total sample will be defined as that partitioning which minimizes the variance of $\hat{e}_o$. Thus, if the experimenter follows this procedure, he will obtain an unbiased minimum variance estimate of e_o, the optimum error probability. Of course, if he finally decides that a particular method is applicable, he can then redesign the corresponding machine with the entire sample size.

OPTIMUM SAMPLE PARTITIONING

We are interested, then, in minimizing the quantity

$$\sigma_{\hat{e}_o}^{\ 2} = E[\hat{e}_o - e_o)^2] = E[\hat{e}_o^{\ 2}] - e_o^{\ 2}, \qquad (18)$$

where $E[x]$ and σ_x^2 denote the expected value and variance of x, respectively.

Let us first digress and consider the biased estimate $\hat{e}$. Since $\hat{e}$ is discrete (it is the proportion of test samples misrecognized), its expected value can be written

$$E[\hat{e}] = \sum \hat{e}p(\hat{e}),$$

where the summation is over all values of $\hat{e}$, and $p(x)$ denotes the probability of x. But

$$p(\hat{e}) = \int p(\hat{e} \mid e)p(e)\, de,$$

where $p(\hat{e} \mid e)$ is the probability of $\hat{e}$ given e, and the integral is over all (continuous) values of e (by definition $e_o \leqq e \leqq 1$). Hence

$$E[\hat{e}] = \sum \hat{e} \int p(\hat{e} \mid e)p(e)de = \int \left[\sum \hat{e}p(\hat{e} \mid e)\right] p(e)de.$$

Let us henceforth consider only the case of random sampling. Then $\hat{e}$ is proportional to a binomially distributed variable ($n\hat{e}$) with parameter e. Therefore the term in brackets, which is the expected value of $\hat{e}$ given the parameter e, is just e. Then

$$E[\hat{e}] = \int ep(e)de = E[e]. \tag{19}$$

$E[e]$ is a function only of the parameters of the problem and the design sample size; it is not a random variable.

We next determine $E[\hat{e}^2]$. By going through a process analogous to the above, and by making use of (19), we obtain

$$\sigma_{\hat{e}}^2 = E[(\hat{e} - E[e])^2] = E[\hat{e}^2] - (E[e])^2 = \frac{E[e(1 - e)]}{n},$$

where n is the size of the test sample. Hence

$$E[\hat{e}^2] = \frac{E[e(1 - e)]}{n} + (E[e])^2. \tag{20}$$

We now determine $E[e]$. Let the optimum machine be described by c different parameters δ_{oi}, $1 \leqq i \leqq c$. The design of the machine consists of estimating the parameters δ_{oi} by making measurements on a set of sample patterns (the design sample). Let the estimates of these parameters be denoted δ_i, $1 \leqq i \leqq c$. Then the error probability e

of the resulting machine is a function of the estimates of the true parameters:

$$e = e(\delta_1, \delta_2, \cdots, \delta_c).$$

One can now expand e in a Taylor series expansion about its minimum point, e_o. Since this is a minimum point, all the coefficients of the linear terms will be zero. If the error deviation $(e - e_o)$ is small, terms above the second order term may be neglected:

$$e \approx e_o + \tfrac{1}{2} \sum_{i=1}^{c} \sum_{j=1}^{c} \left. \frac{\partial^2 e}{\partial \delta_i \partial \delta_j} \right|_{\delta_o} (\delta_i - \delta_{oi})(\delta_j - \delta_{oj}).$$

The expected value of the error for the resulting machine is then

$$E[e] = e_o + \tfrac{1}{2} \sum_{i=1}^{c} \sum_{j=1}^{c} \left. \frac{\partial^2 e}{\partial \delta_i \partial \delta_j} \right|_{\delta_o} E[(\delta_i \, \delta_{oi})(\delta_j - \delta_{oj})].$$

If it is assumed that the estimates are unbiased, i.e., $E(\delta_i) = \delta_{oi}$, then the above equation may be written as

$$E[e] = e_o + \tfrac{1}{2} \sum_{i=1}^{c} \sum_{j=1}^{c} a_{ij} \sigma_{ij} \tag{21}$$

where

$$a_{ij} = a_{ji} = \left. \frac{\partial^2 e}{\partial \delta_i \partial \delta_j} \right|_{\delta_o},$$

σ_{ij} is the covariance of the estimates for δ_{oi} and δ_{oj}, and $\sigma_{ii} = \sigma_i^2$ is the variance of the estimate for δ_{oi}. (21) is valid for small values of the quantity $(e - e_o)$.

It may be worth-while to digress here to a simple example which may help to clarify the definitions of the above terms. Zachary Oglethorpe is not only a crafty fisherman, but is also a good gadgeteer. He has decided to try to build equipment which will determine each day whether he should use a surface bait or a deep water bait in order to catch the maximum number of fish. He has means available to measure the water temperature, the magnitude of surface ripple, and the atmospheric pressure, and therefore decides to use these as his measurements. He denotes values of these measurements by m_1, m_2, and m_3 respectively.

Mr. Oglethorpe has been recording values of these measurements every day for the past six months, and has noted on each day whether he was more successful with surface or deep water bait. He thus has a total sample size of roughly 180 samples, some from one pattern class (surface bait), and some from the other pattern class (deep water bait). Since each sample was taken without *a priori* knowledge of the class to which it belonged, then this constitutes random sampling; that is, the proportion of samples in each class is an estimate of the *a priori* probability of occurrence of that class.

Our crafty fisherman decides to build a decision making, or pattern recognition, machine by building a correlator for each of the two possible decisions (or pattern classes). That is, the machine will make the following two calculations:

$$\text{Surface bait} = \delta_1 m_1 + \delta_2 m_2 + \delta_3 m_3 ,$$

$$\text{Deep water bait} = \delta_4 m_1 + \delta_5 m_2 + \delta_6 m_3 .$$

The class achieving the highest value represents the desired decision. Let us assume that, according to some theory, the optimum values of the δ_i are the means for each measurement within the appropriate pattern class, normalized so that the sum of the squares of the coefficients of each linear form is unity. That is, δ_1 is proportional to the mean water temperature when surface bait should be used, and so forth, and is normalized with δ_2 and δ_3 so that $\delta_1^2 + \delta_2^2 + \delta_3^2 = 1$.

Thus the parameters δ_i completely characterize this pattern recognition machine in that, given values for each δ_i, $1 \leqq i \leqq 6$, the machine may be built. The optimum values for each δ_i are the appropriate normalized means, which are the δ_{oi} of the previous equations. Mr. Oglethorpe obtains estimates of these optimum parameters by taking normalized averages over a portion of the appropriate data. These estimates are the δ_i of the previous equations, and are the actual numbers on which he would base the construction of his machine. Note that, in this case, these estimates are unbiased and efficient, and may very well be independent of each other (e.g., the probability distribution of the water temperature when surface bait should be used may be independent of the values of surface ripple magnitude and atmospheric pressure).

Having thus designed his fisherman's aid with a portion of his data, he now tests it with the remainder of the data to determine its accuracy. He does not want to use it if there is a good probability that it is less accurate than he has found his own intuition to be. This then leads us to the basic problem being studied: How should Zachary Oglethorpe split his total sample between the design and the testing of his machine to obtain the best estimate of the accuracy of the machine? Again, if the estimated accuracy of his machine were sufficient, he would then be wise to redesign it, basing the new design on the entire sample.

We now return to the study of this sample partitioning. Let each parameter be estimated with m samples.* If each of these estimates is an efficient and unbiased estimate, and if the estimates are independent (either because the estimates are statistically independent, or because different samples are used to estimate each), then all $\sigma_{ij} = 0$, $i \neq j$, and all σ_i^2 will be proportional to $1/m$. Hence one can rewrite (21) as

$$E[e] = e_o + \frac{b}{m}, \tag{22}$$

where b is some constant calculated from (21). (Often, $E[e]$ is in the form (22) even if the estimates are not independent.)

Let t be the total sample size, and p be the number of sets of m samples used to design the machine. p is chosen to be the smallest number which insures that $E[e]$ is of the form (22). It is often simply the number of allowable pattern classes, since, of course, parameters of different classes must be estimated with different samples. If n is the test sample size, then

$$t = n + pm. \tag{23}$$

From (19) and (22),

$$E[\hat{e}] = E[e] = e_o + \frac{b}{m}. \tag{24}$$

Consequently, $\hat{e}$ is a biased estimate of e_o. The adjusted estimate $\hat{e}_o$, given by

$$\hat{e}_o = \hat{e} - \frac{b}{m}, \tag{25}$$

is an unbiased estimate of e_o, with variance given by (18). This variance can now be rewritten using (25):

$$\begin{aligned}\sigma_{\hat{e}_o}^2 = E\,[\hat{e}_o^2] - \epsilon_o^2 &= E\left[\left(\hat{e} - \frac{b}{m}\right)^2\right] - e_o^2 \\ &= E[\hat{e}^2] - 2\,\frac{b}{m}\,E[\hat{e}] + \left(\frac{b}{m}\right)^2 - e_o^2.\end{aligned}$$

* This is not always desirable, since some parameters may be easier to estimate than others, or there may be more data available for some parameters than others. However, this condition is assumed here for simplicity, as are the following assumptions of efficiency, unbiasedness, and independence.

From (20) and (24),

$$\sigma_{\hat{e}_o}^{\;2} = \frac{E[e(1-e)]}{n} + (E[e])^2 - 2\frac{b}{m}e_o - \left(\frac{b}{m}\right)^2 - e_o^{\;2}.$$

$$= \frac{E[e(1-e)]}{n} + (E[e])^2 - \left(e_o + \frac{b}{m}\right)^2.$$

Thus, from (24),

$$\sigma_{\hat{e}_o}^{\;2} = \frac{E[e(1-e)]}{n}. \tag{26}$$

If $b/m \ll 1$ (which will certainly be true for any reasonable design), then

$$\sigma_{\hat{e}_o}^{\;2} \approx \frac{E[e(1-e_o)]}{n} = (1-e_o)\frac{e_o + \dfrac{b}{m}}{n} = (1-e_o)\frac{e_o + \dfrac{pb}{t-n}}{n}, \tag{27}$$

where the relation (23) was used.

We wish to choose n such that (27) is minimized. Differentiating (27) and equating to zero, one obtains

$$\frac{e_o t}{pb} = \frac{2\dfrac{n_o}{t} - 1}{\left(1 - \dfrac{n_o}{t}\right)^2}, \tag{28}$$

where n_o is that value of n satisfying (28); it is the optimum test sample size in the sense previously discussed. n_o/t is of course the proportion of the total sample used for the test. One interesting result is immediately obvious: n_o/t must be greater than 0.5 for all cases. The equation (28) is plotted in Fig. 2, from which the following general statements can be made.

1. The proportion of the total sample that should be used to test the machine should never be less than 50 per cent.
2. If $e_o t/pb < 0.1$, then the proportion used for design should be about 50 per cent.
3. The proportion of the total sample that should be used to test the machine becomes larger as:
   a. The total sample size increases,
   b. the error of the optimum machine increases,
   c. the effectiveness of the design increases (pb decreases).

Here $1/pb$ is taken as a measure of the effectiveness of the design,

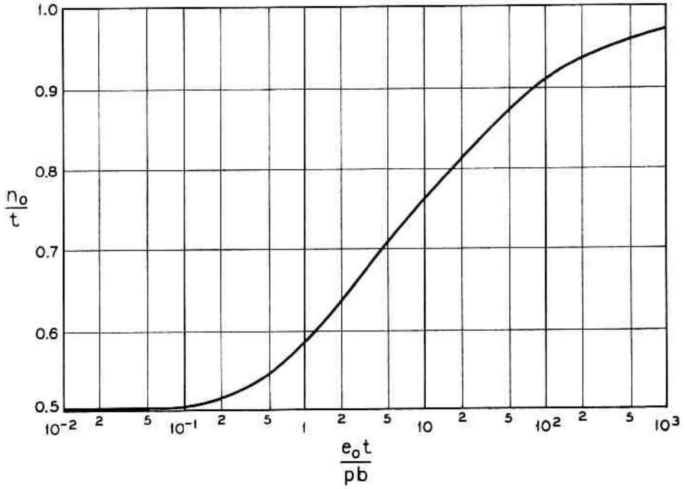


Fig. 2 — Optimum sample partitioning.

since pb is the product of the expected deviation from optimum, $E[e - e_o]$, and the design sample size, pm.

These results indicate just how a sample should be split between the design and test stages of a feasibility study of a pattern recognition method. If the experimenter follows this procedure, he will obtain an estimate $\hat{e}_o$ of e_o which is unbiased and has minimum variance.

The value of this minimum variance can be expressed as

$$\sigma^2_{\hat{e}_{o_{\min}}} = \frac{e_o(1 - e_o)}{n}\left(1 + \frac{1 - \frac{n_o}{t}}{2\frac{n_o}{t} - 1}\right),$$

which was obtained by eliminating pb between (27) and (28). Note that this is the variance that would have been obtained if the optimum machine were tested with n samples, increased by a factor which accounts for the design error.

AN EXAMPLE OF OPTIMUM SAMPLE PARTITIONING

As an illustration of these ideas, consider the following example (perhaps the simplest of the n-dimensional problems). A pattern recognition machine is to be designed using the optimum decision function[15,16] which will distinguish between q classes. The occurrence of each class is equally probable *a priori*, and all costs of misrecognition are the same. The receptor makes a set of k measurements m_j, $1 \leqq j \leqq k$, on each input pattern. It is known that each measurement is normally distributed with variance σ, and that all measurements are independent. Further, it is known that the distances between the mean vectors in measurement space* are all equal. (Consequently, there can be no more than $k + 1$ pattern classes. The tips of the mean vectors are the vertices of a regular polytope.)

It can then be shown that the optimum decision function partitions the measurement space into polytopes which are bounded by those hyperplanes which are the perpendicular bisectors of the line segments joining all pairs of means. The hyperplane separating two classes, say classes 1 and 2, is the set of all points $(x_1, \cdots, x_k)$, represented by the vector $\bar{X}$, which satisfy

$$\bar{x} \cdot (\bar{\mu}_1 - \bar{\mu}_2) = \tfrac{1}{2}(\bar{\mu}_1 \cdot \bar{\mu}_1 - \bar{\mu}_2 \cdot \bar{\mu}_2), \tag{29}$$

where $\bar{\mu}_i$ is the mean vector of class i.[13]

The design procedure consists of estimating each mean vector from a sampling; denote the estimated mean vector for class i by $\bar{x}_i$. The distribution of the estimate of a mean vector from a normal distribution with covariance matrix $[V]$ is also normal with covariance matrix $1/m\ [V]$, where m is the sample size used in the estimate.[17] Since the measurements are independent in this case, then so will be the estimates of the means of the various measurements. Furthermore, each estimate will have a variance of σ^2/m. Consequently, only one set of samples of size m from each pattern class is required to insure that the form (22) is valid, and p is hence equal to the number of allowable pattern classes, q.

It is shown in the Appendix that b is given by

$$b = \frac{q(q-1)}{4} \frac{\Delta\mu}{2\sigma} N\left(\frac{\Delta\mu}{2\sigma}\right),$$

where $\Delta\mu$ is the distance between any pair of mean vectors, and $N(\Delta\mu/2\sigma)$ is the value of the standard normal density function for the variable $\Delta\mu/2\sigma$. The equation (28) then becomes

$$\frac{4e_o t}{q^2(q-1)\frac{\Delta\mu}{2\sigma} N\left(\frac{\Delta\mu}{2\sigma}\right)} = \frac{2\frac{n_o}{t} - 1}{\left(1 - \frac{n_o}{t}\right)^2}. \tag{30}$$

* A geometric interpretation of categorization problems is often useful. By measurement space, we mean a k-dimensional space in which each coordinate represents one of the k receptor measurements. Thus any set of measurements which have been made on an input pattern may be represented as a point in measurement space. The decision function may be thought of as partitioning the measurement space into regions corresponding to the different allowable pattern classes and into rejection regions.

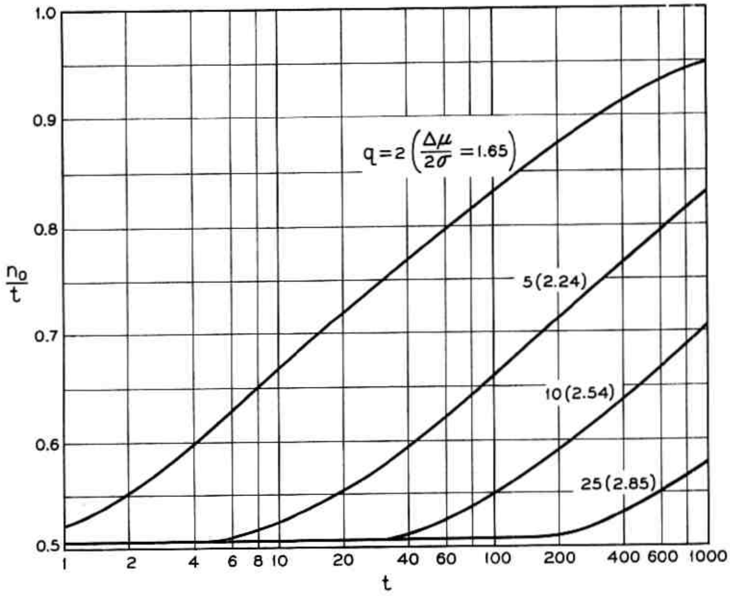


Fig. 3 — Optimum sample partitioning for symmetric Gaussian case.

Some curves representing (30) are plotted in Fig. 3 in which the proportion of the total sample to be used in the test, n_o/t, is shown as a function of t, the total sample size, with the number of allowable pattern classes, q, as a parameter. e_o was held constant at 0.05 (which involves the choosing of the proper value of $\Delta\mu/2\sigma$ for each q).

From Fig. 3 it is seen that, for many cases, the sample should be split evenly between design and test, as one might intuitively suspect. However, there are some drastic deviations from this. For instance, if the categorizer is to separate only two classes, and 1000 samples are available, then only 50 of these should be used to design the machine, and 950 should be used to test it. Consequently, it is seen that intuition may go wrong in some cases.

CONCLUSION

This paper has begun an analysis of some of the problems which arise in the design and analysis of pattern recognition experiments. In Part II, the problem of optimum sample partitioning between the design and test phases of a pattern recognition machine was investigated for the case of a fixed total sample size and no overlap between the design and test samples. The general relation between the optimum partitioning and the total sample size, optimum error rate, and design efficiency was derived. From this, it was apparent that the test sample size should never be smaller than the design sample size. These results are nonparametric in the sense that they do not depend on the detailed structure of the recognition machine. It is only necessary that the deviation of the designed machine from the optimum machine be small, and that the design of the machine be done in such a way that (22) holds.

However, the actual computation of the optimum sample partitioning does depend strongly on the detailed structure of the machine through the quantity b. Since this computation is quite difficult even in the simplest of cases, the interesting question arises as to the possibility of estimating b from the sample. Another interesting phase of this problem which has not been attacked here concerns the case when the design sample and test sample overlap — that is, some of the sample patterns from the design sample are also used in the test sample. In the limit, this reduces to using the total sample for both design and test purposes. In this case, the results of the test are usually not very reliable. Consequently, there may be some sample partitioning with overlap which is better (in the sense discussed in this paper) than for either the case of no overlap or the case of total overlap.

ACKNOWLEDGMENT

I would like to thank Prof. A. E. Laemnel of the Polytechnic Institute of Brooklyn for suggesting this problem, and Messrs. E. Wolman, W. H. Williams, and E. N. Pinson for their very helpful discussions concerning these topics.

APPENDIX

We determine here the coefficient b in (22) for the example discussed in this paper. If the mean vectors are more than about 3σ apart, then only a small error is made if the total error is approximated by adding the errors of each hyperplane taken alone. That is, the integrals on the wrong side of the hyperplane that are counted more than once will be quite small compared to the integrals counted only once.

