THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XLVI MAY–JUNE 1967 NUMBER 5

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published ten times a year by the American Telephone and Telegraph Company, B. S. Gilmer, President, C. E. Wampler, Vice President and Secretary, J. J. Scanlon, Vice President and Treasurer. Checks for subscriptions should be made payable to American Telephone and Telegraph Company and should be addressed to the Treasury Department, Room 2312C, 195 Broadway, New York, N. Y. 10007. Subscriptions $5.00 per year; single copies $1.25 each. Foreign postage $1.08 per year; 18 cents per copy. Printed in U.S.A.

THE BELL SYSTEM

TECHNICAL JOURNAL

VOLUME XLVI — MAY–JUNE 1967 — NUMBER 5

Unit-Cube Expression for Space-Charge Resistance

By S. M. SZE and W. SHOCKLEY*

(Manuscript received November 17, 1966)

A simple analysis shows that the unit-cube conductance is a figure of merit in semiconductor device design theory. The unit-cube conductance, G, is given by $2Kv_d$ where K is the permittivity of the semiconductor and v_d is the limiting drift velocity.

The space-charge resistance, R_{sc}, due to carrier generated under avalanche condition is derived for p-n junctions. It is found that for parallel-plane structure, $R_{sc} = 1/GN$, where N is the number of unit cubes in the depletion region with cube edge equal to the depletion width or $N = A/W^2$ where W is the depletion width and A the junction area. The disturbance in voltage caused by the space-charge effect is given by $I/GN = JW^2/G$ where I and J are the current and current density, respectively. Similar results are obtained for p-n junctions with coaxial-cylinder and concentric-sphere structures.

For silicon, the value of G is approximately 40 μmhos. The transconductance of a silicon surface-controlled avalanche transistor in terms of the unit-cube expression is about 12.5 N μmhos.

A simple analysis of "avalanche resistance" can be given for the limiting case in which carriers are generated at one boundary surface of the depletion region of a p-n junction and travel across the depletion region with a limiting drift velocity v_d. Structures satisfying these conditions can be of the n^+pp^+ form. It will be shown that the quantity

*Stanford University and Bell Telephone Laboratories.

$2Kv_d$ (where K is the permittivity of the semiconductor) is a figure of merit in semiconductor device design theory which limits the performance of space-charge-limited devices. This quantity is a combination of "material constants" similar to $F_B v_d$ (where F_B is "breakdown field") which limits the frequency-power performance of transistors[1,2] and K/σ (σ is the conductivity) which limits the gain-bandwidth product[3] of solid-state devices.

For a structure in which the space-charge layer is bounded by parallel planes of area A and spacing W, it will be shown that the effective space-charge resistance can be interpreted as due to N unit-cube conductances in parallel, where the unit-cube conductance G is given by

$$G = 2Kv_d \tag{1}$$

and N is number of unit cubes in the depletion layer with cube edge equal to the depletion width, or

$$N = A/W^2. \tag{2}$$

The space-charge resistance is then given by

$$R_{sc} = 1/NG. \tag{3}$$

For coaxial-cylinder and concentric-sphere structures, similar results are obtained for the R_{sc}. The number of unit cubes (or curvilinear cubes), however, depends on the radius of the surface upon which avalanche occurs and the length of the cylinder (for the coaxial-cylinder structure). These functional dependences are derived below.

An interesting application of the space-charge resistance and the unit-cube expression is given for a surface-controlled avalanche transistor (SCAT).[4]

I. PARALLEL PLANE STRUCTURE

As represented in Fig. 1(a) the depletion layer of an n^+pp^+ structure extends through the p layer with a doping of N_a, and is bounded by the planes at $x = 0$ and $x = W$. When the applied voltage V is equal to the breakdown voltage V_B the electric field $E(x)$ has its maximum absolute value F_B at $x = 0$ and decreases to $F_B - (qN_aW/K)$ at $x = W$. This insures breakdown at $x = 0$. Furthermore, if $qN_aW/K < 0.9\,F_B$, then the field is everywhere $\geqq F_B/10$, so that holes have their limiting drift velocity v_d all across W.

The space-charge current, I, is given by

$$I = v_d \rho A, \tag{4}$$

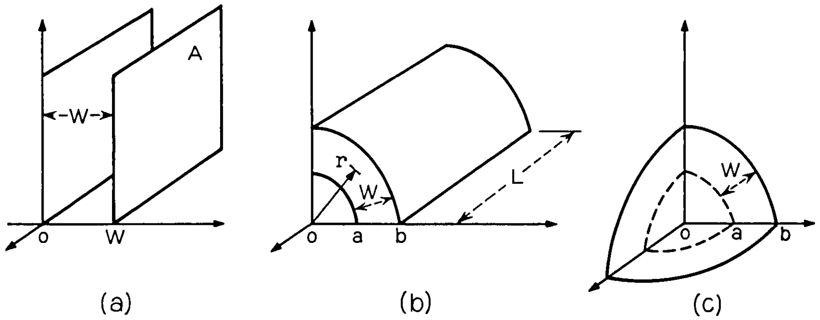


Fig. 1 — p-n junction geometry of (a) parallel plane, (b) coaxial cylinder, and (c) concentric sphere structures.

where ρ is the carrier-charge density and A the area. Since E at $x = 0$ is assumed to be equal to F_B, the disturbance $\Delta E(x)$ in the electric field due to ρ is

$$\Delta E(x) = \frac{Ix}{AKv_d} \tag{5}$$

so that the disturbance in voltage caused by the carriers (i.e., the average field times W) is obtained by integrating $\Delta E(x)$

$$\Delta V_B = I\left(\frac{W^2}{A}\right)\left(\frac{1}{2Kv_d}\right) = \frac{I}{NG}. \tag{6}$$

The total voltage is thus

$$V = V_B + \Delta V_B = \left(F_B W - \frac{qN_a W^2}{2K}\right) + \frac{I}{NG} = V_B + IR_{sc}, \tag{7}$$

which verifies the interpretation of G and N.

II. COAXIAL CYLINDER STRUCTURE

Consider first that the maximum field occurs at the inner surface. As shown in Fig. 1(b) the depletion layer extends through the intrinsic region of an n^+ip^+ coaxial-cylinder structure and is bounded by the cylinders of radii $r = a$ and $r = b$. When $V = V_B$, the electric field $E(r)$ has its maximum absolute value F_B at $r = a$, and decreases to $F_B a/b$ at $r = b$.

The space-charge current per unit length, I/L is given by

$$\frac{I}{L} = 2\pi r \rho v_d \tag{8}$$

so that ρ varies as $1/r$. Integrating Poisson's equation leads to a disturbance $\Delta E(r)$ in the electric field and ΔV_B in the voltage due to ρ given by

$$\Delta E(r) = \frac{\rho(r - a)}{K} = \frac{I}{2\pi K v_d L}\left(1 - \frac{a}{r}\right) \qquad (9)$$

and

$$\Delta V_B = \frac{I}{(2Kv_d)\pi L}\left[b - a - a \ln\left(\frac{b}{a}\right)\right] \equiv \frac{I}{NG} = IR_{sc}\,, \qquad (10)$$

where

$$\frac{1}{N} = \frac{b}{\pi L}\left(1 - \frac{a}{b} - \frac{a}{b}\ln\frac{b}{a}\right) = \frac{\left[2b^2\left(1 - \frac{a}{b} - \frac{a}{b}\ln\frac{b}{a}\right)\right]}{2\pi bL} \equiv \frac{A_c}{2\pi bL}. \qquad (11)$$

A_c is the area on the outer cylinder surface that corresponds to one unit-cube conductance G.

$$\begin{aligned} A_c &\rightarrow (b - a)^2, && \text{for} \quad a \rightarrow b \\ A_c &\rightarrow 2b^2, && \text{for} \quad a \rightarrow 0. \end{aligned} \qquad (12)$$

Equation (10) may be interpreted as the resistance of N unit curvilinear cubes in parallel. These cubes are formed by intersection of equipotential surfaces with the orthogonal family of electric field lines. Each cube has a conductance $(2Kv_d)$, and the number of cubes N is given by (11). The area A_c approaches $(b - a)^2$ when $a \rightarrow b$, and approaches $2(b - a)^2$ when $a \rightarrow 0$, and consequently remains finite even as the inner cylinder approaches a line.

The maximum field may be caused to occur on the outer surface $r = b$ by adjusting the chemical charges in the depletion layer appropriately, such as a p^+pn^+ structure with the pn^+ junction at $r = b$, the p^+p boundary at $r = a$. In this case, the area A_c approaches $(b - a)^2$ when $a \rightarrow b$, and approaches $2b^2 \ln (b/a)$ when $a \rightarrow 0$. Hence, the space-charge resistance has the same value as given by (10) and (11) when $a \rightarrow b$, but approaches infinity as $a \rightarrow 0$.

III. CONCENTRIC SPHERE STRUCTURE

As shown in Fig. 1(c), the depletion layer extends through the intrinsic region of an n^+ip^+ concentric sphere structure and is bounded by the spheres of radii $r = a$ and $r = b$. When $V = V_B$, the electric field $E(r)$ has its maximum value F_B at the inner surface $r = a$, and decreases to $F_B a^2/b^2$ at $r = b$.

The space-charge current is given by

$$I = 4\pi r^2 \rho v_d \ . \tag{13}$$

The quantities $\Delta E(r)$ and ΔV_B are given as follows:

$$\Delta E(r) = \frac{\rho(r-a)}{Kr^2} = \frac{I}{4\pi K v_d r}\left(1 - \frac{a}{r}\right) \tag{14}$$

$$\Delta V_B = \frac{I}{(2Kv_d)2\pi}\left(\ln\frac{b}{a} - 1 + \frac{a}{b}\right) \equiv \frac{I}{NG} = IR_{sc} \ , \tag{15}$$

where

$$\frac{1}{N} = \frac{\left(\ln\frac{b}{a} - 1 + \frac{a}{b}\right)}{2\pi} \equiv \frac{A_s}{4\pi b^2} \tag{16}$$

$$A_s = 2b^2\left(\ln\frac{b}{a} - 1 + \frac{a}{b}\right)$$

$$\begin{aligned} A_s &\to (b-a)^2, \quad &&\text{for} \quad a \to b \\ A_s &\to \infty, &&\text{for} \quad a \to 0 \end{aligned} \tag{17}$$

and A_s is quantity of area on the outer sphere surface that corresponds to one curvilinear unit-cube conductance G. When a is finite or $a \to b$, (15) may be interpreted as the resistance of N unit curvilinear cubes in parallel, where each cube has a conductance $(2Kv_d)$ and the number of cubes N is given by (16). Unlike the n^+ip^+ coaxial-cylinder case, R_{sc} of the concentric-sphere structure approaches infinite resistance as the inner sphere approaches a point.

F_B can occur on the outer sphere for a p^+pn^+ structure, for example. The results of A_s are the same for both limiting cases as given in (17).

An interesting application of the unit-cube expression is that for a surface-controlled avalanche transistor (SCAT)[4] with a total junction perimeter of P and a space-charge layer W. Because the space-charge resistance is finite for an n^+ip^+ coaxial cylinder structure as the inner cylinder approaches a line [see (12)], it is feasible to make calculations for SCAT on the basis of an avalanche line source. There are $N = P/W$ such unit cubes around the edge, and the total transconductance, g_m, is[4]

$$g_m = \left(\frac{2}{\pi}\right) K v_d \left(\frac{P}{W}\right) , \tag{18}$$

Table I—Unit-Cube Expressions ($W \equiv b - a$)

Structures	Parallel plane	Coaxial cylinders		Concentric spheres
N	$\frac{A}{W^2}$	$a \rightarrow b$	$a \rightarrow 0$	$a \rightarrow b$
(No. of unit cubes)		$\frac{2\pi Lb}{W^2}$	$\frac{\pi L}{b}$	$\frac{4\pi b^2}{W^2}$
Cube edge (cm)	W	W	$\sqrt{2b}$	W
R_{sc} (ohms)	$\frac{1}{GN} = \frac{1}{(2Kv_d)N}$			

or

$$g_m = \left(\frac{1}{\pi}\right)(2Kv_dN) \equiv \frac{1}{\pi R_{sc}}. \tag{19}$$

For a silicon SCAT with a device geometry of $P = 1000\ \mu$ and $W = 0.5\ \mu$, there are 2000 unit cubes with cube edge $0.5\ \mu$. The space-charge resistance is 12.5 ohms, and the transconductance is 25,400 μmhos.

Another application is to calculate the voltage disturbance ΔV_B in a Read diode.[5] For a silicon Read diode with drift region of 10 μm and an operating current density of 1000 amp/cm^2, the value of ΔV_B, as obtained from (6), is approximately 25 volts.

A summary of the number of unit cubes and other pertinent quantities is presented in Table I. It has been shown that the unit-cube conductance ($2Kv_d$) is a figure of merit in semiconductor device design theory. The unit-cube expressions are shown to be useful for calculation of the space-charge resistance.

REFERENCES

1. Johnson, E. O., Physical Limitation on Frequency and Power Parameters of Transistors, IEEE International Convention Record, Part 5, March, 1965, pp. 27–34. Also appearing in RCA Review, June, 1965, pp. 163–177.
2. De Loach, B. C., Recent Advances in Solid State Microwave Generators, a chapter of *Advances in Microwaves,* to be published by Academic Press and edited by Leo Young.
3. Rose, A., An Analysis of the Gain-Bandwidth Limitations of Solid State Triodes, RCA Review, December, 1963, pp. 627–640.
4. Shockley, W. and Hooper, W. W., The Surface Controlled Avalanche Transistors, Wescon Meeting, Los Angeles, August, 1964.
5. Read, W. T., A Proposed High-Frequency, Negative-Resistance Diode, B.S.T.J., *37,* March, 1958, pp. 401.

Comparison of *M*-ary Modulation Systems

By IRA JACOBS

(Manuscript received September 22, 1966)

Consideration of large alphabet digital communication systems is of both theoretical and practical interest. Although performance bounds on optimum systems for the Gaussian channel are available, constructive methods for approaching these bounds are unknown, except in a few very special cases. Specific systems have been proposed and evaluated relative to these bounds, but exact evaluation of error probability is generally a difficult numerical task. It is of interest to consider simpler performance criteria which permit comparison of various systems without extensive computation.

An easily evaluated criterion (based on the alphabet size and minimum distance between signal vectors) is shown to yield a simple sufficient condition for one system to be better than another (smaller error probability for the same energy-per-bit). The criterion is applied to orthogonal, biorthogonal, simplex, and more general permutation modulation systems. In addition to comparing the various systems, we consider ways of obtaining good special cases of permutation modulation. Finally, we assess a recently proposed system ("N-orthogonal phase modulation") and show that it is generally inferior to more conventional techniques.

I. INTRODUCTION

The choice of waveforms for communicating over the Gaussian additive noise channel is a classic problem in communication theory. Orthogonal modulation systems (i.e., digital communications in which the alphabet consists of orthogonal waveforms) are known to result in good power efficiency at the expense of poor bandwidth utilization.[1,2] As the alphabet size M is increased, the energy-per-bit E required to achieve a given error probability P_ϵ diminishes, but the information rate to bandwidth ratio (R/W) diminishes even more rapidly. Biorthogonal and simplex modulation afford somewhat improved performance, but are likewise restricted to low values of R/W.

There is considerable interest in finding large alphabet systems which have both good power efficiency and good bandwidth utilization.

Slepian[3] has given bounds on what can be achieved, but constructive techniques for approaching these bounds are generally unknown.

Although computer evaluation is ultimately required for precise knowledge of error probability, it is of interest to consider simpler performance criteria which permit at least a qualitative comparison of various systems without extensive computation. It is the purpose of this paper to demonstrate the utility of the latter approach.

After defining the problem more precisely in Section II, some well-known bounds on the error probability are employed in Section III to obtain a simple analytic criterion for comparing systems in the limit of low P_e. In Section IV this criterion is applied to systems (PSK, FSK, biorthogonal, and simplex) for which extensive exact computations are available and for which the conclusions drawn are already well-known. After these illustrative examples, permutation modulation[4] is considered in Section V and N-orthogonal phase modulation[5,6] in Section VI. It is shown that the former can yield better performance than conventional techniques, but that the latter is generally inferior. Finally, in Section VII limits on our performance criterion, obtained from sphere-packing arguments, are presented.

II. COMMUNICATION SYSTEM MODEL

We consider an M-ary modulation system of equienergy waveforms $S_i(t)$, $i = 1, \cdots, M$, on $(0,T)$, having the correlation matrix

$$\rho_{ij} = \frac{1}{E_s} \int_0^T S_i(t)S_j(t)\,dt. \qquad (1)$$

E_s is the energy of each waveform so that $\rho_{ii} = 1$, and $-1 \leqq \rho_{ij} \leqq 1$. It is conventional[3,7] to define a normalized information rate, $(2 \log_2 M)/n$, where $n \leqq M$ is the rank of the correlation matrix (dimensionality of the signal space). We choose to call this normalized rate the "information to bandwidth ratio, R/W" motivated by the relations

$$\frac{R}{W} = \frac{\log_2 M}{WT} = \frac{2 \log_2 M}{n}, \qquad (2)$$

where the second equality follows if we set $n = 2TW$, which is at least partially justified for large n by the work of Pollak and Landau.[8] For our purposes, the right-hand side of (2) may be considered as the definition of R/W.

It will be assumed that in addition to $\rho_{ii} = 1$ that each row of the correlation matrix can be written as a permutation of the first row. Considering the waveforms as vectors in an n-dimensional linear vector space, this means that each waveform sees an identical environment of neighboring waveforms. This restriction is a desirable one if it is desired to transmit each waveform with equal *a priori* probability. The restriction is satisfied by the various modulation systems mentioned in the introduction.* Slepian[9] has termed such systems "group codes for the Gaussian channel."

It is assumed that the receiver observes a waveform $z(t)$ on the interval $(0,T)$

$$z(t) = S_i(t) + n(t), \tag{3}$$

where $n(t)$ is a sample function from a white Gaussian noise process of spectral density N_o ; i.e.,

$$\langle n(t)n(t')\rangle = \frac{N_o}{2}\,\delta(t' - t). \tag{4}$$

On the basis of this observation we wish to decide with minimum probability of error (P_e) which of the M waveforms was transmitted. The optimum (minimum P_e) receiver is known[10] to consist of M matched filters which give

$$z_j = \int_0^T z(t)S_j(t)\,dt = E_s(\rho_{ij} + x_j), \tag{5}$$

where

$$x_j = \frac{1}{E_s}\int_0^T n(t)S_j(t)\,dt \tag{6}$$

and decision that the kth waveform was transmitted is made if $z_k > z_j$ for all $j \neq k$; i.e., the decision is made on the basis of the largest matched filter output.

From (6), the x_j are zero-mean Gaussian variates with covariance

$$\begin{aligned}\langle x_j x_k\rangle &= \frac{1}{E_s^2}\int_0^T\int_0^T dt\,dt'\,\langle n(t)n(t')\rangle S_j(t)S_k(t')\\ &= \frac{N_o}{2E_s}\,\rho_{jk}\,.\end{aligned} \tag{7}$$

* The only commonly employed M-ary system (known to the author) which does not satisfy this restriction is M-level amplitude modulation.

The error probability of this system is given by*

$$P_e = 1 - \int_{\Omega_i} \cdots \int dx_1 \cdots dx_M \, p(x_1 , \cdots x_M), \tag{8}$$

where $p(x_1 , \cdots , x_M)$ is the multi-variate zero-mean Gaussian distribution with covariance given by (7), and the region of integration Ω_i is defined by the condition

$$\Omega_i = \text{region in which } 1 + x_i > \rho_{ij} + x_j \text{ for all } j \neq i.$$

Clearly P_e is a function of M parameters: E_s/N_o , ρ_{12} , ρ_{13} , $\cdots$, ρ_{1M} , the first of which is a signal-to-noise ratio, the remainder of which describe the correlation properties of the modulation system.

Landau and Slepian[11] have proved the long-conjectured result that P_e is minimized for a given M (but n unrestricted) by the simplex configuration in which the correlation matrix has the form†

$$\rho_{ij} = \begin{cases} 1 & i = j \\ -\dfrac{1}{M-1} & i \neq j \end{cases} \quad \text{simplex.} \tag{9}$$

The rank of this matrix is $n = M - 1$ so that

$$(R/W)_{\text{simplex}} = \frac{2 \log_2 M}{M - 1}. \tag{10}$$

For this case the expression for P_e may be reduced to a single integral[10] and numerical results are readily obtained.[13]

Weber[14] has derived locally optimum configurations when $M/2 \leqq n \leqq M - 1$. For $n = M/2$ a local optimum is the biorthogonal configuration in which the signal vectors are located along the coordinate axes (+ and −) of the n-dimensional vector space such that

$$\rho_{ij} = \begin{cases} 1 & i = j \\ -1 & i = j - (-1)^i \\ 0 & i \neq j, j - (-1)^i \end{cases} \quad \text{biorthogonal.} \tag{11}$$

* Actually this is the error probability assuming the ith signal is transmitted. However, under the assumption of equal *a priori* transmission of all signals, and the permutation property assumed for the correlation matrix, this probability is independent of i and is equal to the system probability of error.

† The "local optimality" of the simplex configuration (*viz.*, that P_e has a local minimum) had been proved previously by Balakrishnan.[12]

The rank of this $M \times M$ matrix is $M/2$ so that

$$\left(\frac{R}{W}\right)_{\text{biorthogonal}} = \frac{4 \log_2 M}{M}. \tag{12}$$

In this case P_e may also be expressed as a single integral which is readily evaluated by machine techniques. Although for a given value of M, biorthogonal modulation requires slightly more energy-per-bit to achieve a given P_e than simplex,* it is noted that (for large M) R/W for biorthogonal is essentially twice that of simplex. Furthermore, for biorthogonal half of the waveforms are the negatives of the remaining half; consequently, $M/2$ rather than M matched filters are required. For these reasons biorthogonal is generally preferred to simplex, and indeed has been employed for deep-space communications.[15]

The disadvantage of both simplex and biothogonal modulation is that good power efficiency is associated with large values of M (as it must be for any modulation system) which from (10) and (12) imply small values of R/W. Weber's[14] results indicate locally optimum systems with R/W between simplex and biorthogonal, but these are then also restricted to relatively small R/W.

Optimum systems (in the sense of minimum P_e) are not known for $n < M/2$. However, bounds on the error probability of optimum systems have been obtained[7] and evaluated.[3] (The upper bound is obtained by random coding arguments, and the lower bound by sphere packing arguments.) These bounds are extremely useful in assessing the performance of specific systems; however, to do so involves explicit evaluation of P_e for the specific systems of interest. This is at best a difficult numerical task. Furthermore, we may find in comparing two systems that one is better if we are interested in $P_e \approx 10^{-3}$, whereas the reverse is true when $P_e \approx 10^{-6}$. Also, in comparing systems with different values of M it may be unrealistic to compare P_e, since P_e is the word error probability, and the systems contain a different number of bits per word. Comparison on the basis of bit error probability involves a difficult conversion from word to bit error probability which involve coding arguments separate from the modulation system performance.[16] For all of these reasons it is desirable to find a simpler criterion than P_e which permits at least a gross comparison of modulation systems.

* If the comparison is made for a fixed R/W rather than M then biorthogonal requires less energy per bit. The simple unqualified statement that simplex is the optimum modulation system is misleading.

III. BOUNDS ON ERROR PROBABILITY

One approach to comparing modulation systems is to obtain lower and upper bounds on the true error probability*

$$P_l \leqq P_e \leqq P_u \tag{13}$$

and to say that system 1 is better than system 2 if $P_{u1} < P_{l2}$.

If two systems are close in performance, the above procedure may not enable us to determine which is better unless the bounds are close. On the other hand, close bounds may be difficult to evaluate and may not lead to a simple performance criterion. We adopt the viewpoint here that it is desirable to have bounds, which although quite loose, lead to a simple sufficient condition for determining when one system is better than another.

Let

$$\rho = \max_{j \neq i} \rho_{ij} = \max_{j > 1} \rho_{1j} . \tag{14}$$

That is, ρ is the largest non-diagonal entry of the correlation matrix. It is readily established that†

$$\Phi\left(-\sqrt{\frac{E_s}{N_o}(1-\rho)}\right) \leqq P_e \leqq (M-1)\Phi\left(-\sqrt{\frac{E_s}{N_o}(1-\rho)}\right), \tag{15}$$

where

$$\Phi(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} dy \exp(-y^2/2). \tag{16}$$

The lower bound is obtained by observing that P_e for an M-ary system can be no less than that of the binary system containing nearest neighbor waveforms. The upper bound follows from

$$P_e \leqq \sum_{j=2}^{M} \Phi\left(-\sqrt{\frac{E_s}{N_o}(1-\rho_{ij})}\right) \leqq (M-1)\Phi\left(-\sqrt{\frac{E_s}{N_o}(1-\rho)}\right) \tag{17}$$

where the first inequality in (17) is a consequence of the symmetry property of the system and the fact that the probability of a union of events is less than the sum of the probabilities of the events. The

* We consider here bounds on word error probability, which, however, may be easily converted to bounds on bit error probability. For example, if a word is in error at least one bit is in error, and at most all bits are in error. Hence, $P_l/\log_2 M$ and P_u are lower and upper bounds on the bit error probability.

† These bounds are generally well known and appear widely in the literature; e.g., Refs. 7, 17, 18, 19. Also, as noted in the previous footnote, these bounds on word error probability are readily converted into bounds on bit error probability.

second inequality in (17) follows simply by observing that a sum of $(M - 1)$ terms is no greater than $(M - 1)$ times the largest term.

In comparing modulation systems with different alphabet size it is more appropriate to consider the energy per bit E rather than the signal energy E_s, where

$$E = E_s/\log_2 M. \tag{18}$$

Indeed, the parameter E/N_0 is an appropriate measure of the power efficiency of a modulation system. The Shannon channel capacity formula requires that $E/N_0 > \log_e 2$ to achieve arbitrarily small P_e, conventional systems generally require values of E/N_0 at least 4 times the Shannon minimum.[2]

In terms of the parameter E/N_0 the error probability bounds may be rewritten

$$\Phi\left(-\sqrt{\frac{E}{N_0}K}\right) \leqq P_e \leqq (M - 1)\Phi\left(-\sqrt{\frac{E}{N_0}K}\right), \tag{19}$$

where

$$K \equiv (1 - \rho)\log_2 M. \tag{20}$$

We will say that system 1 is "better" than system 2 if

$$(M_1 - 1)\Phi\left(-\sqrt{\frac{E}{N_0}K_1}\right) < \Phi\left(-\sqrt{\frac{E}{N_0}K_2}\right). \tag{21}$$

Several conclusions are apparent from (21).

(*i*) In the limit of large E/N_0, $K_1 > K_2$ is sufficient to ensure that system 1 is better than system 2. (If $K_1 > K_2$ we will say that system 1 is "asymptotically better" than system 2.)

(*ii*) If system 1 is asymptotically better than system 2, then there exists a value of E/N_0 above which system 1 is better than system 2. Below this value of E/N_0 our formalism is generally inadequate to determine which system is better. (The critical value of E/N_0 may be obtained by replacing the inequality in (21) by an equality.)

(*iii*) A binary system that is asymptotically better than an M-ary system is always better than the M-ary system.

Thus, we can always determine quite simply which of two systems is asymptotically better, and may, in many special cases, be able to make comparisons at specific E/N_0 of interest.

It should be emphasized that the above comparison is on the basis of the P_e obtained with the two systems when operated at the same average power and information rate. To complete the comparison, the bandwidth requirements of the two systems should also be considered. Thus, the parameter R/W, as well as K should be used in comparing systems.

In the following sections of this paper specific systems will be considered and represented by points on a K, R/W plot. This will enable an immediate comparison of the asymptotic performance of systems having the same R/W. It should be noted that Gilbert[17] used a similar plot in his 1952 paper which addressed the same subject considered here. Gilbert employed a (SNR, R/W) plot in which the effective signal-to-noise ratio (SNR) was obtained for a given P_e by using the upper bound in (19). Since the SNR is related to our $1/K$, better systems correspond to smaller SNR. Our purpose in writing this paper is *not* to argue that our plot is a better way to present the results than Gilbert's. (Indeed, since in general, P_e is much closer to the upper than to the lower bound, his method of comparison is somewhat better, although somewhat less convenient to use.) Our purpose rather is to resurrect these old methods which have been largely discarded since the advent of high-speed computation, and to illustrate their applicability to recently proposed modulation systems.

IV. PHASE, FREQUENCY, BIORTHOGONAL AND SIMPLEX MODULATIONS

4.1 *Phase-Shift Modulation*

For M phasors uniformly spaced on the unit circle, $\rho = \cos 2\pi/M$. Therefore,

$$K = 2 \log_2 M \sin^2 \frac{\pi}{M} \cdot \tag{22}$$

Note that $K = 2$ for both $M = 2$ and $M = 4$* and falls off thereafter.

Since the dimensionality of the signal space is $n = 1$ for $M = 2$ and $n = 2$ for $M > 2$, it follows that R/W is given by

$$\frac{R}{W} = \begin{cases} 2 & \text{for} \quad M = 2, \\ \log_2 M & \text{for} \quad M > 2. \end{cases} \tag{23}$$

* K is maximized (for integer M) when $M = 3$. In practice, it is generally desirable to consider only those values of M which are integer powers of 2 (i.e., each symbol conveys an integer number of bits). We shall restrict our numerical examples to such cases.

Table I—Phase-Shift Modulation

M	K	R/W
2	2	2
4	2	2
8	0.88	3
16	0.30	4
32	0.098	5
64	0.030	6

Table I lists the K and R/W values for phase-shift modulation, and these are denoted by dots in Fig. 1. It is apparent that $M = 2$ and $M = 4$ are asymptotically better than the higher-order systems, and from our previous results this implies that the binary system is always better than the general M-ary case with $M > 4$.* Recall that we are

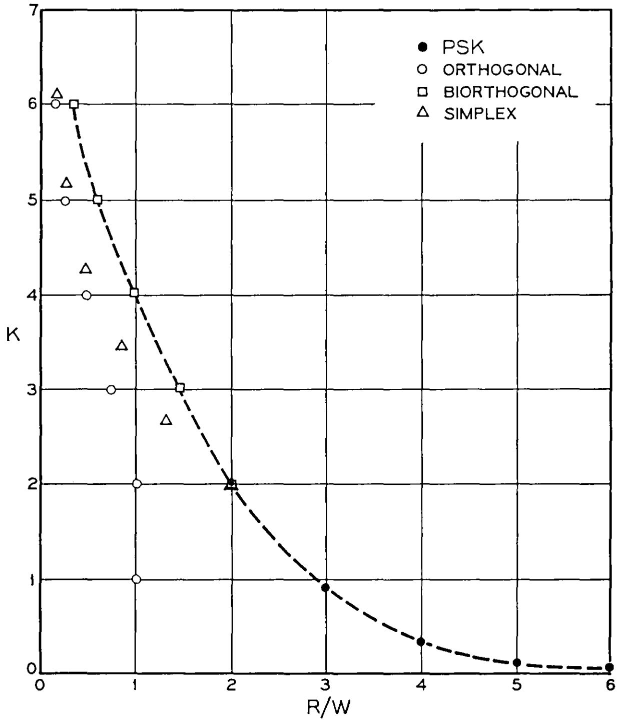


Fig. 1 — $(K,R/W)$ plot for phase-shift, orthogonal, biorthogonal, and simplex modulations.

consistently using the term "better" to mean smaller P_e for a given E/N_0. Large alphabet phase modulation may still be desirable because of the larger R/W.

4.2 *Frequency Shift (Orthogonal) Modulation*

For M orthogonal signals (e.g., frequency-shifted signals with essentially non-overlapping spectra), $\rho = 0$ and

$$K = \log_2 M. \tag{24}$$

The dimensionality of the signal space is the number of orthogonal vectors, $n = M$, so that

$$\frac{R}{W} = \frac{2 \log_2 M}{M}. \tag{25}$$

Table II lists the K and R/W values for orthogonal modulation, and these are denoted by circles in Fig. 1. Larger values of M correspond to systems which are asymptotically better, at the expense, however, of smaller values of R/W. It is clear that binary orthogonal is inferior to binary and quarternary PSK both in terms of a smaller K and smaller R/W.[†]

TABLE II—ORTHOGONAL MODULATION

M	K	R/W
2	1	1
4	2	1
8	3	3/4
16	4	1/2
32	5	5/16
64	6	3/16

4.3 *Biorthogonal Modulation*

A biorthogonal system consists of $M/2$ orthogonal waveforms and their negatives. The maximum correlation coefficient is $\rho = 0$ for $M \geqq 4$, but $\rho = -1$ for $M = 2$. Therefore,

$$K = \begin{cases} 2 & \text{for } M = 2, \\ \log_2 M & \text{for } M \geqq 4 \quad (M \text{ even}). \end{cases} \tag{26}$$

[*] This conclusion is confirmed by the exact calculations of P_e for M-ary PSK by C. R. Cahn.[20]

[†] Binary FSK may still be employed, of course, for simplicity reasons or because the channel phase coherence may not be consistent with phase-shift modulation.

Since $n = M/2$,

$$\frac{R}{W} = \frac{4 \log_2 M}{M}. \tag{27}$$

Table III lists the K and R/W values for biorthogonal modulation, and these are denoted by □'s in Fig. 1. Note that $M = 2$ and $M = 4$ biorthogonal are equivalent, respectively, to binary and quarternary PSK.

Clearly, for fixed R/W, biorthogonal is asymptotically better than orthogonal. For example, consider $M = 4$ orthogonal and $M = 16$ biorthogonal, both of which have $R/W = 1$. From (21), the biorthogonal system is better than the orthogonal system for all $E/N_o > 2.5$, which corresponds to all P_e of practical interest. ($P_e < 3(10)^{-2}$).

4.4 *Simplex Modulation*

In simplex modulation, the M code vectors form a regular simplex in $M - 1$ dimensions. (All vectors are equally spaced from all other vectors. This corresponds to an equilateral triangle in two dimensions, and a regular tetrahedron in three dimensions.) All correlation coefficients are equal and are given by[10,12,13] $\rho = -1/(M - 1)$. Therefore,

$$K = \frac{M}{M - 1} \log_2 M. \tag{28}$$

Since $n = M - 1$,

$$\frac{R}{W} = \frac{2 \log_2 M}{M - 1}. \tag{29}$$

Comparison of (28), (29) with (24), (25) indicates that for large M simplex modulation is essentially identical to orthogonal modulation. Table IV lists the K and R/W values for simplex modulation, and these are denoted by △'s in Fig. 1. A quick glance at Fig. 1 indicates

TABLE III—BIORTHOGONAL MODULATION

M	K	R/W
2	2	2
4	2	2
8	3	3/2
16	4	1
32	5	5/8
64	6	3/8

that depending on the R/W of interest, biorthogonal or PSK modulation offers the best asymptotic performance of the systems considered so far. (The dashed line in Fig. 1 is drawn through these "best" points.) Note that although simplex provides the largest K for a fixed value of M, it does not do so for fixed R/W.*

TABLE IV—SIMPLEX MODULATION

M	K	R/W
2	2	2
4	2.67	1.33
8	3.43	0.86
16	4.26	0.53
32	5.16	0.32
64	6.10	0.19

V. PERMUTATION MODULATION

Slepian[4] has recently described an exceedingly general modulation system (permutation modulation) for which all of the systems considered in the previous section are special cases. The optimum demodulation algorithm is particularly simple, but the actual evaluation of P_e, and the finding of good special cases is somewhat more complex. We restrict ourselves here to a special subclass of permutation modulation. This subclass is suggested both as the simplest generalization of biorthogonal systems, and because perusal of Slepian's results indicate that systems taken from this subclass are amongst the better of the moderate-sized alphabet examples which he considers.

Following Slepian we define an (n,m) permutation modulation system as follows. The time interval T is divided into n subintervals ($n = 2TW$). The first waveform of the alphabet consists of a signal with amplitude unity in the first m subintervals ($m < n$), and zero amplitude in the remaining subintervals. The remainder of the waveforms consist of all possible permutations of the subintervals, allowing also all combinations of plus and minus amplitudes. For example, the (3,2) system contains twelve waveforms which we may represent as

$$(1,1,0),\ (1,-1,0),\ (-1,1,0),\ (-1,-1,0),$$
$$(1,0,1),\ (1,0,-1),\ (-1,0,1),\ (-1,0,-1),$$
$$(0,1,1),\ (0,1,-1),\ (0,-1,1),\ (0,-1,-1).$$

* For the special case $M = 2$, simplex, biorthogonal and PSK are all equivalent.

In general, it is easily seen that the alphabet size M is given by

$$M = 2^m \binom{n}{m}. \tag{30}$$

It is also noted that the special case $(n,1)$ corresponds to biorthogonal modulation.*

This (n,m) modulation clearly satisfies the symmetry requirements of our theory. All members of the alphabet have equal energy† and the correlation matrix has the desired permutation property. It is readily seen that the maximum correlation coefficient is given by

$$\rho = \frac{m-1}{m}. \tag{31}$$

Thus,

$$\begin{aligned} K &\equiv (\log_2 M)(1 - \rho) \\ &= 1 + \frac{1}{m} \log_2 \binom{n}{m} \end{aligned} \tag{32}$$

so that (n,m) modulation always achieves $K > 1$. Also

$$\begin{aligned} \frac{R}{W} &\equiv \frac{2 \log_2 M}{n} \\ &= \frac{2m}{n} K. \end{aligned} \tag{33}$$

Equations (32) and (33) suggest that (n,m) modulation may achieve both large values of K and large R/W, which was not possible with any of the systems described in the previous section.

5.1 $(n,2)$ *Modulation*

Since $m = 1$ leads to biorthogonal modulation which has many desirable properties, it is natural to look next at the special case $m = 2$. From (32) and (33) it follows that for $(n,2)$,

$$K = \tfrac{1}{2}[1 + \log_2 n(n-1)] \qquad (n \geqq 3) \tag{34}$$

and

$$\frac{R}{W} = \frac{4K}{n}. \tag{35}$$

* In Slepian's terminology, the (n,m) modulation described here is a variant II system in which $m_1 = n - m$, $m_2 = m$ and $\mu_1 = 0$, $\mu_2 = 1$.

† With the normalization employed above, the signal energy is m. However, all code words may be multiplied by a constant to achieve any desired E_s.

values of the K and R/W are given in Table V and are plotted as ■'s in Fig. 2. (For reference, Fig. 2 also contains the biorthogonal and PSK results from Fig. 1.) Thus, similar to biorthogonal, as n becomes large K increases but R/W decreases. It is seen from Fig. 2 that $(n,2)$ modulation gives better performance (larger K for a given R/W) than biorthogonal or PSK.*

5.2 *(2m,m) Modulation*

(2,1) corresponds to $M = 4$ biorthogonal, which from our earlier results gives $K = 2$, $R/W = 2$. It is seen from Table V that (4,2) gives $K = 2.30$, $R/W = 2.30$ which corresponds to both better asymptotic performance and better bandwidth utilization. It is apparent from (33) that whenever $n = 2m$, $R/W = K$, and an immediate question is how large can we make these two quantities.

TABLE V — $(n,2)$ MODULATION

n	$M = 2n(n-1)$	K	R/W
3	12	1.79	2.38
4	24	2.30	2.30
5	40	2.66	2.13
6	60	2.95	1.97
7	84	3.19	1.82

With $n = 2m$, it follows from (32) that

$$K = 1 + \frac{1}{m}\log_2\binom{2m}{m}. \tag{36}$$

Use of Stirling's approximation when $m \gg 1$ gives

$$\binom{2m}{m} \approx \frac{1}{\sqrt{\pi m}}\, 2^{2m} \tag{37}$$

so that for large m, $K \rightarrow 3$. It is easily shown that K increases monotonically towards this asymptotic value as m is increased. Thus, $(2m, m)$ modulation does not permit attainment of arbitrarily large values of K, and hence cannot attain arbitrarily low P_e with finite E/N_0. This is consistent with Slepian's statement[4] that permutation modulation cannot approach channel capacity arbitrarily closely at non-zero R/W.

* This is, of course, achieved only at the expense of a larger alphabet size. It may also be noted from Table V that the alphabet size is not generally a power of 2 which may also be a practical disadvantage.

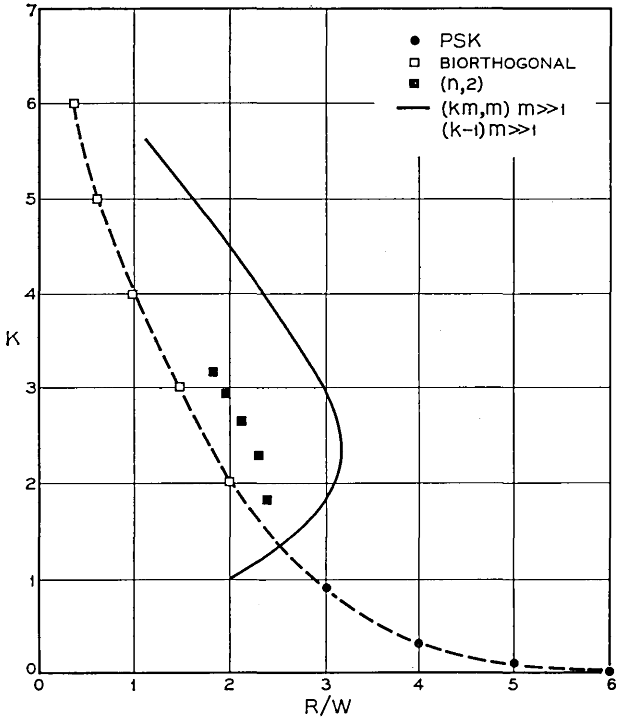


Fig. 2 — $(K,R/W)$ plot for permutation modulation.

5.3 (km,m) *Modulation* $(k > 1)$

As an immediate generalization of the above, consider the more general case $n = km$ where $k > 1$.* Then, from (33)

$$\frac{R}{W} = \frac{2}{k} K \tag{38}$$

and from (32)

$$K = 1 + \frac{1}{m} \log_2 \binom{km}{m}. \tag{39}$$

Again using Stirling's approximation for large m, (assuming also that $(k - 1)\, m \gg 1$)

$$\binom{km}{m} \approx \frac{1}{\sqrt{2\pi m}} \sqrt{\frac{k}{k-1}} \left(\frac{k}{k-1}\right)^{km} (k - 1)^m. \tag{40}$$

[*] Of course k should be chosen so that km is an integer.

Thus, in the limit of large m, for fixed $k > 1$,

$$K \to 1 + k \log_2 \left(\frac{k}{k - 1}\right) + \log_2 (k - 1). \tag{41}$$

For large k, the right-hand-side of (41) increases as log k; however, as seen from (38) R/W diminishes as k^{-1}. The locus of K, R/W values obtained with different values of k (but large m so that the approximation (41) applies) is shown by the solid curve in Fig. 2. In the limit as $k \to 1$ (but m always sufficiently large such that $(k - 1)\, m \gg 1$), $K \to 1$ and $R/W \to 2$. As k increases, both R/W and K increase until $k \approx 1.5$ at which point $R/W \approx 3.2$ and $K = 2.3$. Further increases in k result in a reduction in R/W but continued increase in K.

The above results indicate that (n,m) codes can be found with R/W as large as octary PSK ($R/W = 3$) and with considerably better asymptotic performance.

VI. COMBINED PHASE-SHIFT AND ORTHOGONAL MODULATION

In the previous examples we have compared by approximate methods modulation systems which have already been analyzed exactly. Although perhaps additional insight into the relative performance of these systems has been obtained, many of our conclusions may be inferred from existing exact calculations. We now wish to consider a new system, recently proposed by Reed and Scholtz,[5,6] which (to our knowledge) has not yet been evaluated numerically.

Consider an alphabet M divided into M_f groups, each group containing M_p members. Thus,

$$M = M_f M_p \,. \tag{42}$$

The different groups may be considered to be sufficiently separated in frequency so that waveforms from different groups are orthogonal. Within a group the waveforms have the correlation properties associated with phase-shift modulation. Thus for $M_p \geqq 4$ the maximum correlation coefficient is $\rho = \cos\,(2\pi/M_p)$, and

$$K = 2 \sin^2 \frac{\pi}{M_p} (\log_2 M_p + \log_2 M_f). \tag{43}$$

Since each group requires a two-dimensional sub-space

$$n = 2M_f \,. \tag{44}$$

Thus,

$$R/W = \frac{\log_2 M_f}{M_f} + \frac{\log_2 M_p}{M_f}. \tag{45}$$

In the special case of $M_f = 1$ it is apparent that this system reduces to simple phase-shift modulation (Section 4.1). In the special case of $M_p = 4$ it reduces to the biorthogonal case (Section 4.3). A question of interest then is whether choices of $M_p > 4$, $M_f > 1$ lead to better performance than either phase-shift or biorthogonal modulation.*

In Fig. 3 the K and R/W values (obtained from 43 and 45) are shown for the combined phase orthogonal modulation. The solid curves are for constant values of M_f (noted on the curve); the uppermost point on each such curve corresponds to $M_p = 4$, and each lower point corresponds to M_p increased by a factor of two. The dashed curve goes through the $M_p = 4$ points (biorthogonal). It is apparent from this figure that in this class of systems, for $R/W \leqq 2$, the $M_p = 4$ biorthogonal systems give the largest value of K. For $R/W \geqq 2$, the $M_f = 1$ phase-shift systems give the largest value of K. Thus, *in terms*

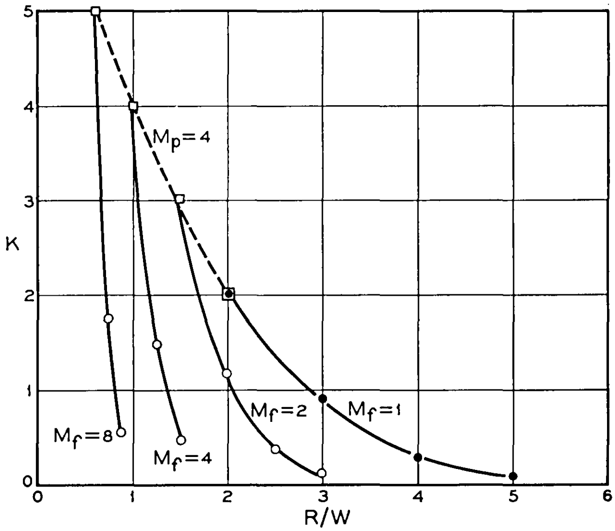


Fig. 3 — $(K,R/W)$ plot for combined phase-shift and orthogonal modulation.

* Reed and Scholtz[5,6] are concerned largely with an algebraic method of generating waveforms with the above correlation properties, rather than in a comparative evaluation of performance.

of asymptotic performance, choice of $M_p > 4$, $M_f > 1$ always gives poorer performance than systems which achieve the same R/W with either $M_p = 4$ (biorthogonal) or $M_f = 1$ (simple phase shift).

For example, consider $M_f = 2$, $M_p = 8$. This yields $R/W = 2$ and $K = 1.17$. However, $R/W = 2$ is also achieved with $M_f = 1$, $M_p = 4$ (quaternary PSK), and for this case $K = 2$. From (21) we can conclude that the latter system is better than the former for all $E/N_o > 2.5$, which includes all P_e of interest. The significance of these results is that we can make this comparison with only a simple slide-rule calculation.

In the above comparison we considered only $M_p \geqq 4$. The case of $M_p = 1$, $M_f > 1$ is the orthogonal modulation previously considered. The case of $M_p = 2$, $M_f > 1$ gives the same performance as biorthogonal but achieves only $\frac{1}{2}$ the R/W and consequently is of little interest. The case of $M_p = 3$, $M_f > 1$ consists of orthogonal combinations of two-dimensional simplexes (equilateral triangles). Reed and Scholtz[6] conjecture that for $M = 3M_f$, the three-phase orthogonal system gives a smaller P_e than any other collection of $3M_f$ signal functions in a space of dimensionality $2M_f$. Although this conjecture may well be true, we wish to point out that if the comparison is made on the basis of fixed R/W (rather than fixed M) then biorthogonal is asymptotically better than three-phase orthogonal. One way of seeing this is by noting that three-phase orthogonal has the same K but smaller R/W than the four-phase (biorthogonal) system of the same dimensionality. To increase the R/W of the three-phase system requires a reduction in K which makes it asymptotically poorer than the corresponding biorthogonal system.

VII. BOUNDS ON K

It has been shown that the $(K,R/W)$ plot provides a useful technique for comparing the performance of various modulation systems. Although our main concern here is in the comparison of specific systems, it is still natural to ask whether there are bounds on what may be achieved in the $(K,R/W)$ plane.

It is apparent from the definition of K

$$K = (1 - \rho) \log_2 M \tag{46}$$

that if no constraint is placed on alphabet size or signal space dimensionality, K can, in principle, be made arbitrarily large for any

R/W. This corresponds to the fact that the Shannon channel capacity formula implies that arbitrarily small P_e may be achieved at all (finite) R/W with finite E/N_o.

If M is held fixed but n is unconstrained, then the maximum K is achieved by the simplex modulation[12] (Section 4.4) for which case $K = [M/(M - 1)] \log_2 M$ and $R/W = (2 \log_2 M)/M - 1$.

Perhaps of more practical interest is the opposite case where the signal space dimensionality n is fixed, but M is unconstrained. Here, sphere-packing arguments may be used to show that[3]

$$M \leqq 2/I_{(1-\rho)/2}\left(\frac{n-1}{2}, \frac{1}{2}\right), \tag{47}$$

where $I_x(p,q)$ is the incomplete beta-function which is extensively tabulated.[21] Thus, for a given ρ and n, an upper bound to M may be calculated from (47). Since $I_x(p,q)$ is monotonic increasing in x, this also gives a lower bound on ρ for fixed M and n. Considered in this latter context we can then determine an upper bound on K with which is associated a given value of $R/W = (2/n) \log_2 M$. This upper bound, K_u, is plotted in Fig. 4 as a function of R/W for $n = 5$ and $n = 10$. Both curves indicate that K_u achieves a maximum value. This is understandable since for large R/W, $1 - \rho$ decreases more rapidly than $\log_2 M$ increases. On the other hand, as R/W decreases, $\log_2 M$ keeps decreasing, whereas $1 - \rho$ is of course always less than 2. Thus, it is not surprising that there exists an R/W at which K_u is a maximum.

It should be noted, however, that K_u is an upper bound which likely cannot be achieved. For example, when $R/W = (2/n) \log_2 2n$, corresponding to $M = 2n$, the optimum configuration is widely conjectured to be the biorthogonal case.[14] The corresponding K and R/W values for biorthogonal with $n = 10$ and $n = 5$ are shown by the points marked (10,1) and (5,1) on the dashed curves of Fig. 4. These points lie well below the upper bounds.

Biorthogonal is a special case ($m = 1$) of the (n,m) permutation modulation considered in Section V. Fig. 4 (dashed curves) shows the K and R/W values for the $(10,m)$ and $(5,m)$ cases. As must be, these curves lie below the upper bounds given by the solid curves.

Finally, we note from Fig. 4 that (n,m) permutation modulation possesses the interesting feature that as m is increased (for a fixed n) a maximum R/W is achieved. Both the properties of the maxima of K_u and the maxima of the R/W of (n,m) modulation are probably worthy of further study.

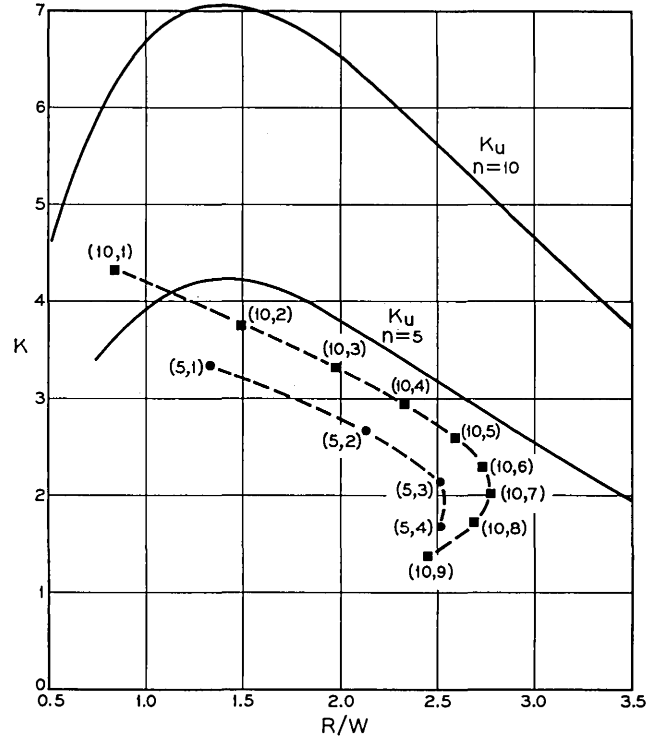


Fig. 4 — Bounds on $(K,R/W)$ for fixed n and comparison with permutation modulation.

VIII. CONCLUSION

The main conclusion to be drawn is that the K, R/W plot provides an exceedingly useful technique for comparing modulation systems. We have restricted ourselves to modulation systems in which the signal alphabet consists of equienergy waveforms for which all rows of the correlation matrix are permutations of a given row. (Geometrically, the alphabet consists of M points on the surface of an n-dimensional sphere such that all points see exactly the same environment.) This class of systems, although somewhat limited, is sufficiently broad to cover most systems of theoretical and practical interest. Given two systems in this class such that $K_1 > K_2$; then in the limit of large E/N_o (low P_e) $P_{e1} < P_{e2}$ for the same E/N_o. Furthermore, we have obtained a simple sufficient condition on the E/N_o above which this inequality is valid. These results are in reality not new.

They are implicit in the results of Shannon[7] and in many other works.[19] What is perhaps new is that many interesting results and comparisons can be obtained by such simple techniques.

Considerably more precise comparisons can of course be made by exact computation of P_e rather than by comparison of K. The latter procedure however is considerably quicker and allows ready consideration of entire classes of systems (e.g., the (n,m) permutation modulation and the combined phase-shift orthogonal modulations considered in the previous sections). The comparisons discussed here are not meant to supplant exact evaluation, but rather as a coarse sieve for delineating systems worthy of more extensive calculation.

REFERENCES

1. Viterbi, A. J., On Coded Phase-Coherent Communications, IRE Trans. Space Elec. Tel., *SET-7,* 1, March, 1961, pp. 3–14 (see also Chapter 7 of *Digital Communications with Space Applications,* by S. W. Colomb, et al., Prentice Hall, 1964).
2. Jacobs, I., Theoretical and Practical Limitations of Weak-Signal Processing Techniques, *Space Research II,* North-Holland Publishing Co., 1961, pp. 413–425.
3. Slepian, D., Bounds on Communication, B.S.T.J., *42,* May, 1963, pp. 681–707.
4. Slepian, D., Permutation Modulation, Proc. IEEE, *53,* March, 1965, pp. 228–236.
5. Reed, I. S. and Scholtz, R. A., N-Orthogonal Phase-Modulated Codes, University of Southern California USCEE Report 134, May, 1965, (Presented at the First IEEE Annual Communications Convention, Boulder, Colorado, June 7–9, 1965.)
6. Reed, I. S. and Scholtz, R. A., N-Orthogonal Phase-Modulated Codes, IEEE Trans. Inform. Theor., *IT-12,* July, 1966, pp. 388–395.
7. Shannon, C. E., Probability of Error for Optimal Codes in a Gaussian Channel, B.S.T.J., *38,* May, 1959, pp. 611–656.
8. Landau, H. J. and Pollak, H. O., Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty-III: The Dimension of the Space of Essentially Time- and Band Limited Signals, B.S.T.J., *41,* July, 1962, pp. 1295–1336.
9. Slepian, D., Group Codes for the Gaussian Channel, Notes prepared for use at Summer School on Coding, Royan, France, August-September 1965.
10. Kotelnikov, V. A., *The Theory of Optimum Noise Immunity,* Translated by R. A. Silverman, McGraw-Hill Book Co., Inc., New York, 1959.
11. Landau, H. J. and Slepian, D., On the Optimality of the Regular Simplex Code, B.S.T.J. *45*, October, 1966, pp. 1247–1272.
12. Balakrishnan, A. V., Signal Selection Theory for Space Communication Channels, *Advances in Communication Systems: Theory and Application,* Vol. 1, Chapter 1, Academic Press, 1965.
13. Nuttall, A. H., Error Probabilities for Equi-Correlated M-ary Signals Under Phase-Coherent and Phase-Incoherent Reception, IRE Trans. Inform. Theor., *IT-8,* July, 1962, pp. 305-314.
14. Weber, C. L., Signal Design for Space Communication Channels, Part I, University of Southern California USCEE Report 129, February, 1965.
15. Sanders, R. W., The Digilock Orthogonal Modulation System, *Advances in Communication Systems: Theory and Application,* ed. by A. V. Balakrishnan, Vol. 1, Chapter 3, Academic Press, 1965.
16. Wolf, J. K., On Comparing N-ary Systems, IRE Trans. Commun. Syst., *CS-10,* June, 1962, pp. 216–217.

17. Gilbert, E. N., A Comparison of Signaling Alphabets, B.S.T.J., *31,* May, 1952, pp. 504–522.
18. Arthurs, E. and Dym, H., On the Optimum Detection of Digital Signals in the Presence of White Gaussian Noise, IRE Trans. Commun. Syst., *CS-10,* December, 1962, pp. 336–372.
19. Wozencraft, J. M. and Jacobs, I. M., *Principles of Communication Engineering,* John Wiley & Sons, Inc., New York, 1965, Chapters 4 and 5.
20. Cahn, C. R., Performance of Digital Phase-Modulation Communication Systems, IRE Trans. Commun. Syst., *CS-7,* May, 1959, pp. 3–6.
21. Pearson, K., *Tables of the Incomplete Beta-Function,* University Press, Cambridge, England, 1934.

Combinatorial Solution to the Problem of Optimal Routing in Progressive Gradings

By V. E. BENEŠ

(Manuscript received June 10, 1966)

The grading or graded multiple proposed by E. A. Gray is a certain kind of one-stage, two-sided, partial access telephone connecting network for switching customers' lines to trunks all having the same destination. Its essential feature is that traffic from lines not having identical access patterns can be offered to a common trunk, and so pooled. In a progressive grading the trunk groups are partially ordered in a hierarchy, i.e., some provide primary routes, others function as secondary routes which handle traffic overflowing from primary routes, as well as originating traffic, etc., up to final routes.

A call which is using an overflow or "later" trunk when it could be using a primary or "earlier" group is said to make a "hole in the multiple". It was recognized early in the development of gradings that such holes were undesirable.

The problem of optimal routing in telephone networks, considered in general in the author's earlier work, is here specialized to progressive gradings. It had been shown that for networks with certain combinatorial properties the optimal choices of routes for accepted calls (so as to minimize the loss under perfect information) could be described in a simple and intuitive way in terms of these properties. The present paper gives a proof that all progressive gradings have such a combinatorial property, associated with the hierarchical nature of the grading. The optimal policy for routing accepted calls is related to the phenomenon of "holes in the multiple", and can be paraphrased in the traditional telephone terminology thus: filling a hole in the multiple is preferable to using a final route, and filling an earlier hole is preferable to filling a later one.

I. INTRODUCTION

The term 'hierarchical' has often been used to describe connecting networks in which the possible routes for a call are ordered, with the order determining the routing decisions in that the earlier routes are hunted over before the later. The Bell System's toll network is often cited as an example of a hierarchical network. Recently, J. H. Weber has used the word 'hierarchical' in a more technical sense to describe trunking networks ". . . 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 route."[1]

In this paper, we consider some ways in which the concept of a hierarchy of routes is relevant to the problem of optimal routing as formulated in previous work.[2] Naturally, such a hierarchy can be relevant to routing only if it is in a suitable way related to those combinatorial properties of the network which distinguish the 'good' from the 'poor' ways of completing calls. (Examples of such properties were given in Ref. 2.) It shall be shown that natural hierarchies associated with certain *gradings* hold the key to the routing problem in these one-stage networks.

It is now known[2] that if a network possesses one of certain combinatorial properties, then this property can be used to describe in a simple way the optimal choices of routes for accepted calls so as to minimize the loss under perfect information. The next natural question is, then, what networks possess some of these properties? We shall prove that the members of an important subclass of connecting networks, that of *progressive gradings,* all have a combinatorial property similar to the strongest of those of Ref. 2; this property is associated with a natural hierarchy of routes, and leads to a solution of the routing problem for accepted calls.

II. GRADINGS

We first discuss and clarify some of the usage and terminology associated with gradings. Since about* 1905 the noun 'grading' and the adjective 'graded' have been used in telephony to describe a certain kind of one-stage two-sided network for connecting customers' lines to trunks all having the same destination. Roughly speaking, a grading has this property: some trunk is such that two lines have access to it which do not have access to the same trunks. The essential feature is that traffic from distinguishable lines (i.e., ones not having identical access patterns) can be offered to a common trunk.

* E. A. Gray proposed the "graded multiple" in 1905, and was granted a patent for it (No. 1002388) in 1911.

It appears, though, that the word 'grading' has been used in a wider sense in Europe than in the United States. In particular, the American usage[3] implies a certain order in the pattern of access that the lines have to the trunks, whereas in the European meaning this implication is absent. The order implicit in the American usage amounts to this: the trunks are partitioned into groups which are so partially ordered that no group has more than one successor in the ordering; a line that has access to one group has access to all groups that follow it in the ordering. (This ordering usually determines the order in which the lines hunt over the trunks.) Thus, e.g., a trunk group with no predecessors in the ordering can be used by exactly one group of lines, for which it is the "primary" route. In one European sense of "grading," however, a trunk group which is the first one hunted over by one line group may be the nth one ($n > 1$) hunted over by some other line group.[4] The distinction drawn here is of some importance, inasmuch as the order structure implicit in the American usage gives rise to a natural hierarchy of routes that is directly relevant to routing, whereas in the more general case this hierarchy is not necessarily present.

Recently, in an effort to establish a uniform terminology, the nomenclature committee of the International Teletraffic Congress decided[5] that the terms 'grading' and 'graded multiple' should be interchangeable, and the structures described in R. I. Wilkinson's paper[3] as graded multiples be called, more specifically, *progressive graded multiples* or *progressive gradings,* the word 'progressive' here referring to the order structure we have described as characteristic of the American usage. The usage recommended by this committee is adopted herein.

Since the present work can be viewed as a continuation of Ref. 2, we take the liberty of assuming familiarity with the notations and concepts used there, and we include only occasional reminders of the meanings of important notions.

III. HIERARCHIES OF ROUTES

It will be convenient to have a notation for *routes.* A route r for a call c is just a way in which c can be put up or realized in a network ν, and so it can be identified with the state in which the only call in progress is c using route r. Thus, a route for c is any element of $\gamma^{-1}(c)$.* We use the variables q and r (over the set L_1 of states with one call in progress) to denote routes.

* We recall that if x is a state, $\gamma(x)$ is the assignment of inlets to outlets realized by x.

By a *hierarchy of routes* we mean a partial ordering $\supseteq$ contained in

$$\bigcup_{c} [\gamma^{-1}(c)]^2.$$

It is apparent that $\supseteq$ can hold only between alternative routes for the same call. (Of course, not every hierarchy of routes is relevant to routing; only those that have a suitable relation to the ways in which calls in progress block new calls will be of interest. The problem is to clarify the meaning of 'suitable'.)

A hierarchy of routes, being a partial ordering of the states with one call in progress, can be extended to, or can induce, a partial ordering of the whole set S of states in several natural ways. Since $\supseteq$ can hold only between alternative routes for the same call, it is reasonable to confine attention to extensions which hold only between states that are equivalent in the sense of $\sim$ in Ref. 2, i.e., are (possibly) different ways of realizing the same assignment. An obvious first candidate for such an extension is given by the condition

$$x \sim y \quad \text{and} \quad r \leqq x,\, q \leqq y,\, r \sim q \quad \text{imply} \quad r \supseteq q. \tag{1}$$

However, we eschew this definition in favor of a stronger one: let us set

$x \supseteq y \equiv x$ is reachable from y by sequentially moving calls in progress from routes that are lower (later) (in the sense of $\supseteq$ on L_1) to routes that are higher (earlier).*

It is intended here not merely that, as in (1), each call have a higher route in x than in y, but that it should be possible to pass from y to x by a sequence of equivalent states each differing from the previous one in that one call has been rerouted on a higher route. This stronger condition is rendered formally by first defining

$$\begin{aligned} x\, Q\, y \equiv |\, x \cap y \,| = |\, x \,| - 1 \text{ and either} \\ x - (x \cap y) \supseteq y - (x \cup y) \text{ or} \\ |\, x \,| = 1 \text{ and } x \supseteq y \end{aligned}$$

and then setting

$$\begin{aligned} \supseteq &= I \cup Q \cup Q^2 \cup \cdots \\ &= \text{transitive closure of } Q. \end{aligned} \tag{2}$$

* In an attitude prejudiced and justified by the principal results (Theorems 1 and 2) we are working toward, we use the words 'lower', 'earlier', and their antonyms so as to suggest consistently that *lower* routes are less desirable than higher, *earlier* ones are preferable to later, etc.

