THE BELL SYSTEM

TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 50 March 1971 Number 3

The Energy and General Translation Force of Cylindrical Magnetic Domains

By A. A. THIELE, A. H. BOBECK, E. DELLA TORRE,* and U. F. GIANOLA

(Manuscript received October 7, 1970)

In this paper we compute the change in the energy of a uniformly magnetized uniaxial platelet produced by the introduction of a cylindrical domain. Differentiation of the energy expression yields the translational force produced by gradients in plate thickness, material composition, or temperature. The force expressions provide a means for estimating the effect of gradients in these parameters on the margins of domain devices. Equating the sum of the gradient produced forces to the drag force yields a general domain velocity expression. The various results are presented in both graphical and tabular form.

I. INTRODUCTION

Magnetic memory and logic devices employing cylindrical domains in uniaxial platelets have recently received considerable attention.[1,2] The theory of the static stability of these domains[3] and its application to cylindrical domain devices[4,5] have been discussed in previous papers. This paper is concerned with the translational forces acting on the domains and their effect on device performance.

* Present address Department of Electrical Engineering, McMaster University, Hamilton, Ontario.

In most device applications, cylindrical domains are propagated by gradients in the applied field.[1,2] Cylindrical domains may, however, be propagated by gradients in any of the independent parameters which determine the total domain energy. These parameters are: the applied field, H; the plate thickness, h; the saturation magnetization, M_s; and the wall energy density, σ_w. The domain radius, r_0, is not an independent parameter for domains in equilibrium but is determined once the other parameters are specified. The gradients in σ_w and M_s may be produced by composition gradients, strain gradients or temperature gradients. Composition or thickness gradients may be used to provide forces which are functions of position only, while temperature or strain gradients may be used to provide time variable forces.

The translational force is obtained by differentiation of the total domain energy expression with respect to position, under the assumption that the gradients in the domain parameters which produce the translational force are sufficiently small that the domain remains circular and stable; consequently the energy expression remains valid. Since this method of computing the force is independent of the detailed stress pattern which produces the domain motion, no estimate of the shape distorting tendency of the various parameters is obtained. Equating the sum of the translational forces to the drag force yields the general velocity equation, and comparing the magnitudes of the various forces yields their effects on device operation.

II. DOMAIN ENERGY

The energy change produced by the introduction of a single isolated circular 180° domain into an infinite plate of uniaxial magnetic material which is otherwise uniformly magnetized along the average plate normal (the z axis) is now calculated. Such a domain configuration is shown in Fig. 1. The assumptions and notation of Refs. 3, 4 and 5 are maintained except that the domain parameters h, H, M_s, σ_w, and r_0 are allowed to be functions of position on the plate. In particular, the following is assumed in the model: the wall has negligible width, the wall is everywhere parallel to the z axis, and the wall energy density is independent of wall orientation or curvature. The values of all parameters are assumed to vary sufficiently slowly that they may be represented by their z-averaged values at the center of the circular domain. Additionally, the applied field is represented by its z component, H, since under the assumptions stated above only this component interacts with the magnetization.

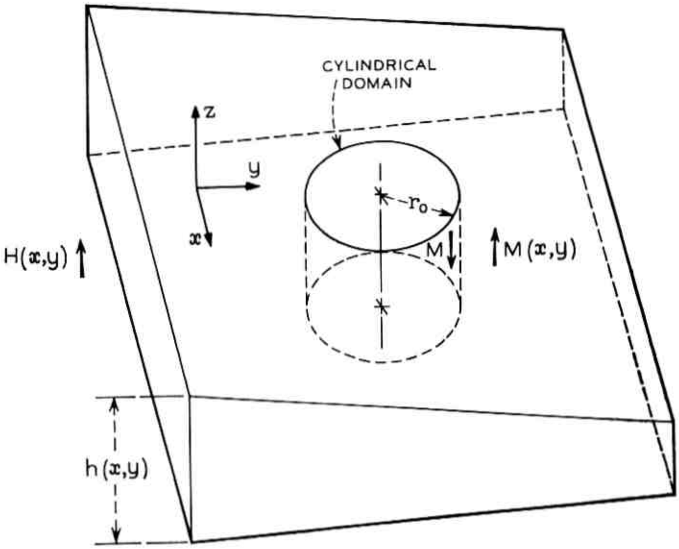


Fig. 1—The cylindrical domain configuration and coordinate system.

The total change in the energy of the system due to the presence of a cylindrical domain is

$$E_T = 2\pi r_0 h \sigma_w + 2M_s H \pi r_0^2 h - 2\pi h^3 (2\pi M_s^2) I(2r_0/h). \tag{1}$$

In this expression the first term, the wall energy, is the product of the wall energy density and the wall area; the second term, the applied field interaction energy, is the product of the magnetization change, $2M_s$, the applied field and the domain volume; and the third term, the internal magnetostatic energy, is the negative of the integral of the generalized radial magnetostatic force of Ref. 3. The internal magnetostatic energy function, $I(2r_0/h)$ is therefore defined as

$$I(2r_0/h) = \int_0^{2r_0/h} F(x)\, dx, \tag{2}$$

where $F(x)$ is the force function defined by equations 33 and 138 of Ref. 3. The lower limit of the integral (2) is chosen so that when the plate is uniformly magnetized, $r_0 = 0$, the domain energy expression (1) is zero. Various closed form and power series representations of $I(d/h)$ are given in the appendix of this paper. This function, which is plotted in Fig. 2 as a function of the diameter-to-thickness ratio, d/h, and is tabulated in Table I, has the asymptotic forms

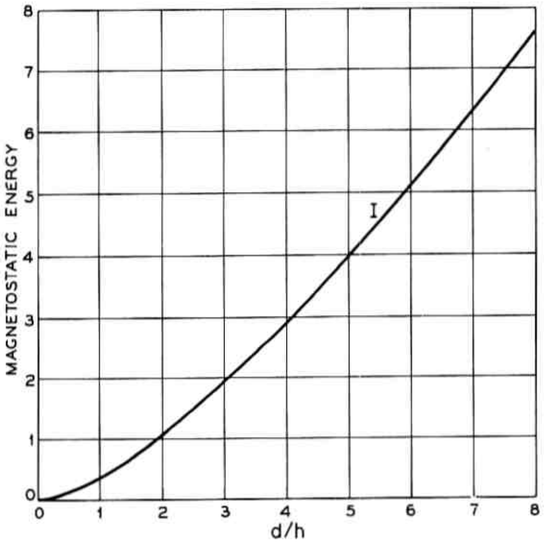


Fig. 2—The internal magnetostatic energy function, I, as a function of the diameter-to-thickness ratio.

$$I\left(\frac{d}{h}\right) \approx \frac{1}{\pi}\frac{d}{h}\left\{-\frac{1}{2} + \frac{1}{32}\left(\frac{h}{d}\right)^2 + \left[1 + \frac{1}{8}\left(\frac{h}{d}\right)^2\right] \ln\left|\frac{4d}{h}\right|\right\}, \quad \frac{h}{d} \ll 1 \tag{3a}$$

and

$$I\left(\frac{d}{h}\right) \approx \frac{d}{h}\left[\frac{1}{2}\left(\frac{d}{h}\right) - \frac{2}{3\pi}\left(\frac{d}{h}\right)^3 + \frac{1}{16}\left(\frac{d}{h}\right)^4\right], \quad \frac{d}{h} \ll 1. \tag{3b}$$

The physical origin of the asymptotic behavior of the I function was discussed in some detail in Section IV of Ref. 3.

The domain energy per unit wall length is

$$E_T/\pi d = (2\pi M_s^2)h^2[l/h - J(d/h)], \tag{4}$$

where $l = \sigma_w/4\pi M_s^2$ is the characteristic material length, where the applied field has been eliminated using the equilibrium condition[3-5]

$$\frac{l}{h} + \frac{d}{h}\frac{H}{4\pi M_s} - F\left(\frac{d}{h}\right) = 0, \tag{5}$$

and where

$$J\left(\frac{d}{h}\right) = 2\frac{h}{d}I\left(\frac{d}{h}\right) - F\left(\frac{d}{h}\right). \tag{6a}$$