Due to the symmetry of the problem, the error associated with each hyperplane for the optimum decision function is identical, and the derivatives of (21) will also be identical for each hyperplane. Since there are

$q(q-1)/2$ hyperplanes, b may be expressed (from (21) and (22)) as

$$\frac{b}{m} = \frac{q(q-1)}{2} \frac{1}{2} \sum_{i=1}^{k} \left[\frac{\partial^2 e_{12}}{\partial \bar{x}_{i1}^2} \bigg|_{\mu_1,\mu_2} + \frac{\partial^2 e_{12}}{\partial \bar{x}_{i2}^2} \bigg|_{\mu_1,\mu_2} \right] \frac{\sigma^2}{m}, \tag{31}$$

where the hyperplane separating classes 1 and 2 is taken as typical, and the independence of the estimates is used. e_{12} is the error associated with this hyperplane, μ_1 and μ_2 are the mean vectors of these classes, and $\bar{x}_1$ and $\bar{x}_2$ are the estimates of the mean vectors.

There is no loss in generality if μ_1 is taken as zero, and all the components of μ_2 $(\mu_{12}, \cdots, \mu_{k2})$ are taken as zero except for μ_{12}. That is,

$$\mu_1 = (0,0, \cdots ,0)$$

$$\mu_2 = (\mu,0, \cdots ,0),$$

where μ_{12} is denoted μ, $\mu > 0$. Consequently, the optimum boundary is given by

$$x_1 = \mu/2.$$

A sampling of size m is taken from each class, and the mean vectors are estimated, giving

$$\bar{x}_1 = (\bar{x}_{11}, \bar{x}_{21}, \cdots, \bar{x}_{k1})$$

$$\bar{x}_2 = (\bar{x}_{12}, \bar{x}_{22}, \cdots, \bar{x}_{k2}).$$

A boundary given by (29) is computed based on the above estimates, and this, together with the other estimated boundaries, determines the structure of the machine.

The error e_1 associated with this particular boundary for class 1 is

$$e_1 = \prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2 dx_j$$

$$\cdot \int_{\xi_1(x_2,\ldots x_{1k})}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2}\left(\frac{x_1}{\sigma}\right)^2 dx_1,$$

where $\xi_1(x_2, \cdots, x_k)$ is the value of x_1 on the boundary, and is given by (from (29))

$$\xi_1(x_2, \cdots, x_n) = -\sum_{i=2}^{k} \frac{\bar{x}_{i1} - \bar{x}_{i2}}{\bar{x}_{11} - \bar{x}_{12}} x_i + \frac{1}{2} \sum_{i=1}^{k} \frac{(\bar{x}_{i1}^2 - \bar{x}_{i2}^2)}{\bar{x}_{11} - \bar{x}_{12}}$$

$$= \frac{\bar{x}_{11} + \bar{x}_{12}}{2} - \frac{1}{2} \sum_{i=2}^{k} \frac{2(\bar{x}_{i1} - \bar{x}_{i2})x_i - (\bar{x}_{i1}^2 - \bar{x}_{i2}^2)}{x_{11} - x_{12}}.$$

Then

$$\frac{\partial e_1}{\partial \bar{x}_{i1}} = \prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp -\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2 dx_j$$

$$\cdot\left(\frac{1}{\sqrt{2\pi\sigma}} \exp -\frac{1}{2}\left(\frac{\xi_1}{\sigma}\right)^2\right)\left(\frac{x_i - \bar{x}_{i1}}{\bar{x}_{11} - \bar{x}_{12}}\right), \quad 2 \leqq i \leqq n.$$

$$\frac{\partial^2 e_1}{\partial \bar{x}_{i1}{}^2} = \prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp -\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2 dx_j \left(\frac{1}{\sqrt{2\pi\sigma}} \exp -\frac{1}{2}\left(\frac{\xi_1}{\sigma}\right)^2\right)$$

$$\cdot\left[\frac{\xi_1}{\sigma^2}\left(\frac{x_i - \bar{x}_{i1}}{\bar{x}_{11} - \bar{x}_{12}}\right)^2 - \left(\frac{1}{\bar{x}_{11} - \bar{x}_{12}}\right)\right], \quad 2 \leqq i \leqq n.$$

$$\left.\frac{\partial^2 e_1}{\partial \bar{x}_{i1}{}^2}\right|_{\mu_1,\mu_2} = \left(\frac{1}{\sqrt{2\pi\sigma}} \exp -\frac{1}{2}\left(\frac{\mu/2}{\sigma}\right)^2\right)$$

$$\cdot\prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp -\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2 dx_j \left[-\frac{x_i}{2\sigma^2} + \frac{1}{\mu}\right]$$

$$= \frac{1}{\sigma} N\left(\frac{\mu}{2\sigma}\right)\left[-\frac{1}{2\sigma^2} E[x_i] + \frac{1}{\mu}\right]$$

$$= \frac{1}{\mu\sigma} N\left(\frac{\mu}{2\sigma}\right), \qquad 2 \leqq i \leqq n,$$

where $N(\mu/2\sigma)$ is the value of the standard normal density function for the variate $\mu/2\sigma$. In a like manner,

$$\left.\frac{\partial^2 e_2}{\partial \bar{x}_{i1}{}^2}\right|_{\mu_1,\mu_2} = -\frac{1}{\mu\sigma} N\left(-\frac{\mu}{2\sigma}\right) = -\frac{1}{\mu\sigma} N\left(\frac{\mu}{2\sigma}\right), \qquad 2 \leqq i \leqq n,$$

where e_2 is the error associated with this boundary for class 2. Since the total error for this boundary is $e_{12} = e_1 + e_2$, then

$$\left.\frac{\partial^2 e_{12}}{\partial \bar{x}_{i1}{}^2}\right|_{\mu_1,\mu_2} = \left.\frac{\partial^2 e_1}{\partial \bar{x}_{i1}{}^2}\right|_{\mu_1,\mu_2} + \left.\frac{\partial^2 e_2}{\partial \bar{x}_{i1}{}^2}\right|_{\mu_1,\mu_2} = 0, \qquad 2 \leqq i \leqq n.$$

A like result holds for $\dfrac{\partial^2 e_{12}}{\partial \bar{x}_{i2}{}^2}$, $2 \leqq i \leqq n$. Going through this same procedure for $\bar{x}_{11}$,

$$\frac{\partial e_1}{\partial \bar{x}_{11}} = -\prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp -\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2 dx_j \left[\frac{1}{2\sigma} N\left(\frac{\xi_1}{\sigma}\right)\right].$$

$$\frac{\partial^2 e_1}{\partial \bar{x}_{11}{}^2} = -\prod_{j=2}^{k} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp -\frac{1}{2}\left(\frac{x_j}{\sigma}\right)^2 dx_j \left[-\frac{1}{4\sigma}\left(\frac{\xi_1}{\sigma^2}\right) N\left(\frac{\xi_1}{\sigma}\right)\right].$$

$$\left.\frac{\partial^2 e_1}{\partial \bar{x}_{11}{}^2}\right|_{\mu_1,\mu_2} = \frac{1}{8}\frac{\mu}{\sigma^3} N\left(\frac{\mu}{2\sigma}\right).$$

Similarly,

$$\left.\frac{\partial^2 e_2}{\partial \bar{x}_{11}^2}\right|_{\mu_1 \cdot \mu_2} = \frac{1}{8}\frac{\mu}{\sigma^3} N\left(\frac{\mu}{2\sigma}\right).$$

Hence

$$\left.\frac{\partial^2 e_{12}}{\partial \bar{x}_{11}^2}\right|_{\mu_1 , \mu_2} = \frac{1}{4}\frac{\mu}{\sigma^3} N\left(\frac{\mu}{2\sigma}\right).$$

It would also be found that

$$\left.\frac{\partial^2 e_{12}}{\partial \bar{x}_{12}^2}\right|_{\mu_1 , \mu_2} = \frac{1}{4}\frac{\mu}{\sigma^3}\ N\left(\frac{\mu}{2\sigma}\right).$$

This analysis is perfectly general for arbitrary mean vectors, providing that μ is merely interpreted as the distance between a pair of mean vectors (all such distances being assumed identical). This distance will henceforth be written $\Delta\mu$ to indicate that it is a difference of means. Therefore, from (31), we find that

$$b = \frac{q(q-1)}{4}\frac{\Delta\mu}{2\sigma}\ N\left(\frac{\Delta\mu}{2\sigma}\right).$$

REFERENCES

1. Fraser, D. A. S., *Statistics: An Introduction*, John Wiley and Sons, Inc., New York, 1960.
2. Mattson, R. L., Master's Thesis, E. E. Dept., M.I.T., May, 1959.
3. Clopper, C. S., and Pearson, E. S., Biometrika, **26,** 1934, p. 404.
4. Pearson, E. J., and Hartley, H. O., *Biometrika Tables for Statisticians*, The University Press, Cambridge, 1954, p. 204.
5. Bowley, A. L., Bull. Int. Stat. Inst., **22,** 1926, p. 1.
6. Neyman, J., J. Royal Stat. Soc., **97,** Pt. 4, 1934, p. 558.
7. Baran, P., and Estrin, G., I.R.E. Wescon Record, Pt. 4, 1960, p. 29.
8. Bledsoe, W. W., and Browning, I., Proc. E.J.C.C., Dec., 1959, p. 225.
9. Bomba, J. S., Proc. E.J.C.C., Dec., 1959, p. 218.
10. Doyle, W., Proc. W.J.C.C., 1960, p. 133.
11. Frishkopf, L. S., and Harmon, L. D., Proc. 4th London Symposium on Information Theory, 1960.
12. Mathews, M. V., and Denes, P., J. Acous. Soc. Amer., **32,** Nov., 1960, p. 1450.
13. Marill, T., and Green, D. M., I.R.E. Trans. on Elec. Comp., **EC-9**, Dec., 1960, p. 472.
14. Sebestyen, G. S., I.R.E. Trans. on Information Theory, **IT-7**, Jan., 1961, p. 44.
15. Middleton, D., and Van Meter, D., J. Soc. Ind. and App. Math., **3,** Sept., 1955, p. 192; **4,** June, 1956, p. 86.
16. Chow, C. K., I.R.E. Trans. on Elec. Comp., **EC-6**, Dec., 1957, p. 247.
17. Anderson, T. W., *An Introduction to Multivariate Statistics*, John Wiley and Sons, New York, 1958.

Lined Waveguide*

By H. G. UNGER

(Manuscript received October 19, 1961)

The existing approximate analysis of wave propagation in lined waveguide is, under practical conditions, limited to linings thinner than 0.025 per cent of the waveguide diameter. A more exact analysis is presented here for the straight and curved waveguide and for all practical linings. In the case of anisotropic or sandwiched linings, the boundary value problem is formulated using wall impedances. The single isotropic lining is taken as an example to prove this formulation useful for typical cases.

The exact analysis shows that neither the thickness nor the permittivity of the lining can increase the phase difference between TM_{11} and TE_{01} beyond a certain limit. The curvature coupling between these two waves is enhanced slightly by the lining.

I. INTRODUCTION

Round waveguide with dielectric lining shows promise as a communication medium.[1] The circular electric wave loss in perfectly straight and round metallic waveguide decreases steadily with frequency. Any deformations of the cross section or curvature of the guide axis degrade these ideal transmission characteristics.

The TM_{1n} waves in particular are coupled by curvature to TE_{0n}, and since they propagate with nearly equal phase velocity, there will be large mode conversion. A dielectric lining close to the wall changes the TM_{1n} waves appreciably with almost no change to TE_{0n}. The phase velocities are now different and, despite curvature coupling, mode conversion stays small.

When the lining is made lossy it will serve still another purpose.[2] Circular electric wave loss is increased only very little, while all other waves suffer an effective dielectric loss. This mode filtering loss will reduce the degrading effects of mode conversion and reconversion.

* This work was performed under a Letter Contract with Bell Telephone Laboratories at the Institut fuer Hoechstfrequenztechnik, Technische Hochschule Braunschweig.

To serve both purposes best, the lining should be made anisotropic or sandwiched from different materials.[3] The circular electric wave loss will then remain very low, yet all other waves will suffer high loss and also curvature loss will stay small.

Wave propagation in straight and curved lined waveguide has been analyzed elsewhere.[1,4] Likewise, imperfections in the lining and cross-sectional deformations have been calculated.[4,5] However, the lining was always assumed to be very thin, and so far only a first-order approximation has been found.

On the other hand, it has been shown both theoretically and experimentally that these approximations do not in general hold for any practical linings.[1] Linings which are designed optimally change wave propagation much more than could be described by a first-order approximation.

An analysis of wave propagation in lined waveguide will be presented here which is sufficiently general and accurate to hold for all practical cases. Sandwiched and anisotropic linings will also be considered.

Circular electric wave transmission is most strongly degraded in curved waveguide. Therefore, the lined waveguide will be assumed to have curvature. Cross-sectional deformations and imperfections of the lining will be analyzed with corresponding accuracy in another paper.[6]

II. NORMAL MODE FIELDS

Normal modes of straight round waveguide with a single isotropic lining have been analyzed before.[1] To adapt this analysis to an investi-

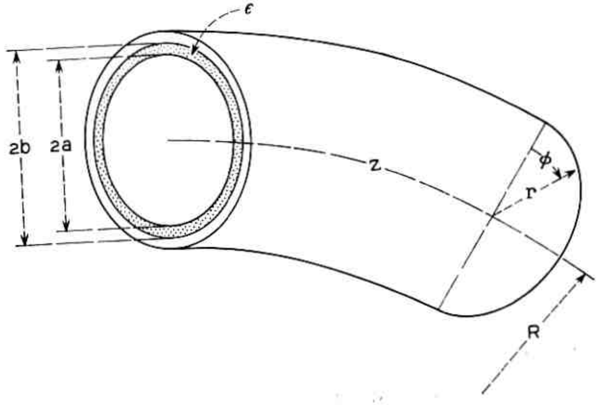


Fig. 1 — Lined waveguide with curvature.

gation of bends in lined waveguide and of waveguides with an irregular lining, the boundary value problem will be repeated and generalized here.

The waveguide structure to be considered is shown in Fig. 1. The dielectric lining will later on be assumed to be heterogeneous or anisotropic or irregular. At present, however, a single uniform and homogeneous lining is assumed.

The electromagnetic field in the waveguide can be derived from two sets of scalar functions T_n and T_n' given by:

$$\left.\begin{aligned} T_n &= N_n\, J_p(\chi_n r)\sin p\varphi \\ T_n' &= N_n\, J_p(\chi_n r)\cos p\varphi \end{aligned}\right\} \quad \text{for} \quad 0 < r < a \tag{1}$$

and

$$T_n = N_n \frac{\chi_n^{\,2}}{\chi_n^{e^2}} J_p(k_n) \frac{H_p^{(2)}(\chi_n^{\,e} r) - cH_p^{(1)}(\chi_n^{\,e} r)}{H_p^{(2)}(k_n^{\,e}) - cH_p^{(1)}(k_n^{\,e})} \sin p\varphi$$

$$\text{for} \quad a < r < b \tag{2}$$

$$T_n' = N_n \frac{\chi_n^{\,2}}{\chi_n^{e^2}} J_p(k_n) \frac{H_p^{(2)}(\chi_n^{\,e} r) - c'H_p^{(1)}(\chi_n^{\,e} r)}{H_p^{(2)}(k_n^{\,e}) - c'H_p^{(1)}(k_n^{\,e})} \cos p\varphi.$$

The T functions satisfy the wave equation:

$$\nabla^2 T = \frac{1}{e_1 e_2}\left[\frac{\partial}{\partial u}\left(\frac{e_2}{e_1}\frac{\partial T}{\partial u}\right) + \frac{\partial}{\partial v}\left(\frac{e_1}{e_2}\frac{\partial T}{\partial v}\right)\right] = -\chi^2 T \tag{3}$$

in a general orthogonal curvilinear coordinate-system (u,v,w), in which the element of length is:

$$ds^2 = e_1^2 du^2 + e_2^2 dv^2 + e_3^2 dw^2. \tag{4}$$

The curved waveguide may be described in these coordinates when, according to Fig. 1:

$$u = r, \qquad v = \varphi, \qquad w = z \tag{5}$$

$$e_1 = 1, \qquad e_2 = r, \qquad e_3 = 1 + \xi \tag{6}$$

where

$$\xi = \frac{r}{R}\cos\varphi. \tag{7}$$

The field components are written in terms of the field functions (1) and (2)

$$\begin{aligned} E_u &= \sum_n V_n \left[\frac{\partial T_n}{e_1 \partial u} + d_n \frac{\partial T_n'}{e_2 \partial v} \right] \\ E_v &= \sum_n V_n \left[\frac{\partial T_n}{e_2 \partial v} - d_n \frac{\partial T_n'}{e_1 \partial u} \right] \\ H_u &= -\sum_n I_n \left[\frac{\partial T_n}{e_2 \partial v} - d_n \frac{h_n^2}{k^2} \frac{\partial T_n'}{e_1 \partial u} \right] \epsilon \\ H_v &= \sum_n I_n \left[\frac{\partial T_n}{e_1 \partial u} + d_n \frac{h_n^2}{k^2} \frac{\partial T_n'}{e_2 \partial v} \right] \epsilon . \end{aligned} \tag{8}$$

Maxwell's equations are:

$$\frac{1}{e_2 e_3} \left[\frac{\partial}{\partial v} (e_3 E_w) - \frac{\partial}{\partial w} (e_2 E_v) \right] = -j\omega\mu_0 H_u \tag{9}$$

$$\frac{1}{e_3 e_1} \left[\frac{\partial}{\partial w} (e_1 E_u) - \frac{\partial}{\partial u} (e_3 E_w) \right] = -j\omega\mu_0 H_v \tag{10}$$

$$\frac{1}{e_1 e_2} \left[\frac{\partial}{\partial u} (e_2 E_v) - \frac{\partial}{\partial v} (e_1 E_u) \right] = -j\omega\mu_0 H_w \tag{11}$$

$$\frac{1}{e_2 e_3} \left[\frac{\partial}{\partial v} (e_3 H_w) - \frac{\partial}{\partial w} (e_2 H_v) \right] = j\omega\epsilon\epsilon_0 E_u \tag{12}$$

$$\frac{1}{e_3 e_1} \left[\frac{\partial}{\partial w} (e_1 H_u) - \frac{\partial}{\partial u} (e_3 H_w) \right] = j\omega\epsilon\epsilon_0 E_v \tag{13}$$

$$\frac{1}{e_1 e_2} \left[\frac{\partial}{\partial u} (e_2 H_v) - \frac{\partial}{\partial v} (e_1 H_u) \right] = j\omega\epsilon\epsilon_0 E_w . \tag{14}$$

μ_0 and ϵ_0 are permeability and permittivity of free space. ϵ is the relative permittivity of the respective cross-sectional part of the waveguide.

Substituting from (8) into (11) and (14) and taking advantage of (3) the longitudinal field components are obtained:

$$\begin{aligned} H_w &= j\omega\epsilon\epsilon_0 \sum_n V_n \, d_n \frac{\chi_n^2}{k^2} T_n' \\ E_w &= j\omega\mu_0 \sum_n I_n \epsilon \frac{\chi_n^2}{k^2} T_n \end{aligned} \tag{15}$$

where $k = \omega\sqrt{\epsilon\epsilon_0\mu_0}$ is the intrinsic propagation constant of the medium in a particular cross-sectional part of the guide. ϵ and k have constant but different values for the different cross-sectional parts of the guide.

In (8), this dependence on the cross-sectional part of the guide is not only true for ϵ and k, but until we learn more, also for d_n.

The quantities d_n, c, c' and the separation constants χ_n and χ_n^e must be chosen so that the boundary conditions of the lined waveguide are satisfied.

These boundary conditions are, at the surface of the lining:

$$E_w^i = E_w^e \tag{16}$$

$$H_w^i = H_w^e \tag{17}$$

$$E_v^i = E_v^e \tag{18}$$

$$H_v^i = H_v^e \tag{19}$$

and at the metal surface

$$E_w = 0 \tag{20}$$

$$E_v = 0. \tag{21}$$

The superscrips i and e indicate the internal and external field components at the surface of the lining.

To satisfy (20):

$$c = \frac{H_p^{(2)}(\rho k_n^e)}{H_p^{(1)}(\rho k_n^e)} \tag{22}$$

where

$$\rho = \frac{b}{a} \tag{23}$$

and

$$\begin{aligned} k_n &= \chi_n a \\ k_n^e &= \chi_n^e a. \end{aligned} \tag{24}$$

To satisfy (21)

$$c' = \frac{H_p^{(2)'}(\rho k_n^e)}{H_p^{(1)'}(\rho k_n^e)}. \tag{25}$$

The prime at the Bessel functions denotes differentiation with respect to the argument.