IV. PROGRESSIVE GRADINGS

In a one-stage connecting network $\nu = (G,I,\Omega,S)$, with I the set of customers' lines (inlets) and Ω that of trunks (outlets), the graph G giving network structure is determined entirely by the access relation A such that

$$lAt \equiv \text{line } l \text{ has access to trunk } t.$$

The set S of states of ν can be represented by the set of all subsets of A which are one-to-one correspondences. The range of x, rng (x), is the set of trunks which are busy in x.

The access relation A can be used to give a simple definition of a progressive grading. We use $X \times Y$ for the Cartesian product of X and Y, i.e., the set of pairs (x,y) with $x \varepsilon X$ and $y \varepsilon Y$. If X is a set, $| X |$ denotes the number of elements of X.

Definition: ν is a *progressive grading* if and only if it is a one-stage network for which there exist partitions Π and Ξ of Ω and I, respectively, and a partial ordering $\geqq$ of Π, such that for T, U, V ε Π and L ε Ξ

(*i*) $(L \times T) \cap A \neq \theta$ implies $(L \times T) \subseteq A$,

(*ii*) $(L \times U) \subseteq A$, $V \geqq U$ imply $(L \times V) \subseteq A$,

(*iii*) $U \geqq T$, $V \geqq T$ imply $U \leqq V$ or $V \leqq U$

(*iv*) $| L | \geqq | \bigcup_{T:(L\times T)\subseteq A} T |$.

The first condition simply says that if a line has access to some trunk from a group T, then all lines in its line group have access to every trunk in T. The second condition says (roughly) that a line with access to a trunk group T has access to all groups that are later than T in the partial ordering. The third condition says that a trunk group is followed (in the partial ordering) by at most one other group; if the "later" groups are thought of as overflow groups, this means that each group has at most one group to which to overflow traffic. Finally, the fourth condition rules out the relatively uninteresting cases in which some line group has access to more trunks *in toto* than there are lines in the group.

It is apparent that if a trunk group T_1 is later than one T_2, then every line with access to T_2 has access to T_1. This is the "progressive" property. In analogy with the intuitions expressed in Ref. 2, it should be better to use an earlier trunk group than a later one, if both are available. Thus, the structure of a progressive grading at once suggests the conjecture that optimal routing will consist of using the early routes in preference to the later or (to anticipate a bit) overflow groups. This conjecture is true and follows from Theorem 2. In traditional telephone terminology (see E. C. Molina's appendix in Ref. 3) it states that filling a hole in the multiple is preferable to using a final route, and that filling an earlier hole is preferable to filling a later one.

A line group L is said to be a *bye* if it has access only to "overflow" trunk groups, i.e., if

$$\inf_{\leqq} \{T{:}LAT\}$$

is not minimal in $\leqq$, where we have written LAT for $(L \times T) \subseteq A$.

It is easily seen that in a progressive grading a hierarchy of routes can be defined by this rule: $r \sqsupseteq q$ if and only if $r \sim q$ and $g(q) \geqq g(r)$, where $g(r)$ is the trunk *group* used by route r.* This is the *natural hierarchy* of routes associated with a progressive grading; here $r \sqsupset q$ if and only if $r \sim q$ and r is on an "earlier" trunk group than q. In this instance, $\geqq$ is also a *simple* ordering on each $g(\gamma^{-1}(\gamma(r)))$. These simple orderings forming the hierarchy of course correspond exactly to the preference relation among routes suggested by the natural intuition (already mentioned) that there is no point in using a later or "overflow" trunk when an earlier one is available, because possibly fewer lines have access to the latter. The relation $\sqsupseteq$ defined above on L_1 extends by (2) to all of S.

V. PARTIAL ORDERING OF PROGRESSIVE GRADINGS

In a proof to be given later we shall use the fact that the set of progressive gradings can be partially ordered by a relation $\sqsupseteq$ according to the following definition of covering: ν_1 covers ν_2 if and only if ν_2 is obtained from ν_1 by removing, for some line group L, either (case 1) a trunk from the first (in $\leqq_1$) trunk group to which L has access together with one line of L if L has access to more than one trunk, or (case 2) *the* trunk to which L has access together with L itself if L has access to exactly one trunk. That is, if ν_1 is defined by partitions Π_1, Ξ_1, a partial ordering $\geqq_1$ of Π_1, and an access relation A_1, then ν_1 covers ν_2 provided that there exist $t \in \Pi_1$ and $l \in L \in \Xi_2$ with

$$T = \inf_{\geqq_1} \{U \in \Pi_1{:}(L \times U) \subseteq A\}$$

such that ν_2 is defined by (case 1)

* Note the shift to the converse.

$$\begin{aligned}
\Pi_2 &= \Pi_1 - \{T\} + \{T - \{t\}\} \\
\Xi_2 &= \Xi_1 - \{L\} + \{L - \{l\}\} \\
\geqq_2 &= \geqq_1 \quad \text{with} \quad T - \{t\} \quad \text{for } T \text{ throughout} \\
A_2 &= A_1 - (I \times \{t\}) - (\{l\} \times \Omega),
\end{aligned}$$

if $T \neq \{t\}$ or $A_1 \cap (L \times \Omega) \not\subseteq (L \times \{t\})$, and by (case 2)

$$\begin{aligned}
\Pi_2 &= \Pi_1 - \{T\} \\
\Xi_2 &= \Xi_1 - \{L\} \\
\geqq_2 &= \geqq_1 - (\Pi_1 \times \{T\}) \\
A_2 &= A_1 - (I \times \{t\}) - (L \times \Omega),
\end{aligned}$$

if $T = \{t\}$ and $A_1 \cap (L \times \Omega) \subseteq (L \times \{t\})$.

For practical purposes a network in which some line group has access to no trunks is in all respects equivalent to the same network with those lines omitted. For this reason the definition of covering was divided into cases 1 and 2, so as to build this equivalence right into the definition.

As we have said, ν_1 covers ν_2 if and only if ν_2 results from ν_1 by ripping out (*i*) some trunk from a "primary" group, (*ii*) a line with access to it, and (*iii*) all crosspoints associated with these terminals, with the proviso that if this leaves some lines with access to no trunks, then these lines are also to be removed. Because of this, there exists a natural or canonical map μ of the states $S(\nu_1)$ of ν_1 into those $S(\nu_2)$ of ν_2, defined roughly by the condition that μx is what is left of x after the line and trunk that define the covering of ν_2 by ν_1 have been ripped out. The canonical map can be defined formally very simply, as follows: A state x of ν_1 is representable as a subset of A_1 which is also a one-to-one correspondence; similarly, a state of ν_2 is just a one-to-one map contained in A_2; what is left of x after the ripping-out process is just

$$\mu x = x \cap A_2 .$$

Thus, if μ corresponds to ripping out line l and trunk t, and $x = \{(l,t)\}$, then $\mu x = \theta =$ zero state. If $x = \{(l,t_1)\}$ or $x = \{(l_1, t)\}$ with $l_1 \neq l$ and $t_1 \neq t$, then again $\mu x =$ zero state. If $x = \{(l_1, t_1)\} \cup y$, with $l_1 = l$ or $t_1 = t$, then $\mu x = \mu y$. It is easy to see that if μ rips out l and t, then μS is isomorphic with the "cone"

$$\{x \varepsilon S{:}x \geqq \{(l,t)\}\},$$

because it does not make any difference whether l and t are present in the system and connected to each other, or are just absent. That is,

μS is essentially the set of states of S that remain available if l is connected to t with a holding-time of $+\infty$.

This notion of a canonical map provides many useful notations. It is convenient to extend the μ-notation as follows: For $T \varepsilon \Pi$

$$\mu T = T \cap \text{range } (A_2) = \begin{cases} T - \{t\} & \text{if} \quad t \varepsilon T \quad \text{and} \quad T \neq \{t\}, \\ T & \text{if} \quad t \not\varepsilon T, \\ \theta & \text{if} \quad T = \{t\}. \end{cases}$$

Clearly, μT is what is left of the trunk group T after the line l and the trunk t associated with μ have been ripped out. Also, we set

$$\mu \geqq \; = \{(\mu T_1 , \mu T_2) : T_1 \geqq T_2 , \mu T_1 \neq \theta, \text{ and } \mu T_2 \neq \theta\}$$

$$\mu \supseteq \; = \{(\mu x, \mu y) : x \supseteq y, \mu x \neq \theta \text{ or } x = \theta, \text{ and } \mu y \neq \theta \text{ or } y = \theta\}.$$

The relation $\mu \geqq$ can be seen to be identical with $\geqq_2$; it is a useful mnemonic; it defines the hierarchy of routes in the "reduced" system ν_2 ; the partial ordering induced in $S(\nu_2)[= \mu S(\nu_1)]$ by this hierarchy is precisely $\mu \supseteq$.

VI. PRELIMINARY RESULTS

In Ref. 2, for a general partial ordering R, the notation

$$\sup_R A_{cx}$$

was used for the *set*

$$\{y : z \varepsilon A_{cx} \text{ implies } yRz\} \cap A_{cx}$$

whenever this set was nonempty. The notation was chosen to denote a *set* of R-maximal elements of A_{cx} , rather than an actual R-maximal element itself, so as not to prejudge the question as to how many there were. It will be shown that if the network ν under study is a progressive grading, and $R = \supseteq =$ natural hierarchy, then unless c is blocked in x (and A_{cx} is empty) A_{cx} always has a $\subseteq$-maximal element which is unique to within equivalence under permutations of lines within their line groups and trunks within their trunk groups.

Let now x be a state and let $c \varepsilon x$ be a call which is not blocked in x. It is apparent that for $y, z \varepsilon A_{cx}$ we have either $g(y - x) \geqq g(z - x)$ or $g(y - x) \leqq g(z - x)$. Hence, there is a $y_0 \varepsilon A_{cx}$ such that

$$g(y_0 - x) \geqq g(w - x)$$

$$y_0 - x \supseteq w - x$$

$$y_0 \, Q \, w$$

$$y_0 \supseteq w$$

for all $w \varepsilon A_{cx}$, and y_0 is unique to within equivalence. (Recall the construction of Q in Section III, and the fact that $\supseteq$ is $\overline{I \cup Q}$.) Hence,

$$\sup_{\supseteq} A_{cx} \quad (\sup A_{cx} \text{ for short when the context permits})$$

exists, and equals $\tau(y_0)$, $\tau(\cdot)$ being the natural homomorphism of S into the quotient $S/(\supseteq \cap \subseteq)$. (See Ref. 2.)

We now consider policies $\varphi(\cdot,\cdot)$ such that

$$\varphi(e,x)\begin{cases} = x - h & \text{if} \quad e \text{ is a hangup } h, \\ \varepsilon \sup A_{cx} & \text{if} \quad e \text{ is a new call } c \text{ not blocked in } x. \end{cases} \tag{3}$$

Such a policy expresses the routing rule of always choosing the earliest available trunk in the natural hierarchy characteristic of a progressive grading.

The relation B (for "better") was defined in Ref. 2 by the condition

$x \, B \, y$ if and only if $x \sim y$ and every call blocked in x is also blocked in y.

By Theorem 1, to be proved shortly, it will follow that $x \supseteq y$ implies $x \, B \, y$, which in turn implies $s(x) \geqq s(y)$. Thus, the policies $\varphi(\cdot,\cdot)$ coincide with the "maximum $s(\cdot)$" policies suggested in Ref. 2. (See Ref. 2 for notations.)

Lemma 1: If the line of c is not involved in the canonical map μ, and $A_{cx} \neq \theta$, then

$$\mu(\sup_{\supseteq} A_{cx}) \subseteq \sup_{\mu\supseteq} A_{c(\mu x)} .$$

Proof: Let l^* be the line of c, and suppose that

$$y \, \varepsilon \sup_{\supseteq} A_{cx} .$$

Let l and t be the line and trunk, respectively, associated with μ. There exists a trunk t^* such that

$$y = x \cup \{(l^*,t^*)\}$$

$$\mu y = \mu x \cup \{(l^*,t^*)\}$$

$$t^* \, \varepsilon \inf_{\geqq} \{T{:}T \nsubseteq \operatorname{rng}(x) \text{ and } l^*AT\}.$$

Let T^* denote the set (trunk group) achieving the infimum on the right. Since t is busy in x and t^* is not, $t \neq t^*$. Thus, $T^* \neq \{t\}$, and $\mu(T^*) \neq \theta$.

We first observe that l^*AT implies $l^*A_2\mu T$, since c is not involved in μ.

We next show that $\mu T \not\subseteq \text{rng}\,(\mu x)$ implies $T \not\subseteq \text{rng}\,(x)$. If not, then there exists $t_1 \;\varepsilon\; \mu T$, hence $\varepsilon\; T$ such that $t_1 \not\varepsilon\; \text{rng}\,(\mu x)$ and $t_1 \;\varepsilon\; \text{rng}\,(x)$. But

$$\text{rng}\,(\mu x) = \text{rng}\,(x) - \{t\}.$$

Hence, $t_1 = t =$ trunk removed by μ. But this is impossible since $t_1 \;\varepsilon\; \mu T$, while $t \not\varepsilon\; \mu T$.

Now $T^* \leqq T$ for every T such that $T \not\subseteq \text{rng}\,(x)$ and l^*AT. From the two previous paragraphs, it follows that

$$(\mu T^*)(\mu \leqq)\mu T$$

for every T such that $\mu T \not\subseteq \text{rng}\,(\mu x)$, $l^*A\mu T$, $\mu T \neq \theta$. That is,

$$\mu T^* = \inf_{\mu \leqq} \{T{:}T \not\subseteq \text{rng}\,(\mu x),\ l^*AT\}.$$

Now $t^* \;\varepsilon\; T^*$, $t^* \neq t$, so $t^* \;\varepsilon\; \mu T^*$. If now $w \;\varepsilon\; A_{c(\mu x)}$, then

$$(w - \mu x)(\mu \subseteq)\{(l^*,t^*)\}$$
$$w(\mu \subseteq)(\mu x \cup \{(l^*,t^*)\})$$
$$w(\mu \subseteq)\mu y.$$

Thus,

$$\mu y \;\varepsilon \sup_{\mu \subseteq} A_{c(\mu x)}\ ,$$

and since y was arbitrary within $\sup\limits_{\supseteq} A_{cx}$, the lemma is proved.

Lemma 2: In a progressive grading, $Q \subseteq B$.

Proof: Let $x\,Q\,y$. This implies that there exists $z \;\varepsilon\; B_x \cap B_y$ such that $x - z \supseteq y - z$, i.e.,

$$g(y - z) \geqq g(x - z).$$

Now let c be a call from line l which is blocked in x but not in y. Then c is not blocked in z either. The only trunk which is busy in x and not in z is that used by the call $\gamma(x - z)$. Thus, since c is blocked in x and not in z, $g(x - z)$ is a trunk group usable for the call c. However, by property (*ii*) of progressive gradings, $\{l\} \times g(y - z) \subseteq A$, i.e., l has access to the group $g(y - z)$ as well. Hence, some trunk of $g(y - z)$ is idle in x, since the call $\gamma(x - z)$ has a choice of routes in state z, one of these being on $g(y - z)$. Thus, c is not blocked in x, and $x\,B\,y$.

Theorem 1: In a progressive grading, the partial ordering $\supseteq$ induced by the natural hierarchy of routes is contained in B.

Proof: Immediate from Lemma 2 and the facts that $\supseteqq$ is the transitive closure of $I \cup Q$, and that B is transitive.

Lemma 3: If $x \supseteqq y$, then x is obtainable from y by moving calls to earlier routes in such a way that each call is moved at most once.

Proof: The result is true if only one move is made. Suppose it to hold if n moves *in toto* are made. Let x be obtainable from y by sequence of $(n + 1)$ moves. The trunk groups available for a given call c form a set simply ordered by $\leqq$, and so can be indexed $1, 2, \cdots$, the $\leqq$-earlier receiving the lower integer. For $c \leqq \gamma(x)$, let $n(c,x)$ be the index of the group used by c in x. Some call c that is moved in obtaining x from y achieves

$$\min \{n(c_1, x) \mid c_1 \text{ moved in getting } x \text{ from } y\}.$$

Starting in state y it is possible to move such a call (once) directly to its route in x, to get a state z in which it is still possible to carry out exactly each of the moves that take y into x except those involving c. These are at most n in number, so each call involved need be moved at most once.

A policy $\varphi(\cdot,\cdot)$ is said to *preserve* a relation $R \subseteq \sim$ if $x \, R \, y$ implies

$$\varphi(e,x) \, R \, \varphi(e,y)$$

for every event e that is either a hangup or a new call not blocked in either x or y. It has been shown in Ref. 3 for a general network that if φ preserves B then it embodies the optimal routing policy for accepted calls.

The main theorem we prove (Section VII) states that a sup A_{cx} policy, i.e., one satisfying (3), preserves $\supseteqq$. The method to be used in the proof of this result is illustrated in part by the following remarks: consider linear arrays $x, y, z, \cdots$ each of n urns, $n \geqq 2$, each urn containing at most one ball, with fewer than n nonempty urns per array. Let $x \supseteqq y$ mean that x is obtainable from y by moving balls to the left. Let φx denote the result of adding a ball in the leftmost empty urn.

Observation: If $x \supseteqq y$, then $\varphi x \supseteqq \varphi y$.

Proof: The result is obviously true for $n = 2$ by enumeration. Let it hold for a given value $n \geqq 2$, and consider arrays x, y of n urns satisfying the hypotheses. Let ψz denote the result of removing the leftmost urn from z, and bz that of adding an urn containing one ball at the left of z. There are two cases: (i) the leftmost urns are empty in both x and

y, or both nonempty in x, y; (*ii*) in y, but not in x, the leftmost urn is empty.

Case (*i*): $\varphi x = b\psi x$, $\varphi y = b\psi y$, $\psi x \supseteq \psi y$; hence, $\varphi x \supseteq \varphi y$.

Case (*ii*): In obtaining x from y some ball moved into the leftmost urn. Obtain z from y by moving just this one ball to the leftmost urn. Then $x \supseteq z \supseteq y$, $\varphi x \supseteq \varphi z$, $\varphi y = b\psi z$, $\varphi z = b\varphi\psi z$. Since $\varphi\psi z$ is obtained from ψy by removing some ball, and replacing it in the leftmost empty urn of the resulting array, we have $\varphi\psi z \supseteq \psi y$, and so $\varphi z \supseteq \varphi y$.

In cases 3 and 4 of the proof of the next theorem, the analog of the inductive index n will be the partial ordering of the set of progressive gradings.

VII. PRINCIPAL RESULT

Theorem 2: In a progressive grading ν let $\supseteq$ be the partial ordering induced by the natural hierarchy of routes in ν, and let φ be a policy with the property that

$$\varphi(c,x) \ \varepsilon \ \sup_{\supseteq} A_{cx} \ , \qquad\qquad c \ \varepsilon \ x, \ c \text{ not blocked in } x.$$

Then φ preserves $\supseteq$.

Proof: The proof is by induction over the partial ordering $\supseteq$ of the set of progressive gradings which is defined by the definition of covering given earlier. A grading ν that is minimal in $\supseteq$ has no "overflow groups", i.e., $\geqq$ = identity relation, so that no trunk group has a successor in the order $\geqq$ characteristic of ν. Thus, ν consists entirely of trunk groups serving line groups on a one-to-one basis, so that for some n

$$A = \bigcup_{i=1}^{n} (L_i \times T_i),$$

where

$$\Xi = \{L_i\ , = 1, \cdots , n\}$$

$$\Pi = \{T_i\ , = 1, \cdots , n\}, \quad \text{with} \quad | \, T_i \, | = 1.$$

In this minimal case $\supseteq$ is the identity relation, and φ obviously preserves it.

As a hypotheses of induction, we now suppose that every progressive grading covered by ν has the property that any sup A_{cx}-policy preserves $\supseteq$. Let now $x \supseteq y$ in ν and let $e \ \varepsilon \ x$. The induction argument will have four cases, the last two of which are analogous to the observation made earlier.

Case 1: $x \supseteq y$, and e is a hangup h. There is a sequence $x = z_1, z_2, \cdots, z_n = y$ with

$$z_j \, Q \, z_{j+1} \qquad j = 1, \cdots, n-1.$$

This sequence indicates how one would get y from x by moving calls to "preferred" routes. By Lemma 3 it is no restriction to assume that no call is rerouted more than once. Let the route of h be r in x and q in y. If h is one of the calls whose route is changed in the above sequence, say to take z_k into z_{k+1} by changing the route of h from r to q, then

$$x - r = z_1 - r, z_2 - r, \cdots, z_k - r = z_{k+1} - q, \cdots, z_n - q = y - q$$

is a sequence which shows that $(x - r) \supseteq (y - q)$. If the route of h is not changed, then $r = q$ and the same conclusion follows.

Case 2: $x \supseteq y$, and $e \; \varepsilon \; x$ is a new call c blocked in x. By Theorem 1, $x \, B \, y$, so c is also blocked in y. Then,

$$A_{cx} = \{x\}, \qquad A_{cy} = \{y\}$$

$$\varphi(c,x) = x \qquad \varphi(c,y) = y$$

$$\varphi(c,x) \supseteq \varphi(c,y).$$

Case 3: $x \supseteq y$, e is a new call c not blocked in either x or y, and the line group L of c is not a bye. Let

$$T = \inf_{\leq} \{S{:}LAS\}.$$

Subcase 3.1: T is full in neither x nor y. Then there exist routes r, q such that $g(r) = g(q) = T$,

$$\varphi(c,x) = x \cup r, \qquad \varphi(c,y) = y \cup q,$$

$r \equiv q$ modulo trunk permutations within T, and clearly

$$\varphi(c,x) \supseteq \varphi(c,y)$$

since c was put up on group T in both cases. To see this, if $x = z_1, z_2, \cdots, z_n = y$ is a sequence with

$$z_j \, Q \, z_{j+1} \qquad j = 1, \cdots, n-1,$$

showing that $x \supseteq y$, then
$x \cup r = z_1 \cup r, z_2 \cup r, \cdots, z_n \cup r = y \cup r, y \cup q$ is a sequence which shows that

$$(x \cup r) \, Q \, (y \cup q).$$

This is because we can assume without loss of generality that the transformations which change y into x reroute a call at most once, and thus move no calls onto T. (Lemma 3.)

Subcase 3.2: T is full in both x and y. Since L is not a bye, there exist $l, m \; \varepsilon \; L$ and $t, u \; \varepsilon \; T$ with

$$(l,t) \; \varepsilon \; x \quad \text{and} \quad (m,u) \; \varepsilon \; y.$$

Because l, m and t, u are respectively interchangeable, i.e., since lines and trunks are permutable within their respective groups, no loss of generality is incurred if it is supposed that $l = m$ and $t = n$. Let μ be the canonical map corresponding to ripping out l and t.

Then ν covers ν_1, where ν_1 is defined by ripping l and t out of ν, i.e., by

$$\mu\Pi = \Pi_1 = \begin{cases} \Pi - \{T\} + \{T - \{t\}\} & \text{in case 1} \\ \Pi - \{T\} & \text{in case 2,} \end{cases}$$

$$\mu\Xi = \Xi_1 = \begin{cases} \Xi - \{L\} + \{L - \{l\}\} & \text{in case 1} \\ \Xi - \{L\} & \text{in case 2,} \end{cases}$$

$$\mu\geqq = \geqq_1 = \begin{cases} \geqq \quad \text{with } T - \{t\} \text{ replacing } T \text{ throughout,} & \text{in case 1} \\ \geqq - (\Pi_1 \times \{T\}), & \text{in case 2,} \end{cases}$$

$$\mu A = A_1 = \begin{cases} A - (I \times \{t\}) - (\{l\} \times \Omega) & \text{in case 1} \\ A - (I \times \{t\}) - (L \times \Omega) & \text{in case 2,} \end{cases}$$

with

$$\text{case 1} \equiv T \neq \{t\} \quad \text{or} \quad A \cap (LT) \not\subseteq (L \times \{t\})$$

$$\text{case 2} \equiv T = \{t\} \quad \text{and} \quad A \cap (LT) \subseteq (L \times \{t\}).$$

The line of c is not involved in μ, and $A_{cx} \neq \theta$, $A_{cy} \neq \theta$. Hence, Lemma 1 gives

$$\mu\varphi(c,x) \; \varepsilon \sup_{\mu\supseteq} A_{c(\mu x)}$$

$$\mu\varphi(c,y) \; \varepsilon \sup_{\mu\supseteq} A_{c(\mu y)} \; .$$

Since $x \supseteq y$, and either both $\mu x = 0$, $\mu y = 0$, or neither, we have

$$(\mu x, \mu y) \; \varepsilon \; \mu\supseteq. \tag{4}$$

Let ξ be a policy for ν_1 with

$$\xi(d,\mu z) \; \varepsilon \sup_{\mu\supseteq} A_{d(\mu z)} \; , \qquad Vd \; \varepsilon \; \mu z. \tag{5}$$

The hypothesis of induction and (4) give

$$\xi(c,\mu x)(\mu\supseteq)\xi(c,\mu y). \tag{6}$$

However, by (6) and (5)

$$\begin{aligned}\mu\varphi(c,x)(\mu\supseteq)&\xi(c,\mu x)\\ (\mu\supseteq)&\xi(c,\mu y)\\ (\mu\supseteq)&\mu\varphi(c,y).\end{aligned}$$

But $\varphi(c,x)$ differs from $\mu\varphi(c,x)$ and $\varphi(c,y)$ from $\mu\varphi(c,y)$, only in having an additional line l and an additional trunk t connected to each other. Hence,

$$\varphi(c,x) \supseteq \varphi(c,y).$$

The argument of subcase 3.2, basic to Theorem 2, can be appreciated by looking at it thus: $x \supseteq y$ means $\exists z_1, \cdots, z_n$ with $z_i \; Q \; z_{i+1}$, $i = 1, \cdots, n-1$, $z_1 = x$, $z_n = y$. Since

$$e = \{(l,t)\} \leqq x \cap y$$

we have $r \leqq z_i$, $i = 1, \cdots, n$ because we can assume that the call using r is not moved as y is transformed into x by moving calls. Thus,

$$(z_i - r) \; Q \; (z_{i+1} - r), \qquad i = 1, \cdots, n-1$$
$$(x - r) \; Q \; (y - r).$$

But the "cone" $\{z{:}z \geqq r\}$ is isomorphic to the states of a grading (ν_1 of the proof) covered by ν and the isomorphism, *viz.*, μ restricted to the cone, has the basic property, for x, y in the cone

$$x \supseteq y \quad \text{if and only if} \quad (\mu x)(\mu\supseteq)(\mu y).$$

Subcase 3.3: T is full in x, but not in y. Since L is not a bye it is $\leqq$-minimal, and hence there exists a call d with $d \leqq \gamma(x) \cap \gamma(y)$ such that d is on T in x is not on T in y, and can be moved to T in state y to give rise to a new state z without rendering impossible any the remaining moves which transform y into x. Thus, $x \supseteq z \supseteq y$. Since $x \cap z \neq \theta$, subcase 3.1 gives $\varphi(c,x) \supseteq \varphi(c,z)$. Further, the route of c in $\varphi(c,z)$ is no higher (later) in $\leqq$ than the one in y left by d as it was moved to T to give rise to state z. Hence, to within equivalence

$$\varphi(c,z) \supseteq \varphi(c,y).$$

Case 4: $x \supseteq y$, e is a new call c not blocked in either of x or y, and the line group of c is a bye. There is at least one other line group L which is not a bye. Let

$$T = \inf_{\leqq} \{S{:}LAS\}.$$

Subcase 4.1: $L \times T \cap x \neq \theta$, $L \times T \cap y \neq \theta$ or $L \times T \cap x = \theta$, $L \times T \cap y = \theta$. Since L is not a bye, there exist l, $m \;\varepsilon\; L$ and t, $u \;\varepsilon\; T$ with

$$(l,t) \;\varepsilon\; x, \qquad (m,u) \;\varepsilon\; y$$

or

$$(l,t) \not\varepsilon\; x, \qquad (m,u) \not\varepsilon\; y.$$

(In the second instance, property (*iv*) of the definition of a progressive grading has been used to conclude that there must be idle lines on L if there are idle trunks on T.)

As in subcase 3.2, no loss of generality is incurred if it is supposed that $l = m$ and $t = u$. Let μ be the canonical map corresponding to ripping out l and t. The argument now continues as in subcase 3.2.

Subcase 4.2: $(L \times T) \cap x \neq \theta$, $(L \times T) \cap y = \theta$. Since L is $\leqq$-minimal, there exists a call d with $d \leqq \gamma(x) \cap \gamma(y)$ such that d is on T in x, is not on T in y, and can be moved to T in state y to give rise to a new state z without rendering impossible any of the remaining moves which transform y into x. Thus, $x \supseteq z \supseteq y$. Since $x \cap z \neq \theta$, subcase 3.1 gives $\varphi(c,x) \supseteq \varphi(c,z)$.

Let l^* be the line of c, r be the route of d in y, $T_d = g(r)$, and

$$T_c = g(\inf_{\leqq} \{S{:}l^*AS, S \not\subseteq \operatorname{rng}(y)\}).$$

Here T_c is the earliest group c could be put on in y. Let also ψ denote the operation of moving d from T_d to T, and for any call f

$$\begin{aligned} A_f &= \{g(r){:}f = \gamma(r)\} \\ &= \{S{:}\text{the line of } f \text{ has access to } S\}. \end{aligned}$$

Case (*i*)*:* $T_d \;\varepsilon\; A_c \cap A_d$, $T_d \leqq T_c$. Then moving d from T_d to T means that c can use T_d in z, so $\varphi(c,z) \supseteq \varphi(c,y)$, because $\varphi(c,z)$ results from $\varphi(c,y)$ by moving first d to T_d and then c to T_d, so actually

$$\varphi(c,z) \supseteq \psi\varphi(c,y).$$

Case (*ii*)*:* $A_c \cap A_d = \theta$, or $A_c \cap A_d \neq \theta$ and either

$$T_c, T_d \;\varepsilon\; A_c \,\Delta\, A_d,$$

or

$$T_d \varepsilon A_c \cap A_d , \qquad T_c \varepsilon A_c \Delta A_d ,$$

or

$$T_c \varepsilon A_c \cap A_d , \qquad T_d \varepsilon A_c \Delta A_d ,$$

or

$$T_c , T_d \varepsilon A_c \cap A_d , \qquad T_c < T_d .$$

In all these cases $\psi\varphi(c,y) = \psi(c,\psi y) = \varphi(c,z)$, whence

$$\varphi(c,z) \supseteq \varphi(c,y).$$

REFERENCES

1. Weber, J. H., Some Traffic Characteristics of Communication Networks with Automatic Alternate Routing, B.S.T.J., *41,* July, 1962, pp. 1201–1247.
2. Beneš, V. E., Programming and Control Problems Arising from Optimal Routing in Telephone Networks, B.S.T.J., *45,* November, 1966, pp. 1373–1438; Abstract in SIAM J. Control, *4,* 1966, pp. 6–18.
3. Wilkinson, R. I., The Interconnection of Telephone Systems—Graded Multiples, B.S.T.J., *10,* October, 1931, pp. 531–564.
4. Syski, R., *Introduction to Congestion Theory in Telephone Systems,* Oliver and Boyd, Edinburgh and London, 1960.
5. Weber, J. H., private communication.

Integral Equation for Simultaneous Diagonalization of Two Covariance Kernels

By T. T. KADOTA

(Manuscript received January 20, 1967)

Let $K_1(s,t)$ and $K_2(s,t)$, $-T \leqq s, t \leqq T$, be real, symmetric, continuous and strictly positive-definite kernels, and denote by K_1 and K_2 the corresponding integral operators. Let $x(t)$ be a sample function of either of two zero-mean processes with covariances $K_1(s,t)$ and $K_2(s,t)$. We prove a generalized version of the following: If the integral equation

$$(K_2\psi_i)(t) = \lambda_i(K_1\psi_i)(t), \qquad -T \leqq t \leqq T,$$

has formal solutions λ_i and $\psi_i(t)$ which may contain δ-functions, and if $\{K_1\psi_i\}$ forms a complete set in $\mathfrak{L}_2[-T,T]$, then (i) the two kernels have the following simultaneous diagonalization:

$$K_1(s,t) = \sum_i (K_1\psi_i)(s)(K_1\psi_i)(t),$$

$$K_2(s,t) = \sum_i \lambda_i(K_1\psi_i)(s)(K_1\psi_i)(t),$$

uniformly on $[-T,T] \times [-T, T]$, and (ii) the sample function has an expansion

$$x(t) = \sum_i (x,\psi_i)(K_1\psi_i)(t)$$

in the stochastic mean, uniformly in t, and the coefficients are simultaneously orthogonal, i.e.,

$$E_1\{(x,\psi_i)(x,\psi_j)\} = \delta_{ij} , \qquad E_2\{(x,\psi_i)(x,\psi_j)\} = \lambda_i \, \delta_{ij} ,$$

where (x,ψ_i) is obtained by formally integrating $\psi_i(t)$ against $x(t)$.

I. INTRODUCTION

Let $K_1(s,t)$ and $K_2(s,t)$, $-T \leqq s, t \leqq T$, be real, symmetric, continuous and strictly positive-definite kernels, and denote by K_1 and

K_2 the integral operators with kernels $K_1(s,t)$ and $K_2(s,t)$. We have previously[1] established that, if $K_1^{-\frac{1}{2}}K_2K_1^{-\frac{1}{2}}$ is a densely defined and bounded operator on $\mathfrak{L}_2$ (the space of all square-integrable functions on $[-T,T]$) and if its extension to the whole of $\mathfrak{L}_2$ has eigenvalues λ_i and complete orthonormal eigenfunctions $\varphi_i(t)$, $i = 0, 1, \cdots$, then the two kernels have the following simultaneous diagonalization:

$$\begin{aligned} K_1(s,t) &= \sum_i (K_1^{\frac{1}{2}}\varphi_i)(s)(K_1^{\frac{1}{2}}\varphi_i)(t), \\ K_2(s,t) &= \sum_i \lambda_i(K_1^{\frac{1}{2}}\varphi_i)(s)(K_1^{\frac{1}{2}}\varphi_i)(t) \end{aligned} \tag{1}$$

uniformly on $[-T,T] \times [-T,T]$. In addition, if $x(t)$ is a sample function of either of two (separable and measurable) zero-mean processes with covariances $K_1(s,t)$ and $K_2(s,t)$ with associated measures P_1 and P_2, then

$$x(t) = \sum_i \eta_i(x)(K_1^{\frac{1}{2}}\varphi_i)(t) \tag{2}$$

in the stochastic mean, uniformly in t. Moreover,*

$$E_1\{\eta_i(x)\eta_j(x)\} = \delta_{ij}, \qquad E_2\{\eta_i(x)\eta_j(x)\} = \lambda_i\, \delta_{ij},$$

where†

$$\eta_i(x) = \lim_{n\to\infty} (x, K_1^{-\frac{1}{2}}\varphi_{in}) \tag{3}$$

in the stochastic mean, and $\{\varphi_{in}\}$ is any sequence of functions in the domain of $K_1^{-\frac{1}{2}}$ such that $\lim \| \varphi_i - \varphi_{in} \| = 0$.[1,2] Furthermore, if the two kernels have continuous $2r$th derivatives $(\partial^{2r}/\partial s^r \partial t^r)K_p(s,t)$, $p = 1, 2$, then (1) and (2) can be differentiated term-by-term r times while retaining the same senses of convergence.[1]

We remarked in Ref. 1 that, if φ_i is in the domain of $K_1^{-\frac{1}{2}}$, $\psi_i = K_1^{-\frac{1}{2}}\varphi_i$ satisfies the integral equation

$$(K_2\psi_i)(t) = \lambda_i(K_1\psi_i)(t), \qquad -T \leqq t \leqq T, \tag{4}$$

and

$$\begin{aligned} \eta_i(x) &= (x,\psi_i) \quad \text{a.s. (almost surely)}, \\ (K_1^{\frac{1}{2}}\varphi_i)(t) &= (K_1\psi_i)(t). \end{aligned} \tag{5}$$

Slepian (private communication) has long conjectured that, if (4) admits formal solutions λ_i and ψ_i, $i = 0, 1, \cdots$, where ψ_i may contain

* E_p, $p = 1, 2$, denotes the expectation with respect to P_p.

† For any f, $g \in \mathfrak{L}_2$, (f,g) denotes the inner product of f and g, and $\| f \|$ the norm of f.

δ-functions and their derivatives, then the expansion coefficients and functions of (2) are given by formally substituting such ψ_i into (5).* This conjecture, proved here, is significant since it provides a concrete means of obtaining the expansions (1) and (2). To illustrate the point, consider the following pair of covariance kernels:

$$K_1(s,t) = e^{-\alpha|s-t|}, \qquad K_2(s,t) = e^{-\beta|s-t|}.$$

For this pair, (4) admits the following formal solutions[5]

$$\begin{aligned} \tilde{\psi}_{2k}(t) &= \cos\theta_k t + \frac{\cos\theta_k T}{\alpha+\beta}\,[\delta(t-T) + \delta(t+T)], \\ &\qquad\qquad\qquad\qquad\qquad k = 0, 1, \cdots, \\ \tilde{\psi}_{2k+1}(t) &= \sin\hat{\theta}_k t + \frac{\sin\hat{\theta}_k T}{\alpha+\beta}\,[\delta(t-T) - \delta(t+T)], \end{aligned} \tag{6}$$

corresponding to

$$\lambda_{2k} = \frac{\beta}{\alpha}\,\frac{\alpha^2+\theta_k^2}{\beta^2+\theta_k^2}, \qquad \lambda_{2k+1} = \frac{\beta}{\alpha}\,\frac{\alpha^2+\hat{\theta}_k^2}{\beta^2+\hat{\theta}_k^2},$$

where θ_k and $\hat{\theta}_k$ are positive solutions of

$$\begin{aligned} (\alpha+\beta)\theta_k \tan\theta_k T &= \alpha\beta - \theta_k^2, \\ -(\alpha+\beta)\hat{\theta}_k \operatorname{ctn}\hat{\theta}_k T &= \alpha\beta - \hat{\theta}_k^2, \end{aligned} \tag{8}$$

respectively, indexed in ascending order. Thus, formally,

$$\begin{aligned} (x,\tilde{\psi}_{2k}) &= \int_{-T}^{T} x(t)\cos\theta_k t\,dt + \frac{\cos\theta_k T}{\alpha+\beta}\,[x(T) + x(-T)], \\ (x,\tilde{\psi}_{2k+1}) &= \int_{-T}^{T} x(t)\sin\hat{\theta}_k t\,dt + \frac{\sin\hat{\theta}_k T}{\alpha+\beta}\,[x(T) - x(-T)], \end{aligned} \tag{9}$$

$$\begin{aligned} (K_1\tilde{\psi}_{2k})(t) &= \frac{2\alpha}{\alpha^2+\theta_k^2}\cos\theta_k t, \\ (K_2\tilde{\psi}_{2k+1})(t) &= \frac{2\alpha}{\alpha^2+\hat{\theta}_k^2}\sin\hat{\theta}_k t. \end{aligned} \tag{10}$$

Through a direct calculation, we previously[5] established that

(*i*) $K_1^{-\frac{1}{2}}K_2K_1^{-\frac{1}{2}}$ is densely defined and bounded,

(*ii*) its extension has eigenvalues λ_i given by (7) and complete

* Similar conjectures have been made elsewhere.[3,4]

orthonormal eigenfunctions φ_i given as

$$\varphi_i = c_i \ \text{l.i.m.} \sum_{j=0}^{n} \mu_{1j}^{\frac{1}{2}}(\psi_i \, , f_{1j})f_{1j} \ ,$$

(iii) $\eta_i = c_i(x,\tilde{\psi}_i)$ a.s.,* $K_1^{\frac{1}{2}}\varphi_i = c_i K_1 \tilde{\psi}_i$, which verifies Slepian's conjecture for this example. Here c_i is a normalization constant given by

$$c_{2k} = \left[\frac{2\alpha}{\alpha^2 + \theta_k^2}\left(T + \frac{(\alpha + \beta)\alpha\beta}{\theta_k^4 + (\alpha^2 + \beta^2)\theta_k^2 + \alpha^2\beta^2}\right)\right]^{-\frac{1}{2}},$$

$$c_{2k+1} = c_{2k} \, \big|_{\theta_k = \hat{\theta}_k} \ ,$$

(that is, c_{2k+1} is obtained by replacing θ_k with $\hat{\theta}_k$ in c_{2k}), μ_{pj} and f_{pj} , $p = 1, 2,` j = 0, 1, \cdots$, are the eigenvalues and orthonormal eigenfunctions of K_p , and $(\psi_i \, , f_{1j})$ is defined analogously to (9).

In this paper we prove the generalization of *(i)*, *(ii)*, and *(iii)*, starting with abstract kernels $K_1(s,t)$ and $K_2(s,t)$ and a generalized version of the integral equation (4).

II. MAIN RESULT

Theorem: Let $K_p(s,t)$, $p = 1, 2$, $-T \leqq s, t \leqq T$, be real, symmetric, strictly positive-definite kernels with continuous 2rth derivatives $(\partial^{2r}/\partial s^r \partial t^r)K_p(s,t)$. If there exist sequences of real numbers $\{a_{ilm}\}$, $\{t_m\}$: $-T \leqq t_m \leqq T$, and $\{\lambda_i\}$:

$$0 < b_1 \leqq \lambda_i \leqq b_2 \ , \qquad i = 0, 1, \cdots , \tag{11}$$

for some constants b_1 and b_2 , and sequences of square-integrable functions $\{\psi_{il}\}$, which satisfy the equation

$$\begin{aligned}\sum_{l=0}^{r}\left[\int_{-T}^{T}\left(\frac{\partial^l}{\partial t^l} K_2(s,t)\right)\psi_{il}(t) \, dt + \sum_{m=1}^{q} a_{ilm} \frac{\partial^l}{\partial t^l} K_2(s,t) \, \bigg|_{t=t_m}\right] \\ = \lambda_i \sum_{l=0}^{r}\left[\int_{-T}^{T}\left(\frac{\partial^l}{\partial t^l} K_1(s,t)\right)\psi_{il}(t) \, dt + \sum_{m=1}^{q} a_{ilm} \frac{\partial^l}{\partial t^l} K_1(s,t) \, \bigg|_{t=t_m}\right], \\ -T \leqq s \leqq T,\end{aligned} \tag{12}$$

such that the right-hand side of (12) forms a complete set in $\mathfrak{L}_2$, then

(i) $K_1^{-\frac{1}{2}}K_2K_1^{-\frac{1}{2}}$ is a densely defined and bounded operator on $\mathfrak{L}_2$,

(ii) its extension to the whole of $\mathfrak{L}_2$ has eigenvalues and complete orthonormal eigenfunctions, which are the λ_i and

$$\varphi_i(s) = \sum_{l=0}^{r}\left[(K_{10l}^{\frac{1}{2}}\psi_{il})(s) + \sum_{m=1}^{q} a_{ilm}K_{10l}^{\frac{1}{2}}(s,t_m)\right] , \tag{13}$$

* This portion is proved in a separate article.[6]

(*iii*) *η_i and $K_1^{\frac{1}{2}}\varphi_i$ of* (2) *can be given, respectively, by*

$$\eta_i(x) = \sum_{l=0}^{r} \left[(x^{(l)},\psi_{il}) + \sum_{m=1}^{q} a_{ilm} x^{(l)}(t_m) \right] \quad \text{a.s.} \tag{14}$$

and by the right-hand side of (12) *without λ_i. Here, $K^{\frac{1}{2}}_{p0l}$, $p = 1, 2$, denotes an integral operator whose kernel is defined as*

$$K^{\frac{1}{2}}_{p0l}(s,t) = \sum_{i} \mu^{\frac{1}{2}}_{pi} f_{pi}(s) f_{pi}^{(l)}(t) \qquad l = 0, 1, \cdots, r, \tag{15}$$

in the mean in s, uniformly in t.

Remarks:

(*i*) $K^{\frac{1}{2}}_{p0l}(s,t)$ of (15) is well defined since

$$\sum_{i=0}^{\infty} \mu_{pi} f_{pi}^{(k)}(s) f_{pi}^{(l)}(t) = \frac{\partial^{k+l}}{\partial s^k\, \partial t^l} K_p(s,t), \qquad p = 1, 2, \tag{16}$$

uniformly in (s,t).[7] It follows from this that (15) converges in the mean in (s,t) as well. Hence, from Fubini's theorem, $K^{\frac{1}{2}}_{10l}(s,t)$ is a square-integrable function of t for almost every s. Thus, $\varphi_i(s)$ of (13) is well defined. We assume without loss of generality that φ_i, $i = 0, 1, \cdots$, are normalized.

(*ii*) For the example in Section I, $r = 0$, $q = 2$, $t_1 = T$, $t_2 = -T$, and

$$\psi_{2k,0}(t) = c_{2k} \cos \theta_k t, \qquad \psi_{2k+1,0}(t) = c_{2k+1} \sin \hat{\theta}_k t,$$

$$a_{2k,0,1} = a_{2k,0,2} = c_{2k} \frac{\cos \theta_k T}{\alpha + \beta},$$

$$a_{2k+1,0,1} = -a_{2k+1,0,2} = c_{2k+1} \frac{\sin \hat{\theta}_k T}{\alpha + \beta},$$

$$b_1 = \frac{\alpha}{\beta}, \qquad b_2 = \frac{\beta}{\alpha},$$

the right-hand side of (12) without λ_i is given by (10), and completeness of $\{\cos \theta_k t, \sin \hat{\theta}_k t\}$ follows from (18) and a gap-and-density theorem.[8]

III. PROOF OF THEOREM

For notational simplicity, we write K_{pkl}, $p = 1, 2$, for the integral operator whose kernel is

$$K_{pkl}(u,v) = \frac{\partial^{k+l}}{\partial u^k\, \partial v^l} K_p(u,v), \qquad k, l = 0, 1, \cdots, r.$$

K_{p00} and $K^{\frac{1}{2}}_{p00}$ are abbreviated as before by K_p and $K^{\frac{1}{2}}_p$, respectively.

(*i*) For any f, $g \epsilon \mathfrak{L}_2$,

$$(K_{p0k}^{\frac{1}{2}}f, K_{p0l}^{\frac{1}{2}}g) = (f, K_{pkl}g), \tag{17}$$

$$K_{p0l}^{\frac{1}{2}}g = \underset{n\to\infty}{\text{l.i.m.}} \sum_{j=0}^{n} \mu_{pj}^{\frac{1}{2}} f_{pj}(f_{pj}^{(l)}, g). \tag{18}$$

To prove (17), note

$$\begin{aligned}(K_{p0k}^{\frac{1}{2}}f, K_{p0l}^{\frac{1}{2}}g) &= \iiint_{-T}^{T} f(s)g(t)K_{p0k}^{\frac{1}{2}}(u,s)K_{p0l}^{\frac{1}{2}}(u,t)\,ds\,dt\,du \\ &= \int_{-T}^{T}\int f(s)g(t) \sum_{j} \mu_{pj} f_{pj}^{(k)}(s) f_{pj}^{(l)}(t)\,ds\,dt \\ &= (f, K_{pkk}g),\end{aligned}$$

where the second equality follows from the mean convergence of (15) and the third from the uniform convergence of (16). To prove (18), consider

$$\left\| K_{p0l}^{\frac{1}{2}}g - \sum_{j=0}^{n} \mu_{pj}^{\frac{1}{2}} f_{pj}(f_{pj}^{(l)}, g) \right\|^2 = (g, K_{pll}g) - \sum_{j=0}^{n} \mu_{pj}(f_{pj}^{(l)}, g)^2,$$

which vanishes as $n \to \infty$ since (16) converges uniformly in (s,t).

(*ii*) $K_2^{-\frac{1}{2}}K_1^{\frac{1}{2}}$ and $K_1^{-\frac{1}{2}}K_2^{\frac{1}{2}}$ are densely defined and bounded on $\mathfrak{L}_2$.

To prove this, apply $K_2^{-\frac{1}{2}}$ on both sides of (12) and use (18) to obtain

$$\sum_{l=0}^{r}\left[K_{20l}^{\frac{1}{2}}\psi_{il} + \sum_{m=1}^{q} a_{ilm}K_{20l}^{\frac{1}{2}}(\cdot, t_m)\right] = \lambda_i K_2^{-\frac{1}{2}}K_1^{\frac{1}{2}}\varphi_i .$$

Then, for each i,

$$\begin{aligned}&\lambda_i^2 \,\| K_2^{-\frac{1}{2}}K_1^{\frac{1}{2}}\varphi_i \|^2 \\ &\quad= \sum_{k,l=0}^{r}\Bigg\{(K_{20k}^{\frac{1}{2}}\psi_{ik}, K_{20l}^{\frac{1}{2}}\psi_{il}) \\ &\qquad + \sum_{m=1}^{q}[a_{ilm}(K_{20l}^{\frac{1}{2}}(\cdot, t_m), K_{20k}^{\frac{1}{2}}\psi_{ik}) + a_{ikm}(K_{20k}^{\frac{1}{2}}(\cdot, t_m), K_{20l}^{\frac{1}{2}}\psi_{il})] \\ &\qquad + \sum_{m,n=1}^{q} a_{ilm}a_{ikn}(K_{20l}^{\frac{1}{2}}(\cdot, t_m), K_{20k}^{\frac{1}{2}}(\cdot, t_n))\Bigg\} \\ &\quad= \sum_{k,l=0}^{r}\Bigg\{\Bigg(\psi_{ik}, K_{2kl}\psi_{il} + \sum_{m=1}^{q} a_{ilm}K_{2kl}(\cdot, t_m)\Bigg) \\ &\qquad + \sum_{n=1}^{q} a_{ikn}\Bigg[(K_{2kl}\psi_{il})(t_n) + \sum_{m=1}^{q} a_{ilm}K_{2kl}(t_n, t_m)\Bigg]\Bigg\}\end{aligned}$$

$$= \lambda_i \sum_{k,l=0}^{r} \left[\left(K_{10k}^{\frac{1}{2}}\psi_{ik}\,,\, K_{10l}^{\frac{1}{2}}\psi_{il} + \sum_{m=1}^{q} a_{ilm}K_{10l}^{\frac{1}{2}}(\cdot\,,t_m)\right)\right.$$

$$\left. + \sum_{n=1}^{q} a_{ikn}\left(K_{10k}^{\frac{1}{2}}(\cdot\,,t_n),\, K_{10l}^{\frac{1}{2}}\psi_{il} + \sum_{m=1}^{q} a_{ilm}K_{10l}^{\frac{1}{2}}(\cdot\,,t_m)\right)\right]$$

$$= \lambda_i \,\|\, \varphi_i \,\|^2,$$

where the second equality follows from (17) and (18), the third from k time differentiation of (12) and from (17) and (18), and the last from (13). Hence, with φ_i being normalized,

$$\|\, K_2^{-\frac{1}{2}}K_1^{\frac{1}{2}}\varphi_i \,\|^2 = \frac{1}{\lambda_i}\,, \qquad i = 0, 1, \cdots .$$

Now $\{\varphi_i\}$ is complete since the right-hand side of (12) without λ_i, which forms a complete set by hypothesis, is equal to $K_1^{\frac{1}{2}}\varphi_i$, and $K_1^{\frac{1}{2}}$ is strictly positive-definite. Hence, from (11), $K_2^{-\frac{1}{2}}K_1^{\frac{1}{2}}$ is densely defined and bounded.

To prove that $K_1^{-\frac{1}{2}}K_2^{\frac{1}{2}}$ is also densely defined and bounded, define $\hat{\varphi}_i$ as the normalized right-hand side of (13) with the subscript 1 replaced by 2. Completeness of $\{\hat{\varphi}_i\}$ is similarly deduced via (12). Now, by following the same procedure with the roles of K_1 and K_2 interchanged, we obtain

$$\|\, K_1^{-\frac{1}{2}}K_2^{\frac{1}{2}}\hat{\varphi}_i \,\|^2 = \lambda_i\,, \qquad i = 0, 1, \cdots .$$

Then, the assertion follows immediately from (11).

(*iii*) The ranges of $K_1^{\frac{1}{2}}$ and $K_2^{\frac{1}{2}}$ are equal, namely,

$$K_1^{\frac{1}{2}}(\mathcal{L}_2) = K_2^{\frac{1}{2}}(\mathcal{L}_2).$$

To prove this, denote by L and M the extensions to the whole of $\mathcal{L}_2$ of $K_2^{-\frac{1}{2}}K_1^{\frac{1}{2}}$ and $K_1^{-\frac{1}{2}}K_2^{\frac{1}{2}}$ respectively, which exist as a result of (*ii*). Since the domains of $K_2^{\frac{1}{2}}L$ and $K_1^{\frac{1}{2}}M$ are $\mathcal{L}_2$, which is also the domains of $K_1^{\frac{1}{2}}$ and $K_2^{\frac{1}{2}}$, we have

$$K_1^{\frac{1}{2}} = K_2^{\frac{1}{2}}L, \qquad K_2^{\frac{1}{2}} = K_1^{\frac{1}{2}}M.$$

Then, from the first equality, $K_1^{\frac{1}{2}}(\mathcal{L}_2) \subset K_2^{\frac{1}{2}}(\mathcal{L}_2)$, while, from the second, $K_2^{\frac{1}{2}}(\mathcal{L}_2) \subset K_1^{\frac{1}{2}}(\mathcal{L}_2)$. Hence, the assertion holds.

(*iv*)

$$K_{20l}^{\frac{1}{2}}(\cdot\,,t) = \operatorname*{l.i.m.}_{n\to\infty} \sum_{j=0}^{n} K_2^{\frac{1}{2}}f_{1j}f_{1j}^{(l)}(t), \qquad -T \leqq t \leqq T, \tag{19}$$

$$K_{20l}^{\frac{1}{2}}g = \operatorname*{l.i.m.}_{n\to\infty} \sum_{j=0}^{n} K_2^{\frac{1}{2}}f_{1j}(f_{1j}^{(l)}, g), \qquad g \in \mathcal{L}_2\,. \tag{20}$$

To prove (19), note first that f_{1j}, $j = 0, 1, \cdots$, are in the domain of $K_2^{-\frac{1}{2}}$ as a result of (*iii*) and also that $(K_2^{-\frac{1}{2}} f_{1i}, K_2^{\frac{1}{2}} f_{1j}) = \delta_{ij}$ from orthonormality of $\{f_{1j}\}$. Thus, $\{K_2^{-\frac{1}{2}} f_{1j}\}$ and $\{K_2^{\frac{1}{2}} f_{1j}\}$ form a pair of mutually reciprocal bases of $\mathfrak{L}_2$. Hence,

$$K_{20l}^{\frac{1}{2}}(\cdot, t) = \underset{n\to\infty}{\text{l.i.m.}} \sum_{j=0}^{n} K_2^{\frac{1}{2}} f_{1j} (K_2^{-\frac{1}{2}} f_{1j}, K_{20l}^{\frac{1}{2}}(\cdot, t)). \tag{21}$$

But from (15)

$$(K_2^{-\frac{1}{2}} f_{1j}, K_{20l}^{\frac{1}{2}}(\cdot, t)) = \sum_{i=0}^{\infty} (f_{1j}, f_{2i}) f_{2i}^{(l)}(t), \qquad l = 0, 1, \cdots, r, \tag{22}$$

uniformly in t. Now, since $\{f_{2i}\}$ is an orthonormal basis of $\mathfrak{L}_2$,

$$f_{1j} = \underset{n\to\infty}{\text{l.i.m.}} \sum_{i=0}^{n} (f_{1j}, f_{2i}) f_{2i}.$$

But, according to (22), the right-hand side converges uniformly. Hence, the above partial sum must converge uniformly to f_{1j}. Suppose for some k, $0 \leqq k < r$,

$$f_{1j}^{(k)}(t) = \sum_{i=0}^{\infty} (f_{1j}, f_{2i}) f_{2i}^{(k)}(t) \tag{23}$$

uniformly in t. Then, from (22),

$$f_{1j}^{(k+1)}(t) = \sum_{i=0}^{\infty} (f_{1j}, f_{2i}) f_{2i}^{(k+1)}(t)$$

uniformly in t.[9] Hence, by induction, (23) holds for every k, $0 \leqq k \leqq r$. Therefore, from (22),

$$(K_2^{-\frac{1}{2}} f_{1j}, K_{20l}^{\frac{1}{2}}(\cdot, t)) = f_{1j}^{(l)}(t), \qquad l = 0, 1, \cdots, r.$$

Then, (19) follows from (21) and the above.

To prove (20), we expand $K_{20l}^{\frac{1}{2}} g$ relative to $\{K_2^{\frac{1}{2}} f_{1j}\}$:

$$K_{20l}^{\frac{1}{2}} g = \underset{n\to\infty}{\text{l.i.m.}} \sum_{j=0}^{n} (K_2^{-\frac{1}{2}} f_{1j}, K_{20l}^{\frac{1}{2}} g) K_2^{\frac{1}{2}} f_{1j},$$

and note from (18) and (23) that

$$(K_2^{-\frac{1}{2}} f_{1j}, K_{20l}^{\frac{1}{2}} g) = \sum_{i=0}^{\infty} (f_{1j}, f_{2i})(f_{2i}^{(l)}, g) = (f_{1j}^{(l)}, g).$$

(*v*) To prove (*i*) of the theorem, we note from (*ii*) and (*iii*) that $K_1^{-\frac{1}{2}} K_2^{\frac{1}{2}}$ is everywhere-defined and bounded on $\mathfrak{L}_2$. Hence, its adjoint $(K_1^{-\frac{1}{2}} K_2^{\frac{1}{2}})^*$ is also everywhere-defined and bounded. Now, for any

$f \epsilon \mathfrak{L}_2$ and $g \epsilon \mathfrak{D}(K_1^{-\frac{1}{2}})$, the domain of $K_1^{-\frac{1}{2}}$, we have $(K_1^{-\frac{1}{2}}K_2^{\frac{1}{2}}f,g) = (f,K_2^{\frac{1}{2}}K_1^{-\frac{1}{2}}g)$. Thus, $K_2^{\frac{1}{2}}K_1^{-\frac{1}{2}}g = (K_1^{-\frac{1}{2}}K_2^{\frac{1}{2}})^*g$, $g \epsilon \mathfrak{D}(K_1^{-\frac{1}{2}})$. Hence, $K_2^{\frac{1}{2}}K_1^{-\frac{1}{2}}$ is bounded. Since $\mathfrak{D}(K_1^{-\frac{1}{2}})$ is dense in $\mathfrak{L}_2$, we conclude that $K_1^{-\frac{1}{2}}K_2K_1^{-\frac{1}{2}}$ is densely defined and bounded.

(*vi*) To prove (*ii*) of the theorem, define

$$\varphi_{in}(t) = \sum_{j=0}^{n} \mu_{1j}^{\frac{1}{2}} \sum_{l=0}^{r} \left[(\psi_{il}, f_{1j}^{(l)}) + \sum_{m=1}^{q} a_{ilm} f_{1j}^{(l)}(t_m) \right] f_{1j}(t), \tag{24}$$

and note $\varphi_{in} \epsilon \mathfrak{D}(K_1^{-\frac{1}{2}})$ and $\lim_{n\to\infty} \| \varphi_i - \varphi_{in} \| = 0$. Then

$$\begin{aligned} \underset{n\to\infty}{\text{l.i.m.}}\, K_2K_1^{-\frac{1}{2}}\varphi_{in} &= \underset{n\to\infty}{\text{l.i.m.}} \sum_{j=0}^{n} \sum_{l=0}^{r} \left[(\psi_{il}, f_{1j}^{(l)}) + \sum_{m=1}^{q} a_{ilm} f_{1j}^{(l)}(t_m) \right] K_2 f_{1j} \\ &= \sum_{l=0}^{r} \left[K_{20l}\psi_{il} + \sum_{m=1}^{q} a_{ilm} K_{20l}(\cdot, t_m) \right] \\ &= \lambda_i K_1^{\frac{1}{2}} \varphi_i , \end{aligned}$$

where the second equality follows from (19), (20), (15) and (18), and third from (12) and (13). Now denote by Q the extension of $K_1^{-\frac{1}{2}}K_2K_1^{-\frac{1}{2}}$ to the whole of $\mathfrak{L}_2$. Then,

$$K_1^{\frac{1}{2}}Qf = \underset{n\to\infty}{\text{l.i.m.}}\, K_2K_1^{-\frac{1}{2}}f_n$$

for any $f \epsilon \mathfrak{L}_2$ and $\{f_n\}: f_n \epsilon \mathfrak{D}(K_1^{-\frac{1}{2}})$, $\lim \| f - f_n \| = 0$, since

$$\| K_1^{\frac{1}{2}}Qf - K_2K_1^{-\frac{1}{2}}f_n \| \leqq \| K_1^{\frac{1}{2}}Q(f - f_n) \| + \| (K_1^{\frac{1}{2}}Q - K_2K_1^{-\frac{1}{2}})f_n \|$$

which vanishes as $n \to \infty$. Therefore, $Q\varphi_i = \lambda_i\varphi_i$. Lastly, since $\{\varphi_i\}$ is complete in $\mathfrak{L}_2$, $\{\lambda_i\}$ constitutes the entire spectrum of Q.

(*vii*) To prove (*iii*) of the theorem, note from (3), (24) and (*vi*) that

$$\eta_i(x) = \underset{n\to\infty}{\text{l.i.m.}} \sum_{j=0}^{n} \sum_{l=0}^{r} \left[(\psi_{il}, f_{1j}^{(l)}) + \sum_{m=1}^{q} a_{ilm} f_{1j}^{(l)}(t_m) \right] (f_{1j}, x).$$

Now

$$\begin{aligned} E_1 \left| (x^{(l)}, \psi_{il}) - \sum_{j=0}^{n} (x, f_{1j})(f_{1j}^{(l)}, \psi_{il}) \right|^2 \\ = (\psi_{il}, K_{1ll}\psi_{il}) - \sum_{j=0}^{n} \mu_{1j}(\psi_{il}, f_{1j}^{(l)})^2, \end{aligned}$$

$$E_1 \left| x^{(l)}(t) - \sum_{j=0}^{n} (x, f_{1j}) f_{1j}^{(l)}(t) \right|^2 = K_{1ll}(t,t) - \sum_{j=0}^{n} \mu_{1j}[f_{1j}^{(l)}(t)]^2,$$

both of which vanish as $n \to \infty$ by virtue of (16). Also, with the use of (17) and (18),

$$E_2 \left| (x^{(l)}, \psi_{il}) - \sum_{j=0}^{n} (x, f_{1j})(f_{1j}^{(l)}, \psi_{il}) \right|^2$$

$$= (\psi_{il}\,, K_{2ll}\psi_{il}) - 2 \sum_{j=0}^{n} (\psi_{il}\,, f_{1j}^{(l)})(f_{1j}\,, K_{20l}\psi_{il})$$

$$+ \sum_{j,k=0}^{n} (\psi_{il}\,, f_{1j}^{(l)})(\psi_{il}\,, f_{1k}^{(l)})(f_{1j}\,, K_2 f_{1k})$$

$$= \left\| K_{20l}^{\frac{1}{2}}\psi_{il} - \sum_{j=0}^{n} K_2^{\frac{1}{2}} f_{1j}(f_{1j}^{(l)}, \psi_{il}) \right\|^2,$$

$$E_2 \left| x^{(l)}(t) - \sum_{j=0}^{n} (x, f_{1j}) f_{1j}^{(l)}(t) \right|^2 = K_{2ll}(t,t) - 2 \sum_{j=0}^{n} f_{1j}^{(l)}(t)(K_{2l0} f_{1j})(t)$$

$$+ \sum_{j,k=0}^{n} f_{1j}^{(l)}(t) f_{1k}^{(l)}(t)(f_{1j}\,, K_2 f_{1k}) = \left\| K_{20l}^{\frac{1}{2}}(\cdot\,, t) - \sum_{j=0}^{n} K_2^{\frac{1}{2}} f_{1j} f_{1j}^{(l)}(t) \right\|^2,$$

both of which vanish as $n \to \infty$ by virtue of (19) and (20). Therefore, upon combination of the above results, (14) is proved.

REFERENCES

1. Kadota, T. T., Simultaneous Diagonalization of Two Covariance Kernels and Application to Second-order Stochastic Processes, submitted for publication in SIAM J. Appl. Math.
2. Root, W. L., Singular Gaussian Measures in Detection Theory, Proc. Symp. on Time Series Analysis, edited by M. Rosenblatt, John Wiley & Sons, New York, 1963, pp. 292–315.
3. Yaglom, A. M., On the Equivalence and Perpendicularity of Two Gaussian Measures in Function Space, Proc. Symp. on Time Series Analysis, edited by M. Rosenblatt, John Wiley & Sons, New York, 1963, pp. 327-348.
4. Huang, R. Y. and Johnson, R. A., Information Capacity of Time Continuous Channels, IRE Trans. Inform. Theory, *IT-8*, September, 1962, pp. 191–205.
5. Kadota, T. T., Simultaneously Orthogonal Expansion of Two Stationary Gaussian Processes-Examples, B.S.T.J., *45*, September, 1966, pp. 1071–1096.
6. Kadota, T. T. and Shepp, L. A., On the Best Finite Set of Linear Observables for Discriminating Two Gaussian Signals, IEEE Trans. Inform. Theory, April, 1967.
7. Kadota, T. T., Term-by-term Differentiability of Mercer's Expansion, Proc. Am. Math. Soc., *18*, February, 1967, pp. 69–72.
8. Levinson, N., Gap and Density Theorems, Am. Math. Soc. Colloquium Publications, *26*, Am. Math. Soc., Providence, Rhode Island, 1940, p. 3, Theorem II.
9. Apostol, T. M., *Mathematical Analysis,* Addison-Wesley, Reading, Massachusetts, 1957, p. 403.