TABLE I—MAGNETOSTATIC ENERGY, FORCE AND STABILITY FUNCTIONS

d/h	I	$F_H = F$	S_0	S_2	$F_h = 3J$	F_M
0.00	0.	0.	0.	0.	0.	0.
0.10	0.0048	0.0939	0.0059	0.0007	0.0060	0.0979
0.20	0.0184	0.1765	0.0215	0.0028	0.0225	0.1915
0.30	0.0398	0.2493	0.0442	0.0063	0.0474	0.2809
0.40	0.0680	0.3137	0.0716	0.0111	0.0787	0.3662
0.50	0.1023	0.3708	0.1017	0.0172	0.1149	0.4474
0.60	0.1419	0.4216	0.1332	0.0243	0.1544	0.5246
0.70	0.1864	0.4672	0.1648	0.0323	0.1962	0.5980
0.80	0.2352	0.5083	0.1960	0.0411	0.2394	0.6679
0.90	0.2879	0.5455	0.2262	0.0505	0.2832	0.7343
1.00	0.3442	0.5794	0.2552	0.0603	0.3271	0.7975
1.10	0.4037	0.6104	0.2829	0.0705	0.3708	0.8576
1.20	0.4662	0.6390	0.3093	0.0809	0.4139	0.9150
1.30	0.5315	0.6655	0.3343	0.0914	0.4564	0.9698
1.40	0.5993	0.6901	0.3579	0.1020	0.4980	1.0221
1.50	0.6694	0.7130	0.3804	0.1126	0.5386	1.0721
1.60	0.7418	0.7345	0.4016	0.1231	0.5783	1.1200
1.70	0.8163	0.7547	0.4218	0.1336	0.6169	1.1660
1.80	0.8927	0.7737	0.4410	0.1439	0.6546	1.2101
1.90	0.9710	0.7917	0.4592	0.1541	0.6912	1.2525
2.00	1.0510	0.8087	0.4765	0.1642	0.7268	1.2933
2.10	1.1327	0.8249	0.4931	0.1741	0.7615	1.3326
2.20	1.2160	0.8404	0.5089	0.1838	0.7952	1.3705
2.30	1.3008	0.8551	0.5240	0.1933	0.8280	1.4071
2.40	1.3870	0.8692	0.5385	0.2027	0.8599	1.4424
2.50	1.4746	0.8827	0.5524	0.2119	0.8910	1.4766
2.60	1.5635	0.8956	0.5657	0.2209	0.9212	1.5098
2.70	1.6537	0.9081	0.5786	0.2297	0.9507	1.5418
2.80	1.7451	0.9200	0.5909	0.2383	0.9794	1.5729
2.90	1.8377	0.9316	0.6028	0.2468	1.0074	1.6031
3.00	1.9314	0.9427	0.6143	0.2551	1.0347	1.6325
3.20	2.1221	0.9639	0.6362	0.2712	1.0872	1.6887
3.40	2.3169	0.9837	0.6566	0.2866	1.1374	1.7420
3.60	2.5155	1.0024	0.6759	0.3015	1.1853	1.7926
3.80	2.7178	1.0201	0.6940	0.3158	1.2310	1.8407
4.00	2.9235	1.0368	0.7112	0.3295	1.2749	1.8867
4.20	3.1324	1.0526	0.7275	0.3428	1.3170	1.9306
4.40	3.3445	1.0678	0.7430	0.3556	1.3574	1.9727
4.60	3.5595	1.0822	0.7578	0.3679	1.3962	2.0130
4.80	3.7773	1.0960	0.7720	0.3799	1.4337	2.0518
5.00	3.9978	1.1092	0.7855	0.3914	1.4698	2.0891
5.20	4.2209	1.1219	0.7985	0.4026	1.5047	2.1250
5.40	4.4465	1.1341	0.8109	0.4134	1.5383	2.1596
5.60	4.6745	1.1458	0.8229	0.4239	1.5709	2.1931
5.80	4.9048	1.1572	0.8345	0.4341	1.6025	2.2255
6.00	5.1374	1.1681	0.8456	0.4439	1.6331	2.2568
6.20	5.3721	1.1787	0.8564	0.4535	1.6628	2.2872
6.40	5.6088	1.1889	0.8668	0.4629	1.6916	2.3166
6.60	5.8476	1.1988	0.8769	0.4720	1.7196	2.3452
6.80	6.0883	1.2084	0.8866	0.4808	1.7468	2.3730
7.00	6.3310	1.2177	0.8961	0.4894	1.7733	2.3999
7.20	6.5754	1.2268	0.9053	0.4978	1.7991	2.4262
7.40	6.8217	1.2356	0.9142	0.5060	1.8242	2.4518
7.60	7.0696	1.2442	0.9229	0.5140	1.8488	2.4767
7.80	7.3193	1.2525	0.9314	0.5218	1.8727	2.5010
8.00	7.5706	1.2606	0.9396	0.5295	1.8960	2.5247

The function J has the asymptotic forms

$$J\left(\frac{d}{h}\right) \approx \frac{1}{\pi}\left\{-\frac{3}{2} - \frac{1}{32}\left(\frac{h}{d}\right)^2 + \left[1 + \frac{3}{8}\left(\frac{h}{d}\right)^2\right] \ln\left|\frac{4d}{h}\right|\right\}, \qquad \frac{h}{d} \ll 1 \tag{6b}$$

and

$$J\left(\frac{d}{h}\right) \approx \frac{2}{3\pi}\left(\frac{d}{h}\right)^2 - \frac{1}{8}\left(\frac{d}{h}\right)^3, \qquad \frac{d}{h} \ll 1. \tag{6c}$$

The function $J(d/h)$, the normalized total magnetostatic energy per unit wall length, is plotted in Fig. 3 together with the force function, $F(d/h)$ and the stability functions $S_0(d/h)$ and $S_2(d/h)$ which have the asymptotic forms

$$F\left(\frac{d}{h}\right) \approx \frac{1}{\pi}\left\{\frac{1}{2} + \frac{3}{32}\left(\frac{h}{d}\right)^2 + \left[1 - \frac{1}{8}\left(\frac{h}{d}\right)^2\right] \ln\left|\frac{4d}{h}\right|\right\}, \qquad \frac{h}{d} \ll 1 \tag{7a}$$

and

$$F\left(\frac{d}{h}\right) \approx \frac{d}{h} - \frac{2}{\pi}\left(\frac{d}{h}\right)^2 + \frac{1}{4}\left(\frac{d}{h}\right)^3, \qquad \frac{d}{h} \ll 1, \tag{7b}$$

$$S_0\left(\frac{d}{h}\right) = \frac{1}{\pi}\left\{-\frac{1}{2} + \frac{13}{32}\left(\frac{h}{d}\right)^2 + \left[1 - \frac{3}{8}\left(\frac{h}{d}\right)^2\right] \ln\left|\frac{4d}{h}\right|\right\}, \qquad \frac{h}{d} \ll 1 \tag{8a}$$

and

$$= \frac{2}{\pi}\left(\frac{d}{h}\right)^2 - \frac{1}{2}\left(\frac{d}{h}\right)^3, \qquad \frac{d}{h} \ll 1 \tag{8b}$$

and

$$S_2\left(\frac{d}{h}\right) = \frac{1}{\pi}\left\{-\frac{11}{6} - \frac{17}{96}\left(\frac{h}{d}\right)^2 + \left[1 + \frac{5}{8}\left(\frac{h}{d}\right)^2\right] \ln\left|\frac{4d}{h}\right|\right\}, \quad \frac{h}{d} \ll 1 \tag{9a}$$

and

$$= \frac{2}{9\pi}\left(\frac{d}{h}\right)^2 - \frac{1}{48}\left(\frac{d}{h}\right)^5, \qquad \frac{d}{h} \ll 1. \tag{9b}$$

Numerical values of all these functions are given in Table I.

Since from the figure and the asymptotic forms of the functions, (6), (8) and (9), $J(d/h)$ lies roughly midway (with respect to diameter) between $S_0(d/h)$ and $S_2(d/h)$ and since the condition for domain stability is[3-5] $S_0(d/h) > l/h > S_2(d/h)$, then a platelet of arbitrary thickness may always be biased such that the introduction of a domain

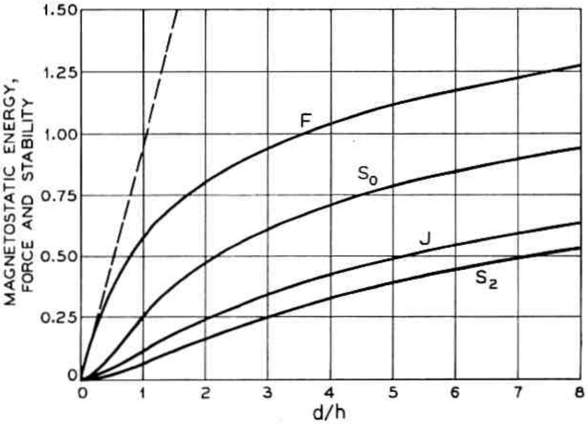


Fig. 3—The magnetostatic energy per unit wall length function, J, the magnetostatic force function, F, and the magnetostatic stability functions, S_0 and S_2, as functions of the diameter-to-thickness ratio.

produces either a positive, or negative, or zero change in the total energy. The force function, $F(d/h)$, was included in the figure so that the bias fields yielding these conditions may be determined. The values of the diameters and bias fields of zero energy determined using Fig. 3 or Table I together with (4) and (5) are found to agree with those obtained previously.[6] In the case of a domain having the preferred dimensions,[4,5] $l/h = 0.2500$ and $d/h = 2.000$ (corresponding to an applied field of $H/4\pi M_s = 0.279$), then $J(2.000) = 0.2422$. Since J is nearly equal to l/h the total energy is nearly zero. Under the preferred conditions and in a platelet in which $4\pi M_s = 100$ Gauss and $d = 10$ microns, the absolute value of the wall energy and the total magnetostatic energy change are each approximately 0.2 times the rest energy of the electron. [Note that the terms in (4) may not be identified as wall energy and magnetostatic contributions because the equilibrium condition was used to eliminate the applied field.]

III. THE TRANSLATIONAL FORCE

The translational force is given by

$$\begin{aligned}\mathbf{F} &= -\nabla E_T,\\ &= -\left(\frac{\partial E_T}{\partial h}\right)_{H,M_s,\sigma_w,r_0}\nabla h - \left(\frac{\partial E_T}{\partial H}\right)_{h,M_s,\sigma_w,r_0}\nabla H\end{aligned}$$

$$- \left(\frac{\partial E_T}{\partial M_s}\right)_{h,H,\sigma_w,r_0} \nabla M_s - \left(\frac{\partial E_T}{\partial \sigma_w}\right)_{h,H,M_s,r_0} \nabla \sigma_w$$

$$- \left(\frac{\partial E_T}{\partial r_0}\right)_{h,H,\sigma_w,M_s} \nabla r_0. \tag{10}$$

In this expression h, H, M_s and σ_w are considered to be independent variables (functions of position on the platelet) and r_0 is a dependent function of these variables determined by the equilibrium condition (5). Since the fundamental equilibrium condition is $(\partial E_T/\partial r_0)_{h,H,\sigma_w,M_s} = 0$, the last term in equation (10) may be dropped. Evaluating the remaining terms using equations (1) and (2) yields

$$\begin{aligned}\mathbf{F} = -[&2\pi r_0\sigma_w + 2M_sH\pi r_0^2 - 6\pi h^2(2\pi M_s^2)I(2r_0/h)\\ &+ 4\pi r_0 h(2\pi M_s^2)F(2r_0/h)]\nabla h - [2M_s\pi r_0^2 h]\nabla H\\ &- [2H\pi r_0^2 h - 4\pi h^3(2\pi M_s)I(2r_0/h)]\nabla M_s\\ &- [2\pi r_0 h]\nabla \sigma_w.\end{aligned} \tag{11}$$