The condition of E_w being continuous across the surface of the lining is satisfied by virtue of the formulation of the T-functions in (1) and (2). To satisfy (17)

$$d_n^i = d_n^e.$$

d_n is therefore independent of the cross-sectional area of the guide and needs no superscript.

To satisfy (18)

$$\frac{1}{d_n} = \frac{k_n k_n^{e2}}{pk^2a^2(\epsilon - 1)} \left[\frac{J_p'(k_n)}{J_p(k_n)} - \frac{k_n}{k_n^e} \frac{H_p^{(2)'}(k_n^e) - c'H_p^{(1)'}(k_n^e)}{H_p^{(2)}(k_n^e) - c'H_p^{(1)}(k_n^e)} \right]. \tag{26}$$

The remaining condition (19) leads to the following (characteristic) equation of the lined waveguide:

$$\begin{aligned} &\left[\frac{J_p'(k_n)}{J_p(k_n)} - \frac{k_n}{k_n^e} \frac{H_p^{(2)'}(k_n^e) - c'H_p^{(1)'}(k_n^e)}{H_p^{(2)}(k_n^e) - c'H_p^{(1)}(k_n^e)} \right] \\ &\qquad \cdot \left[\frac{J_p'(k_n)}{J_p(k_n)} - \epsilon \frac{k_n}{k_n^e} \frac{H_p^{(2)'}(k_n^e) - cH_p^{(1)'}(k_n^e)}{H_p^{(2)}(k_n^e) - cH_p^{(1)}(k_n^e)} \right] \\ &\qquad\qquad = p^2(\epsilon - 1)^2 \frac{h_n^2 a^2 k^2 a^2}{k_n^2 k_n^{e4}}. \end{aligned} \tag{27}$$

The characteristic equation (27), together with

$$\begin{aligned} k_n^2 &= (\omega^2 \epsilon_0 \mu_0 - h_n^2)a^2 \\ k_n^{e2} &= (\omega^2 \epsilon\epsilon_0 \mu_0 - h_n^2)a^2 \end{aligned} \tag{28}$$

determines the separation constants k_n and k_n^e.

The transverse field components of any two different modes are orthogonal to each other in that[7]

$$\begin{aligned} &\frac{1}{V_n I_m} \int_S (E_{tn} \times H_{tm}) \, dS \\ &\qquad = \int_S \epsilon \left[\left(\frac{\partial T_n}{e_1 \partial u} + d_n \frac{\partial T_n'}{e_2 \partial v} \right) \left(\frac{\partial T_m}{e_1 \partial u} + d_m \frac{h_m^2}{k^2} \frac{\partial T_m'}{e_2 \partial v} \right) \right. \\ &\qquad \left. + \left(\frac{\partial T_n}{e_2 \partial v} - d_n \frac{\partial T_m}{e_1 \partial u} \right) \left(\frac{\partial T_m}{e_2 \partial v} - d_m \frac{h_m^2}{k^2} \frac{\partial T_m'}{e_1 \partial u} \right) \right] dS = \delta_{nm}. \end{aligned} \tag{29}$$

The integration is to be extended over the entire cross section of the guide. The quantity δ_{nm} is the Kronecker delta. To satisfy (29) for $n = m$ requires N_n to have a certain value, which is to be determined from (29).

III. GENERALIZED TELEGRAPHIST'S EQUATIONS FOR CURVED WAVEGUIDE

All quantities in (8) and (15) have now been determined except the current and voltage coefficients I_n and V_n. To find relations for them,

(8) is substituted for the field components into Maxwell's equations. Then

$$-e_3\left(\frac{\partial T_m}{e_2\partial v} - d_m\,\frac{h_m^{\,2}}{k^2}\,\frac{\partial T_m{}'}{e_1\partial u}\right)\epsilon \text{ times } (9)$$

is added to

$$e_3\left(\frac{\partial T_m}{e_1\partial u} + d_m\,\frac{h_m^{\,2}}{k^2}\,\frac{\partial T_m{}'}{e_2\partial v}\right)\epsilon \text{ times } (10)$$

and the result is integrated over the cross section. Using the orthonormality condition (29), the wave equation (3), and the boundary conditions (16), (19), and (20), one obtains:

$$\begin{aligned}\frac{dV_m}{dw} + j\,\frac{h_m^{\,2}}{\omega\epsilon_0}\,I_m = j\omega\mu_0\sum_n I_n \Bigg\{&\int_S \xi\,\frac{\chi_n^{\,2}\chi_m^{\,2}}{k^2}\,\epsilon^2 T_n T_m\,dS \\ &- \int_S \xi\epsilon^2\Bigg[\left(\frac{\partial T_n}{e_2\partial v} - d_n\,\frac{h_n^{\,2}}{k^2}\,\frac{\partial T_n{}'}{e_1\partial u}\right)\left(\frac{\partial T_m}{e_2\partial v} - d_m\,\frac{h_n^{\,2}}{k^2}\,\frac{\partial T_m{}'}{e_1\partial u}\right) \\ &+ \left(\frac{\partial T_n}{e_1\partial u} + d_n\,\frac{h_n^{\,2}}{k^2}\,\frac{\partial T_n{}'}{e_2\partial v}\right)\left(\frac{\partial T_m}{e_1\partial u} + d_m\,\frac{h_m^{\,2}}{k^2}\,\frac{\partial T_m{}'}{e_2\partial v}\right)\Bigg]dS\Bigg\}.\end{aligned} \tag{30}$$

Similarly

$$-e_3\left(\frac{\partial T_m}{e_1\partial u} + d_m\,\frac{\partial T_m{}'}{e_2\partial v}\right) \text{ times } (12)$$

is added to

$$-e_3\left(\frac{\partial T_m}{e_2\partial v} - d_m\,\frac{\partial T_m{}'}{e_1\partial u}\right) \text{ times } (13)$$

and the result is integrated over the cross section:

$$\begin{aligned}\frac{dI_m}{dw} + j\omega\epsilon_0 V_m = j\omega\epsilon_0\sum_n V_n \Bigg\{&\int_S \xi\epsilon\; d_n d_m\,\frac{\chi_n^{\,2}\chi_m^{\,2}}{k^2}\,T_n{}'T_m{}'\,dS \\ &- \int_S \xi\epsilon\Bigg[\left(\frac{\partial T_n}{e_2\partial v} - d_n\,\frac{\partial T_n{}'}{e_1\partial u}\right)\left(\frac{\partial T_m}{e_2\partial v} - d_m\,\frac{\partial T_m{}'}{e_1\partial u}\right) \\ &+ \left(\frac{\partial T_n}{e_1\partial u} + d_n\,\frac{\partial T_n{}'}{e_2\partial v}\right)\left(\frac{\partial T_m}{e_1\partial u} + d_m\,\frac{\partial T_m{}'}{e_2\partial v}\right)\Bigg]dS\Bigg\}.\end{aligned} \tag{31}$$

Equations (30) and (31) are the generalized telegraphist's equations for the curved waveguide with a dielectric lining.

Introducing traveling waves

$$\begin{aligned} V_m &= \sqrt{K_m}(a_m + b_m) \\ I_m &= \frac{1}{\sqrt{K_m}}(a_m - b_m) \end{aligned} \tag{32}$$

with

$$K_m = \frac{h_m}{\omega\epsilon_0}$$

the more convenient form in terms of amplitudes of forward (a_m) and backward (b_m) traveling waves is obtained.

$$\begin{aligned} \frac{da_m}{dw} + jh_m a_m &= j\sum_n (c_{mn}{}^+ a_n + c_{mn}{}^- b_n) \\ \frac{db_m}{dw} - jh_m b_m &= -j\sum_n (c_{mn}{}^+ b_n + c_{mn}{}^- a_n) . \end{aligned} \tag{33}$$

The coupling coefficients in (33) are

$$\begin{aligned} c_{mn}{}^{\pm} = & \\ & \pm \frac{\omega^2\mu_0\epsilon_0}{2\sqrt{h_m h_n}} \int_S \xi\epsilon^2 \left[\left(\frac{\partial T_n}{e_1\partial u} + d_n \frac{h_n{}^2}{k^2} \frac{\partial T_n{}'}{e_2\partial v} \right) \left(\frac{\partial T_m}{e_1\partial u} + d_m \frac{h_m{}^2}{k^2} \frac{\partial T_m{}'}{e_2\partial v} \right) \right. \\ & \qquad \left. + \left(\frac{\partial T_n}{e_2\partial v} - d_n \frac{h_n{}^2}{k^2} \frac{\partial T_n{}'}{e_1\partial u} \right) \left(\frac{\partial T_m}{e_2\partial v} - d_m \frac{h_m{}^2}{k^2} \frac{\partial T_m{}'}{e_1\partial u} \right) \right] dS \\ & \qquad - \frac{\sqrt{h_m h_n}}{2} \int_S \xi\epsilon \left[\left(\frac{\partial T_n}{e_1\partial u} + d_n \frac{\partial T_n{}'}{e_2\partial v} \right) \left(\frac{\partial T_m}{e_1\partial u} + d_m \frac{\partial T_m{}'}{e_2\partial v} \right) \right. \\ & \qquad \left. + \left(\frac{\partial T_n}{e_2\partial v} - d_n \frac{\partial T_n}{e_1\partial u} \right) \left(\frac{\partial T_m}{e_2\partial v} - d_m \frac{\partial T_m{}'}{e_1\partial u} \right) \right] dS \\ & \pm \frac{\omega^2\mu_0\epsilon_0}{2\sqrt{h_m h_n}} \int_S \xi\epsilon^2 \frac{\chi_n{}^2 \chi_m{}^2}{k^2} T_n T_m \, dS \\ & \qquad + \frac{\sqrt{h_m h_n}}{2} \int_S \xi\epsilon \, d_n d_m \frac{\chi_n{}^2 \chi_m{}^2}{k^2} T_n{}' T_m{}' \, dS . \end{aligned} \tag{34}$$

To analyze circular electric wave propagation, it is sufficient to consider only coupling between circular electric and other waves. Let m denote the TE_{om} wave; then

$$T_m = 0$$

and (34) reduces to

$$\begin{aligned} c_{mn}^{\pm} = \pm \frac{d_m h_n^2}{2\sqrt{h_m h_n}} \int_S \xi\epsilon \left(\frac{\partial T_n}{e_2 \partial v} - d_n \frac{h_n^2}{k^2} \frac{\partial T_n'}{e_1 \partial u} \right) \frac{\partial T_m'}{e_1 \partial u} \, dS \\ + \frac{1}{2} \sqrt{h_m h_n} \, d_m \left[\int_S \xi\epsilon \left(\frac{\partial T_n}{e_2 \partial v} - d_n \frac{\partial T_n'}{e_1 \partial u} \right) \frac{\partial T_m'}{e_1 \partial u} \, dS \right. \\ \left. + \frac{d_n}{\omega^2 \mu_0 \epsilon_0} \int_S \xi \chi_n^2 \chi_m^2 T_n T_m \, dS \right]. \end{aligned} \tag{35}$$

To find the z-dependence of the wave-amplitudes a_m and b_m for certain initial conditions, requires the solution of the generalized telegraphist's equations (33).

They are a system of simultaneous first-order and linear differential equations and can be solved by standard methods. From this point of view, (33) with (35) and all the preceding definitions represent the formal solution to propagation of circular electric waves in round waveguide with a single uniform lining. But this formal solution is still to be reduced to a practical form accessible for numerical evaluation. Also, heterogeneous or anisotropic linings have yet to be considered.

IV. A FIELD APPROXIMATION IN THE LINING

Before proceeding any further, an approximation will be made, which is justified for all practical linings in round waveguide for circular electric wave transmission. In all practical cases, the lining is thin compared to the radius of the guide

$$\rho - 1 \equiv \delta \ll 1. \tag{36}$$

Furthermore, it can be seen from (35) that there is only coupling between TE_{om} waves and waves of first circumferential order, i.e. $p = 1$. For these waves it may be safely assumed that

$$p^2 \ll \chi_n^{\epsilon 2} r^2 \tag{37}$$

for all

$$a < r < b$$

within the lining, since $| \chi_n^{\epsilon} r | \geqq | k_n^{\epsilon} |$ and, according to (28),

$$k_n^{\epsilon 2} = k^2 a^2 \left[\epsilon - \frac{h_n^2}{k^2} \right].$$

Under conditions (36) and (37), the wave functions (2) for the lining may be simplified, by replacing the Hankel-functions by their asymptotic expressions. The result is:

$$\begin{aligned} T_n &= N_n \frac{k_n^2}{k_n^{e2}} J_p(k_n) \frac{\sin\left(\rho - \frac{r}{a}\right) k_n^e}{\sin(\rho - 1)k_n^e} \sin p\varphi \\ T_n' &= N_n \frac{k_n^2}{k_n^{e2}} J_p(k_n) \frac{\cos\left(\rho - \frac{r}{a}\right) k_n^e}{\cos(\rho - 1)k_n^e} \cos p\varphi. \end{aligned} \tag{38}$$

The characteristic equation may now also be simplified to:

$$\left(y_n - \frac{1}{k_n^e} \tan \delta\, k_n^e\right)\left(y_n + \frac{\epsilon}{k_n^e} \cot \delta\, k_n^e\right) = p^2(\epsilon - 1)^2 \frac{h_n^2 k^2 a^4}{k_n^4 k_n^{e4}} \tag{39}$$

where

$$y_n = \frac{J_p'(k_n)}{k_n J_p(k_n)}. \tag{40}$$

Likewise, using (39) and the simplified form of (26), the factor d_n may be written:

$$d_n = \frac{k_n^2 k_n^{e2}}{p(\epsilon - 1)h_n^2 a^2}\left(\frac{\epsilon}{k_n^e} \cot \delta\, k_n^e + y_n\right). \tag{41}$$

The orthonormality condition (29) determines the normalization factor N_n. Using the asymptotic expressions (38) for the field functions within the lining, the integration in (29) results in:

$$\begin{aligned} \frac{\pi}{2} N_n^2 k_n^2 J_p^2(k_n) &\left\{\left(1 + d_n \frac{2h_n^2}{k^2}\right)\left(1 - \frac{p^2}{k_n^2} + k_n^2 y_n^2 + 2y_n\right)\right. \\ &\quad - 2\, d_n \frac{p}{k_n^2}\left[\left(1 + \frac{h_n^2}{k^2}\right) - \left(1 + \frac{h_n^2}{\epsilon k^2}\right)\frac{\epsilon k_n^4}{k_n^{e4}}\right] \\ + \frac{k_n^2}{2k_n^{e3}} &\left.\left(\epsilon \frac{2\delta k_n^e + \sin 2\delta k_n^e}{\sin^2 \delta\, k_n^e} + d_n^2 \frac{h_n^2}{k^2} \frac{2\delta k_n^e - \sin 2\delta\, k_n^e}{\cos^2 \delta\, k_n^e}\right)\right\} = 1. \end{aligned} \tag{42}$$

For circular electric waves with $p = 0$, the integration results in:

$$\begin{aligned} \pi N_m^2\, d_m^2 k_m^2 J_0^2(k_m) \frac{h_m^2}{k^2} &\left[1 + k_m^2 y_m^2 + 2y_m \right. \\ &\left. + \frac{k_m^2}{k_m^{e3}} \frac{2\delta k_m^e - \sin 2\delta\, k_m^e}{2 \cos^2 \delta\, k_m^e}\right] = 1. \end{aligned} \tag{43}$$

The asymptotic expressions (38) have also been used to calculate the coupling coefficients (35) for circular electric waves. Instead of y_n, another abbreviation

$$x_n = \frac{J_1(k_n)}{k_n\, J_0(k_n)} \tag{44}$$

has been used to facilitate numerical evaluation:

$$\begin{aligned} c_{mn}{}^{\pm} = {} & \frac{\pi}{2} \sqrt{h_m h_n}\, \frac{k_m{}^2\, d_m N_n N_m}{k_m{}^2 - k_n{}^2}\, J_1(k_n)\, J_0(k_m)\, \frac{a}{R} \cdot \Bigg\{ \left(1 \pm \frac{h_m}{h_n}\right) \Bigg[1 + d_n \\ & + d_n \frac{h_n{}^2}{k^2} \left(\frac{2k_n{}^2}{k_n{}^2 - k_m{}^2} - \frac{1}{x_n} \right) - \left(1 + d_n \frac{k_n{}^2 + k_m{}^2 \left[1 - 2\, \frac{\chi_n{}^2}{k^2} \right]}{k_n{}^2 - k_m{}^2} \right) \frac{x_m}{x_n} \\ & - d_n k_n{}^2 x_m + \epsilon \frac{k_n{}^2}{k_n{}^{e2}} \frac{k_m{}^2 - k_n{}^2}{k_n{}^{e2} - k_m{}^{e2}} \Bigg(1 - \frac{k_n{}^e}{k_m{}^e} \frac{\tan k_m{}^e\, \delta}{\tan k_n{}^e\, \delta} \\ & - d_n k_n{}^e \left[1 - \frac{k_n{}^2}{\epsilon k^2 a^2} \right] \tan k_n{}^e\, \delta + d_n \frac{k_n{}^{e2}}{k_m{}^e} \tan k_m{}^e\, \delta \Bigg) \Bigg] \\ & + d_n \frac{\chi_n{}^2}{k^2} \Bigg[\left(1 \pm \frac{h_m}{h_n} \frac{k_n{}^{e2}}{k_m{}^{e2}} \right) \frac{k_n{}^2 - k_m{}^2}{k_n{}^{e2} - k_m{}^{e2}}\, k_m{}^e \tan k_m{}^e\, \delta \\ & \qquad + \left(k_m{}^2 \pm \frac{h_m}{h_n} k_n{}^2 \right) x_m \pm \frac{h_m}{h_n} \left(\frac{x_m}{x_n} - 1 \right) \Bigg] \Bigg\} . \end{aligned} \tag{45}$$

Equations (39) to (45) reduce the problem to as simple analytical expressions as seems to be possible at this time. Any further simplification would only be brought about by replacing the trigonometric functions and the Bessel-functions by their Taylor series expansions for small arguments, respectively arguments k_n close to the roots of Bessel-functions for the empty guide. Such simplification, however, would lead to a first-order approximation for very thin lining, which has been studied in detail elsewhere.[1]

The present aim is for a better approximation. Therefore, numerical methods starting with expressions (39) through (45) will have to do the rest.

To this end the characteristic equation (39), which in implicit form determines the separation-constant k_n, will first have to be solved. For a lossless or low-loss lining, the relative permittivity is real or may be assumed real. Then all the quantities in (39) are real. An iterative procedure for this solution has been found earlier.[1]

For the sake of completeness it will be repeated here using the present symbols. Equation (39) may be written as:

$$\cot \delta k_n^e = \frac{F}{2}\left(1 \pm \sqrt{1 + \frac{4}{F^2\epsilon}}\right) \tag{46}$$

where

$$F = \frac{1}{k_n^e y_n}\left[1 + (\epsilon - 1)\frac{2p^2 h_n^2 k^2 a^4}{\epsilon\, k_n^4\, k_n^{e2}}\right] - \frac{k_n^e y_n}{\epsilon}. \tag{47}$$

A value for the relative permittivity ϵ of the lining is now specified. For a free-space propagation constant $k = 2\pi/\lambda$, a wave propagation constant h_n is assumed and in calculating

$$\begin{aligned} k_n^2 &= k^2a^2 - h_n^2a^2 \\ k_n^{e2} &= \epsilon k^2a^2 - h_n^2a^2 \end{aligned} \tag{48}$$

for lack of knowing the true radius of the lining, a is replaced by b, the radius of the guide. Using (47), a first approximation for the relative thickness is found. In general, according to the two signs of the square root in (46), there will be two such values of relative thickness which will lead to the same propagation constant h_n.

The first approximation for δ is used to correct the radius a of the lining in (47) and (48). The calculation is then repeated. Since for small values of δ, a change in δ affects the right-hand side only slightly, this method converges rapidly.