Principles of Design of Magnetic Devices for Attitude Control of Satellites

By M. S. GLASS

(Manuscript received December 29, 1966)

Magnetic devices mounted within an orbiting satellite interact with the earth's magnetic field and produce torque to modify the attitude or angular adjustment of the satellite axis of spin. The satellite environment dictates that these devices be designed for minimum weight or minimum power consumption, or a suitable compromise between these two minima. Principles of design of magnetic devices to satisfy these requirements are developed in this paper. The resulting design equations and charts enable the ready optimization of design and selection of preferred materials. While most of this work was directed initially at the Telstar® satellite project, the design charts and formulas are found useful in other areas of magnet design. Methods of magnetic measurement devised for the satellite are discussed.

I. INTRODUCTION

Satellites with directional instrumentation, such as the antenna system of the communications satellites, require attitude control to keep this instrumentation properly on target. For example, a spin imparted to the satellite at time of launch gives it a sort of gyroscopic stability. However, complete attitude control requires some available torque to correct the direction of the spin axis.

In the orbiting satellite the earth's gravitational field is balanced by centrifugal force, leaving the earth's magnetic field as a convenient means for interaction torque. Suitable interaction with the earth's magnetic field can be set up by electromagnets, or by air-core coils of large area, either of which can be turned on or off at will to provide attitude correction as needed. Small permanent magnets can be designed and installed to cancel out residual magnetic moment in the satellite, which if permitted to interact with the earth's field could cause precession of the spin axis. Other miscellaneous torque applications of magnets in the satellite have been proposed and investigated.

Limitations of payload and of available power in the satellite generally make it necessary to design with quantitative accuracy and to optimize the factors which control weight and power consumption. To this end, the magnet designer may select from various geometries of magnet and coil and from various available materials. This selection and optimization is facilitated by the use of suitable design formulas and charts. In this paper, we review the derivation and illustrative use of such formulas and charts. While the work reported here has been aimed specifically at certain problems of the *Telstar*® satellite, it is evident that the technique of magnet design presented here is applicable to any similar set of problems.

For the convenience of the magnet designer who buys magnets and magnet wire by the pound, measures them in feet, inches, or mils, and measures torque in pound-inches, all of the derived design formulae and graphs are built around the practical units (inches, pounds, oersteds, gauss, etc.). This avoids the necessity of converting units, which is time consuming and can lead to costly errors. There is included for convenience a table of the most frequently used conversion factors (Table I).

II. QUANTITATIVE DESIGN OF AIR-CORE COIL FOR TORQUE

The torque characteristic of the air-core coil is derived from the galvanometer formula which, in some textbooks, is written in MKS units:

$$NIA \text{ (ampere-turn-meter}^2\text{)} = \frac{10^7}{4\pi}\,\frac{T_o}{H_a} \text{ (weber-meters)} \qquad (1)$$

Table I—Conversion Factors

1 unit pole (emu)	$= 4\pi$ maxwells
1 oersted	$= \frac{10^3}{4\pi}$ ampere turns per meter
	$= 2.02$ ampere turns per inch
1 weber-meter	$= \frac{10^{10}}{4\pi}$ emu
	$= \frac{8.85 \times 10^3}{4\pi}$ lb-in per oersted
1 newton-meter	$= 10^7$ dyne-cm
	$= 8.85$ lb-in

TABLE II — CHARACTERISTICS OF WINDINGS

	Copper	Aluminum
R(ohms)	$\frac{0.75 \times 10^{-6} N^2 P}{A}$	$\frac{1.21 \times 10^{-6} N^2 P}{A}$
E(volts)	$\frac{0.75 \times 10^{-6} NP(NI)}{A}$	$\frac{1.21 \times 10^{-6} NP(NI)}{A}$
W(watts)	$\frac{0.75 \times 10^{-6} P(NI)^2}{A}$	$\frac{1.21 \times 10^{-6} P(NI)^2}{A}$
Wgt.(lbs.)	$0.321\ AP$	$0.0983\ AP$
(Power × Wgt.)	$0.241 \times 10^{-6} P^2(NI)^2$	$0.119 \times 10^{-6} P^2(NI)^2$

NI: Required ampere turns
N: Number of turns used
A: Cross section area of winding (inch^2)
(N times the section area of a single turn)
P: Average length of turn in winding (inch)

and may be written in practical units:

$$NIA(\text{ampere-turn-inch}^2) = 1.667 \times 10^6 \frac{T_o}{H_a} \text{ lb-in/oersted}). \qquad (2)$$

Here T_o is the maximum torque exerted on the coil when its axis is perpendicular to the field H_a, and NIA is the required product of ampere turns and area enclosed by the coil to deliver that amount of torque.

It is convenient to set up a table of formulas from which one may translate the geometry and ampere-turn characteristics of the coil into power and weight requirements. The power and weight will depend upon the winding material used, but practical considerations usually limit this to copper or aluminum. So one may take the weight and resistivity characteristics of copper and aluminum from handbook tables and with the aid of Ohm's Law derive the formulas of Table II. Using (2) and Table II, one may estimate readily the power and weight of an air-core coil to satisfy specified torque requirements. It is evident that copper has the advantage in lower power consumption, but that aluminum offers a greater advantage in weight reduction. If power and weight are of about equal importance, then the power-weight product should be minimized. It is evident that aluminum has a factor-of-two advantage over copper in this characteristic.

III. QUANTITATIVE DESIGN OF MAGNETIZED BARS FOR TORQUE

A magnetized bar, either a permanent magnet or the core of an electromagnet, displays a moment, or normalized torque, proportional to the product of the volume of the bar by the intrinsic induction within the bar. The magnetic moment, M_m, is identified as normalized torque in the familar equation

$$M_m = \frac{T_o}{H_a} \qquad \text{(MKS)}, \tag{3}$$

and the relation between magnetic moment, intrinsic induction in the bar, and the geometry of the bar is given by another familiar equation

$$M_m = \frac{B - H}{4\pi} A \cdot S \qquad \text{(emu)} \tag{4}$$

in which $B - H$ is the intrinsic induction, A is the cross-section area of the bar at the median plane, and S is the effective distance between poles. For a magnet of length l and diameter d, one may define a shortening factor, $R_S = S/l$, which evaluates the effect of recession of the poles, and rewrite (4) as

$$M_m = \frac{B - H}{4\pi} AlR_S \; . \tag{5}$$

The shortening factor, R_S, has been evaluated by Okoshi.[1] Okoshi's values are plotted as a function of l/d of the bar in Fig. 1, with a

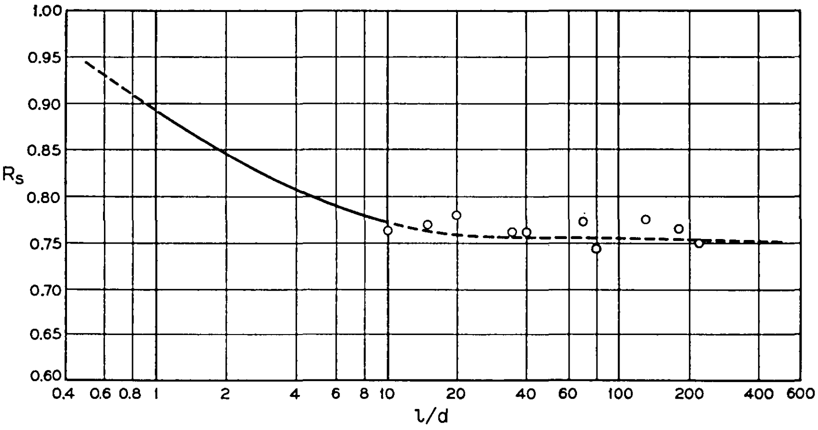


Fig. 1 — Effective shortening of magnets with increasing aspect ratio.

broken line extrapolation guided by experimental data. If one combines (3) and (5) after conversion to practical units, the result is

$$\frac{T_o}{H_a} \text{ (lb-in per oersted)} = 1.55 \times 10^{-6}(B - H)AlR_S \qquad (6)$$

and the required volume of bar to produce a specified magnetic moment is given by rearrangement of (6)

$$\text{Vol (in}^3\text{)} = \frac{1}{R_S}\frac{0.866 \times 10^6}{B - H}\frac{T_o}{H_a}. \qquad (7)$$

3.1 *Optimum Bar Shape—The Load Factor*

When a bar of ferromagnetic material is placed in a field of strength H_o, it assumes a state of magnetization which is commonly described by the equation

$$H = H_o - D_B(B - H)$$

which may also be written

$$D_B = \frac{H_o - H}{B - H}. \qquad (8)$$

Here H_o is the applied magnetizing field, and B and H describe the condition of magnetization within the bar. The demagnetizing factor, D_B, which is partially defined by (8) is used to express the dependence of the intrinsic induction within the bar upon the aspect ratio (l/d) of the bar. It has been tabulated and charted as a function of l/d by Okoshi,[1] Bozorth and Chapin,[2] and others. These sources agree upon the value of D_B for long slender magnets. For shorter magnets, where there is some disagreement, we find the Okoshi data to be in agreement with experiment.

In plotting the characteristics of magnetic materials we normally plot the intrinsic induction ($B - H$) or the flux density (B) as the dependent variable, and the field strength (H, or H_o) as the independent variable. Hence, the ratio

$$\frac{B - H}{H_o - H} = \frac{1}{D_B} = U \qquad (9)$$

becomes the slope of the generalized load line of the magnetized bar. This reciprocal of the demagnetizing factor is found useful in numerous magnetic calculations and possibly deserves a name and symbol of its own. We have elected to call it the loading factor, with the symbol U, and have plotted it as a function of l/d in Fig. 2. In the electromagnet core operating below saturation, H is generally negligibly small compared with either B or H_o, and the expression for the loading factor reduces to $U = B/H_o$. In the permanent magnet, H_o disappears and the expression for loading factor becomes $U = (B - H)/(-H)$. For long magnets, ($l/d > 5$) H is negligibly small compared with B, and the loading factor is further simplified to $U = B/(-H)$. In this restricted form the loading factor is identified with the "permeance coefficient" and similar terms used in the literature of permanent magnets.

In Fig. 3 we illustrate the application of the loading factor to the analysis of permanent magnets and electromagnets. For this illustration each is assumed to have $l/d \cong 33$ so that the loading factor, $U \cong 400$. In the permanent magnet (Remendur) the magnetomotive force

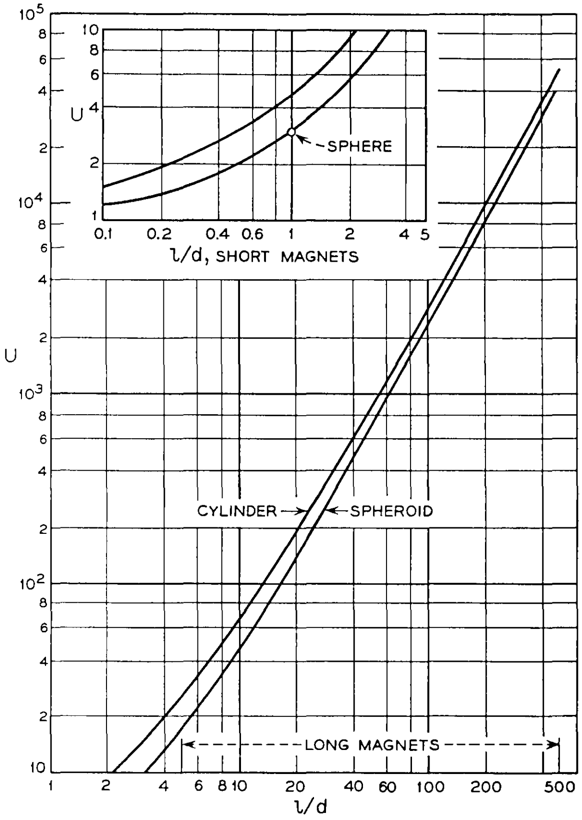


Fig. 2 — Variation of load factor with aspect ratio.

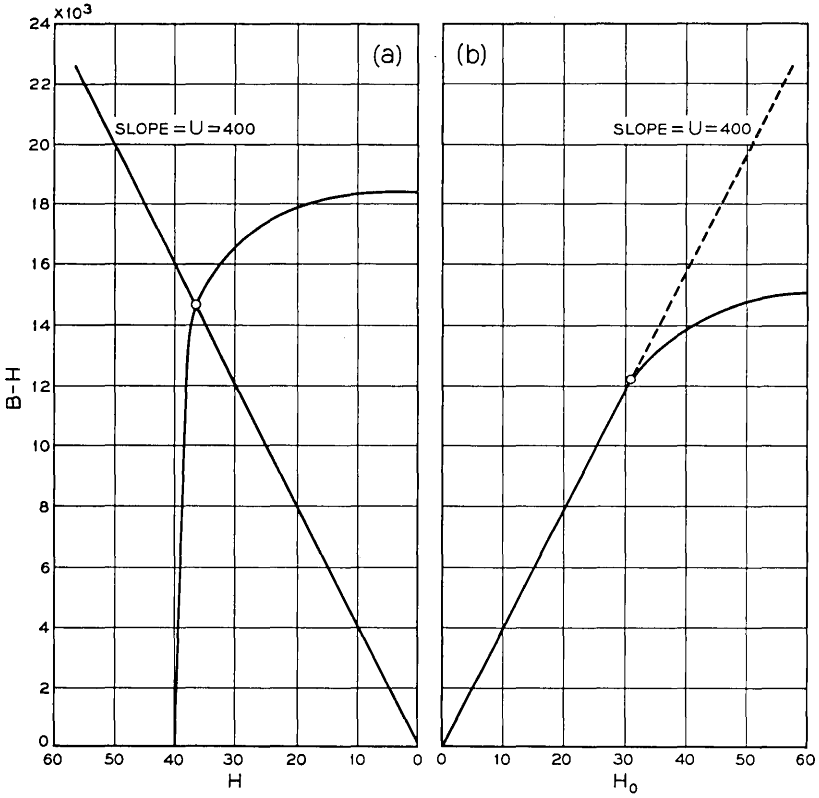


Fig. 3 — Application of load factor to magnet design: (a) permanent magnet (Remendur 38), (b) electromagnet (Permalloy 45 core).

is generated within the magnet and varies with the loading of the magnet as indicated by the B, H curve. (Here H is sufficiently small so that B is indistinguishable from $B - H$). The operating point is determined by the intersection of the B, H curve with the load line of slope U. This is a fixed point for a particular magnet with a particular condition of magnetization. The conditions of Fig. 3 were chosen to match the characteristics of Remendur. In the electromagnet when operated below saturation, H is negligibly small so the load line represents the relation between flux B in the bar, and applied field, H_o, up to the region (around $B = 10{,}000$ for Permalloy 45) where the core material starts saturation, and the characteristic starts deviating from the straight load line.

3.2 *Design of Permanent Magnets for High Torque-Weight Ratio*

The weight of the magnet in pounds is

$$W_m = Al\rho, \tag{10}$$

in which ρ is density of magnet material in lbs/in^3. One may combine (10) with (6) to obtain

$$\frac{T_o}{W_m H_a} = \frac{1.155 \times 10^{-6}}{\rho} (B - H) R_S \ . \tag{11}$$

Here the left-hand side of (11) is the normalized torque-weight ratio. The design objective is to maximize this ratio.

The dependence of the operating point of the permanent magnet upon the load factor, U, has been illustrated in Fig. 3. On similar charts one may plot intrinsic induction $(B - H)$ as a function of field (H) for various magnet materials as in Fig. 4. Values of B and H for these plots may be derived readily from the regular demagnetization curves supplied by magnet manufacturers. Then, for a particular value of

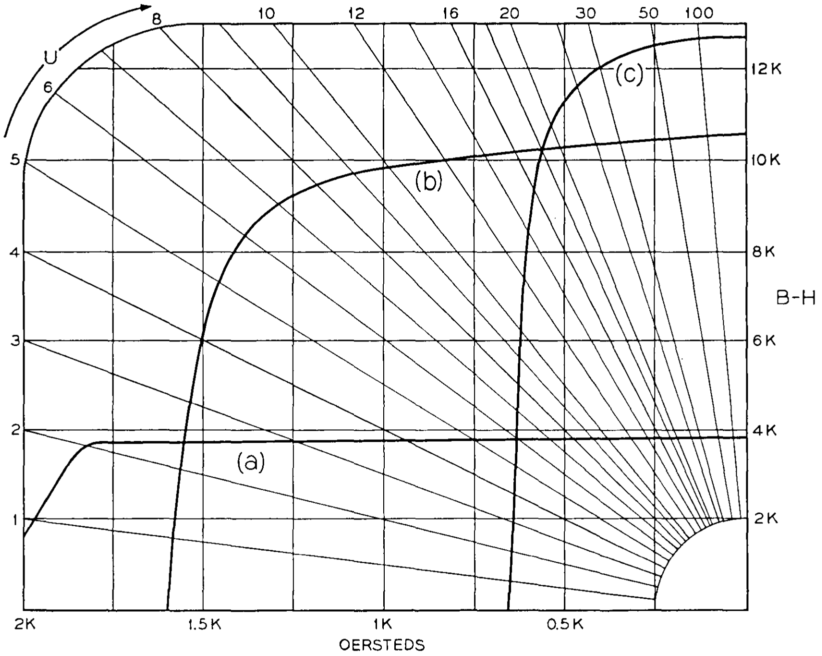


Fig. 4 — Intrinsic induction of permanent magnets; (a) directional grain ceramic, (b) Alnico 9, (c) Alnico 5.

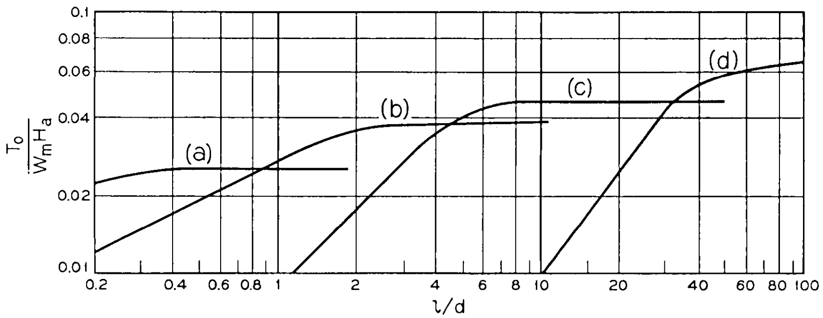


Fig. 5 — Variation of torque-weight ratio with aspect ratio; (a) directional grain ceramic, (b) Alnico 9, (c) Alnico 5, (d) Remendur 38.

l/d one may pick off the corresponding value of U from Fig. 2, and using this as the slope of the load line may find the value of $(B - H)$ for a particular magnet material at the point of intersection. This value of $(B - H)$ and the value of ρ appropriate to the material may be inserted in (11) to give the normalized torque-weight ratio. For example, at $l/d = 4$, $U \cong 17.5$. The load line of that slope intersects the intrinsic induction curve for Alnico 5 at $(B - H) \cong 10{,}000$. Inserting this value and the value of R_S in (11) and using $\rho = 0.26$ for Alnico, one obtains

$$\frac{T_o}{W_m H_a} = 3.6 \times 10^{-2}. \qquad (12)$$

Repeating this procedure for various values of l/d and for various materials, one can assemble the necessary data to plot the curves of Fig. 5.

It is evident that for each magnet material there is a value of l/d above which the torque-weight ratio is essentially constant, and below which the torque-weight ratio falls off rapidly with decreasing l/d. This follows the shape of the demagnetization curves of Fig. 4. This value of l/d is large for magnets having low coercivity, and small for magnets having high coercivity. One would normally design the magnet to operate on the flat part of the characteristic to obtain high torque-weight ratio.

3.3 *Design of Electromagnets for Torque*

The electromagnet is assumed to consist of a cylindrical core of ferromagnetic material with a solenoid wound around it. The design formula for the core is the same as that for the permanent magnet and is given in (7). This gives the required volume to produce a specified moment, operating the core at a specified value of flux density. The ampere-turn requirements are derived as follows.

In terms of equivalent ampere turns, the applied field is given by

$$H_o = \frac{\mathrm{NI}}{2.02l}. \tag{13}$$

If, as is usually the case in the electromagnet, H is negligibly small compared with H_o, then one may combine (8) and (13) to obtain

$$\mathrm{NI} = 2.02lD_B(B - H). \tag{14}$$

If one multiplies each side of (14) by $(Al)^{\frac{1}{3}}$ and collects terms, one obtains

$$\mathrm{NI} = 2.02 \frac{l}{(Al)^{\frac{1}{3}}} D_B(B - H)(Al)^{\frac{1}{3}}. \tag{15}$$

But,

$$\frac{l}{(Al)^{\frac{1}{3}}} = (l/d)^{\frac{2}{3}}(\pi/4)^{-\frac{1}{3}}. \tag{16}$$

Combining (15), (16), and (6)

$$\mathrm{NI} = 1.9 \times 10^6(Al)^{-\frac{2}{3}} \frac{T_o}{H_a} \frac{1}{R_S} D_B(l/d)^{\frac{2}{3}}. \tag{17}$$

Since D_B and R_S are functions of aspect ratio (l/d) one may define an aspect ratio factor,

$$F_a = D_B(R_S)^{-1}(l/d)^{\frac{2}{3}} \tag{18}$$

and chart it as a function of (l/d) as in Fig. 6. Then combining (17) and (18), gives

$$\mathrm{NI} = 1.9 \times 10^6 \frac{F_a}{(Al)^{\frac{2}{3}}} \frac{T_o}{H_a}. \tag{19}$$

For any proposed geometry of an electromagnet with specified value of magnetic moment (T_o/H_a) the ampere-turn requirement may be determined from (19) and then translated into power and weight requirements by reference to Table II.

3.4 *The Semi-Permanent Magnet*

A permanent magnet material with low coercivity and high remanence, as exhibited by Remendur in Fig. 3, offers the inviting possi-

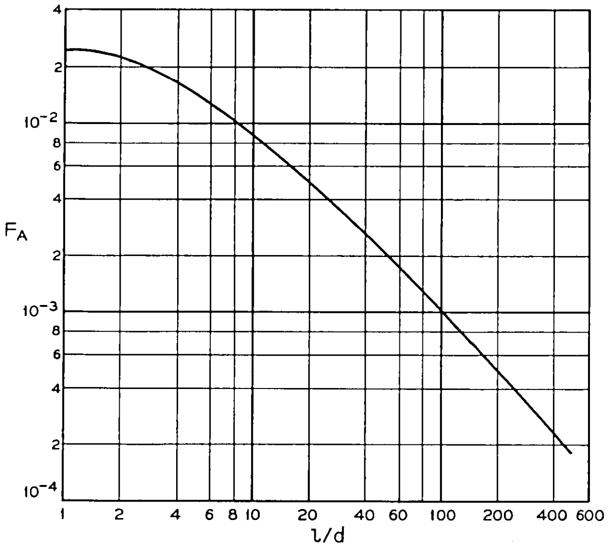


Fig. 6 — The aspect ratio function (F_a).

bility of easy magnetization and reversal of field by means of short pulses of current through a winding. Between pulses it acts as a permanent magnet, with polarity determined by the direction of the preceding pulse. Thus, it provides a switchable field with very low expenditure of power. It requires, however, for complete demagnetization, or "knock-down," much more sophisticated circuitry. Also, the certainty of complete demagnetization from an applied pulse or series of pulses, depends to a considerable extent upon the preceding history of the magnet. For this reason, it is not likely to replace readily the simple air-core coil or electromagnet unless the available power is so severely limited as to justify the added circuit development effort.

IV. INTERCOMPARISON—AIR-CORE COILS AND ELECTROMAGNETS

In the preceding sections we have developed design formulas and design graphs which enable us to estimate with fair quantitative accuracy the size and weight of various magnetic structures to satisfy torque requirements as specified. Intercomparison of the air-core coil and electromagnet offers an interesting illustration of the use of these techniques. We consider a typical example which assumes a spherical satellite of 45 inches effective diameter in which there is required an available magnetic moment of 0.2 lb-in per oersted which can be turned on or off at will. It is further assumed that an upper limit of nine pounds weight and twelve watts power consumption is to be imposed upon the magnetic circuitry.

First, we assume that an air-core coil is laid out around the equator of the satellite to enclose maximum area, and that this area is

$$A = \frac{\pi}{4}(45)^2 = 1590 \text{ in}^2.$$

Substituting this value of A in (2) gives the required ampere turns,

$$\text{NI} = \frac{1.667 \times 10^6}{1.590 \times 10^3} 0.2 = 210.$$

The average length per turn of winding is $45\pi = 141$ inches. Using the formula for aluminum from Table II we can show that the power × weight product is

$$\text{power} \times \text{weight} = 0.119 \times 10^{-6}(141)^2(210)^2 = 105.$$

If we use the total weight allowance of nine pounds for the winding then the required power is

$$\text{power} = \frac{105}{9} = 11.7 \text{ watts}.$$

This is within the permitted 12 watts, so we have shown that it is feasible to use an equatorial coil.

Turning now to the design of the electromagnet, one inserts $T_o/H_a = 0.2$ and $(B - H) = 10{,}000$ in (7) to show that the required volume of core material is

$$\text{vol} = \frac{0.866 \times 10^6}{10^4} 0.2 = 17.32 \text{ in}^3.$$

Assuming density of 0.26 lbs/in^3, the core will weigh 4.5 pounds.

The characteristics of the winding, however, are closely dependent upon the aspect ratio of the core. To illustrate this point we consider two shapes, one to be 10 inches long, the other to be 45 inches long to just fit in along the spin axis of the satellite. For the 10-inch core,

$$\frac{\pi}{4}(10)d^2 = 17.32; \qquad d = 1.485''; \qquad l/d = 6.73; \qquad F_a = 0.1.$$

Substituting these values in (19) gives the required ampere turns,

$$\text{NI} = \frac{1.9 \times 10^6}{(17.32)^{\frac{2}{3}}}(0.2)(0.1) = 5640.$$

If we assume the average diameter of the winding is 1.6 inches then the average length of turn, $P = 5.05$ inches. We insert these numbers into the power × weight formula for aluminum in Table II and obtain

$$\text{power} \times \text{weight} = 0.119 \times 10^{-6}(5.05)^2(5.64)^2 \times 10^6 = 97.$$

If we let the winding weigh 4.5 pounds to use up the residue of the weight allotment, then the power requirement is 21.5 watts. Since the maximum allowable power dissipation is 12 watts, it is evident that the 10-inch electromagnet, as described, cannot satisfy the requirements.

For the 45-inch core:

$$\frac{\pi}{4}(45)d^2 = 17.32; \qquad d = 0.7''; \qquad l/d = 64; \qquad F_a = 0.012$$

and substituting these numbers in (19) gives the required ampere turns,

$$\text{NI} = \frac{1.9 \times 10^6}{(17.32)^{\frac{2}{3}}}(0.2)(0.012) = 680.$$

If we assume that the average diameter of the winding is 0.85 inch, then the average length of turn, $P = 2.67$ inches, and

$$\text{power} \times \text{weight} = 0.119 \times 10^{-6}(2.67)^2(680)^2 = 0.39.$$

So we may use 0.5 pound of winding to bring the total weight only to five pounds, and the required power will be only 0.78 watt. This illustrates the advantage of the long slender electromagnet over the short one for purposes of producing torque.

It is evident that the specified conditions of the example can be satisfied by the equatorial coil or by the long slender electromagnet. On the basis of the calculated results one might well prefer the electromagnet, which satisfies the requirements with a substantial saving of power and weight. However, other factors must be considered. It is not likely to be convenient to mount the magnet full length along the spin axis because of interference with other equipment. In a core of this length, a very small amount of residual magnetization after removal of current will result in a considerable magnetic moment, rather than the desired zero magnetic moment which is characteristic of the de-energized air-core coil. The weight distribution of the electromagnet along the spin axis decreases the spin stability, while the weight distribution of the equatorial coil enhances the spin stability of the satellite. For these and other reasons the equatorial coil remains a favored method of attitude control in the communications satellite.

V. OPTIMUM DESIGN OF PERMANENT MAGNETS FOR FRICTION DAMPING

There have been various proposals to provide friction damping of roll or precession of the spin axis by mounting a small magnet within a hollow spherical enclosure attached to the satellite. The magnet would tend to maintain its alignment in the earth's field and to provide damping through friction contact with the interior of the sphere. For this application, if it exists, or for any similar application, one would wish to design for maximum normalized torque and minimum normalized period of oscillation in the field and within the confines of the sphere.

5.1 *Design for Maximum Moment within Limiting Spheres*

Referring to (6) and dividing through by D^3 where D is diameter of sphere in inches, and $D^3 = (d^2+l^2)^{3/2}$,

$$\frac{T_o}{H_a D^3} = 0.91 \times 10^{-6} l/d[1 + (l/d)^2]^{-\frac{3}{2}}[B - H]R_S \, . \tag{20}$$

The relation expressed by (20) is displayed in Fig. 7 for the same magnet materials for which the torque-weight relation was shown in Fig. 5.

5.2 *Design for Minimum Period of Oscillation within Circumscribed Sphere*

A magnet used to damp out roll or precession should have a natural period of oscillation in the earth's field much shorter than the period of the motion it is to damp out. This would suggest a magnet designed to have minimum period of oscillation with a spherical enclosure.

The moment of inertia of the cylindrical magnet around a diameter through its equator is given by

$$M_i = \frac{\pi}{4} \frac{\rho}{g^*} l d^2 \left[\frac{d^2}{16} + \frac{l^2}{12}\right] \tag{21}$$

and dividing through by D⁵,

$$\frac{M_i}{D^5} = 42.6 \times 10^{-6} \rho(l/d) \frac{[3 + 4(l/d)^2}{[1 + (l/d)^2]^{5/2}} . \tag{22}$$

* $g = 384$ in/sec/sec.

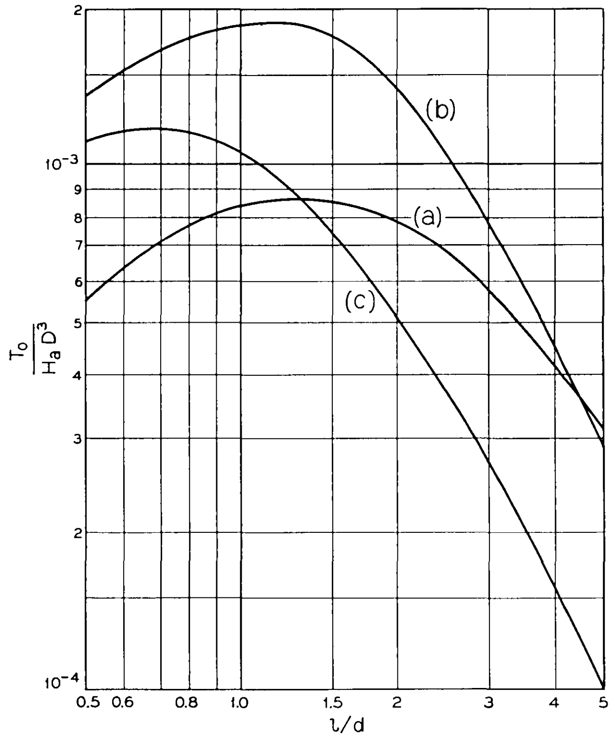


Fig. 7 — Normalized moment within limiting sphere; (a) ceramic, (b) Alnico 9, (c) Alnico 5.

Combining (20) and (22) and collecting terms, one obtains,

$$\frac{M_i}{T_o}\frac{H_a}{D^2} = \frac{46.8\rho[3 + 4(l/d)^2]}{R_S[B - H][1 + (l/d)^2]}. \tag{23}$$

The period of oscillation is given by

$$\tau = 2\pi\sqrt{\frac{M_i}{T_o}}\ \text{sec}. \tag{24}$$

Combining (23) and (24) yields,

$$\frac{\tau\sqrt{H_a}}{D} = 42.8\sqrt{\frac{\rho[3 + 4(l/d)^2]}{R_S[B - H][1 + (l/d)^2]}}. \tag{25}$$

This normalized period of oscillation is displayed graphically as a function of l/d in Fig. 8. In designing a magnet for friction damping

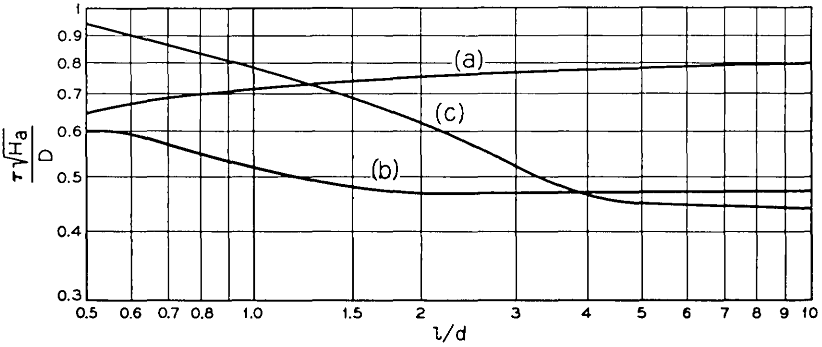


Fig. 8 — Period of oscillation within limiting sphere; (a) ceramic, (b) Alnico 9, (c) Alnico 5.

one would probably select the best compromise between maximum torque displayed in Fig. 7 and minimum period of oscillation as shown in Fig. 8. This would suggest the use of Alnico 9 and design for $l/d \cong 1.5$.

VI. CHARACTERISTICS OF SPHERICAL AND SPHEROIDAL MAGNETS

The spheroids are a family of solids the surfaces of which are generated by ellipses revolving around an axis. Revolution around a major axis generates a prolate spheroid for which $l/d > 1$. Revolution around a minor axis generates an oblate spheroid for which $l/d < 1$. Revolution of a circle around a diameter generates a sphere for which $l/d = 1$.

Values of load factor, U, for spheroids are plotted in Fig. 1. The volume of the spheroid is only two thirds that of a cylinder having the same l and d, so (11) becomes, for spheroids,

$$\frac{T_o}{W_m H_a} = \frac{1.733 \times 10^{-6}(B - H)R_S}{\rho}. \tag{26}$$

From solutions of (26) one may plot curves for normalized torque-weight ratio. In Fig. 9 we show a curve for spheroids of Alnico together with a curve for cylinders of Alnico borrowed from Fig. 5. While the spheroids show a somewhat better torque-weight ratio than the cylinders, it is doubtful whether the advantage is sufficient to offset the added cost of shaping and mounting.

The sphere might have unique advantages mounted in a spherical enclosure for friction damping. For the sphere of diameter D, one may rewrite (6)

$$\frac{T_o}{H_a} = 1.155 \times 10^{-6}(B - H)\frac{\pi}{4}\cdot D^3 R_S\ . \tag{27}$$

Dividing through by D^3 and inserting value of R_S, gives

$$\frac{T_o}{H_a D^3} = 0.81 \times 10^{-6}(B - H). \tag{28}$$

This is an expression for the total normalized torque that can be packed into a specified spherical enclosure. Solutions of (28) for various magnet materials are collected in Table III. The moment of inertia of the sphere is

$$M_i = 0.1MD^2 = 0.1\frac{\pi}{6}D^5\frac{\rho}{g}. \tag{29}$$

Combining (27) and (29)

$$\frac{M_i H_a}{T_o} = \frac{0.1(\pi/6)D^5\rho}{1.155 \times 10^{-6}(\pi/4)(B - H)D^3 g R_S} = \frac{150\rho D^2}{(B - H)R_S}. \tag{30}$$

Combining (24) and (30)

$$\frac{\tau\sqrt{H_a}}{D} = 24.5\pi\sqrt{\frac{\rho}{(B - H)R_S}}. \tag{31}$$

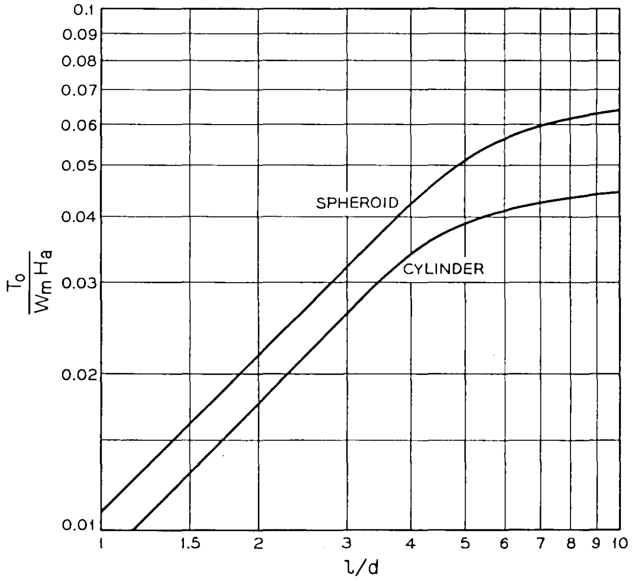


Fig. 9 — Comparison, spheroidal and cylindrical magnets of Alnico 5.

Table III—Magnetic Characteristics of Spheres

Material	ρ	$(B - H)$	$\dfrac{T_o}{W_m H_a}$	$\dfrac{T_o}{H_a D^3}$	$\dfrac{\tau\sqrt{H_a}}{D}$
Ceramic	0.15	3700	3.81×10^{-2}	3.0×10^{-3}	0.514
Alnico 9	0.26	4600	2.73×10^{-2}	3.74×10^{-3}	0.613
Alnico 5	0.26	1900	1.13×10^{-2}	1.54×10^{-3}	0.955

Equations (26), (28), and (31) define for the sphere the same normalized quantities which are plotted for the cylinder in Figs. 5, 7, and 8. Solutions of these equations for various magnet materials are listed in Table III. The combination of the table and the three figures gives all the information required to select the preferred material and geometry for a specified application and to arrive at a quantitative design of the magnet.

VII. SATELLITE MAGNETIC MEASUREMENTS

Satellites with spin stabilization introduce two magnetic measuring problems—measurement of "drag" and measurement of residual moment. The "drag" results from eddy currents induced in the rotating metal shell of the satellite by the earth's magnetic field. The energy dissipated in these eddy currents must be derived from the rotational energy of the satellite, and there results a decay of the spin rate. One wishes to evaluate the rate of this decay to forecast when the spin rate will fall below the minimum required for stability. The moment measurement is to detect any residual magnetic moment perpendicular to the axis of spin which will interact with the earth's magnetic field to induce precession of the spin axis. After an accurate measurement this moment is canceled out by mounting in the satellite a small permanent magnet of equal moment and opposite polarity. Both measurements—drag and moment—can be made conveniently with a specially designed coil array.

7.1 *The Telstar® Coil Array*

The drag test requires a reasonably uniform field over the volume of the satellite. A paper analysis reveals that this can be provided by an array of coils of reasonable size with a particular distribution of ampere turns. Two coils, each of radius r, and of N turns are spaced $\pm r/4$ from an assumed zero point on a common axis. Two other coils, each of radius r, and 7N/3 turns, are spaced $\pm r$ from the assumed zero along the common axis. The coils are connected in series to run at the same current so that the outer coils have effectively 7N/3 times as many ampere turns as the inner pair. The arrangement of coils and the resulting distribution of field along the axis are shown in Fig. 10.

It was established by measurements that the region of uniform field extended out radially from the axis to include a spherical volume, the radius of which is roughly two thirds the radius of the coils. Hence, an array of coils five feet in diameter easily provided uniform field over the volume of the satellite. (If a conventional Helmholtz array were used the coils would have to be about 10 feet in diameter to achieve reasonably uniform field over the same volume.) This array was mounted on a turntable and rotated around the satellite which was supported by a calibrated torque suspension. From the result of this drag measurement it was possible to calculate the rate of decay of satellite spin in the earth's magnetic field.

7.2 *Measurements of Magnetic Moment*

The magnetic moment perpendicular to the spin axis of the satellite was measured by rotating it within a coil array similar to the one used for drag tests except that the windings were connected to an integrating fluxmeter. One reasons intuitively that if a magnetic object is aligned parallel to the axis of the coils and rotated 180°, it will give a deflection of the integrating fluxmeter proportional to the moment, and that the proportionality constant will be unaffected by the position

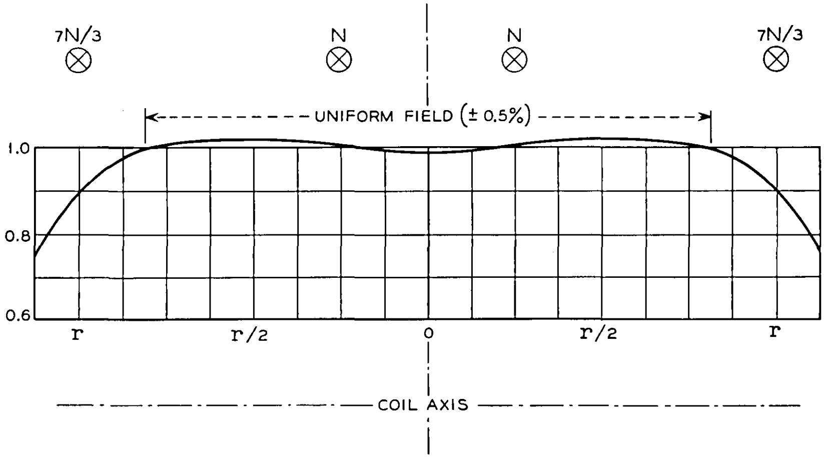


Fig. 10 — Field distribution along axis of *Telstar*® array.

of the magnet in the array as long as it is within the volume in which the array produces uniform field. This intuitive reasoning has been confirmed by various measurements. The proportionality constant for the array is established by calibration with a small air-coil, of known NIA for which the moment can be calculated from (2). A two-to-one scale down of the array has proved to be convenient for bench measurements of magnetic moment of small magnetic objects.

VIII. ACKNOWLEDGMENT

The author is indebted to Mr. L. Rongved for stimulating discussions of the dynamics of the orbiting satellite.

APPENDIX

Definition of Symbols

The following letter symbols have been adopted for use in this paper.

A	= section area of magnet or winding.
B	= flux density.
$(B - H)$	= intrinsic induction,
d	= diameter of magnet.
D	= diameter of enclosing sphere.
D_B	= demagnetizing factor.
F_a	= aspect ratio factor, as defined in text.
g	= gravity (384 in/sec/sec).
H	= field strength in magnet.
H_a	= ambient field, or field of interaction.
H_o	= applied magnetizing field.
l	= length of magnet.
M	= mass.
M_i	= moment of inertia.
M_m	= magnetic moment.
NI	= ampere turns.
R_S	= shortening ratio, from recession of poles.
T_o	= torque between magnet and perpendicular field.
τ	= period of mechanical oscillation.
U	= load factor, reciprocal of demagnetizing factor.

REFERENCES

1. Okoshi, Takanori, Demagnetizing Factors of Rods and Tubes Computed from Analog Measurements, J.A.P., *36,* August, 1965, pp. 2382–2387.
2. Bozorth, R. M., and Chapin, D. M., Demagnetizing Factors of Rods, J.A.P., *13,* May, 1942, pp. 320–326.

Nonlinear Optical Coefficients

By F. N. H. ROBINSON

(Manuscript received January 18, 1967)

We consider, from a number of different viewpoints, the tensor coefficients which describe second harmonic generation, optical rectification, and the Pockels or linear electro-optic effect in acentric crystals. Stationary perturbation theory is used to calculate the low-frequency limit of the intrinsic electronic nonlinearity neglecting all effects due to local fields or lattice polarization. Solid methane is used as an example and the result used to estimate the coefficient in hexamethylene tetramine. The calculated result is within a factor of 2 of the experimental figure. The method is susceptible to further refinement and, since it requires only a knowledge of ground state wave functions, and is essentially very simple, it appears to offer a useful approach to the calculation of the coefficients.

The classical anharmonic oscillator model is briefly covered and the model is related to a quantal treatment. We find that the anharmonic potential used in the model is directly related to the actual crystalline potential. It can also be related to the charge distribution in the electronic ground state.

Local field corrections and the effects of lattice polarization are presented. These alter the nonlinear properties in a simple and obvious way, but one which has been misunderstood in some of the literature.

Our results form a theoretical background to Miller's empirical rule relating the nonlinear coefficients to the linear susceptibilities. An extensive table of Miller-reduced tensor coefficients collated from the published literature is presented.

Finally, we draw together some of the threads of the previous sections. An appendix deals with the vexing question of definitions.

I. INTRODUCTION

Second harmonic generation, optical rectification, and the linear electro-optic effect are particular aspects of a process in which two fields, $E_j^\beta e^{i\beta t}$ and $E_k^\gamma e^{i\gamma t}$, generate a polarization

$$P_i^\alpha e^{i\alpha t} = d_{ijk}^{\alpha\beta\gamma} E_j^\beta e^{i\beta t} E_k^\gamma e^{i\gamma t}. \tag{1}$$

Our concern is with the tensor coefficients $d_{ijk}^{\alpha\beta\gamma}$ which (Nye[1]) necessarily vanish in centric (centrosymmetric) crystals and which, in acentric crystals, are subject to symmetry restrictions, (Kleinman[2]) which often leave only one or two independent components.

Experimentally, the values of the allowed components of d in different materials and at different frequencies range from about 2.10^{-10} esu (cm/stat volt) to about 6.10^{-5} esu. This range may be contrasted with the linear optical susceptibility χ which is between 0.1 and 0.3 for the vast majority of materials and only quite exceptionally exceeds unity. There is, however, a connection between the tensor d and χ which is expressed by an important empirical rule due to Miller.[3] If we write $d_{ijk}^{\alpha\beta\gamma}$ as

$$d_{ijk}^{\alpha\beta\gamma} = \chi_{ii}^{\alpha}\chi_{jj}^{\beta}\chi_{kk}^{\gamma}\Delta_{ijk} , \tag{2}$$

where χ_{ii}^{α} is the ii component of χ at a frequency α, and if we have chosen a principal axis system for χ, then the allowed components of Δ_{ijk} for all effects and all materials are similar in magnitude. We shall see in a later section that for very many materials in both the visible and 10 μ region of the spectrum (Patel[4]), Δ_{ijk} is near 3×10^{-6} esu. No materials with Δ above 20×10^{-6} esu have yet been found and very few are known to have a value below 0.2×10^{-6} esu. In the case of $NH_4H_2PO_4$ where the best measurements of s.h.g., optical rectification and the electro-optic effect are available (Francois,[5] Ward,[6] Carpenter[7]) the value of Δ_{123} from all three effects is 3×10^{-6} esu within the experimental error of 15 percent. The fact that s.h.g., a purely optical effect, leads to the same value of Δ as rectification and the electro-optic effect indicates quite clearly that the basic mechanism of the nonlinearities is common to all three effects and must therefore reside in the electronic motion of the system. In the next section, we shall concentrate on this aspect of the problem and neglect the effects of local fields and lattice polarization.

A number of authors (see Section IV for references) have given quantal treatments of the optical nonlinearities whose end result is an expression for the coefficients $d_{ijk}^{\alpha\beta\gamma}$ in terms of sums of rather inaccessible matrix elements. Useful as these expressions are, in establishing some of the general properties of the coefficients, they are not a practical step on the road to calculating the coefficients from other empirical quantities. At the other extreme, the classical anharmonic oscillator model has been used to demonstrate some of the qualitative features of nonlinear behavior (see Section III). This treatment, though simple and appealing, suffers from the defect that the relation between the parameters of the model and those of the real system is obscure. In Section IV, we shall remedy this defect and show that the two approaches are closely related.

First, however, we give an approximate method of calculating the low-frequency limit of the coefficients from stationary perturbation theory in a form in which it has been successfully applied to the linear properties of n electron atoms (see, e.g., Dalgarno[8]).

II. MAGNITUDE OF THE COEFFICIENTS

At low frequencies, i.e., well below any electronic resonance we can use stationary perturbation theory to calculate, to arbitrary order in the applied field $\underset{\sim}{E}$ the energy W of the ground state. The polarization is then given by

$$P_i = -\frac{\partial W}{\partial E_i}. \tag{3}$$

We shall assume that we are dealing with a crystal containing N identical atoms or molecules in unit volume whose individual ground state energies are w, so that $W = Nw$.

If H_o is the Hamiltonian of an unperturbed molecule, its Hamiltonian in the field $\underset{\sim}{E}$ is

$$H = H_o + h = H_o - e\underset{\sim}{E}\cdot\underset{\sim}{R}, \tag{4}$$

where

$$e\underset{\sim}{R} = e\sum_{m=1}^{n}\underset{\sim}{r}^m \tag{5}$$

is the dipole moment operator of the molecule, and the sum extends over all n valence electrons. We can neglect the core electrons because of their high binding energies. If we expand w in increasing order in $\underset{\sim}{E}$ as

$$w = w_0 + w_1 + w_2 + w_3\ , \quad \text{etc.}, \tag{6}$$

the term w_1 gives the permanent dipole moment of the molecule, w_2 gives the linear susceptibility and w_3 gives a polarization quadratic in $\underset{\sim}{E}$ which leads to the desired nonlinear coefficients. The electric field will perturb the state function ψ and we shall write the perturbed function as either

$$\psi = \psi_0 + \psi_1 + \psi_2 + \cdots \quad \text{or} \quad |\ \rangle = |0\rangle + |1\rangle + |2\rangle + \cdots. \tag{7}$$

Knowledge of ψ or $|\ \rangle$ to first order in $\underset{\sim}{E}$ is sufficient to determine w_1, w_2, and w_3 for

$$\begin{aligned} w_1 &= \langle 0|\, h\, |0\rangle \\ w_2 &= \langle 0|\, \hat{h}\, |1\rangle \\ w_3 &= \langle 1|\, \hat{h}\, |1\rangle. \end{aligned} \tag{8}$$

Moreover, the correct value of ψ_1 or $|\, 1\rangle$ is determined by the requirement that it minimize w. Thus, we can obtain $|\, 1\rangle$ by a variational procedure and the only element of choice left to us is that of the trial wave function.

Minimizing w is equivalent (see Dalgarno and Lewis[9]) to the simpler problem of minimizing

$$\langle 1|\, \hat{H}_o\, |1\rangle + 2\langle 0|\, \hat{h}\, |1\rangle, \tag{9}$$

where the notation $\hat{H}_o$ or $\hat{h}$ means $H_o - \langle 0 \mid H_o \mid 0\rangle$ or $h - \langle 0 \mid h \mid 0\rangle$.

As a trial function, we take

$$|1\rangle = \lambda \hat{h}\, |0\rangle \tag{10}$$

so that (9) becomes

$$\lambda^2 \langle 0|\, \hat{h}\hat{H}_o\hat{h}\, |0\rangle + 2\lambda\langle 0|\, \hat{h}^2\, |0\rangle. \tag{11}$$

The minimization with respect to λ gives

$$\lambda = -\frac{\langle 0|\, \hat{h}^2\, |0\rangle}{\langle 0|\, \hat{h}\hat{H}_o\hat{h}\, |0\rangle} \equiv -\frac{\langle 0|\, \hat{h}^2\, |0\rangle}{\langle 0|\, h\hat{H}h\, |0\rangle}. \tag{12}$$

The unperturbed Hamiltonian of the system is of the form

$$H_o = -\frac{\hbar^2}{2m}\sum_{m=1}^{n} \nabla_m^2 + V_o \tag{13}$$

and so, in the denominator of (12),

$$\hat{H}_o h = h\hat{H}_o + \frac{\hbar^2}{m} e\underline{E}\cdot \sum_m \nabla_m \quad . \tag{14}$$

Thus,

$$\langle 0|\, h\hat{H}_o h\, |0\rangle = -\frac{e^2\hbar^2}{m}\int \psi_o \sum_{m'} \underline{E}\cdot\underline{r}_{m'}\underline{E}\cdot \sum_m \nabla_m\psi_o\, d\tau, \tag{15}$$

where $d\tau$ is an element of configuration space and we have used $\hat{H} \mid 0\rangle = 0$. Equation (15) can be written as

$$\langle 0|\, h\hat{H}_o h\, |0\rangle = -\frac{e^2\hbar^2}{2m}\underline{E}\cdot\int \sum_m \nabla_m\psi_o^2 \sum_{m'} \underline{E}\cdot\underline{r}_{m'}\, d\tau, \tag{16}$$

and integrated by parts, to give

$$\langle 0|\, h\hat{H}_o h \,|0\rangle = +\frac{e^2\hbar^2}{2m}\,\underline{E}\cdot\underline{E}\,\sum_m \int \psi_o^2\, d\tau, \tag{17}$$

where the discarded first integration part vanishes at the limits, if these are infinite, or if they are the boundary of a cell in periodic lattice, provided only that $\underline{E}$ does not vary appreciably within a cell (dipole approximation).

If we are dealing with isolated atoms $\int \psi_o^2\, d\tau = 1$ and we have

$$\langle 0|\, h\hat{H}_o h \,|0\rangle = +n\,\frac{e^2\hbar^2}{2m}\,\underline{E}\cdot\underline{E}, \tag{18}$$

a somewhat unfamiliar form of the sum rule. If, on the other hand, we are dealing with overlapping molecules in a periodic lattice, the variational problem is to minimize the contribution to w from a single cell of the lattice. Thus, in (11) and all succeeding equations, the integrals implied by the expectation values are to be taken only over the interior of a cell. This will also apply to all integrals involved in evaluating $w_2 = \langle 1 \mid \hat{h} \mid 0\rangle$ and $w_3 = \langle 1 \mid \hat{h} \mid 1\rangle$. In this case (18) remains unchanged. This can be shown to be a general consequence of time reversal invariance and the commutation rule

$$(p,\, q) = i\hbar. \tag{19}$$

We now have

$$|1\rangle = \lambda\hat{h}\,|0\rangle = +\frac{2m\langle 0|\,\hat{h}^2\,|0\rangle}{ne^2\hbar^2E^2}\,\hat{h}\,|0\rangle \tag{20}$$

or

$$|1\rangle = +\frac{2me\langle 0|\,(\underline{E}\cdot\hat{\underline{R}})^2\,|0\rangle}{n\hbar^2E^2}\,\underline{E}\cdot\hat{\underline{R}}\,|0\rangle. \tag{21}$$

From this we obtain the second-order energy

$$w_2 = -\frac{2}{na_o}\,\frac{\langle(E\cdot\hat{R})^2\rangle^2}{E^2}, \tag{22}$$

where $a_o = \hbar^2/me^2 = 0.53$ Å. If we let $E = E_x$, 0, 0, and $\underline{R} = X, Y, Z$ this gives

$$w_2 = -\frac{2}{na_o}\,\langle\hat{X}^2\rangle^2 E_x^2 \tag{23}$$

and the atomic polarizability is

$$\alpha = \frac{4}{na_o} \langle \hat{X}^2 \rangle^2. \tag{24}$$

For the H atom, this gives $\alpha = 4a_0^3$ instead of the correct value $4.5a_0^3$, while for the helium atom, taking an effective nuclear charge $Z = 27/16$ gives 1.8×10^{-25} ccs. The experimental value is 2.1×10^{-25} ccs. In general, (24) is a lower limit to α, if we evaluate $\langle \hat{X}^2 \rangle$ correctly as the expectation value of the mean square moment of all the electrons. If the electrons are uncorrelated

$$\langle \hat{X}^2 \rangle = n\langle \hat{x}^2 \rangle, \tag{25}$$

where $\langle \hat{x}^2 \rangle$ refers to one electron. We used this procedure in helium since the two electrons are in orthogonal spin states and are automatically uncorrelated. In more complicated atoms correlation exists and almost always results in

$$\langle \hat{X}^2 \rangle < n\langle \hat{x}^2 \rangle \tag{26}$$

since electrons repel each other. Thus, while (24) is a lower limit we cannot say anything about the sign of the error in

$$\alpha = \frac{4n}{a_o} \langle \hat{x}^2 \rangle^2. \tag{27}$$

We note, in passing, that, in a solid with overlapping molecules, α the polarizability is large. This leads to an element of instability in the situation for as α increases the screening of the coulomb potential becomes more effective and the electrons less localized leading to a further increase in α and eventually metallic behavior. For this reason, most materials, which are not regular insulators, are metals. Those rare materials which have values of $N\alpha$ appreciably greater than 0.3 ($n > 2.2$) owe their existence to a rather delicate balance of forces.

The third-order energy is

$$w_3 = \langle 1 | \hat{h} | 1 \rangle = -\left(\frac{2m}{n\hbar^2 E^2}\right)^2 \langle (E \cdot \hat{R})^2 \rangle^2 e^3 \langle (E \cdot \hat{R})^3 \ . \tag{28}$$

In most cases α is very nearly isotropic and we have

$$\tfrac{1}{2}\alpha E^2 = -w_2 = \left(\frac{2me^2}{n\hbar^2 E^2}\right) \langle (E \cdot \hat{R})^2 \rangle^2 \tag{29}$$

so that

$$w_3 = -\frac{\alpha}{na_o e} \langle (E \cdot \hat{R})^3 \rangle. \tag{30}$$

With N molecules in unit volume this gives a nonlinear coefficient

$$d_{ijk} = \frac{3N\alpha}{a_o e}\frac{T_{ijk}}{n} = \frac{3\chi}{a_o e}\frac{T_{ijk}}{n}, \tag{31}$$

where

$$T_{ijk} = \langle \hat{R}_i \hat{R}_j \hat{R}_k \rangle = \langle R_i R_j R_k \rangle - \langle R_i \rangle \langle R_j R_k \rangle - \langle R_j \rangle \langle R_k R_i \rangle - \langle R_k \rangle \langle R_i R_j \rangle + 2\langle R_i \rangle \langle R_j \rangle \langle R_k \rangle. \tag{32}$$

Equation (31) is the central result of this section. It expresses d_{ijk} in terms of the linear (corrected for local fields) susceptibility χ and a cubic moment (third-order semi-invariant) of the electronic distribution in the ground state.

If we neglect overlap and, for simplicity, also assume that the electrons are uncorrelated so that $T_{ijk} = nt_{ijk}$ where t_{ijk} refers to a single electron we have

$$d_{ijk} = \frac{3\chi}{a_o e} t_{ijk} \tag{33}$$

and T_{ijk} is now apart from numerical factors the octupole moment of the charge distribution. If the electron density in the molecule is $\rho(\underset{\sim}{r})$

$$T_{ijk} = \iiint \hat{r}_i \hat{r}_j \hat{r}_k \rho(\underset{\sim}{r})\, d^3 r. \tag{34}$$

If we account for local fields through a Lorentz correction the correct value of χ to insert in (33) is obtained from

$$\frac{n^2 - 1}{n^2 + 2} = \frac{4\pi}{3}\chi \tag{35}$$

and the observed value of d_{ijk} (see Section V) is

$$d_{ijk}^{\text{obs}} = \left(\frac{n^2 + 2}{3}\right)^3 d_{ijk}\,. \tag{36}$$

At first sight (33) seems to imply that d is proportional to χ in conflict with Miller's rule. However, χ depends on $N/n\langle \hat{R}^2 \rangle^2 \approx nN\langle \hat{r}^2 \rangle^2$ and N is inversely proportional to $\langle r^2 \rangle^{\frac{3}{2}}$ so that $\chi \approx nr$, while $d \approx r^4$. Thus, d is in fact more nearly proportional to χ^4 than χ.

We now consider as an example, the tetrahedral molecule methane CH_4, which crystallizes in the tetrahedral space group $F\bar{4}3m$ with a lattice parameter ≈6 Å and a molar volume of 32 ccs. If we take Cartesian axes along the sides of the cubic cell, the bonds point in the 111, and tetrahedrally related, $1\bar{1}\bar{1}$, $\bar{1}1\bar{1}$, $\bar{1}\bar{1}1$, directions. From symmetry, there is only one independent component of d_{ijk}, in which all the subscripts are unequal.

The shortest c-c distance is 4.2 Å and from the size of the free molecule we conclude that overlap is unimportant.

Turner, Saturno, Hauk and Parr[10] have used one center wave functions to calculate the electronic density ρ in the molecule. From this we can obtain t_{123} using (34).

The result is

$$t_{ijk} = 0.5 \times 10^{-24} \text{ cm}^3 \tag{37}$$

and this is not very sensitive to the limits of integration. The experimental molar susceptibility of CH_4 is 1.6 ccs and so $\chi = 0.05$. Thus, if we neglect correlations between the eight valence electrons we have from (33)

$$d_{123} = 3 \times 10^{-9} \text{ esu.} \tag{38}$$

In a similar way, neglecting correlations, we can calculate the molar susceptibility from (27). Turner *et al*'s charge density leads to

$$\langle x^2 \rangle = 3.3 \times 10^{-17} \text{ cm}^2$$

and so, with eight valence electrons, we obtain $\alpha = 6.5 \times 10^{-24}$ and $\chi =$ a molar susceptibility of 3.9 ccs, rather over twice the experimental value.

This is a clear indication that the electrons are correlated. However, the correlation enters twice in χ but only once in d (since we have expressed d in terms of the experimental χ). Thus, d problaby lies between 2×10^{-9} and 3×10^{-9} esu.

To see whether 3×10^{-9} esu is a reasonable value for d_{123} we compute the Miller reduced tensor $d/\chi^3 = \Delta_{123} = 24 \times 10^{-6}$ esu.

This is quite exceptionally high. Most materials have allowed components of Δ near 3×10^{-6} esu and only one coefficient in $LiNbO_3$ (9×10^{-6}) and Δ_{123} in hexamethylene tetramine (15×10^{-6}) approach this value.

However, we believe it is in fact not far wrong. In most materials geometric factors conspire to reduce d by various factors of cos θ and the atomic groups are in the first instance less aspherical than CH_4. In CH_4 the effects of every electron are directly additive.

Hexamethylene tetramine (HMT), the other exception to Miller's rule, is, like methane, a tetrahedral molecule $N_4(CH_2)_6$ in a tetrahedral $I\bar{4}3m$ crystal. The 4 nitrogen atoms form the 111 $1\bar{1}\bar{1}$, $\bar{1}1\bar{1}$, $\bar{1}\bar{1}1$, corners of a regular tetrahedron and the CH_2 groups occupy the edges but the N–C–N bonds are bent outwards in such a way that all the angles are very nearly tetrahedral. The carbon atoms occupy the six sites 2, 0, 0 etc. (see Kitaigorodskii[11]).

The refractivity of the molecule as a whole can be very satisfactorily accounted for by a system of additive bond refractions. (See LeFevre[12] for a review of bond refractions.) The three basic units are 12 C–H bonds pointing in the $\bar{1}$, $\bar{1}$, $\bar{1}$, and related directions, 4 nonbonding orbitals on the nitrogen atoms pointing in the 111 and related directions, and 12 N–C bonds in the 111 directions.

Since the refractivities are additive, these units appear to act independently in determining the molar refractivity. Le Fevre (*loc. cit.*) gives values of $R = (4\pi/3)L\alpha$ (where L is Avagadro's number) of 2.8 ccs for each unbonded nitrogen pair, 1.65 ccs for each C–H bond and 0.62 ccs for each N–C bond. Thus, the N–C bonds make a rather small contribution to χ, and probably even less to d since they have an approximate inversion centre at the centre of the bond (C and N are similar atoms as compared with C and H). We therefore neglect them.

The 12 CH bonds in the $\overline{111}$ direction are roughly equivalent to 3 methane molecules in the molecular volume 105 ccs, and further the electrons will be less correlated than in methane. Thus, their contribution to d_{123} is

$$d_{123}^{\mathrm{Me}} = -\tfrac{3}{105}\cdot\tfrac{32}{1}\cdot 3 \times 10^{-9} = -2.7 \times 10^{-9} \text{ esu.} \tag{39}$$

To calculate the effect of the nonbonding nitrogen electrons we assume that they occupy SP^3 hybrid orbitals directed along 111 etc. with Slater radial wave functions Ar exp $(-2.5r/2a_o)$. It is then straightforward to show that for *one* electron

$$t_{xyz} = -0.055 \times 10^{-24} \text{ cm}^3. \tag{40}$$

The contribution of the 4 nitrogen atoms to χ is

$$\chi^N = \frac{3}{4\pi}\cdot\frac{4 \times 2.8}{105} = 0.0255$$

and so

$$d_{123}^N = -1.7 \times 10^{-9} \text{ esu.} \tag{41}$$

Thus, the total value of d_{123} is -4.5×10^{-9} esu. This could be slightly increased by the effects of atomic overlap, and possibly by contributions from the N–C bonds. It could be either increased or decreased by electron correlations on individual CH_2 groups. The experimental values for the electro-optic effect Heilmeyer[13] and second harmonic generation, Heilmeyer, Ockman, Braunstein, and Kramer,[14] when corrected for local field effects using a Lorentz factor, both give $d = \pm 8.2 \times 10^{-9}$ esu. Thus, our calculation is within a factor 2 of the observed value.

This method is therefore capable, in simple cases, of predicting the magnitude of d rather successfully. Moreover, the experimental value of d for HMT suggests that we were correct in assuming that CH_4 will have an anomalously large reduced tensor Δ_{123} .