Eliminating the applied field using the equilibrium condition (5) and rearranging yields

$$\begin{aligned}\mathbf{F} = \pi\, dh^2(2\pi M_s^2)\{&-[l/h - F_h(d/h)]\nabla h/h + [l/h - F_H(d/h)]\nabla H/H\\ &+ [l/h + F_M(d/h)]\nabla M_s/M_s - 2[l/h]\nabla\sigma_w/\sigma_w\},\end{aligned} \tag{12}$$

where

$$F_h\left(\frac{d}{h}\right) = 6\left(\frac{h}{d}\right)I\left(\frac{d}{h}\right) - 3F\left(\frac{d}{h}\right) = 3J\left(\frac{d}{h}\right) \tag{13a}$$

$$\approx \frac{1}{\pi}\left\{-\frac{9}{2} - \frac{3}{32}\left(\frac{h}{d}\right)^2 + \left[3 + \frac{9}{8}\left(\frac{h}{d}\right)^2\right]\ln\left|\frac{4d}{h}\right|\right\}, \quad \frac{h}{d} \ll 1, \tag{13b}$$

$$\approx \frac{2}{\pi}\left(\frac{d}{h}\right)^2 - \frac{3}{8}\left(\frac{d}{h}\right)^3, \quad \frac{d}{h} \ll 1, \tag{13c}$$

$$F_H(d/h) = F(d/h), \tag{14}$$

and

$$F_M\left(\frac{d}{h}\right) = 4\left(\frac{h}{d}\right)I\left(\frac{d}{h}\right) - F\left(\frac{d}{h}\right), \tag{15a}$$

$$= \frac{1}{\pi}\left\{-\frac{5}{2} + \frac{1}{32}\left(\frac{h}{d}\right)^2 + \left[3 + \frac{5}{8}\left(\frac{h}{d}\right)^2\right]\ln\left|\frac{4d}{h}\right|\right\}, \quad \frac{h}{d} \ll 1, \tag{15b}$$

$$= \frac{d}{h} - \frac{2}{3\pi}\left(\frac{d}{h}\right)^2 + \frac{1}{64}\left(\frac{d}{h}\right)^5, \quad \frac{d}{h} \ll 1. \tag{15c}$$

Figure 4 shows plots of F_h, F_M and $F = F_H$ as functions of the normalized domain diameter. Numerical values of these functions are included in Table I. From the figure and the asymptotic forms of (7), (8), (13) and (14), it is seen that the functions are positive and that both F_h and F_H are greater than $S_0(d/h)$. From these properties and the stability requirement $S_0(d/h) > l/h > S_2(d/h)$, it is seen that for any stable domain the thickness gradient and magnetizàtion gradient contributions to the force are in the direction of the gradient while the field gradient and wall energy gradient contributions to the force are in the direction opposite to the gradient.

The absolute value of each of the terms in equation (12) is small but measureable. If H_0 and H_2 are defined as the collapse and elliptical runout fields respectively, then in the case of a gradient in the applied field where the stability limits are known roughly (see Ref. 5), the force produced on a domain for which $d/h = 2$, $h/l = 4$, $\nabla H = (H_0 - H_2)/d$, $d = 100$ microns and $4\pi M_s = 100$ Gauss is approximately 6×10^{-3} dynes. While such absolute force measurements could possibly be carried out, a more relevant experiment for device applications is the balancing of field gradients against gradients in the other quantities. At the present time, no such measurements have been completed. However, the directions of the applied field gradient force, the wall energy gradient force, and the thickness gradient force have been verified. The sign of the H gradient force is verified in everyday device operation.

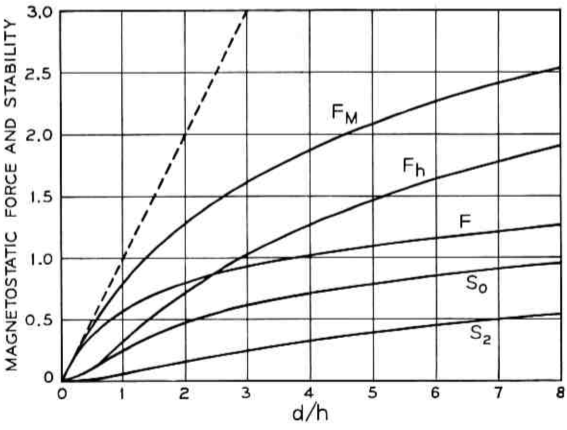


Fig. 4—The transverse magnetostatic force functions, F_M, F_h, and F_H ($F = F_H$) and the magnetostatic stability functions S_0 and S_2, as functions of the diameter-to-thickness ratio.

The sign of σ_w gradient force was checked using a temperature gradient in a sample of $Sm_{0.55}Tb_{0.45}FeO_3$. According to the data of F. C. Rossol,[7] at room temperature the wall energy temperature coefficient is approximately +1.5%/C° and the magnetization is only a slightly decreasing function of temperature. The σ_w gradient force thus dominates in this material, and we observed that the domains move away from heated regions towards cooler regions as predicted. The sign of the M_s gradient force was verified using a temperature gradient in a sample of $Gd_{2.3}Tb_{0.7}Fe_5O_{12}$ garnet. In this material M_s increases at 3%/C° at room temperature while the wall energy is approximately constant. The M_s gradient force thus dominates, and we observed that the domains moved towards heated regions as predicted. The force direction measurements were completed in tapered platelets of orthoferrite in which we observed that the domains move towards the thick end of a platelet when restraints were removed.

It should be noted that in carrying out these experiments it is important to obtain low coercivity materials. This is especially true in measuring the thickness gradient force when $\nabla h/h$ is small.

The drag force acting on a domain propagating with uniform velocity $\mathbf{v}_d$ from equation 58 of Ref. 5 is given by

$$\mathbf{F}_d = -\frac{\pi}{2} dh\, M_s \left[\frac{8}{\pi} H_c + \frac{2}{\mu_w} |\mathbf{v}_d| \right] \frac{\mathbf{v}_d}{|\mathbf{v}_d|}, \tag{16}$$

where H_c is the wall motion coercivity and μ_w is the wall motion mobility. Equating to zero the sum of the gradient force (12) and the drag force (16) yields the velocity equation for an otherwise freely propagating domain. In order to avoid the vector sum in this equation, it is convenient to assume that all the gradients lie in the same direction and are positive. It is also convenient to assume that the gradients are uniform so that their magnitudes may be expressed in terms of the maximum parameter difference across the domain divided by the domain diameter, $\nabla X = \Delta X/d$. Under these assumptions the velocity equation is

$$\frac{8}{\pi} \frac{H_c}{4\pi M_s} + \frac{2\,|\mathbf{v}_d|}{\mu_w 4\pi M_s} = C_h \frac{\Delta h}{h} - C_H \frac{\Delta H}{H} + C_M \frac{\Delta M_s}{M_s} - C_\sigma \frac{\Delta \sigma_w}{\sigma_w}, \tag{17a}$$

where

$$C_h = -\frac{h}{d} \left[\frac{l}{h} - F_h\left(\frac{d}{h}\right) \right], \tag{17b}$$

$$C_H = -\frac{h}{d} \left[\frac{l}{h} - F_H\left(\frac{d}{h}\right) \right], \tag{17c}$$

$$C_M = \frac{h}{d}\left[\frac{l}{h} + F_M\!\left(\frac{d}{h}\right)\right], \tag{17d}$$

and

$$C_\sigma = 2\,\frac{h}{d}\,\frac{l}{h} = \frac{l}{r_0} \tag{17e}$$

are called transverse force coefficients and are all positive.

For a domain having the diameter $d = 8l$ in a plate of thickness $4l$, the velocity equation becomes

$$\frac{8}{\pi}\,\frac{H_c}{4\pi M_s} + \frac{2\,|\mathbf{v}_d|}{\mu_w 4\pi M_s}$$
$$= 0.238\,\frac{\Delta h}{h} - 0.279\,\frac{\Delta H}{H} + 0.772\,\frac{\Delta M_s}{M_s} - 0.250\,\frac{\Delta\sigma_w}{\sigma_w}. \tag{18}$$

If the magnitude of the sum of the gradients is not sufficiently large, no motion takes place. For example, when a domain having dimensions such that equation (18) applies is subjected to a field gradient, the magnitude of the field difference across the domain must satisfy the condition $\Delta H/H > 9.1 \times H_c/4\pi M_s$ before any motion can take place.

The transverse force coefficients are plotted as functions of the normalized thickness h/l in Fig. 5 for the bias condition $d = (d_0 d_2)^{\frac{1}{2}}$ where d_0 and d_2 are the collapse and elliptical runout diameters respectively. For small thicknesses the asymptotic forms of C_h, C_M, and C_σ are proportional to (l/h) exp $(-\pi l/h)$ and the asymptote of C_H is proportional to exp $(-\pi l/h)$. For large thicknesses C_H and C_M approach unity, and C_σ and C_h approach the asymptote $(8/3\pi)^{\frac{1}{2}}(l/h)^{\frac{1}{2}}$.

Some caution must be exercised in interpreting equation (17) and Fig. 5. First, the stability of moving domains has only been investigated for the case of gradients in the applied field and then only incompletely (see Ref. 5). Another problem is that drive gradients which are applied from the surface and which must obey Laplace's equation, such as field gradients and temperature gradients, decrease exponentially into the platelet. (This does not apply to volume heating such as laser heating.) For any given value of these gradients at the surface the maximum z averaged value of these gradients which may be applied thus decreases according to the inverse first power of the plate thickness in thick plates. This consideration shows that the use of platelets having a thickness no greater than the preferred thickness of $4l$ is thus more strongly preferred for achieving a high domain velocity than is indicated by inspection of Fig. 5 alone.

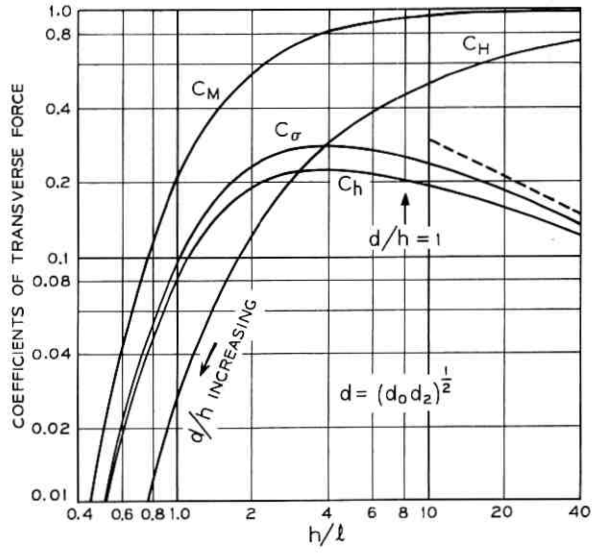


Fig. 5—The transverse force coefficients C_h, C_H, C_M, and C_σ for the bias condition $d = (d_0d_2)^{\frac{1}{2}}$, as functions of the normalized thickness, h/l.

IV. SIGNIFICANCE FOR DEVICE APPLICATIONS

Equation (17) and Fig. 5, when properly interpreted, show again that there is a preferred plate thickness for achieving high bit rates in devices and that this thickness is again $4l$. Equation (18) shows that, for plates of this thickness, wall energy, magnetization and applied field gradients produce transverse forces of similar magnitude. Figure 5 shows that this is also true for plates having a thickness in the neighborhood of $4l$. There is therefore no preferred thickness with respect to favoring applied field gradients over temperature or composition gradients.