For typical values of b/λ and ϵ the phase constants of four normal modes have been plotted in Fig. 2.* The modes shown in Fig. 2 degenerate into TE_{11} TM_{11} TE_{12} and TM_{12} when the lining is very thin. Of all the modes, these four are most strongly coupled to TE_{01} by curvature. The broken lines represent first-order approximations as they have been found earlier.[1] Note that the first-order approximations hold only for extremely thin linings. In the case of the TE_{11} wave, in particular, the lining should be less than 0.05 per cent of the waveguide radius. Here the first-order approximations are of no use whatsoever.

Note also that the TM_{11} phase constant does not increase as expected from the first-order approximation. The curve levels off, and eventually

* These and most of the other numerical results have been obtained by H. P. Kindermann.[8]

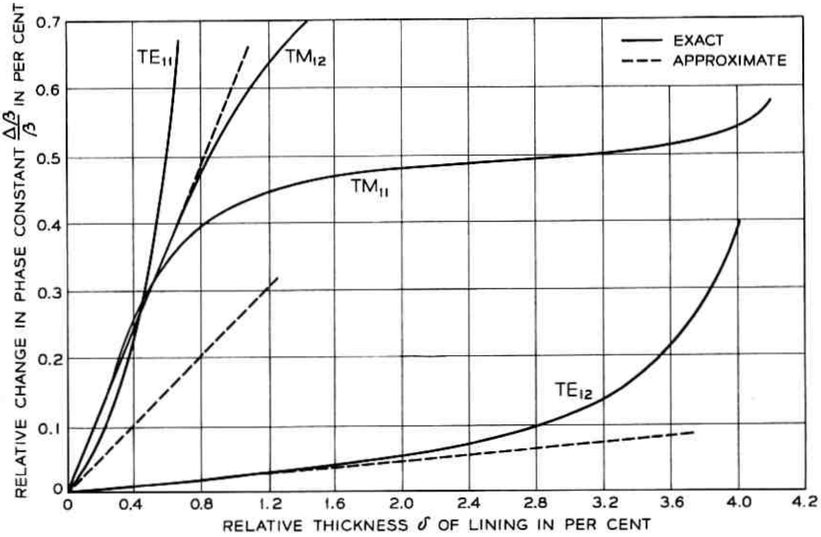


Fig. 2 — Change in phase constant of normal modes in lined waveguide, $b/\lambda = 4.70$, $\epsilon = 2.5$ — solid line exact; broken line approximate.

a heavier lining will not change the TM$_{11}$ phase. According to Fig. 3, at higher frequencies the TM$_{11}$ phase levels off at even lower values. Also, a higher permittivity will not change these relations.

This is, of course, very unfortunate since to reduce curvature losses the TM$_{11}$ phase should differ most from the TE$_{01}$ phase.

Having solved the characteristic equation, curvature coupling is

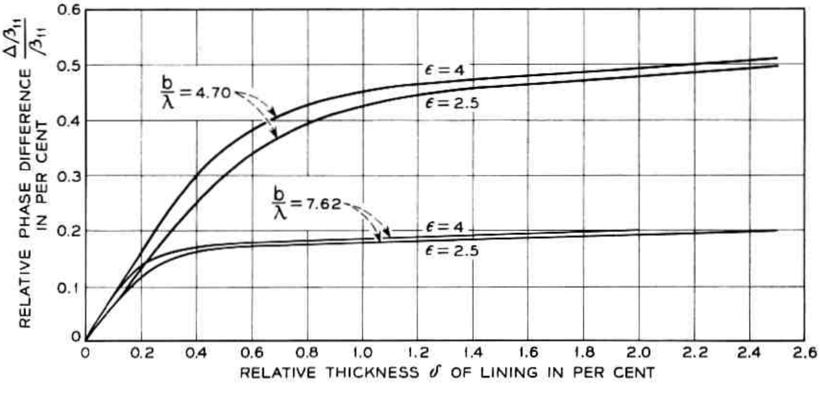


Fig. 3 — The phase constant of TM$_{11}$ is, over a wide range, nearly independent of δ and ϵ.

obtained by substituting numerical values into (45). For the TE_{01} characteristics first-order approximations may be substituted. Even a very heavy lining does not change these characteristics very much. For example, with $\delta = 0.03$ and $\epsilon = 2.5$ the relative change in phase constant of TE_{01} according to this approximation is only:

$$\frac{\Delta\beta_m}{\beta_m} = 2.10^{-4}.$$

The coefficients of curvature coupling between TE_{01} and TE_{11}, TM_{11} and TE_{12} are plotted in Fig. 4. Note again that any first-order approximations hold only for extremely thin linings.

The coupling between TE_{01} and TM_{11} is at first increased by the lining and then stays almost constant. The increase in TE_{01}-TM_{11} coupling disagrees with another first-order approximation.[4] The present result has to be considered correct, however, since the corresponding shielded helix waveguide curvature coupling is about equally enhanced.[9]

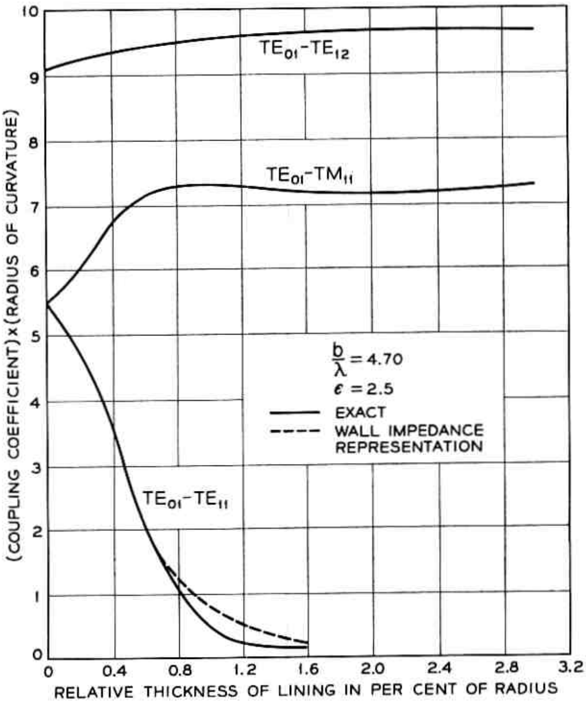


Fig. 4 — Coefficient of curvature coupling in lined waveguide $b/\lambda = 4.70$, $\epsilon = 2.5$ — solid line exact; broken line is wall impedance representation.

Curvature coupling between TE_{01} and TE_{11} decreases substantially in lined waveguide as is shown by the solid line in Fig. 4. The broken line will be referred to later. TE_{01}-TE_{12} coupling is nearly independent of the lining.

V. A WALL IMPEDANCE REPRESENTATION

The preceding analysis of lined waveguide considers only the simplest case, of a single isotropic and uniform lining. Yet this analysis could only be reduced to analytical expressions which are still quite involved and require a lot of computation for numerical evaluation. It is even more difficult to analyze a waveguide with a more complicated lining by the same methods. Further simplifications are necessary to facilitate the analysis of anisotropic or heterogeneous linings.

Such simplifications are brought about when the effects of the lining are described by wall impedances which the lining presents to the waveguide interior. Looking in radial direction, two wall impedances may be defined which are associated with the two different polarizations of the fields:

$$Z_w = -\frac{E_w}{H_v}; \qquad Z_v = \frac{E_v}{H_w}. \tag{49, 50}$$

For a mode n represented by one term n of the series expansions (8) and (15), these wall impedances can be expressed by the field functions (38):

$$Z_w = j\frac{\chi_n^e}{\omega\epsilon\epsilon_0}\,\frac{1}{\cot k_n^e\delta + d_n\dfrac{p}{k_n^e}\dfrac{h_n^2}{k_e^2}} \tag{51}$$

$$Z_v = j\frac{\omega\mu_0}{\chi_n^e}\left[\tan k_n^e\delta - \frac{p}{d_n k_n^e}\right]. \tag{52}$$

For circular symmetric modes $p = 0$ and

$$Z_{w0} = j\frac{\chi_n^e}{\omega\epsilon\epsilon_0}\tan k_n^e\,\delta \tag{53}$$

$$Z_{v0} = j\frac{\omega\mu_0}{\chi_n^e}\tan k_n^e\,\delta. \tag{54}$$

The characteristic equation can be derived with these wall impedances. Instead of satisfying the boundary conditions (16) through (19), the two conditions (49) and (50) will now be substituted. The ratio

$-(E_w/H_v)$ and E_v/H_w of the field components in the waveguide interior adjacent to the lining will be required to be equal to the wall impedances Z_w and Z_v presented by the lining:

$$-\frac{E_w}{H_v} = \frac{1}{j\omega\epsilon_0\, a\left(y_n - d_n\, \frac{p}{k_n^2}\, \frac{h_n^2}{k^2}\right)} = Z_w \tag{55}$$

$$\frac{E_v}{H_w} = j\,\frac{k^2a^2}{\omega\epsilon_0 a}\left(y_n - \frac{p}{k_n^2\, d_n}\right) = Z_v\,. \tag{56}$$

After the factor d_n has been eliminated from (55) and (56) one equation, the characteristic equation, remains:

$$\left(y_n + \frac{j}{\omega\epsilon_0 a Z_w}\right)\left(y_n + j\,\frac{Z_v}{\omega\mu_0 a}\right) = \frac{p^2}{k_n^4}\,\frac{h_n^2}{k^2}\,. \tag{57}$$

This equation is still exact to the same order as are the wall impedances. If expressions like (51) and (52) for the wall impedance were substituted, the same equation as before, that is (39), would result.

Instead of (51) and (52) wall impedance values for circular symmetric modes as given by (53) and (54) will now be substituted into (57). But for the rest, the circumferential index p will be kept general in (57). One obtains:

$$\left(y_n - \frac{1}{k_n^e}\tan\delta\, k_n^e\right)\left(y_n + \frac{\epsilon}{k_n^e}\cot\delta\, k_n^e\right) = \frac{p^2}{k_n^4}\,\frac{h_n^2}{k^2}\,. \tag{58}$$

This expression is quite similar to (39).

The left-hand sides of (58) and (39) are identical, and there is only a small difference on the right-hand sides of these equations. The right-hand side of (39), for example, may be written as:

$$\frac{h_n^2p^2}{k^2k_n^4}\,\frac{k^4a^4(\epsilon - 1)^2}{k_n^{e4}} = \frac{h_n^2p^2}{k^2k_n^4}\,\frac{(\epsilon - 1)^2}{\left(\epsilon - \frac{h_n^2}{k^2}\right)^2}\,. \tag{59}$$

All modes of interest are those which have nearly the same propagation constant h_n as the circular electric wave. Since under practical circumstances the circular electric wave propagates nearly as in free space, h_n will also be nearly equal to k. If under these circumstances the right-hand side of (39) according to (59) is compared with the right-hand side of (58), the difference is found to be very small indeed.

To determine the normal modes in a straight waveguide with a single dielectric lining, it therefore seems well justified to use (58) as characteristic equation instead of (39).

As further confirmation, this approximation (58) has been solved numerically and compared with solutions of (39).

Let δ' be the thickness of the lining. Solving (58) will give a certain phase constant for a particular mode. To obtain the same phase constant from (39), the thickness has to be δ. Plotted in Fig. 5 for the three modes is the relative error $(\delta' - \delta)/\delta$ that results from using (58) instead of the exact form (39).

An analytic expression for this error can be given for a very thin lining, when the modes in lined waveguide may be considered first-order perturbations of modes in metallic waveguide. Under these conditions for TM modes

$$\frac{\delta' - \delta}{\delta} = 0$$

and for TE_{pn} modes

$$\frac{\delta' - \delta}{\delta} = -\frac{k_{n0}^{\ 2}}{(\epsilon - 1)k^2a^2}\left[2 - \frac{k_{n0}^{\ 2}}{k^2a^2}\right]$$

where $J_p'(k_{n0}) = 0$.

In the numerical example of Fig. 5 the error is largest for the TE_{12} mode and very thin lining. For the other two modes the error stays generally below 1 per cent.

The wall impedance representation, therefore, holds for single isotropic linings of any practical dimensions.

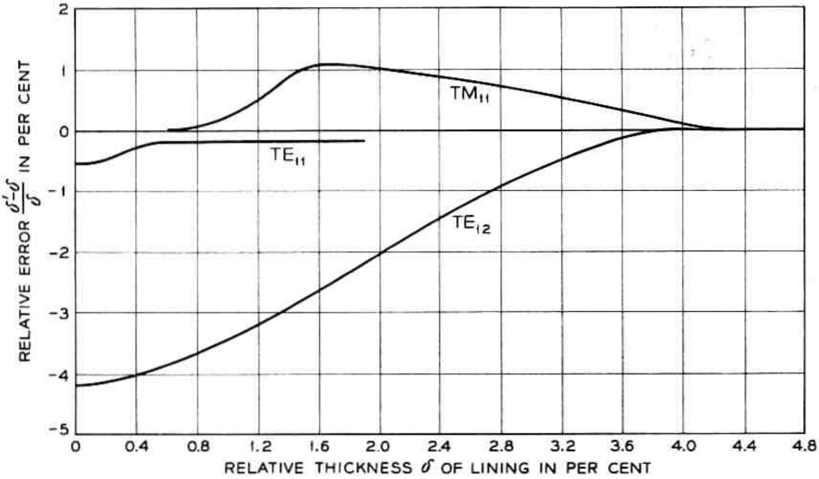


Fig. 5 — For the same phase constant, the thickness of the lining is δ according to (39) and δ' according to (58), $b/\lambda = 4.70$, $\epsilon = 2.5$.

The characteristic equation (58) for the waveguide with a single isotropic lining is not the only form to which the more general expression (57) can be reduced. It can be utilized in much more general cases. As long as we are able to determine the wall impedance Z_w and Z_v, we can use (57) to determine the normal modes of the structure for any lining such as anisotropic lining or heterogeneous linings.

The example of a single isotropic lining has taught us that it is sufficient to use wall impedance values of the corresponding circular symmetric modes. It is relatively easy to find these wall impedances even for quite complicated jacket structures. We will then be able to determine the normal modes characteristics of waveguides with such complicated jacket structures.

To make full use of the wall impedance representation in our analysis of the curved waveguide, some approximations are necessary for the coefficient of curvature coupling (35) and the normalization factor N_n.

To obtain these two quantities, products of field components and other functions had to be integrated over the total cross section. The range of integration included the lining.

In our present representation we do not explicitly determine the field-distributions within the lining, but the effect of the lining is taken into account by only considering the input impedances as seen from the waveguide interior. In this representation we therefore cannot calculate the contributions to the various integrals by extending them over the lining. We will consider the effect of neglecting these contributions.

Under practical circumstances, the area of the lining is always very much smaller than the total cross section. The components of the magnetic field, since they are continuous across a dielectric boundary, are of the same order of magnitude within the lining as in the empty space of the waveguide. Except for a possible change of the order ϵ, the same is true for the components of the electric field.

In summary, then, the integrals of products of field components over the area of the lining are always very much smaller than the corresponding integrals over the total cross section. They consequently might be neglected.

Under these circumstances, (42) reduces to

$$\begin{aligned}\frac{\pi}{2} N_n^{\,2} k_n^{\,2} J_p^{\,2} (k_n) \cdot \Bigg[\left(1 + d_n^{\,2} \frac{h_n^{\,2}}{k^2}\right) \\ \cdot \left(1 - \frac{p^2}{k_n^{\,2}} + 2\, y_n + k_n^{\,2}\, y_n^{\,2}\right) - 2\left(1 + \frac{h_n^{\,2}}{k^2}\right) d_n \frac{p}{k_n^{\,2}} \Bigg] = 1\end{aligned} \tag{60}$$

and (43) for circular electric waves to:

$$\pi\, N_m^{\ 2}\, d_m^{\ 2}\, k_m^{\ 2}\, J_0^{\ 2}\,(k_m)\,\frac{h_m^{\ 2}}{k^2}\,(1 + 2y_m + k_m^{\ 2}\, y_m^{\ 2}) = 1. \tag{61}$$

For the coefficient of curvature coupling we get instead of (45)

$$\begin{aligned} c_{mn}^{\ \pm} = {} & \frac{\pi}{2}\sqrt{h_m\, h_n}\,\frac{d_m\, k_m^{\ 2}\, N_n\, N_m}{k_m^2 - k_n^2}\, J_1\,(k_n)\, J_0\,(k_m)\,\frac{a}{R} \\ & \cdot\Bigg\{\left(1 \pm \frac{h_m}{h_n}\right)\Bigg[1 + d_n + d_n\,\frac{h_n^{\ 2}}{k^2}\left(\frac{2k_n^{\ 2}}{k_n^{\ 2} - k_m^{\ 2}} - \frac{1}{x_n}\right) \\ & - \left(1 + d_n\,\frac{k_n^{\ 2} + k_m^{\ 2}\left(1 - 2\,\frac{\chi_n^{\ 2}}{k^2}\right)}{k_n^2 - k_m^2}\right)\frac{x_m}{x_n} - d_n\, k_n^{\ 2}\, x_m\Bigg] \\ & + d_n\,\frac{\chi_n^{\ 2}}{k^2}\left[\left(k_m^{\ 2} \pm \frac{h_m}{h_n}\, k_n^{\ 2}\right) x_m \pm \left(\frac{x_m}{x_n} - 1\right)\frac{h_m}{h_n}\right]\Bigg\}. \end{aligned} \tag{62}$$

The factor d_n in all these equations can be determined from (55) or (56). For example from (55) we get:

$$d_n = \frac{k_n^{\ 2}}{p}\,\frac{k^2}{h_n^{\ 2}}\left(y_n + \frac{j}{\omega\epsilon_0 a Z_w}\right). \tag{63}$$

After the characteristic equation (57) has been solved for a particular combination of wall impedance values Z_w and Z_v, all the other quantities and eventually the coefficient of curvature coupling $c_{mn}^{\ \pm}$ can be found by straightforward evaluation of (60) to (63). The wall impedances Z_w and Z_v may of course be determined from the circular symmetric field components.

The approximations which have been made to obtain (60) to (63) have been examined more closely by numerical evaluation. The coefficient of curvature coupling has been calculated using these equations and compared with the plots in Fig. 4. For TM_{11}-TE_{01} and TE_{12}-TE_{01} coupling the differences are small enough not to show in Fig. 4. The wall impedance representation fails only for TE_{01}-TE_{11} coupling and $\delta > 0.8$ per cent. The corresponding coefficient of curvature coupling is shown by the broken line in Fig. 4. Fortunately the coupling is then so small and the phase constants of the waves differ so much that there is no significant interaction between TE_{01} and TE_{11}.

The explanation of why the two methods of calculation result in different values in just this case is as follows:

The TE_{11} wave in round waveguide is most strongly modified by the lining. Even in a thin lining, the TE_{11} field tends to be concentrated within the lining. In the present numerical example, it takes only a relative thickness of $\delta = 0.4$ per cent for the radial propagation constant k_n to become imaginary and consequently the TE_{11} fields to be evanescent towards the center of the guide. When this happens because of very weak TE_{11} fields within the guide, the curvature coupling to TE_{01} will be very weak too.

The wall impedance representation fails for calculating the coupling, because it entirely neglects any field interaction within the lining, which is more and more significant for TE_{11} and a thick lining. This phenomenon is limited to TE_{11} ; coupling to all other modes is accurately described by the wall impedance representation.

VI. WALL IMPEDANCE OF ANISOTROPIC AND HETEROGENEOUS LININGS

It has now been established that the wall impedance representation is a useful method in analyzing wave propagation in straight and curved sections of lined waveguide. To use this method for waveguides with anisotropic or heterogeneous linings we need to know the wall impedances of these linings.

6.1 *Anisotropic Lining*

Flock coating shows promise as a lining for circular electric waveguide. Resistive fibers of the flock are parallel to the electric field of unwanted modes but perpendicular to circular electric fields. A flock coat is anisotropic, and in an (x,y,z) system identified by

$$u = x, \qquad av = y, \qquad w = z \tag{64}$$

it may be described by the permittivity tensor

$$\| \epsilon \| = \begin{Vmatrix} \epsilon_x & 0 & 0 \\ 0 & \epsilon_z & 0 \\ 0 & 0 & \epsilon_z \end{Vmatrix} . \tag{65}$$

Wall impedances of circular symmetric modes are used in the wall impedance representation. In our present system of coordinates, circular symmetry corresponds to $\partial/\partial y = 0$.