The fact that the division of a complex molecule such as HMT into simple components leads to a reasonable value for d leads us to hope that a similar procedure will be possible in other cases. It might then be possible to assign empirical values of d to basic components such as the C–H bond or the N: nonbonding pair, and to combine these additively (with a proper attention to geometry) to predict the values of d for even more complex molecules. This would not be surprising since a similar procedure (see LeFevre *loc. cit.*) works very satisfactorily for the linear susceptibilities.

It is then obvious that large nonlinear effects will only result, if the molecule contains polarizable groups disposed in an arrangement which results in a ground state of, far from inversion, symmetry. The large value of Δ in HMT results from the fortunate coincidence that the most polarizable components are themselves strongly asymmetric and so oriented that their effects are additive. The much smaller values of Δ commonly observed can then be explained as partly due to no group in the crystal being quite so asymmetric as N: or CH_2 in HMT and partly due to unfavorable geometric relations between the groups. For example, if our approach is correct we should expect the analogous compound adamantane $(CH)_4(CH_2)_6$ in which the nitrogens are replaced by CH groups with the CH bond along 111, etc. to have a d_{123} appropriate to 2 $(= 3 - 1)$ CH_4 molecules in 105 ccs, i.e., $d_{123} \approx 2 \times 10^{-9}$ esu or about half the value for HMT.

Exceptionally small values of Δ will occur in materials where most of the molecule possesses local inversion symmetry, so that only a fraction of the molecule contributes to d, while the whole molecule contributes to χ. We shall consider an example of this in a later section.

Overlap between adjacent molecules is neccessarily bound to lend further uncertainty to the calculation in materials with a pronounced band structure, but it seems possible that rough approximations should be obtainable from, for example, the relation between bandgap and the corresponding separation in the isolated atoms. In fact, since what we actually require is T_{ijk}/n, which, if the electrons are uncorrelated, is simply

$$\frac{T_{ijk}}{n} = \frac{\int^{\mathrm{Cell}} \hat{r}_i \hat{r}_j \hat{r}_k \rho(\underline{r})\, d^3r}{\int^{\mathrm{Cell}} \rho(\underline{r})\, d^3r} \tag{42}$$

we may expect that this factor will, to some extent, be self-cancelling.

Finally, we may remark that very much better approximations to d_{ijk} can obviously be made if we know the ground state wave function explicitly and also use more sophisticated trial wave functions in the variational calculation. It is at first sight surprising that a knowledge of the ground state wave function alone is sufficient to determine χ and d which, in the more usual treatments involve the properties of excited states. However, we should remember that a knowledge of the exact ground state wave function is, except in pathological circumstances, sufficient to determine the unperturbed Hamiltonian; thus, the whole spectrum of states.

III. THE CLASSICAL ANHARMONIC OSCILLATOR MODEL

Although the considerations of the preceding section are sufficient to determine the magnitude of d at low frequencies, they offer little guide to the variation of d with frequency and, if recast in terms of time dependent perturbation theory they tend to lose their attractive simplicity. In the next section we shall show that a more familiar form of time dependent theory leads to results which can be represented in terms of a classical anharmonic oscillator model. Here, we discuss the properties of the model itself.

We assume that unit volume of the material contains N optical electrons which move in a potential

$$V = \tfrac{1}{2} m \Omega_i^2 x_i^2 + V_{ijk} x_i x_j x_k , \tag{43}$$

where a sum over repeated subscripts is implied. The potential V_{ijk} obviously satisfies $V_{ijk} = V_{ikj}$, etc.

In a field $E_i^\beta e^{i\beta t}$ the equation of motion is

$$\ddot{x}_i + \Omega_i^2 x_i + 3\,\frac{V_{ijk}}{m}\, x_j x_k = \frac{e}{m} E_i^\beta e^{i\beta t} \tag{44}$$

and the linear response obtained by neglecting V_{ijk} is

$$x_i^{(1)} = \frac{\dfrac{e}{m} E_i^\beta e^{i\beta t}}{\Omega_i^2 - \beta^2}. \tag{45}$$

There will be a similar response to a field $E_i^\gamma e^{i\gamma t}$ and, if we introduce these responses back into the nonlinear term in (44) we obtain a response at the sum frequency $\alpha = \beta + \gamma$ given by

$$x_i^{(2)} = -\frac{3V_{ijk}e^2}{m^3}\frac{1}{\Omega_i^2 - \alpha^2}\frac{1}{\Omega_j^2 - \beta^2}\frac{1}{\Omega_k^2 - \gamma^2}(E_j^\beta E_k^\gamma + E_k^\gamma E_j^\beta)e^{i\alpha t}. \quad (46)$$

The resulting polarization is $Nex_i^{(2)}$ and so the nonlinear coefficient is

$$d_{ijk}^{\alpha\beta\gamma} = -\frac{3V_{ijk}Ne^3}{m^3}\frac{1}{\Omega_i^2 - \alpha^2}\frac{1}{\Omega_j^2 - \beta^2}\frac{1}{\Omega_k^2 - \gamma^2}. \quad (47)$$

Thus, the symmetry of d_{ijk} mimics that of V_{ijk} if we neglect the resonance denominators.

The linear susceptibility obtained from (45) is the familiar expression

$$\chi_{ii}^\omega = \frac{Ne^2}{m}\frac{1}{\Omega_i^2 - \omega^2} \quad (48)$$

and so if we express $d_{ijk}^{\alpha\beta\gamma}$ as

$$d_{ijk}^{\alpha\beta\gamma} = \chi_{ii}^\alpha \chi_{jj}^\beta \chi_{kk}^\gamma \Delta_{ijk} \quad (49)$$

the reduced Miller tensor is

$$\Delta_{ijk} = -\frac{3V_{ijk}}{N^2e^3}, \quad (50)$$

which is frequency independent and has the same symmetry as V_{ijk}.

If we assume that V_{ijk} is electrostatic in origin its order of magnitude will be e^2/d^4 where d is an atomic spacing and we shall also have $Nd^3 \approx 1$. Thus,

$$| \Delta_{ijk} | \approx 3\frac{d^2}{e}. \quad (51)$$

With d equal to 2 Å this is 2.5×10^{-6} esu, about the mean value of Δ for most materials. In a later section, we shall give another estimate of Δ.

The potential $V_{ijk}x_ix_jx_k$ distorts the shape of the ground state of the harmonic oscillator and as a result the system acquires a cubic moment t_{ijk} [defined in (32)] which we now calculate.

Let $| 0\rangle$ represent the unperturbed ground state wave function in the absence of the anharmonic term and $| p\rangle$ be an excited state, then the perturbed wave function is

$$| \rangle = |0\rangle - \sum_p{}' \frac{\langle p|\ V\ |0\rangle}{\hbar\omega_p}|p\rangle. \quad (52)$$

The expectation values of even operators such as $\langle x_i^2 \rangle$, $\langle x_i x_j \rangle$ are unaltered by V, while the expectation value of an odd operator such as x_i or $x_i x_j x_k$ is given by

$$\langle x_i x_j x_k \rangle = -2 \sum_p{}' \frac{\langle 0 |\ x_i x_j x_k\ | p \rangle \langle p |\ V\ | 0 \rangle}{\hbar \omega_p}. \tag{53}$$

It will suffice if we calculate t_{ijk} with $i \neq j \neq k$. Since $\langle x_i x_j \rangle = 0$ if $i \neq j$ we only require $\langle x_i x_j x_k \rangle$ and contributions to this come only from the $6 = 3!$ terms in V with $i \neq j \neq k$. The only state which contributes to the sum is $| p \rangle = | 1, 1, 1 \rangle$ with an energy $\hbar(\Omega_1 + \Omega_2 + \Omega_3)$. The matrix element is

$$\langle 0 |\ x_1 x_2 x_3\ | 111 \rangle = \left(\frac{\hbar}{2m} \right)^{\frac{3}{2}} \left(\frac{1}{\Omega_1 \Omega_2 \Omega_3} \right)^{\frac{1}{2}}$$

and so

$$t_{123} = \langle x_1 x_2 x_3 \rangle = -12 \left(\frac{\hbar}{2m} \right)^3 \frac{V_{123}}{\hbar \Omega_1 \Omega_2 \Omega_3 (\Omega_1 + \Omega_2 + \Omega_3)}.$$

It is straightforward to show that a similar result

$$t_{ijk} = -\frac{3}{2} \left(\frac{\hbar}{m} \right)^3 \frac{V_{ijk}}{\hbar \Omega_i \Omega_j \Omega_k (\Omega_i + \Omega_j + \Omega_k)} \tag{54}$$

holds for all the components of t_{ijk}.

If we substitute this relation in (47) and take the limit as $\alpha\beta\gamma \to 0$ we obtain

$$d_{ijk} = 2 \frac{3\chi}{a_o e} t_{ijk}. \tag{55}$$

This is twice the value obtained in (33) because there we treated t_{ijk} as a fixed property of the ground state which was then perturbed by E; whereas here we have considered an even ground state perturbed by E and a fixed potential.

Thus, if the real system has a cubic moment t_{ijk} in the ground state, the equivalent anharmonic oscillator model requires an anharmonic potential

$$V'_{ijk} = -\frac{m^3 \Omega^4}{\hbar^2} t_{ijk} \tag{56}$$

and this will result in a cubic moment $t'_{ijk} = \frac{1}{2} t_{ijk}$ in the oscillator ground state.

In the real crystal t_{ijk} may be an accessible quantity. It obviously is in molecular crystals of strongly covalent compounds such as CH_4. But, in ionic crystals it may be more sensible to consider the ions as spheres perturbed by a crystal potential V_{ijk}^{c}. In a later section we shall see that there is a simple relation between the model potential and V_{ijk}^{c}.

The classical anharmonic oscillator model has previously been used by Bloembergen,[15] Garrett and Robinson[16] and Kurtz[17] to give a qualitative account of nonlinear phenomena. The latter authors also discuss in some detail its relation to Miller's rule.

Obviously, the model is the nonlinear analogue of the classical harmonic oscillator model used with such success for the last 60 years in the discussion of linear behavior such as dispersion, and, just as the harmonic model is directly related to the results of a quantum mechanical treatment, we may expect the anharmonic oscillator to have a similar basis. In the next section we explore this relation.

IV. TIME DEPENDENT QUANTAL TREATMENT

A number of authors Bloembergen,[15] Armstrong, Bloembergen, Ducuing and Pershan,[18] Butcher and McLean,[19] Kelley,[20] Cheng and Miller[21] and Ward[22] have given rigorous quantal treatments of optical nonlinearities in solids. We select an expression due to Armstrong, et al (*loc. cit.*) which expresses the nonlinear coefficients in terms of the energies $\hbar\omega_p$ of excited states and the matrix elements $\langle 0 \mid x_i \mid p \rangle$, $\langle p \mid x_i \mid q \rangle$, etc. of the dipole operator between states. The ground state is $\langle 0 \mid$.

This expression is valid, either for an assembly of N isolated atoms in unit volume or, in the dipole approximation, for a real solid where the wave functions overlap. In the latter case, the solid must be divided into cells of the periodic lattice, and N is then the density of cells, while the matrix elements are to be evaluated only over the interior of a cell. The periodicity of the lattice ensures that contributions from parts of the wave function outside a cell cancel in the crystal as a whole.

To avoid a plethora of subscripts we let each of x, y, and z serve to represent any one of the components and we can then write the expression for d as

$$d_{xyz}^{\alpha\beta\gamma} = \frac{Ne^3}{\hbar^2} \sum_p \sum_q \left\{ x_{op}y_{pq}z_{qo} \frac{\omega_p\omega_q + \alpha\gamma}{(\omega_p^2 - \alpha^2)(\omega_q^2 - \gamma^2)} \right.$$

$$\left. + y_{op}z_{pq}x_{qo} \frac{\omega_p\omega_q + \beta\alpha}{(\omega_p^2 - \beta^2)(\omega_q^2 - \alpha^2)} + z_{op}x_{pq}y_{qo} \frac{\omega_p\omega_q - \gamma\beta}{(\omega_p^2 - \gamma^2)(\omega_q^2 - \beta^2)} \right\}. \quad (57)$$

This expression vanishes if the states $|\, 0\rangle$, $|\, p\rangle$, etc. have a definite parity, its value therefore depends on the existence of matrix elements whose presence is contingent on the absence of inversion symmetry. For this reason, it is almost impossible to make an informed guess about its magnitude or behavior.

An analogous expression for the linear susceptibility is

$$\chi_{xy}^{\omega} = \frac{2Ne^2}{\hbar} \sum_p \frac{\omega_p x_{op} y_{po}}{\omega_p^2 - \omega^2} \tag{58}$$

and in both expressions an operator x is to be understood as the total operator for the contents of a cell, i.e., the sum of the individual electron operators. Of course, we can neglect the core (nonvalence) electrons on the grounds that they are too tightly bound to contribute to the optical properties.

A familiar approximation to χ is obtained if we note that in (58) the variation of the summand with $|\, p\rangle$ is almost exclusively due to the matrix elements. These not only obey selection rules, but also decrease rapidly in magnitude as the state $|\, p\rangle$ increases in energy, and therefore overlaps the ground state less and less. For example, in the H atom with a $1S$ ground state the matrix element x_{op} vanishes unless p is one of the states $2P$, $3P$, etc. Moreover, as we go from the $2P$ state to the $8P$ state $x_{op}x_{po}$ decreases by over a hundredfold. At the same time, ω_p changes by less than 30 percent. Thus, except near a resonance, we can treat ω_p as a constant Ω, somewhere near the first allowed transition and write (3.2) as

$$X_{xy}^{\omega} = \frac{2Ne^2\Omega}{\hbar(\Omega^2 - \omega^2)} \sum_p{}' x_{op} y_{po} \tag{59}$$

where the primed sum excludes $p = 0$. Now

$$\begin{aligned}\sum{}' x_{op} y_{po} \equiv \sum x_{op} y_{po} - x_{oo} y_{oo} &\equiv (xy)_{oo} - x_{oo} y_{oo} \\ &\equiv \langle (x - \langle x\rangle)(y - \langle y\rangle)\rangle\end{aligned} \tag{60}$$

where a $\langle\ \rangle$ denotes a ground state expectation value. Thus,

$$\chi_{xy}^{\omega} = \frac{2Ne^2\Omega}{\hbar(\Omega^2 - \omega^2)} \langle (x - \langle x\rangle)(y - \langle y\rangle)\rangle. \tag{61}$$

We shall not pursue the further manipulations of (61) using the sum rule which lead back to (27) but we remark that in many cases a form such as (61) for χ, involving a single Sellmeier or classical oscillator term, gives an excellent account of optical dispersion, and that when applied to the hydrogen atom, with $\hbar\Omega$ set equal to $\frac{3}{8}e^2/a_o$ the $1S - 2P$ energy, it leads to a value of χ at low frequencies

$$\chi = \tfrac{16}{3}a_o^3$$

which exceeds the correct value 4.5 a_0^3 by 32/27 or 18 percent.

Before we can adopt a similar procedure with the nonlinear coefficient we must first satisfy ourselves that there is no essential difference between a sum with three matrix elements and one with two. In χ all matrix elements terminate on $|\,0\rangle$ but in d it is quite possible in a term such as $x_{op}y_{pq}z_{qo}$, with $p \approx q$ corresponding to highly excited states, of great spatial extent, that the term y_{pq} may be large enough to compensate for the smallness of $x_{op}z_{qo}$. If this were the case it would be possible for the exact value of the sum to depend critically on cancellations between large terms involving highly excited states. The removal of the frequencies ω_p, etc. as a single average would then have disastrous results on the sum.

We will advance three arguments why this is unlikely. Consider first an even higher-order calculation, that of the fourth-order Stark shift of the ground state of atomic hydrogen due to a field F. In atomic units this is given by an exact calculation (Dalgarno[23]) as

$$W^{(4)} = -\tfrac{3555}{64}F^4 \approx -56F^4. \tag{62}$$

We can also express $W^{(4)}$ (Dalgarno, *loc. cit.*) as

$$W^{(4)} = -\sum_p{}' \sum_q{}' \sum_r{}' \frac{x_{op}\hat{x}_{pq}\hat{x}_{qr}x_{ro}}{\omega_p\omega_q\omega_r} + \sum_p{}' \frac{x_{op}x_{po}}{\omega_p} \sum{}' \frac{x_{op}x_{po}}{\omega_p^2}.$$

Our procedure treats $\omega_p\omega_q$ and ω_r as a single constant Ω and leads to

$$W^{(4)} = -\frac{F^4}{\Omega^3}\{\langle(x - \langle x\rangle)^4\rangle - 2\langle(x - \langle x\rangle)^2\rangle^2\}. \tag{63}$$

For the H atom $\langle x\rangle = 0$, $\langle x^2\rangle = 1$ a.u. and $\langle x^4\rangle = 9/2$ a.u. so that, if we set $\Omega = 3/8$ a.u., the $1S - 2P$ energy difference

$$W^{(4)} = -\tfrac{1280}{27}F^4 \approx -48F^4. \tag{64}$$

This is close to the correct result (36), but despite the fact that we have taken the lowest possible value of Ω it is too small. This is a clear indication that some cancellation of higher terms, which we have aggravated by our cavalier treatment of ω_p, etc., is occurring. This is not surprising for, if in the triple sum we consider the lowest possible sequence of levels $1S$ $2P$ $2S$ $2P$ $1S$ for which $\omega_p = \omega_q = \omega_r = \Omega$ the product of the matrix elements is 5 a.u. while for the sequence $1S$ $8P$ $8S$ $8P$ $1S$ the product is 10 a.u. made up of a contribution of 0.0033 from the two $1S$ $8P$ elements and 3000 from the $8S$ $8P$ elements.

However, this is not quite so serious as it appears, for in a real solid no matrix element can exceed the linear dimensions of a cell say 5 a.u. so that the product in the low transition would remain at 5 a.u. while the product for the upper transition would be reduced to 0.08.

Our final argument is empirical. If cancellations between large terms are critically important, the relevant feature of our procedure is the change in the ratio ω_p/ω_q it causes for highly excited neighboring states. In hydrogen the ratio of the $1S - 8P$ energy to the $1S - 7P$ energy is 1.005 and we replace this by unity. In a time dependent theory resonance denominators appear, and if the sum is really so critically balanced, we expect the observed quantity, in this case the hyperpolarizability, to vary rapidly with frequency when $\omega^2 \approx 0.005\ \Omega^2$, i.e., at a frequency 10 times lower than the first absorption edge. In nonlinear optics, no such variation is observed until one of the frequencies approaches much more closely (about 70 percent) to the absorption edge (Chang, Ducuing, and Bloembergen[24]).

Taken together these arguments give us reasonable grounds for hoping that the sums will not bite us if we remove ω_p, etc. from under the summation sign.

In the sum in (57) there is no restriction on p or q, in particular terms with either $p = 0$ or $q = 0$ occur. These will lead to trouble if we attempt to approximate the sums as they stand. We therefore first segregate all such terms. Let $\{\ \}$ denote the entire summand in (57), then

$$\begin{aligned}\sum_p \sum_q \{\ \} = \sum_p{}' \sum_q{}' \{\ \} &- x_{oo}\frac{\gamma}{\alpha}\sum_r \frac{y_{or}z_{ro}}{\omega_r^2 - \gamma^2} - z_{oo}\frac{\alpha}{\gamma}\sum_r{}' \frac{x_{or}y_{ro}}{\omega_r^2 - \alpha^2} \\ &- y_{oo}\frac{\alpha}{\beta}\sum_r \frac{z_{or}x_{ro}}{\omega_r^2 - \alpha^2} - x_{oo}\frac{\beta}{\alpha}\sum_r{}' \frac{y_{or}z_{ro}}{\omega_r^2 - \beta^2} \\ &+ z_{oo}\frac{\beta}{\gamma}\sum_r \frac{x_{or}y_{ro}}{\omega_r^2 - \beta^2} + y_{oo}\frac{\gamma}{\beta}\sum_r{}' \frac{z_{or}x_{ro}}{\omega_r^2 - \gamma^2}. \qquad (65)\end{aligned}$$

Three single sums remain unprimed, but because $\alpha = \beta + \gamma$ the terms with $r = 0$ cancel and so we may regard all the sums as primed.

We can now remove ω_p, ω_q and ω_r as a single average Ω, and this leads to an expression containing terms such as

$$\begin{aligned}\sum_p{}' \sum_q{}' x_{op}y_{pq}z_{qo} = \sum_p \sum_q x_{op}y_{pq}z_{qo} &- x_{oo}\sum_r y_{or}z_{ro} \\ &- z_{oo}\sum_r x_{or}y_{ro} + 2x_{oo}y_{oo}z_{oo}\ .\end{aligned}$$

Each of the sums on the right is now a ground state expectation value. When all the terms are collected together we obtain

$$d_{xyz}^{\alpha\beta\gamma} = \frac{Ne^3}{\hbar^2} \frac{\Omega^2(3\Omega^2 + \beta\gamma - \alpha^2)}{(\Omega^2 - \alpha^2)(\Omega^2 - \beta^2)(\Omega^2 - \gamma^2)} \{\langle xyz\rangle - \langle x\rangle\langle yz\rangle - \langle y\rangle\langle zx\rangle - \langle z\rangle\langle xy\rangle + 2\langle x\rangle\langle y\rangle\langle z\rangle\} \qquad (66)$$

which we can also write as

$$d_{xyz}^{\alpha\beta\gamma} = \frac{Ne^3}{\hbar^2} D(\Omega, \alpha, \beta, \gamma) T_{xyz} \qquad (67)$$

in terms of the, by now, familiar cubic moment T_{xyz}. This expression bears an obvious resemblance to (61) for χ.

Our expression (66) or (67) would be very nearly exact if all the optical levels had very nearly the same energy. It would then correspond to the fictitious two level system (see Refs. 15, 16, 18) often used to obliterate some of the intractable features of (57). Unlike this model, however, our expression retains the geometry of the system implicit in the selection and sum rules.

Equation (66) is possibly valid up to a frequency where one of α, β, or γ approaches the first allowed transition frequency. At somewhat lower frequencies, it is legitimate to drop the term $\beta\gamma - \alpha^2$ in the numerator. This then allows us to make a further generalization at no increase in complexity.

By removing ω_p and ω_q from (57) as a single average we have tacitly neglected the possibility that the system might be birefringent. We can remedy this by noting that in (57) each frequency ω_p or ω_q is uniquely associated with a matrix element such as x_{op} or z_{qo} which terminates on the ground state $|\ 0\rangle$ and therefore also appears in χ. Thus, we can consistently introduce three averages Ω_x, Ω_y, and Ω_z associated with correspondingly polarized transitions. If we follow this process through all its tedious ramifications, we find that, except in the term $\beta\gamma - \alpha^2$ which we are omitting, it leads to the surprisingly simple result that $D(\Omega,\alpha,\beta,\gamma)$ is replaced by

$$D(\underline{\Omega}, \alpha, \beta, \gamma) = \frac{\Omega_x\Omega_y\Omega_z(\Omega_x + \Omega_y + \Omega_z)}{(\Omega_x^2 - \alpha^2)(\Omega_y^2 - \beta^2)(\Omega_z^2 - \gamma^2)}. \qquad (68)$$

Thus,

$$d_{xyz}^{\alpha\beta\gamma} = \frac{Ne^3}{\hbar^2} \frac{\Omega_x\Omega_y\Omega_z(\Omega_x + \Omega_y + \Omega_z)}{(\Omega_x^2 - \alpha^2)(\Omega_y^2 - \beta^2)(\Omega_z^2 - \gamma^2)} T_{xyz} . \qquad (69)$$

If we compare this with the result for the classical anharmonic oscillator obtained by combining (54) with (47) we see that they are identical except for a factor 2 which once again arises because in one case we assumed that T_{xyz} was a fixed parameter while in the other it was V_{xyz}.

We now see that the classical model is equivalent to the quantal treatment, except near a resonance, in the following sense.

If we construct the model, by choosing Ω_x, Ω_y, and Ω_z to give the correct linear properties then we must choose the anharmonic term in the potential to produce a cubic moment in the ground state of the model equal to $\frac{1}{2}$ the corresponding moment in the real system. The dynamical properties of the two systems are then equivalent and the model can be used to treat more complicated systems where the quantal treatment is too difficult.

We now consider the relation of V'_{xyz} to the actual potential responsible for the existence of T_{xyz}. Obviously, the relation is obtained by requiring that both potentials yield the same cubic moment, one in the model, the other in the real system. In this case, there will be no factor of 2.

For simplicity, we consider only a system which is isotropic before the application of the anharmonic potential. Further, we restrict ourselves to atoms in which there is only one valence electron. Our results will, however, be directly applicable to atoms with more electrons if we can neglect electron correlations.

We already have an expression for the oscillator (54) which we can write as

$$T_{ijk} = -4a^6 \frac{V'_{ijk}}{\hbar\omega_o}, \tag{70}$$

where ω_o, the classical frequency, also corresponds to the first allowed transition, and

$$a = \left(\frac{\hbar}{2m\omega_o}\right)^{\frac{1}{2}} \tag{71}$$

is a measure of the extent of the system in one dimension. The direct proportionality between T_{ijk} and the corresponding component of V_{ijk} occurs because the oscillator Schrödinger equation separates in Cartesian coordinates. In general, as we shall show, it will only hold if $V = V_{ijk}x_ix_jx_k$, the crystal potential, satisfies Laplace's equation.

We will consider a more general potential of the form

$$V = \sum_{nlm} V^m_{nl} r^n P^m_l(\theta, \varphi), \tag{72}$$

where P_l^m is an associated Legendre polynomial normalized to unity. This potential satisfies Laplace's equation only if $n \equiv l$.

If the unperturbed ground state wave function is ψ_0 the first-order correction ψ_1 due to V satisfies

$$(H_o - E_o)\psi_1 + (V - E_1)\psi_0 = 0.$$

Since T is an odd moment we need only consider odd terms in V (in fact only $l = 1$ and $l = 3$) and for these E_1 vanishes since ψ_0 has definite parity.

We let

$$\psi_1 = f\psi_0 \tag{73}$$

and then

$$(H_0 - E_0)f\psi_0 = -V\psi_0$$

but, since

$$H_o = -\frac{\hbar^2}{2m}\nabla^2 + V_o\,,$$

this leads to

$$\nabla^2 f + 2\nabla f\cdot\nabla \log \psi_0 = \frac{2m}{\hbar^2} V. \tag{74}$$

Now ψ_0 is a function of r alone and so we can write

$$f = \sum_{nlm} V_{nl}^m \alpha_{nl}(r) P_l^m(\theta, \varphi), \tag{75}$$

where $\alpha_{nl}(r)$, which does not depend on m, satisfies

$$\frac{1}{r^2}\frac{\partial}{\partial r} r^2 \frac{\partial \alpha}{\partial r} - \frac{l(l+1)}{r^2}\alpha + 2\frac{\partial \alpha}{\partial r}\frac{\partial \log \psi_0}{\partial r} = \frac{2m}{\hbar^2} r^n. \tag{76}$$

The perturbed ground state is therefore,

$$\psi = \{1 + \sum_{nlm} V_{nl}^m \alpha_{nl}(r) P_l^m(\theta, \varphi)\}\psi_0(r) \tag{77}$$

and in this new ground state we can easily evaluate expectation values such as

$$\langle r^\nu P_\lambda^{-\mu}\rangle = \sum_n V_{n\lambda}^\mu \beta_{\nu n\lambda}\,, \tag{78}$$

where

$$\beta_{\nu n\lambda} = \int_0^\infty r^\nu \alpha_{n\lambda}(r)\psi_0^2(r) r^2\, dr. \tag{79}$$

In evaluating T_{ijk} we shall need $\langle x_i \rangle$, $\langle x_i^2 \rangle$, and $\langle x_i x_j x_k \rangle$. The even moments are unchanged by V and we obtain the odd moments by expanding x_i and $x_i x_j x_k$ in terms of Legendre polynomials and powers of r.

We omit most of the gruesome details of the calculation, and further restrict V to contain only terms of the type

$$V = V_{ijk} x_i x_j x_k + X_i x_i = \sum_m V_{33}^m r^3 P_3^m + V_{31}^m r^3 P_1^m + V_{11}^m r P_1^m . \tag{80}$$

The term in V_{31}^m which does not satisfy Laplace's equation is necessary to obtain the most general form of the cubic part of the potential $V_{ijk} x_i x_j x_k$. This contains 10 independent parameters while P_3^m has only 7. The missing 3 are supplied by P_1^m.

If this term is absent, we have

$$\nabla^2 V = 0 \tag{81}$$

and then

$$S_i \equiv V_{iii} + V_{ijj} + V_{ikk} = 0. \tag{82}$$

With all the terms present we obtain

$$\langle x_i \rangle = \tfrac{2}{5} \beta_{131} S_i + \tfrac{2}{3} \beta_{111} X_i \tag{83}$$

$$\langle x_i^3 \rangle = \tfrac{4}{35} \beta_{333} V_{iii} + 3\{\tfrac{2}{25} \beta_{331} - \tfrac{4}{175} \beta_{333}\} S_i + \tfrac{2}{5} \beta_{311} X_i \tag{84}$$

$$\langle x_i x_j^2 \rangle = \tfrac{4}{35} \beta_{333} V_{ijj} + \{\tfrac{2}{25} \beta_{331} - \tfrac{4}{175} \beta_{333}\} S_i + \tfrac{2}{15} \beta_{311} X_i \tag{85}$$

$$\langle x_i x_j x_k \rangle = \tfrac{4}{35} \beta_{ijk} V_{ijk} , \tag{86}$$

where it is to be understood that $i \neq j \neq k$.

If we let

$$\gamma = \langle x_i^2 \rangle, \tag{87}$$

then since $\langle x_i x_j \rangle = 0$, $i \neq j$, and $\langle x_i \rangle \langle x_j \rangle \langle x_k \rangle$ is third order in V, we obtain

$$\begin{aligned} T_{iii} = \tfrac{4}{35} \beta_{333} V_{iii} + 3\{\tfrac{3}{25} \beta_{331} - \tfrac{4}{175} \beta_{333} - \tfrac{2}{5} \gamma \beta_{131}\} S_i \\ + 3\{\tfrac{2}{15} \beta_{311} - \tfrac{2}{3} \gamma \beta_{111}\} X_i \end{aligned} \tag{88}$$

$$\begin{aligned} T_{ijj} = \tfrac{4}{35} \beta_{333} V_{ijj} + \{\tfrac{2}{25} \beta_{331} - \tfrac{4}{175} \beta_{333} - \tfrac{2}{5} \gamma \beta_{131}\} S_i \\ + \{\tfrac{2}{15} \beta_{311} - \tfrac{2}{3} \gamma \beta_{111}\} X_i \end{aligned} \tag{89}$$

$$T_{ijk} = \tfrac{4}{35} \beta_{333} V_{ijk} . \tag{90}$$

For a harmonic oscillator

$$\left.\begin{array}{lll} \beta_{333} = -35\,\dfrac{a^6}{\hbar\omega_o}, & \beta_{331} = -85\,\dfrac{a^6}{\hbar\omega_o}, & \beta_{311} = -15\,\dfrac{a^4}{\hbar\omega_o} \\ \beta_{133} = -5\,\dfrac{a^4}{\hbar\omega_o}, & \beta_{131} = -15\,\dfrac{a^4}{\hbar\omega_o}, & \beta_{111} = -3\,\dfrac{a^2}{\hbar\omega_o} \\ & \gamma = a^2 & \end{array}\right\} \tag{91}$$

and it is easy to check that the coefficients of S_i and X_i vanish, so that we recover (70).

If V satisfies Laplace's equation $S_i = 0$ and in the absence of an internal field X_i every component is given by

$$T_{ijk} = \tfrac{4}{35}\beta_{333}V_{ijk}\,. \tag{92}$$

Thus, in this case T_{ijk} and the reduced Miller tensor Δ_{ijk} have the same symmetry as V_{ijk}. Therefore, since $S_i = 0$ we have

$$\Delta_{iii} + \Delta_{ijj} + \Delta_{ikk} = 0. \tag{93}$$

If i is an axis of 3-fold or higher symmetry, $\Delta_{ijj} = \Delta_{ikk}$ and so, for example,

$$\Delta_{333} = -2\Delta_{311}\,. \tag{94}$$

This relation is rather well obeyed by the coefficients for the 6-mm crystals listed in the Table I. Signs are available only for the electro-optic

TABLE I

Material	Wavelength μ	$\Delta_{333} \times 10^6$ esu	$\Delta_{311} \times 10^6$ esu	Ratio
		Linear Electro-optic		
ZnO	optical	1.5	−0.8	−2.1
ZnS	optical	0.9	−0.45	−2.0
CdS	optical	1.2	−0.55	−2.2
		Second Harmonic		
ZnO	1.06	3.3	1.1	±3.0
ZnS	10.6	4.9	2.45	±2.0
CdS	1.06	3.2	1.6	±2.0
CdS	10.6	5.4	3.3	±1.6
CdSe	10.6	4.8	2.4	±2.0

coefficients and so the s.h.g. results represent moduli only. References to the experimental data are given in conjunction with later tables. In the case of the electro-optic data, the experimental figure is for Δ_{113} and we have assumed that Kleinman's rule (Kleinman[2]) holds and that this is equal to Δ_{311}. Except for s.h.g. in ZnO the ratio is -2 within the experimental error.

If, on the other hand, the sole perturbation in V is the field X_i, we have

$$T_{iii} = 3T_{ijj} = \tfrac{2}{5}\{\beta_{311} - 5\gamma\beta_{111}\} \tag{95}$$

and the expected ratio is $+3$. In crystals where both terms occur in V with arbitrary strength, any value of the ratio is possible. This is observed in the ferro-electric crystals $BaTiO_3$ ratio $+\frac{1}{2}$ and $LiNbO_3$ where it is $+1.7$ for the electro-optic effect and ± 11 for s.h.g. It is perhaps somewhat surprising that the ratio is so exactly -2 in the 6-mm crystals since this is a polar point group and an internal field X_3 is not forbidden by symmetry.

If V does not satisfy Laplace's equation, (it need only satisfy Poisson's equation) there is no direct relation between the components of T_{ijk} and those of V_{ijk} even in the absence of a field X_i, although, since xyz is a spherical harmonic, we still have

$$T_{123} = \tfrac{4}{35}\beta_{333}V_{123} \,. \tag{96}$$

However, since the coefficients of S_i vanish for the harmonic oscillator we may expect them to be small in other cases. We gain some support for this view by considering the hydrogen atom for which

$$\left.\begin{aligned} &\beta_{333} = -\left(\frac{105}{8}\right)^2 \frac{a_o^6}{\hbar\omega_1}, \quad \beta_{331} = -\frac{35385}{256}\frac{a_o^6}{\hbar\omega_1}, \quad \beta_{311} = -\frac{1485}{64}\frac{a_o^4}{\hbar\omega_1} \\ &\beta_{133} = -\frac{1305}{128}\frac{a_o^4}{\hbar\omega_1}, \quad \beta_{131} = -\frac{315}{16}\frac{a_o^4}{\hbar\omega_1}, \quad \beta_{111} = -\frac{81}{32}\frac{a_o^2}{\hbar\omega_1} \\ &\qquad\qquad \gamma = a_o^2 \end{aligned}\right\}, \tag{97}$$

where as usual $a_o = \hbar^2/me^2$ and $\hbar\omega_1 = 3/8(e^2/a_o)$ is the first allowed transition $(1S - 2P)$ energy.

The coefficient of V_{ijk} in each term of T_{ijk} is then $-315/16(a_o^6/\hbar\omega_1)$ while the coefficient of S_i in T_{iii} is a factor 23/200 smaller. In T_{ijj} it is 23/600 smaller. Thus, in the absence of X_i the non-Laplacean terms in V cause no more than an 11 percent departure from the relation

$$T_{ijk} = -\frac{315}{16}\frac{a_o^6}{\hbar\omega_1} V_{ijk} \,. \tag{98}$$

Since we expect the dominant terms in V to satisfy Laplace's equation it appears that T_{ijk} , Δ_{ijk} and the model potential V'_{ijk} will be very nearly proportional to the corresponding terms in V.

The potential V' required in the model is related to the crystal potential by

$$\beta^{\text{osc}}_{333} V'_{ijk} = \beta_{333} V_{ijk} \ . \tag{99}$$

For a hydrogen atom this gives $V'_{ijk} \approx 5V_{ijk}$ thus, insofar as real atoms behave like hydrogen atoms, a model with the same spatial extent $a \approx a_o$ and the same first allowed transition $\omega_0 \approx \omega_1$ will require a potential roughly five times as strong as the actual potential. This reflects the obvious fact that a harmonic oscillator is a stiffer system with more sharply localized ($\psi \approx e^{-r^2}$) wave functions than an atom ($\psi \approx e^{-r}$).

We have now shown that, with an appropriate choice of parameters a classical anharmonic oscillator model is a very good approximation to the intrinsic electronic nonlinearities of real systems.

In the next section, we use the model to consider the effect of lattice polarizability which we have so far neglected.

V. LOCAL FIELDS AND LATTICE POLARIZATION

We have already remarked in the introduction that the seat of the nonlinearities resides in the electronic motion. It is, however, considerably modified by local field corrections and in the case of optical rectification and the linear electro-optic effect by lattice polarization.

Miller's rule states that $d^{\alpha\beta\gamma}_{ijk}$ is proportional to the product of the observed linear susceptibilities X^{α}_{ii} , etc. at the appropriate frequencies. If one of these is dc we are to take the actual dc susceptibility and not the extrapolated long wavelength limit of the optical susceptibility.

At first sight, it seems plausible that this is simply the effect of internal fields, which cause the local field experienced by an atom to be greater than the applied field. We now examine this hypothesis and show that it is inadequate.

Microscopic calculations yield the polarization of single atoms due to local fields. In the linear case, if we have N atoms per unit volume of polarizability α

$$P = N\alpha E_l \tag{100}$$

and the local field is related to the applied field E by

$$E_l = E + \Gamma P. \tag{101}$$

In some cases the Lorentz value of $\Gamma = 4\pi/3$ is applicable and we then obtain the well-known relation between the refractive index n, or the dielectric constant $\epsilon = n^2$ and α.

$$\frac{n^2 - 1}{n^2 + 2} = \frac{4\pi}{3} N\alpha = \frac{R}{V}, \tag{102}$$

where V is the molar volume and R is the molar refractivity.

In general,

$$P = N\alpha(E + \Gamma P) = \frac{N\alpha}{1 - \Gamma N\alpha} E \tag{103}$$

and the observed susceptibility is

$$\chi = \frac{N\alpha}{1 - \Gamma N\alpha} = \frac{E_l}{E} N\alpha \tag{104}$$

while

$$E_L = (1 + \Gamma\chi)E = \frac{E}{1 - \Gamma N\alpha}. \tag{105}$$

In nonlinear optics the two driving fields E_j^β and E_k^γ are obviously modified according to (105) but, as Armstrong, Bloembergen, Ducuing and Pershan [18] have shown, there is a further factor in P. This arises because the nonlinear polarization

$$p_i^\alpha = d_{ijk}^{\alpha\beta\gamma}(E_j^\beta)_{\text{local}}(E_k^\gamma)_{\text{local}}, \tag{106}$$

produced directly on the atoms, further polarizes the surrounding medium.

We have

$$P_i^\alpha = p_i^\alpha + \Gamma N\alpha P_i^\alpha, \tag{107}$$

so that

$$P_i^\alpha = \frac{p_i^\alpha}{1 - \Gamma N\alpha} = (1 + \Gamma\chi_{ii}^\alpha)p_i^\alpha. \tag{108}$$

Thus, if $d_{ijk}^{\alpha\beta\gamma}$ is the (calculated) intrinsic coefficient, the observed coefficient is

$$D_{ijk}^{\alpha\beta\gamma} = (1 + \Gamma\chi_{ii}^\alpha)(1 + \Gamma\chi_{jj}^\beta)(1 + \Gamma\chi_{kk}^\gamma)d_{ijk}^{\alpha\beta\gamma}. \tag{109}$$

Therefore, even if d does not vary with χ, D will do so. This is, however, not enough to explain the observed variation of D with χ. For example, in semiconductors it is very likely that Γ is small if not zero and yet the measured values of D appear to obey Miller's rule and be proportional to χ^3. Thus, the intrinsic coefficient d itself must have a similar dependence on X,

If we write

$$D^{\alpha\beta\gamma}_{ijk} = \chi^{\alpha}_{ii}\chi^{\beta}_{jj}\chi^{\gamma}_{kk}\Delta_{ijk} \tag{110}$$

in terms of the measured susceptibilities (i.e., $n^2 - 1$), which is the content of Miller's rule, and then use (104) to express D in terms of the atomic polarizabilities we obtain

$$D^{\alpha\beta\gamma}_{ijk} = (1 + \Gamma\chi^{\alpha}_{ii})(1 + \Gamma\chi^{\beta}_{jj})(1 + \Gamma\chi^{\gamma}_{kk})N^3\alpha^{\alpha}_{ii}\alpha^{\beta}_{jj}\alpha^{\gamma}_{kk}\Delta_{ijk} \tag{111}$$

so that from (109)

$$d^{\alpha\beta\gamma}_{ijk} = N\alpha^{\alpha}_{ii}N\alpha^{\beta}_{jj}N\alpha^{\gamma}_{kk}\Delta_{ijk} \ . \tag{112}$$

Thus, the reduced tensor is the same whether or not we apply local field corrections as long as we do it consistently. To obtain a more or less constant value of Δ we must have d varying as α^3.

Since Δ for $NH_4H_2PO_4$ derived from the purely optical s.h.g. effect agrees with Δ from the quasi-static electro-optic effect to within 10 percent, although the values of d differ by a factor of 12 and in $BaTiO_3$ the two values of Δ_{311} are within a factor 2 while the d's differ by 300 it is clear that lattice polarization has a direct effect in d not described by local field terms.

We repeat that optical nonlinearities have an electronic origin. Electrons in atoms do not move in a harmonic potential. Second harmonic generation, which can only involve electronic motion, is much the same in covalent organic materials, ionic crystals and ferro-electrics. Large values of $d^{2\omega}$ are associated exclusively with large refractive indices. Thus, nonlinearities in the ionic motion play a secondary role in nonlinear optics; however important they may be in determining the ferro-electric properties.

We shall attempt to construct a model, just sufficiently general to exhibit the gross features of ferro-electric behavior, and show that it modifies the nonlinear optical behavior exactly as predicted by Miller's rule. The model is not put forward as an explanation of ferro-electricity although it has a venerable past in that connection, but as a demonstration that a simple system with singular dielectric properties behaves in a way consistent with Miller's rule.

In Fig. 1, we illustrate a moderately realistic one-dimensional model in which electrons of mass m are coupled to ions of mass M in a lattice. Forces act between like and unlike particles and of these by far the

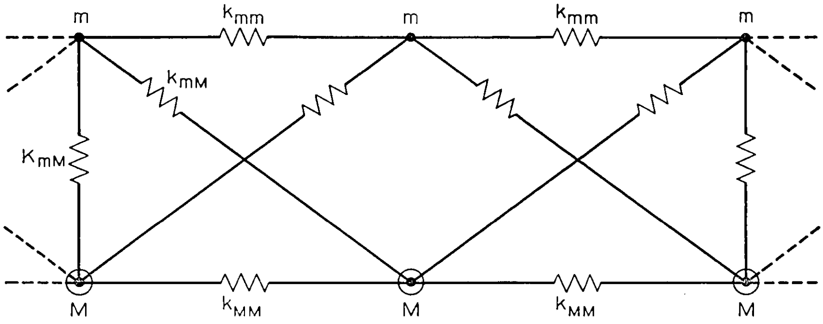


Fig. 1 — A realistic one-dimensional model.

strongest is K_{mM} which is responsible for the electronic optical spectrum. The remaining forces determine the lattice spectrum. The important nonlinearities are associated with K_{Mm}. The linear behavior of this model is formidably complicated and we therefore assume that its salient features are already evident in the much simpler model of Fig. 2.

The electron of mass $m_1 = m$ is coupled to the ion of mass $m_2 = M$ by the force constant k_{12} which replaces K_{mM}. It is anharmonic. The electron and the ion are also coupled to rigid supports representing the rest of the crystal by forces k_1 and k_2. It is as though we had gone directly from the Born-Von Karman theory of specific heats to the Einstein theory without mentioning Debye.

Let x_1 be the displacement of the electron of charge e_1 and x_2 that of the ion of charge e_2. We shall assume that the potential energy is

$$\varphi = \tfrac{1}{2}k_1x_1^2 + \tfrac{1}{2}k_2x_2^2 + \tfrac{1}{2}k_{12}(x_1 - x_2)^2 + v_{12}(x_1 - x_2)^3 \qquad (113)$$

so that the anharmonic term is exclusively associated with the "atomic" binding of the electron to its parent ion. It will be convenient to define $v_{21} = -v_{12}$. The equation of motion in a field $E^\beta e^{i\beta t}$ is then

$$m_i\ddot{x}_i + k_ix_i + k_{12}(x_i - x_j) + 3v_{ij}(x_i - x_j)^2 = e_iE^\beta e^{i\beta t} \qquad (114)$$

and the linear response neglecting v_{ij} is

$$x_i^{(1)} = \frac{(k_j - m_j\beta^2)e_i + k_{12}(e_1 + e_2)}{(k_1 + k_{12} - m_1\beta^2)(k_2 + k_{12} - m_2\beta^2) - k_{12}^2}E^\beta e^{i\beta t}. \qquad (115)$$

With N units in unit volume, the polarization is

$$P^\beta = N(e_1x_1^{(1)} + e_2x_2^{(1)})$$

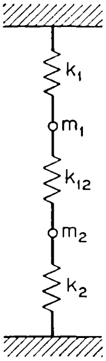


Fig. 2 — A simplified one-dimensional model.

and so

$$\chi = N\frac{e_1^2(k_2 - m_2\beta^2) + e_2^2(k_1 - m_1\beta^2) + k_{12}(e_1 + e_2)^2}{(k_1 + k_{12} - m_1\beta^2)(k_2 + k_{12} - m_2\beta^2) - k_{12}^2}. \tag{116}$$

At an optical frequency ω well above $(k_2/m_2)^{\frac{1}{2}}$ the ionic resonance

$$\chi^\omega \approx \frac{Ne_1^2}{k_1 + k_{12} - m_1\omega^2}, \tag{117}$$

while at dc

$$\chi^0 = N\frac{k_2e_1^2 + k_1e_2^2 + k_{12}(e_1 + e_2)^2}{k_1k_2 + k_{12}k_1 + k_{12}k_2}. \tag{118}$$

To obtain the sum frequency polarization due to two fields $E^\beta e^{i\beta t}$ and $E^\gamma e^{i\gamma t}$ we substitute the linear responses back into the nonlinear term in (114). The result is a nonlinear coefficient

$$d^{\alpha\beta\gamma} = -3v_{12}f(\alpha)f(\beta)f(\gamma), \tag{119}$$

where

$$f(\alpha) = \frac{e_1(k_2 - m_2\alpha^2) - e_2(k_1 - m_1\alpha^2)}{(k_1 + k_{12} - m_1\alpha^2)(k_2 + k_{12} - m_2\alpha^2) - k_{12}^2}, \quad \text{etc.} \tag{120}$$

If we express $d^{\alpha\beta\gamma}$ as $\chi^\alpha\chi^\beta\chi^\gamma\Delta$ we have

$$\Delta = -\frac{3v_{12}}{N^2e_1^3}\, g(\alpha)g(\beta)g(\gamma), \tag{121}$$

where

$$g(\alpha) = \frac{k_2 - m_2\alpha^2 - \frac{e_2}{e_1}(k_1 - m_1\alpha^2)}{k_2 - m_2\alpha^2 + \left(\frac{e_2}{e_1}\right)^2(k_1 - m_1\alpha^2) + k_{12}\left(1 + \frac{e_2}{e_1}\right)^2}, \quad \text{etc.} \qquad (122)$$

We note first that if, as seems most reasonable, $e_2 = -e_1$ then $g(\alpha) = g(\beta) = g(\gamma) = 1$. In any case at optical frequencies $g(\omega) \approx 1$ for all reasonable values of e_2/e_1 and at dc

$$g(0) = \frac{k_2 - \frac{e_2}{e_1}k_1}{k_2 + \left(\frac{e_2}{e_1}\right)^2 k_1 + k_{12}\left(1 + \frac{e_2}{e_1}\right)^2} \qquad (123)$$

which is also near unity if $e_2 \approx -e_1$. Thus, to all intents

$$\Delta \approx -\frac{3v_{12}}{N^2e^3} \qquad (124)$$

which is exactly the result obtained by neglecting the ionic motion.

Thus, Δ is an intrinsic electronic property and the effect of ionic motion is entirely contained in its effect on χ. We note, however, that in some ferro-electrics, where the departure from inversion symmetry is both small and temperature dependent, Δ will also be temperature dependent.

If k_1, k_{12}, and k_2 are all positive, the dc susceptibility is greater than the low frequency limit of χ^ω but not dramatically so. There is, however, no reason why one of these constants should not be negative. Negative compliances are familiar in classical mechanics, a well-known example is the common automatic door stop which exhibits a positive compliance as the door is first opened but a negative compliance when the door is almost fully open. The force between atoms as a whole in a lattice exhibits a positive compliance but if we separate this force into nuclear-nuclear and electron-electron repulsion and nuclear-electron attraction, it is quite reasonable to assume that at the equilibrium distance the latter component has a negative compliance.

It is immaterial which term in (113) we take as negative although on physical grounds it seems most suitable to take k_1 and this is also a convenient choice.

Provided that

$$k_1k_2 + k_{12}k_1 + k_{12}k_2 > 0 \qquad (125)$$

or

$$\eta = -k_1 \frac{k_2 + k_{12}}{k_2 k_{12}} < 1 \tag{126}$$

the system remains in stable equilibrium at $x_1 = x_2 = 0$.

The natural resonance ω_1 and ω_2 of the system satisfy

$$m_1 m_2 \omega_1^2 \omega_2^2 = k_1 k_2 + k_1 k_{12} + k_2 k_{12} \tag{127}$$

and so as $\eta \rightarrow 1$ one of these frequencies $\rightarrow 0$. At the same time the dc susceptibility (for simplicity we take $-e_2 = e_1 = e$)

$$\chi^0 = \frac{Ne^2}{k_{12}} \frac{1 - \eta \dfrac{k_{12}}{k_{12} + k_2}}{1 - \eta} \tag{128}$$

becomes infinite, while the low-frequency limit of the optical susceptibility remains finite.

If η exceeds unity there is a spontaneous polarization limited only by terms such as ωx_2^4 which we have failed to include in φ.

All this is reminiscent of ferro-electric behavior if η is temperature dependent and the Curie point corresponds to $\eta = 1$.

The inclusion of a term ωx_2^4 in φ will, in fact, make η temperature dependent, for the effect of this term is to replace k_2 by an effective value for low-frequency displacements

$$k_2' \approx k_2 + 6\omega \overline{x_2^2} = k_2(1 + \lambda T), \tag{129}$$

where $\overline{x_2^2}$ is the mean square thermal displacement. As a result if η_o is the value at $T = 0$ we have

$$\eta = \eta_0 \left(1 - \lambda T \frac{k_{12}}{k_2 + k_{12}}\right) \tag{130}$$

and so

$$\chi^0 = \frac{Ne^2}{k_{12}} \frac{\dfrac{k_2}{k_{12}} + \lambda\left(\dfrac{k_{12}}{k_{12} + k_2} T - T_o\right)}{\lambda(T - T_o)}, \tag{131}$$

if we define T_o as the temperature at which $\eta = 1$. This is of course a crude approximation to a Curie-Weiss Law.

By ascribing all the temperature dependence to changes in k_2, it is obvious from (117) that X^ω is temperature independent. For η to be equal to unity, it is not necessary for $-k_1$ to be of the same magnitude as k_{12}, all we require [see (126)] is that $-k_1$ be near k_2. Thus, from (117), we do not expect any very anomalous values of χ^{ω} in ferro-electrics, except in so far as materials with a high electronic polarizability are more likely to be ferro-electric.

We have now shown that it is possible to incorporate in our model features which lead to quite different behavior for the optical and dc dielectric constants without either invalidating Miller's rule or even changing the value of Δ which is essentially a purely electronic property.

We should, therefore, expect the temperature variation of $D_{ijk}^{\alpha\beta\gamma}$ to correspond to that of $\chi_{ii}^{\alpha}\chi_{jj}^{\beta}\chi_{kk}^{\gamma}$ and this is well borne out by the measurements of Zwicker and Scherrer[25] of the electro-optic coefficients and Bass, Franken, and Ward[26] of the optical rectification coefficients in the dihydrogen phosphates. Both coefficients are directly proportional to the dc dielectric constant which obeys a Curie Weiss Law.

In KDP there is almost no change of the s.h.g. coefficient (Van de Ziel and Bloembergen[27]) with temperature above or below the Curie point, in accord with our expectations, but at the Curie point there is a small discontinuous change. In an orthorhombic coordinate system $d_{311}^{2\omega}$ and $d_{322}^{2\omega}$ are equal above T_o but below T_o, $d_{311}^{2\omega}$ increases and $d_{322}^{2\omega}$ decreases while at the same time $\chi_{11} - \chi_{33}$ decreases and $\chi_{22} - \chi_{33}$ increases. With a constant Δ this is not compatible with Miller's rule.

However, at the transition there is a change in crystal class in which a_1 increases and a_2 decreases, Jona and Shirane.[28] It is not unreasonable to assume that this increases T_{311} and decreases T_{322} by more than enough to compensate for the changes in χ_{11} and χ_{22}.

VI. MILLER'S RULE

The classical anharmonic oscillator model, which we have shown to be a good approximation to the behavior of a real system, leads directly to that part of Miller's rule which refers to the geometric properties and frequency dependence of the nonlinear coefficients in a single material. We have also in (51) advanced a crude argument to show that Δ will not vary much from material to material.

When we examine the experimental data we shall see that the allowed components of Δ are between 1×10^{-6} and 6×10^{-6} esu for most materials but that there are a few materials with significantly higher values and a number with values as low as 0.1×10^{-6} esu.

In most cases, these exceptional values have a rather simple explanation and we have therefore to explain a constancy of Δ to within a factor of about 10.

Neglecting the effects of lattice polarization and local field corrections, which we have shown are irrelevant, the results for the classical anharmonic oscillator model are, from (50) and (56),

$$\Delta_{ijk} = 3\,\frac{T_{ijk}}{a_o e \chi^2}. \tag{132}$$

This is also the result from the static perturbation treatment of Section II.

If we use

$$\chi = 4N\,\frac{\langle x^2\rangle^2}{a_o} = \tfrac{4}{9}N\,\frac{\langle r^2\rangle^2}{a_o}\,, \tag{133}$$

we arrive at

$$\Delta_{ijk} = \frac{243}{16}\,\frac{a_o}{e}\,\frac{T_{ijk}}{N^2\langle r^2\rangle^4}. \tag{134}$$

Now the volume occupied by the oscillator is both $1/N$ and 8 $\langle r^2\rangle^{\frac{1}{2}}$ and so

$$\Delta_{ijk} \approx 10^4\langle r^2\rangle^{\frac{1}{2}}\,\frac{T_{ijk}}{\langle r^2\rangle^{\frac{3}{2}}}\ \text{esu} \tag{135}$$

where we have inserted numerical values for a_o and e. This expresses Δ_{ijk} as the product of a scale factor $\langle r^2\rangle^{\frac{1}{2}}$ and a dimensionless shape factor T/r^3. Whether we assign to each oscillator the volume per valence electron, per atom or per group of atoms, $\langle r^2\rangle^{\frac{1}{2}}$ is likely to be between 0.75 and 3 Å; so that Δ will be sensibly constant near 3×10^{-6} esu, if the shape factor is of the order of 0.01 to 0.05. We have from Turner, Saturno, Hank and Parr's[10] results for CH_4 a shape factor of 0.05, and so this range of shape factors is not unreasonable. It corresponds to a linear distortion $0.02^{\frac{1}{3}} \approx 25$ percent. It is also not unreasonable that the distortion should be of this general order, wherever it is allowed by symmetry. We may speculate that much smaller values of T/r^3 would imply very weak interatomic forces and much larger values would lead to a structure unstable relative to a more symmetric arrangement.

Thus, qualitatively, the relative constancy of Δ reflects relatively constant shape factors, although we can hardly claim that this is more than a sophisticated form of dimensional analysis. It does, however, suggest that Δ is determined primarily by the geometric properties of the molecular and crystal structure.

Large values of Δ will occur only when the molecules themselves depart considerably from inversion symmetry and are arranged in the crystal in such a way that the effects of individual parts of the molecule are additive. Small values of Δ will occur when sections of the molecule have local near inversion symmetry or when their disposition in the crystal favors the cancellation of effects from different atomic groupings. However the molecules are arranged in the lattice, Δ will be small if the molecules themselves have near inversion symmetry, or consist of uncoupled parts with the same property.

In Tables II, III, and IV, we present 50 values of Δ derived from

TABLE II—SECOND HARMONIC COEFFICIENTS

Units of d 10^{-9} esu Units of Δ 10^{-6} esu

Material	Class	λ_μ	d_{123}	Δ					Ref.
HMT = $N_4(CH_2)_6$	$\bar{4}3m$	1.06	30	17					1
ZnS	$\bar{4}3m$	1.06	153	3.5					3
ZnS	$\bar{4}3m$	10.6	146	4.5					2
ZnSe	$\bar{4}3m$	1.06	200	2.5					3
ZnSe	$\bar{4}3m$	10.6	370	6.6					2
ZnTe	$\bar{4}3m$	1.06	660	2.9					3
ZnTe	$\bar{4}3m$	10.6	440	3.6					2
CdTe	$\bar{4}3m$	10.6	800	7					2
GaP	$\bar{4}3m$	1.06	525	1.3					4
GaP	$\bar{4}3m$	1.06	255	0.6					3
GaAs	$\bar{4}3m$	1.06	1,500	1					4
GaAs	$\bar{4}3m$	10.6	1,760	3.7					2
InAs	$\bar{4}3m$	10.6	2,000	3.2					2
					d_{321}	Δ			
KH_2PO_4	$\bar{4}2m$	1.06	3	3.6	3	3.6			5
KD_2PO_4	$\bar{4}2m$	1.06	2.7	3.2	2.7	3.2			5
KH_2AsO_4	$\bar{4}2m$	1.06	3.4	2.6	3.2	2.5			4
$NH_4H_2PO_4$	$\bar{4}2m$	1.06	2.9	3.15	3.	3.15			5
			d_{333}	Δ	d_{311}	Δ	d_{113}	Δ	
ZnO	6mm	1.06	43	3.3	13	1.1	14	1.1	4
ZnS	6mm	1.06	84	1.9					3
ZnS	6mm	10.6	180	4.9	90	2.45	102	2.7	2
CdS	6mm	1.06	186	3.2	96	1.6	105	1.8	2
CdS	6mm	10.6	210	5.4	126	3.3	138	3.6	2
CdSe	6mm	1.06	500	3.4					3
CdSe	6mm	10.6	260	4.8	136	2.4	148	2.6	2
$BaTiO_3$	4mm	1.06	42	1.0	111	2.45	105	2.35	5
							d_{222}	Δ	
$LiNbO_3$	3m	1.06	250	9	36	1.1	19	0.55	6
$LiNbO_3$	3m	1.152			32	1.05	15	0.45	6
			d_{111}	Δ					
SiO_2	32	1.06	2.5	1.9					4
$AlPO_4$	32	1.06	2.5	2.2					4
Se	32	10.6	380	2.1					2
Te	32	10.6	25,400	4.3					7

Table III—Optical Rectification Coefficients

Units of d 10^{-9} esu Units of Δ 10^{-6} esu

Material	Class	λ_μ	d_{123}	Δ				Ref.
ZnTe	$\bar{4}3m$	0.694 1.06	3650 1040	14 5				8 8
				d_{321}	Δ			
KH_2PO_4	$\bar{4}2m$	0.694			50	3.2		8
KD_2PO_4	$\bar{4}2m$	0.694			105	2.9		8
$NH_4H_2PO_4$	$\bar{4}2m$	0.694	132	3.0				9
			d_{333}	Δ	d_{311}	Δ		
CdS	6mm	0.694	700	7	900	9		8

published s.h.g. data, 7 values from optical rectification data and 50 from electro-optic data. Definitions and conventions are discussed in the appendix and a separate list of references is given for the data in the appendix. Probable errors vary from measurement to measurement. It is probably safe to say that no measurement has an accuracy better than ±10 percent and in many cases the probable error is greater. That for the s.h.g. data at 10.6 μ is 30 percent and except for ADP the rectification data is only good to a factor of 3. In addition, a few materials have discordant results reported by different groups and this suggests that, especially in the case of crystals which are difficult to grow, the data should be regarded rather critically. In the case of CuCl, Sterzer, Blattner and Miniter[29] have constructed a modulator whose behavior is consistent with the higher value of the electro-optic coefficient. This casts some doubt on the low value for CuBr reported in conjunction with CuCl. In the case of the linear *e-o* coefficient in HMT, Heilmeyer's[13] value $d_{123} = 32 \times 10^{-9}$ esu is the most recent and reliable.

The average of all the s.h.g. data is $\Delta = 3.3 \times 10^{-6}$ esu, and only two coefficients d_{123} in HMT and d_{333} in $LiNbO_3$ exceed 6×10^{-6} esu by more than the probable error. One coefficient d_{222} in $LiNbO_3$ is unambiguously less than 1×10^{-6} esu. The sole accurate rectification coefficient, Δ_{123} in NH_4HPO_4, is 3×10^{-6} esu which is remarkably close to the values 3.15 and 3.2×10^{-6} esu for s.h.g. and the electro-optic effect.

Whereas the s.h.g. data have a rather compact distribution about 3×10^{-6} esu the linear *e-o* data are more straggled. The mean value is 2.3×10^{-6} esu but there are a considerable number of materials with $\Delta < 1 \times 10^{-6}$ esu. The difference between the averages $\bar{\Delta}_{shg}$ and $\bar{\Delta}_{eo}$ is not due to the different materials in the two lists, it persists if we

Table IV—Electro-optic Coefficients

Units of d 10^{-9} esu Units of Δ 10^{-6} esu

Material	Class	λ	d_{123}	Δ								Ref.
HMT = $N_4(CH_2)_6$	$\bar{4}3m$	0.5	32	14								10
HMT = $N_4(CH_2)_6$	$\bar{4}3m$		6	2.3								11
HMT = $N_4(CH_2)_6$	$\bar{4}3m$		55	21								12
$Bi_4(GeO_4)_3$	$\bar{4}3m$		22	0.8 assumed $\epsilon = 6$								13
Sodalite	$\bar{4}3m$		9.5	1.8								14
CuCl	$\bar{4}3m$		28	0.75								16
CuCl	$\bar{4}3m$		110	3								15
CuBr	$\bar{4}3m$		22	0.4 assumed $\epsilon = 10$								16
ZnS	$\bar{4}3m$	0.65	74	0.9								17
ZnSe	$\bar{4}3m$	0.55	120	0.8								18
ZnTe	$\bar{4}3m$	0.60	440	1.5								19
GaP	$\bar{4}3m$	0.63	150	0.3								20
GaAs	$\bar{4}3m$	1.02	215	0.3								21
$NaClO_3$	23	0.59	2.5	0.6								22
$K_2Mg_2(SO_4)_3$	23		$<.26$	<0.1 assumed $\epsilon = 6$								23
$(NH_4)_2Mn_2(SO_4)_3$	23		4.3	0.5 assumed $\epsilon = 9$								24
$(NH_4)_2Cd_2(SO_4)_3$	23		5.7	0.6								24
$NaVO_2(CH_3COO)_3$	23		5.3	1.3 } assumed $\epsilon = 6$								25
$Na_3SbS_4 \cdot 9H_2O$	23		10	2								26
Tren Chloride	23		9.5	2.7								27
			d_{333}	Δ	d_{113}	Δ						
ZnO	6mm	0.63	50	1.5	26	0.8 } $\frac{d_{113}}{d_{333}}$ neg.						28
ZnS	6mm	0.63	67	0.9	34	0.45						28
CdS	6mm	0.63	110	1.2	48	0.55						28
Material	Class	λ	d_{123}	Δ	d_{321}	Δ						Ref.
KH_2PO_4	$\bar{4}2m$	0.55	+65	4.0	−50	1.7 constant stress						29
KH_2PO_4	$\bar{4}2m$		60	3.7		constant strain						30
KD_2PO_4	$\bar{4}2m$		160	4.0		constant stress						31
KH_2AsO_4	$\bar{4}2m$		77	3.7	84	1.7 constant stress						32
RbH_2AsO_4	$\bar{4}2m$		92	3.5		constant stress						32
$NH_4H_2PO_4$	$\bar{4}2m$		+55	4.4	−146	3.4 constant stress						32,29
$NH_4H_2PO_4$	$\bar{4}2m$		36	3.2		constant strain						29
			d_{333}	Δ	d_{311}	Δ	d_{113}			Δ	$\frac{d_{113}}{d_{333}}$	
$BaTiO_3$	4mm	0.63	1000	1			320			0.3	+	33
					3.3×10^4	1.4		constant strain				34
					6.6×10^4	1.9		constant stress				35
			d_{333}	Δ	d_{311}	Δ	d_{113}	Δ	d_{222}	Δ	$\frac{d_{113}}{d_{333}}$	
$LiNbO_3$	3m	0.63	860	4.3	840	2.5	280	1.2	110	0.3	+	36
			d_{111}	Δ								
SiO_2	32	0.5	3.2	0.9								22
$K_2S_2O_6$	32	0.55	1.4	0.4 } assumed $\epsilon = 6$								11
$S_rS_2O_6 \cdot H_2O$	32	0.55	0.65	0.15								11
$C_6H_{12}O_6NaBr \cdot H_2O$	32	0.55	0.65	0.15								11
$CsC_4H_4O_6$	32	0.55	7	1.4								11
			d_{312}	Δ								
$C(CH_2OH)_4$	$\bar{4}$		9.7	7								37

eliminate all materials not common to both lists and may therefore, be either a real effect or a systematic error.