In temperature dependent materials in which the domain tends to move towards the high temperature direction the domain will tend to follow a laser beam initially placed at its center. In materials having the opposite temperature characteristic, the domains may be pushed by a laser beam. Gradients in thickness or composition may be used to define domain tracks in order to increase margins with respect to spurious field or temperature gradients. An immediately useful application of equation (18) is in the calculation of the effect on device margins of the heat produced by domain generators and detectors.

APPENDIX

The Internal Magnetostatic Energy Function

Various representations of the internal magnetostatic energy function, $I(d/h)$, are given in this appendix. Since $F(d/h)$ is non-negative (Ref. 3, Fig. 3 and equation 138), the integral definition of $I(d/h)$ (2) implies that $I(d/h)$ is positive and monotonic increasing as can be seen in Fig. 1. (Neither function is defined in the present case for negative values of the argument.) The definition of $I(d/h)$ also implies the differential equation, $dI(x)/dx = F(x)$, with the boundary condition $I(0) = 0$. The internal magnetostatic function may be written as

$$I\left(\frac{d}{h}\right) = -\frac{1}{3\pi}\frac{d}{h}\left[2\frac{d^2}{h^2} + \left(1 - \frac{d^2}{h^2}\right)U\left(\frac{h^2}{d^2}\right) - \left(1 + \frac{h^2}{d^2}\right)V\left(\frac{h^2}{d^2}\right)\right], \tag{19}$$

where as in equations (84) and (85) of Ref. 3

$$U(x) \equiv 2(x + 1)^{\frac{1}{2}}E\left(\frac{1}{1 + x}\right), \tag{20a}$$

and

$$V(x) \equiv 2(x + 1)^{-\frac{1}{2}}K\left(\frac{1}{1 + x}\right), \tag{20b}$$

where $K(m)$ and $E(m)$ are the complete elliptic integrals of the first and second kind respectively and the argument is in the m of Ref. 8. Differentiation of equation (19) with respect to d/h and combining terms using equations (86) and (87a) of Ref. 3 verifies that the derivative of $I(d/h)$ is indeed $F(d/h)$ (equation 138 of Ref. 3). Substitution of the series expansions of U and V (equations 96 and 97 of Ref. 3) verifies that $I(0) = 0$.

Writing $I(d/h)$ in terms of the L_1 function (Ref. 3 Sec. A.4) eliminates the negative power of d/h,

$$I\left(\frac{d}{h}\right) = \frac{1}{3\pi}\frac{d}{h}\left\{\left(\frac{d}{h}\right)^2\left[U\left(\frac{h^2}{d^2}\right) - 2\right] - \frac{1}{2}L_1\left(\frac{h^2}{d^2}\right) + V\left(\frac{h^2}{d^2}\right)\right\}, \tag{21}$$

and the expression in terms of $F(d/h)$ and $S_0(d/h)$ (equations 138 and 139 of Ref. 3) is also sometimes useful,

$$I\left(\frac{d}{h}\right) = \frac{1}{3}\frac{d}{h}\left\{2F\left(\frac{d}{h}\right) + S_0\left(\frac{d}{h}\right)\left[1 + \frac{h^2}{d^2}\right] - \frac{2}{\pi}\right\}. \tag{22}$$

The power series expansions of $I(d/h)$ may be determined either by substituting the power series expansions for U, V and L_1 (Ref. 3, Sections A.3 and A.4) into equation (21) or by term-by-term integration of $F(d/h)$ with the integration constants being determined from the lowest order terms of the power series expansions of U and V. The result is

$$I\left(\frac{d}{h}\right) = \frac{1}{\pi}\left\{\left[-\frac{1}{2}\left(\frac{d}{h}\right) + \frac{1}{32}\left(\frac{h}{d}\right) + \frac{1}{96}\left(\frac{h}{d}\right)^3 - \frac{109}{24576}\left(\frac{h}{d}\right)^5 + \cdots\right] + \left[\left(\frac{d}{h}\right) + \frac{1}{8}\left(\frac{h}{d}\right) - \frac{1}{64}\left(\frac{h}{d}\right)^3 + \frac{5}{1024}\left(\frac{h}{d}\right)^5 + \cdots\right]\ln\left|\frac{4d}{h}\right|\right\} \tag{23a}$$

$$= \frac{1}{2}\left(\frac{d}{h}\right)^2 - \frac{2}{3\pi}\left(\frac{d}{h}\right)^3 + \frac{1}{16}\left(\frac{d}{h}\right)^4 - \frac{1}{128}\left(\frac{d}{h}\right)^6 + \frac{5}{2048}\left(\frac{d}{h}\right)^8 + \cdots . \tag{23b}$$

REFERENCES

1. Bobeck, A. H., Fischer, R. F., Perneski, A. J., Remeika, J. P., and Van Uitert, L. G., "Application of Orthoferrites to Domain-Wall Devices," IEEE Trans. Magnetics, *5*, No. 3 (September 1969), pp. 544–553.
2. Perneski, A. J., "Propagation of Cylindrical Magnetic Domains in Orthoferrites," IEEE Trans. Magnetics, *5*, No. 3 (September 1969), pp. 554–557.
3. Thiele, A. A., "The Theory of Cylindrical Magnetic Domains," B.S.T.J., *48*, No. 10 (December 1969), pp. 3287–3335.
4. Thiele, A. A., "Theory of the Static Stability of Cylindrical Domains in Uniaxial Platelets," J. Appl. Phys., *41*, No. 3 (March 1970), pp. 1139–1145.
5. Thiele, A. A., "Device Implications of the Theory of Cylindrical Magnetic Domains," B.S.T.J., this issue, pp. 725–773.
6. Cape, J. A., and Lehman, G. W., "Magnetic Bubble Domain Interactions," Solid State Commun., *8*, No. 16 (August 1970), pp. 1303–1306.
7. Rossol, F. C., "Temperature Dependence of Rare-Earth Orthoferrite Properties Relevant to Propagating Domain Device Applications," IEEE Trans. Magnetics, *5*, No. 3 (September 1969), pp. 562–565.
8. *Handbook of Mathematical Functions,* Abramowitz, M., and Stegun, J. A., editors, Nat. Bureau of Standards Appl. Math. Series 55, Washington, D. C., 1966, pp. 589–626.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

Device Implications of the Theory of Cylindrical Magnetic Domains*

By A. A. THIELE

(Manuscript received October 7, 1970)

This paper applies the theory of cylindrical magnetic domains[1] *to cylindrical domain devices. The stability conditions are examined as bounds to the region of possible device operation and it is found that the plate thickness, $h = 4l$, and the domain diameter, $d = 8l$, where l is the ratio of the wall energy per unit area to 4π times the saturation magnetization squared, are preferred values. When the effects of wall coercivity and mobility are examined, it is found that the preferred plate thickness and domain diameter are even more strongly preferred, that the wall motion coercivity should be less than one percent of 4π times the saturation magnetization, and that a domain coercivity and mobility may be defined. Consideration of the Néel temperature and the desired absolute domain size in addition to the static stability conditions shows that domain materials having some antiferromagnetic character and induced uniaxial anisotropy are preferred. Where appropriate, domain methods for measuring material parameters are described.*

I. INTRODUCTION

The application of cylindrical magnetic domains or "bubbles" to memory and logic devices has recently received considerable attention.[2–6] Such domain devices may operate in a continuum of modes ranging from the wall motion coercivity dominated mode to the "hard bubble" mode. In the coercivity dominated mode, applied fields determine the domain configuration which is then maintained by coercivity. In the hard bubble mode the coercivity must be sufficiently low that the domains have a well-defined size and shape permitting the move-

* Portions of this article were presented at the "Fifteenth Annual Conference on Magnetism and Magnetic Materials" Philadelphia, Pennsylvania, on Nov. 20, 1969. (See Ref. 2.)

ment of individual domains as distinct entities. A previous paper developed the theory of static stability of cylindrical domains in materials having zero coercivity.[1] The present work applies this theory to the determination of the preferred conditions for construction of devices operating in the hard bubble mode. It is found that specifying an operating domain diameter determines preferred values for the plate thickness, magnetization, anisotropy constant and the maximum allowable value of the wall motion coercivity.

Figure 1 shows the model for the domain structure from which the static stability theory was developed. The coordinate system and symbols used here are the same as in Ref. 1 except for the addition of a few symbols such as μ_w, the wall mobility; H_c, the wall motion coercivity; $\mathbf{v}_d$, the domain velocity; and $\Delta\sigma_w$ the variation in wall energy. The model represents a single isolated domain in a plate of magnetic material of uniform thickness, h, and an infinite extent in the plane, $r_f = \infty$. Everywhere within the material the magnetization has a uniform (saturation) magnitude, M_s, directed along the upward plate normal (the z direction) within the domain and along the downward plate normal elsewhere within the material. The domain wall is assumed to have negligible width and the domain wall energy density, σ_w, is initially taken to be independent of both wall orientation and curvature. The domain wall is cylindrical in the sense that it everywhere contains a line parallel to the plate normal. Under these assump-

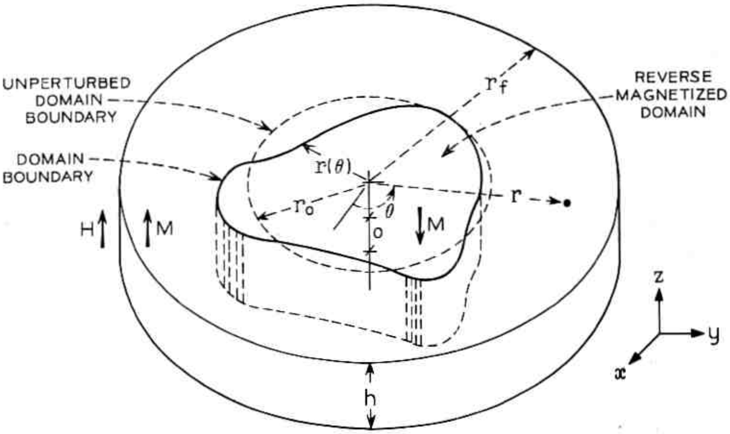


Fig. 1—Domain configuration and coordinate system.

tions only the component of the spatially uniform applied field which lies along the plate normal will interact with the domain so that only this component is considered. This component is denoted by H and is taken positive when directed upward, the direction tending to collapse the domain.