Assuming furthermore a z-dependence of e^{-jhz}, Maxwell's equations may be written in the following form:

$$\begin{aligned}
j\,h\,E_y &= j\omega\mu H_x \\
-j\,h\,E_x - \frac{\partial E_z}{\partial x} &= -j\omega\mu H_y \\
\frac{\partial E_y}{\partial x} &= -j\omega\mu H_z \\
j\,h\,H_y &= j\,\omega\,\epsilon_x\,E_x \\
-j\,h\,H_x - \frac{\partial H_z}{\partial x} &= j\,\omega\,\epsilon_z\,E_y \\
\frac{\partial H_y}{\partial x} &= j\,\omega\,\epsilon_z\,E_z\,.
\end{aligned} \tag{66}$$

Eliminating the field components E_x, E_y and H_x, H_y from (66) we get:

$$\frac{\partial^2 E_z}{\partial x^2} + (\omega^2\mu\,\epsilon_x - h^2)\frac{\epsilon_z}{\epsilon_x}E_z = 0 \tag{67}$$

$$\frac{\partial^2 H_z}{\partial x^2} + (\omega^2\,\mu\,\epsilon_z - h^2)\,H_z = 0. \tag{68}$$

The general solution of these equations has the form

$$Ae^{j\chi x} + Be^{-j\chi x}$$

where for (67)

$$\chi = \chi_x\sqrt{\frac{\epsilon_z}{\epsilon_x}} = \sqrt{\frac{\epsilon_z}{\epsilon_x}\,(\omega^2\mu\,\epsilon_x - h^2)}$$

and for (68)

$$\chi = \chi_z = \sqrt{\omega^2\mu\epsilon_z - h^2}$$

are the propagation constants in x- or radial direction.

A wave traveling in positive x-direction or outwardly in the cylindrical system is represented by the second term. For such waves

$$\frac{\partial}{\partial x} = -j\chi$$

and we obtain from Maxwell's equations and (67)

$$\begin{gathered} j\,\omega\,\epsilon_z\,E_z = -j\,\chi_x\sqrt{\frac{\epsilon_z}{\epsilon_x}}\,H_y \\ Z_z = \frac{\chi_x}{\omega\sqrt{\epsilon_x\epsilon_z}} \end{gathered} \tag{69}$$

and from (68)

$$\begin{gathered} -j\omega\mu\,H_z = -j\,\chi_z\,E_y \\ Z_y = \frac{\omega\mu}{\chi_z}\,. \end{gathered} \tag{70}$$

Z_z and Z_y are the wave impedances of an anisotropic medium. The wall impedances of the anisotropic lining are input impedances of a radial transmission line of length $(b - a)$ short-circuited at the end. In our present approximation:

$$Z_w = j\,\frac{\chi_x}{\omega\sqrt{\epsilon_x\epsilon_z}}\tan\chi_x\sqrt{\frac{\epsilon_z}{\epsilon_x}}\,(b - a) \tag{71}$$

$$Z_v = j\,\frac{\omega\mu}{\chi_z}\tan\chi_z\,(b - a). \tag{72}$$

To make Z_w and Z_v constants of the waveguide, independent of a particular mode, we consider only modes which are sufficiently far from cutoff to propagate nearly as in free space. Then

$$\begin{gathered} \chi_x = \frac{2\pi}{\lambda}\sqrt{\frac{\epsilon_x}{\epsilon_0} - 1} \\ \chi_z = \frac{2\pi}{\lambda}\sqrt{\frac{\epsilon_z}{\epsilon_0} - 1}\,. \end{gathered} \tag{73}$$

For circular electric waves only Z_v enters the boundary condition. Note that in (72) Z_v is independent of ϵ_x. Resistive components in the flock coat will cause a loss factor only of ϵ_x. Such resistive components leave the circular electric wave loss unaffected.

6.2 *Double Lining*

A base layer of dissipative material and a top layer of low-loss material provide mode filtering for TE_{01} transmission and reduce TE_{01} loss in bends.[3]

Let the base lining have a relative permittivity ϵ_b and thickness $b - a$, the top lining ϵ_t and $a_1 - a$.

Input impedances of the base lining are

$$Z_{w1} = j \frac{\chi_1}{\omega \epsilon_b \epsilon_0} \tan \chi_1(b - a), \qquad Z_{v1} = j \frac{\omega\mu}{\chi_1} \tan \chi_1(b - a)$$

where

$$\chi_1 = k \sqrt{\epsilon_b - 1}.$$

These input impedances are transformed by the second lining according to ordinary transmission line theory

$$Z_w = j \sqrt{\frac{\mu_0}{\epsilon_0}} \frac{\sqrt{\epsilon_t - 1}}{\epsilon_t} \frac{\epsilon_t \sqrt{\epsilon_b - 1} \tan \varphi_b + \epsilon_b \sqrt{\epsilon_t - 1} \tan \varphi_t}{\epsilon_b \sqrt{\epsilon_t - 1} - \epsilon_t \sqrt{\epsilon_b - 1} \tan \varphi_b \tan \varphi_t} \tag{74}$$

$$Z_v = j \sqrt{\frac{\mu_0}{\epsilon_0}} \frac{1}{\sqrt{\epsilon_t - 1}} \frac{\sqrt{\epsilon_t - 1} \tan \varphi_b + \sqrt{\epsilon_b - 1} \tan \varphi_t}{\sqrt{\epsilon_b - 1} - \sqrt{\epsilon_t - 1} \tan \varphi_b \tan \varphi_t} \tag{75}$$

where

$$\varphi_b = 2\pi \sqrt{\epsilon_b - 1} \frac{b - a_1}{\lambda}$$

and

$$\varphi_t = 2\pi \sqrt{\epsilon_t - 1} \frac{a_1 - a}{\lambda}.$$

For a thin base layer $\varphi_b \ll 1$ and

$$Z_v = j \sqrt{\frac{\mu_0}{\epsilon_0}} \frac{\dfrac{2\pi}{\lambda}(b - a_1) + \dfrac{\tan \varphi_t}{\sqrt{\epsilon_t - 1}}}{1 - \dfrac{2\pi}{\lambda}(b - a_1) \sqrt{\epsilon_t - 1} \tan \varphi_1}. \tag{76}$$

Note that in (76) Z_v is independent of the permittivity in the base layer. Any loss in the base layer will not significantly raise circular electric wave loss.

VII. CONCLUSION

In previous first-order approximations, normal modes of lined waveguide were considered perturbed modes of plain waveguide, and coefficients of curvature couplings were assumed the same as in metallic waveguide. In some respects these approximations hold only for extremely thin linings, thinner for example than 0.05 per cent of the waveguide radius. The present more exact analysis shows that the TE_{11} wave has a phase constant much higher than would be expected from these approximations. Neither the thickness nor the permittivity of the lining can increase the phase difference between TM_{11} and TE_{01} beyond a certain limit. The phase difference eventually is almost independent of δ and ϵ and is small for high frequency.

Curvature coupling between TE_{11} and TE_{01} is substantially smaller in lined waveguide than in plain waveguide, while it is nearly independent of the lining between TE_{12} and TE_{01}. Between TM_{11} and TE_{01}, however, it first increases and then stays constant.

Waveguides with sandwiched or anisotropic linings may be analyzed by using a wall impedance representation. Wall impedances which the lining presents to fields of circular symmetry may be used in this analysis. They may easily be calculated for flock coatings and double linings. The wall impedance representation is found to be accurate for all typical cases.

REFERENCES

1. Unger, H. G., Circular Electric Wave Transmission in a Dielectric-Coated Waveguide, B.S.T.J., **36,** Sept., 1957, p. 1253.
2. Unger, H. G., Round Waveguide with Lossy Lining, Proc. of Symposium on Millimeter Waves, New York, 1959.
3. Unger, H. G., Round Waveguide with Double Lining, B.S.T.J., **39,** Jan., 1961, p. 161.
4. Noda, Ken-ichi, Circular Electric Wave Transmission through Hybrid-Mode Waveguide, Review of the Electrical Communication Lab. (Tokyo), **8,** No. 9–10, Sept.–Oct., 1960.
5. Katsenelenbaum, B. Z., The Effect of a Dielectric Film on the Attenuation of H_{01} Waves in a Straight Nearly Circular Waveguide, Radiotekhnika i Electronika, 1958, p. 38.
6. Kreipe, H. L., and Unger, H. G., Imperfections in Lined Waveguide, to be published.
7. Bressler, A. D., Joshi, G. H., and Marcuvitz, N., Orthogonality Properties for Modes in Passive and Active Uniform Waveguides. J. Appl. Phys., **29,** p. 794–799.
8. Kindermann, H. P., Kruemmungskopplung im Rundhohlleiter mit Dielektrischem Wandbelag. Diplom-Thesis submitted to 'Institut fuer Hoechstfrequenztechnik,' Technische Hochschule Braunschweig.
9. Unger, H. G., Helix Waveguide Theory and Application, B.S.T.J., **37,** Nov., 1958, p. 1599.

Some Traffic Characteristics of Communications Networks with Automatic Alternate Routing

By J. H. WEBER

(Manuscript received August 23, 1961)

As a first step in the investigation of communications networks with automatic alternate routing, a simulator has been prepared using the IBM 7090 high-speed digital computer. The simulator is capable of being applied to a large class of networks, the principal restrictions being that blocked calls are cleared, and no congestion or delay is encountered at the switching points. Although the first version of the simulation program requires that the alternate routing plan be fixed in advance (i.e., before a run), the program design is such that traffic-dependent alternate routing doctrines can easily be provided.

The simulator has so far been used to examine the behavior of small networks of various sizes, configurations, and alternate routing doctrines under normal and abnormal conditions of load. Several criteria are introduced and used to evaluate the relative performance of different networks, leading to conclusions regarding the merits of certain alternate routing procedures and the areas of profitable application of the networks studied. The overload capabilities of these networks are of particular interest and are examined in some detail.

I. INTRODUCTION

The recent rapid expansion of long-distance communications facilities to serve increasing civilian and military demands, along with the evolution of cheaper trunking facilities and more sophisticated switching techniques, continues to bring the problem of network design and engineering to the attention of communications engineers. Although methods have been developed for engineering certain types of networks for the most economical distribution of trunking facilities, several critical problems remain.

One of these is the lack of understanding of the behavior of alternate routing networks under overload conditions, whether the overload is local or system wide. Local disasters, such as storms, earthquakes, etc., have caused severe deterioration of service in certain regions owing to increased loads directed toward the affected area. At other times, such as Christmas Day in the United States, the pattern of traffic shifts radically, again causing serious overloads and long delays in completing calls. Finally, some concern is felt for the behavior of the system under the impact of some widespread disaster, where overloads may appear everywhere simultaneously.

Such considerations lead in turn to two questions. First, how shall networks be designed to be efficient during normal operation and yet not deteriorate catastrophically under overloads, and second, given a network design, can the switching pattern be altered for the duration of an overload to improve performance, and if so, how?

Another problem is our present inability to engineer any but the limited class of alternate routing networks of a "hierarchical" nature which have been widely used in the Bell System and elsewhere.

Since no analytic techniques appeared to be available or soon forthcoming to answer these questions, a simulation study was undertaken in the hope that some insights might be provided into the operation of such networks which would be helpful in their design and in the development of theoretical models to predict their characteristics.

Accordingly, a program was written for the IBM 7090 computer which enables various alternate routing philosophies to be simulated and compared. In line with the general nature of the problem being studied, the program was designed to accept a large variety of networks and be easily expandable to encompass more sophisticated alternate routing procedures as they evolve.

The capabilities and limitations of the simulator are outlined in some detail in the next section, followed by a description of the first experiment using the program. Finally the results are presented and analyzed, and some general characteristics of alternate routing networks of the types studied are set forth.

II. SIMULATOR CHARACTERISTICS

Although many of the problems which arise when alternate routing networks are overloaded are caused by switching delays and shortages, it was decided, as a first step, to consider only the effects of trunking, since the switching problems are unique to particular systems, and would in any event considerably complicate both the program and interpretation of the results. Accordingly, the program was constructed under the following restrictions:

(1) No blocking or delay is introduced by any switching point.

(2) Calls which do not receive service immediately are cleared from the system and do not return. (If setup time is assumed negligible, and there are no delays, the lack of retrials is not likely to materially affect the nature of the results.)

If the network is considered to consist of nodes (corresponding to switching centers) and links connecting them (corresponding to direct trunk groups), then each link may be assigned an originating traffic, a trunk group size and an alternate routing pattern. In addition, calls which overflow the direct route and are to be alternate routed may be assigned a directionality, or originating node, which allows one of two alternate routing configurations to be hunted over, depending upon the direction of the call. Every trunk group is a "two way" group, so no direction need be assigned to calls which are carried on the direct route.

The simulator will accept systems which contain as many as 63 switching points, each of which may be connected to any other switching point by up to 511 trunks. Calls which do not find an available trunk in the direct route may overflow through one of two sets of up to 63 specified routes, depending upon the direction of the call. Each alternate route may contain as many as 7 links, which implies switching through up to 6 intermediate nodes. (A modification of the program allows the alternate routes to be chosen on a "step by step" basis, where the first node in the alternate route chain is specified, and the call proceeds from node to node according to the alternate routing specification at the last node through which the call was switched. "Ring-around-the-Rosy" and "Shuttling" are prevented by keeping records of where the call has already been switched and not allowing it to use the same node twice.) It should be emphasized that the program as described here requires that the entire alternate route be specified at the originating link, and failure to connect on any link of an alternate route allows an entirely new route to be selected.

An over-all maximum size of the system, set by the limitations of the computer memory, is

$$(13 \times \text{Number of links}) + \text{Total number of trunks} \\ + \text{Total number of alternate routes} = 22{,}013.$$

For example, if a system has 40 nodes, (and therefore 780 links), and if there is a total of 3900 trunks in the system, (corresponding to an average of five trunks per link), then 7973 alternate routes, or about 10 per link, can be specified. This is a rather large system, and in fact the simulator allows for experimentation with a wide range of possible trunk, node, and alternate route configurations.

In order to estimate the reliability of the simulation results, outputs may be printed out at subintervals of the run. Furthermore, since the system starts empty, equilibrium may be attained before records are kept by running the program for a number of subintervals and discarding their records. The number of subintervals to be printed out, as well as the number to be discarded, can be specified in the input, along with the average holding time per call, the total time the simulation should be run in holding times, and indications as to what sort of alternate routing scheme is to be used. The holding times of all calls are assumed to be exponentially distributed with identical means, and traffic levels are varied by altering the average time between calls offered to each link. Pseudo-random numbers to specify the input are generated by a multiplicative congruential technique which gives a cycle of 2^{33} numbers before a repeat. The random number generator is not cleared after every simulation, so that if several experiments are made successively, they will not utilize the same sequence of random numbers. Thus runs can be repeated identically if desired, or, alternatively, a different set of random numbers can be used for the same system configuration by the simple expedient of reordering the experiments.

The information which is printed out, in addition to that derived directly from the inputs, (number of nodes, number of trunks and loads per link, alternate routing patterns, etc.) is as follows:

(1) An estimate of the probability of loss (blocking) from each link; i.e., the proportion of calls directly offered to a specific link which are unable to be served at all.

(2) An estimate of the probability of direct overflow; i.e., the proportion of calls which overflow the direct route, although they may be served on an alternate route.

(3) Number of calls offered to each link, both directly and as an alternate route.

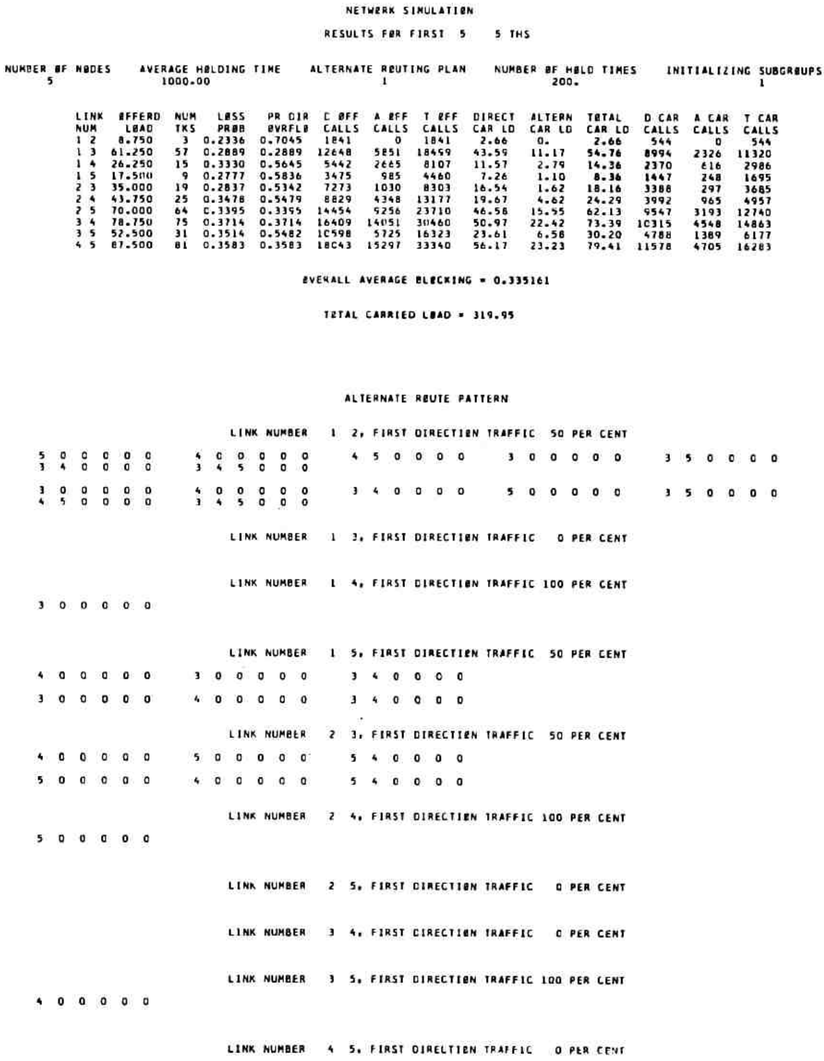
(4) Load in erlangs carried by each link, both from direct and alternate routed traffic.

(5) Calls carried by each link, both from direct and alternate routed traffic.

(6) Over-all average blocking; i.e., $\sum_{i=1}^{n} a_i p_i / \sum_{i=1}^{n} a_i$ where a_i is the load offered to link i, and p_i is the blocking experienced by a_i.

(7) Total carried load, which is really total calls multiplied by holding times. That is, a call is assumed to provide the number of erlangs its holding time would represent, regardless of how many links are used. This quantity is derived indirectly, by multiplying the total offered load by one minus the overall average blocking.

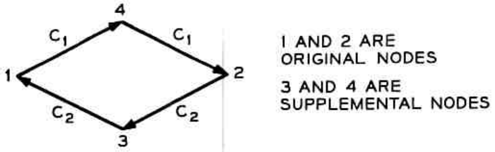
For moderate sized systems this program will process calls at the rate of about 1,200,000 calls per hour. A sample output is shown in Fig. 1.