A few materials [e.g., $SrS_2O_6 \cdot H_2O$, $C_6H_{12}O_6NaBr \cdot H_2O$ and $K_2Mg_2(SO_4)_3$] have very low values of Δ. The latter is especially interesting since the isomorphous $(NH_4)_2Cd_2(SO_4)_3$ and $(NH_4)_2Mn_2(SO_4)_3$ salts have somewhat larger values. The ammonium cadmium salt is known to be ferro-electric at very low temperatures and the ammonium manganese salt is also suspected of ferro-electricity (Jona and Shirane[28]). More significant, perhaps, is the fact that the divalent ions have very nearly regular octahedral coordination (Zemann and Zemann[30]) and so form a unit with near inversion symmetry and contribute little to d. The main contribution comes from the monovalent ions and their irregularly placed neighbors. The difference between the potassium and ammonium salts would then be due to the difference in the polarizability of the two ions. For K^+ the refractivity is 2.45 ccs and for NH_4^+ it is 4.05 ccs (see Le Fevre[12]). If this enters d as a cube the expected ratio of the d coefficients would be 4.5. The observed value is greater than about 5. Note also that NH_4^+ itself lacks inversion symmetry.

The tabulated values of Δ show that Miller's rule is an excellent rough guide to the probable value of d. If the component is allowed by symmetry

$$| d_{ijk}^{\alpha\beta\gamma} | \approx 3 \times 10^{-6} X_{ii}^{\alpha} X_{jj}^{\beta} X_{kk}^{\gamma} \text{ esu.} \tag{136}$$

However, the rule by itself is not infallible. Occasionally, a value of d much higher than that predicted by (136) will occur. In some cases, (e.g., d_{333} in $LiNbO_3$) this is accompanied by a very low value of another coefficient and it is then plausible to assume that this is due to a particularly critical geometric configuration. In other cases, (e.g., HMT) it is quite clearly due to the coincidence of a number of favorable factors. The atoms, the molecule and the crystal all have the same symmetry and moreover, as we saw in Section 1, all the separate contributions to d have the same sign. Thus, it is likely that the value of $\Delta \approx 15 \times 10^{-6}$ esu for HMT represents something of an upper limit to what is possible.

More often (136) will overestimate d. This is especially likely to occur if the molecules themselves, or large sections of the molecule have near inversion symmetry, but it may also occur if the crystal structure itself departs only very slightly from a centro-symmetric structure.

VII. CONCLUSION

If reasonably good ground state wave functions are available, the direct perturbation method of Section II seems most suitable as a basis for calculating the magnitudes of the coefficients. It gives the intrinsic nonlinear coefficient

$$d^{a}_{ijk} = \frac{3\chi^{a}}{a_o e}\frac{T_{ijk}}{n} \tag{137}$$

in terms of the intrinsic low-frequency limit of the optical susceptibility χ^{a} and a cubic moment in the ground state. If the electrons are uncorrelated, this can be replaced by

$$d^{a}_{ijk} = \frac{3\chi^{a}}{a_o e} t_{ijk} \tag{138}$$

and t_{ijk} can be obtained from the charge distribution. From (138) we obtain the reduced tensor

$$\Delta_{ijk} = \frac{3}{a_o e}\frac{t_{ijk}}{(\chi^{a})^{2}} \tag{139}$$

and we can then incorporate this directly in Miller's rule using the observed susceptibilities χ^{α}_{ii}, etc. to obtain $d^{\alpha\beta\gamma}_{ijk}$.

This continuation of the basic perturbation calculation with Miller's rule appears to be the most straight-forward approach to the coefficients. Apart from the cubic moment t_{ijk} it involves only experimental quantities.

The analogy with the classical anharmonic oscillator established in Section IV seems most likely to be fruitful in qualitative discussions of the general behavior of the coefficients. It appears to have both empirical and theoretical justification.

Obviously, on this basis further generalizations of Miller's rule are possible. For example, we might expect the fourth rank tensor $d^{\alpha\beta\gamma\delta}_{ijkl}$ which describes induced second harmonic generation, the Kerr effect, etc. to satisfy a relation of the form

$$d^{\alpha\beta\gamma\delta}_{ijkl} = X^{\alpha}_{ii}X^{\beta}_{jj}X^{\gamma}_{kk}X^{\delta}_{ll}\Delta_{ijkl} \; . \tag{140}$$

A calculation based on fourth-order perturbation theory and a lavish use of sum rules leads to

$$\Delta_{ijkl} \approx 1.5 \times 10^{4}\,\frac{a_o}{e^{2}}\,\langle r^{2}\rangle^{\frac{3}{2}}\,\frac{Q_{ijkl}}{\langle r^{2}\rangle^{2}} \; , \tag{141}$$

where $\langle r^{2}\rangle$ is the mean square radius of the charge distribution and Q_{ijkl} is the semi-invariant

$$Q_{ijkl} = \langle x_i x_j x_k x_l\rangle - 2\langle x_i x_j\rangle\langle x_k x_l\rangle \tag{142}$$

if we assume that all odd moments vanish.

If we take $\langle r^2 \rangle^{\frac{1}{2}}$ as 1 Å this gives

$$\Delta_{ijkl} \approx 3 \times 10^{-10} \frac{Q_{ijkl}}{\langle r^2 \rangle^2} \text{ esu.} \tag{143}$$

We have seen that in the lower-order processes T/r^3 is of the order of 2×10^{-2}. This does not imply that Q/r^4 is of the order $(2 \times 10^{-2})^{4/3} \sim 5 \times 10^{-3}$ for whereas T is nonzero only because of asymmetric molecular and intermolecular forces, Q is nonzero even for free atoms or ions. For example, in the hydrogen atom $Q_{iiii}/\langle r^2 \rangle^2 = 5/18$ and $Q_{iijj}/\langle r^2 \rangle^2 = 1/9$ so that we expect Δ to be of the order of 3×10^{-11} to 10^{-10} esu. For calcite with $\chi_{\text{optical}} = 0.1$ and $\chi_{\text{dc}} = 0.55$ this gives a value of d between 3×10^{-14} and 10^{-13} esu. Bjorkholm and Siegman[31] have measured 3×10^{-14} esu.

We have seen that the reduced tensor Δ_{ijk} is proportional to the cubic moment,

$$T_{ijk} = \langle \hat{x}_i \hat{x}_j \hat{x}_k \rangle$$

and it is therefore clearly symmetric in all its indices. This is in agreement with Kleinman's[2] hypothesis and follows from the origin of the nonlinear behavior in the electronic motion.

Finally, we remark that nothing increases d like large values of the linear susceptibilities yet, although, the values of most allowed reduced tensor components Δ_{ijk} are near 3×10^{-6} esu they can vary by a factor 100:1. The molecular geometry will often indicate which end of the range is likely to apply to a particular material.

VIII. ACKNOWLEDGMENT

It is a pleasure to record many helpful discussions with C. G. B. Garrett, I. Kaminow, E. Turner, and L. Snyder. I am especially indebted to S. Kurtz for his continual interest and comments which have helped to clarify my ideas at every stage of the work described in this paper.

APPENDIX

We have throughout adopted a notation, originally introduced by Bloembergen[15] and his colleagues, in which two fields with complex time dependence $E_j^{\beta} e^{i\beta t}$ and $E_k^{\gamma} e^{i\gamma t}$ produce a polarization $P_i^{\alpha} e^{i\alpha t}$ at the algebraic sum frequency $\alpha = \beta + \gamma$ according to

$$P_i^{\alpha} e^{i\alpha t} = d_{ijk}^{\alpha\beta\gamma} E_j^{\beta} E_k^{\gamma} e^{i(\beta+\gamma)t}. \tag{144}$$

If the actual fields vary as cos $\omega_1 t$ and cos $\omega_2 t$ there will be terms in P at $\omega_1 + \omega_2$ and $\omega_1 - \omega_2$ obtained from (144) by letting $\beta = \omega_1, \bar{\omega}_1$, etc. where $\bar{\omega}_1 = -\omega_1$.

This notation has several advantages in theoretical calculations for much the same reason that the use of complex numbers simplifies ac circuit theory, and for much the same reason it has a number of disadvantages in calculating numerical values. For this reason, it has not gained general acceptance by experimentalists who tend to use a number of different notations, some of which, especially in electro-optics, are of respectable antiquity. The difference between the two notations introduces various factors of 2. These are independent of the general reluctance of physicists to state unequivocally whether they are using peak or rms fields. In particularly fertile ground, these various factors can luxuriate and blossom as factors of 8 in the final answer. We use peak fields in all our definitions.

If the applied field is

$$\underline{F}(t) = (0,\ F_2 \cos \omega t,\ F_3 \cos \Omega t), \tag{145}$$

it has components $E_2^{\bar{\omega}} = E_2^{\omega} = \frac{1}{2}F_2$, $E_3^{\bar{\Omega}} = E_3^{\Omega} = \frac{1}{2}F_3$ and the 1 component of $\underline{P}(t)$ is

$$\begin{aligned} P_1(t) = {} & \tfrac{1}{4}\{d_{123}^{\omega\,\Omega} F_2 F_3 e^{i(\omega+\Omega)t} + cc\} + \tfrac{1}{4}\{d_{132}^{\Omega\omega} F_3 F_2 e^{i(\omega+\Omega)t} + cc\} \\ & + \tfrac{1}{4}\{d_{123}^{\omega\,\bar{\Omega}} F_2 F_3 e^{i(\omega-\Omega)t} + cc\} + \tfrac{1}{4}\{d_{132}^{\Omega\bar{\omega}} F_3 F_2 e^{i(\omega-\Omega)t} + cc\}, \end{aligned} \tag{146}$$

where we have used $d_{123}^{abc} = d_{123}^{\overline{abc}}$ and suppressed the first superscript, which is always the algebraic sum of the second and third superscripts.

We can also write (146) as

$$\begin{aligned} P_1(t) = {} & \tfrac{1}{2}\{d_{123}^{\omega\,\Omega} + d_{132}^{\Omega\omega}\} F_2 F_3 \cos(\omega + \Omega)t \\ & + \tfrac{1}{2}\{d_{123}^{\omega\,\bar{\Omega}} + d_{132}^{\Omega\bar{\omega}}\} F_2 F_3 \cos(\omega - \Omega)t. \end{aligned} \tag{147}$$

If $\omega = \Omega$, this gives

$$P_1(t) = \tfrac{1}{2}\{d_{123}^{2\omega\,\omega\,\omega} + d_{132}^{2\omega\,\omega\,\omega}\} F_2 F_3 \cos 2\omega t + \tfrac{1}{2}\{d_{123}^{0\,\omega\,\bar{\omega}} + d_{132}^{0\,\bar{\omega}\,\omega}\} F_2 F_3\,. \tag{148}$$

Now the usual experimental definition would be

$$P_1(t) = (d_{123}^{2\omega} + d_{132}^{2\omega}) F_2 F_3 \cos 2\omega t + (d_{123}^{0} + d_{132}^{0}) F_2 F_3 \tag{149}$$

and so we see that

$$\text{s.h.g.} \quad d_{ijk}^{2\omega} = \tfrac{1}{2} d_{ijk}^{2\omega\,\omega\,\omega} \tag{150}$$

$$\text{rectification} \quad d_{ijk}^{0} = \tfrac{1}{2} d_{ijk}^{0\,\omega\,\bar{\omega}}\,. \tag{151}$$

If we let $\Omega = 0$, we have

$$P_1(t) = (d_{123}^{\omega\omega 0} + d_{132}^{\omega 0 \omega})F_2F_3 \cos \omega t. \tag{152}$$

The experimental definition reads

$$P_1(t) = d_{123}^{\omega}F_2F_3 \cos \omega t \tag{153}$$

i.e.,

$$\delta\chi_{12} = d_{123}^{\omega}F_3 , \tag{154}$$

so that it is possible to contract the last two suffices according to the scheme

$$11 \rightarrow 1 \;\; 22 \rightarrow 2 \;\; 33 \rightarrow 3, \quad 32 = 23 \rightarrow 4, \quad 31 = 13 \rightarrow 5, \quad 12 = 21 \rightarrow 6. \tag{155}$$

Thus, $d_{ip}^{2\omega}$ $(i = 1 \cdots 3,\ p = 1 \cdots 6)$ represents $d_{ijk}^{2\omega}$ and, for example, $d_{25} \equiv d_{231} = d_{213}$.

$$\begin{aligned} P_1 \cos 2\omega t &= (d_{123}^{2\omega} + d_{132}^{2\omega})F_2F_3 \cos 2\omega t \\ &= 2d_{123}^{2\omega}F_2F_3 \cos 2\omega t = d_{14}^{2\omega}\cdot 2F_2F_3 \cos 2\omega t. \end{aligned} \tag{156}$$

It is therefore common to define the "vector" $\mathfrak{F}$

$$\mathfrak{F} \equiv \mathfrak{F}_1 , \mathfrak{F}_2 , \mathfrak{F}_3 , \mathfrak{F}_4 , \mathfrak{F}_5 , \mathfrak{F}_6 = F_1^2 , F_2^2 , F_3^2 , 2F_2F_3 , 2F_3F_1 , 2F_1F_2 \tag{157}$$

so that

$$P_i = \sum_{p=1}^{6} d_{ip}\mathfrak{F}_p . \tag{158}$$

With this notation $d_{16}^{2\omega} = d_{123}^{2\omega} = d_{132}^{2\omega} \neq d_{123}^{2\omega} + d_{132}^{2\omega}$ but also $d_{11}^{2\omega} = d_{111}^{2\omega}$.

In the electro-optic case, the subscripts referring to optical fields can be contracted

$$2d_{ijk}^{\omega\omega 0} = d_{ijk}^{\omega} = d_{pk}^{\omega} \qquad p = 1 \cdots 6, \qquad k = 1.3. \tag{159}$$

The alternative ordering (155) leads to d_{kp}^{ω}.

Note that in this case, since d operates on two distinct fields, one optical the other dc, there is no possibility of constructing a "vector" such as $\mathfrak{F}$. The sum implied in the definition of d_{pk}^{ω} is

$$\delta\chi_p = \delta\chi_{ij} = \sum_{k=1}^{3} d_{pk}^{\omega}E_k^{dc}. \tag{160}$$

Electro-optic data are often presented as coefficients r_{pk} in the susceptibility ellipsoid. If n is the refractive index (assumed isotropic),

$$r_{pk} = -\frac{4\pi}{n^4} d_{pk}^{\omega} = -\frac{4\pi}{n^4} d_{ijk}^{\omega} = -\frac{8\pi}{n^4} d_{ijk}^{\omega\omega 0}. \tag{161}$$

The dimensions of d and r are those of an inverse field.

In the MKS system, the units are meters per volt, in the cgs system they are centimeters per stat-volt. One MKS unit is 3×10^4 esu and so numerical values of d in esu are the larger numbers.

We have not discussed the influence of a mixed use of rms and peak fields but we note that if rms fields are used throughout the values of the coefficients will all be $\sqrt{2}$ times larger than if peak fields are used throughout. No one is, however, likely to use an rms dc field.

Experimental values of the electro-optic coefficients are usually expressed in absolute units and the only ambiguity that can occur is associated with whether the measurements were made at constant stress (unclamped) or constant strain (clamped). It is safe to assume that constant stress is implied by the absence of any definite statement to the contrary.

Second harmonic coefficients are sometimes given in absolute units but more often relative to the coefficient $d_{321}^{2\omega}$ in KH_2PO_4 . An absolute measurement of this by Ashkin, Boyd, and Diedzic[32] gave

$$d_{321}^{2\omega} = \tfrac{1}{2} d_{321}^{2\omega\omega\omega} = 3 \times 10^{-9} \text{ esu},$$

but this is now believed to be too large. The most recent measurements, Francois[5], Bjorkholm,[33] give

$$d_{321}^{2\omega} = \tfrac{1}{2} d_{321}^{2\omega\omega\omega} = 1.38 \times 10^{-9} \text{ esu} \pm 12 \text{ percent}$$

for the coefficient in $NH_4H_2PO_4$. Relative measurements show that it is identical in KDP and ADP. We have used a rounded off, compromise value

$$\text{KDP } d_{321}^{2\omega} = \tfrac{1}{2} d_{321}^{2\omega\omega\omega} = 1.5 \times 10^{-9} \text{ esu} \tag{162}$$

in compiling the tables. It affects all values of $d^{2\omega}$ at optical frequencies but not at 10.6 μ.

It will be apparent that in comparing theory or experiment with experiment, considerable care is needed to be sure that like definitions are being compared with like.

REFERENCES

1. Nye, J. F., *Physical Properties of Crystals,* Oxford, 1964.
2. Kleinman, D. A., Nonlinear Dielectric Polarization in Optical Media, Phys. Rev., *126,* 1962, p. 1977.
3. Miller, R. C., Optical Second Harmonic Generation in Piezo-Electric Crystals, Appl. Phys. Lett., *5,* 1964, p. 17.

4. Patel, C. K. N., Optical Second Harmonic Generation in the Infrared, Phys. Rev. Lett., *16,* 1966, p. 613.
5. Francois, G. E., Optical Nonlinearity of Ammonium Dihydrogen Phosphate, Phys. Rev., *143,* 1966, p. 597.
6. Ward, J. F., Optical Rectification Coefficient of Ammonium Dihydrogen Phosphate, Phys. Rev., *143,* 1966, p. 569.
7. Carpenter, R. O'B., Electro-optical Effect in Crystals, J. Opt. Soc. Am., *40,* 1950, p. 225.
8. Dalgarno, A., Atomic Polarisabilities and Shielding Factors, Adv. in Physics, *11,* 1962, p. 281.
9. Dalgarno, A. and Lewis, J. T., Equivalence of Variational and Perturbation Calculations, Proc. Phys. Soc., *69,* 1956, p. 628.
10. Turner, A. G., Saturno, A. F., Hauk, P., and Parr, R. G., Description of the Distribution of Electrons in Methane, J. Chem. Phys., *40,* 1964, p. 1919.
11. Kitáigorodskii, A. I., *Organic Chemical Crystallography,* Consultants Bureau, New York, 1957.
12. Le Fevre, R. J. W., Molecular Refractivity and Polarizability, Adv. in Phys. Organic Chem., *2,* 1964, p. 1.
13. Heilmeyer, G. H., Dielectric and Electro-optic Properties of Hexamine, Appl. Optics, *3,* 1964, p. 1281.
14. Heilmeyer, G. H., Ockman, N., Braunstein, R., and Kramer, D. A., Second Harmonic Generation and the Electro-optic Effect in Hexamine, Appl. Phys. Lett., *5,* 1964, p. 229.
15. Bloembergen, N., *Non-Linear Optics,* Benjamin, New York, 1965.
16. Garrett, C. G. B. and Robinson, F. N. J., Miller's Phenomenological Rule for Computing Nonlinear Susceptibilities, J. Quantum Electronics, *QE-2,* 1966, p. 328.
17. Kurtz, S., Proc. Intl. Conf., Ferroelectrics, Prague, 1966, Czech. Acad. Sci.
18. Armstrong, J. A., Bloembergen, N., Ducuing, J., and Pershan, P., Light Waves in a Nonlinear Dielectric, Phys. Rev., *127,* 1962, p. 1918.
19. Butcher, P. N. and McLean, T. P., The Nonlinear Constitutive Relation in Solids at Optical Frequencies, Proc. Phys. Soc., *81,* 1963, p. 219.
20. Kelley, P. L., Nonlinear Effects in Solids, J. Phys. Chem. Solids, *24,* 1963, p. 607.
21. Cheng, H. and Miller, P. B., Nonlinear Optical Theory in Solids, Phys. Rev., *134,* 1964, p. A683.
22. Ward, J. F., Calculation of Nonlinear Optical Susceptabilities, Rev. Mod. Phys., *37,* 1965, p. 1.
23. Dalgarno, A., in *Quantum Theory I,* ed. Bates Academic Press, New York, 1961.
24. Chang, R. K., Ducuing, J., and Bloembergen, N., Dispersion of the Optical Nonlinearity in Semiconductors, Phys. Rev. Lett., *15,* 1965, p. 415.
25. Zwicker, B. and Scherrer, P., Helv. Phys. Acta., *17,* 1944, p. 346.
26. Bass, M., Franken, P., and Ward, J. F., Optical Rectification, Phys. Rev., *138,* 1963, p. A534.
27. Van der Ziel, J. P. and Bloembergen, N., Temperature Dependence of Optical Harmonic Generation in Ferro-Electrics, Phys. Rev., *135,* 1964, p. A1662.
28. Jona, F. and Shirane, G., Ferroelectric Crystals, Macmillan, New York, 1962.
29. Sterzer, F., Blattner, D., and Miniter, S. J., Coprous Chloride Light Modulators, Opt. Soc. Am., *54,* 1964, p. 62.
30. Zemann, A. and Zemann, J., Structure of Langbeinite, Acta. Crysta., *10,* 1957, p. 409.
31. Bjorkholm, J. E. and Siegman, A. E., Accurate cw Measurements of Optical Second-Harmonic Generation in Ammonium Dihydrogen Phosphate and Calcite, Phys. Rev., *154,* February, 1967, p. 851.
32. Ashkin, A., Boyd, G. D., and Diedzic, J. M., Phys. Rev. Lett., *11,* 1963. p. 14.
33. Bjorkholm, J. E., Optical Second Harmonic Generation Using a Focused Gaussian Laser Beam, Phys. Rev., *142,* February, 1966, pp. 126–136.

REFERENCES TO EXPERIMENTAL DATA

1. Heilmeyer, G. H., Ockman, N., Braunstein, R., and Kramer, D. A., Second Harmonic Generation and the Electro-optic Effect in Hexamine, Appl. Phys. Lett., *5,* 1964, p. 229.
2. Patel, C. K. N., Optical Second Harmonic Generation in the Infra Red, Phys. Rev. Lett., *16,* 1966, p. 613.
3. Soref, R. A. and Moos, H. W., Optical Second Harmonic Generation in ZnS—CdS and CdS—CdSe Alloys, J. Appl. Phys., *35,* 1964, p. 2152.
4. Miller, R. C., Optical Second Harmonic Generation in Piezo-electric Crystals, Appl. Phys. Lett., *5,* 1964, p. 17.
5. Miller, R. C., Kleinman, D. A., and Savage, A., Quantitative Studies of Optical Harmonic Generation, Phys. Rev. Lett., *11,* 1963, p. 146.
6. Boyd, G. D., Miller, R. C., Nassau, K., Bond, W. L., and Savage A., $LiNbO_3$—An Efficient Phase Matchable Nonlinear Material; Appl. Phys. Lett., *5,* 1964, p. 234; Miller, R. C. and Savage, A., Temperature Dependence of the Optical Properties of Ferro-electric $LiNbO_3$ and $LiTaO_3$, Appl. Phys. Lett., *9,* 1966, p. 169.
7. Patel, C. K. N., Efficient Phase Matched Harmonic Generation in Tellurium, Phys. Rev. Lett., *15,* 1965, p. 1027.
8. Bass, M., Franken, P., and Ward, J. F., Optical Rectification, Phys. Rev., *138,* 1963, p. A534.
9. Ward, J. F., Optical Rectification Coefficient of Ammonium Dihydrogen Phosphate, Phys. Rev., *143,* 1966, p. 569.
10. Heilmeyer, G. H., Dielectric and Electro-optical Properties of Hexamine, Appl. Opt., *3,* 1964, p. 1281.
11. Sliker, T. R., Linear Electro-optic Effects, J. Opt. Soc. Am., *54,* 1964, p. 1281.
12. McQuaid, R. W., Pockels Effect of Hexamethylene-tetramine, Appl. Opt., *2,* 1963, p. 320.
13. Nitsche, R., Crystal Growth and Electro-optic Effect of Bismuth Germanate, J. Appl. Phys., *36,* 1965, p. 2385.
14. Shaldin, Yu. U., Electro-optic Effect of Sodalite, Sov. Phys. Crystallog., *10,* 1966, p. 484.
15. West, C. D., Electro-optic and Related Properties of Crystals with the Zinc Blend Structure, J. Opt. Soc. Am., *43,* 1953, p. 335.
16. Belyaev, L. M., Dobrzhanskii, G. F., and Shaldin, Yu. U., Electro-optical Properties of CuCl and CuBr, Sov. Phys. Solid State, *6,* 1965, p. 2988.
17. Namba, S., Electro-optical Effect of Zinc Blend, J. Opt. Soc. Am., *51,* 1961, p. 76.
18. McQuaid, R. W., Proc. IRE *50,* 1962, p. 2484; *ibid. 51,* 1963, p. 470.
19. Sliker, T. R. and Jost, J. M., Electro-optic Effect and Refractive Index of ZnTe, J. Opt. Soc. Am., *56,* 1966, p. 130.
20. Thornber, K. K., Kurtzig, A. J., and Turner, E. H., unpublished work.
21. Turner, E. H. and Kaminow, I. P., J. Opt. Soc. Am., *53,* 1963, p. 523.
22. Landolt-Börnstein, Zahlenwerte und Funktionen II Band 8 Teil, Optische Konstanten, Springer, Berlin.
23. Thornber, K. K. and Turner, E. H., unpublished work.
24. Buhrer, C. F. and Ho, L., Electro-optic Effect in $(NH_4)_2$ $Cd_2(SO_4)_3$ and $(NH_4)_2$ $Mn_2(SO_4)_3$, Appl. Opt., *3,* 1964, p. 314.
25. Warner, J., Robertson, D. S., and Parfitt, H. T., Electro-optic effect of Sodium Uranyl Acetate, Phys. Lett., *19,* 1965, p. 479.
26. Buhrer, C. F., Ho, L., and Zucker, J., Electro-optic Effect in Optically Active Crystals, Appl. Opt., *3,* 1964, p. 517.
27. Buhrer, C. F., Electro-optic Effect in Tren Chloride, Appl. Opt., *4,* 1965, p. 545.
28. Turner, E. H., unpublished work.
29. Carpenter, R. O'B., Electro-optic Sound on Film Modulator, J. Opt. Soc. Am., *25,* 1953, p. 1145.
30. Billings, B. H., *Optics in Metrology,* Pergamon Press, Oxford, 1960.

31. Sliker, T. R. and Burlage, S. R., Dielectric and Optical Properties of $K D_2 PO_4$, J. Appl. Phys., *34,* 1963, p. 1837.
32. Ott, J. H. and Sliker, T. R., Electro-optic Effects in KH_2PO_4 and its Isomorphs, J. Opt. Soc. Am., *54,* 1964, p. 1442.
33. Kaminow, I. P., Barium Titanate Light Phase Modulator, Appl. Phys. Lett., *7,* 1965, p. 123; *ibid., 8,* 1966, p. 54.
34. Johnston, A. R., Strain-Free Electro-optic Effect in $Ba\ Ti\ O_3$, Appl. Phys. Lett., *7,* 1965, p. 195.
35. Johnston, A. R. and Weingart, J. M., Low-Frequency Electro-optic Effect in $Ba\ Ti\ O_3$, J. Opt. Soc. Am., *55,* 1965, p. 828.
36. Turner, E. H., High-Frequency Electro-optic Coefficients of Lithium Niobate, Appl. Phys. Lett., *8,* 1966, p. 303.
37. Bloch, O. G., Zheludev, I. S., and Shamburov, U. A., Electro-optic Effect in Penta-erythritol, Sov. Phys. Crystallog., *8,* 1963, p. 37.

A High-Capacity Digital Light Deflector Using Wollaston Prisms

By W. J. TABOR

(Manuscript received December 27, 1965)

A high-capacity digital light deflector (DLD) using Wollaston prisms as the passive elements is described. It is shown that, for a 1-cm aperture, approximately $4(10)^6$ resolvable positions with a crosstalk ratio of 17 dB are theoretically possible. A manually-operated model was constructed that gave $\frac{1}{4}(10)^6$ resolvable positions with a crosstalk ratio of 20 to 28 dB. The output positions of the model showed resolution approximately equal to that set by diffraction theory.

The problems associated with imperfect modulators are discussed and the characteristics of three different schemes of operation are calculated. Results from experiments with one such scheme, the reflection mode of operation, are given. They compare favorably with the calculations.

I. INTRODUCTION

A digital light deflector (DLD) is a device that can switch a light beam to a number of distinguishable positions and has been previously described by a number of authors.[1–4] Such a device can be made from a number of modulators and passive deflectors. The modulator, for this application, is one that is capable of switching the sense of polarization, and the deflector unit is a passive element which has different optical paths corresponding to the two senses of polarization. A basic unit of a DLD is shown in Fig. 1. It has been previously

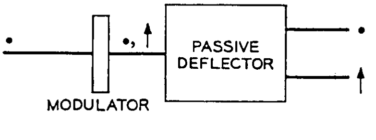


Fig. 1 — Basic unit in a digital light deflector.

shown[1–3] that n such units in series can generate 2^n distinguishable positions.

A modulator can be made from any material which can become birefringent with the application of an external signal. A minimum of π retardation is needed in order to switch the sense of polarization. Examples of modulators that have been considered for this application are Kerr cells,[5] stressed plate shutters,[6] and crystals such as KDP[3] and KTN[7,8] which exhibit an electro-optic effect. The Kerr cells, although very fast, cannot be used at high repetition rates because of heating difficulties. Stressed plate shutters, since they depend on mechanical strain, are limited to lower frequencies. The most attractive modulator materials are the electro-optic crystals and are the ones that are being seriously considered for the DLD.

The passive deflectors that have been considered for this application are uniformly thick sections of properly oriented uniaxial crystals such as calcite,[2,3] prisms of the same materials,[1] and Wollaston prisms.[4] The uniformly thick pieces of calcite are used with converging light for the maximum number of resolvable positions[2,3] but with such use suffer from aberrations that are caused by the variation of angle in the converging beam. A converging beam of light passing through a thick piece of calcite oriented for a displaced beam shows aberrations which for the most part appear like astigmatism. Prisms, when used with plane waves, can deviate the angle of the plane wave without distortion, and therefore, a DLD using prisms can give results that are limited only by diffraction theory. A DLD using prisms also uses much less birefringent material than one based on uniformly thick pieces of the same material. A disadvantage of simple prisms is that the difference in angle between the two oppositely polarized beams is only a small variation superimposed on the much larger normal type prism deflection. This difficulty can be minimized if the prisms are immersed in an oil whose index of refraction is near that of the prism. Wollaston prisms do not have the deflection associated with simple prisms—instead, the only deviation is that between the two oppositely polarized beams—and are therefore well suited for DLD use.

In this paper, the design of a high-capacity DLD using Wollaston prisms will be discussed along with experimental results which will show that this system does lead to densities limited primarily by diffraction effects. The problem of an imperfect modulator is also discussed, and calculations on several systems are given which relate the signal to background ratio to the modulation efficiency.

II. CAPACITY OF A DLD

Since a Wollaston prism is used to deflect a beam in angle, it is convenient to think of the operation of a DLD in terms of angular space. Later it will be convenient to place a lens after the DLD which will focus the beams of light, each with a different angle with respect to the axis of the lens, into corresponding points on the image plane.

The capacity of a DLD is determined by the values of two angles. One is the largest angle that is allowed in the system and the other is the minimum angular separation between adjacent positions. The capacity of the DLD is then just the square of this ratio. The minimum value is determined by either diffraction effects or imperfections in the optical system, and the largest value is determined by the maximum angular aperture of the system. First, let us consider the lower limit set only by diffraction theory. The light emerging from a circular aperture illuminated by a uniformly intense plane wave will have a spread of angles that is caused by diffraction. The intensity of the light as a function of angle is given by the well-known Airy function (Fig. 2).[9] The smallest deflection angle in a DLD must be sufficiently large so that the deflected beam must be resolved from the undeflected one. If we set the criterion that the two beams should be separated in angle such that in the far field the first dark ring of each beam overlap, then the minimum angle is given by $2.44\ \lambda/D$ where λ is the wavelength of light and D is the diameter of the circular aperture (Fig. 2).

It is possible to estimate the crosstalk, e.g., the ratio of light within the first dark ring to the light within a circle of equal size as the first dark ring but displaced, by examining the Airy function. The light within the first dark ring contains 84 percent of the total energy,[9] and a ring displaced by one diameter (corresponding to a separation of the two directions of $2.44\ \lambda/D$) falls within an annulus of $1.22\ \lambda/D$ to $3.66\ \lambda/D$ which contains 10.6 percent of the total energy.[9] Since a circle can be surrounded by six circles of the same diameter, this displaced ring contains somewhat less than $\frac{1}{6}$ (10.6 percent) $=$ 1.77 percent. The crosstalk ratio is then $84/1.77 = 47$ or 16.7 dB. For a separation between the beams of twice the above, i.e., $4.88\ \lambda/D$, the crosstalk ratio can be estimated to be 210 or 23.2 dB.

If, in addition to diffraction effects, the wavefront is distorted further by some aberration, then the focused spot size will increase and thereby decrease the capacity of the DLD. The amount of wavefront distortion depends somewhat on the type of aberration, but for wavefront distortion of $\lambda/4$ or less the increase in the focal spot size is not very significant.[10] The aberrations in the DLD will result from inhomogeneities in the material and from poorly worked surfaces. It should be emphasized that the value of $\lambda/4$ is the maximum variation allowed after passing through the entire DLD, and therefore, the maximum variation for any individual unit is much less. For a DLD with 10^6 resolvable positions, the total number of units would be 20, i.e., 2^{20} is approximately 10^6; and therefore the maximum wavefront error in any unit should be less than $\lambda/4\sqrt{20} \cong \lambda/20$ where by using the square root we have

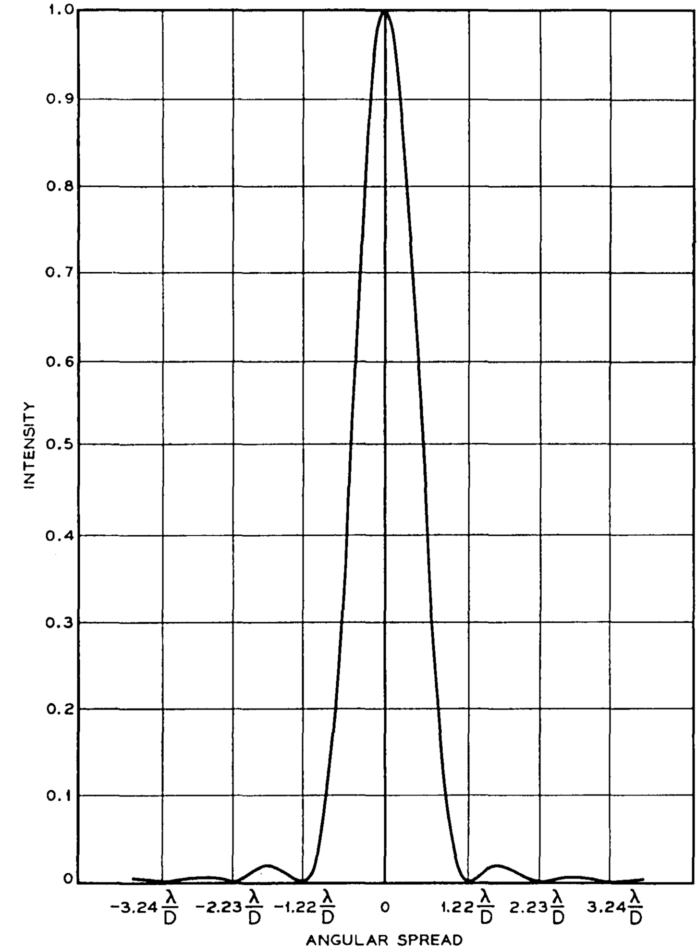


Fig. 2 — Far-field diffraction at a circular aperture (the Airy pattern).

assumed random irregularities. The requirement that a modulator have an optical distortion of less than $\lambda/20$, not allowing for any imperfections in the remainder of the optics, is extremely difficult and will probably represent a serious problem for some time to come. The high requirements placed on individual components is a direct result of the large number of such elements that must be placed in series for the complete DLD.

The maximum angular aperture of a DLD is limited by a number of effects: (*i*) the response of the Wollaston prisms, (*ii*) the walk-off of the beam as it is deflected to larger and larger angles, (*iii*) the angular aperture of the modulators, and (*iv*) the angular aperture of the output lens. These limitations will now be considered in more detail.

The deviation angle of a Wollaston prism is not constant but is a function of the incident angle (see Appendix A) and at some angle the deviation will vary sufficiently such that the array of angles is no longer uniformly spaced. Calculations based on the equations in Appendix A indicate that if the Wollaston prism with the smallest deviation is placed first in the DLD and the next largest second, etc., for a total of 20 stages and a maximum angle of deviation of 8°, the array is uniformly spaced to within 10 percent.

As the beam traverses through the DLD, the deflection angle can become larger and larger and unless the apertures of the prisms and modulators are very large, the beam will eventually strike the sides of the apparatus. It is clear that the Wollaston prism with the largest deviation angle should be placed last in the DLD in order to minimize the spreading of the beams. With this arrangement approximately ½ of the beam is intercepted by the apparatus for a 20-stage deflector with a maximum deviation angle of 8°. This latter figure is not a result applicable in all cases because it depends on the specific lengths of the elements in a DLD.

The relative retardation in a modulator is also a function of the incident angle. How rapidly this function varies depends on the type of modulator. In KDP and similar crystals the angular aperture is very small (less than 1 minute of arc for 30 dB extinction between crossed polarizers for a crystal thickness of 0.089 inches[11]) since these crystals are uniaxial, with a large birefringence, and will, therefore, give large relative retardation for even small angles away from the crystal axis. Techniques are reported that compensate the birefringence[11,12] in KDP but the degree of success is not made clear. Because of the limited angular aperture, crystals in the class with KDP are not considered attractive for the DLD. In CuCl and other crystals in this class, the angular aperture can be very large because they are cubic crystals in the field-free case and are therefore optically isotropic. Angular apertures of $\pm 25°$ have been reported[13] for this material. KTN is also a cubic crystal, but the electro-optic effect in this crystal is quadratic in contrast to the linear effect in most other materials useful as modulators. Therefore, KTN is usually biased by a dc voltage in order to reduce the value of the modulation voltage, and this bias reduces the angular aperture of the modulator; however, as shown in Appendix B, the angular aperture can still be $\pm 10°$ for reasonable bias fields.

The lens at the output of the DLD will focus the beams to points on the image plane. If this lens is not perfect, the spots will be larger than that calculated by diffraction theory, and the capacity of the DLD will be reduced. Since the choice of lens will depend on the application of the DLD, it is not possible to state very precisely what the angular aperture could be; however, $\pm 10°$ seems reasonable for most applications.

The calculations on the capacity of the DLD have been based on a plane wave of uniform intensity resulting in an Airy pattern in the far field. By placing a filter in such a beam which attenuates the light as a function of the radial distance, it is possible to greatly reduce the energy in the rings at an expense of slightly increasing the size of the central disk.[14] If such a filter could be effectively incorporated in the DLD, the crosstalk between resolvable positions could be greatly reduced without significantly reducing the overall capacity.

As an illustrative example, we will calculate the number of resolvable positions assuming that the minimum angle is set by diffraction theory which corresponds to 2.44 λ/D for 16.7-dB crosstalk and to 4.88 λ/D for 23.2-dB crosstalk, and that the maximum angle is $\pm 10°$. For a wavelength of 6000Å and an aperture of 1 cm, this corresponds to a two-dimensional array of approximately $4(10)^6$ resolvable positions with a crosstalk ratio of 16.7 dB and $(10)^6$ positions with a crosstalk of 23.2 dB. These two cases imply 22 basic units in the first case and 20 units in the latter. If one can learn to use much larger angles, then the number of positions will increase; however, on the other hand, if components are optically imperfect so that diffraction-limited performance is not possible, then these numbers will be reduced.

Thus far the DLD has only been considered in conjunction with a diffraction-limited beam; however, images may also be transmitted through the deflector. Since it takes a number of diffraction-limited points to make up an image, the capacity of a DLD in terms of images will obviously be less.

III. PERFORMANCE OF A MANUALLY-OPERATED DLD

At the present time it is not possible to construct a large capacity DLD using electro-optic modulators since these materials are not available in the quantity and quality required. In order to check the performance of this system, it is necessary to replace the electro-optic modulators with half-wave plates and thereby replace electronic activation with mechanical rotation.

A system as shown in Fig. 3 was constructed consisting of 18 mica half-wave plates, 7 pair of quartz Wollaston prisms, and 2 pair made from calcite. The aperture of the system was 18 mm and the wavelength was 6328 Å. The smallest deviation angle in the system was 1 minute and the largest was 4°; the smallest deflection angle corresponds to 8.3 λ/D, which is a separation somewhat larger than that considered earlier in this paper. The aperture of the pinhole was (0.001 inches which is larger, by a factor of approximately 4, than that required to give a diffraction-limited divergence to the wave emerging from lens 1. This system when used with an aperture of this size should be considered to be deflecting an image rather than operating with a diffraction-limited beam. The purpose of using a spot of this size is that the ratio of light in the central disk to that in the diffraction rings is much higher than when a diffraction-limited beam is used, hence the crosstalk between adjacent positions should decrease when compared to the diffraction-limited case.

Fig. 4 is a picture of a focal plane taken with this apparatus. It shows the $2^{18} \approx \frac{1}{4}(10)^6$ resolvable positions. This picture was taken by setting each half-wave plate in the halfway position so that light

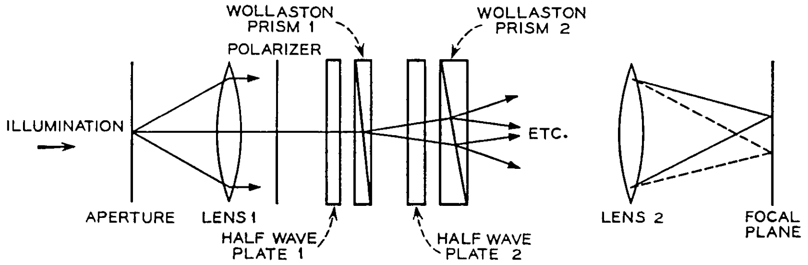


Fig. 3 — Arrangement of elements for a high capacity DLD.

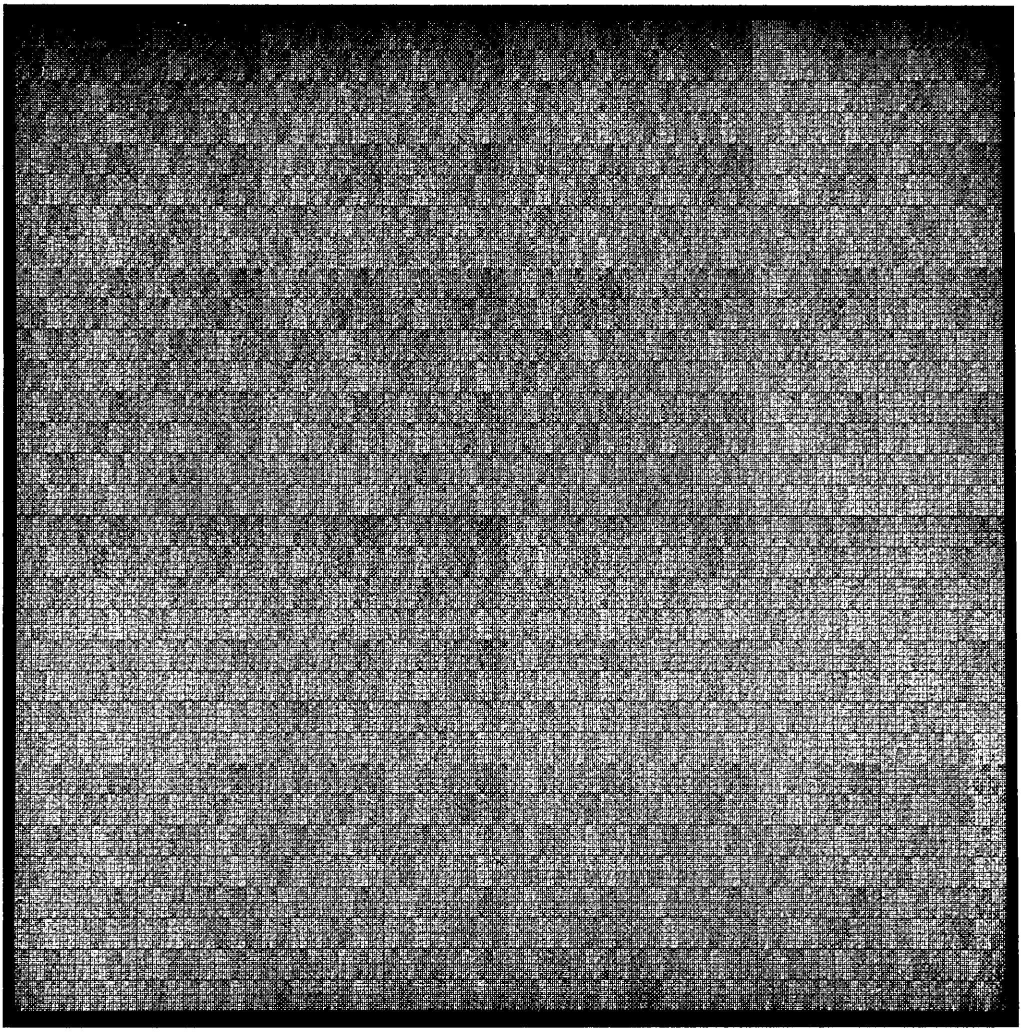


Fig. 4 — $2^{18} \approx \frac{1}{4}(10)^6$ resolvable positions of the experimental apparatus.

was divided equally into both polarizations. In this way all 2^{18} positions are simultaneously illuminated. Fig. 5 is an enlargement of an arbitrarily-selected subsection of Fig. 4 and shows the resolution much more clearly.

Fig. 6 is an enlargement of a single position taken under two cases: (*i*) the pattern on the top illustrates the focused beam with the DLD removed from the system, and (*ii*) the pattern on the bottom is the same focused beam with the DLD in the system. The degradation of the pattern on the bottom is a result of the optical imperfections in the many elements that make up the DLD. A comparison of the two patterns shows that the DLD did not increase the size of the central disk by an appreciable factor but did make it much more irregular. Fig. 6 was overexposed in order to show the weaker diffraction rings much more clearly.

The crosstalk ratio between adjacent positions was measured by first setting the DLD so that only one position was present at the focal plane. A 0.001-inch aperture was placed at the focal point, adjusted in position for maximum light transmission, and the amount of light was measured by a detector. The aperture was then moved to an adjacent position, and the amount of light passing through the opening was again measured. The ratio of these two numbers is the crosstalk. The measurements were made for various settings of the DLD and the values ranged from 20 to 28 dB. The range in the measurements is presumably due to imperfections in the optical system which cause the focused spot to be irregular and unsymmetric in shape. The irregular shape is also evident from Fig. 6.

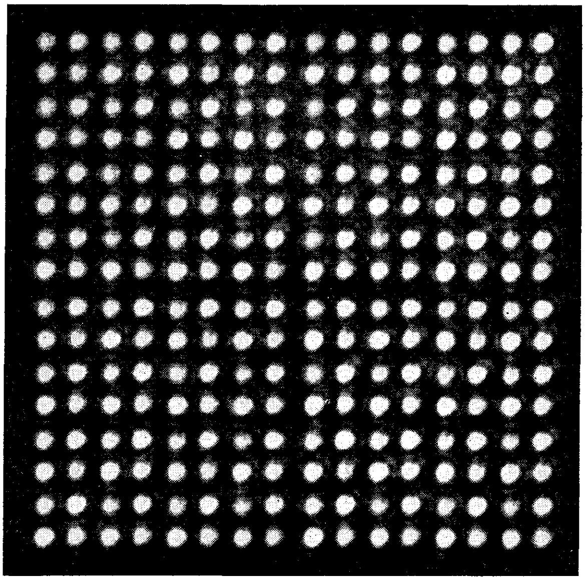


Fig. 5 — An enlargement of a section of Fig. 4.

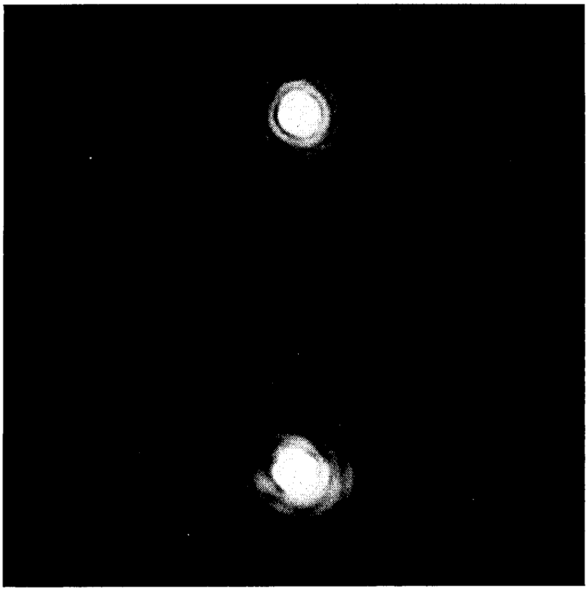


Fig. 6 — The degradation of the focused beam pattern by the DLD. (The pattern on the top was taken without the DLD in the system and that on the bottom with the DLD in the system.)

The performance of this system is probably worse than the $\lambda/4$ tolerance discussed previously in the paper but is probably not much worse than a wave or so; this latter figure was not measured directly but was estimated from diagrams which show spot patterns as a function of various aberrations.[10,15]

IV. DLD PERFORMANCE RESULTING FROM IMPERFECT MODULATORS

It is anticipated that an electronically activated modulator will be the weak link in DLD performance for some time to come; it therefore is important to know how an inefficient modulator will affect the performance of a DLD. In this study, we assume that the Wollaston prisms in the DLD are perfect and that the modulator can be characterized by a single term, E, which is defined as

$$E = \frac{a}{b} = \frac{\text{light intensity in the desired polarization}}{\text{light intensity in the undesired polarization}} \qquad (1)$$

with

$$a + b = 1.$$

A perfect modulator by this definition has an infinite efficiency.

With perfect modulators the image plane would have one bright spot at the desired position and the remaining 2^n–1 positions would be completely dark. With imperfect modulators, and for simplicity we assume that they are all imperfect to the same degree, some light will fall on every position. The resulting intensity distribution on the focal plane has been studied by others[16,17] and is also given in Appendix C. It is shown there that the intensities in the focal plane of an n unit DLD can be generated by the expansion of $a^n[1 + (1/E)]^n$ where the first term a^n gives the intensity of the desired position, the second term $n(a^n/E)$ implies n positions with intensity a^n/E, the third term $[n(n–1)/2](a^n/E^2)$ implies $n(n–1)/2$ positions of intensity a^n/E^2, etc. The sum of all the coefficients in the expansion of $(1 - 1/E)^n$ is equal to 2^n so that each position in the focal plane can be assigned to one of these terms.

It can also be established that the polarization of the even powers of E, i.e., a^n, a^n/E^2, a^n/E^4, etc. have the opposite sense of polarization from that of the odd powers of E, i.e., a^n/E, a^n/E^3, etc. This result can be determined from the basic definition of the efficiency (1).

The requirement on the modulator efficiency is determined by the particular application of the DLD. If the deflector will be used to accomplish localized heating or welding, to supply energy for a switch, or to be used as a printout or display, then the ratio of the intensity at the desired location to that at the next highest position is important. This ratio must be sufficiently large so that the intensity at the desired location must be great enough to cause the reaction, and yet the intensity at the next brightest position must be less than that to cause the reaction. For example, it is possible that for processes that depend on heating 10 dB is a sufficient ratio, whereas for a visual display 30 dB or greater may be necessary so that the eye will not be confused by multiple images. The intensity in the desired position, as previously stated, is a^n and that in the next highest case is a^n/E for the opposite polarization and a^n/E^2 for the same polarization. Therefore, this ratio is $1/E$ when there is no polarization selection and $1/E^2$ when polarization selection is used. In general, it should always be possible to eliminate the opposite polarization so that the first troublesome term will be a^n/E^2.

The DLD can also be used as a memory device.[2] In this application a memory is placed in the focal plane which is constructed such that at each of the focused points an opaque or transparent spot is present. This code is suitable for a binary organized memory where, for example, the opaque position can represent 0 and the transparent position can represent 1. The nature of the position can be read by placing a detector behind the memory and then directing a light beam to the desired location. The presence of an opaque position is determined by no response at the detector, and similarly a transparent position will result in a positive response.

When the modulators are imperfect, the undesired beams of light will also strike the memory plane and have some chance of reaching the detector with the possibility of causing erroneous results. For error-free operation the detector must receive more light when the DLD is addressed to a transparent position than when it is addressed to an opaque position, and for future reference let us call the ratio of these two intensities the signal-to-background ratio, R. The least light that can reach the detector when the DLD is addressed to a transparent position is a^n, i.e., the main beam alone, while the most light that can reach the detector when the DLD is addressed to an opaque position is $I_{\text{tot}} - a^n$, i.e., all the light except for the main beam. The minimum signal-to-background ratio is then

$$R_{\text{min}} = \frac{a^n}{I_{\text{tot}} - a^n}. \tag{2}$$

The R_{min} ratio defined by (2) is not an unreasonable minimum in that a memory plane can be designed to give ratios very close to the values calculated by this equation.

Three different ways of interrogating the memory will be discussed, and for each case calculations will be made for the signal-to-background ratios. The first case is where all of the light is allowed to strike the memory plane; the second uses a polarization selection before the memory so that only light polarized in the same sense as the main beam will reach the memory plane; and the third is the reflection mode of operation. To prevent confusion the subscripts 1, 2, 3 will be used on the R_{min} ratios for the three cases mentioned.

4.1 *Case 1—All Light From The DLD Allowed to Reach The Memory*

In this case $I_{\text{tot}} = 1$ since $a + b = 1$. Therefore,

$$(R_{\text{min}})_1 = \frac{a^n}{1 - a^n} = \frac{1}{\left(1 + \frac{1}{E}\right)^n - 1}$$

$$= \frac{1}{\frac{n}{E} + \frac{n(n-1)}{2!}\left(\frac{1}{E^2}\right) + \frac{n(n-1)(n-2)}{3!}\left(\frac{1}{E^3}\right) + \cdots}. \tag{3}$$

The signal-to-background ratio for this case is plotted as a function of modulator efficiency, E, for several values of n in Fig. 7. It shows that for an $n = 20$ DLD with an $(R_{\min})_1$ ratio of 5 dB, a modulator efficiency of 18.6 dB is required.

4.2 *Case 2—Polarization Selection Before The Memory*

If polarization selection is used after the DLD, which would require an additional modulator and polarizer, the odd powers of E can be cancelled from (2) and the $(R_{\min})_2$ ratio becomes

$$(R_{\min})_2 = \frac{1}{\dfrac{n(n-1)}{2!}\left(\dfrac{1}{E^2}\right) + \dfrac{n(n-1)(n-2)(n-3)}{4!}\left(\dfrac{1}{E^4}\right) + \cdots}. \tag{4}$$

This ratio, $(R_{\min})_2$, is plotted as a function of efficiency in Fig. 8. It shows that for $n = 20$ and $(R_{\min})_2 = 5$ dB, a modulator with an efficiency of 14.0 dB is required. This case represents an improvement of 4.6 dB over the first case.

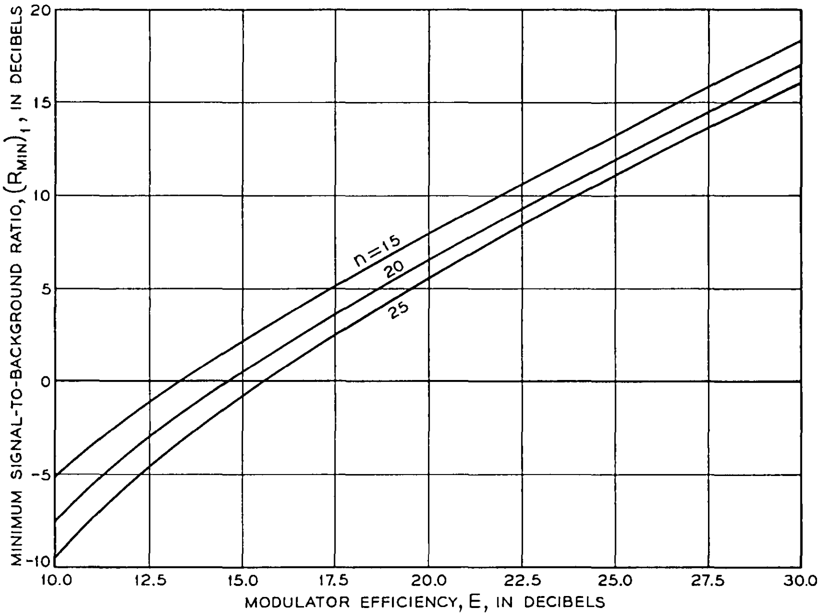


Fig. 7 — Minimum detector ratio versus modulator efficiency for a DLD where the detector is placed behind the focal plane.

4.3 *Case 3—Reflection Mode of Operation*

An alternate way[18] of reading the memory plane is shown in Fig. 9. In this case, the light is reflected from a mirror located just behind the memory and is redirected through the DLD to be detected after passing through a second aperture. The second aperture eliminates a large part of the background and therefore the ratio, $R_{\min}$, for the same modulator efficiency is considerably improved. The derivation of the $R_{\min}$ ratio for this reflecting mode of operation, $(R_{\min})_3$, is given in Appendix D and only the result is shown here:

$$(R_{\min})_3 = \frac{1}{n\left(\frac{1}{E^2}\right) + \frac{n(n-1)}{2}\left(\frac{1}{E^4}\right) + \cdots}. \tag{5}$$

$(R_{\min})_3$ is plotted as a function of E in Fig. 10. With this mode of operation a modulator with an efficiency of only 9.1 dB is required to give an $(R_{\min})_3$ ratio of 5 dB or better. The reflection mode of operation therefore represents an improvement, when measured by reduced requirements on the modulator, of 9.5 dB over the first case and 4.9 dB

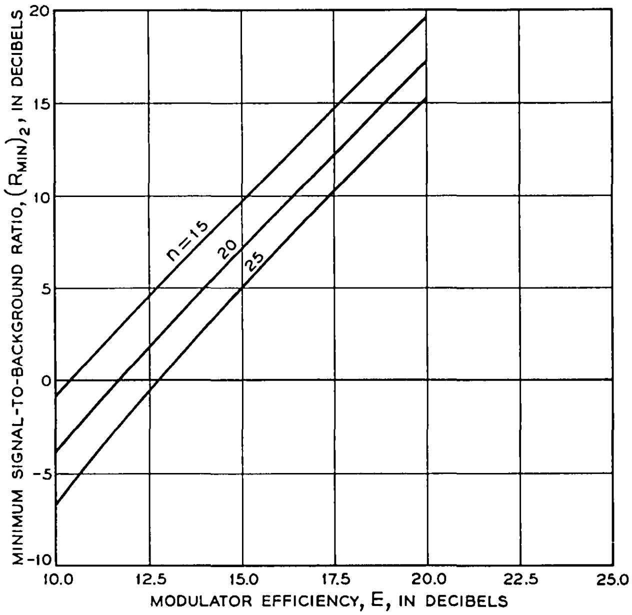


Fig. 8 — Minimum detector ratio vs modulator efficiency for a DLD with the detector placed behind the focal plane and with polarization selection.

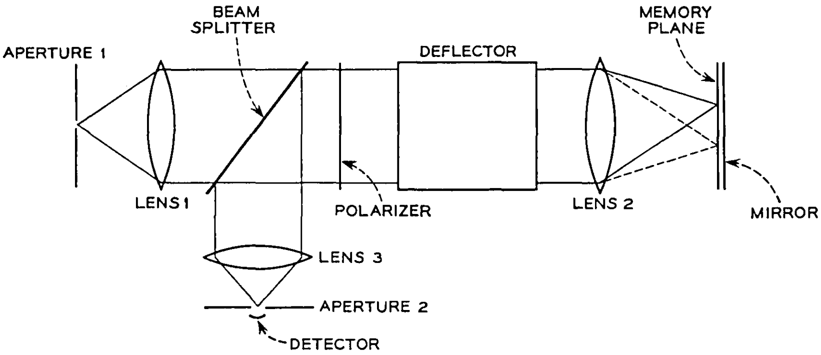


Fig. 9 — Arrangement of elements for the DLD using the reflection mode of operation.

over the second case. It should be emphasized that the improvement ratios given here are strongly dependent on the choice of the $R_{\min}$ ratio, which was taken to be 5 dB in this paper. For larger $R_{\min}$ ratios the improvement in the E ratio would be even greater and vice versa.

A disadvantage of the reflection mode is that a low f-number lens must be used at the output of the DLD. The reason for this is that the light reflected from the plane mirror must enter the output end of the DLD, and this requires that the focal plane be approximately ½ the linear dimension of the aperture of the DLD. One can show that this requires a lens of f:1.5 or so if the angular spread of the DLD is ±10°. If a lens is designed to have a spherical focal surface and a spherical mirror is used as the reflector, then the lens can have any f number.

Any light reflected from the surfaces in the DLD when used in the reflection mode can be prevented from entering the aperture near the detector by giving a slight tilt to the elements that make up the DLD. It is possible to choose an angle such that no reflection is centered on the aperture.

V. EXPERIMENTS WITH THE MANUALLY-OPERATED DLD EQUIPPED WITH POOR MODULATORS

The experiments that will be described in this section make use of the n = 18 manually-operated DLD where the half-wave plates have been substituted for the electro-optic modulators. This is the same apparatus as used in Section III.

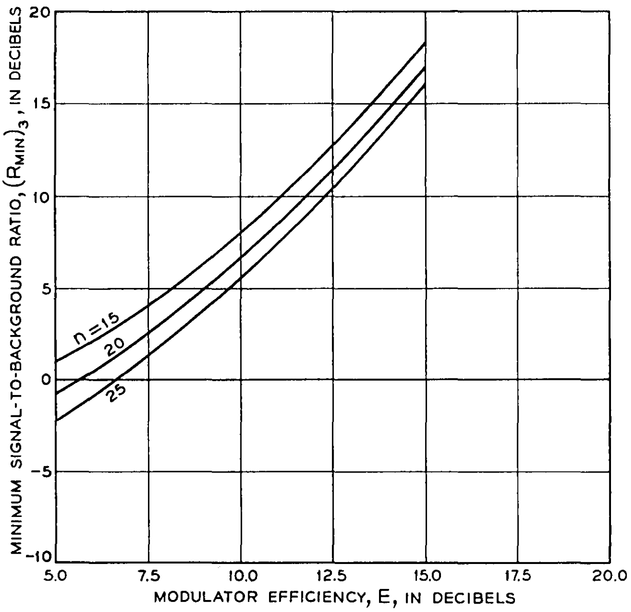


Fig. 10 — Minimum detector ratio versus modulator efficiency using the reflection mode of operation.

Ideal half-wave plates have the property that if the angle between the polarization direction and the axis of the half-wave plate is θ then the plane of polarization emerging from the plate will be rotated by 2θ from its original direction. For maximum efficiency, the half-wave plates are oriented at $\theta = 0$ if no switching action is desired and at $\theta = 45°$ if the other polarization is desired. The practical maximum efficiency for the split mica plates used in this experiment ranged from 30 to 40 dB, which is high enough to give almost perfect DLD behavior. In order to simulate poor modulators, the wave plates are set at $\theta = \epsilon$ for the predominantly unswitched case and at $\theta = 45° - \epsilon$ for the predominantly switched case. The angle ϵ can then be adjusted to achieve any degree of modulator efficiency.

To illustrate the behavior of a DLD under the influence of poor modulators, the half-wave plates were set for $E = 10$ dB, and a picture, which is shown in Fig. 11, was taken at the focal plane. This picture shows some characteristics which will now be enumerated:

(*i*) The desired spot is shown as the brightest point in the upper left-hand quadrant and is vertically polarized.

(*ii*) Some of the 18 points whose intensity is $1/E$ of the main beam are shown in the lower right-hand quadrant and are horizontally polarized.

(*iii*) Some of the 153 points whose intensity is $1/E^2$ of the main beam are shown in the upper left-hand quadrant and are vertically polarized.

(*iv*) Some of the 816 points whose intensity is $1/E^3$ of the main beam are shown in the lower right-hand quadrant and are horizontally polarized.