The assumptions implicit in this domain model are not all independent. The interrelation of the assumptions and the dependence of the assumptions on the domain geometry and the material parameters are discussed on pp. 3312–3318 of Ref. 1. The validity of the assumptions will not be discussed further here except to note that in order for domains of the type to be considered here to exist, domain nucleation and wall-width considerations require roughly that

$$K_u > 2\pi M_s^2 , \tag{1}$$

where K_u is the uniaxial anisotropy constant, and that the validity of the approximations generally improves as K_u is increased above this minimum value.

The domain radius function, $r_b(\theta)$, which is expanded in the series

$$r_b(\theta) = r_0 + \Delta r_0 + \sum_{n=1}^{\infty} \Delta r_n \cos [n(\theta - \theta_n - \Delta\theta_n)], \tag{2}$$

describes the domain shape in the plane. The Δr_n and $\Delta\theta_n$ describe a variation in domain size and shape from a circular domain of radius $r_b(\theta) = r_0$ and the θ_n describe the direction of the variation. (The $\Delta\theta_n$ have significance only for second variations.) Since only near circular domains are of interest here, the condition

$$|r_0| \gg |\Delta r_0| + \sum_{n=1}^{\infty} n\, |\Delta r_n| \tag{3}$$

is imposed to assure that the radius is single valued and smooth.

The first and second variations of the total domain energy with respect to the Δr_n and $\Delta\theta_n$ determine the domain equilibrium and stability conditions. The total domain energy,

$$E_T = E_W + E_H + E_M, \tag{4}$$

is the sum of three terms. The wall energy, E_W, is the product of the wall energy density and the wall area. The applied field interaction energy, E_H, is proportional to the product of the domain volume and the external field interaction energy. The last term, E_M, is the internal magnetostatic energy of the domain. The energy variation has the form

$$\Delta E_T = \sum_{n=0}^{\infty} \left[\left(\frac{\partial E_T}{\partial r_n} \right)_0 \Delta r_n + \left(\frac{\partial E_T}{\partial \theta_n} \right)_0 \Delta \theta_n \right]$$
$$+ \frac{1}{2} \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \left[\left(\frac{\partial^2 E_T}{\partial r_n\, \partial r_m} \right)_0 \Delta r_n\, \Delta r_m + 2 \left(\frac{\partial^2 E_T}{\partial r_n\, \partial \theta_m} \right)_0 \Delta r_n\, \Delta \theta_m \right.$$
$$\left. + \left(\frac{\partial^2 E_T}{\partial \theta_n\, \partial \theta_m} \right)_0 \Delta \theta_n\, \Delta \theta_m \right] + O_3 \tag{5}$$

where in the energy derivatives the independent variables, the Δr_n and $\Delta \theta_n$, have been written as r_n and θ_n for compactness, the zero subscript indicates that the derivatives are to be evaluated for a strictly circular domain, $r_b(\theta) = r_0$, and O_3 indicates terms of order three or higher. The first derivatives are the negative of the generalized forces, $-(\partial E_T/\partial r_n)_0$ and $-(\partial E_T/\partial \theta_n)_0$ being respectively the radial and angular generalized forces of rotational periodicity n. The second derivatives form the stiffness matrix of the system with $(\partial^2 E_T/\partial r_n\, \partial r_m)_0$ being the $(n,\ m)$ element of the radial submatrix and $(\partial^2 E_T/\partial r_n\, \partial \theta_m)_0$ and $(\partial^2 E_T/\partial \theta_n\, \partial \theta_m)_0$ being the corresponding elements of the mixed and angular submatrices respectively. In Ref. 1 the derivatives of the total energy with respect to the Δr_n and $\Delta \theta_n$ were obtained by differentiating the integrals which form the terms of equation (4) and evaluating the resulting integrals for the case of a strictly circular domain.

In the present work it is convenient to write the energy variation expression in a normalized form in which: energy is measured in units of $4(2\pi M_s^2)(\pi h^3)$, the equilibrium condition has been used to eliminate the applied field from the second variations (the stiffness matrix is of interest here only for a domain in equilibrium), the first- and second-order variations which are identically zero and deleted, and the nonzero generalized forces and stiffness matrix elements are written as functions of dimensionless variables, the applied field measured in units of the magnetization, $H/4\pi M_s$, and ratios of the plate thickness, h, the domain diameter, $d = 2r_0$, and the characteristic length of the material,

$$l = \sigma_w/4\pi M_s^2\ . \tag{6}$$

This normal form of the energy variation is [Ref. 1, equation (68)]

$$\frac{\Delta E_T}{4(2\pi M_s^2)(\pi h^3)} = \left[\frac{l}{h} + \frac{d}{h} \frac{H}{4\pi M_s} - F\left(\frac{d}{h} \right) \right] \frac{\Delta r_0}{h}$$

$$+\frac{1}{2}\left\{-\left(2\frac{h}{d}\right)\left[\frac{l}{h} - S_0\left(\frac{d}{h}\right)\right]\left(\frac{\Delta r_0}{h}\right)^2\right.$$

$$\left.+\sum_{n=2}^{\infty}(n^2 - 1)\left(\frac{h}{d}\right)\left[\frac{l}{h} - S_n\left(\frac{d}{h}\right)\right]\left(\frac{\Delta r_n}{h}\right)^2\right\} + O_3 \qquad (7)$$

where F and the S_n are called the force and stability functions respectively and are written in terms of the complete elliptic integral of the first kind,

$$K(m) \equiv \int_0^{\pi/2} (1 - m\sin^2\theta)^{-\frac{1}{2}}\,d\theta, \qquad (8)$$

the complete elliptic integral of the second kind,

$$E(m) \equiv \int_0^{\pi/2} (1 - m\sin^2\theta)^{\frac{1}{2}}\,d\theta \qquad (9)$$

and the L_n functions which are auxiliary functions introduced for convenience,

$$L_0(x) = 0 \qquad (10a)$$

$$L_1(x) = 4[(x+1)^{\frac{1}{2}}E((1+x)^{-1}) - x(x+1)^{-\frac{1}{2}}K((1+x)^{-1})] \qquad (10b)$$

$$L_{n+1}(x) = \frac{1}{2n+1}[4n(2x+1)L_n(x) - (2n-1)L_{n-1}(x)$$

$$- 16nx(x+1)^{-\frac{1}{2}} \times K((1+x)^{-1})], \qquad n \geqq 1. \qquad (10c)$$

The force and stability functions are

$$F(x) = \frac{2}{\pi}x^2[(1+x^{-2})^{\frac{1}{2}}E((1+x^{-2})^{-1}) - 1], \qquad (11)$$

$$S_0(x) = F(x) - x\frac{\partial}{\partial x}F(x), \qquad (12a)$$

$$= -\frac{1}{2\pi}x^2[L_1(x^{-2}) - L_1(0)], \qquad (12b)$$

and

$$S_n(x) = -\frac{1}{n^2-1}\frac{1}{2\pi}x^2[L_n(x^{-2}) - L_1(x^{-2}) - L_n(0) + L_1(0)],$$

$$n \geqq 2. \qquad (13)$$

Figure 3 of Ref. 1 plots $F(d/h)$ and $S_n(d/h)$ for $d/h \leqq 6$ and $n \leqq 10$.

Appendices A and B of Ref. 1 give methods of computation and series expansions of these functions while Section IV discusses the physical interpretation of the terms in the energy variation expansion [equation (5) here].

Section II of this paper discusses the domain size and stability implications of the energy variation expansion (7). This discussion yields the conditions for the existence of cylindrical domains in the total absence of dissipative processes and these existence conditions in turn place several restrictions upon device design. The section concludes with a discussion of domain size and domain ellipticity in the presence of anisotropic wall energy. Section III considers the effects of dissipative processes (wall motion coercivity and mobility) on domain existence and movement, the relation of domain mobility to wall mobility, and the limiting conditions under which the hard bubble mode may be achieved. Section IV combines the results obtained here and in Ref. 1 to obtain several relations between material parameters and device performance.

II. CYLINDRICAL DOMAIN SIZE, SHAPE AND STABILITY AT EQUILIBRIUM

This section examines the domain equilibrium and stability conditions and some of their implications.

2.1 *The Equilibrium Condition*

The domain is in equilibrium when all of the first-order variations of the total energy with respect to the Δr_n and $\Delta \theta_n$ are zero. In the variation expansion (7), all the first-order energy variations except the variation with respect to Δr_0 (representing a variation in domain size with no variation in shape) are identically zero. Since the domain is initially assumed to be a circular cylinder and there are no forces tending to deform it, the domain is in equilibrium when it is a circular cylinder having a diameter which is a solution to the normalized force equation,

$$\frac{l}{h} + \frac{d}{h}\frac{H}{4\pi M_s} - F\left(\frac{d}{h}\right) = 0. \tag{14}$$

The (normalized) generalized forces appearing in this equation have a one to one correspondence with the terms of the energy sum (4): the first term being produced by the wall energy, the second by the applied field, and the third by the internal magnetostatic energy. The normalized wall force, $-l/h = -\sigma_w/4\pi M_s^2 h$ always tends to collapse the domain (σ_w and h are positive) and is independent of domain diameter. Each of the normalized generalized forces may be converted to an equivalent field by multiplying by $h4\pi M_s/d$. The equivalent wall field is $\sigma_w/2r_0M_s$ so that the wall field is proportional to the product of the wall energy and the wall curvature. The normalized applied field force, $-(d/h)(H/4\pi M_s)$, tends to collapse the domain for positive applied fields and its magnitude is proportional to the domain diameter. In this case the equivalent field is the applied field, H. The normalized internal magnetostatic force is $+ F(d/h)$. The force function, F, is positive with positive first derivative and negative second derivative at all points. It approaches d/h for small domain diameters and $\pi^{-1} \ln | 4e^{\frac{1}{2}}d/h |$ for large domain diameters (In Ref. 1 see equation (138) and Fig. 3). Since F is positive, the internal magnetostatic force always tends to expand the domain. The internal magnetostatic force and the wall force are thus oppositely directed. Therefore equilibrium is attained for any given diameter by adjusting the applied field to a value which compensates for the difference in magnitude of these two forces, the sign of the field depending on which force is dominant. The equivalent field for the case of the internal magnetostatic force, $4\pi M_s \cdot (h/d)F(d/h)$, is the z averaged z component of the internal demagnetizing field or the internal magnetostatic scalar potential difference between the top and the bottom of the plate divided by the plate thickness.[1] This field approaches $4\pi M_s$ for small domain diameters and $4M_s(h/d) \ln | 4e^{\frac{1}{2}}d/h |$ for large domain diameters.