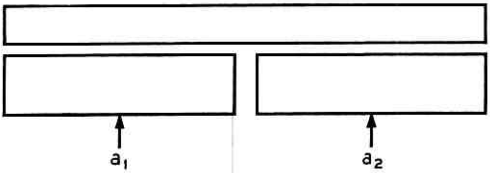
```
                                            NETWØRK SIMULATIØN

                                        RESULTS FØR FIRST  5    5 THS

NUMBER ØF NØDES     AVERAGE HØLDING TIME     ALTERNATE RØUTING PLAN     NUMBER ØF HØLD TIMES     INITIALIZING SUBGRØUPS
      5                  1000.00                       1                       200.                        1

          LINK  ØFFERD  NUM    LØSS    PR DIR  C ØFF  A ØFF  T ØFF  DIRECT  ALTERN  TØTAL   D CAR  A CAR  T CAR
          NUM    LØAD   TKS    PRØB    ØVRFLØ  CALLS  CALLS  CALLS  CAR LD  CAR LD  CAR LD  CALLS  CALLS  CALLS
          1 2    8.750    3  0.2336  0.7045   1841      0   1841    2.66   0.        2.66    544      0    544
          1 3   61.250   57  0.2889  0.2889  12648   5851  18459   43.59  11.17     54.76   8994   2326  11320
          1 4   26.250   15  0.3330  0.5645   5442   2665   8107   11.57   2.79     14.36   2370    616   2986
          1 5   17.500    9  0.2777  0.5836   3475    985   4460    7.26   1.10      8.36   1447    248   1695
          2 3   35.000   19  0.2837  0.5342   7273   1030   8303   16.54   1.62     18.16   3388    297   3685
          2 4   43.750   25  0.3478  0.5479   8829   4348  13177   19.67   4.62     24.29   3992    965   4957
          2 5   70.000   64  C.3395  0.3395  14454   9256  23710   46.58  15.55     62.13   9547   3193  12740
          3 4   78.750   75  0.3714  0.3714  16409  14051  30460   50.97  22.42     73.39  10315   4548  14863
          3 5   52.500   31  0.3514  0.5482  1C598   5725  16323   23.61   6.58     30.20   4788   1389   6177
          4 5   87.500   81  0.3583  0.3583  18C43  15297  33340   56.17  23.23     79.41  11578   4705  16283

                                    ØVERALL AVERAGE BLØCKING = 0.335161

                                         TØTAL CARRIED LØAD = 319.95



                                          ALTERNATE RØUTE PATTERN

                             LINK NUMBER   1  2, FIRST DIRECTIØN TRAFFIC  50 PER CENT
5 0 0 0 0 0       4 0 0 0 0 0       4 5 0 0 0 0       3 0 0 0 0 0       3 5 0 0 0 0
3 4 0 0 0 0       3 4 5 0 0 0

3 0 0 0 0 0       4 0 0 0 0 0       3 4 0 0 0 0       5 0 0 0 0 0       3 5 0 0 0 0
4 5 0 0 0 0       3 4 5 0 0 0

                             LINK NUMBER   1  3, FIRST DIRECTIØN TRAFFIC   0 PER CENT


                             LINK NUMBER   1  4, FIRST DIRECTIØN TRAFFIC 100 PER CENT
3 0 0 0 0 0


                             LINK NUMBER   1  5, FIRST DIRECTIØN TRAFFIC  50 PER CENT
4 0 0 0 0 0       3 0 0 0 0 0       3 4 0 0 0 0
3 0 0 0 0 0       4 0 0 0 0 0       3 4 0 0 0 0

                             LINK NUMBER   2  3, FIRST DIRECTIØN TRAFFIC  50 PER CENT
4 0 0 0 0 0       5 0 0 0 0 0       5 4 0 0 0 0
5 0 0 0 0 0       4 0 0 0 0 0       5 4 0 0 0 0

                             LINK NUMBER   2  4, FIRST DIRECTIØN TRAFFIC 100 PER CENT
5 0 0 0 0 0


                             LINK NUMBER   2  5, FIRST CIRECTIØN TRAFFIC   0 PER CENT


                             LINK NUMBER   3  4, FIRST CIRECTIØN TRAFFIC   0 PER CENT


                             LINK NUMBER   3  5, FIRST DIRECTIØN TRAFFIC 100 PER CENT
4 0 0 0 0 0


                             LINK NUMBER   4  5, FIRST DIRELTIØN TRAFFIC   0 PER CENT
```

Fig. 1 — Sample computer output.

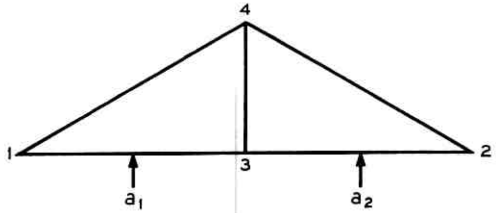
Clearly, the flexible nodal structure upon which the program is based allows certain things to be done by expending extra nodes which are not directly programmed. For example, if it is desired to have two one-way groups in a link, it is necessary merely to assign no trunks to the direct link, and have each of the alternate route patterns contain one node, which is assigned for the purpose. So, a single link would then have 4 nodes and 4 links with trunks assigned, as shown below.



Another possible use is the simulation of progressive graded multiples. Suppose it is desired to simulate the simple multiple shown below.



This is clearly equivalent in terms of loads carried and blocking to the following nodal structure.



In this analogue, a_1 has links 1-4 and 3-4 as an alternate route, while a_2 overflows through 2-4 and 3-4. Thus, if links 1-4 and 2-4 are provided with more trunks than 3-4, they can introduce no blocking and the system corresponds to the graded multiple above where link 3-4 is equivalent to the common group. This sort of flexible structure, then, appears to be useful in many ways, and may in fact come to have application beyond its original intent.

This program was primarily designed as a tool for the evaluation of alternate routing networks and as an aid in formulating principles for their design and administration, although one of the purposes was to assist in the solution of real problems as they arose. Accordingly, studies have been begun with the aim of codifying classes of networks and determining the significant parameters, advantages and disadvantages of each.

III. ANALYSIS OF NETWORKS

As a first step in the study of alternate routing communications networks it is desired to compare the behavior of several alternative configurations under normal and overload conditions. The variables which seem most likely to be significant in determining network performance are:

(1) Number of switching points in the network.
(2) Overall size of the network, perhaps best described as a measure of the average load or number of trunks per link.
(3) Alternate routing procedure used. Thus, a system which allows all traffic to overflow in some specified manner will probably perform differently than one which considers some routes to be "high usage" and from which traffic is alternate routed, and others to be "finals," from which no alternate routing is permitted.
(4) Type of overload encountered. A uniform (system wide) overload, for example, may cause a behavior quite different from an overload which is confined to a particular portion of the network.

In order to estimate the performance of networks when these parameters vary, and yet keep the results simple enough so that they can be easily interpreted, eight different networks were studied, each having two different alternate routing plans. In each case both uniform and nonuniform overloads were considered.

The eight networks studied consisted of two networks with three, two with four, two with five, and two with six nodes. The loads were adjusted so that the average load per link varied from three and one half erlangs per link in the most lightly loaded network to about 28 erlangs per link in the most heavily loaded configuration.

For purposes of convenience, the following terminology will subsequently be used:

(1) A *link* is a connection between two nodes, which may have any number of trunks, including zero.
(2) A *node* is a switching center, characterized by two or more links terminating at it.
(3) If network A is *larger* than network B, it has more nodes.
(4) If network A is *heavier* than network B, it has more offered erlangs per link on the average.

(5) A *hierarchical* alternate route network is one in which at least some of the trunk groups are *high usage*; i.e., traffic which is not carried can be overflowed to other groups, at least some of which are *finals*, which have no alternate routes.
(6) A *symmetrical* alternate route network has only high usage groups.
(7) A *simple* network has only final routes; i.e., no alternate routing is allowed.

The procedure which was followed in all cases was to postulate loads offered to each link in the network. For the sake of generality, these loads are ordinarily unequal, although in some cases equal loads are used in places where it is thought that this will not prejudice the results. Each network was engineered for a loss of 1 per cent on the worst link for simple, symmetrical and hierarchical networks. Loads were then applied corresponding to (a) 25 per cent, 50 per cent, 75 per cent, and 100 per cent uniform overload and (b) 25 per cent, 50 per cent, 75 per cent, and 100 per cent overload on all routes terminating at node 1. The selection of node 1, of course, is quite arbitrary, but this choice appears to be immaterial in the symmetrical case, and is likely to be most typical for hierarchical networks. (The heaviest loads in the hierarchical networks were reserved for the final routes, since this will allow most effective use of the hierarchy and seems to correspond to actual practice.) The simulations ran for 200 holding times for heavy networks and 1000 holding times for light networks, with an additional 20 per cent of this time (i.e., 40 or 200 holding times) discarded at the beginning of each run to remove the initial transient. Results were printed out at five subintervals of the total run, and examined to determine that the initial transient had been removed and the run was long enough to yield results sufficiently accurate for the purposes of this study.

Sketches of the networks are shown in Figs. 2 to 9 along with tables indicating the link loads and trunks assigned for each of the alternate routing doctrines. (Two sketches of each network are provided, one of which can be easily related to a symmetrical alternate routing philosophy while the other suggests a hierarchical doctrine. The dashed lines in the hierarchical sketch denote high usage groups, while the solid lines represent final routes. Since all links are high usage in the symmetrical system this distinction is not needed, and identical solid lines were used throughout.) The simple networks were engineered using the Erlang B tables, while the hierarchical networks were engineered using conventional methods with the sort of hierararchy used in the Bell System, allowing about 0.7 erlangs (25 ccs) on the last trunk in a high usage group. The (hierarchical) configurations were then checked experimentally using the simulator, and adjustments were made where required. The symmetrical

NETWORK PARAMETERS, 3 NODES — LIGHT

Link Number	Engineered Loads-Erlangs	Engineered Trunks		
		Simple	Symmetrical	Hierarchical
1–2	2	7	6	1
1–3	4	10	8	13
2–3	6	13	11	16
Total	12	30	25	30
Ave/Link	4	10	8.35	10

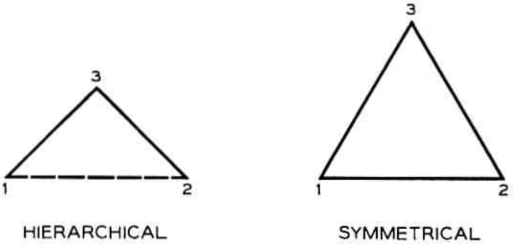


Fig. 2 — Three-node network models with table of link loads and trunk assignments for light loading.

NETWORK PARAMETERS, 3 NODES — HEAVY

Link Number	Engineered Loads-Erlangs	Engineered Trunks		
		Simple	Symmetrical	Hierarchical
1–2	5	11	11	4
1–3	10	18	16	21
2–3	15	24	21	27
Total	30	53	48	52
Ave/Link	10	17.65	16	17.35

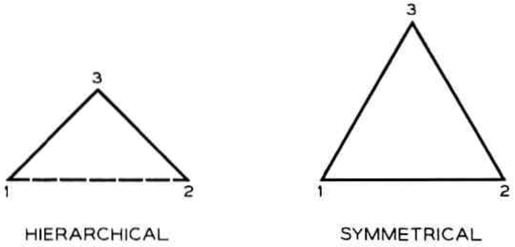
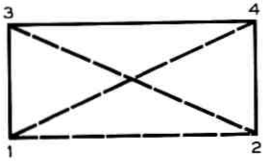


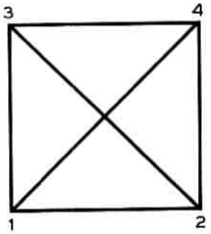
Fig. 3 — Three-node network models with table of link loads and trunk assignments for heavy loading.

Network Parameters, 4 Nodes — Light

Link Number	Engineered Loads-Erlangs	Engineered Trunks		
		Simple	Symmetrical	Hierarchical
1–2	1	5	4	0
1–3	4	10	8	13
1–4	2	7	5	2
2–3	3	8	6	3
2–4	5	11	9	15
3–4	6	13	10	17
Total	21	54	42	50
Ave/Link	3.5	9	7	8.33



HIERARCHICAL

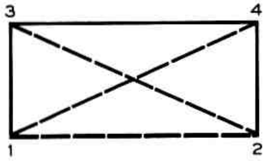


SYMMETRICAL

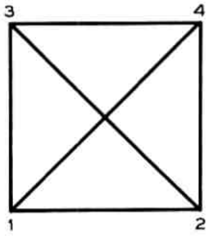
Fig. 4 — Four-node network models with table of link loads and trunk assignments for light loading.

Network Parameters, 4 Nodes — Heavy

Link Number	Engineered Loads-Erlangs	Engineered Trunks		
		Simple	Symmetrical	Hierarchical
1–2	5	11	12	3
1–3	20	30	27	37
1–4	10	18	18	9
2–3	15	24	22	14
2–4	25	36	32	43
3–4	30	42	38	52
Total	105	161	149	158
Ave/Link	17.5	26.8	24.8	26.35



HIERARCHICAL



SYMMETRICAL

Fig. 5 — Four-node network models with table of link loads and trunk assignments for heavy loading.

NETWORK PARAMETERS, 5 NODES — LIGHT

Link Number	Engineered Loads-Erlangs	Engineered Trunks		
		Simple	Symmetrical	Hierarchical
1–2	1	5	4	0
1–3	5	11	9	17
1–4	2	7	5	1
1–5	2	7	5	1
2–3	3	8	6	2
2–4	4	10	7	3
2–5	6	13	10	19
3–4	6	13	10	21
3–5	4	10	7	4
4–5	7	14	11	23
Total	40	98	74	91
Ave/Link	4	9.8	7.4	9.1

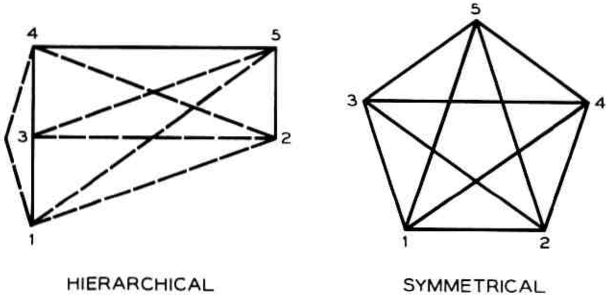


Fig. 6 — Five-node network models with table of link loads and trunk assignments for light loading.

networks were designed to allow each parcel of traffic to overflow through all other nodes in turn and were engineered entirely with the simulator by trial and error. A first estimate of trunk quantities was made using a fixed differential between the load in erlangs and the number of trunks in each link, and corrections were then made as required.

Having established this framework, or procedure for evaluation, a critical question is, What criteria can be used to compare the performance of various types of networks? It is desirable to take account of the efficiency (carried load per dollar of investment) of the network at all times, as well as the grades of service which are provided to each group of customers. Although grade of service here can no longer be interpreted as the small percentage of blocked calls that is ordinarily encountered at normal engineered loads, it is nevertheless incumbent upon the network

NETWORK PARAMETERS, 5 NODES — HEAVY

Link Number	Engineered Loads-Erlangs	Engineered Trunks		
		Simple	Symmetrical	Hierarchical
1–2	5	11	13	3
1–3	35	47	43	57
1–4	15	24	23	15
1–5	10	18	18	9
2–3	20	30	28	19
2–4	25	36	33	25
2–5	40	53	48	64
3–4	45	58	53	75
3–5	30	42	38	31
4–5	50	64	58	81
Total	275	383	355	379
Ave/Link	27.5	38.3	35.5	37.9

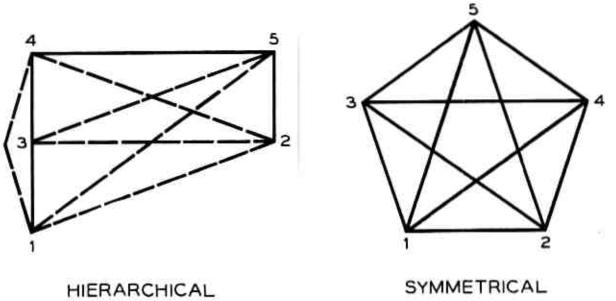


Fig. 7 — Five-node network models with table of link loads and trunk assignments for heavy loading.

designer to consider the extent to which service is degraded on any particular link. Similarly, one would expect the efficiency under overload conditions to be higher than that encountered during normal operation, but the relative efficiencies of networks using various alternate routing doctrines (to carry the same loads) may be rather different. It is clear, for example, that if a call uses a trunk in each of two links, there is a possibility of lower network efficiency being obtained than if it used a trunk in only one link. It is one of the purposes of this study to determine at what overload point such a loss in efficiency takes place, and what if any remedial action can be taken.

Thus, two rather different criteria appear to be important, one of which is essentially an economic variable, and the other a service variable. They are both further complicated by the fact that the first depends on the relative costs of trunks in different links, and the second

NETWORK PARAMETERS, 6 NODES — LIGHT

Link Number	Engineered Loads-Erlangs	Engineered Trunks		
		Simple	Symmetrical	Hierarchical
1–2	1	5	4	0
1–3	5	11	8	16
1–4	1	5	4	1
1–5	3	8	7	3
1–6	2	7	5	2
2–3	2	7	5	2
2–4	6	13	9	18
2–5	3	8	6	3
2–6	4	10	7	4
3–4	4	10	7	4
3–5	6	13	9	21
3–6	4	10	7	4
4–5	5	11	8	5
4–6	7	14	11	23
5–6	7	14	11	24
Total	60	146	108	130
Ave/Link	4	9.75	7.2	8.67

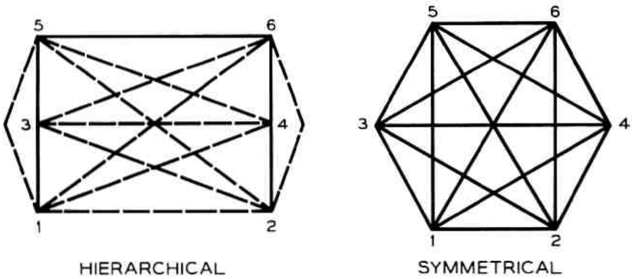


Fig. 8 — Six-node network models with table of link loads and trunk assignments for light loading.

has a different value for every parcel of traffic in the network. In order to simplify these complexities and reduce the number of variables which enter into the measure of performance, only the worst blocking in the network will be considered as the service criterion. This is, of course, conservative, and reflects the difficulties which might occur when alternate routing is canceled and a small parcel of traffic has no trunks in the direct path. The blocking on such a parcel would then be unity, and it would quickly be noticed that a parcel of traffic is isolated.

The problem of assigning costs to trunks is more difficult, of course, since there is no apparent logical worst or best case. Thus the assumptions made here for the relative costs are quite arbitrary and oversimpli-

NETWORK PARAMETERS, 6 NODES — HEAVY

Link Number	Engineered Loads-Erlangs	Engineered Trunks		
		Simple	Symmetrical	Hierarchical
1–2	5	11	15	3
1–3	40	53	48	67
1–4	10	18	19	9
1–5	20	30	28	22
1–6	15	24	24	13
2–3	10	18	19	9
2–4	40	53	48	67
2–5	20	30	28	18
2–6	25	36	33	27
3–4	30	42	38	27
3–5	45	58	53	82
3–6	30	42	38	30
4–5	35	47	43	35
4–6	50	64	58	90
5–6	50	64	58	97
Total Ave/Link	425 28.3	590 39.4	550 36.7	596 39.7

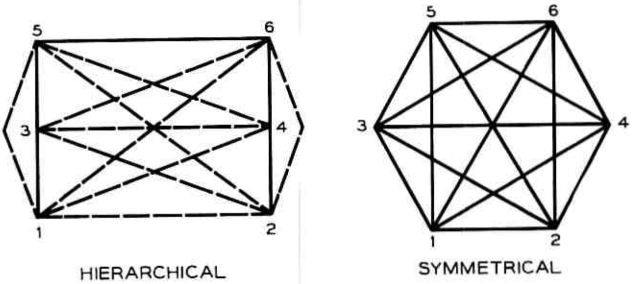


Fig. 9 — Six-node network models with table of link loads and trunk assignments for heavy loading.

fied, but may still be useful in evaluating network performance. Two different assumptions will be made. The first is that all trunks have the same cost. This might be a reasonably realistic assumption in a network where the designers are likely to think of symmetrical alternate routing doctrines. In effect, it states that the distance between any two nodes is not sufficiently different from the distance between any other two nodes to materially affect the cost of trunking facilities between them. Although this may appear to represent an unrealistic geographical situation, it may not be too far in error if the economics of long haul, large cross section trunking facilities are considered. In such systems, the terminal costs make up a large portion of the total trunk cost, and these, of course, are independent of the length of haul. The second assumption is that some routes are half as expensive as the others. For example, in Fig. 4, routes 1–3 and 2–4 are each considered to be half as long as each of the other routes in the network, all of which are assumed to be virtually the same length (all lengths here, of course, refer to costs, which are ordinarily roughly proportional to lengths). This assumption is geographically reasonable, and is, in fact, the kind of layout that is often encountered and which may well have prompted the development of hierarchical alternate routing procedures. Although neither of these weighting schemes may exactly represent an actual case, using each assumption in its logical place may yield more realistic comparative results than would using the same assumption throughout.