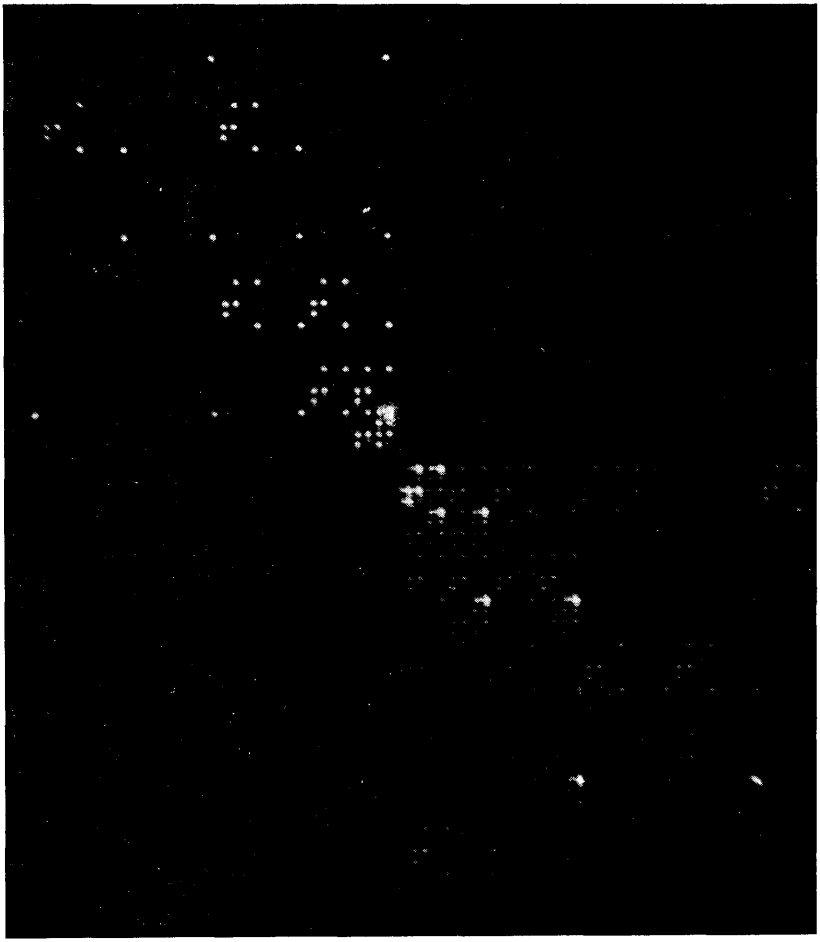


Fig. 11 — Focal plane intensity distribution with modulators set at E = 10 dB.

(*v*) The points in the upper left-hand quadrant are vertically polarized and those in the lower right-hand quadrant are horizontally polarized. A polarization selector in this case would eliminate the entire lower side. It will always eliminate the side that is opposite to the one that contains the main beam.

(*vi*) There are no points with significant intensity ($> 1/E^{n/2}$) in either the upper-right or lower-left quadrant.

As all of the focal plane is not shown in this figure, some of the background positions are missing in order that an enlargement could be presented. In this exposure there is a total of approximately 4.6 times more light energy in the background than in the main beam.

The performance of the reflection mode of operation was compared to the theoretical calculation by making use of the properly misoriented half wave plates. For two reasons (5) cannot be directly used: (*i*) In this experiment a memory was used that contained only one opaque position and for such a case the signal-to-background ratio using this equation is not accurate, and (*ii*) equation (5) assumes that the opaque positions are also perfectly absorbing which is not valid for this experiment in that the opaque position reflected 4 percent of the incident power. A signal-to-background ratio for this particular experiment can be calculated as follows. When the DLD is addressed to a transparent position the light reaching the mirror is I_{tot} and the light striking the mirror when the DLD is addressed to the opaque position is $I_{\text{tot}} - (1 - \Gamma)a^n$ where Γ is the power reflection coefficient. The ratio of these two values is

$$(R_{\text{exp}})_3 = \frac{I_{\text{tot}}}{I_{\text{tot}} - (1 - \Gamma)a^n} = \frac{I_{\text{tot}} - (1 - \Gamma)a^n}{I_{\text{tot}} - (1 - \Gamma)a^n} + \frac{(1 - \Gamma)a^n}{I_{\text{tot}} - (1 - \Gamma)a^n}$$

$$= 1 + \frac{(1 - \Gamma)a^n}{\dfrac{a^n}{(R_{\text{min}})_3} + \Gamma a^n} = 1 + \frac{(1 - \Gamma)}{\dfrac{1}{(R_{\text{min}})_3} + \Gamma}. \qquad (6)$$

The calculated and experimental values of $(R_{\text{exp}})_3$ are summarized in Table I. The first column lists the modulator efficiency, the second column contains the calculated $(R_{\text{exp}})_3$ and the third column lists the measured values.

The agreement between the calculated and measured quantities agree very well and indicate that the behavior of the reflection mode of operation is adequately understood.

TABLE I

Modulator efficiency	Signal-to-background ratio for the reflection mode experiment, $(R_{exp})_3$	
dB	Calculated in dB	Measured in dB
30	14.0	14
20	13.8	14
10	7.0	8

VI. PARALLELLING THE OUTPUT

In the DLD thus far discussed, only one memory in the output focal plane is used (see Fig. 3), and therefore the memory is read one bit at a time. For some applications it may be advantageous to parallel the output as shown in Fig. 12 in order to increase the bit capacity of the DLD. With this scheme the number of bits read with each setting of the DLD is equal to the number of memory planes; the memory now corresponds to one with word organization. Fig. 12 shows four such memory planes but by going to a three-dimensional array it is possible to parallel 30 to 40 such planes and still use only one output lens providing the maximum angular aperture of the DLD is limited to a total angle of 12° or so. If additional lenses are employed, then any number of memory planes can be incorporated.

This scheme of paralleling is directly applicable to cases 1 and 2, which were discussed in Section IV, but will not work for the reflection mode since there is no way to distinguish between different memory planes when the light is redirected through the DLD. In order to parallel the output of the reflection mode, three different schemes have been devised: (*i*) to use different wavelengths for each memory plane and separate the colors before and after the DLD,[19] (*ii*) to modulate a monochromatic beam at each memory plane with a different frequency and then to separate the different frequencies after reflection through the DLD,[20] and (*iii*) to arrange the memory planes to have different distances between the output of the DLD and the memory plane and to use a short pulse of light; the different planes can now be read since each plane will return the pulse to the detector at a different time.[21]

One difficulty that can arise in the paralleling schemes is that very low *f*-number lenses must be used to refocus the output plane into repeated images (Fig. 13). The first lense placed after the DLD can

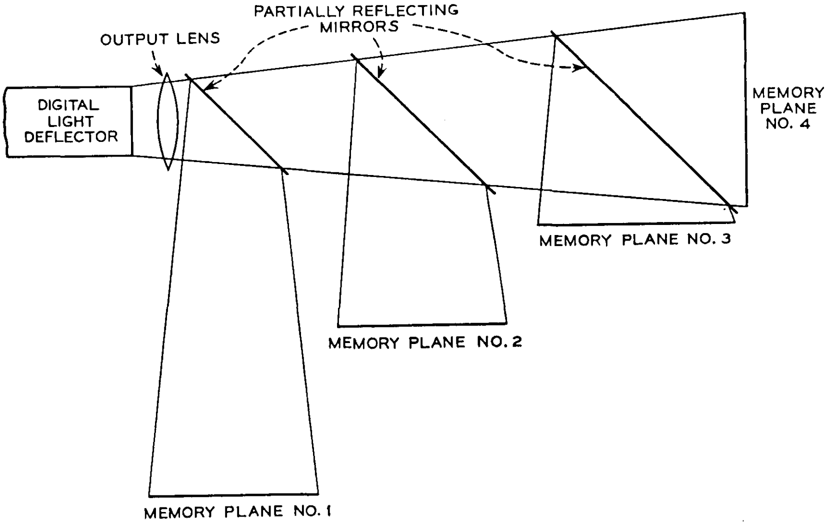


Fig. 12 — One method of paralleling the output of the DLD.

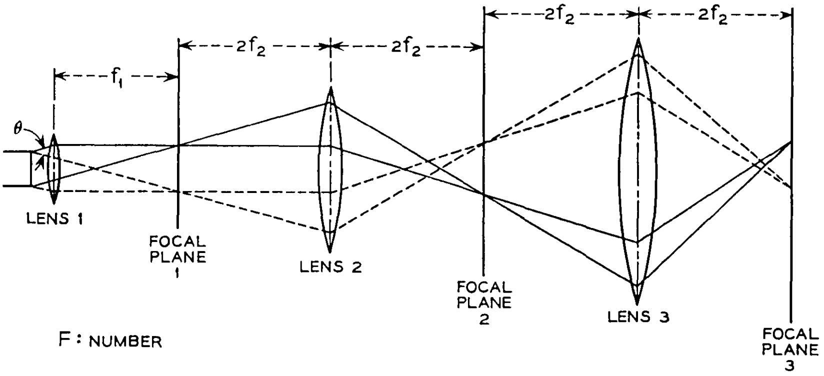
have a reasonable f number since the light beams are still confined to the aperture of the DLD. The second lens must have f:1 or so if the output of the DLD has a total angle of approximately 12°. Additional lenses must have even lower f numbers (Fig. 13). The practical solution to this problem is to perform all of the paralleling within the first focal length. This scheme will not work with the reflection mode where the time of flight varies for each memory plane since one lens implies only one distance between it and the various memory planes.

VII. MEMORY MEDIA

The problem of reading a memory has been discussed in earlier sections of this paper; we will now consider the problem, which is again primarily the result of imperfect modulators, of using the DLD to write into a memory. Two general types of materials will be considered for use as a memory medium. One is a medium where the process is linear in terms of total exposure, i.e., the effect on the medium of n pulses of light of intensity I/n, each lasting for a time ΔT, is the same for any value of n; and the second is one which has a threshold in terms of the light intensity. An example of the first type is a photographic film and of the second is a memory based on a transparent ferrimagnetic garnet at its compensation temperature.[22]

A linear medium has the disadvantage, for this application, in that it integrates the light striking its surface. Therefore, when the photographic film is being exposed by a DLD with poor modulators, the problem of the background light must be considered. This problem is different from that considered in Section IV because in that case the DLD was set for one address and the question was asked what is the light intensity distribution over the whole focal plane. In this case, we ask what is the total amount of light energy striking one position on the memory plane when the DLD is addressed to all of the positions. When the DLD is set for one address, the total exposure over the whole plane is $a^n[1 + (1/E)]^n \Delta T$ where ΔT is the duration of the exposure. This result is evident from the discussion in Section IV. When one sits at a position and the DLD is addressed to all positions and dwells at each position for the same time ΔT, the total exposure at that position is also $a^n[1 + (1/E)]^n \Delta T$. This latter result has been previously published[17] and is also proven in Appendix C.

The following calculations, which represent worst cases for writing, can be performed. The first case that will be considered is the situation where all of the light is allowed to strike the memory plane and the second case makes use of polarization selection. In both cases it will be assumed that all points will be addressed, except for one.



F : NUMBER

$$\text{LENS 1} = \frac{D}{f_1}$$

$$\text{LENS 3} = \frac{1}{2\left[\left(\frac{3f_1}{2f_2} + 1\right)\theta + \frac{1}{F_1}\right]}$$

$$\text{LENS 2} = \frac{1}{2\left[\left(\frac{f_1}{2f_2} + 1\right)\theta + \frac{1}{F_1}\right]}$$

$$\text{LENS N} = \frac{1}{2\left[\left(\frac{2N-3}{2}\,\frac{f_1}{f_2} + 1\right)\theta + \frac{1}{F_1}\right]}$$

Fig. 13 — Lens requirements for re-imaging the output focal plane of the DLD.

7.1 *Case 1—All Light from the DLD Allowed to Reach the Memory*

In this case the light striking any of the addressed positions will have an exposure of nearly $a^n(1 + 1/E)^n\Delta T$ and the exposure at the one position that was not addressed will be $[a^n(1 + 1/E)^n - a^n]\Delta T$, and the ratio of these two values is

$$M_1 = \frac{a^n\left(1 + \frac{1}{E}\right)^n}{a^n\left(1 + \frac{1}{E}\right)^n - a^n} = \frac{a^n\left(1 + \frac{1}{E}\right)^n - a^n}{a^n\left(1 + \frac{1}{E}\right)^n - a^n} + \frac{a^n}{a^n\left(1 + \frac{1}{E}\right)^n - a^n} \qquad (7)$$

$$= 1 + (R_{\min})_1 .$$

7.2 *Case 2—Polarization Selection Before the Memory*

In the same way as in the first case one can derive

$$M_2 = 1 + (R_{\min})_2 . \qquad (8)$$

The exposure ratios M_1 and M_2 are plotted as a function of modulator efficiency in Figs. 14 and 15. These plots can be used to determine the minimum modulator efficiency required for a certain exposure ratio.

From the exposure ratio and the properties of the medium, e.g., photographic film, the density ratios of the positions can be calculated. These two ratios do not have to be the same, as a material such as photographic film can be very linear in terms of exposure but the exposure vs film density can be very nonlinear.

For a theshold medium the problem is much simpler. The only requirement is that the most intense beam must be greater than threshold and that the next highest position be less than threshold. As mentioned in Section IV, this ratio is $1/E$ when no polarization selection is used and $1/E^2$ when polarization selection is used.

VIII. ANALOG CONTRASTED WITH DIGITAL DEFLECTION

The same large number of resolvable positions as described in this paper could, in principle, be achieved by means of an analog deflection; an example of this is a prism of an electro-optic material with electrodes placed on the parallel surfaces. It is necessary to induce an increase of 2π retardation along the base of such a prism in order to deflect the beam by one resolvable position. Therefore, for a 10^6 position deflector it is necessary to have $2(10)^3\pi$ along the X bank and

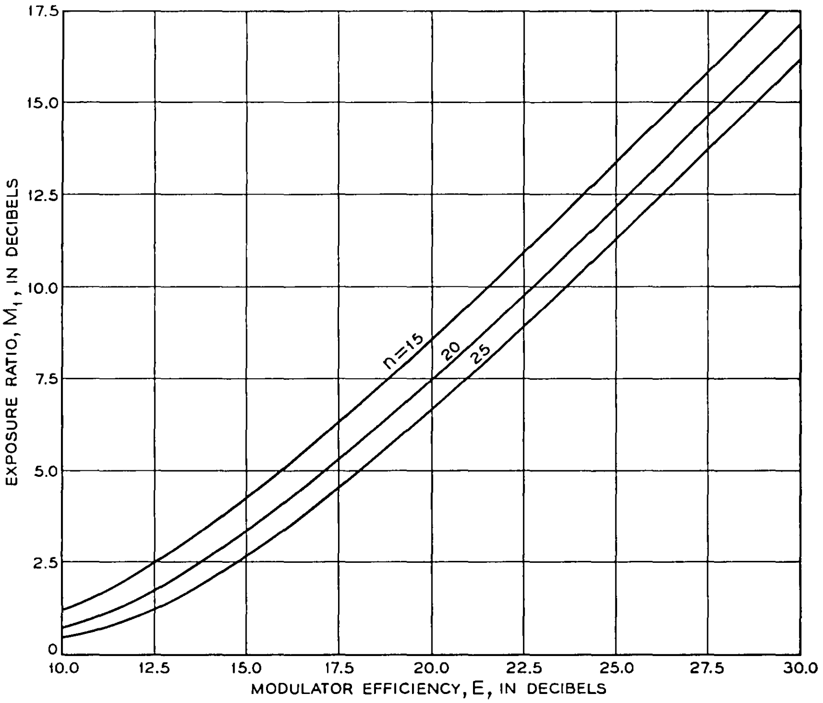


Fig. 14 — Exposure ratio vs modulator efficiency for a DLD where all of the light is allowed to reach the memory plane.

$2(10)^3\pi$ along the Y bank for a total of $4(10)^3\pi$ total retardation. With a DLD, a total retardation of 20π can accomplish the same number of resolvable positions. Therefore, it is evident that the DLD makes very efficient use of the variable retardation. The reason for this efficiency is that the DLD makes use of the fixed retardation in the passive elements whereas the analog deflector must generate all of the retardation. In addition, the DLD can be designed for any separation between the beams and still not require any more than the 20π variable retardation. The analog reflector, on the other hand, cannot separate the beams any further without supplying additional retardation.

IX. CONCLUSION

The construction and characteristics of a high-capacity DLD have been described, and it has been demonstrated that the number of

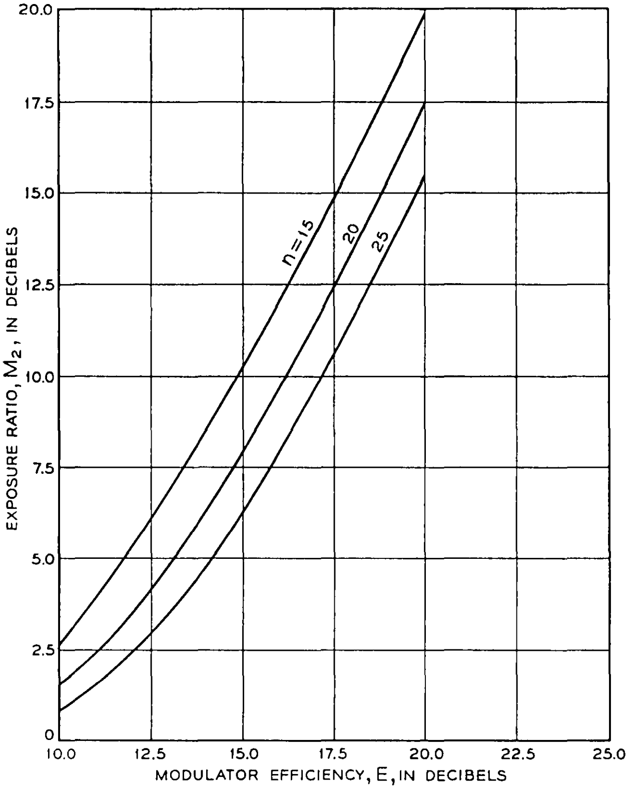


Fig. 15 — Exposure ratio vs modulator efficiency for a DLD where polarization selection is used.

resolvable positions that can be attained is reasonably close to that allowed by diffraction theory. The effect of imperfect modulators on the performance of the DLD has also been discussed.

The discussion presented here does not mention the problems associated with the high-speed switching of an electro-optic modulator, which is a problem that must be solved if the DLD is to have broad application. This problem has been studied by S. K. Kurtz.[23]

X. ACKNOWLEDGMENTS

The author is greatly indebted to H. E. D. Scovil and K. D. Bowers for valuable discussions involving the applications and overall performance requirements of the digital-light deflector.

The author is also indebted to J. T. Sibilia and J. E. Geusic for their contributions to many of the solutions associated with the reflection mode of operation. He also wishes to acknowledge discussions with R. G. Smith on some of the problems associated with imperfect modulators and to J. G. Skinner on some of the optical problems.

APPENDIX A

Optical Properties of a Wollaston Prism

The formulas necessary to trace the two wave normals through a Wollaston prism in a direction as shown in Fig. 16 are given below:

$$\begin{aligned} \sin \beta &= \frac{1}{n_o} \sin \gamma \\ \tan \alpha &= \frac{n_o}{n_e} \tan \beta \\ \sin (\theta + \epsilon) &= \frac{\sin \beta}{\sin \alpha} \sin (\theta + \alpha) \\ \sin \alpha &= n_o \sin \epsilon \end{aligned} \tag{9}$$

. . . .

$$\begin{aligned} \sin \beta &= \frac{1}{n_o} \sin \gamma \\ \sin (\theta + \delta) &= \frac{n_o}{n_e} \sin (\theta + \beta) \\ \sin b &= n_e \sin \delta, \end{aligned} \tag{10}$$

where the symbols are defined in Fig. 16.

A useful approximate formula for calculating the total deviation angle of a Wollaston prism, Δ, $(\Delta = a + b)$ for perpendicular incidence, $\gamma = 0$ is

$$\Delta = (\,|\,a\,|\, + \,|\,b\,|\,)_{\gamma=0} = 2\,|\,n_o - n_e\,|\,\tan \theta + \cdots . \tag{11}$$

The variation of Δ with respect to a variation in γ at perpendicular incidence, $(\partial\Delta/\partial\gamma)_{\gamma=0}$, can be calculated from (9) and (10) to be

$$\left(\frac{\partial \Delta}{\partial \gamma}\right)_{\gamma=0} = \cos \theta \left\{ \frac{\cos \delta}{\cos b \cos (\theta + \delta)} - \frac{\cos \epsilon}{\cos a \cos (\theta + \epsilon)} \right\}, \tag{12}$$

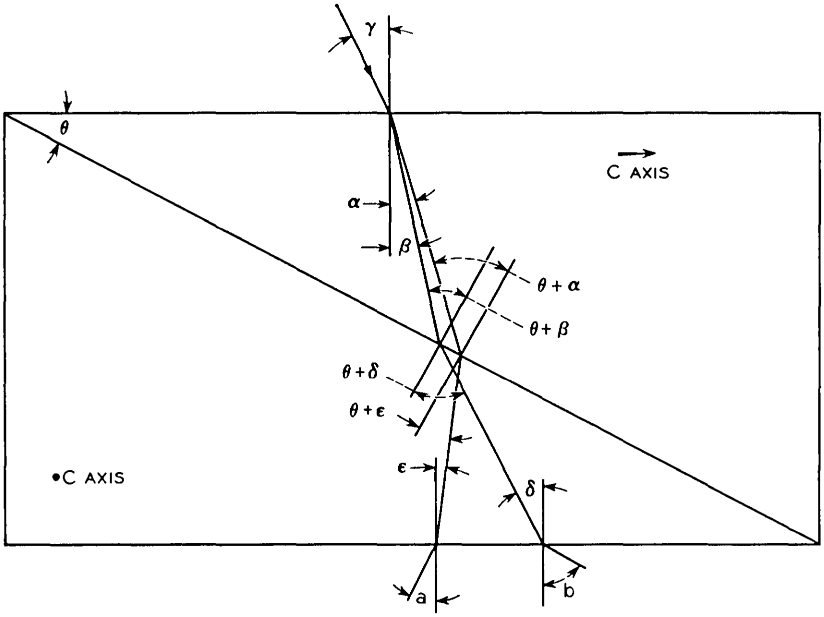


Fig. 16 — Diagram for wave normal paths in a Wollaston prism.

which to a good approximation can be reduced to

$$\left(\frac{\partial\Delta}{\partial\gamma}\right)_{\gamma=0} = (n_o - n_e)\left(\frac{1}{n_e} + \frac{1}{n_o}\right) \tan^2 \theta. \tag{13}$$

APPENDIX B

Angular Aperture of a Modulator

This calculation is valid for materials that are cubic, and therefore optically isotropic, in the absence of an electric field and become uniaxial, with the optic axis parallel to the electric field, in the presence of an electric field.

Fig. 17 describes the placement of the crystal with respect to the incident radiation. The xy plane is the first surface of the modulator, and the second surface is parallel to the first and passes through the point Z equals $-T$. The induced C axis of the crystal is parallel to the y axis. The light ray makes an angle γ with the z axis, and the intersection of the plane of incidence with the xy plane makes an angle α with the x axis. The relative retardation between the extraordinary and ordinary ray can be calculated to be

$$R(e - o) = \left\{ n_e \left[1 - \frac{\cos^2 \alpha \sin^2 \gamma}{n_e^2} - \frac{\sin^2 \alpha \sin^2 \gamma}{n_o^2} \right]^{\frac{1}{2}} - n_o \left[1 - \frac{\sin^2 \gamma}{n_o^2} \right]^{\frac{1}{2}} \right\} \frac{2\pi T}{\lambda_o}, \tag{14}$$

where n_o and n_e are the ordinary and extraordinary indices of refraction and λ_o is the free-space wavelength of the light. Equation (14) reduces the familiar $(n_e - n_o)(2\pi T/\lambda_o)$ for perpendicular incidence.

Since n_o and n_e are nearly the same in this case, we will expand (14) in terms of powers of $(n_e - n_o)$ and drop terms containing $(n_e - n_o)^2$ and higher. We will also expand sin γ using a power series in γ. The result of these substitutions is

$$\begin{aligned} R(e - o) = \Big\{ 1 &- \frac{1}{n_o^2}(\cos^2 \alpha - \tfrac{1}{2})\gamma^2 \\ &- \left[\frac{1}{3n_o^2}(\cos^2 \alpha - \tfrac{1}{2}) - \frac{1}{n_o^4}\left(\frac{\cos^2 \alpha}{2} - \frac{1}{8} \right) \right] \gamma^4 \\ &+ \left[\frac{2}{45n_o^2}(\cos^2 \alpha - \tfrac{1}{2}) - \frac{2}{3n_o^4}\left(\frac{\cos^2 \alpha}{2} - \frac{1}{8} \right) \right. \\ &\left. + \frac{1}{n_o^6}\left(\frac{3 \cos^2 \alpha}{8} - \frac{1}{16} \right) \right] \gamma^6 \Big\} \frac{2\pi T(n_e - n_o)}{\lambda_o}. \end{aligned} \tag{15}$$

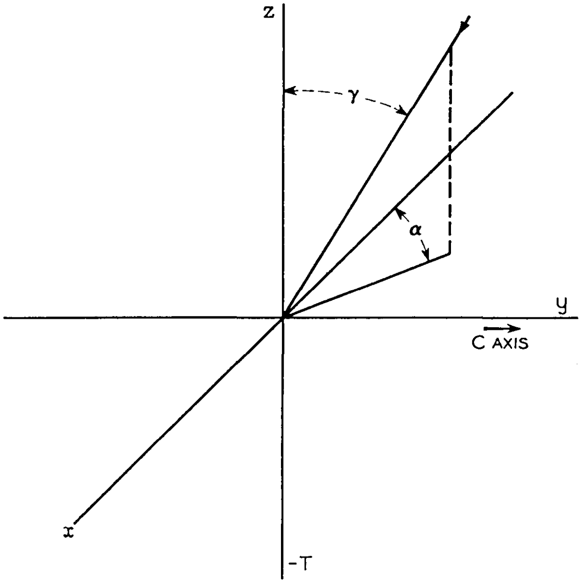


Fig. 17 — Coordinate axes showing the placement of the uniaxial crystal with respect to the incident radiation.

Equations (14) and (15) can be used to describe the interference pattern obtained with a uniaxial crystal whose C axis is parallel to the crystal surface is placed between crossed polarizers. This pattern can be observed in any standard text on optics.[24]

In order that a modulator switch the sense of polarization, it is necessary that the retardation be changed by π. In general, the retardation will be changed from $N\pi$ to $(N + 1)\pi$ with the application of an electric field. The retardation as a function of incident angle for the largest retardation, $(N + 1)\pi$, using terms only up to γ^2, becomes

$$R = (N + 1)\pi\left\{1 + \frac{1}{n_o^2}(\cos^2 \alpha - \tfrac{1}{2})\gamma^2\right\}. \tag{16}$$

The angular aperture of a modulator is determined by the value of γ where the change in retardation from π becomes serious enough to cause unwanted behavior. If we call this change in retardation ΔR, then the value of γ that corresponds to this ΔR can be calculated from (16), i.e.,

$$\gamma = \left[\frac{n_o^2 \Delta R}{(N + 1)\pi(\cos^2 \alpha - \frac{1}{2})}\right]^{\frac{1}{2}}. \tag{17}$$

From (17) one can see that the angular aperture for a given material is inversely proportional to $\sqrt{N + 1}$ so that we can write

$$\gamma_{\text{biased}} = \frac{\gamma_{\text{unbiased}}}{\sqrt{N + 1}}, \tag{18}$$

i.e., to the same degree of performance the angular aperture of a modulator biased to $N\pi$ is decreased by the factor $1/\sqrt{N + 1}$ of the unbiased case.

For a material with a linear electro-optic effect, it is most sensible to use $N = 0$ in order to use the lowest voltages; this will result in the maximum angular aperture. For a quadratic material like KTN, it is sometimes more efficient to use a biasing dc voltage in order to reduce the modulation voltage. In that case, an N of 10 or 20 might be used. If we decide that the maximum retardation error is $\Delta R = 0.0206\pi$, corresponding to a minimum extinction of 30 dB between polarizers, then the angular aperture of KTN is ±26° in the unbiased case, ±7.8° for $N = 10$, and ±5.7° for $N = 20$.

The angular aperture of biased KTN can be increased by placing a properly oriented positive uniaxial crystal such as quartz in series with the biased KTN modulator. This technique can be used to eliminate the terms in γ^2 from the total retardation and thereby increase the angular aperture to approximately that of the unbiased case.

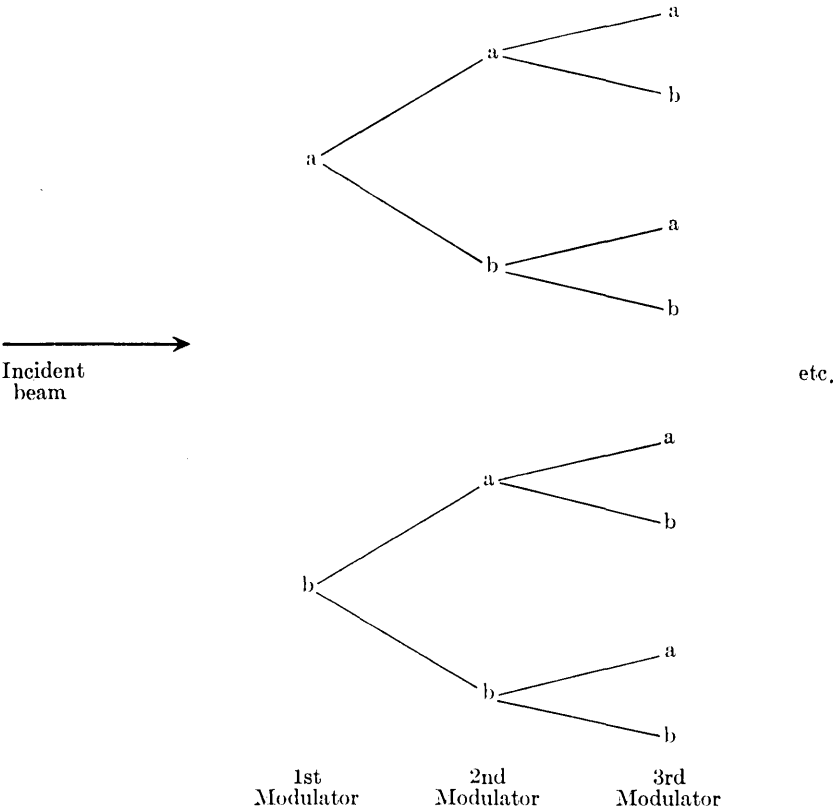
APPENDIX C

Intensity Distribution in a DLD

Assume that we have a DLD consisting of n stages made up of modulators with an efficiency, $E = a/b$, $a + b = 1$. We again assume that only the modulators are imperfect, and that every modulator can be characterized by the same efficiency.

Any light beam incident to a modulator is broken up into two beams, an "a" beam for the desired polarization and a "b" beam for the oppositely polarized position. A table can be made up which lists the total number of paths through the DLD (Table II). Since we have a choice at each modulator as to whether an a or b path is taken, the different paths are characterized by all possible combinations of the

TABLE II



a and b terms. Let us consider the paths containing an r number of b terms and an $(n - r)$ number of a terms. The total number of possible ways of grouping these terms is given by

$$\frac{n!}{(n - r)!\,r!} \tag{19}$$

and the associated intensity of these beams is $a^{(n-r)}b^r$. The total number of all paths is then given by summing r through its range

$$\sum_{r=0}^{n} \frac{n!}{(n - r)!\,r!}\, a^{(n-r)} b^r \tag{20}$$

where the intensity terms have been included. These terms can also be generated by

$$(a + b)^n \tag{21}$$

since the terms $n!/(n - r)!r!$ are also the coefficients of the binomial expansion.

From the definition of E, (21) can also be written

$$a^n\left(1 + \frac{1}{E}\right)^n. \tag{22}$$

To derive (22), the DLD was set for one address and the intensity at each point was determined. We now ask what are the different intensities that arrive at a particular position when the DLD is addressed to all possible positions.

In order to address the DLD to every position, the state of the modulators can be arranged according to Table III. In this table, 0 means no change in the state of polarization and 1 refers to a change

TABLE III

Modulator number			
etc.	$(n - 2)$	$(n - 1)$	n
	0	0	0
	0	0	1
	0	1	0
	0	1	1
	1	0	0
	1	0	1
	1	1	0
	1	1	1
	etc.	etc.	etc.

TABLE IV

Address of main beam and setting of DLD, A_0	00101
Address of the position at which we wish to compare intensity, A_n	00010
Intensity division at each modulator between A_0 and A_n	$1 \;\; 1 \;\; (b/a) \; (b/a) \; (b/a)$
Intensity at A_0	a^5
Intensity at A_n	a^2b^3

to the opposite sense of polarization. Table III is a partial listing. The complete table is made by first writing the nth column which consists of alternating 0's and 1's for a total number of 2^n entries; the $(n - 1)$th column is written by entering pairs (2^1) of the 0's and 1's for a total of 2^n; the $(n - 2)$th column by entering 2^2 of 0's and 1's, etc. The sequence of addresses in Table III would place the main beam once at each location on the focal plane. The 0's and 1's that appear in any horizontal row is the address of that beam.

We must now be able to compare intensities between that of the main beam, which we shall call A_o, and some arbitrary position, which we shall call A_n. Table IV illustrates the technique. Table IV was constructed by using a rule that sets the intensity ratio at 1 for modulators that have the same setting and b/a at modulators that have different settings.

We now wish to determine all of the intensities at some arbitrary position, say $A_n = \ldots 010$ while the DLD is addressed to all positions. Using Table III which lists all of the addresses and Table IV which illustrates the comparison rule, we can construct a table (Table V) which lists the intensity ratios at $A_n = \ldots 010$.

Table V is similar to Table III in appearance in that for any vertical column a 0, 1 in Table III is changed into 1, b/a or b/a, 1 to make Table V.

Table V is one that lists all possible combinations of the entries 1, b/a and therefore is calculable from the same general formula as that deduced for Table II, i.e.,

$$\left(1 + \frac{b}{a}\right)^n. \tag{23}$$

It is evident that Table V would not change, except for a different ordering of the horizontal rows, no matter what the particular address of A_n. Therefore, (23) is the same for all points A_n.

Equation (23) will list the intensity ratios between A_o and some point A_n. If we multiply though by the intensity of $A_o = a^n$, then (22) becomes

$$(a + b)^n \tag{24}$$

which is the desired result.

TABLE V—LIST OF INTENSITY RATIOS AT POINT ···010 WHEN DLD IS ADDRESSED TO ALL POSITIONS

Modulator number			
etc.	$(n - 2)$	$(n - 1)$	n
	1	b/a	1
	1	b/a	b/a
	1	1	1
	1	1	b/a
	b/a	b/a	1
	b/a	b/a	b/a
	b/a	1	1
	b/a	1	b/a
	etc.	etc.	etc.

APPENDIX D

*R Ratio For The Reflection Mode Of Operation**

A beam of light traveling through the DLD breaks up into 2^n exit beams due to the imperfect modulators. These intensities are given by the terms in the expansion of $(a + b)^n$ as shown in Appendix C. We now need to ask how much of the light comes back through the second aperture after being reflected from the focal plane (see Fig. 9).

Let us consider one term of the expansion of $(a + b)^n$, say $a^{n-r}b^r$. We state that this system is reciprocal and that if $a^{n-r}b^r$ of the incident beam exits the DLD, then if unity power were directed through the DLD in exactly the opposite direction the same fraction of power, i.e., $a^{n-r}b^r$, will pass through the aperture.

Thus, for an n unit DLD there will be 2^n exit terms and each of these terms, for example $a^{n-r}b^r$, will generate one term that contributes

*This appendix represents the results of calculations performed jointly by J. T. Sibilia and the author.

to the intensity at the second aperture; and for the example above that corresponds to $(a^{n-r}b^r)^2$. The total number of terms that exit through the second aperture is then the sum of the squares of the exit terms and can be generated by $(a^2 + b^2)^n$.

Before we can add up the 2^n terms in the second aperture, we must know something about the relative phases of the terms. Each of the 2^n exit terms in the expansion of $(a + b)^n$ traverses through a different optical path length in the DLD. The reason for this is because light traverses through some of the prisms as an ordinary ray and others as an extraordinary ray, and the combinations of such paths are different for each of the 2^n exit beams. Thus, unless the DLD has been specifically designed to the contrary, each path has a different phase delay in passing through the DLD.

A term in the expansion $(a^2 + b^2)^n$ such as $a^{2(n-r)}b^{2r}$ represents an E field of $[a^{2(n-r)}b^{2r}]^{\frac{1}{2}}$ and a phase factor φ. Consider the sum of all the $n!/(n - r)!r!$ terms of the type $a^{2(n-r)}b^{2r}$

$$[a^{2(n-r)}b^{2r}]^{\frac{1}{2}}(\varphi_1 + \varphi_2 + \cdots + \varphi_{n!/(n-r)!r!}). \tag{25}$$

The intensity of the sum of all $a^{2(n-r)}b^{2n}$ terms is given by the square of (25)

$$a^{2(n-r)}b^{2r}(\varphi_1 + \varphi_2 + \cdots \varphi_{n!/(n-r)!r!})^2. \tag{26}$$

A series of phase terms such as in (26) can add as follows:

$$(\varphi_1 + \varphi_2 + \cdots + \varphi_{n!/(n-r)!r!})^2 = \frac{n!}{(n-r)!\,r!} \text{ for random phases}$$

$$\left[\frac{n!}{(n-r)!\,r!}\right]^2 \text{ for the same phase.}$$

As explained earlier, all of the phases are, in general, different, and so we will use the random phase addition. The intensity at the second aperture is then the sum of all of the terms in $(a^2 + b^2)^n$.

The R ratio, the ratio of the light from the main beam, a^{2n}, to the light from the remaining positions, $(a^2 + b^2)^n - a^{2n}$, is then

$$R_3 = \frac{a^{2n}}{(a^2 + b^2)^n - a^{2n}}$$

$$= \frac{1}{\left(1 + \frac{b^2}{a^2}\right)^n - 1} = \frac{1}{n\frac{1}{E^2} + \frac{n(n-1)}{2}\frac{1}{E^4} + \cdots} \tag{27}$$

which is the same as that used in Section IV.

REFERENCES

1. Schmidt, U. J., *Optical Processing of Information,* Spartan Books, Inc., Baltimore, 1963, p. 98.
2. Nelson, T. J., Digital Light Deflection, B.S.T.J., *43,* May, 1964, p. 821.
3. Kulcke, W., Harris, T. J., Kosanke, K., and Max, E., IBM J. Res. Develop., *8,* 1964, p. 64.
4. Tabor, W. J., Use of Wollaston Prisms for a High-Capacity Digital Light Deflector, B.S.T.J., *43,* May, 1964, p. 1153.
5. Schmidt, U. J., Phys. Lett., *12,* 1964, p. 205.
6. Hauser, S. H., Smith, L. S., Marlowe, D. G., and Yoder, P. R., Jr., J. Appl. Op. *2,* 1963, p. 1175.
7. Geusic, J. E., Kurtz, S. K., Van Uitert, L. G., and Wemple, S. H., Appl. Phys. Lett., *4,* 1964, p. 141.
8. Chen, F. S., Geusic, J. E., Kurtz, S. K., Skinner, J. G., and Wemple, S. H., Proc. IEEE, *52,* 1964, p. 1258.
9. Born, M. and Wolf, E., *Principles of Optics,* Pergamon Press, 1959, p. 394.
10. Cagnet, M., Francon, M., and Thrierr, J. C., *Atlas of Optical Phenomena,* Prentice-Hall, 1962.
11. Billings, B. H., J. Opt. Soc. Am., *39,* 1949, p. 802.
12. Wiley, C. A., U. S. Patent No. 2,780,958.
13. Sterzer, F., Blattner, D., Miniter, S. J., Opt. Soc. Am., *54,* 1963, p. 62.
14. Lansraux, G., Rev. d'Optique, *32,* 1953, p. 475.
15. Born, M. and Wolf, E., *Principles of Optics,* Pergamon Press, 1959, p. 467.
16. Kulcke, W., Kosanke, K., Max, E., and Fleisher, H., Use of Optical Masers in Displays and Printers, Third Quarterly Report, February 24 through May 23, 1964, Contract No. DA36-039-AMC-00118(E).
17. Lee, R. K., Jr., Moskowitz, F., J. Appl. Op., *3,* 1964, p. 1305.
18. Sibilia, J. T., Tabor, W. J., unpublished work.
19. Skinner, J. G., Increasing the Memory Capacity of the Digital Light Deflector by "Color Coding," B.S.T.J., *45,* April, 1966, pp. 597–608.
20. Seidel, H., unpublished work.
21. Bowers, K. D. and Varnerin, L. J., unpublished work.
22. Chang, J. T., Dillon, J. F., Jr., and Gianola, U. F., J. Appl. Phys. *36,* 1965, p. 1110.
23. Kurtz, S. K., Design of an Electro-Optic Polarization Switch for a High-Capacity High-Speed Digital Light Deflection System, B.S.T.J., *45,* October, 1966, pp. 1209–1246.
24. Jenkins, F. A. and White, H. E., *Fundamentals of Optics,* McGraw-Hill Book Co. Inc., 1957, Third Ed., Fig. 27M(c), p. 569.

Transistor Distortion Analysis Using Volterra Series Representation

By S. NARAYANAN

(Manuscript received January 3, 1967)

Intermodulation distortion due to nonlinear elements in transistors is analyzed using Volterra series representation. It is shown that this technique is well suited for the analysis of transistor distortion where the nonlinearities are small but frequency dependent. An ac transistor model incorporating four nonlinearities is briefly described. The nonlinear nodal equations of the model are successively solved by expressing nodal voltages in terms of the Volterra series expansion of the input voltage. Based on this analysis, a digital computer program has been developed which computes the second and the third harmonic distortion for a given set of input frequencies and transistor parameters. The results compare favorably with measured values. This method also enables the derivation of closed form ac expressions for a simplified model; these expressions show the dependence of distortion on frequency, load and source impedances, bias currents and voltages, and the parameters of the transistor. The technique is also extended to cascaded transistors, and simplified expressions for the overall distortion in terms of the distortion and gain of individual transistors are derived. Finally, a few pertinent practical applications are discussed.

I. INTRODUCTION

Solid-state long-haul analog communication systems are being designed for higher frequencies to meet the growth in demand. One of the more critical and significant problems facing the system designer is intermodulation noise arising from transistor nonlinearities. Thus, an analysis of transistor distortion at higher frequencies is a practical problem; this paper investigates the transistor distortion using the Volterra series as an analysis tool.

Transistor distortion has been investigated in some detail previously. Many authors have considered the exponential nonlinear relation between emitter current and emitter-to-base voltage which is important at low currents.[1,2,3,4,5] The effect of frequency on this nonlinear source alone has been reported.[5] Three nonlinearities (exponential, avalanche, and h_{FE} at dc) have been examined by Riva, Beneteau and Dallavolta.[6] For currents up to 20 mA and frequencies up to 100 kHz, Meyer[7,8,9] has developed a more accurate and complex model obtaining the nonlinearities from h-parameters. However, he takes into account the frequency dependence by assuming that the h-parameters can be written as $h' + j\omega h''$. Moreover, he does not take into account avalanche distortion, nor has he extended the model to higher currents (100 mA) and frequencies (20 MHz). The model described here considers four nonlinearities; they are, exponential, avalanche, h_{FE}, and collector capacitance nonlinearities. These nonlinearities are superimposed on a linear ac equivalent circuit.[10,11] Much of the initial development of the model with three nonlinearities was done by Thomas.[10]

The transistor model is analyzed using a Volterra series representation; this series is a generalization of the power series. In a now classic report, Wiener applied this analysis technique[12] to find the response of a nonlinear device to noise.[13] Bose has carried the theory further.[14] Following a series of lectures by Wiener,[15] the theoretical framework, higher-dimensional transforms, and optimization with Gaussian inputs were considered by Brilliant,[16] George,[17] and Chesler,[18] respectively. Barrett[19] has treated statistical inputs. The synthesis problem has been examined by Van Trees,[20] who also applied the method to phase-locked loops.[21] The technique has been extended to discrete systems,[22,23,24] and a class of time-variant systems.[24,25] More recently the theory of the convergence of the series has been treated.[26] This work relies more on George's work on the higher-dimensional transform theory.[17]

Even though much work has been done in this area, the Volterra series has not found a wide application in solving nonlinear system problems due to several reasons; if the rate of convergence is not rapid, the higher-degree terms, which are cumbersome to handle, cannot be neglected; hence, it cannot conveniently represent gross nonlinearities. It is not simple to invert the multidimensional transforms to the time domain, and it is not a useful technique to determine the stability of a nonlinear differential equation.

The Volterra series method does, however, offer certain distinct advantages in analyzing transistor distortion. Since transistor distortion is frequency dependent, the power series is inadequate to characterize it; the Volterra series does indeed represent frequency dependent systems. The nonlinearities in the transistors under consideration are extremely small so that the second- and third-degree terms suffice to characterize them. Since the output corresponding to sinusoidal input signals is of interest, there is no need to find the inverse of the higher-dimensional transforms; the output can be expressed in terms of the transform of the kernel. The higher-dimensional transforms of the kernel are complex numbers when $s_i = j\omega_i$, where s_i is the complex variable in the transform domain; hence, these kernels can be numerically evaluated using the computer (see Section IV). Moreover, for a slightly simpler model closed form ac expressions can be derived. Since the kernels retain phase information, this approach will be useful for the AM-to-PM conversion problem at IF frequencies. Finally, in an amplifier two or more transistors are cascaded; the nonlinear behavior of such cascaded transistors is a significant problem. The Volterra series approach can be easily extended to study such cascaded transistors.

II. AN INTRODUCTION TO VOLTERRA SERIES REPRESENTATION

A brief exposition of Volterra series with pertinent reference to the problem under consideration is presented below. For further details the reader is referred to the references cited.

Consider a simple memoryless nonlinear system described by the following power series; let $y(t)$ be the output and $x(t)$ the input; the system is represented by

$$y(t) = c_1 x(t) + c_2[x(t)]^2 + c_3[x(t)]^3, \tag{1}$$

where c_1, c_2, c_3 are constants. For a time-invariant system with memory (capacitors and inductors in an electrical network), the linear term $\{c_1 x(t)\}$ is replaced by the convolution integral ($x(t) = 0; t < 0$)

$$y_1(t) = \int_0^t c_1(t - \tau)x(\tau)\, d\tau. \tag{2}$$

In the transform domain, (2) may be written

$$Y_1(s) = C_1(s)X(s). \tag{3}$$

This transform domain representation of the system $[C_1(s)]$ has been an invaluable aid to the communication engineers since it brings into focus the frequency behavior of the system.

A generalization of the second-degree term, $c_2[x(t)]^2$, is the double convolution integral

$$y_2(t) = \int_0^t \int_0^t c_2(t - \tau_1 , t - \tau_2) \prod_{i=1}^{2} x(\tau_i)\, d\tau_i . \tag{4}$$

The output depends on the past values of the input; the above expression involves a product of the input with itself, thus representing a quadratic system. $c_2(t - \tau_i , t - \tau_2)$ is known as the second-degree Volterra kernel.

A two-dimensional Laplace transform can be defined for (4) after introducing dummy variables t_1 and t_2. As shown in Appendix A, (4) becomes

$$Y_2(s_1 , s_2) = C_2(s_1 , s_2) \prod_{i=1}^{2} X(s_i). \tag{5}$$

When two sinusoidal signals at frequencies f_a and f_b are applied (Appendix A), the output at the harmonic frequency $f_a \pm f_b$ is given by $[| C_2(f_a \pm f_b) | \cos (2\pi(f_a \pm f_b)t + \phi_{a\pm b})]$. Since in general $C_2(f_a , f_b)$ will not be equal to $C_2(f_a , -f_b)$, different values of distortion at different harmonic frequencies are directly reflected in the kernel. Moreover, as in the power series case, the $2f$ product is less by a factor of two.

Likewise, the third-degree term $[c_3(x(\tau))^3]$ can be generalized to a triple convolution integral;

$$y_3(t) = \int_0^t \int_0^t \int_0^t C_3(t - \tau_1 , t - \tau_2 , t - \tau_3) \prod_{i=1}^{3} x(\tau_i)\, d\tau_i . \tag{6}$$

In the transform domain (6) may be written

$$Y_3(s_1 , s_2 , s_3) = C_3(s_1 , s_2 , s_3) \prod_{i=1}^{3} X(s_i). \tag{7}$$

The magnitude of the signal at the harmonic frequency $f_a + f_b - f_c$ due to the three fundamental signals at f_a, f_b and f_c is given by $| C_3(f_a , f_b , -f_c) |$. The constants like 1/4 for a '$3f_a$' product are the same as obtained from the power series approach.

Later in the paper (in Section IV) the cascade relations in the transform domain are frequently used; their physical significance is discussed in detail in Section VI. (See also Fig. 1.) The cascade formulae and the procedure for deriving them are given in Appendix A.

The second and third harmonic distortion are defined as the second and third harmonic power in dBm, respectively, when the fundamental power at the output of the transistor is at zero dBm (one milliwatt). In the analysis of the model in Section IV, the output voltage is expressed in terms of a Volterra series of the input voltage. Thus, the kernels $C_1(s_1)$, $C_2(s_1 , s_2)$, and $C_3(s_1 , s_2 , s_3)$ are the voltage transfer ratios; for a given load R_L, the second and the third harmonic distortion in dBm are given by the following expressions:

$$M_{2_{a\pm b}} = 20 \log \frac{1}{2} \frac{|\,C_2(f_a \pm f_b)\,|\,\sqrt{10^{-3}R_L}}{|\,C_1(f_a)\,|\,|\,C_1(\pm f_b)\,|} \qquad (8)^*$$

$$M_{3_{a\pm b\pm c}} = 20 \log \frac{1}{4} \frac{|\,C_3(f_a \pm f_b \pm f_c)\,|\,10^{-3}R_L}{|\,C_1(f_a)\,|\,|\,C_1(\pm f_b)\,|\,|\,C_1(\pm f_c)\,|}. \qquad (9)^*$$

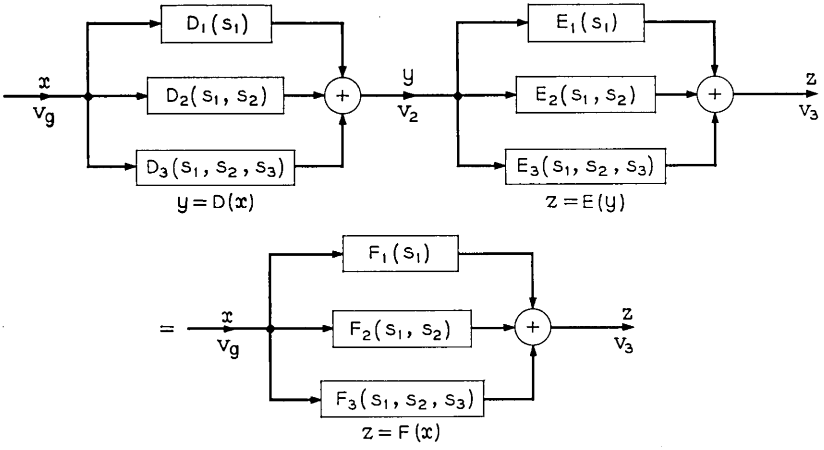


Fig. 1—Two cascaded systems.

III. THE JUNCTION TRANSISTOR NONLINEAR MODEL

A model is a simple but realistic representation of a physical phenomenon in terms of measurable parameters such that the phenomenon can be analyzed, and controlled if possible. The linear equivalent circuit of a transistor is one such example. In reality, several elements of the transistor equivalent circuit are not linear but are linearized versions of nonlinear functions; they are the first-degree terms of the Taylor's series expansion of the nonlinear functions. Hence, a logical way to develop the nonlinear model is to consider the second- and third-degree terms of the Taylor's series expansion; thus, the emitter resistance (exponential nonlinearity), current gain (h_{FE} and avalanche nonlinearity), and the collector capacitance (collector capacitance nonlinearity) have been represented by nonlinear voltage dependent current generators whose parameters are higher-degree Taylor's series terms. This approach has another advantage in that it is difficult to measure the nonlinearities since they are small; but, it is not too difficult to measure the overall functions and to curve fit with the known theoretical

* The factors $\frac{1}{2}$ and $\frac{1}{4}$ normalize the distortion to $2f$ and $3f$ products.

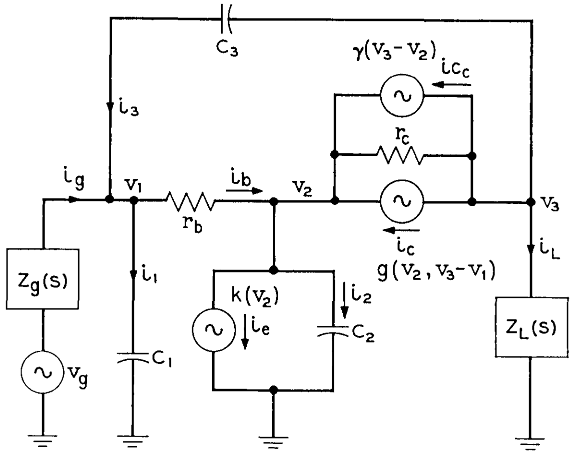


Fig. 2 — Common-emitter nonlinear equivalent circuit.

or empirical relations. These nonlinearities are superimposed on the linear equivalent circuit (Fig. 2). The nonlinearities are described next.

3.1 *Exponential Nonlinearity*

The emitter current, I_E, is related to the emitter voltage, V_2, by the exponential relation

$$I_E = A\left[\exp\left(q\frac{V_2}{KT}\right) - 1\right] + B, \tag{10}$$

where K = Boltzmann's constant,

q = electron charge in coulombs,

T = Temperature in degrees Kelvin,

and A and B are constants which depend on the transistor parameters (Ref. 27; p. 181, p. 249). An experimental curve of the emitter current I_E and the emitter-to-base voltage V_{be} is shown in Fig. 3. This nonlinearity is expressed as a voltage-dependent current generator by a Taylor's series expansion of (10) as follows:

$$i_e = K(v_2) = K_1v_2 + K_2v_2^2 + K_3v_2^3, \tag{11}$$

where the Taylor's series coefficients are derived in terms of known parameters, the emitter resistance r_{e_1} and the emitter bias current I_E; i.e.,

$$K_1 = \frac{1}{r_{e_1}}\,; \quad K_2 = \frac{1}{2(r_{e_1})^2}\frac{1}{I_E}\,; \quad K_3 = \frac{1}{6}\frac{1}{(I_E)^2}\frac{1}{(r_{e_1})^3}. \tag{12}$$

3.2 *Avalanche and h_{FE} Nonlinearity*

The collector current is a nonlinear function of the emitter current at higher values of current (h_{FE} nonlinearity) and of the collector-to-base voltage at higher values of voltage (avalanche nonlinearity).[27] h_{FE}, the ratio of I_C to I_B, is plotted as a function of collector current I_C in Fig. 4. It is seen that the following empirical relation[6] matches the experimental result (Fig. 4):

$$h_{FE} = \frac{h_{FE\,\max}}{1 + a \log^2 \dfrac{I_C}{I_{C\,\max}}}, \tag{13}$$

where $h_{FE\,\max}$ is the maximum value of h_{FE}, $I_{C\,\max}$ is the value of I_C at which $h_{FE\,\max}$ occurs, and a is a constant.

The avalanche nonlinearity is due to avalanche multiplication which occurs at higher collector-to-base voltage. It is determined from the collector characteristic which is a plot of collector current (I_C) and collector-to-emitter voltage (V_{CE}), (Fig. 5). The empirical Miller's avalanche multiplication factor is given by

$$\frac{1}{1 - \left(\dfrac{V_{CE}}{V_{CEO}}\right)^n}, \tag{14}$$

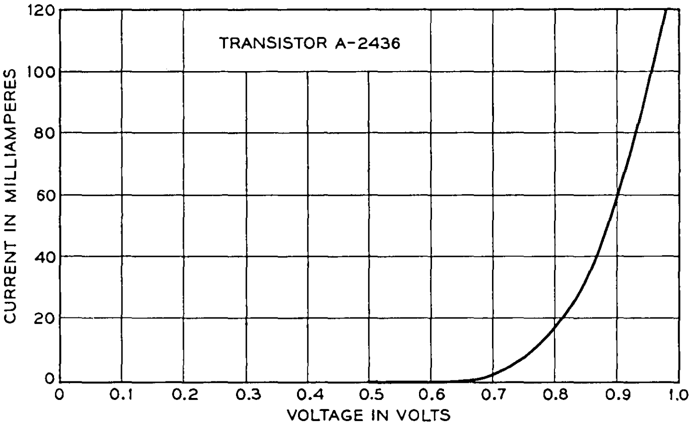


Fig. 3 — Exponential nonlinearity — measured curve.

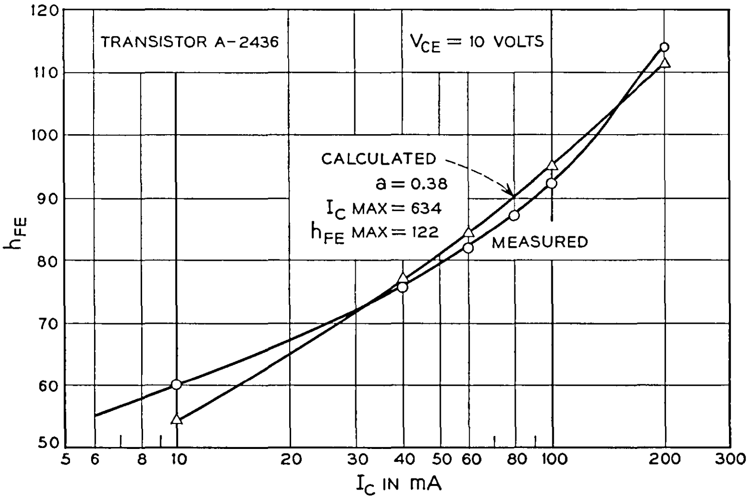


Fig. 4—h_{FE} nonlinearity — calculated and measured curves.

where V_{CEO} is the sustained voltage, and the exponent n is determined by experiment. From expressions (13) and (14), the ratio $I_C v_S I_E$ is given by

$$\frac{I_C}{I_E} = \frac{h_{FE\,\max}}{1 + h_{FE\,\max} + a \log^2\left(\dfrac{I_C}{I_{C\,\max}}\right)} \cdot \frac{1}{1 - \left(\dfrac{V_{CB}}{V_{CBO}}\right)^n}, \qquad (15)$$

where $V_{CBO} = V_{CEO}/n\sqrt{1 - \alpha}$ and $V_{CB} \cong V_{CE}$.

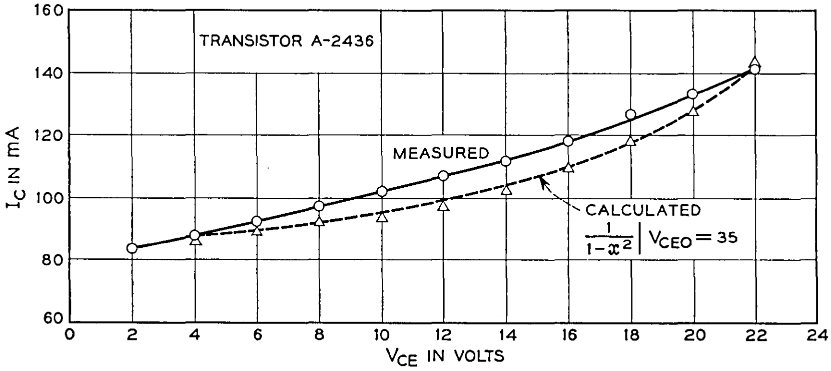


Fig. 5—Avalanche nonlinearity — calculated and measured curves.

The ac i_c can be expressed in terms of i_e and $v_{cb}[v_3 - v_1]$ by a Taylor's series expansion of (15). Since i_e is a function of emitter voltage v_2, i_c is represented by a current generator $g(v_2, v_3 - v_1)$; for convenience in notation it is separated into a linear term $g_1(v_2, v_3 - v_1)$, a second-degree term $g_2(v_2, v_3 - v_1)$ and a third-degree term $g_3(v_2, v_3 - v_1)$. The linear term equals $\hat{M}_0(\alpha_1 K_1)v_2 + \hat{M}_1(v_3 - v_1)$. The second-degree term is given by $\alpha_2 \hat{M}_0 K_1^2 (v_2)^2 + m_2(v_3 - v_1)^2 + (\alpha_1 \hat{M}_1)K_1 v_2(v_3 - v_1)$. The coefficients α_1, α_2, $\hat{M}_1$, m_2, etc., and the third-degree term are given in Appendix B.

3.3 *Collector Capacitance Nonlinearity*

The collector capacitance is a nonlinear function of collector-to-base voltage (V_{CB}) since the depletion layer width is a function of V_{CB}. The exact functional relationship is determined by plotting the common-base imaginary part of h_{22} as a function of collector-to-base voltage (V_{CB}) as shown in Fig. 6.[9] It is evident from the figure that C_c follows the 1/3 voltage law (Ref. 19; Equation 5-96);

$$C_c = k(V_{CB})^{-1/3}. \tag{16}$$

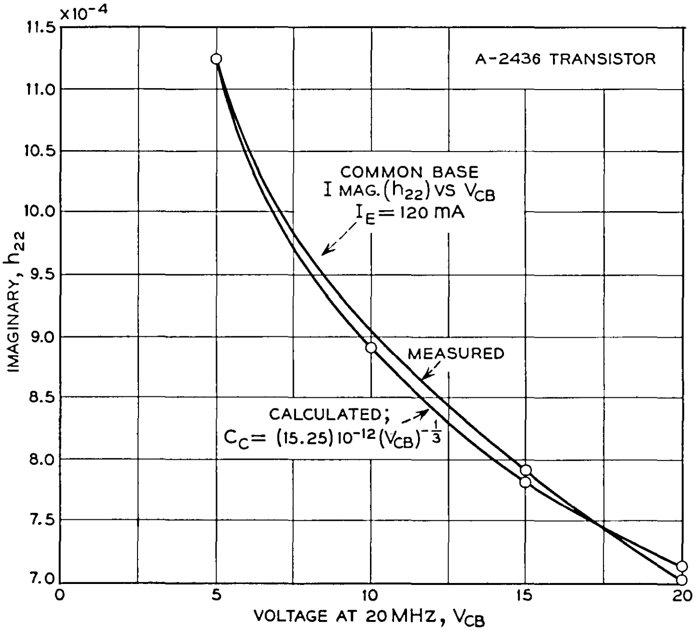


Fig. 6 — Collector capacitance nonlinearity — calculated and measured curves.

This nonlinearity is represented as a frequency (differentiation) and voltage-dependent current generator as follows:

$$\begin{aligned} i_{c_c} &= \gamma(v_3 - v_1) \\ &= \gamma_1 \frac{d}{dt}(v_3 - v_1) + \gamma_2 \frac{d}{dt}(v_3 - v_1)^2 + \gamma_3 \frac{d}{dt}(v_3 - v_1)^3, \end{aligned} \tag{17}$$

where $\gamma_1 = C_c$, and where γ_2 and γ_3 are known from (16).

The above nonlinear current generators are incorporated in the linear equivalent circuit as shown in Fig. 2. The linear equivalent circuit parameters are obtained from the equivalent circuit characterization. They can, for example, be computed from the h-parameters at different frequencies. In general, the distortion is not a critical function of the linear parameters. (Figs. 14 to 17).

All the nonlinear coefficients (K_2, α_2, m_2, etc.) are easily obtained from a simple computer program. The parameters to be specified along with typical values for transistor type A-2436 are listed in Appendix C.

IV. THE VOLTERRA KERNELS FOR THE NONLINEAR MODEL

The Volterra series method is applied to the model to compute the second and the third harmonic distortion. The voltage at each node is a nonlinear frequency-dependent function of the input voltage. Each nodal voltage is expressed by a Volterra series expansion of the generator voltage; since the nonlinearities are small only three terms are considered. The kernels at each node are determined from Kirchoff's current equations.

4.1 *Nodal Equations*

The Kirchoff's current law is applied at each node; the currents are next expressed in terms of the generator voltage v_g, the three nodal voltages v_1, v_2, and v_3, and the known linear and nonlinear parameters. The impedances are represented by their transforms and o denotes that it operates on the voltage across it. The nodal equations are given below.

$$\frac{1}{Z_g(s)}\, o\,(v_g - v_1) + (sC_3)\, o\,(v_3 - v_1) = (sC_1)\, o\, v_1 + \frac{1}{r_b}\, o\,(v_1 - v_2), \tag{18}$$

$$\begin{aligned} \frac{1}{r_b}\, o\,(v_1 - v_2) = K(v_2) + (sC_2)\, o\, v_2 + \left(\frac{1}{r_c}\right) o\,(v_2 - v_3) - \gamma(v_3 - v_2) \\ - g(v_2, v_3 - v_1), \end{aligned} \tag{19}$$

$$-\gamma(v_3 - v_2) + \frac{1}{r_c}(v_2 - v_3) - g(v_2\,, v_3 - v_1) = (sC_3)\ o\ (v_3 - v_1) + \left(\frac{1}{Z_L(s)}\right) o\ v_3\ , \tag{20}$$

where $K(v_2)$, $\gamma(v_2 - v_3)$ and $g(v_2\,, v_3 - v_1)$ are the nonlinear current generators.