Solutions to the equilibrium problem may be discussed either in terms of the equivalent fields as was done by A. H. Bobeck[3] or in terms of the generalized force equation (14) with the preferred method depending on the specific application. In the present case, the generalized force equation will be used since the general properties of the solutions of the equilibrium problem may be easily obtained by straight line constructions on a plot of F.

Figure 2 shows examples of such constructions (along with stability constructions which will be explained later). The equilibrium construction consists of first locating the point on the vertical axis whose ordinate is l/h, then drawing a straight line through this point whose numerical slope is $H/4\pi M_s$. The line so constructed thus represents the first two terms in equation (14) so that its intersections with F (if any) are the equilibrium points. The dashed line asymptotic to the force function at the origin has numerical slope one and therefore provides a reference for estimating the magnitude of applied fields. In each of the constructions of Fig. 2, l/h is 0.3 and this point on the vertical axis is

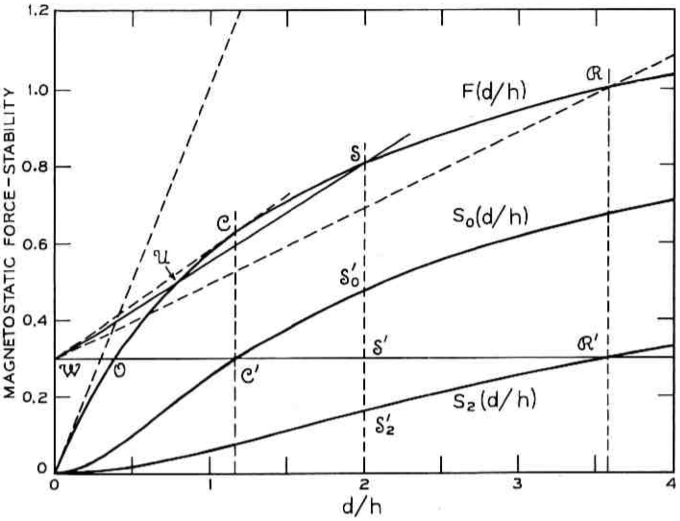


Fig. 2—Construction of solutions to the force equation for $l/h = 0.3$, $d = d_0$, $d = d_2$ and $d = 2h$.

denoted by 𝒲. The symbol 𝒲 was used since l/h is the wall energy per unit area in units of twice the magnetostatic self-energy per unit surface area of an infinite plate of uniform thickness, h, when it is uniformly magnetized in a direction perpendicular to the surface of the plate.

For the first example, consider the line 𝒲ℛ′ for which $l/h = 0.3$, $H/4\pi M_s = 0$. This line has one intersection with F at 𝒪($d/h = 0.378$). Since the slope of F is everywhere positive, there will in general be only one intersection with the construction line for any zero or negative value of $H/4\pi M_s$ independent of the magnitude of l/h. That this single solution is unstable may be appreciated by considering small variations in the solution diameter from its value at 𝒪: Increasing the diameter of the domain from the diameter at 𝒪 increases the magnitude of the internal magnetostatic force term while the other terms remain constant. The total force tending to expand the domain thus increases from zero as the domain expands from the solution diameter. The domain is thus unstable with respect to expansion. An entirely similar argument shows that the domain is unstable with respect to collapse. It is easily seen in general that for a fixed solution diameter the solution will become more unstable (the destabilizing force increasing faster with increasing diameter) as the field is made more negative. Stable solutions to the equilibrium problem exist only for fields aiding the wall field in tending to collapse the domain although one unstable solution exists for any magnitude of negative applied field.

Because of the logarithmic behavior of the force function at large domain diameters, a radially stable solution to the force equation will appear (in addition to the radially unstable solution) when a vanishingly small positive bias field is applied. (Radial stability is determined in the same way as in the preceding paragraph. The existence of radial stability does not assure total domain stability as will be seen.) Since the F function has a negative second derivative everywhere, there can be no solutions in addition to a single radially unstable and a single radially stable solution. This situation is illustrated by the line $\mathcal{WS}$ in Fig. 2. The line represents $l/h = 0.3$, $H/4\pi M_s = 0.254$ for which the unstable solution, denoted by $\mathcal{U}$, occurs at $d/h = 0.767$ and the stable solution, denoted by $\mathcal{S}$, occurs at $d/h = 2.000$. Bobeck[7] has observed experimentally the unstable solution under static conditions which permit the existence of the stable solution. This was done by applying a short duration bias field pulse which reduced the domain diameter from the stable solution diameter at $\mathcal{S}$ to the unstable solution diameter at $\mathcal{U}$ and then returning the bias field to its original value just as the unstable solution was attained. The domain having the unstable solution diameter was then stabilized sufficiently by coercivity to allow it to persist.

If the applied field is increased from a value for which there are two solutions, then, the solutions approach each other until at some field value they coalesce. This point at which the solutions coalesce is a point of radial metastability since, at this point, the construction line is tangent to F. Thus, the solution is neither stable nor unstable to lowest order by the small variation argument of the preceding paragraphs. There are no solutions for applied fields greater than the field of radial metastability. Since in this case the inward forces dominate at all diameters, the domain collapses. (Since F everywhere lies below the asymptote d/h, the domains always collapse for applied fields greater than $4\pi M_s$.) In Fig. 2 where $l/h = 0.3$, the collapse point, $\mathcal{C}$, occurs at $H/4\pi M_s = 0.283$, $d/h = 1.16$. The sequence of the types of solutions which occur as H is varied depends only on the invariant signs of the slope and curvature of F. Therefore for a fixed value of l/h this sequence is independent of the value of l/h. From the slope and curvature properties of F the uniqueness of the collapse diameter and field for a given value of l/h may also be shown. However, the discussion of both this uniqueness and the detailed behavior of the domain diameter as a function of applied field are conveniently postponed until general stability is discussed.

The device implications of this subsection may be summarized by noting that cylindrical domains exist only in the presence of an applied field applied in a direction tending to collapse the domain and having a magnitude less than $4\pi M_s$.

2.2 *General Stability*

The sign of the second variation of the total domain energy produced by a weak variation in shape characterizes the stability of a cylindrical domain. Since only domains in equilibrium are of interest here, attention is restricted to variations of such domains from their equilibrium size and strictly circular shape. Since an arbitrary weak variation is describable by the expansion (2) with condition (3), the stability problem is reduced to the study of the coefficients of the terms in the energy expansion which are quadratic in the Δr_n and $\Delta \theta_n$. These coefficients are defined (up to some constant factor) as the stiffness matrix elements of the system, and these stiffness matrix elements may clearly be classified as either radial, mixed, or angular stiffness matrix elements.

In the equilibrium energy expansion (7) the only nonzero quadratic coefficients are the coefficients of the $(\Delta r_n)^2$, $n \neq 1$. As required by the cylindrical symmetry of the system, the domain is completely metastable with respect to angle. This is indicated formally in (7) by the absence of any nonzero terms in $\Delta\theta_n \Delta\theta_m$ or $\Delta r_n \Delta\theta_m$. (The angular and mixed stiffness submatrices are identically zero.) Since no terms in $\Delta r_n \Delta r_m$ for $m \neq n$ appear in equation (7), (the radial stiffness matrix is diagonal) the variation amplitudes are quasi-normal-modes of the system. Nonzero terms of order three and higher in the variation amplitudes and angles prevent the variation amplitudes being true normal modes. Finally, since the energy of a domain in an infinite plate is independent of the domain position, the translational stiffness of the domain is zero and therefore since the variation Δr_1 corresponds to lowest order to a translation (see Ref. 1 Sec. 4.2.2), the coefficient of $(\Delta r_1)^2$ in equation (7) is zero.

Since the second-order energy variation with respect to an arbitrary small amplitude weak variation is simply the sum of the energy variations from each normal component of the variation, the study of the stability of circular cylindrical domains reduces to the study of the stability of the domains with respect to size, Δr_0, and shape, Δr_2 to Δr_∞ . The sign of the corresponding stiffness matrix element determines the stability or instability of the domain with respect to the variation of a particular Δr_n. From equation (7), the nonzero normal stiffness matrix elements are

$$\frac{h^2}{4(2\pi M_s^2)(\pi h^3)}\left(\frac{\partial^2 E_T}{\partial r_0^2}\right)_0 = -\left(2\,\frac{h}{d}\right)\left[\frac{l}{h} - S_0\left(\frac{d}{h}\right)\right] \tag{15a}$$

and

$$\frac{h^2}{4(2\pi M_s^2)(\pi h^3)}\left(\frac{\partial^2 E_T}{\partial r_n^2}\right)_0 = (n^2 - 1)\left(\frac{h}{d}\right)\left[\frac{l}{h} - S_n\left(\frac{d}{h}\right)\right], \qquad n \geqq 2. \tag{15b}$$

The quantities in square brackets are termed stability coefficients. The domain is stable with respect to an arbitrary variation in shape when all of the stiffness matrix elements are positive. From equation (15), this occurs when the $n = 0$ coefficient is negative and all the other stability coefficients are positive,

$$[l/h - S_0(d/h)] < 0 \tag{16a}$$

and

$$[l/h - S_n(d/h)] > 0, \qquad n \geqq 2. \tag{16b}$$

or

$$S_0(d/h) > l/h > S_2(d/h). \tag{17}$$

Since the stability functions have the property,

$$S_{n+1}(d/h) < S_n(d/h), \tag{18}$$

(at least up to $n = 10$ see Ref. 1) the condition for total stability reduces to

$$S_0(d/h) > l/h > S_2(d/h). \tag{19}$$

Since domain stability with respect to Δr_0 and Δr_2 assures total stability, attention is largely restricted to the $n = 0$ and $n = 1$ coefficients. The variations Δr_0 and Δr_2 are termed radial and elliptical variations respectively.