Having reduced the parameters for evaluation to two (worst blocking and load carried per dollar of investment), they can be combined into one by the following argument. Both of these parameters, which will be called B (blocking) and E (efficiency) generally increase with increasing loads (although E may occasionally decrease in a non-simple network). A large value of E is generally desirable, but, of course, a large value of B is not. In fact, quite the reverse is true, and so a high value of $(1 - B)$ is a desirable goal. Furthermore, the two factors will increase under different circumstances. For example, a highly efficient network may readily yield a very high value of E, but will also cause very high blocking. Thus B will be high and $(1 - B)$ low. Conversely, a loosely engineered network is likely to provide good service under overloads, yielding a relatively low B and high $(1 - B)$, but in turn be inefficient, with a low E. In both of these cases, the product $E(1 - B)$, will take on some intermediate value. Accordingly, a figure of merit for networks, called the Performance Measure, will be defined as $M = E(1 - B)$. This number may be dimensioned to lie between zero and one and will pass through a maximum as the load is increased. At engineered levels it will essentially represent the network efficiency, and as the load is increased it will indicate when service or efficiency or both are deteriorating. A high M is clearly a mark of a well performing system, with efficient trunk usage and at least tolerable service, while a low or rapidly decreasing M will show a system which is either being inefficiently used or is providing poor service or both. If $M = 0$ an intolerable situation exists; i.e., either some parcel of traffic is unable to be served or no load is being carried by the network. If a network can be designed to be efficient under normal conditions and to maximize M during moderate or partial overloads, and steps can be taken to prevent too rapid a degradation of this quantity during severe overloads, then it is a reasonable assumption that this design will be satisfactory for its purposes. That is, it will provide efficient communications facilities at normal loads, and will also allow the continuation of at least tolerable service between all points under adverse conditions.

One more modification was made in the results before analysis was undertaken. Since large trunk groups are more efficient than small ones, the group size would naturally affect the values of E and M, both at normal loads and under overload conditions. Accordingly, E and M were then plotted for each network relative to the E and M of the corresponding simple networks. The simple network was chosen as a convenient reference point, since it is easily engineered and requires no complicated switching equipment for implementation. Thus, one of the strong changes in efficiency which is not caused by the alternate routing pattern is largely removed, permitting comparisons among the latter to be more readily made.

IV. RESULTS

In order to investigate the effects of various parameters on network behavior, the network efficiencies, E, and performance measures, M, were calculated for engineered loads and for the various overload conditions which were tested. The relative values of E and M (relative to a simple network designed to carry the same loads) were then plotted versus the degree and kind of overload experienced by the network. Thus Fig. 11 shows the values of M for symmetrical networks, for loads ranging from engineered to 100 per cent overload where the overloads occur uniformly throughout the network. Fig. 13 exhibits the same information for hierarchical networks. Figs. 12 and 14 show the behavior of M with load when the overloads occur only on those links which terminate at node 1 with all other loads remaining at engineered levels. Finally, Figs. 15 through 18 are graphs of efficiency (E) versus load for the same situations as pertain to Figs. 11 through 14. The points from which the (smoothed) curves were plotted are shown in Figs. 12 to 18. They are omitted in Fig. 11 for the sake of clarity.

In order to keep the comparisons between symmetrical and hierarchical networks on a somewhat realistic basis, it is necessary to make some adjustment for the probable differences in geography which are likely to encourage consideration of one or the other type of network. Accordingly, as was mentioned above, certain trunks were considered to cost twice as much as others, which introduced a weighting factor into the values of E and hence into M as well. For example, in the 4 node case shown in Fig. 4, trunks in links 1-3 and 2-4 were considered to be only half as expensive as trunks in the remaining four links in the network. This reduced the cost of the trunk plant to 0.805 times the value which would result if all trunks were assumed to be of equal value in the case of a simple network, and to 0.720 times its former value for the hierarchical case. Thus the relative efficiency of the hierarchical network is increased by a factor of 0.805/0.720 or 1.119. This sort of adjustment was made in all calculations relating to hierarchical networks, while all trunks in symmetrical networks were assumed to be of equal value. The weighting factors obtained for the various networks are tabulated in Fig. 10.

The symmetrical networks which were studied in detail operated in the following fashion. Traffic which was blocked from any link was overflowed to an alternate route consisting of two links in tandem. If the call was blocked on this route it was offered to still another two-link route, and so on until all such routes were exhausted. No call was permitted to use a route which required more than two links in tandem. Some experiments were made on networks which allowed three tandem links to be used, but is was found that they were at best marginally more efficient than the two-link maximum network at engineered loads and deteriorated much more violently under overloads. Therefore, they are not considered further in this paper. The order of selection of alternate routes was arranged to approximately equalize the load overflowed to each route. Although this is probably not the most efficient arrangement, it should be adequate to illustrate the behavior of symmetrical networks.

The hierarchical networks operated in a manner similar to the Bell System toll network, with the difference that whereas in the Bell System the routes are selected link by link, in the simulation the routes are entirely preselected at the originating node. If a network is drawn as shown in the hierarchical sketches in Figs. 2 to 9, the route selection is made by

OVERALL SYSTEM CHARACTERISTICS

Case	Engd. Load/ Link	Trunk Value Adjustment for Hierarchy	Average Number of Links/Call					
			Engineered Load		100% Uniform Overload		100% Local Overload	
			Hier.	Symmet.	Hier.	Symmet.	Hier.	Symmet.
3 Nodes, light	4	1.194	1.155	1.014	1.156	1.128	1.203	1.124
4 Nodes, light	3.5	1.119	1.181	1.034	1.185	1.229	1.251	1.184
5 Nodes, light	4	1.181	1.288	1.074	1.305	1.305	1.372	1.210
6 Nodes, light	4	1.194	1.263	1.117	1.328	1.363	1.324	1.245
3 Nodes, heavy	10	1.121	1.076	1.034	1.073	1.164	1.138	1.180
4 Nodes, heavy	17.5	1.064	1.088	1.040	1.121	1.271	1.140	1.202
5 Nodes, heavy	27.5	1.071	1.100	1.062	1.147	1.309	1.151	1.274
6 Nodes, heavy	28.3	1.085	1.134	1.074	1.205	1.364	1.203	1.322

Fig. 10 — Over-all system characteristics.

hunting up the hierarchy, starting from the terminating office, in the distant region, and then down the hierarchy in the home region. (A region can be thought of as all centers whose final routes ultimately terminate at the same highest level switching center.) Thus, for example, in the four node network shown in Fig. 4, the alternate routes for traffic offered to link 1-2 are:

For half the traffic, 1-4-2, 1-3-2, 1-3-4-2; and for the remainder of the traffic, 1-3-2, 1-4-2, 1-3-4-2.

4.1 *General Observations*

In order to draw conclusions from this study as to the relative merits of various types of alternate routing systems under different load conditions, Figs. 10 to 18 will be examined and the significance of the results discussed.

As a very first step, a cursory examination of all figures reveals the following:

(1) The relative effectiveness of alternate routing networks, whether measured by E or M, tends to decrease with overload, with the decrease occurring more rapidly under uniform than under local overload. Although in some cases the network remains superior to a simple network even for 100 per cent overload, the relative performance at such overloads is almost always poorer than at engineered loads. This is due, of course, to the fact that the average number of links per call increases with overload, causing a decrease in efficiency which may outweigh the gains yielded by the larger effective access provided by the alternate route system. (See Fig. 10.)

(2) Light networks (those with less traffic), gain more from alternate routing than do heavy networks. This seems to occur because systems designed for large parcels of traffic use large efficient groups. Thus providing alternate routes in heavy networks, which increases the effective access somewhat, does not materially increase the efficiency, while the degradation caused by using several links per call is nevertheless present. In lighter networks, the increase in efficiency owing to the larger effective access is substantial, overriding the degradation and causing a considerable gain in effectiveness.

Perhaps to this list should be added:

(3) As mentioned above, symmetrical systems do not appear to benefit from allowing more than two links in tandem to be used by any call. This effect is apparently caused by the decrease in

efficiency which results from using many links per call overriding the gain yielded by increased access. In this situation, of course, the increase in link occupancy may be substantial, while the increase in effective access is likely to be small.

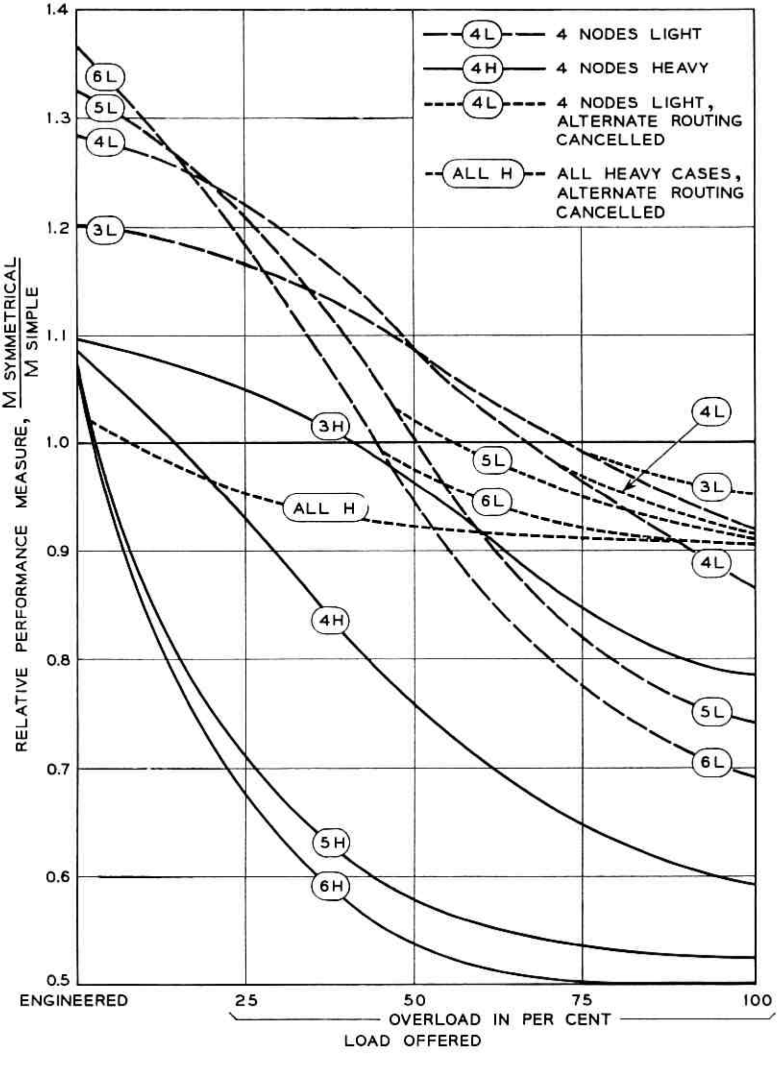


Fig. 11 — Relative performance measure of symmetrical networks under uniform overloads.

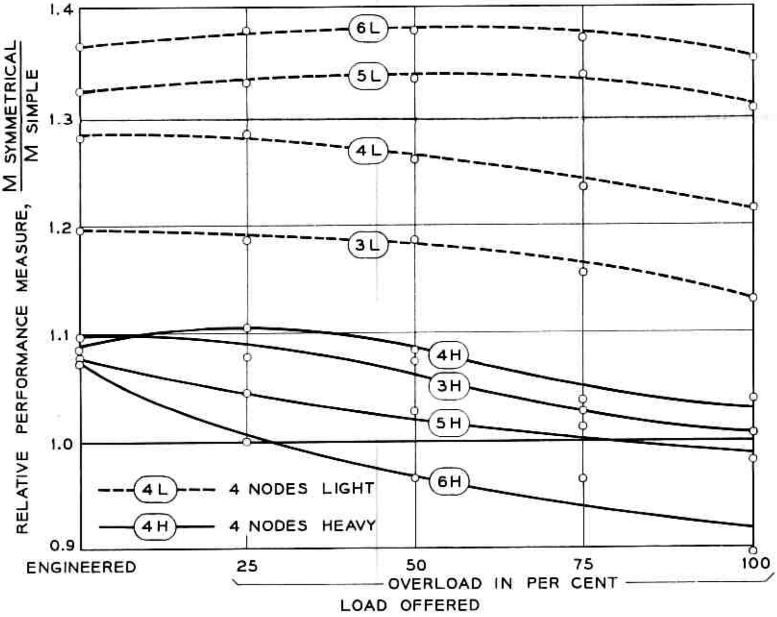


Fig. 12 — Relative performance measure of symmetrical networks under local overloads.

4.2 *Symmetrical Networks*

The curves shown in Figs. 11 through 18, when studied closely, reveal much information regarding the characteristics of the networks considered. In Fig. 11, the high relative performance measure of light symmetrical networks at engineered loads and the rapid decline as the load is increased uniformly is clearly indicated. The heavier networks exhibit a lower relative value of M at engineered loads and also decline rapidly, bringing their performance measure down to very low relative values at high overloads. Such a rapid decrease in M, it would appear, would make it impracticable to install symmetrical systems in many actual applications, were it not for the fact that M can be kept relatively high by canceling alternate routing at some appropriate point. The dotted lines in Fig. 11 indicate the relative performance measure if alternate routing is canceled, and it is clear that this factor can be kept above 0.9, regardless of the size of the network and even for 100 per cent overload. In any event, it does appear that for extremely heavy networks the decline is so precipitous that this method of alternate routing might well prove to be inapplicable. Fig. 12, however, illustrates the real

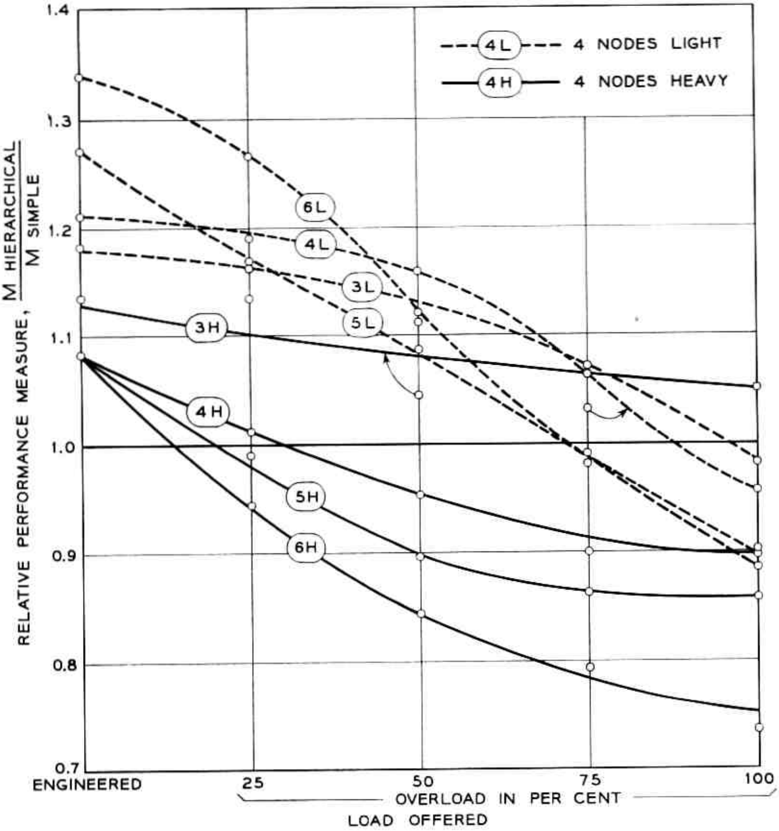


Fig. 13 — Relative performance measure of hierarchical networks under uniform overloads.

strength of the symmetrical routing doctrine. The relative performance measure is shown to be almost constant under local overloads, and remains above unity for all but the largest, heaviest networks. Furthermore, this sort of alternate routing structure is likely to be quite efficient at engineered loads in systems where call setup time and switching delays are no longer negligible, since it generally uses a relatively small number of links per call, as evidenced in Fig. 10.

A symmetrical network structure then can be devised which has the following characteristics:

(1) Performance measure (and thus efficiency) are high at engineered loads.

(2) Local overloads are well tolerated, with the network remaining

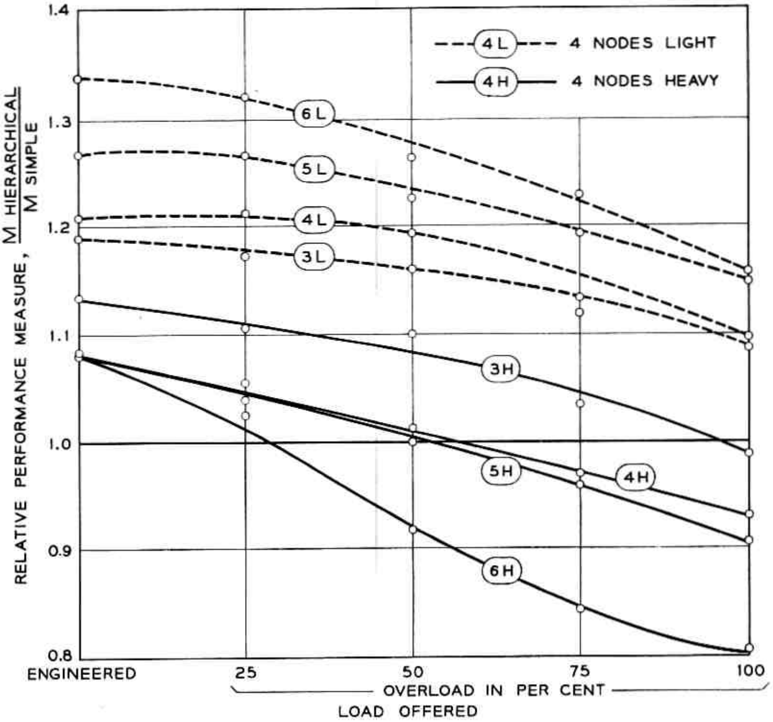


Fig. 14 — Relative performance measure of hierarchical networks under local overloads.

efficient and not allowing any parcel of traffic to suffer excessive blocking.

(3) If alternate routing can be canceled at the appropriate point, then the performance measure can be maintained at a tolerable level even under severe uniform overloads.

(4) The average number of links per call is quite low at engineered loads, increasing rapidly as overloads are applied.

An important practical question in (3), however, is whether a network control can be devised to cancel alternate routing easily, and how the control can determine the degree of overload. Another disadvantage of such networks is the unavailability, at present, of any but the very crudest methods of trunk engineering. However, this type of network is, in principle, capable of satisfying the four points listed above, all of which are desirable and often are difficult to attain concurrently.

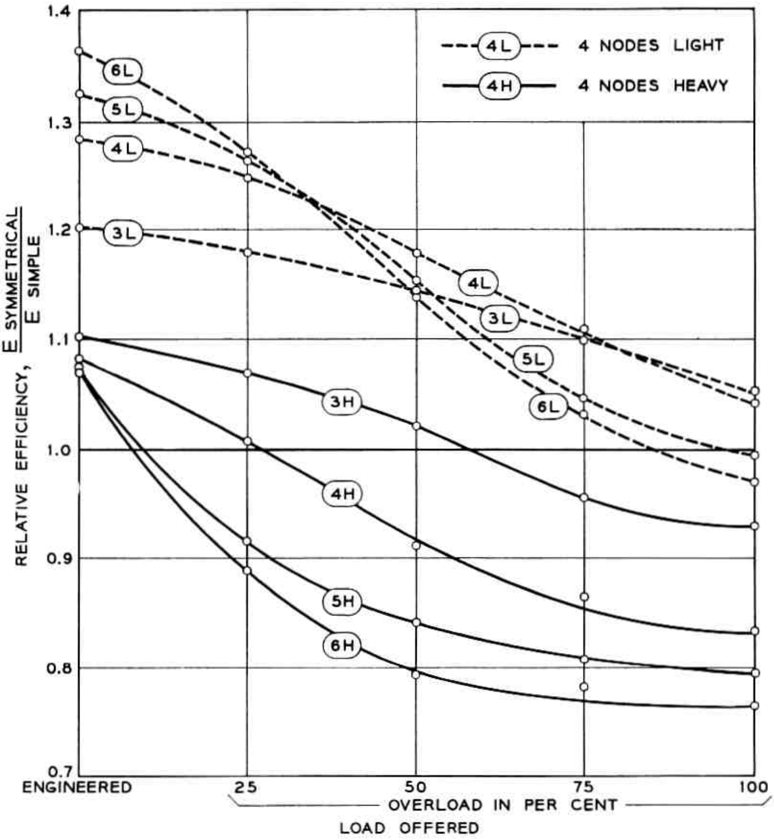


Fig. 15 — Relative efficiency of symmetrical networks under uniform overloads.