4.2 *Solution Using Volterra Series*

Since each nodal voltage is to be expressed in terms of three Volterra kernels, there are nine unknown Volterra kernels to be determined from the three equations. The problem of solving for nine unknowns from three equations is resolved by noting that the polynomials x, x^2 and x^3 are linearly independent; hence, each degree term is separately and successively solved. The linear kernels are first determined; then the second-degree kernels are determined in terms of the linear kernels; lastly, the third-degree kernels are evaluated in terms of the first- and second-degree kernels.

Let $A_1(s)$, $B_1(s)$, $C_1(s)$ denote the transforms of the linear kernels at nodes one, two and three, respectively. From the nodal equations (18) to (20), the following vector matrix equation is derived.

$$\begin{Bmatrix} \dfrac{1}{Z_g(s)} \\ 0 \\ 0 \end{Bmatrix} = P_E(s) \begin{Bmatrix} A_1(s) \\ B_1(s) \\ C_1(s) \end{Bmatrix}, \tag{21}$$

where

$$P_E(s) = \begin{bmatrix} \dfrac{1}{Z_g(s)} + s(C_3 + C_1) + \dfrac{1}{r_b} & -\dfrac{1}{r_b} & -sC_3 \\ -\dfrac{1}{r_b} + m_1 & \dfrac{1}{r_b} + sC_2 + \dfrac{1}{r_c} + K_1(1 - \alpha) + s\gamma_1 & -\dfrac{1}{r_c} - m_1 - s\gamma_1 \\ -\ sC_3\ -\ m_1 & -\dfrac{1}{r_c} + \alpha_1 K_1 - s\gamma_1 & \dfrac{1}{r_c} + sC_3 + \dfrac{1}{Z_L(s)} + m_1 + s\gamma_1 \end{bmatrix}. \tag{22}$$

Equation (21) is solved by inverting matrix $P_E(s)$ and post-multiplying by the vector

$$\begin{Bmatrix} \dfrac{1}{Z_g(s)} \\ 0 \\ 0 \end{Bmatrix}.$$

For a given frequency $s = j\omega$, the computation is done numerically.

The second-degree terms are equated next in (18) to (20). There are two types of second-degree terms; those arising from the unknown second-degree kernels [for example, $(s_1 + s_2)C_1A_2(s_1, s_2)$] and those arising from the known nonlinear coefficients and the known linear kernels [for example, $K_2 \Pi_{i=1}^2 B_1(s_i)$]. The terms associated with the unknown second-degree kernels are the same as were associated with the unknown linear kernels in (21), but at the harmonic frequency $(s_1 + s_2)$. The following vector matrix equation is obtained for the second-degree kernels:

$$\left\{\begin{array}{l} \qquad\qquad o \\ [\hat{g}_2(B_1, C_1 - A_1) \\ +\gamma_2(C_1 - B_1) - K_2(B_1)] \\ [-\hat{g}_2(B_1, C_1 - A_1) \\ \qquad\qquad -\hat{\gamma}_2(C_1 - B_1)] \end{array}\right\} = P_E(s_1 + s_2) \left\{\begin{array}{c} A_2(s_1, s_2) \\ B_2(s_1, s_2) \\ \\ C_2(s_1, s_2) \\ \end{array}\right\}, \tag{23}$$

where $\hat{g}_2$ and $\hat{\gamma}_2$ represent the second harmonic contribution due to $g_2(v_2, v_3 - v_1)$ and $\gamma_2(v_3 - v_2)$; hence,

$$\begin{aligned} \hat{g}_2(B_1, C_1 - A_1) = [\alpha_1 K_2 + \alpha_2 K_1^2] \prod_{i=1}^{2} B_1(s_i) & \\ + \frac{\alpha_1 M_1 K_1}{2} [B_1(s_1)[C_1(s_2) - A_1(s_2)] & \\ + B_1(s_2)[C_1(s_1) - A_1(s_1)]] & \\ + m_2 \prod_{i=1}^{2} [C_1(s_i) - A_1(s_i)] & \end{aligned} \tag{24}$$

$$\hat{K}_2(B_1) = K_2 \prod_{i=1}^{2} B_1(s_i) \tag{25}$$

$$\hat{\gamma}_2(C_1 - B_1) = K_2(s_1 + s_2)\gamma_2 \prod_{i=1}^{2} [C_1(s_i) - B_1(s_i)]. \tag{26}$$

$P_E(s_1 + s_2)$ is the matrix $P_E(s)$ with s replaced by $(s_1 + s_2)$.

The vector on the left side of (23) is known. Thus, the unknown kernels are determined by inverting the matrix $P_E(s_1 + s_2)$ and post-multiplying by the vector on the left-hand side of (23). When $s_1 = j\omega_b$, the inversion of the matrix and the post multiplication by the vector can be done numerically.

The procedure for obtaining the third-degree kernels is almost the same; the significant difference is that the vector on the left side not only contains terms arising from the third-degree nonlinear parameters but also includes second-degree coefficients which give rise to third-degree terms by the interaction of the first- and the second-degree kernels. These interaction terms are denoted by $\hat{K}_{23}$, $\hat{g}_{23}$, $\hat{\gamma}_{23}$, respectively. For example, $K_3(B_1) = K_3 \Pi_{i=1}^{3} B_1(s_i)$, whereas $K_{23} = 2k_2B_1(s_1) B_2(s_2 , s_3)$ which shows the interaction of the first- and the second-degree kernels. The third-degree kernels are derived from the following equations:

$$\left\{\begin{matrix} o \\ [\hat{g}_3(B_1 , C_1 - A_1) + \hat{\gamma}_3(C_1 - B_1) \\ \qquad - \hat{K}_3(B_1) + \hat{g}_{23} + \hat{\gamma}_{23} + \hat{K}_{23}] \\ [-\hat{g}_3(B_1 , C_1 - A_1) - \hat{\gamma}_3(C_1 - B_1) \\ \qquad - \hat{g}_{23} - \hat{\gamma}_{23}] \end{matrix}\right\}$$

$$= P_E(s_1 + s_2 + s_3) \begin{Bmatrix} A_3(s_1 , s_2 , s_3) \\ B_3(s_1 , s_2 , s_3) \\ C_3(s_1 , s_2 , s_3) \end{Bmatrix} , \qquad (27)$$

where $\hat{g}_3$, $\hat{g}_{23}$ are given in Appendix B.

A computer program has been developed which calculates the kernels and the second and the third harmonic distortion. It uses existing programs to invert the matrix $P_E(s)$. The nonlinear coefficients are computed from the known and measured parameters. Computed and measured results at different currents are given in Fig. 7. The program has been extended to common-base and common-collector configurations.

V. SIMPLIFIED DISTORTION EXPRESSIONS, THEIR PHYSICAL SIGNIFICANCE AND COMPARISON WITH EXPERIMENTAL RESULTS

Another advantage of the Volterra series method is that it permits derivation of closed-form expressions for second and third harmonic distortion. These equations show the interaction between the various nonlinear parameters and the effect of frequency.

The model includes the base resistance (r_b), the emitter resistance (r_{e_1}), the diffusion capacitance (C_2), the load (R_L) and the source impedances $Zg(s)$, and three nonlinearities, namely, exponential, avalanche, and h_{FE} nonlinearities. In the computer program C_{bc} , C_{be} ,

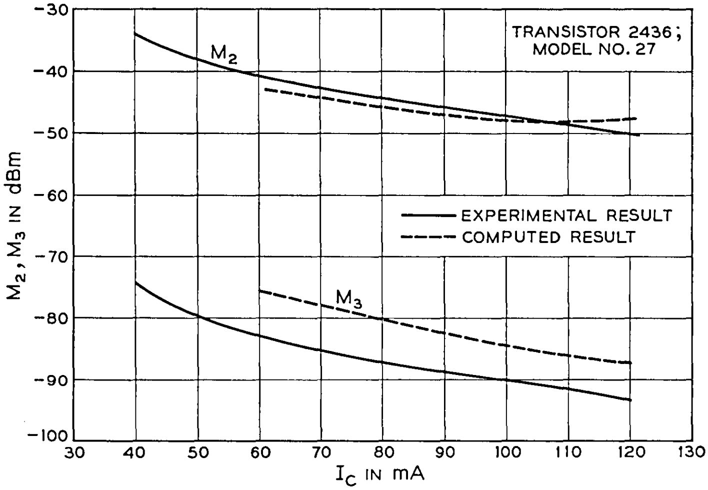


Fig. 7 — Comparison of experimental and computed results.

r_c, C_c, m_1 and collector capacitance nonlinearity have been taken into account. The expressions given below are for the common-emitter configuration.

5.1 *The Second Harmonic Distortion Term*

The second harmonic distortion in dBm (8) is given by

$$M_{2_{a\pm b}} \approx 20 \log \frac{1}{2}\sqrt{\frac{10^{-3}}{R_L}} \left| \left[\frac{(r_b + Z_g(\underline{s}))\cdot(K_1 + \underline{s}C_2) + 1}{(r_b + Z_g(\underline{s}))\cdot[K_1(1-\alpha_1) + \underline{s}C_2] + 1}\right] \right.$$

$$\cdot\left[\frac{\alpha_2}{(\alpha_1)^2} - \hat{M}_1\left(R_L + \frac{\underline{s}C_2}{2}\frac{r_b}{\alpha_1 K_1}\right) + m_2 \prod_{i=1}^{2}\left(R_L + \frac{s_i C_2 r_b}{\alpha_1 K_1}\right)\right.$$

$$\left.\left. + \frac{K_2}{\alpha_1 K_1^2}\left(\frac{(Z_g(\underline{s}) + r_b)\cdot\underline{s}C_2 + 1}{(Z_g(\underline{s}) + r_b)\cdot(K_1 + \underline{s}C_2) + 1}\right)\right]\right| \tag{28}$$

where $s_1 = j\omega_a$, $s_2 = \pm j\omega_b$ and $\underline{s} = j\omega_a \pm j\omega_b$.

5.2 *The Third Harmonic Distortion Term*

In the third harmonic distortion term given below, the interaction terms due to the first- and the second-degree kernels have not been included mainly to reduce the complexity; in certain cases, they may be significant.

$$M_{3_{a\pm b\pm c}} \approx 20 \log \frac{1}{4}\frac{10^{-3}}{R_L} \left| \left[\frac{(r_b + Z_g(\underline{s}))\cdot(\underline{s}C_2 + K_1) + 1}{(r_b + Z_g(\underline{s}))\cdot(K_1(1-\alpha_1) + \underline{s}C_2) + 1}\right] \right.$$

$$\cdot\left[\frac{\alpha_3}{(\alpha_1)^3} + \hat{M}_1\left(R_L^2 + \frac{2}{3}\frac{R_L\underline{s}C_2r_b}{\alpha_1K_1} + \frac{1}{3}\frac{\overline{s_is_j}C_2^2r_b^2}{(\alpha_1K_1)^2}\right) - m_3\prod_{i=1}^{3}\left(R_L + \frac{s_iC_2r_b}{\alpha_1K_1}\right)\right.$$
$$+ \frac{K_3}{(\alpha_1K_1)^3}\frac{\alpha_1[(r_b + Z_g(\underline{s}))(\underline{s}C_2) + 1]}{(r_b + Z_g(\underline{s}))\cdot(\underline{s}C_2 + K_1) + 1} + \frac{2\alpha_2K_2}{(\alpha_1)^3K_1^2}$$
$$\left. - \frac{\alpha_2\hat{M}_1}{(\alpha_1)^3}\left[\alpha_1R_L + \frac{\underline{s}}{3}\frac{C_2r_b}{(K_1)^3}\right]\right]\Bigg|, \tag{29}$$

where $s_1 = j\omega_a$, $s_2 = j\omega_b$, $s_3 = \pm j\omega_c$ and $\underline{s} = s_1 + s_2 + s_3$ and $\overline{s_is_j} = s_1s_2 + s_2s_3 + s_3s_1$.

5.3 *Physical Interpretation of the Distortion Terms*

The interaction of different nonlinearities and their dependence on load impedance, source impedance, bias currents, bias voltage and frequency is indeed somewhat complex. However, the closed form expressions derived above give a general qualitative picture which will be discussed now.

5.3.1 *Effect of Frequency*

It is important to know the effect of frequency on distortion. The distortion depends not only on the frequencies of the fundamental tones but also on the harmonic frequency of interest. As shown in Fig. 8,

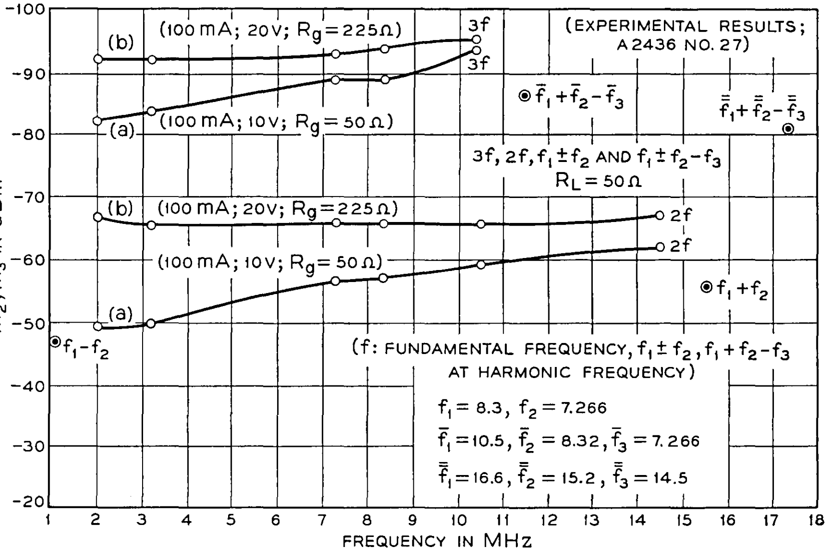


Fig. 8 — Variation of M_2, M_3 with frequency.

M_2 due to $a + b$ product is better than M_2 of $a - b$ product by 10 dB with the two tones at 8.32 and 7.266 MHz. These measurements were made with the transistor biased at 100 mA, 10V and with $R_L = 50\Omega$ and $R_g = 50\Omega$. In curve (a) of Fig. 8, the fundamental tone was increased from 2 MHz to 10.5 MHz and signals at $2f$ and $3f$ were measured. It is seen that both M_2 and M_3 improved with increase in frequency. A theoretical explanation on the basis of dominant terms (in this range of parameter values) in (28) and (29) is given below. In (28) as well as in (29) the terms in brackets are multiplied by a frequency-dependent term

$$\frac{(r_b + R_g)(K_1 + \underline{s}C_2) + 1}{(r_b + R_g)(K_1(1 - \alpha_1) + \underline{s}C_2) + 1}.$$

In this range of frequency ($\underline{s}$ = harmonic frequency), if $K_1(1 - \alpha_1) \leqq |\underline{s}C_2|$ and if $|(r_b + R_g)\underline{s}C_2| > 1$ but $K_1 > |(\underline{s}C_2)|$, then the above term reduces to $K_1/\underline{s}C_2$ which decreases with increase in frequency. However, the avalanche terms ($\hat{M}_2$, $\hat{M}_3$, etc.) involve the terms $\underline{s}C_2$ [in (28) and (29)] and S_iC_2 in the numerator. Thus, if the avalanche terms are dominant, as at higher voltages, there should be no net contribution due to avalanche terms alone. The exponential terms $[K_2/(K_1)^2$ and $K_3/(K_1)^3]$ are multiplied by the factor

$$\frac{(r_b + R_g)\cdot\underline{s}C_2 + 1}{(r_b + R_g)\cdot(K_1(1 - \alpha_1) + \underline{s}C_2) + 1}.$$

This term is independent of frequency if $(\underline{s}C_2(r_b + R_g) + 1) > 1$. Thus the above discussion shows that distortion will improve with increase in frequency at lower voltages and if $|\underline{s}C_2(r_b + R_g) + 1| < 1$. To verify this statement, the voltage was increased to 20 volts and the input resistance changed to 225Ω. The plots of M_2 and M_3 with frequency, as measured, are given in curves labeled (b) in Fig. 8. It is seen that M_2 and M_3 do not improve with increase in frequency. The small improvement can be attributed to the h_{FE} terms.

In general, increase in frequency increases distortion; this is especially true for the common base configuration. But as shown above, for certain ranges of frequency and certain values of source impedance, distortion can improve with frequency.

5.3.2 *Effect of Load Resistance, R_L*

The load resistance is an external parameter which the circuit designer can vary; hence, it is useful to know its effect on distortion. The second and the third harmonic terms are multiplied by $1/\sqrt{R_L}$ and $1/R_L$

terms, respectively; it shows that the distortion can be reduced by increasing R_L. However, the avalanche terms $\hat{M}_1$, $\hat{M}_2$, and $\hat{M}_3$ are multiplied by the R_L term, so that increasing R_L will increase the contribution from the avalanche terms. Thus, an increase in R_L may increase distortion or reduce it due to cancellation. (The contribution from the collector capacitance terms also increases with increase in load (R_L).) Because of the above interaction, for a given set of parameters and input frequencies and the harmonic frequency of interest there exists an optimum load R_L; this, of course, can be determined using the computer program. In Fig. 9, the measured values of M_2 and M_3 at different values of R_L are plotted; in both cases increasing R_L reduces distortion until the optimum value is reached and then distortion increases with increase in R_L.

5.3.3 *Effect of Source Impedance, $Z_g(\underline{s})$*

Source impedance is another important external parameter. The source impedance affects the exponential nonlinearities K_2/K_1^2 in (28) and K_3/K_1^3 in (29) by the factor

$$\frac{(r_b + Z_g(s))sC_2 + 1}{(r_b + Z_g(\underline{s}))\cdot[sC_2 + K_1(1 - \alpha_1)] + 1}.$$

At low frequencies, this nonlinearity is reduced by the factor $1/[(1 - \alpha)(R_g + r_b)K_1 + 1]$. Thus, an increase in R_g will reduce distortion from this source. However, the contribution from other nonlinearities are increased by

$$\frac{1 + K_1(R_g + r_b)}{1 + K_1(R_g + r_b)(1 - \alpha_1)}.$$

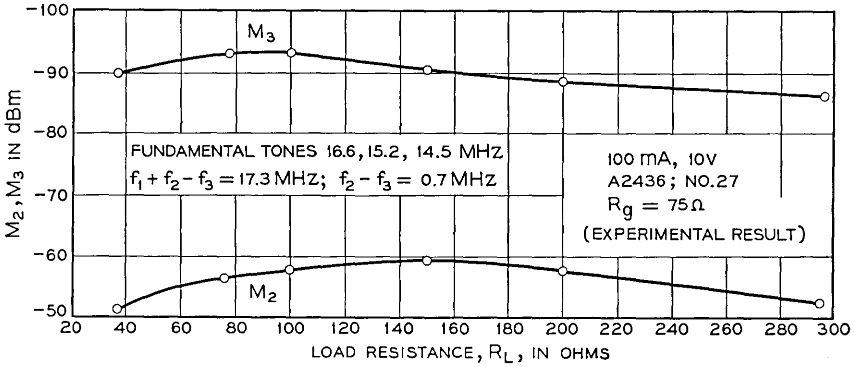


Fig. 9 — Variation of M_2, M_3 with load resistance.

It is seen from this expression that if $K_1(R_g + r_b)(1 - \alpha_1)$ is greater than 1, the other nonlinearities are not affected by the increase in R_g. Thus, if the exponential term is dominant, increasing R_g reduces distortion at low frequencies. At higher harmonic frequencies if $| \underline{s}C_2(Z_g(\underline{s}) + r_b) | > 1$, the distortion terms are independent of the source impedance since the $[r_b + Z_g(\underline{s})]$ term in the numerator and denominator cancel. This is well illustrated in the measured results of Fig. 10. The second harmonic frequency being 0.7 MHz, $| \underline{s}C_2(R_g + r_b) |$ is not much greater than one up to $R_g = 100\Omega$; hence, the second harmonic distortion improves with increase in source resistance up to 140Ω. Further increase in R_g does not cause much change in distortion. The third harmonic frequency is 17.3 MHz; hence, a change in R_g does not affect M_3 appreciably. $(| (\underline{s}C_2) \cdot (R_g + r_b) | > 1)$

5.3.4 *Effect of Bias Current*

Increase in bias current usually reduces distortion due to the following reasons. The increase in emitter bias current reduces the exponential terms

$$\frac{K_2}{(K_1)^2}\left(\alpha \frac{1}{I_E}\right) \quad \text{and} \quad \frac{K_3}{(K_1)^3}\left[\alpha \frac{1}{(I_E)^2}\right].$$

Fig. 11 shows the effect of bias current on h_{FE} terms; α_2 decreases with increase in I_C by

$$\frac{1}{I_C} \log\left(\frac{I_C e}{I_{C\,\max}}\right);$$

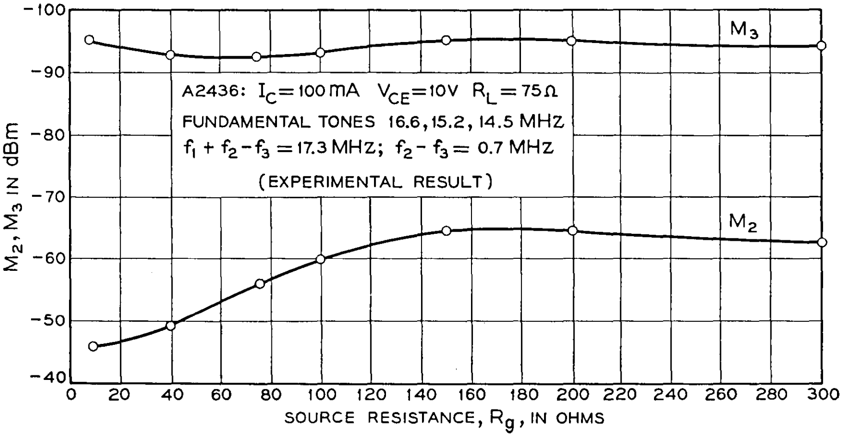


Fig. 10—Variation of M_2, M_3 with source resistance.

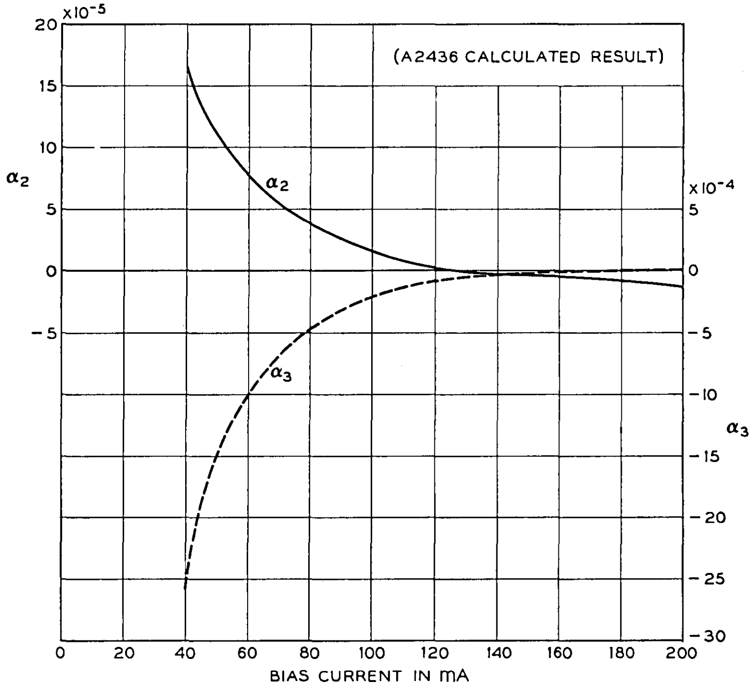


Fig. 11 —Variation of α_2 , α_3 with bias current.

it becomes zero at $I_C = I_{C\,\max}/e$, and then becomes negative, and increases with further increase in I_C . The coefficient α_3 decreases with bias current I_E . Thus, in general, an increase in bias current has the effect of reducing both the second and third harmonic distortion (Fig. 7) (at least until $\alpha_2 = 0$).

5.3.5 *Effect of Bias Voltage*

Whereas exponential and h_{FE} terms are functions of bias current, the avalanche and collector capacitance nonlinearities are affected by the bias voltage. The coefficient $\hat{M}_2$ increases with the voltage; but $\hat{M}_1$ and $\hat{M}_3$ increase much more rapidly (Fig. 12). (Both the collector capacitance nonlinear coefficients γ_2 , γ_3 decrease with the increase in bias voltage.) The effects of change in bias voltage are especially pronounced at higher load resistance since avalanche (and collector capacitance) terms become dominant. The third harmonic distortion decreases more with the increase in voltage (Fig. 13) than the second harmonic distortion does.

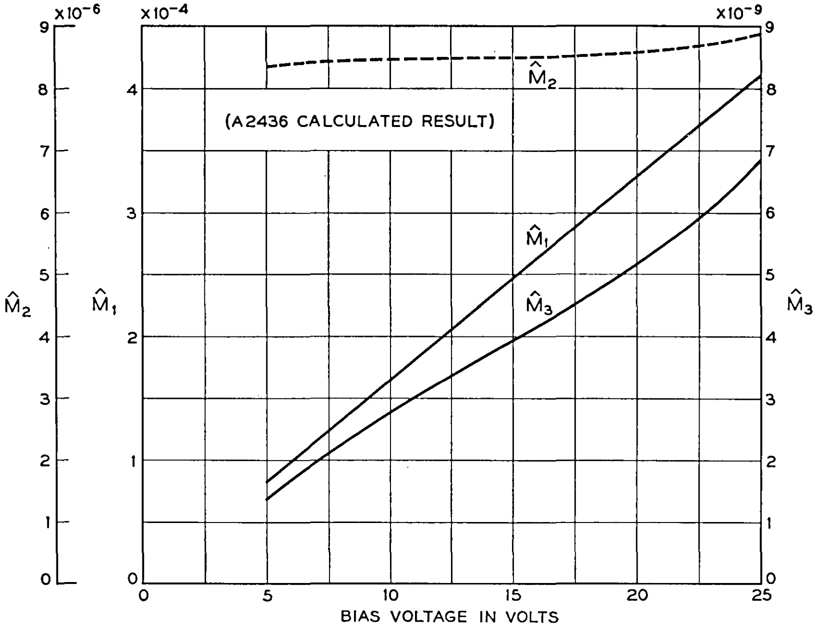


Fig. 12—Variation of $\hat{M}_1$, $\hat{M}_2$, $\hat{M}_3$ with bias voltage.

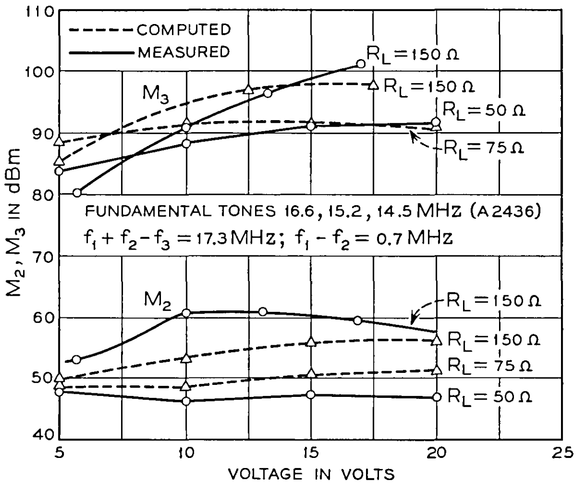


Fig. 13—Variation of M_2, M_3 with voltage at 100 mA.

The physical significance of the closed form expressions has been qualitatively discussed. Precise quantitative estimates can and have been obtained using the computer program. For example, the effect of varying linear parameters by fifty percent of their original values was studied. The results show that the distortion does not critically depend on the linear parameters (Figs. 14 to 17). The other transistor parameters such as $I_{C\,\max}$, V_{CEO} , n, etc., can also be varied.

VI. ANALYSIS OF CASCADED TRANSISTORS

It is often stated that in a multi-stage amplifier, the output stage alone determines the over-all distortion. Even though this statement is true to some extent, it is frequently found in practice that the effects

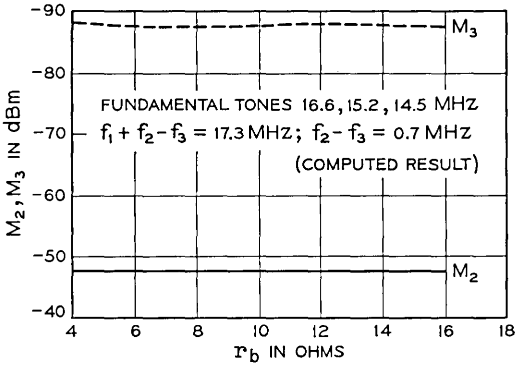


Fig. 14—Variation of M_2, M_3 with r_b.

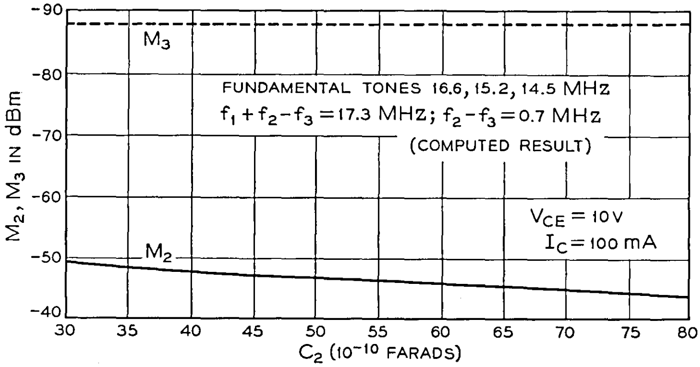


Fig. 15—Variation of M_2, M_3 with C_2.

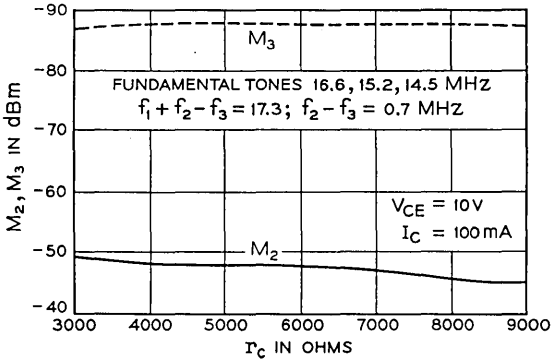


Fig. 16 — Variation of M_2, M_3 with r_c.

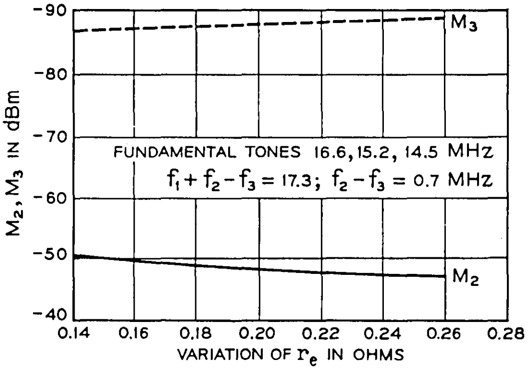


Fig. 17 — Variation of M_2, M_3 with r_e.

of the previous stages cannot be ignored and sometimes the previous stage is dominant. This is especially true if both minimum noise figure (which requires lower bias current) and modulation requirements are to be met by a two stage amplifier. Two analysis tools based on Volterra series are presented here which enable the study of such cascaded stages.

The first approach makes use of the cascaded formulae mentioned earlier; this method illustrates the cascade phenomenon and permits derivation of simple cascade rules.

Consider two cascaded transistors (Fig. 1); let the output voltage (v_2) of the first transistor be denoted by $D(v_g)$; the output voltage (v_3) of the second stage by $E(v_2)$ and $F(v_g)$. The aim is to compute the kernels $F_1(s_1)$, $F_2(s_1, s_2)$, $F_3(s_1, s_2, s_3)$ knowing D and E. To calculate $D(s_1)$, etc., it is necessary to know the load impedance of the first stage which is the input impedance of the second transistor. This can be computed; thus, for a given generator impedance and bias conditions, $D(s_1)$, $D_2(s_1, s_2)$, $D_3(s_1, s_2, s_3)$ can be determined. $E(s_1)$, $E_2(s_1, s_2)$ and $E_3(s_1, s_2, s_3)$ can be computed for a given load and bias conditions with $R_g = 0$ (voltage v_2 is directly impressed across the second). Now expression v_3 in terms of v_g is given by

$$v_3 = F(v_g) = E(v_2) = E(D(v_g)) = (E \ o \ D)(v_g). \tag{30}$$

It is seen that F is related to E and D by the cascade formulae whose physical significance is discussed below.

6.1 *Linear Term*

The linear term is given by

$$F_1(s) = D_1(s)E_1(s) \tag{31}$$

which states that the overall gain in dB is the gain of the first stage in dB plus the gain of the last stage in dB.

6.2 *Second Harmonic Term*

The second-degree kernel is given by

$$F_2(s_1, s_2) = E_1(s_1 + s_2)D_2(s_1, s_2) + E_2(s_1, s_2) \prod_{i=1}^{2} D_1(s_i). \tag{32}$$

The first term of the formula states that a given harmonic product from the first transistor $D_2(j\omega_a \pm j\omega_b)$ is amplified by the second transistor at the harmonic frequency $E_1(j\omega_a \pm j\omega_b)$. The second term shows that the two fundamental tones are amplified by the first transistor $[D_1(j\omega_a)D_1(\pm j\omega_b)]$ and then the second transistor acts on these tones to produce distortion $E_2(j\omega_a, \pm j\omega_b)$.

Equation (32) is related to the second harmonic distortion (M_2) as follows:

$$M_2 = 20 \log \frac{\sqrt{10^{-3}R_L}}{2} \left| \frac{F_2(s_1, s_2)}{F_1(s_1)F_1(s_2)} \right| \tag{33}$$

$$= 20 \log \frac{\sqrt{10^{-3}R_L}}{2} \left| \left[\frac{D_2(s_1, s_2)}{\prod_{i=1}^{2} D_1(s_i)} \right] \frac{E_1(s_1 + s_2)}{E_1(s_1)E_1(s_2)} + \frac{E_2(s_1, s_2)}{\prod_{i=1}^{2} E_1(s_i)} \right| . \tag{34}$$

The second term is the second harmonic distortion of the last stage. The first term expresses the contribution from the first stage; it approximately equals

$$\begin{bmatrix}\text{First stage second harmonic} \\ \text{distortion in dBm}\end{bmatrix} - \begin{bmatrix}\text{gain of the last stage} \\ \text{in dB}\end{bmatrix}. \tag{35}$$

This shows that if the gain of the last stage is high, the contribution from the first stage is small. Equation (35) is approximate in two respects; it neglects the frequency effects and the phase addition of the contributions from the first and the second stage. In (35), the second stage gain in question is actually the ratio

$$\frac{E_1(s_1 + s_2)}{E_1(s_1)E_1(s_2)}$$

which involves the two fundamental and the harmonic frequencies. As an example, a shaping network which was introduced increased the gain (18 dB) at the harmonic frequency (0.7) MHz) and decreased the gain at fundamental tones 15.2 MHz (8 dB) and 14.5 MHz (8 dB) with the result the overall distortion was poorer by 34 dB.

6.3 *Third Harmonic Distortion Term*

The third harmonic kernel $F_3(s_1, s_2, s_3)$ is given by

$$F_3(s_1, s_2, s_3) = E_1(s_1 + s_2 + s_3)D_3(s_1, s_2, s_3) + 2E_2(s_1, s_2 + s_3) \cdot D_1(s_1)D_2(s_2, s_3) + E_3(s_1, s_2, s_3) \prod_{i=1}^{3} D_1(s_i). \tag{36}$$

The first term shows that the third harmonic product of the first stage $[D_3(s_1, s_2, s_3)]$ is amplified by the last stage at the harmonic frequency $[E_1(s_1 + s_2 + s_3)]$. The second term is the interaction term; it arises when the second-degree kernel of the last stage $[E_2(s_1, s_2 + s_3)]$ acts on the sum of the fundamental $[D_1(s_1)]$ and the second harmonic output of the first stage $[D_2(s_2, s_3)]$. The last term shows that the second stage third-degree kernel $[E_3(s_1, s_2, s_3)]$ acts on the fundamental tones amplified at the respective frequencies by the first stage $[D_1(s_1)D_1(s_2)D_1(s_3)]$.

From (36), the overall third harmonic distortion is related to that of the individual transistors by

$$M_3 = 20 \log \frac{1}{4} 10^{-3} R_L \left| \left[\frac{E_1(s_1 + s_2 + s_3)}{\prod_{i=1}^{3} E_1(s_i)}\right] \left[\frac{D_3(s_1, s_2, s_3)}{\prod_{i=1}^{3} D_1(s_i)}\right] + 2\left[\frac{E_2(s_1, s_2 + s_3)}{\prod_{i=1}^{2} E_1(s_i)}\right] \frac{1}{E_1(s_3)} \left[\frac{D_2(s_2, s_3)}{\prod_{i=1}^{2} D_1(s_i)}\right] + \left[\frac{E_3(s_1, s_2, s_3)}{\prod_{i=1}^{3} E_1(s_i)}\right] \right|. \tag{37}$$

The first term is the contribution from the third harmonic term of the first stage; it is given approximately by

$$\begin{bmatrix}\text{Third harmonic distortion}\\ \text{of the first stage in dBm}\end{bmatrix} - 2\begin{bmatrix}\text{Gain of the last}\\ \text{stage in dB}\end{bmatrix}. \qquad (38)$$

The interaction term approximately equals

$$\begin{bmatrix}\text{Second harmonic distortion}\\ \text{of the first stage in dBm}\end{bmatrix} + \begin{bmatrix}\text{Second harmonic distortion}\\ \text{of the second stage in dBm}\end{bmatrix}$$

$$+ 6 \text{ dB} - \begin{bmatrix}\text{Gain of the last}\\ \text{stage in dB}\end{bmatrix}. \qquad (39)$$

The third term in (37) is the third harmonic distortion of the last stage in dBm.

It is seen that the effect of the first stage and the interaction term can be reduced by increasing the gain of the last stage. Equation (39) illustrates that the second harmonic distortion of each stage should be good. This may become a limitation if the first stage is biased at lower currents.

In the above simplified expressions [(38) and (39)] phase addition and frequency effects have not been considered. In (38), 2 (gain in dB) actually represents

$$20 \log \left| \frac{E_1(j\omega_a \pm j\omega_b \pm j\omega_c)}{E_1(j\omega_a)E_1(\pm j\omega_b)E_1(\pm j\omega_c)} \right|.$$

In (39) the second harmonic distortion is to be measured with two tones, one at the fundamental and the other at the harmonic frequency

$$\frac{E_2(s_1\ , s_2 + s_3)}{E_1(s_1)E_1(s_2 + s_3)}$$

and then multiplied by the ratio of the gain

$$\frac{E_1(s_2 + s_3)}{E_1(s_1)}.$$

Moreover, the kernel must be made symmetrical by taking the average of three possible combinations.

Thus, the simplified expressions (35), (38), and (39) are exact if the transistor performance is not frequency dependent; in general, they can be used to get a qualitative picture. Equations (34) and (37) are indeed exact and take into account the frequency dependence. The computer program is being extended to calculate (34) and (37).

An alternate approach to calculate the distortion of cascaded stages is to analyze the nonlinear equivalent circuit of cascaded transistors using the nodal technique illustrated in Section IV. The nodal equations are derived first; next each nodal voltage is expressed in terms of the Volterra series of the input voltage; the resulting vector matrix equations are successively solved. Since the procedure is similar, the details are omitted.

Two common-collector stages were cascaded using this approach. (Fig. 18). The measured values at 120 mA, 10 V with 75 ohm source and load impedances were −87 dBm and −112 dBm for the second and the third harmonic distortion, respectively. The computed distortion values are −88.7 dBm for second and −116.6 dBm for third harmonic distortion. Thus, good agreement with experimental result is obtained.

The cascade formulas are simple, physically meaningful and yield rules of thumb to judge the effect of the first stages. The nodal approach is more complicated. However, the advantage of the nodal approach is that it is general and can be used for an amplifier. For example, a cascade of common-emitter and common-collector stages involves five nodes; if shunt feedback is used at the input and at the output, the same program can be used to analyze this amplifier. (Cascade formulas do not take feedback into account.) In general, the nodal approach can be extended to study frequency-dependent nonlinear network with n nodes, if the nonlinearities are small.

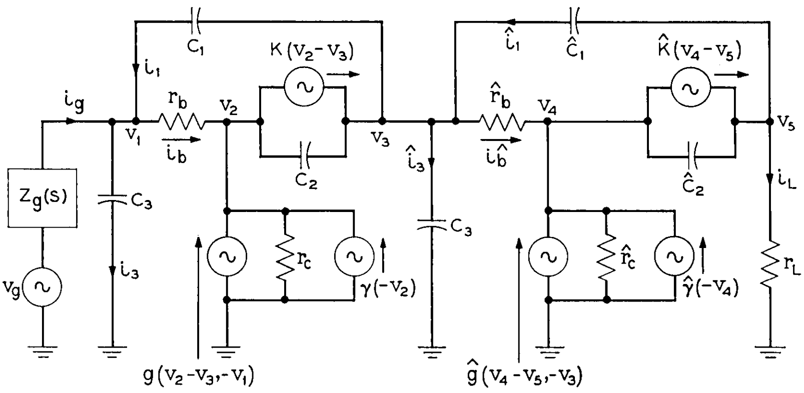


Fig. 18 — Common-collector — common-collector nonlinear equivalent circuit.

VII. ENGINEERING APPLICATIONS

A few pertinent practical applications of the work are described below. These results were either first predicted by the model and then verified in the laboratory or first experimentally observed and then confirmed by analysis.

In the initial design of L4 repeater a common-emitter—common-emitter—common-collector configuration[28] was used in the power amplifier. The third harmonic modulation performance was not as good as desired. This led first to the study of the output common-collector stage. As shown in Fig. 19, the increase in source impedance increases the distortion of the common-collector stage. Since the preceding common-emitter stage output impedance is high, the common-collector performance was not optimum. Secondly, the preceding common-emitter stage was studied because the gain of the common-collector stage is low. (see Section VI) As shown in Fig. 20, increase in load impedance beyond optimum R_L degrades its performance radically. Since the common-collector input impedance is high, the common-emitter stage performance was not optimum either. Thus, in the redesign work by Ken Tantarelli, the common-collector output stage is not being used.

Another interesting application feature was the improvement in modulation performance of the common-emitter stage with increase in voltage. As shown in Fig. 8, it is a function of load impedance, and at about 150Ω, maximum improvement was obtained.

New coaxial systems are currently being studied to operate at higher frequencies. Different configurations have been examined for the output stage. The model showed that common-collector and common-base performance is poorer with an increase in frequency and thus the use of these stages as output stages was questioned (unless transistors with

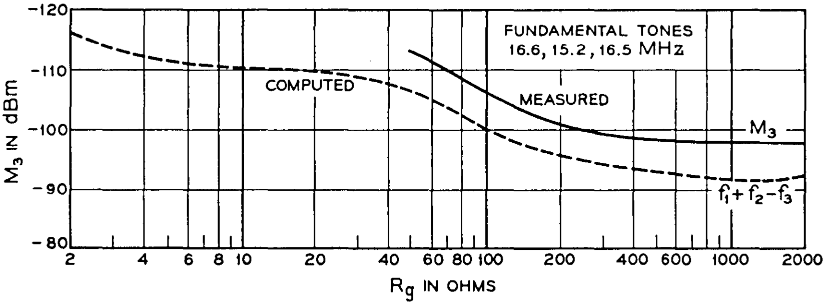


Fig. 19—Common-collector; M_3 variation with source impedance; I_c = 120 mA; V_{ce} = 10V; R_L = 75 Ω.

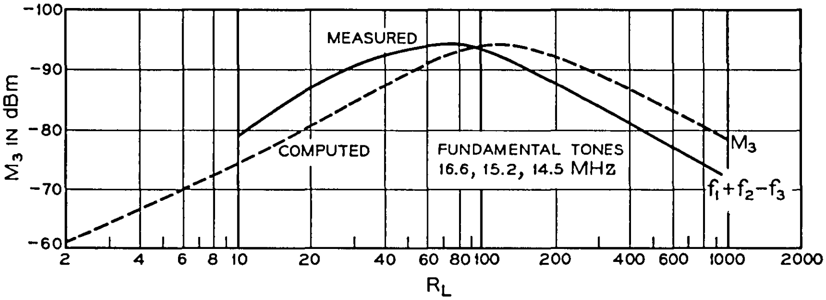


Fig. 20—Common-emitter; M_3 variation with load impedance; $I_c = 120$ mA; $V_{ce} = 10V$; $R_g = 75\ \Omega$.

higher f_t's are available). Recently when a new, high-frequency modulation test set was built, experiments confirmed the prediction. The third harmonic coefficient ($M3$) for $a + b - c$ product was 8 dB poorer at 36.5 MHz (due to signals at 36.5 MHz, 40.1 MHz, and 43.1 MHz) compared to the value at 17.3 MHz (due to signals at 14.5, 15.2, and 16.6 MHz). The common-emitter configuration modulation performance suffered only about one dB degradation.

VIII. CONCLUSION AND ACKNOWLEDGMENT

This paper has presented a useful analysis tool for investigating the frequency-dependent nonlinear behavior of transistors. A digital program for all the three configurations has been prepared. The results obtained compare favorably with experimental results. The closed-form expressions yield a qualitative picture of distortion. The Volterra series proves useful in examining cascaded transistors; a few rules of thumb are derived and a general nodal analysis which can be extended to cascaded stages with feedback is developed. The practical applications cited show that the technique can be useful in the computer-aided optimal design of linear transistor feedback amplifiers.

The author acknowledges gratefully the cooperation he received from Lee C. Thomas who is responsible for much of the initial development of the model; he has been particularly concerned about the cancellation mechanism which he observed in the analog computer simulation. The author wishes to thank Mr. Jack Huang, who first suggested the model, Miss J. A. Nicosia, who wrote the program, Mr. F. Kelcourse for many useful discussions, and Mr. R. E. Maurer for reading the draft. The author would also like to thank Dr. F. H. Blecher for his encouragement and continued interest and guidance.

APPENDIX A

A.1 *Higher-Dimensional Transforms*

The second-degree case is illustrated as an example. From (4),

$$y_2(t) = \int_0^t \int_0^t c_2(t - \tau_1 , t - \tau_2) \prod_{i=1}^{2} x(\tau_i) \, d\tau_i \, . \tag{40}$$

If the system is physically realizable, $c_2(t - \tau_1 , t - \tau_2) = 0$, for $\tau_i > t$. Hence, the limits can be extended to ∞.

$$y_2(t) = \int_0^\infty \int_0^\infty c_2(t - \tau_1 , t - \tau_2) \prod_{i=1}^{2} x(\tau_i) \, d\tau_i \, . \tag{41}$$

Introducing dummy variables t_1 and t_2, the two-dimensional transform is taken

$$\begin{aligned} Y_2(s_1 , s_2) &= \int_0^\infty \int_0^\infty y_2(t_1 , t_2) \exp(-s_1 t_1) \exp(-s_2 t_2) \, dt_1 \, dt_2 \\ &= \int_0^\infty dt_1 \int_0^\infty dt_2 \left[\int_0^\infty d\tau_1 \int_0^\infty d\tau_2 \, c_2(t_1 - \tau_1 , t_2 - \tau_2) \right. \\ &\qquad \left. \cdot \prod_{i=1}^{2} x(\tau_i) \, d\tau_i \right] \exp(-s_1 t_1) \exp(-s_2 t_2). \end{aligned} \tag{42}$$

Substituting $t_1 - \tau_1 = m_1$, $t_2 - \tau_2 = m_2$, and using the fact that $c_2(m_1 , m_2) = 0$ for $m_i < 0$ yields

$$\begin{aligned} Y_2(s_1 , s_2) &= \int_0^\infty dm_1 \int_0^\infty dm_2 \int_0^\infty d\tau_1 \int_0^\infty d\tau_2 \, c_2(m_1 , m_2) x(\tau_i) x(\tau_2) . \\ &\quad \cdot \exp(-s_1 m_1) \exp(-s_1 \tau_1) \exp(-s_2 m_2) \exp(-s_2 \tau_2) \end{aligned} \tag{43}$$

$$= C_2(s_1 , s_2) \, X(s_1) \, X(s_2). \tag{44}$$

A.2 *The Output of the Kernels to Sinusoidal Inputs*

For the second-degree case, consider two sinusoidal signals at frequencies f_a and f_b. The input $x(\tau)$ equals,

$$x(\tau) = \left[\frac{\exp(j\omega_a \tau) + \exp(-j\omega_a \tau)}{2} \right] + \left[\frac{\exp(j\omega_b \tau) + \exp(-j\omega_b \tau)}{2} \right]. \tag{45}$$

From (41)

$$y(t) = \int_0^\infty d\tau_1 \int_0^\infty d\tau_2 \, c_2(t - \tau_1 , t - \tau_2)$$

$$\cdot\left[\frac{\exp(j\omega_a\tau_1)+\exp(-j\omega_a\tau_1)}{2}+\frac{\exp(j\omega_b\tau_1)+\exp(-j\omega_b\tau_1)}{2}\right]$$

$$\cdot\left[\frac{\exp(j\omega_a\tau_2)+\exp(-j\omega_a\tau_2)}{2}+\frac{\exp(j\omega_b\tau_2)+\exp(-j\omega_b\tau_2)}{2}\right]. \quad (46)$$

Considering one cross term only,

$$\int_0^\infty d\tau_1 \int_0^\infty d\tau_2\, c_2(t-\tau_1\,,\, t-\tau_2)$$

$$\cdot\tfrac{1}{4}\exp(j\omega_a\tau_1)\exp(j\omega_b\tau_2). \quad (47)$$

Substituting $m_1 = t - \tau_1$, $m_2 = t - \tau_2$ and carrying out the integration yields

$$\tfrac{1}{4}C_2(j\omega_a\,;\, j\omega_b)\exp[j(\omega_a+\omega_b)t]. \quad (48)$$

This term occurs twice as does its complex conjugate.

Hence, the output due to the $a + b$ term alone is

$$y_{a+b}(t) = |\, C_2(j\omega_a\,,\, j\omega_b)\,|\cos[(\omega_a+\omega_b)t+\varphi_{a+b}]. \quad (49)$$

The $2\omega_a$ term and its conjugate occur only once in (46); hence, it is 6 db better. The response of the third harmonic kernel to three sinusoidal inputs is similarly treated.

A.3 *Cascade Relations*

For the system shown in Fig. 1, the cascade formula are given below. The cascade relations can be symbolically written as

$$Z = F(x) = E(y) = E(D[x]) = (E\ o\ D)(x). \quad (50)$$

The formula are

$$F_1(s_1) = E_1(s_1)D_1(s_1) \quad (51)$$

$$F_2(s_1\,,\, s_2) = E_1(s_1+s_2)D_2(s_1\,,\, s_2) + E_2(s_1\,,\, s_2)\prod_{i=1}^{2} D_1(s_1) \quad (52)$$

$$F_3(s_1\,,\, s_2\,,\, s_3) = E_1(s_1+s_2+s_3)D_3(s_1\,,\, s_2\,,\, s_3)$$

$$+\, 2E_2(s_1\,,\, s_2+s_3)D_1(s_1)D_2(s_2\,,\, s_3) + E_3(s_1\,,\, s_2\,,\, s_3)\prod_{i=1}^{3} D_1(s_i). \quad (53)$$

A physical interpretation of the formula for cascaded transistors is given in Section VI. The procedure for deriving the cascade relation is as follows: the output $Z(t)$ of the last stage is expressed in terms of the Volterra series of its input. (Only two terms are considered)

$$Z(t) = \int_0^\infty e_1(t - \tau_1) y(\tau_1)\, d\tau_1 + \int_0^\infty \int_0^\infty e_2(t - \tau_1, t - \tau_2) \prod_{i=1}^{2} y(\tau_i)\, d\tau_i . \tag{54}$$

The output of the first stage $y(t)$ is related to its input by

$$y(\tau) = \int_0^\infty d_1(\tau - \sigma_1) x(\sigma_1)\, d\sigma + \int_0^\infty \int_0^\infty d_2(\tau - \sigma_1, \tau - \sigma_2) \prod_{i=1}^{2} x(\sigma_i)\, d\sigma_i \tag{55}$$

Substituting (55) in (54), terms of the same degree are collected; as an example, the first second-degree term equals

$$\int d\tau\, e_1(t - \tau) \iint d_2(\tau - \sigma_1, t - \sigma_2) \prod_{i=1}^{2} x(\sigma_i)\, d\sigma_i . \tag{56}$$

Taking the two-dimensional transforms yields

$$E_1(s_1 + s_2) D_2(s_1, s_2) \prod_{i=1}^{2} X(s_i). \tag{57}$$

APPENDIX B

The Nonlinear Parameters

From (15),

$$I_c = f(I_E) h(V_{CB}). \tag{58}$$

A two-dimensional Taylor's series expansion of (58) is taken; i_e is expressed by $K(v_2)$ and $v_{CB} = v_3 - v_1$. Hence,

$$i_c = g(v_2, v_3 - v_1) = g_1(v_2, v_3 - v_1) + g_2(v_2, v_3 - v_1) + g_3(v_2, v_3 - v_1), \tag{59}$$

where

$$g_1(v_2, v_3 - v_1) = \alpha_1 \hat{M}_0 K_1 v_2 + \hat{M}_1(v_3 - v_1), \tag{60}$$

$$g_2(v_2, v_3 - v_1) = \alpha_2 \hat{M}_0 K_1^2 (v_2)^2 + m_2(v_3 - v_1)^2 + \alpha_1 \hat{M}_0 K_2 (v_2)^2 + \alpha_1 \hat{M}_1 K_1 (v_2)(v_3 - v_1), \tag{61}$$

and

$$g_3(v_2, v_3 - v_1) = \alpha_3 \hat{M}_0 (K_1)^3 (v_2)^3 + m_3(v_3 - v_1)^3 + \alpha_1 \hat{M}_2 K_1 v_2 (v_3 - v_1)^2 + \alpha_2 \hat{M}_1 K_1^2 v_2^2 (v_3 - v_1) + \alpha_1 \hat{M}_0 K_3 (v_2)^3 + 2\alpha_2 \hat{M}_0 K_1 K_2 (v_2)^3 + \alpha_1 \hat{M}_1 K_2 (v_2)^2 (v_3 - v_1). \tag{62}$$

The avalanche coefficients are given below.

$$\hat{M}_0 = \frac{1}{1 - \left(\dfrac{V_{CB}}{V_{CBO}}\right)^n} \tag{63}$$

$$\hat{M}_1 = (\hat{M}_0)' = \frac{n(V_{CB})^{n-1}}{(V_{CBO})^n} (\hat{M}_0)^2 \tag{64}$$

$$\hat{M}_2 = \tfrac{1}{2}(\hat{M}_1)' = \tfrac{1}{2}(n - 1) \frac{\hat{M}_1}{V_{CB}} + \frac{(\hat{M}_1)^2}{\hat{M}_0} \tag{65}$$

$$\hat{M}_3 = \tfrac{1}{3}(\hat{M}_2)' = \tfrac{2}{3}\hat{M}_2\left(\frac{(n-1)}{2V_{CB}} + \frac{2\hat{M}_1}{\hat{M}_0}\right) - \frac{\hat{M}_1}{3}\left[\frac{(n-1)}{2(V_{CB})^2} + \left(\frac{\hat{M}_1}{\hat{M}_0}\right)^2\right]\cdot \tag{66}$$

The coefficients m_1, m_2, and m_3 equal $m_i = I_C(\hat{M}_i/\hat{M}_0)$; $i = 1, 2, 3$, where I_C is the collector dc bias current.

The h_{FE} coefficients are given below:

$$\alpha_1 = \frac{h_{FE\,\max}}{h_{FE\,\max} + 1 + a \log^2 \dfrac{I_C}{I_{C\,\max}} + 2a \log e \log \dfrac{I_C}{I_{C\,\max}}} \tag{67}$$

$$\alpha_2 = -\frac{1}{2I_C} \frac{(\alpha_1)^3}{h_{FE\,\max}} 2a \log e\left[\log \frac{I_C}{I_{C\,\max}} + \log e\right] \tag{68}$$

$$\alpha_3 = \frac{\alpha_1}{6}\left[\frac{-2\alpha_2}{I_C} + 12 \frac{(\alpha_2)^2}{(\alpha_1)^2} - \frac{1}{(I_C)^2} \frac{(\alpha_1)^3}{h_{FE\,\max}} 2a(\log e)^2\right]\cdot \tag{69}$$

The collector capacitance coefficients are given by

$$\gamma_1 = k(V_{CB})^{-\frac{1}{3}} \tag{70}$$

$$\gamma_2 = \frac{-1}{6} k(V_{CB})^{-\frac{4}{3}} \tag{71}$$

$$\gamma_3 = \frac{k}{27} (V_{CB})^{-\frac{7}{3}}. \tag{72}$$

From (62) for $g_3(v_2\,; v_3 - v_1)$, $\hat{g}_3$ is obtained by replacing $B_1(s_i)$ for v_2 and $C_1(s_i) - A_1(s_i)$ for $(v_3 - v_1)$; moreover, the kernel must also be symmetrical. Since the procedure is the same as for $\hat{g}_2$ it is omitted. The interaction terms are given below:

$$\begin{aligned}\hat{g}_{23} = {} & 2\alpha_2\hat{M}_0K_1^2\, \overline{B_1(s_1)B_2(s_2\,, s_3)} \\ & + 2m_2\, \overline{[C_1(s_1) - A_1(s_1)][C_2(s_2\,, s_3) - A_2(s_2\,, s_3)]} \\ & + \alpha_1\hat{M}_1K_1\, \overline{B_2(s_1\,, s_2)[C_1(s_3) - A_1(s_3)]}\end{aligned}$$

$$+ \alpha_1 \hat{M}_1 K_1 \overline{B_1(s_1)[C_2(s_2\,, s_3) - A_2(s_2\,, s_3)]} \tag{73}$$

$$\hat{\gamma}_{23} = 2\gamma_2 \overline{[C_1(s_1) - B_1(s_1)][C_2(s_2\,, s_3) - B_2(s_2\,, s_3)]} \tag{74}$$

where—denotes symmetrical kernel.

APPENDIX C

A2436 is an n-p-n silicon transistor with overlay type of construction. Its f_T ranges from 800 to 1000 MHz. It is a power transistor with current capability of 1 amp and can handle 2.2 watts of power.

Typical parameter values for transistor type 2436 27 at 120 mA, 10V are given below:

I_c = 0.12 amps.
r_b = 13.6 ohms
r_c = 5200 ohms
C_1 = $(6)10^{-12}$ farads
C_2 = $(3.97)10^{-9}$ farads
C_3 = $(9.2)10^{-12}$ farads
Z_g = 50 ohms
Z_L = 50 ohms
V_{CB} = 10 volts
V_{CBO} = 350 volts
n = 2
re_1 = 0.2165 ohms
a = 0.38
$h_{FE\ \max}$ = 122
$I_{C\ \max}$ = 0.633 amps.

REFERENCES

1. Lotsch, H., Third Order Distortion and Cross-Modulation in a Grounded Emitter Transistor Amplifier, IRE Trans. Audio, March, 1961, pp. 49–56.
2. Mallinckrodt, A. J. and Gardner, F. M., Distortion in Transistor Amplifiers, IRE Trans. Electron Devices, July, 1963, pp. 288–289.
3. Krislov, Y. D., Nonlinear and Cross-Modulation Distortion in Transistor Amplifiers, Telecommun. Radio Eng., March, 1965, pp. 97–103.
4. Lotsch, H., Survey of Nonlinear Distortions in Transistor Stages Including Cross-Modulation, Arch. Ebek Vebertragung 14, No. 5, 1960, pp. 204–216.
5. Reynolds, J., Nonlinear Distortions and Their Cancellation in Transistors, IEEE Electron Devices, November, 1965.
6. Riva, G. M., Beneteau, P. J., and DallaVolta, E., Amplitude Distortion in Transistor Amplifier, Proc. IEEE, No. 3, March, 1964, pp. 481–490.
7. Meyer, N. I., Nonlinear Distortion in Transistor Class A Amplifiers at Low and Medium Frequencies, Proc. IEEE, May, 1959, pp. 481–489.
8. Meyer, N. I., On the Variation of Transistor Small Signal Parameters with Emitter Current and Collector Voltage, J. Electron. Control, *4,* 1958, p. 305.

9. Meyer, N. I., *Nonlinear Distortion and Small Signal Parameters of Alloyed Junction Transistors,* Danish Science Press Ltd., Copenhagen, 1960.
10. Thomas, L. C., Broadband linearization of Transistor Amplifiers, Presented at International Solid State Circuits Conference, February, 1967.
11. Narayanan, S., Transistor Distortion Analysis Using Volterra Series Representation, (Oral presentation at International Conference on Communications, June, 1967).
12. Volterra, V., *Theory of Functionals and of Integral and Integrodifferential Equations,* Dover Publications, New York, 1959.
13. Wiener, N., Response of a Nonlinear Device to Noise, MIT, Radiation Laboratory, Cambridge, Mass., Report No. 129, (V-16), April, 1942.
14. Bose, A. G., A Theory of Nonlinear Systems, Technical Report 309, Research Laboratory of Electronics, MIT, May 15, 1956.
15. Wiener, N., *Nonlinear Problems in Random Theory,* MIT Press, Cambridge, Mass., 1958.
16. Brilliant, M. B., Theory of the Analysis of Nonlinear Systems, Technical Report 345, Research Laboratory of Electronics, MIT, March 3, 1958.
17. George, D. A., Continuous Nonlinear Systems, Technical Report 355, Research Lab of Electronics, MIT, July 22, 1959.
18. Chesler, D. A., Nonlinear Systems with Gaussian Inputs, Technical Report 366, Research Lab of Electronics, MIT, February 15, 1960.
19. Barrett, J. F., The Use of Functionals in the Analysis of Nonlinear Physical Systems, Statistical Advisory Unit, Report No. 1/57, Ministry of Supply, Great Britain, 1957.
20. VanTrees, H. L., *Volterra Series Representation of Optimum Nonlinear Control Systems*, MIT Press, Cambridge, Mass., 1962.
21. VanTrees, H. L., Functional Techniques for the Analysis of the Nonlinear Behavior of Phase-Locked Loops, Proc. IEEE, *52,* No. 8, August, 1964, pp. 894–911.
22. Alper, P., A Consideration of Discrete Volterra Series, 1965, Joint Automatic Control Conference, Rensselaer Polytechnic Institute, Troy, N. Y., June, 1965, p. 96.
23. Bush, A. M., Some Techniques for the Synthesis of Nonlinear Systems, Sc-D Thesis, Dept. of Electrical Engineering, MIT, Cambridge, Mass., May, 1965.
24. Narayanan, S., Transform Methods for Special Nonlinear Systems, Ph.D. Thesis, Carnegie Tech., Pittsburgh, Pa., May, 1965.
25. Flake, R. H., Volterra Series Representation of Time-Varying Nonlinear Systems, Proc. Second International Congress of IFAC on Automatic Control Based, Switzerland, Paper No. 408/1, 1963.
26. Ku, Y. H. and Wolf, A. A., Volterra-Wiener Functionals for the Analysis for Nonlinear Systems, J. Franklin Inst., *281,* No. 1, January, 1966, pp. 9–26.
27. Phillips, A. B., *Transistor Engineering,* McGraw-Hill Book Co., New York, 1962.
28. Kelcourse, F. C. and Labbe, L. P., Transistor Feedback Amplifiers for 0.5 m/c-20 m/c Long Haul Coaxial Cable Transmission System, IEEE NEREM Conference, 1964.

Coding for Numerical Data Transmission*

By M. M. BUCHNER, JR.

(Manuscript received June 2, 1966)

This paper considers the effectiveness of error-correcting codes for the transmission of numerical data. In such a situation, errors in the numerically most significant positions of a message are of greater consequence than are errors in the less significant positions. A measure of transmission fidelity based upon the average magnitude by which the numbers delivered to the destination differ from the transmitted numbers is developed and is referred to as the average numerical error (ANE). Codes are compared by comparing the ANE that results from their use.

Significant-bit codes are defined and the ANE resulting from their use is determined. For constant-symbol-rate transmission, the relative effectiveness of various coding schemes is analyzed when the error probability in the channel is small. The ANE resulting from the use of certain specific codes is numerically evaluated and compared.

I. INTRODUCTION

The usual approach to coding is to ignore the actual meaning of the transmitted symbols and to represent them in a purely statistical manner. As a result, all message errors are assumed to be equally costly and codes have been sought that simply reduce the probability that a message is received in error.

While this may be appropriate for the transmission of some types of data, there are situations in which other criteria of goodness are of greater merit. If, for example, one is interested in the transmission of the temperature of a satellite, the probability that a particular observation is transmitted incorrectly may have little direct relation to system performance whereas a measure of the average magnitude by which the received data differ from the data actually transmitted could prove useful.

[*] The material presented in this paper is based upon the dissertation, Coding for Numerical Data Transmission, submitted by the author to The Johns Hopkins University in conformity with the requirements for the degree Doctor of Philosophy.