The numerical value of the stability coefficients and in particular the conditions under which the stability coefficients change sign, is determined by a graphical construction, an example of which is now given. The construction of the radial and elliptical stability coefficients for the case $l/h = 0.3$ is shown in Fig. 2. The elevation of the horizontal line $\mathcal{W}\mathcal{R}'$ represents the value of l/h. In order to determine the stability coefficients it is necessary to specify an operating diameter which, in turn, is determined by the applied field. When the domain diameter is greater than the diameter at the intersection of the horizontal line and S_2 at $\mathcal{R}'$ (applied fields less than that represented by the slope of the line $\mathcal{W}\mathcal{R}$) the elliptical stability coefficient is negative and the domain is unstable with respect to Δr_2 . The domain diameter at the intersection of the horizontal line and S_2 is called the diameter of elliptical metastability and the corresponding field (the field represented by the slope of the line $\mathcal{W}\mathcal{R}$) is termed the field of elliptical metastability. When the diameter is decreased below the diameter of elliptical metastability (by increasing the field above the field of elliptical metastability) the domain becomes stable with respect to elliptical deformation. When the domain diameter is decreased to $d = 2h$ (the corresponding field being represented by the slope of the line $\mathcal{W}\mathcal{S}$) the magnitude of the radial stability coefficient (corresponding to the length of $\mathcal{S}'\mathcal{S}_0'$) is approximately equal to the magnitude of the elliptical stability coefficient (corresponding to the length of $\mathcal{S}'\mathcal{S}_2'$).

Increasing the field to the value corresponding to the slope of the line $\mathcal{W}\mathcal{C}$, decreases the domain diameter further to the value at the intersection of the horizontal construction line with $\mathcal{S}_0$ at $\mathcal{C}'$, so that the radial stability coefficient is zero and the force equation construction line is tangent to the force curve. Increasing the field above the field value corresponding to the slope of $\mathcal{W}\mathcal{C}$ decreases the diameter even further so that the radial stability coefficient becomes positive, the radial stiffness matrix element becomes negative, there are no solutions to the force equation and the domain collapses.

Since the construction line for the force equation is tangent to the force curve at the diameter of radial metastability, it is possible to construct the S_0 curve from the F curve by plotting the loci of points whose ordinate is l/h and whose abscissa is the diameter at which the construction line is tangent to the F curve. Figure 3 illustrates four points of such a construction. Conversely, given the initial slope of F, the S_0 curve may be used to construct F.

In Appendix A of Ref. 1 (see also Fig. 2 here) it was shown that the stability functions are in general monotonic increasing functions of the normalized domain diameter, d/h, and monotonic decreasing functions of the radial periodicity, n (at least up to $n = 10$). From these properties of the S_n functions and the properties of the solution to the force equation discussed in Section 2.1, the following general stability properties may be deduced for any fixed value of l/h: When there is no applied field, the domain diameter is infinite, and the domain is unstable with respect to all variations for which $n \geqq 2$.

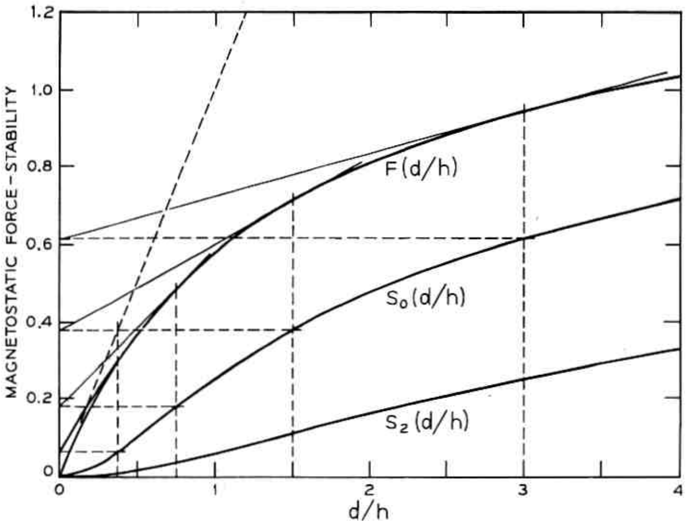


Fig. 3—Construction of the $S_0(d/h)$ function as a sequence of collapse diameter solutions.

When the applied field is small, the diameter is finite and the domain is stable with respect to all variation for which n is equal or greater than some n_m, while remaining unstable with respect to variations of r_n for which $2 \leqq n < n_m$. When the applied field is between the values of elliptical and radial metastability, the domain is stable with respect to all variations. Finally, for applied fields greater than the field of radial metastability, the domain collapses. (As noted previously the domain collapses for applied fields greater than the fields of radial metastability for any value of the domain diameter.)

Note that except for S_1 each S_n forms the boundary between the regions of stability and instability with respect to the corresponding Δr_n and that metastability with respect to each Δr_n $(n \neq 1)$ occurs only along the boundary between the stable and unstable regions.

Even though the radial and elliptical stability functions bound the region of total domain stability, the threefold and fourfold stability functions, S_3 and S_4, lie quite close to S_2 and it might be expected that when the bias field on a cylindrical domain is reduced in the presence of a small stabilizing coercivity the domain might run out into an initially three or fourfold figure. Such runouts have indeed been observed.[8]

When the bias field on an initially circular stable domain, represented for example by the line $\mathcal{WS}$ in Fig. 2, is decreased, $r_b(\theta)$ will increase uniformly for all θ maintaining a circular shape until the point $\mathcal{R}$ is reached. Increasing the domain radius beyond this value causes the domain to become unstable with respect to elliptical variations. Since for small variations, the variations in the expansion used here are normal modes, the breakaway to instability will not be coupled to any pure radial motion. Thus, when the breakaway occurs $r_b(\theta)$ will actually decrease along one half the length of the wall. Experiments in platelets which have just enough coercivity so that the initiation of the breakaway is observable confirm this somewhat surprising prediction. (See Ref. 4, pp. 1916–1917. The sequence of photographs on these pages were taken with an increasing bias field. The corresponding sequence for a decreasing bias field is similar.)

2.3 *Restrictions Placed on the Possible Region of Device Operation by Stability Considerations*

Since the diameters of radial and elliptical metastability (and the corresponding applied fields) are the boundaries of the region of total stability, they are the boundaries of the region of possible device operation in the hard bubble mode. It is easily appreciated by inspection of equation (15) that the domain will be metastable with respect to a particular Δr_n $(n \neq 1)$ if and only if the corresponding stability coefficient is zero,

$$l/h - S_n(d/h) = 0. \tag{20}$$

Since the S_n are monotonic increasing functions of d/h, a diameter of metastability is uniquely defined for each value of n and characteristic length to thickness ratio,

$$d_n/h = S_n^{-1}\,(l/h), \qquad n \neq 1. \tag{21}$$

The corresponding applied fields of metastability are then defined using the force equation (14) by

$$\frac{H_n\left(\frac{l}{h}\right)}{4\pi M_s} = \frac{h}{d_n}\left[F\left(\frac{d_n}{h}\right) - \frac{l}{h}\right]. \tag{22}$$

While l/h is a normalized wall energy, it is also, obviously, the reciprocal of the thickness measured in units of the characteristic length. Thus, the d_n and H_n are functions of the normalized thickness, h/l. In the remaining topics of this section it will be appropriate to think of this parameter as the normalized thickness, since in a given material it is the thickness which is the accessible variable.

In principle, the values of the d_n and H_n for each value of h/l could be obtained by graphical construction. Figure 2 may be taken as an example of such a construction for $h/l = 3.33$ and $n = 0$ and 2. In this construction the intersection at $\mathcal{C}'$ represents $d_0(0.3) = 1.16\text{h}$, the intersection at $\mathcal{R}'$ represents $d_2(0.3) = 3.58h$, the slope of $\mathcal{WC}$ represents $H_0(0.3) = 0.283(4\pi M_s)$ and the slope of $\mathcal{WR}$ represents $H_2(0.3) = 0.196(4\pi M_s)$. (The values quoted here were obtained numerically.)

To better appreciate the way in which the requirements of static stability bound the possible region of device operation, several diameter and applied field functions of the normalized plate thickness have been plotted and their asymptotic forms in the limit of very thick or very thin plates have been computed.

The values used in plotting the functions were obtained in the following way: The implicit equation for d_0 and d_2 [equation (20)] was inverted numerically using the expressions for $S_0(d/h)$ [equation (12)] and $S_2(\text{d/h})$ [equation (13)] to obtain d_0/h and d_2/h as functions of h/l. (Parametric plotting may be used only for functions of a single d_n.) The desired functions were then computed from d_0/h, d_2/h and the expressions for the H_n [equation (22)] and $F(d/h)$ [equation (11)]. (See also Ref. 1, Sections A.1 and A.8.) The asymptotes of the various curves are obtained using the expansions for $F(d/h)$, $S_0(d/h)$ and $S_2(d/h)$ from Appendix B of Ref. 1, with the force equation where necessary (14) and the d_n defining equation (20). The small d/h expansions are carried to order $(d/h)^2$ and the large d/h expansions are carried only to the lowest order constant and log terms in order to facilitate the algebraic inversion of equation (20).

Since the validity of some of the assumptions of the theory increases with increasing d/h, each plot includes an arrow indicating the direction of increasing d_n/h along the plotted curve or curves. The marks at $d_0/h = 1$ and $d_2/h = 1$ indicate the points at which the theory becomes definitely suspect. It has been found experimentally, however, that the theory does give reasonably consistent results for values of d/h which are somewhat less than one.