4.3 *Hierarchical Networks*

In Fig. 13, the relative performance measure of hierarchical networks under uniform overloads is shown. A comparison with Fig. 11 indicates that the relative M is higher at engineered loads for symmetrical than for hierarchical light networks and not too different for heavy networks, although the decline with uniform overload is more rapid in the former case. In the hierarchical system, however, the relative M cannot be increased by complete cancellation of alternate routing, since this increases the blocking on some parcels of traffic which are offered to high usage groups to a high level. Fig. 13 then shows that, although light

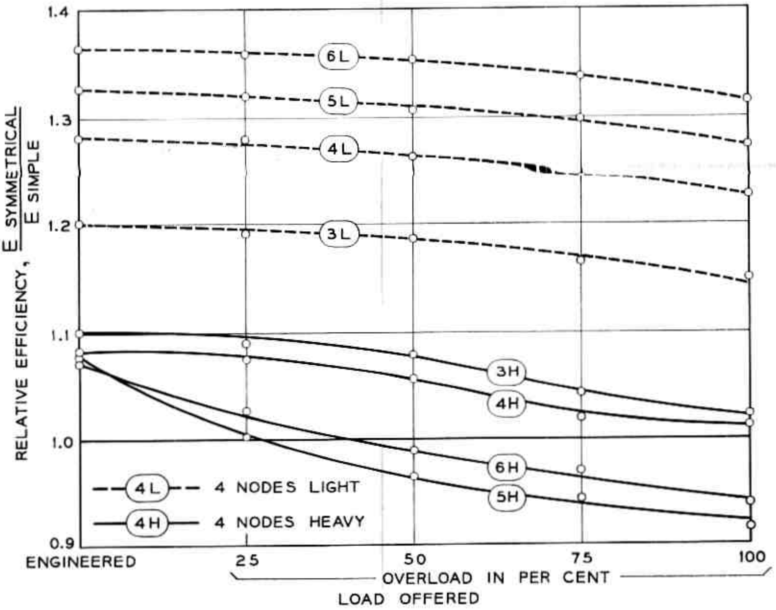


Fig. 16 — Relative efficiency of symmetrical networks under local overloads.

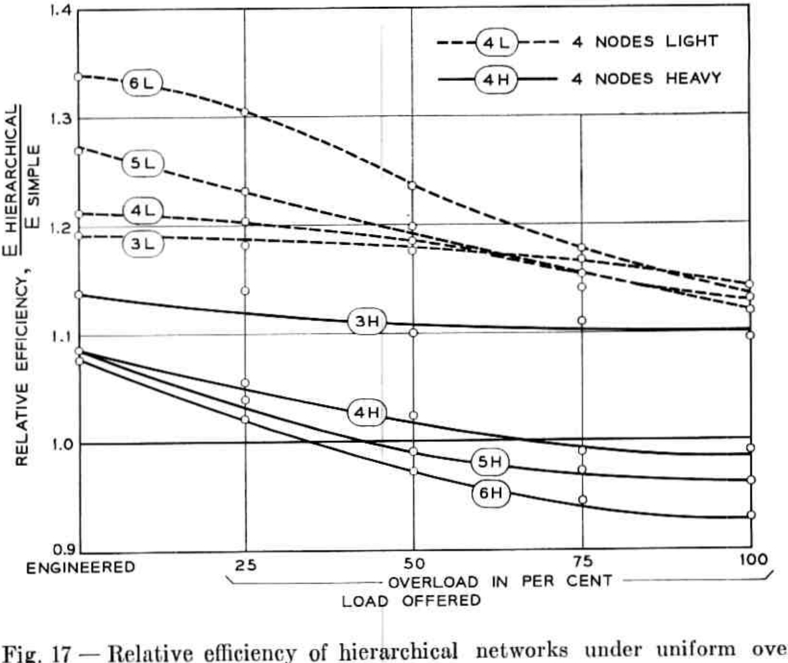


Fig. 17 — Relative efficiency of hierarchical networks under uniform overloads.

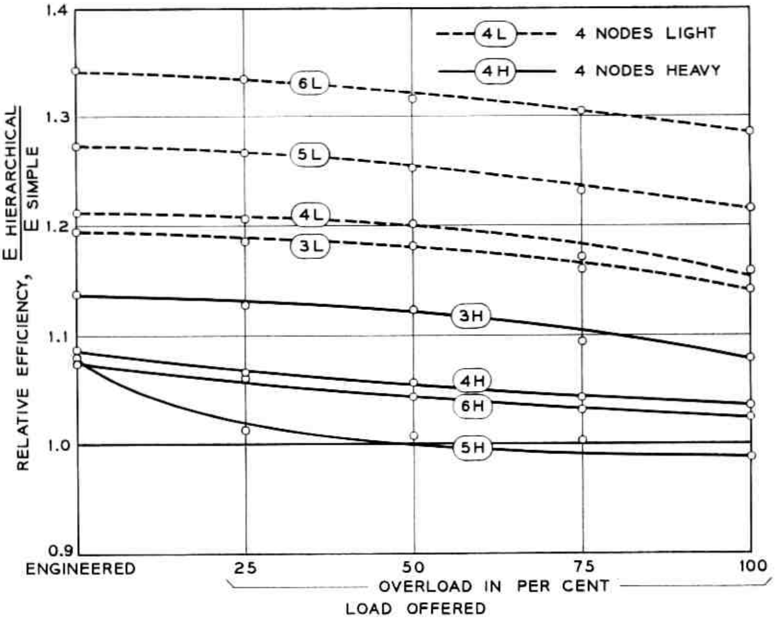


Fig. 18 -- Relative efficiency of hierarchical networks under local overloads.

networks retain their effectiveness up to 100 per cent overload, large heavy networks show a decline with uniform overloads to a quite low value of relative *M*.

The behavior of such networks under local overloads is shown in Fig. 14. In these circumstances the relative performance measure declines slowly from the value at engineered load as the local overload is increased. The decline is sufficiently gradual to enable the lighter networks to retain an *M* greater than unity for all overloads considered. The heavier networks, however, are unable to do this, and the value of relative *M* for the worst network declines almost to 0.8 for the greatest local overload.

The essential operating characteristics of networks of this basic design then appear to be as follows:

(1) The performance measure (and thus efficiency) tend to be high at engineered loads (if the variation in trunk costs is taken into account).

(2) The performance measure declines at a moderate rate under uniform overload, reaching rather low values for large, heavy

networks. No simple corrective measures are available to improve the situation but no special measures are needed to prevent catastrophic performance degradation. (Certain more complicated corrective measures, such as selective cancellation of alternate routing, might prove effective, but this sort of procedure was not studied.)

(3) Local overloads are moderately well tolerated, with the relative performance measure showing a gradual decline with increasing load, and dipping below unity for some cases.

(4) A relatively large number of links are used per call at engineered loads, and this number increases gradually with overloads.

This is then essentially a moderately well behaving network, providing neither superlative nor intolerable service at any level of load. It requires no complex controls to keep operating reasonably well, and is relatively simple to implement without the need for sophisticated switching equipment at the tandem points. In a real system, with switching delays and appreciable call setup time, however, this type of network may behave badly under overloads, since some calls use many links in tandem, and therefore can tie up a great deal of equipment when processing a call, even though the call is not completed. In fact, the large number of links used per call in hierarchical systems even at engineered loads is a source of inefficiency in such systems.

An apparent peculiarity in all the curves is the superiority of large light networks over small light networks at engineered loads, with the situation reversing as the load increases, so that at 100 per cent overload the small light networks are generally superior. A qualitative explanation of this would again involve the average number of links per call, which increases more rapidly in large networks than in small ones. The heavy networks do not exhibit this effect at all, and the larger heavy networks always appear to perform less well than the smaller ones. Since the larger (heavy) networks were designed to be more heavily loaded than the smaller ones, however, (see Fig. 10), this effect is more likely to be a result of network load than size.

4.4 *Efficiency Curves*

Figs. 15 to 18 portray the network efficiencies in the several cases studied. In general, these curves display a somewhat shallower slope than the corresponding curves for M. This implies that as the load is increased, not only does the relative network efficiency decrease, but the blocking encountered by the most poorly served group of customers also increases more rapidly when alternate routing is in effect than when it is not. The only exception to this is the symmetrical system under local overload (Figs. 12 and 16). In this case the relative efficiency and the relative performance measure decline at about the same rate, which implies that in this system the blocking remains at essentially the same level whether a symmetrical or simple system is in use. This is an important consideration in favor of symmetrical networks, particularly since both efficiency and performance measure remain reasonably high for all types of overloads considered.

V. CONCLUSIONS

The foregoing discussion of various types of alternate routing networks may be of use in determining whether alternate routing structures should be incorporated into particular switching systems and, if so, of what sort they should be. Many of these factors have long been known and used by network designers, and the present study should provide additional documentation. In the case of factors not previously considered, this study may provide justification for their incorporation into future designs. Some of these are as follows:

(1) If the overload capability of the system is not important, some sort of alternate routing system is almost certainly justified on economic grounds.

(2) If *local* overload capability is important, then strong consideration should be given to a symmetrical alternate routing network, since this configuration allows the blocking to be kept to a minimum under local overloads while retaining a high network efficiency.

(3) If *uniform* overload capability is an important consideration, then alternate routing structures should be contemplated with caution, but can still be used if the average load per link is small and appropriate action, such as cancellation of alternate routing (either uniformly or selectively) can be taken as required.

(4) If the average load per link is small, alternate routing almost always is advantageous, while if it is large, the advantage is sometimes questionable.

(5) If the initial efficiency is an important criterion, then the selection of the type of alternate routing may well depend upon the geography of the particular system. Thus, in certain situations, where small towns communicate primarily with nearby cities, a hierarchical structure may be preferable, while if there is a large group of

approximately equal-sized cities spread over the country, then a symmetrical system could prove to be superior.

(6) If switching equipment is expensive or call setup time is long, symmetrical networks may prove to be superior to hierarchical structures at engineered loads regardless of the geography. This would come about because of the large number of links per call, and hence the large amount of switching equipment used by hierarchical networks. Clearly, long setup time in this case would lead to inefficient trunk usage, since trunks in one link would be held while the call progressed along a multi-link path.

(7) Although not shown specifically in these results, a multi-alternate route structure provides service protection, which a simple layout does not, and a well connected symmetrical network is likely to be less vulnerable to damage than a hierarchical system.

Most actual systems, of course, must be designed to be efficient at engineered loads, and yet must also be able to accept either uniform or local overloads without excessive degradation of service. Furthermore, real networks usually serve many small towns communicating primarily with larger cities, which in turn communicate with each other on a roughly equal basis. Therefore, the network designer must decide which of these often conflicting criteria are most important, and develop a system which satisfies these as closely as it can within the limitations imposed by the switching and signaling equipment and the available methods of trunk engineering. It is quite likely that the best system in most situations is some combination of symmetrical and hierarchical networks, not necessarily of the particular kinds studied here. Furthermore, the advent of electronic switching systems and high speed signaling devices has made alternate routing doctrines which are dependent on the state of the system feasible, and these may well prove to be superior than any system with a completely prespecified alternate routing structure. However, an analysis of the basic characteristics of simpler networks is likely to be useful in predicting the behavior and influencing the design of specific, more complex systems. It was this potential application which motivated the studies described in this paper.

Contributors to this Issue

M. R. AARON, B.S. in E.E., 1949 and M.S. in E.E., 1951, University of Pennsylvania; Bell Telephone Laboratories, 1951—. He first worked on analysis, design and synthesis of transmission networks for L3 and submarine cable systems. From 1954 to 1956 he supervised a group concerned with design of networks for the L3 system. Since 1956 he has been in charge of a group engaged in systems analysis of PCM. Member I.R.E.

LEE G. BOSTWICK, B.S.E.E., 1922, University of Vermont; American Telephone and Telegraph Company, 1922–26; Bell Telephone Laboratories, 1926–61 (Ret.). Mr. Bostwick's first assignment was on early Bell System public address and program transmission projects. After becoming a member of the Laboratories in 1926, he participated in research initially on free field acoustic measuring techniques and later on the advance development of electroacoustic instruments including wide range loud speakers for theatres. During World War II, he contributed to the development of electrodynamic forms of underwater sound projectors, underwater acoustic measuring techniques, and sonar systems. During the years following he worked in apparatus development on the vibrational mechanics of switching apparatus and on tuned reed filters and selectors. Fellow, Acoustical Society of America, American Association for Advancement of Science; member I.R.E., American Institute of Physics, Phi Beta Kappa.

CHARLES J. BYRNE, B.S.E.E., 1957, Rensselaer Polytechnic Institute; M.S., 1958, California Institute of Technology; Bell Telephone Laboratories, 1958—. At the Laboratories he has investigated fast transistor logic, instrument noise in seismometers, and synchronization of digital systems. Member I.R.E., Sigma Xi, Eta Kappa Nu, Tau Beta Pi.

S. GELLER, A.B., 1941, and Ph.D., 1949, Cornell University; DuPont Postdoctoral Fellow at Cornell, 1949–50; Du Pont Company, 1950–52; Bell Telephone Laboratories, 1952—. At the Laboratories he has specialized in studies of crystal structure, with emphasis on crystal chemistry studies and the relation of the properties of crystals to their structures. He is one of the American co-discoverers of ferrimagnetic garnets, and took part in work which led to the discovery of Nb_3Sn, an intermetallic compound used in a superconductor electromagnet. Member American Crystallographic Association, American Physical Society, Mineralogical Society of America, Summit Association of Scientists (of the Research Society of America), Sigma Xi, Phi Kappa Phi.

A. JAY GOLDSTEIN, B.S., 1948, and M.A., 1951, Pennsylvania State University; Ph.D., 1955, Massachusetts Institute of Technology; faculty, Polytechnic Institute of Brooklyn, 1954–57; Bell Telephone Laboratories, 1957—. At the Laboratories he first engaged in research on information theory and on noise problems. More recently he has been concerned with quantization, timing and synchronization problems in pulse code modulation systems. Member American Mathematical Society, Sigma Xi.

JAMES R. GRAY, B.S. in E.E., 1954, and M.S.E., 1955, University of Florida; Bell Telephone Laboratories, 1955—. He was first engaged in repeater design for pulse code modulation systems. Since 1958 he has concentrated on PCM transmission impairment studies.

WILBUR H. HIGHLEYMAN, B.E.E., 1955, Rensselaer Polytechnic Institute; M.S., 1957, Massachusetts Institute of Technology; D.E.E., 1961, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. At the Laboratories he first engaged in the problem of character recognition. More recently, he has been concerned with the development of data communication equipment and the study of new devices and techniques for data communication problems. He presently serves as a lecturer at the Polytechnic Institute of Brooklyn. Member Tau Beta Pi, Eta Kappa Nu, Sigma Xi, I.R.E.

JOHN C. IRVIN, B.A., 1949, Miami University; M.A., 1953, and Ph.D., 1957, University of Colorado; Bell Telephone Laboratories, 1957—. He first engaged in the study of properties of silicon, including photoelastic investigations and the variation of resistivity as a function of impurity doping and temperature. For the past two years he has been concerned with the development of varactor diodes, including germanium, silicon and gallium-arsenide models. Member American Physical Society, Phi Beta Kappa, Sigma Xi, Omicron Delta Kappa.

BELA JULESZ, Dipl. in Electrical Engineering, 1950, Budapest (Hungary) Technical University; Kandidat in Technical Sciences, 1956, Hungarian Academy of Sciences; Telecommunication Research Institute (Budapest) 1950–56; Bell Telephone Laboratories, 1956—. He first worked on problems of network theory and microwave systems. At the Laboratories he was first engaged in studies of systems for reducing television bandwidth. At present, Dr. Julesz is working in visual research, particularly on problems of depth perception and pattern recognition. Member I.R.E., A.A.A.S., Psychonomic Society, Optical Society of America.

HENRY KATZ, B.S. in Chemical Engineering, 1948, Drexel Institute of Technology; M.S. in Chemistry, 1955, University of Pennsylvania; Bell Telephone Laboratories, Summer, 1959. Mr. Katz is studying toward the Ph.D. degree at the University of Pennsylvania, where he is working on a crystal structure problem. Member American Chemical Society, American Crystallographic Association.

PAUL KISLIUK, B.S., 1943, Queens College; M.A., 1947, and Ph.D., 1952, Columbia University; Brookhaven Laboratory, 1947–48; Bell Telephone Laboratories, 1952—. At the Laboratories, he has engaged in research in contact and surface physics. His studies have included problems at relay contacts and adsorption of gases on metals. Most recently his work has concerned optical masers. Member American Physical Society, Sigma Xi.

DAVID A. KLEINMAN, S.B., 1946 and S.M., 1947, Massachusetts Institute of Technology; Ph.D., 1952, Brown University; Brookhaven National Laboratory, 1949–53; Bell Telephone Laboratories, 1953—. Mr. Kleinman has worked in the areas of neutron scattering in solids, semiconductor electronics, electron energy bands, and the infrared properties of crystals, and is currently working on problems related to the optical maser. Member American Physical Society.

JOAN E. MILLER, A.B., 1953, Mount Holyoke College; M.A., 1956, Indiana University; Bell Telephone Laboratories, 1957—. Miss Miller has engaged in speech analysis and synthesis, computer simulation of speech transmission, and experiments on depth perception. Member Acoustical Society of America.

J. A. Morrison, B.Sc., 1952, King's College, London University; Sc.M., 1954 and Ph.D., 1956, Brown University; Bell Telephone Laboratories, 1956—. He has been engaged in mathematical research involving mostly differential and integral equations arising in a variety of fields, including electromagnetic problems, multi-velocity electron beams and plasmas, nonlinear diffusion and space charge processes, signal theory and satellite orbits. Member American Mathematical Society, Sigma Xi.

Ian M. Ross, B.A., 1948, Gonville and Caius College; Ph.D., 1952, Cambridge University; Bell Telephone Laboratories, 1952—. He has specialized in the research and development of a wide variety of semiconductor devices. He is director of exploratory and intermediate development of transistors, diodes and other semiconductor components. His laboratory is responsible for the study of radiation damage to semiconductor devices used in satellites, and for the specific design of satellite solar cells. Senior member I.R.E.

David Slepian, University of Michigan, 1941–43; M.A., 1947, and Ph.D., 1949, Harvard University; Bell Telephone Laboratories, 1950—. He has been engaged in mathematical research in communication theory, switching theory, and theory of noise, as well as various aspects of applied mathematics. He has been mathematical consultant on a number of Laboratories' projects. During the academic year 1958–59, he was Visiting Mackay Professor of Electrical Engineering at the University of California at Berkeley. Member A.A.A.S., American Mathematical Society, Institute of Mathematical Statistics, I.R.E., Society of Industrial and Applied Mathematics, U.R.S.I. Commission 6.

Hans-Georg Unger, Dipl. Ing., 1951 and Dr. Ing., 1954, Technische Hochschule, Braunschweig (Germany); Siemens and Halske (Germany), 1951–55; Bell Telephone Laboratories, 1956—. His work at Bell Laboratories has been in research in waveguides, especially circular electric wave transmission. He is now on leave of absence from Bell Laboratoires while professor of electrical engineering at the Technische Hochschule in Braunschweig. Senior member I.R.E.; member German Communication Engineering Society.

Joseph H. Weber, B.E.E., 1952, Rensselaer Polytechnic Institute M.S.E., 1959, George Washington University; Hazeltine Electronics Corp., 1952–53; U.S. Navy 1953–56; Bell Telephone Laboratories, 1956—. At the Laboratories, he has been engaged in telephone traffic studies and systems engineering of electronic switching systems. He presently heads a group concerned with traffic analysis, programming and simulation for the Universal Integrated Communications System (UNICOM) under development for the U.S. Signal Corps. Member A.I.E.E., I.R.E., Operations Research Society of America, Association for Computing Machinery, Sigma Pi Sigma.