This paper develops a criterion of transmission fidelity for numerical data transmitted over a binary symmetric channel based upon the average numerical error which occurs. Significant-bit codes are defined and the average numerical error resulting from their use is determined for a binary symmetric channel with independent errors. For constant-symbol-rate transmission, the relative effectiveness of various coding schemes is analyzed when the probability that a symbol is received in error is small. In order to obtain a feeling for the utility of coding, the average numerical error resulting from certain specific codes is numerically evaluated.

II. PRELIMINARIES

Throughout this paper, the channel is taken to include all operations performed upon the symbols during transmission. A binary symmetric channel is defined to be a binary channel such that

(*i*) the channel always gives one of the binary symbols at its output,
(*ii*) the probability that any particular sequence of errors occurs is independent of the symbols transmitted.

In some sections, we shall consider a binary symmetric channel with independent errors. This is a binary symmetric channel for which the errors occur independently with probability p where $0 \leqq p \leqq \frac{1}{2}$ and $p = 1 - q$.

The elements of the Galois field of two elements are denoted by 0 and 1. Let the symbol $\oplus$ denote component by component modulo 2 addition of vectors (or n-tuples) whose components are field elements. The set of all such vectors forms a vector space Γ of dimension n over the field of two elements. Because a field element can be viewed as a vector with one component, $\oplus$ will also be used to denote the addition of field elements.

A binary group code V is a subset of Γ which forms a group. Over the field of two elements, any set of n-tuples that forms a group is indeed a vector space. Therefore, a binary group code V forms a subspace of Γ. The dimension of V is k.

The implementation of a binary group code can be viewed in the following manner. The encoder receives k binary information symbols (called a message) from the source and determines from the message $(n - k)$ binary parity check symbols (called an ending). The message and ending may be interleaved or transmitted sequentially to form a block of length n (called a code vector). The decoder operates upon the blocks of n binary symbols coming from the channel in an attempt to correct transmission errors and provides k binary symbols at its output. The notation (n,k) is used to denote such a code.

Consider the message $(m_k, m_{k-1}, \cdots, m_1)$. The code vector used to transmit this message will have $m_k, m_{k-1}, \cdots, m_1$ in the k information positions. The $(n - k)$ parity check positions that form the ending are denoted by $e_1, e_2, \cdots, e_{n-k}$. The order in which the information positions and the parity check positions are arranged for transmission is arbitrary.

Let H denote the parity check matrix for a binary group code. H is an $(n - k) \times n$ matrix whose entries are field elements. An n-tuple v is a code vector if and only if

$$v\tilde{H} = 0, \tag{1}$$

where $\tilde{H}$ denotes the transpose of H. H can be written in a form such that each column of H that corresponds to a parity check position in a code vector is a distinct weight* one $(n - k)$-tuple. When this is done, let $C_l(1 \leqq l \leqq k)$ denote the column in H that is in the position that corresponds to position m_l in a code vector.

For a binary symmetric channel, the order in which symbols are transmitted can affect code performance. For the binary symmetric channel with independent errors, the order in which symbols are transmitted does not affect performance. In the latter case, we can write H as

$$H = (C_k, C_{k-1}, \cdots, C_1 I_{n-k}), \tag{2}$$

where I_{n-k} denotes the $(n - k) \times (n - k)$ identity matrix.

III. FORMULATION OF A CRITERION OF CODING EFFECTIVENESS

A system for transmitting observations performed upon some physical process over a binary channel is shown in Fig. 1. So that the relationship between the observed numbers and the code will be clear, a general formulation will be presented.

If each quantization step is of uniform size, the quantizer output can be represented as $A + Bi$ where A and B are constants and the integer i indicates the quantization level. The "source scale-to-binary converter" receives $A + Bi$ from the quantizer and transmits i to the encoder. The "binary-to-source scale converter" receives some integer j from the decoder and delivers $A + Bj$ to the destination.

* The weight of a vector v is the number of nonzero components in v and is denoted by $w[v]$. The distance between two vectors u and v is $w[u \oplus v]$.

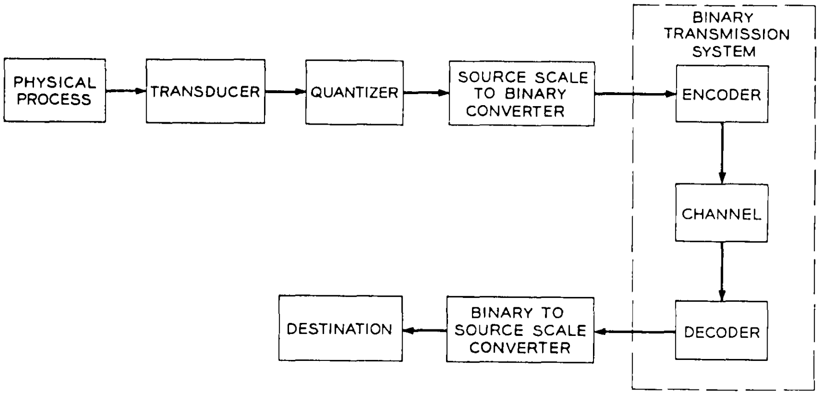

PHYSICAL PROCESS
TRANSDUCER
QUANTIZER
SOURCE SCALE TO BINARY CONVERTER
BINARY TRANSMISSION SYSTEM
ENCODER
CHANNEL
DECODER
BINARY TO SOURCE SCALE CONVERTER
DESTINATION

Fig. 1 — System model.

Let Pr $\{j \mid i\}$ be the probability of receiving j at the decoder output when i served as the encoder input and let Pr $\{i\}$ be the probability that i is sent. The average numerical error (ANE) that occurs is

$$\text{ANE} = \sum_i \sum_j \mid (A + Bj) - (A + Bi) \mid \Pr \{j \mid i\} \Pr \{i\}. \qquad (3)$$

If all values of i are equally likely to be observed and if the range for i is $0 \leqq i \leqq 2^k - 1$, Pr $\{i\} = 2^{-k}$. The range for j is thus $0 \leqq j \leqq 2^k - 1$ and (3) becomes

$$\text{ANE} = \frac{B}{2^k} \sum_{i=0}^{2^k-1} \sum_{j=0}^{2^k-1} \mid j - i \mid \Pr \{j \mid i\}.$$

Because B is a constant not dependent upon the particular coding scheme implemented, B may be set equal to 1 when comparing the effectiveness of different codes. Accordingly, we shall consider the expression

$$\text{ANE} = \frac{1}{2^k} \sum_{i=0}^{2^k-1} \sum_{j=0}^{2^k-1} \mid j - i \mid \Pr \{j \mid i\}. \qquad (4)$$

For a specified value of k, a given coding scheme is considered perferable to some other coding scheme if the ANE resulting from the implementation of the given code is less than the ANE resulting from the alternative code.

The code enters (4) through the terms Pr $\{j \mid i\}$. Thus, for a binary symmetric channel, the ANE will, in general, be dependent not only upon the error statistics of the channel but also upon the order in which the symbols are transmitted.

It is possible to simplify (4) to an expression that involves terms of the form Pr $\{j \mid 0\}$ exclusively. This reduces the number of terms by a factor of 2^k and demonstrates that knowledge of the error probabilities conditional upon zero being sent is sufficient to evaluate the ANE. However, it is necessary to develop some notation and to present two lemmas before proceeding to simplify (4). The proofs of the lemmas are omitted because the lemmas follow from the group property of the code.

When the integer i is to be sent, let us assume that the message ultilized is the k-bit binary representation of i (which is denoted by $B(i)$) such that

$$B(i) = (m_k , m_{k-1} , \cdots , m_1),$$

where

$$i = m_k \cdot 2^{k-1} + m_{k-1} \cdot 2^{k-2} + \cdots + m_1 .$$

The ending $E_i = (e_1 , e_2 , \cdots , e_{n-k})$ required to encode $B(i)$ is chosen so that the resulting code vector $C(i)$ satisfies (1).

Lemma 1: For any values of the integers i and j, $0 \leqq i \leqq 2^k - 1$ and $0 \leqq j \leqq 2^k - 1$, there exists an integer l such that Pr $\{j \mid i\} =$ Pr $\{l \mid 0\}$ where $B(l) = B(i) \oplus B(j)$ and $0 \leqq l \leqq 2^k - 1$.

Lemma 2: Let $B(l) = B(i) \oplus B(j)$ as in Lemma 1. For fixed $i(0 \leqq i \leqq 2^k - 1)$, as j successively takes on the values $0, 1, 2, \cdots , 2^k - 1$, l takes on each of the values in the range $0 \leqq l \leqq 2^k - 1$ once and only once.

Theorem 1: Let all messages be equally likely to be transmitted and let the channel be binary symmetric (but not necessarily with independent errors). For these conditions, the average numerical error is

$$\text{ANE} = \sum_{j=1}^{k} 2^{j-1} \sum_{i=2^{j-1}}^{2^j-1} \Pr \{i \mid 0\}. \tag{5}$$

Proof: By Lemmas 1 and 2, for each value of i and for a specified value of l, there will be a unique integer j_l such that Pr $\{j_l \mid i\} =$ Pr $\{l \mid 0\}$ where $B(l) = B(i) \oplus B(j_l)$. From (4),

$$\text{ANE} = \frac{1}{2^k} \sum_{l=1}^{2^k-1} \sum_{i=0}^{2^k-1} \mid j_l - i \mid \Pr \{l \mid 0\}, \tag{6}$$

where we have used the fact that $\mid j_l - i \mid = 0$ when $l = 0$.

For each value of l $(1 \leqq l \leqq 2^k - 1)$, we wish to determine $\sum_{i=0}^{2^k-1} \mid j_l - i \mid$. Let $\alpha(0 \leqq \alpha \leqq k - 1)$ be the largest integer such that

$2^\alpha \leqq l$. Define i' and j_l' as

$$B(j_l) = B(i') \oplus B(2^\alpha) \tag{7a}$$

$$B(i) = B(j_l') \oplus B(2^\alpha) \quad \text{or} \quad B(j_l') = B(i) \oplus B(2^\alpha). \tag{7b}$$

Then

$$B(i) \oplus B(j_l) = B(i') \oplus B(j_l') = B(l). \tag{7c}$$

Because $l > 0$, $i \neq j_l$ and $i' \neq j_l'$. Suppose $i > j_l$. Then $j_l = i' - 2^\alpha$ and $j_l' = i - 2^\alpha$ by (7). It follows that $j_l - i = -2^\alpha - i + i'$ and $j_l' - i' = -2^\alpha + i - i'$. Conversely, if $i < j_l$, $j_l = i' + 2^\alpha$ and $j_l' = i + 2^\alpha$. Thus, $j_l - i = 2^\alpha - i + i'$ and $j_l' - i' = 2^\alpha + i - i'$.

Therefore,

$$|\, j_l - i \,| + |\, j_l' - i' \,| = |\, 2^\alpha + i - i' \,| + |\, 2^\alpha - i + i' \,|. \tag{8}$$

But $B(i) = B(2^\alpha) \oplus B(l) \oplus B(i')$ by (7). Thus, $|\, i - i' \,| < 2^\alpha$ and, from (8),

$$|\, j_l - i \,| + |\, j_l' - i' \,| = 2 \cdot 2^\alpha.$$

Because of the symmetries involved,

$$2 \sum_{i=0}^{2^k-1} |\, j_l - i \,| = \sum_{i=0}^{2^k-1} |\, j_l - i \,| + |\, j_l' - i' \,| = 2^k \cdot 2^{\alpha+1}.$$

Thus, (6) becomes

$$\text{ANE} = \sum_{l=1}^{2^k-1} 2^\alpha \Pr\{l \mid 0\}$$

or

$$\text{ANE} = \sum_{\alpha=0}^{k-1} 2^\alpha \sum_{l=2^\alpha}^{2^{\alpha+1}-1} \Pr\{l \mid 0\}. \qquad \text{QED}$$

In (5), notice that $\Pr\{0 \mid 0\}$ does not appear and that the terms $\Pr\{i \mid 0\}$ are not weighted linearly in i but that the weighting coefficients go in steps as powers of 2 with several conditional probabilities having the same weighting coefficient. Notice that the weighting coefficient for $\Pr\{i \mid 0\}$ is 2^{j-1} where $(j - 1)$ is the largest power of 2 in i. All errors with the same coefficient are of the same seriousness and a good code must reduce these sets of probabilities rather than simply minimize the probability that a few very large errors occur.

Because the set of messages $B(i)$ $(2^{j-1} \leqq i \leqq 2^j - 1)$ gives rise to the set of conditional probabilities whose weighting coefficient in the ANE expression is 2^{j-1}, we shall call these messages the j-level messages and the corresponding conditional probabilities, Pr $\{2^{j-1} \mid 0\}$ through Pr $\{2^j - 1 \mid 0\}$, the j-level conditional probabilities. The 0-level message is defined to be $B(0)$ and the 0-level conditional probability to be Pr $\{0 \mid 0\}$.

The j-level messages have the following interesting characteristics.

(*i*) Component m_j in each message is 1.

(*ii*) Components $m_i (j + 1 \leqq i \leqq k)$ in each message are 0.

(*iii*) Every possible $(j - 1)$-tuple occurs once and only once as components m_1 through m_{j-1} of some j-level message.

For a perfect error-correcting code used with a binary symmetric channel with independent errors, it is possible to compute the j-level conditional probabilities and thus the ANE from a knowledge of the weight distribution of the code vectors on each level (these weight distributions have been referred to as level weight structures.)[1] The problem of efficiently computing the level weight structures from knowledge of the parity check matrix has been discussed previously.[1]

IV. SIGNIFICANT-BIT CODES

In order to permit the error-correcting capabilities of a code to correspond somewhat to the significance of the information positions, it is possible to formulate a type of code which uses a subcode to protect the $(k - k_0)$ most significant positions of a message and simply transmits the remaining symbols unprotected. The name significant-bit code (SB code) is used for this type of code. An SB code is specified by the parity check matrix H_{SB} and the ANE resulting from the use of an SB code is ANE_{SB} .

The code utilized to protect the $(k - k_0)$ most significant information positions will be named the base code. Because it is confined to the $(k - k_0)$ most significant positions, we can abstract the base code and study it as a separate entity. Accordingly, the base code vectors are $(n - k_0)$-tuples of which the first $(k - k_0)$ positions are the base messages.

Although the concept of SB codes is applicable to any binary symmetric channel, we shall assume independent errors in the following analysis. Thus, from (2), the base code is specified by the base parity check matrix H_{B} where

$$H_{\text{B}} = (C'_{k-k_0} , C'_{k-k_0-1} , \cdots , C'_1 I_{n-k}).$$

In this case, the code vector $C(i) = B(i) \mid E_i$ where the symbol $\mid$ indicates that $C(i)$ can be partitioned into the k-tuple $B(i)$ and the

$(n - k)$-tuple E_i. Let $B(i)$ be partitioned so that $B(i) = B'(i') \mid B''(i'')$ where $B'(i')$ denotes the $(k - k_0)$ most significant positions of $B(i)$ and $B''(i'')$ denotes the k_0 least significant positions of $B(i)$. Then

$$C(i) = B'(i') \mid B''(i'') \mid E_i .$$

The range for i' is $0 \leqq i' \leqq 2^{k-k_0} - 1$ and for i'' is $0 \leqq i'' \leqq 2^{k_0} - 1$.

Let $\mathrm{Pr_B}\,\{i' \mid j'\}$ denote the probability of receiving i' when j' is sent using the base code. By Theorem 1, the ANE for the base code ($\mathrm{ANE_B}$) is

$$\mathrm{ANE_B} = \sum_{j=1}^{k-k_0} 2^{j-1} \sum_{i'=2^{j-1}}^{2^j-1} \mathrm{Pr_B}\,\{i' \mid 0\}. \tag{9}$$

Because the base code is used exclusively to protect the $(k - k_0)$ most significant information positions, H_{SB} must have the form

$$H_{\mathrm{SB}} = (C'_{k-k_0}\,,\ C'_{k-k_0-1}\,,\ \cdots\,,\ C'_2\,,\ C'_1\ \underbrace{0 \cdots 0}_{\substack{k_0\\ \text{columns}}}\ I_{n-k})$$

where 0 is used to represent an all-zero column of H_{SB} and where the $C'_l (1 \leqq l \leqq k - k_0)$ are the columns of H_{B}. The coset leaders[2] in the standard array[2] for the SB code must be obtained from the coset leaders in the standard array for the base code by expanding the base coset leaders in length to n-tuples by inserting k_0 zeros in information positions 1 through k_0 of the expanded vectors. Because all vectors in column i of the standard array for the SB code will have $B''(i'')$ in information positions 1 through k_0,

$$\Pr\,\{i \mid 0\} = p^{w[B''(i'')]} q^{k_0 - w[B''(i'')]}\, \mathrm{Pr_B}\,\{i' \mid 0\}. \tag{10}$$

We shall now show that $\mathrm{ANE_{SB}}$ can be expressed in terms of the properties of the base code.

Theorem 2: Let the base code be defined as above. For a binary symmetric channel with independent errors and when all messages are equally likely to be transmitted,

$$\mathrm{ANE_{SB}} = \mathrm{Pr_B}\,\{0 \mid 0\} \sum_{j=1}^{k_0} 2^{j-1} p q^{k_0-j} + 2^{k_0}\,\mathrm{ANE_B} . \tag{11}$$

Proof: Define

$$\mathrm{ANE}' = \sum_{j=1}^{k_0} 2^{j-1} \sum_{i=2^{j-1}}^{2^j-1} \Pr\,\{i \mid 0\}$$

and

$$\mathrm{ANE}'' = \sum_{j=k_0+1}^{k} 2^{j-1} \sum_{i=2^{j-1}}^{2^j-1} \Pr\{i \mid 0\}.$$

From Theorem 1, $\mathrm{ANE}_{\mathrm{SB}} = \mathrm{ANE}' + \mathrm{ANE}''$.

Let us first analyze ANE′. For $1 \leqq j \leqq k_0$, the sum of the j-level conditional probabilities is

$$\sum_{i=2^{j-1}}^{2^j-1} \Pr\{i \mid 0\} = \sum_{i''=2^{j-1}}^{2^j-1} p^{w[B''(i'')]} q^{k_0-w[B''(i'')]} \Pr_{\mathrm{B}}\{0 \mid 0\}$$

where we have used (10) and realized that $i' = 0$ for all messages on this level. Because every $(j-1)$-tuple occurs as components m_1 through m_{j-1} of some j-level message and $m_j = 1$ in every j-level message, there are

$$\binom{j-1}{w[B''(i'')]-1}$$

messages of weight $w[B''(i'')]$ on the j-level. Thus,

$$\begin{aligned}\sum_{i=2^{j-1}}^{2^j-1} \Pr\{i \mid 0\} &= \Pr_{\mathrm{B}}\{0 \mid 0\} \sum_{l=1}^{j} \binom{j-1}{l-1} p^l q^{k_0-l} \\ &= \Pr_{\mathrm{B}}\{0 \mid 0\} p q^{k_0-j}\end{aligned}$$

and

$$\mathrm{ANE}' = \Pr_{\mathrm{B}}\{0 \mid 0\} \sum_{j=1}^{k_0} 2^{j-1} p q^{k_0-j}.$$

Now consider ANE″. On level $k_0 + \xi$ $(1 \leqq \xi \leqq k - k_0)$, i has the range $2^{k_0+\xi-1} \leqq i \leqq 2^{k_0+\xi} - 1$. Divide this range into $2^{\xi-1}$ sets of consecutive integers each of size 2^{k_0}. Let the integer δ index these sets where $0 \leqq \delta \leqq 2^{\xi-1} - 1$. For a particular value of δ, as i increases from $2^{k_0+\xi-1} + \delta 2^{k_0}$ to $2^{k_0+\xi-1} + (\delta + 1)2^{k_0} - 1$, $i' = 2^{\xi-1} + \delta$ and i'' runs through the range $0 \leqq i'' \leqq 2^{k_0} - 1$. Thus, using (10),

$$\begin{aligned}&\sum_{i=2^{k_0+\xi-1}+\delta 2^{k_0}}^{2^{k_0+\xi-1}+(\delta+1)2^{k_0}-1} \Pr\{i \mid 0\} \\ &\qquad = \sum_{i''=0}^{2^{k_0}-1} p^{w[B''(i'')]} q^{k_0-w[B''(i'')]} \Pr_{\mathrm{B}}\{2^{\xi-1} + \delta \mid 0\}.\end{aligned}$$

As i'' runs through the range $0 \leqq i'' \leqq 2^{k_0} - 1$, each possible k_0-tuple occurs once and only once. Therefore,

$$\begin{aligned}\sum_{i=2^{k_0+\xi-1}+\delta 2^{k_0}}^{2^{k_0+\xi-1}+(\delta+1)2^{k_0}-1} \Pr\{i \mid 0\} &= \Pr\nolimits_{\mathrm{B}}\{2^{\xi-1}+\delta \mid 0\} \sum_{l=0}^{k_0} \binom{k_0}{l} p^l q^{k_0-l} \\ &= \Pr\nolimits_{\mathrm{B}}\{2^{\xi-1}+\delta \mid 0\}.\end{aligned} \tag{12}$$

Because of the manner in which the sets were chosen, ANE″ can be expanded as

$$\mathrm{ANE}'' = \sum_{\xi=1}^{k-k_0} 2^{k_0+\xi-1} \sum_{\delta=0}^{2^{\xi-1}-1} \sum_{i=2^{k_0+\xi-1}+\delta 2^{k_0}}^{2^{k_0+\xi-1}+(\delta+1)2^{k_0}-1} \Pr\{i \mid 0\}. \tag{13}$$

Substituting (12) into (13), we obtain

$$\mathrm{ANE}'' = 2^{k_0} \sum_{\xi=1}^{k-k_0} 2^{\xi-1} \sum_{i'=2^{\xi-1}}^{2^{\xi}-1} \Pr\nolimits_{\mathrm{B}}\{i' \mid 0\}$$

which, from (9), is exactly 2^{k_0} $\mathrm{ANE}_{\mathrm{B}}$. QED

Notice that the situation $k = k_0$ can be included in this formulation if we define $\mathrm{ANE}_{\mathrm{B}} = 0$ and $\Pr_{\mathrm{B}}\{0 \mid 0\} = 1$ when $k = k_0$. Thus, uncoded transmission can be regarded as an SB code in which $k = k_0$.

The interpretation of (11) is interesting. The quantity $\sum_{i=1}^{k_0} 2^{i-1} p q^{k_0-i}$ is the ANE that results from the uncoded transmission of k_0-tuples. Thus, $\mathrm{ANE}_{\mathrm{SB}}$ is the ANE for uncoded transmission of k_0-tuples weighted by $\Pr_{\mathrm{B}}\{0 \mid 0\}$ plus 2^{k_0} times $\mathrm{ANE}_{\mathrm{B}}$.

(11) enables the computation of $\mathrm{ANE}_{\mathrm{SB}}$ from the properties of the base code. Because the base code involves messages of length $(k - k_0)$, it is easier to analyze than the entire SB code.

V. CONSTANT-SYMBOL-RATE TRANSMISSION

Consider two error-correcting codes which are denoted as V_1 and V_2. Let V_1 be an (n_1, k) code and V_2 be an (n_2, k) code where n_1 may or may not be equal to n_2. Let ε_1 denote the minimum weight of the n_1-tuples that are not coset leaders in the standard array for V_1. Similarly, let ε_2 denote the minimum weight of the n_2-tuples that are not coset leaders in the standard array for V_2.

For a binary symmetric channel with independent errors, $\Pr\{i \mid 0\}$ for V_1 is

$$\Pr\{i \mid 0\} = \sum_{j=\varepsilon_1}^{n_1} \tau_{ij} p^j q^{n_1-j}$$

where τ_{ij} is the number of n_1-tuples of weight j in the column headed by $C(i)$ in the standard array for V_1. Thus, for V_1, the average numerical error (ANE_1) is

$$\mathrm{ANE}_1 = \sum_{j=\varepsilon_1}^{n_1} \sigma_j p^j q^{n_1-j},$$

where

$$\sigma_j = \sum_{l=1}^{k} 2^{l-1} \sum_{i=2^{l-1}}^{2^l-1} \tau_{ij} .$$

Similarly, for V_2,

$$\mathrm{ANE}_2 = \sum_{j=\varepsilon_2}^{n_2} \gamma_j p^j q^{n_2-j},$$

where the γ_j are the appropriate constants.

However,

$$\mathrm{ANE}_1 \rightarrow \sigma_{\varepsilon_1} p^{\varepsilon_1} q^{n_1-\varepsilon_1} \quad \text{as} \quad p \rightarrow 0$$

and

$$\mathrm{ANE}_2 \rightarrow \gamma_{\varepsilon_2} p^{\varepsilon_2} q^{n_2-\varepsilon_2} \quad \text{as} \quad p \rightarrow 0.$$

Thus, for p sufficiently small, if $\varepsilon_1 > \varepsilon_2$, $\mathrm{ANE}_1 < \mathrm{ANE}_2$ and V_1 results in less ANE than V_2.

The minimum weight of the vectors that are not coset leaders in an SB code is 1. Thus, consider two SB codes denoted by V_{SB1} and V_{SB2} where V_{SB1} is an (n_1, k) code and V_{SB2} is an (n_2, k) code. V_{SB1} protects the $(k - k_{01})$ most significant positions and V_{SB2} protects the $(k - k_{02})$ most significant positions of a message. By reasoning analogous to that above, for p small, if $k_{01} < k_{02}$ and if the base codes used in V_{SB1} and V_{SB2} correct all weight one errors, then V_{SB1} results in less ANE than V_{SB2}.

We thus have the following ranking of codes for p small. The ranking (in order of increasing effectiveness) assumes that the schemes are compared for the same value of k.

(*i*) Uncoded transmission.
(*ii*) An SB code protecting $(k - k_0)$ positions where $k \neq k_0$.
(*iii*) An SB code protecting $(k - k_0 + k')$ positions where $k' > 0$.
(*iv*) An e-error-correcting code where $e \geqq 1$.
(*v*) An $(e + e')$-error-correcting code where $e' > 0$.

To obtain a feeling for the utility of coding for numerical data transmission over a binary symmetric channel with independent errors, the ANE resulting from certain codes for $k = 26$ will be evaluated for constant-symbol-rate transmission. Ref. 3 contains similar information for $k = 1, 4$ and 11.

Let ANE_{UC} denote the ANE when no coding is used. Contrary to the concept of code equivalence that is obtained under the assumption that all errors are equally costly (i.e., when probability of message error is used as the measure of code performance), the ordering of the columns of the parity check matrix can affect code performance. Thus, for the (31, 26) perfect single error-correcting code (PSEC code), every ordering of the columns of the parity check matrix could yield a distinct ANE. Upper and lower bounds on the ANE for this code are obtained in Ref. 3 and are denoted herein as ANE_{UB} and ANE_{LB}, respectively.

By numerical computation, the ordering in (14) was found to result in as small an ANE as any other ordering tried. The number actually tried was by necessity a small fraction of all possible orderings of the 26 columns. However, notice that C_{12} through C_{26} each have a one in the same position thus assuring us that the number of weight three code vectors on levels 12 through 26 will be the theoretical minimum for this code (by Theorem 9 in Ref. 3). For values of p that are of primary interest (less than 10^{-1}), this assures us that it is not possible to find a different ordering that will result in a significantly better performance (although there are other orderings that in fact give equal performance). Let ANE_{P} denote the ANE that results from the code specified in (14).

$$H_{\text{P}} = \left[\begin{array}{cccccccccccccccccccccccccccc}
1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0& \\
1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&0&0&0&0& \\
1&1&1&1&0&0&0&0&1&1&1&1&0&0&0&1&1&1&1&0&0&0&1&1&1&0&I_5 \\
1&1&0&0&1&1&0&0&1&1&0&0&1&1&0&1&1&0&0&1&1&0&1&1&0&1& \\
1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&1&0&1&0&1&0&1&1&0&1&1&
\end{array}\right]. \qquad (14)$$

If the columns of (14) are regarded as the 5-bit binary representations of integers, then the ordering from left to right corresponds to decreasing integer value (with powers of two omitted because they appear in I_5). Similar ordering was observed to be preferable for the (15, 11) PSEC code[3] and, by exhaustive search, actually found to be as good as any other ordering for the (7, 4) PSEC code[3].

Table I compares ANE_{LB}, ANE_{UB} and ANE_{P}. For convenience (and so that the values given will agree with the data plotted in Figs. 2, 3, and 4), the ANE has been normalized by dividing by $2^{26} - 1$ (i.e., the full-scale value).

The following SB codes are considered. For each, H_{B} and the notation used for the resulting ANE in Figs. 2, 3, and 4 is given. Theorem 2 permits the computation of the ANE for these codes from a knowledge of the base code.

TABLE I—VALUES OF $\mathrm{ANE}_{\mathrm{LB}}$, $\mathrm{ANE}_{\mathrm{UB}}$ AND $\mathrm{ANE}_{\mathrm{P}}$ DIVIDED BY $2^{26} - 1$

p	$\mathrm{ANE}_{\mathrm{LB}}$	$\mathrm{ANE}_{\mathrm{P}}$	$\mathrm{ANE}_{\mathrm{UB}}$
10^{-5}	$0.41992 \cdot 10^{-8}$	$0.41994 \cdot 10^{-8}$	$0.42993 \cdot 10^{-8}$
10^{-4}	$0.41920 \cdot 10^{-6}$	$0.41931 \cdot 10^{-6}$	$0.42932 \cdot 10^{-6}$
10^{-3}	$0.41212 \cdot 10^{-4}$	$0.41310 \cdot 10^{-4}$	$0.42329 \cdot 10^{-4}$
10^{-2}	$0.34850 \cdot 10^{-2}$	$0.35659 \cdot 10^{-2}$	$0.36894 \cdot 10^{-2}$
10^{-1}	$0.90222 \cdot 10^{-1}$	0.10446	0.12817

Base Code 1: (3, 1) PSEC code.

$$H_B = \begin{bmatrix} \begin{matrix} 1 \\ 1 \end{matrix} & I_2 \end{bmatrix}$$

The ANE is denoted as $\mathrm{ANE}_{(3,1)}$.

Base Code 2: (5, 1) perfect double error-correcting code.

$$H_B = \begin{bmatrix} \begin{matrix} 1 \\ 1 \\ 1 \\ 1 \end{matrix} & I_4 \end{bmatrix}$$

The ANE is denoted as $\mathrm{ANE}_{(5,1)}$.

Base Code 3: This base code uses independent (3, 1) PSEC codes to protect the two most significant information positions.

$$H_B = \begin{bmatrix} \begin{matrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{matrix} & I_4 \end{bmatrix}$$

Because the codes are used independently, the required conditional probabilities for the base code can be readily calculated. The ANE is denoted as $\mathrm{ANE}_{(3,1),(3,1)}$.

Base Code 4: (7, 4) PSEC code.

$$H_B = \begin{bmatrix} \begin{matrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{matrix} & I_3 \end{bmatrix}$$

The ANE is denoted as $\mathrm{ANE}_{(7,4)}$.

Base Code 5: This base code uses a (3, 1) PSEC code to protect the most significant information position and a (7, 4) PSEC code to protect the next four most significant information positions.

$$H_{\mathrm{B}} = \left[\begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & \\ 1 & 0 & 0 & 0 & 0 & \\ 0 & 1 & 1 & 1 & 0 & I_5 \\ 0 & 1 & 1 & 0 & 1 & \\ 0 & 1 & 0 & 1 & 1 & \end{array}\right]$$

The ANE is denoted as $\mathrm{ANE}_{(3,1),(7,4)}$.

Base Code 6: (15, 11) PSEC code.

$$H_{\mathrm{B}} = \left[\begin{array}{cccccccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & I_4 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & \end{array}\right]$$

The ANE is denoted as $\mathrm{ANE}_{(15,11)}$.

Figs. 2, 3, and 4 present $\mathrm{ANE}_{\mathrm{UC}}$, $\mathrm{ANE}_{\mathrm{UB}}$, $\mathrm{ANE}_{\mathrm{LB}}$, $\mathrm{ANE}_{\mathrm{P}}$, and the ANE of the SB codes considered. In each case, the ANE has been normalized by dividing by $2^{26} - 1$. For clarity, logarithmic scales are used as p decreases from 10^{-1} until p becomes sufficiently small so that the results for small p apply.

The following observations can be made for constant-symbol-rate transmission.

(*i*) Improvements in transmission fidelity are obtainable by the utilization of codes. It should be noted that no one code is the most desirable for all p $(0 < p < \frac{1}{2})$ and in some cases the codes that are best for small p turn out to be less effective than uncoded transmission for the larger values of p.

(*ii*) For $k = 26$, it can be shown that the probability that a message is received in error when the PSEC code is used is less (for $0 < p < \frac{1}{2}$) than the probability that a message is received in error using any of the SB codes considered. Thus, under the criterion of minimizing the probability that a message is received in error, the PSEC code is preferable to any of the SB codes considered.

However, when the ANE is used as a measure of code effectiveness for numerical data transmission, we observe that the SB codes are preferable to the PSEC code for certain values of p. Thus, when comparing codes, the ranking obtained using probability of message error as the performance index may not correspond to the ranking obtained using ANE as an index. We can conclude that probability of message error and ANE are not equivalent measures of code performance and

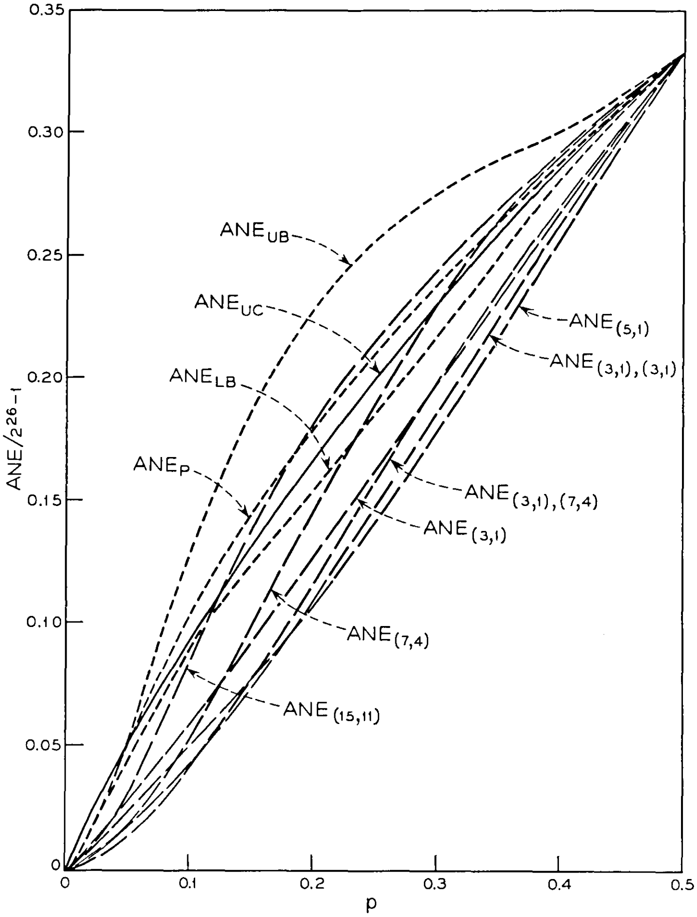


Fig. 2 — Constant-symbol-rate transmission; $k = 26$.

that, in some cases, the ANE can be reduced by using a code whose probability of message error is not minimal.

(*iii*) For $k = 26$, consider the relative performance of the PSEC code and the SB codes. When p is small, the PSEC code will be effective because it can correct all single errors (the only type that have much probability of occurring) whereas a single error in certain positions of an SB code will result in a message error. For larger values of p, there is an increasing chance that an error pattern will occur which the PSEC code cannot correct. The SB codes become effective in this situation. If multiple errors occur during transmission such that the errors occurring in the $(k - k_0)$ most significant information positions and the check positions form an error pattern correctable by the base code, this will be corrected leaving any errors in the k_0 least significant information positions uncorrected. Therefore, the most costly portion of a large number of error patterns can be corrected. As p increases, the number of positions in the base code must decrease so that uncorrectable error patterns in the positions covered by the base code have a sufficiently small probability of occurrence so that the base code can operate effectively. In other words, as p increases, more and more protection must be provided for the significant bits so that the most costly errors are prevented.

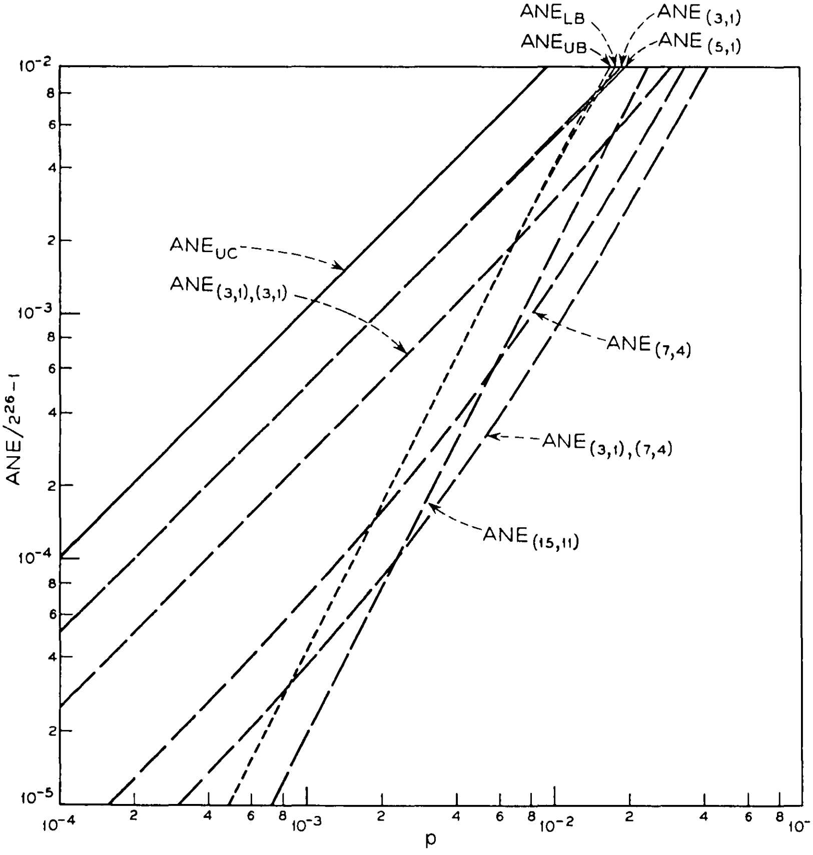


Fig. 3 — Constant-symbol-rate transmission; $k = 26$.

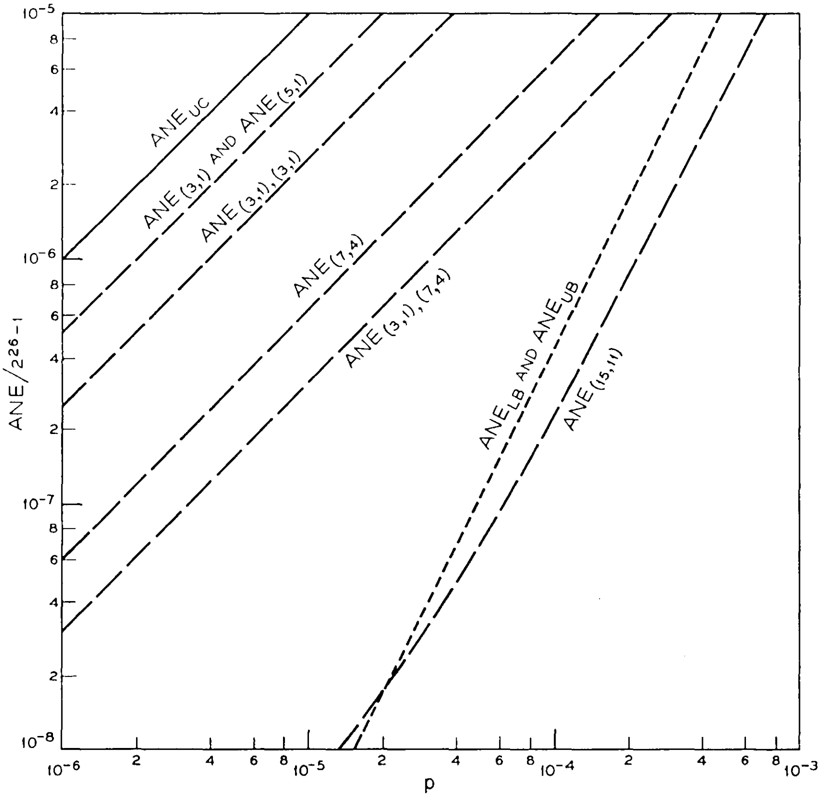


Fig. 4 — Constant-symbol-rate transmission; $k = 26$.

(*iv*) For p small, the ANE from uncoded transmission is approximately $(2^k - 1)p$. For small p, the ANE as a fraction of full scale for uncoded transmission is thus very nearly independent of k.

REFERENCES

1. Buchner, M. M., Jr., Computing the Spectrum of a Binary Group Code, B.S.T.J., *45*, March, 1966, pp. 441–449.
2. Slepian, D., A Class of Binary Signaling Alphabets, B.S.T.J., *35*, January, 1956, pp. 203–234.
3. Buchner, M. M., Jr., Coding for Numerical Data Transmission, Ph.D. Dissertation, The Johns Hopkins University, Baltimore, Maryland, 1965.

Contributors to This Issue

VÁCLAV E. BENEŠ, A.B., 1950, Harvard College; M.A. and Ph.D., 1953, Princeton University; Bell Telephone Laboratories, 1953—. Mr. Beneš has been engaged in mathematical research on stochastic processes, traffic theory, and servomechanisms. In 1959–60 he was visiting lecturer in mathematics at Dartmouth College. He is the author of General Stochastic Process in the Theory of Queues (Addison-Wesley, 1963), and of Mathematical Theory of Connecting Networks and Telephone Traffic (Academic Press, 1965). Member, American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, SIAM, Mind Association, Phi Beta Kappa.

MORGAN M. BUCHNER, JR., B.E.S., 1961, Ph.D., 1965, The Johns Hopkins University; Bell Telephone Laboratories, 1965—. At Bell Telephone Laboratories, Mr. Buchner was engaged in a study of impulse noise in an effort to understand its characteristics and its effects upon data communications. At present, he is on a military leave of absence and is serving as a Lieutenant in the U. S. Army Electronics Command, Fort Monmouth, N. J. Member, IEEE, Tau Beta Pi, Sigma Xi, Eta Kappa Nu.

MYRON S. GLASS, M.S. in Physics, University of Chicago, 1926; Bell Telephone Laboratories, 1926—. Mr. Glass has been engaged in the development of electron devices and in applied magnetics and optics. He is currently supervisor of a group in the Optical Device Department engaged in the development of gas lasers and laser mirrors. Member, AAAS; senior member, IEEE.

IRA JACOBS, B.S. Physics, 1950, City College of New York; M.S., 1952, Ph.D., 1955, Purdue University; Bell Telephone Laboratories, 1955—. Mr. Jacobs has been engaged in studies of electromagnetic wave propagation in nonuniform and anisotropic media, radar cross-section and antenna analyses, and in missile guidance and detection systems. He is currently Head of the Military Communications Research Department in the Detection Systems Laboratory. His current activities are largely in the field of communication theory. During the summers of 1964 and 1966 he was a member of Institute for Defense Analyses study groups considering satellite multiple access and signal processing. He has served as Project Engineer on a study of weak-signal communication techniques and is presently Project Engineer on a deep-space communication study. Senior member, IEEE; member, American Physical Society, American Association for the Advancement of Science, Phi Beta Kappa, Sigma Xi, Sigma Pi Sigma.

T. T. KADOTA, B.S., 1953, Yokahama National University (Japan); M.S., 1956, Ph.D., 1960, University of California (Berkeley); Bell Telephone Laboratories, 1960—. Mr. Kadota has been engaged in the study of noise theory with application to optimum detection theory. Member, Sigma Xi.

SUNDARAM NARAYANAN, B. Tech., 1960, Indian Institute of Technology, Kharagpur (India); M.S., 1963, Ph.D., 1965, Carnegie Institute of Technology; Bell Telephone Laboratories, 1965—. Mr. Narayanan is with the Coaxial Systems Studies group and is primarily concerned with nonlinear distortion mechanism in transistors and in transistor feedback amplifiers. Member, Sigma Xi.

F. N. H. ROBINSON, B.A., Cambridge, 1946; M.A., Oxford, 1950; D. Phil., Oxford, 1955. At Oxford since 1950, where he teaches physics, Mr. Robinson's research has mostly been in the fields of very low temperature and nuclear orientation. As a visitor to Bell Telephone Laboratories in 1954–55, he worked on noise in electron beams. Since then he has been a frequent visitor to the Laboratories and spent the year 1965–66 there when he worked on non-linear optics.

W. SHOCKLEY, B.Sc., 1932, California Institute of Technology; Ph.D., 1936, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1936–1955, 1965—. Dr. Shockley is best known as the inventor of the junction transistor. For this and other contributions, to transistor physics, he received the 1956 Nobel Prize in Physics jointly with his two former colleagues at Bell Telephone Laboratories, John Bardeen and Walter H. Brattain. During World War II, on leave of absence from Bell, he served as Director of Research for the Navy's Anti-Submarine Warfare Operations Research Group and as expert consultant for the Office of the Secretary of War. He returned to Bell Laboratories after the war and became director of the solid state physics research program. In 1953, he was named Director of Transistor Physics Research. During this period he made many contributions to solid-state physics particularly in connection with the transistor. In 1955 he left Bell Telephone Laboratories to join Beckman Instruments Inc. where he established the Shockley Semiconductor Laboratory in Palo Alto, California, for research, development and production of new transistor and other semiconductor devices. In 1965 Dr. Shockley returned to Bell Telephone Laboratories in the capacity of Executive Consultant. He presently holds the position of Alexander M. Poniatoff Professor of Engineering Science at Stanford University. More than 70 United States Patents have been granted for his inventions. Medal for Merit, Office of the Secretary of War, 1946; Air Force Association Citation of Honor, 1951; Morris Liebmann Memorial Prize, IRE, 1952; Oliver E. Buckley Solid State Physics Prize, American Physical Society, 1953; U. S. Army Certificate of Appreciation, 1953; Comstock Award, National Academy of Sciences, 1954; Holley Medal, American Society of Mechanical Engineers, 1963; Wilhelm-Exner Medal, Oesterreichischer Gewerbeverein of Austria, 1963. Honorary Doctorates from the University of Pennsylvania, 1955; Rutgers University, 1956; and Gustavus Adolphus College, 1963. Consultant, Scientific Advisory Panel of U. S. Army, Air Force Scientific Advisory Board; Fellow, IEEE, American Physical Society, American Academy of Arts and Sciences; Member, President's Science Advisory Committee Panel on Scientific and Technical Manpower, American Institute of Physics, Sigma Xi, Tau Beta Pi.

Simon M. Sze, B.S., 1957, National Taiwan University, Taiwan, China; M.S., 1960, University of Washington; Ph.D., 1963, Stanford University; Bell Telephone Laboratories, 1963—. Mr. Sze has been concerned with the study of semiconductor device physics. At present he is engaged in studies of metal-insulator-semiconductor devices and interface states. Member, Sigma Xi, IEEE.

William J. Tabor, B.S. (Chemistry), 1953, Rensselaer Polytechnic Institute; A.M. (Physics), 1954, Ph.D. (Chemical Physics), 1957, Harvard University; U. S. Army 1957–1959; Bell Telephone Laboratories, 1959—. His work at Bell Laboratories included research and development of microwave masers, the design of the maser for the *Telstar*® ground station, and investigation of light deflection techniques. He is currently involved in a study of domain wall motion in magnetic media.

B.S.T.J. BRIEFS

Realizability Conditions for the Impedance Function of the Lossless Tapered Transmission Line—A Critique

By E. N. PROTONOTARIOS

(Manuscript received March 6, 1967)

I. INTRODUCTION

In a recent brief[1] in the B.S.T.J., Zador presents, without proof, realizability conditions for the input impedance of the lossless tapered transmission line terminated in unit resistance. Upon a careful examination of the brief, it appears that the conditions are not accurate. The following analysis clarifies this point and, incidentally, provides alternatives to Zador's necessary conditions.

Consider a nonuniform line (Fig. 1) with inductance per unit length $\mathcal{L}(x)$ and capacitance per unit length $\mathcal{C}(x)$ such that (to follow Zador)

$$\mathcal{L}(x)\mathcal{C}(x) = 1.$$

Let $V(x,s)$ and $I(x,s)$ be the voltage and current along the line with polarities as indicated in Fig. 1. The equations of the line are

$$\frac{dV(x,s)}{dx} = -s\mathcal{L}(x)I(x,s)$$

$$\frac{dI(x,s)}{dx} = -s\mathcal{C}(x)V(x,s).$$

Eliminating $I(x,s)$ and taking into account that $\mathcal{L}(x) = 1/\mathcal{C}(x)$ we get

$$\frac{d}{dx}\left(\mathcal{C}(x)\,\frac{dV(x,s)}{dx}\right) = s^2\mathcal{C}(x)\,V(x,s).$$

Note also that

$$I(x,s) = -\frac{\mathcal{C}(x)}{s}\,\frac{dV(x,s)}{dx}.$$

Hence, we can identify Zador's $y(x,s)$ and $c(x)$ with $V(x,s)$ and $\mathcal{C}(x)$, respectively. From the reference polarities of the voltages and currents in Fig. 1, we see that for a unit resistance termination at $x = 0$ we must

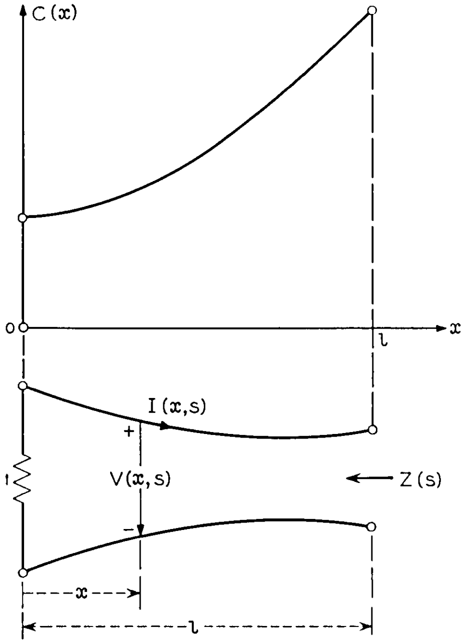


Fig. 1.—Lossless tapered transmission line.

have

$$V(0,s) = -I(0,s).$$

Hence, if we impose the condition (following Zador)

$$y(0,s) = V(0,s) = -i$$

then for unit resistance termination we should have

$$y'(0,s) = \frac{dV(0,s)}{dx} = \frac{sV(0,s)}{c(0)} = -\frac{is}{c(0)}.$$

The driving point impedance, for any termination, should read

$$Z(s) = \frac{s}{c(l)} \frac{y(l,s)}{y'(l,s)}.$$

Thus, the signs are wrong in Ref. 1. This is not the crucial error however.

In this paper, we will show that the difficulties in Zador's paper arise from the following facts:

(*i*) He does not consider the matched line. Unmatched lines tend to have almost periodic behavior for large real frequencies and hence the network functions do not have limits at infinity. This point will be made more precise in the sequel.

(*ii*) Multiplication of $Z(j\omega)$ by $\exp(-2jl\omega)$ in property (*iii*) of the necessity statement introduces periodic behavior at infinity *even in the matched case.*

(*iii*) Physical meaning has not been attached to the N_i and D_i. These should obviously be identified with the well-known ABCD parameters to correct (*ii*) of the necessity conditions.

II. COMMENTS ON ZADOR'S BRIEF

Property (*iii*) in the necessity statement does not appear to be true as stated. One can easily construct many counter examples.

Example 1: The uniform line with (following Zador's notation) $c(x) = 1$ and length $l = 1$, terminated in a 1-ohm resistor. Obviously $c(x)$ satisfies the conditions stipulated by Zador, i.e., $c(x)$ is positive and continuously differentiable in the interval $0 \leqq x \leqq 1$. Clearly the driving point impedance is

$$Z(j\omega) = 1.$$

Therefore,

$$f(\omega) = \operatorname{Re} \exp(-2jl\omega)Z(j\omega) = \cos 2\omega.$$

Clearly $\cos 2\omega$ does not have a limit for $\omega \rightarrow \pm\infty$.

Consider now a less trivial counter example.

Example 2: The exponential line terminated in a unit resistance. With Zador's notation $c(x) = \exp 2x$, and $l = 1$. In this case by solving Zador's (1) with the subsequent boundary conditions (appropriately corrected) we find

$$Z(j\omega) = \frac{A(j\omega) + B(j\omega)}{C(j\omega) + D(j\omega)}, \tag{1}$$

where

$$A(j\omega) = \frac{1}{e}\left\{\cos\sqrt{\omega^2 - 1} + \frac{\sin\sqrt{\omega^2 - 1}}{\sqrt{\omega^2 - 1}}\right\} \tag{2}$$

$$B(j\omega) = \frac{1}{e}\left\{j\omega\frac{\sin\sqrt{\omega^2 - 1}}{\sqrt{\omega^2 - 1}}\right\} \tag{3}$$

$$C(j\omega) = e\left\{j\omega\frac{\sin\sqrt{\omega^2 - 1}}{\sqrt{\omega^2 - 1}}\right\} \tag{4}$$

$$D(j\omega) = e\left\{\cos\sqrt{\omega^2 - 1} - \frac{\sin\sqrt{\omega^2 - 1}}{\sqrt{\omega^2 - 1}}\right\}. \tag{5}$$

It turns out that

$$R = \operatorname{Re} Z(j\omega) = \frac{1}{D^2 - C^2} = \frac{1}{K(\omega)} \tag{6}$$

$$X = \operatorname{Im} Z(j\omega) = -j\frac{BD - AC}{D^2 - C^2} = -\frac{2\omega\sin^2\sqrt{\omega^2 - 1}}{K(\omega)(\omega^2 - 1)}, \tag{7}$$

where

$$\frac{K(\omega)}{e^2} = \left\{\cos\sqrt{\omega^2 - 1} - \frac{\sin\sqrt{\omega^2 - 1}}{\sqrt{\omega^2 - 1}}\right\}^2 + \frac{\omega^2\sin^2\sqrt{\omega^2 - 1}}{\omega^2 - 1}. \tag{8}$$

Hence,

$$\begin{aligned} f(\omega) &= \operatorname{Re}\exp(-2j\omega)Z(j\omega) = R\cos 2\omega + X\sin 2\omega \\ &= \frac{1}{K(\omega)}\left[\cos 2\omega - \frac{2\omega\sin^2\sqrt{\omega^2 - 1}\sin 2\omega}{\omega^2 - 1}\right]. \end{aligned} \tag{9}$$

Obviously $f(\omega)$ does not possess a limit for $\omega \to \pm\infty$.

Example 3: Consider now the class of transmission lines which have a positive bounded and twice differentiable $c(x)$ in the interval $0 \leqq x \leqq l$. It can be shown (see e.g., Ref. 2) that the ABCD parameters satisfy the following asymptotic relations, for ω large:*

$$A(j\omega) = \sqrt{\frac{c(0)}{c(l)}}\cos l\omega + O\left(\frac{1}{\omega}\right) \tag{10}$$

$$B(j\omega) = j\frac{\sin l\omega}{\sqrt{c(0)c(l)}} + O\left(\frac{1}{\omega}\right) \tag{11}$$

$$C(j\omega) = j\sqrt{c(0)c(l)}\sin l\omega + O\left(\frac{1}{\omega}\right) \tag{12}$$

$$D(j\omega) = \sqrt{\frac{c(l)}{c(0)}}\cos l\omega + O\left(\frac{1}{\omega}\right). \tag{13}$$

These results follow from the classical theory of the asymptotic behavior of the eigenfunctions of Sturm-Liouville problems.[3] The WKBJ method is a related subject. Schelkunoff has discussed these

* The line is driven at the point $x = l$. The product of the inductance per unit length and the capacitance per unit length is assumed to be unity.

matters in an elementary way in at least one of his textbooks[4] (he does not include the $O(1/\omega)$ term).

If the line is terminated at $x = 0$ with a resistance R_0, we have for the driving point impedance

$$Z(j\omega) = \frac{R_0 A(j\omega) + B(j\omega)}{R_0 C(j\omega) + D(j\omega)}. \tag{14}$$

Substituting from (10), (11), (12), and (13) we find that for large ω

$$Z(j\omega) = \frac{R_0 c(0)}{c(l)} \left[1 + j \frac{(1 - R_0^2 c^2(0)) \sin l\omega}{R_0 c(0) \{\cos l\omega + j R_0 c(0) \sin l\omega\}} + O\left(\frac{1}{\omega}\right) \right] \tag{15}$$

and

$$R = \operatorname{Re} Z(j\omega) = \frac{R_0 c(0)}{c(l)} \left[1 + \frac{(1 - R_0^2 c^2(0)) \sin^2 l\omega}{1 - (1 - R_0^2 c^2(0)) \sin^2 l\omega} + O\left(\frac{1}{\omega}\right) \right] \tag{16}$$

$$X = \operatorname{Im} Z(j\omega) = \frac{(1 - R_0^2 c^2(0)) \sin 2l\omega}{2c(l)[1 - (1 - R_0^2 c^2(0)) \sin^2 l\omega]} + O\left(\frac{1}{\omega}\right). \tag{17}$$

Hence, if $R_0 c(0) \neq 1$, $Z(j\omega)$, $\operatorname{Re} Z(j\omega)$, and $\operatorname{Im} Z(j\omega)$ do not have limits for $\omega \rightarrow \pm\infty$.

Similarly, $f(\omega) = \operatorname{Re} \exp(-2jl\omega) Z(j\omega)$ does not have a limit for $\omega \rightarrow \pm\infty$. When $R_0 c(0) = 1$, i.e., when the line is "locally matched" at $x = 0$, we have

$$Z(j\omega) = \frac{1}{c(l)} + O\left(\frac{1}{\omega}\right) \tag{18}$$

$$R = \operatorname{Re} Z(j\omega) = \frac{1}{c(l)} + O\left(\frac{1}{\omega}\right) \tag{19}$$

$$X = \operatorname{Im} Z(j\omega) = O\left(\frac{1}{\omega}\right). \tag{20}$$

In this case,

$$f(\omega) = \operatorname{Re} \exp(-2jl\omega) Z(j\omega) = \frac{1}{c(l)} \cos 2l\omega + O\left(\frac{1}{\omega}\right). \tag{21}$$

Clearly, $f(\omega)$ does not have the asymptotic behavior stipulated by Zador; it does not even have a limit (because of the $\cos 2l\omega$ term).

Note that the asymptotic formulas (10), (11), (12), and (13) are also valid for a continuous positive $c(x)$ which is *piecewise* twice differentiable. This can be proven by partitioning the line at the discontinuity points and finding the overall ABCD matrix by multiplying the ABCD matrices of the sections of the line which now have a twice differentiable $c(x)$.

Hence, property *(iii)* of Zador's necessity statement could be replaced by the following: If *(i)* $c(x)$ is a positive continuous and piecewise twice differentiable function of the real variable x, *(ii)* the line is terminated in a unit resistance and $c(0) = 1$, then the following relation is valid for large ω:

$$Z(j\omega) = \frac{1}{c(l)} + O\left(\frac{1}{\omega}\right). \tag{22}$$

Another substitute will be discussed in the following. Let $\rho(j\omega)$ be the voltage reflection coefficient at $x = l$ for the unit resistance terminated line, then

$$Z(j\omega) = \frac{1}{c(l)} \frac{1 + \rho(j\omega)}{1 - \rho(j\omega)}. \tag{23}$$

For a $c(x)$ which is continuous and twice differentiable in the interval $0 \leqq x \leqq l$ with

$$\begin{gathered} c(0) = 1 \\ \frac{dc(0)}{dx} = \frac{dc(l)}{dx} = 0 \end{gathered} \tag{24}$$

we can see, using Schelkunoff's results on wave propagation in stratified media,[5] that for ω large

$$\rho(j\omega) = O\left(\frac{1}{\omega^2}\right). \tag{25}$$

From (23) we have in general for $| \rho(j\omega) | < 1$

$$Z(j\omega) = \frac{1}{c(l)} \{1 + 2\rho(j\omega) + 2\rho^2(j\omega) + \cdots\}. \tag{26}$$

Hence, using (25) we get

$$Z(j\omega) = \frac{1}{c(l)} + O\left(\frac{1}{\omega^2}\right) \tag{27}$$

for large ω.

To generalize (following Schelkunoff[5]) if $c(0) = 1$ and the first n derivatives of $c(x)$ are continuous functions of x and vanish at the boundaries then for large ω

$$\rho(j\omega) = O\left(\frac{1}{\omega^{n+1}}\right) \tag{28}$$

and therefore,

$$Z(j\omega) = \frac{1}{c(l)} + O\left(\frac{1}{\omega^{n+1}}\right). \tag{29}$$

Property (*ii*) in the necessity statement of Zador is also wrong.

Proof: The input impedance of the unit-resistance terminated line may be written, in terms of the ABCD parameters, as follows:

$$Z(s) = \frac{A(s) + B(s)}{C(s) + D(s)} = \frac{Q(s)}{P(s)}. \tag{30}$$

Consider a line with a twice differentiable $c(x)$. In this case $A(s)$, $B(s)$, $C(s)$, and $D(s)$ are entire functions of order 1 and type l (see Ref. 2), i.e.,

$$\begin{aligned} A(s) &\approx c_1 e^{ls} \\ B(s) &\approx c_2 e^{ls} \\ C(s) &\approx c_3 e^{ls} \\ D(s) &\approx c_4 e^{ls} \end{aligned} \tag{31}$$

(where c_1, c_2, c_3, c_4 are positive constants) for real $s \to +\infty$. Note also that

$$\begin{aligned} A(s) &= A(-s) \qquad & D(s) &= D(-s) \\ B(s) &= -B(-s) \qquad & C(s) &= -C(-s) \end{aligned} \tag{32}$$

and

$$AB - CD \equiv 1. \tag{33}$$

In order to find Zador's representation with the N_i, D_i ($i = 1,2$) functions we should be able to find an entire function $\varphi(s) \not\equiv 0$ such that when we multiply both the numerator and denominator of $Z(s)$ in (30) by this entire function, we get functions N_i, D_i ($i = 1,2$) with the properties stipulated by Zador.

We will have

$$N_1(s) = \text{Ev}\,[Q(s)\varphi(s)] = A(s)\,\frac{\varphi(s) + \varphi(-s)}{2} + B(s)\,\frac{\varphi(s) - \varphi(-s)}{2} \tag{34}$$

$$N_2(s) = \text{Odd}\,[Q(s)\varphi(s)] = A(s)\,\frac{\varphi(s) - \varphi(-s)}{2} + B(s)\,\frac{\varphi(s) + \varphi(-s)}{2}. \tag{35}$$

Similarly,

$$D_1(s) = D(s)\,\frac{\varphi(s) + \varphi(-s)}{2} + C(s)\,\frac{\varphi(s) - \varphi(-s)}{2} \tag{36}$$

$$D_2(s) = D(s)\,\frac{\varphi(s) - \varphi(-s)}{2} + C(s)\,\frac{\varphi(s) + \varphi(-s)}{2}. \tag{37}$$

Hence,

$$N_1(s)D_1(s) - N_2(s)D_2(s) = \varphi(s)\varphi(-s). \tag{38}$$

From (34), (35), (36), and (37) it follows that the functions $(\varphi(s) + \varphi(-s))/2$ and $(\varphi(s) - \varphi(-s))/2$ should be of type 0 in order that Zador's N_i and D_i be of type l. Consequently, the functions $\varphi(s)$ and $\varphi(-s)$ themselves are of type 0. Therefore, it is impossible to find an $\varphi(s)$ such that $\varphi(s)\varphi(-s) = \exp 2ls$ as Zador stipulates. So property (*ii*) in Zador's necessity statement could be replaced by

$$N_1D_1 - N_2D_2 = k^2,$$

where k is a constant. Then N_i , D_i $(i = 1,2)$ are proportional to the ABCD parameters with proportionality factor k.

From the above it follows that the sufficiency part as stated is inaccurate. It might be possible to alter the sufficiency conditions to make them valid. In this case a proof must be given. The author has done related work[6] on realizability conditions for nonuniform RC lines and is familiar with the difficulties involved in proving sufficiency conditions of this form.

Finally, Zador's conjectures do not have an obvious physical interpretation and hence they should be justified.

REFERENCES

1. Zador, P. L., Realizability Conditions for the Impedance Function of the Lossless Tapered Transmission Line, B.S.T.J., *45*, November, 1966, pp. 1667–1669.
2. Protonotarios, E. N. and Wing, O., Analysis and Intrinsic Properties of the General Nonuniform Transmission Line, To appear IEEE Trans. Microwave Theor. Tech. March, 1967.
3. Ince, E. L., *Ordinary Differential Equations*, Dover Publications, Inc., New York, 1956.
4. Schelkunoff, S. A., *Electromagnetic Fields*, Blaisdel, New York, 1963.
5. Schelkunoff, S. A., Remarks Concerning Wave Propagation in Stratified Media, Commun. Pure Appl. Math. June, 1951, pp. 117–128.
6. Protonotarios, E. N., On the analysis and Synthesis of the Nonuniform RC Distributed Network, Doctoral Dissertation, Columbia University, May, 1966.