Figure 4 is a plot of the diameters of radial and elliptical metastability measured in units of the characteristic length, $d_0/l = (h/l)(d_0/h)$ and $d_2/l = (h/l)(d_2/h)$, as functions of the thickness measured in units of the characteristic length, h/l. These diameters have the asymptotic values

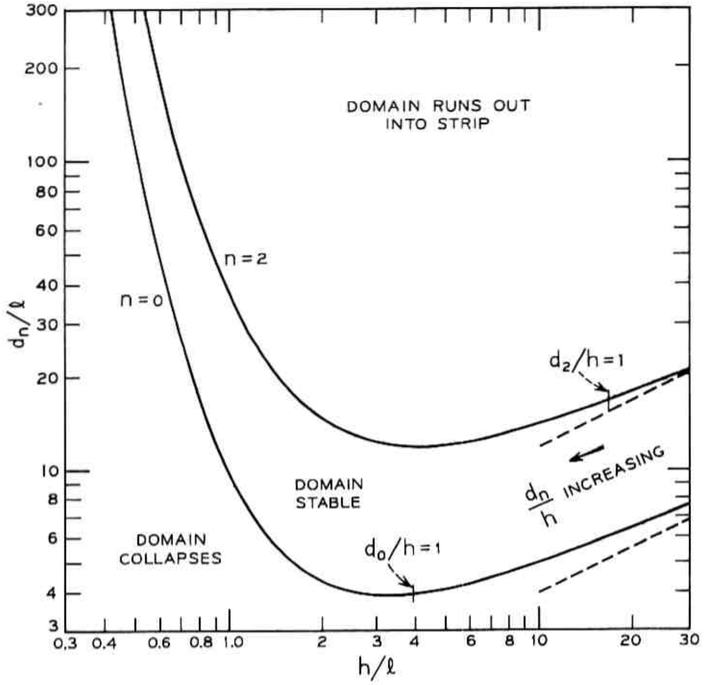


Fig. 4—Diameters of radial and elliptical metastability, d_0 and d_2, in units of the characteristic length, l, as a function of the thickness, h, in units of l.

$$d_0/l \approx 0.412(h/l) \exp (\pi l/h), \qquad h/l \ll 1, \tag{23a}$$

and

$$d_0/l \approx 1.253(h/l)^{\frac{1}{2}}, \qquad h/l \gg 1, \tag{23b}$$

for the collapse diameter, and

$$d_2/l \approx 1.564(h/l) \exp (\pi l/h), \qquad h/l \ll 1, \tag{24a}$$

and

$$d_2/l \approx 3.760(h/l)^{\frac{1}{2}}, \qquad h/l \gg 1, \tag{24b}$$

for the elliptical runout diameter. The uppermost curve in Fig. 5 is a plot of the ratio of the diameter of elliptical metastability to the diameter of radial metastability, $d_2/d_0 = (d_2/h)/(d_0/h) = (d_2/l)/(d_0/l)$, as a function of the h/l. This ratio has the aymptotes

$$d_2/d_0 \approx 3.794, \qquad h/l \ll 1 \tag{25a}$$

and

$$d_2/d_0 \approx 3.000, \qquad h/l \gg 1 \tag{25b}$$

which are easily obtainable from equations (23) and (24). Figure 6 is a plot of the relative variation of the geometric mean of the diameters of radial and elliptical metastability with respect to a relative variation in plate thickness $\partial \ln|(d_0d_2)^{\frac{1}{2}}|/\partial \ln|h|$ as a function of h/l. This curve has the asymptotes

$$\partial \ln |(d_0d_2)^{\frac{1}{2}}|/\partial \ln |h| \approx 1.000 - 3.142(h/l)^{-1}, \qquad h/l \gg 1, \tag{26a}$$

and

$$\partial \ln |(d_0d_2)^{\frac{1}{2}}|/\partial \ln |h| \approx 0.500, \qquad h/l \gg 1. \tag{26b}$$

Inspection of these limiting diameter plots and asymptotic expressions yields the following: In a given material, the minimum stable

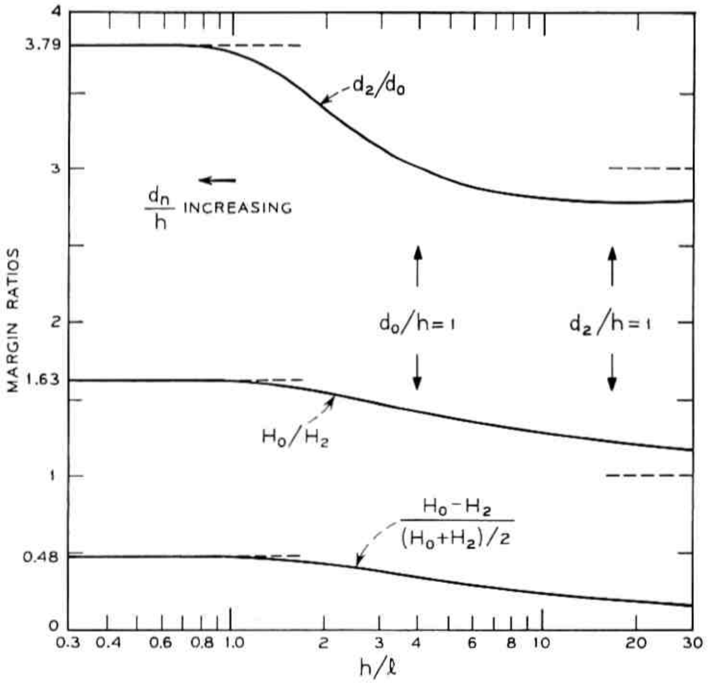


Fig. 5—Diameter and applied field margin ratios, d_2/d_0, H_0/H_2, and $2(H_0 - H_2)/(H_0 + H_2)$, as functions of the thickness, h, measured in units of the characteristic length, l.

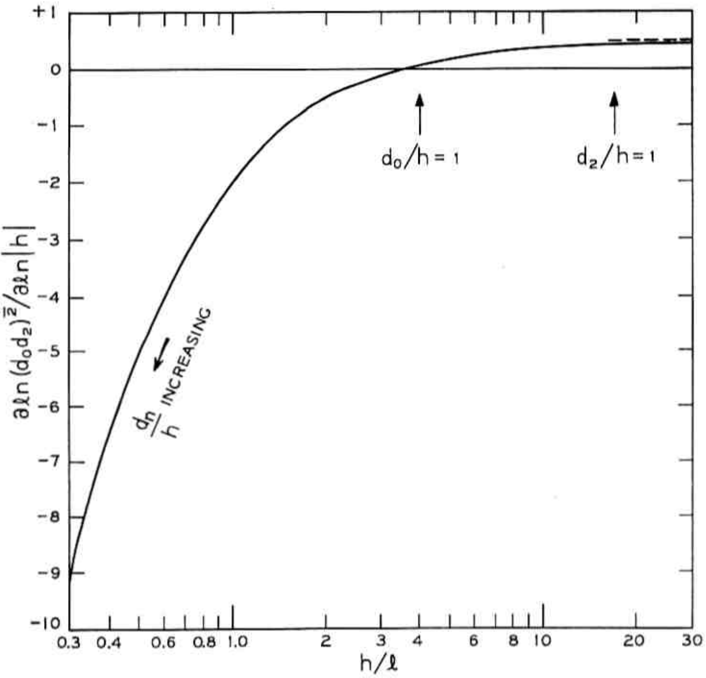


Fig. 6—Logarithmic derivative of the domain diameter with respect to plate thickness, $\partial \ln d / \partial \ln |h|$, for the bias condition $d = (d_0 d_2)^{\frac{1}{2}}$ as a function of the plate thickness, h, measured in units of the characteristic length, l.

diameter attainable for any plate thickness is $\approx 3.9l$ at a thickness of $\approx 3.3l$. The minimum diameter of elliptical instability is $\approx 12l$ attained at a thickness of $\approx 4.2l$. For any given plate thickness, the range of diameters over which the domain is stable is small, the ratio of the diameter of elliptical metastability to the diameter of radial metastability being roughly equal to three for any plate thickness. The domain diameter is thus for all practical purposes determined once the plate thickness and characteristic length are known. The increase in domain diameter with increasing plate thickness in thick plates is quite mild being according to a square root law. The increase in domain diameter with decreasing plate thickness is, on the other hand, exponential for a thin plate. This variation is so rapid that the magnitude of the relative variation of the diameter of a centrally biased domain, $d = (d_0 d_2)^{\frac{1}{2}}$ with respect to a relative variation in thickness increases according to the inverse of the thickness.

Figure 7 is a plot of the applied fields of radial and elliptical metastability, $H_0/4\pi M_s = (h/d_0)[F(d_0/h) - l/h]$ and $H_2/4\pi M_s = (h/d_2)[F(d_2/h) - l/h]$, as functions of h/l. These curves have the asymptotes

$$H_0/4\pi M_s \approx 0.772 \exp(-\pi l/h), \qquad h/l \ll 1, \tag{27a}$$

and

$$H_0/4\pi M_s \approx 1.000 - 1.596(h/l)^{-\frac{1}{2}}, \qquad h/l \gg 1, \tag{27b}$$

for the collapse field and

$$H_2/4\pi M_s \approx 0.475 \exp(-\pi l/h), \qquad h/l \ll 1, \tag{28a}$$

and

$$H_2/4\pi M_s \approx 1.000 - 2.660(h/l)^{-\frac{1}{2}}, \qquad h/l \gg 1, \tag{28b}$$

for the elliptical runout field. The two lowermost curves in Fig. 5 are

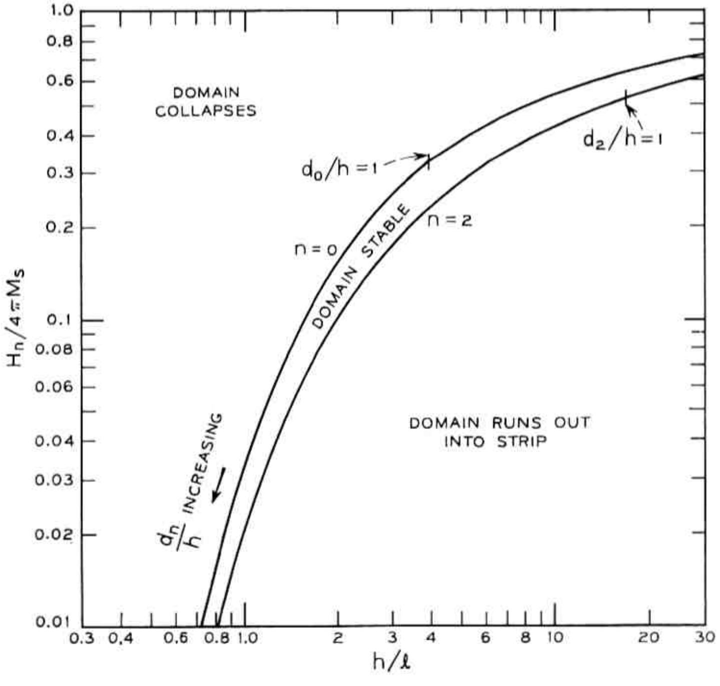


Fig. 7—Applied fields of radial and elliptical metastability, H_0 and H_2, in units of $4\pi M_s$ as a function of thickness, h, measured in units of the characteristic length, l.

plots of the ratio of the collapse field to the elliptical runout field and the ratio of the difference of these fields to their average as functions of h/l. The asymptotes of these ratios are

$$H_0/H_2 \approx 1.626, \qquad h/l \ll 1, \tag{29a}$$

and

$$H_0/H_2 \approx 1.000 + 1.064(h/l)^{-\frac{1}{2}}, \qquad h/l \gg 1, \tag{29b}$$

for the field ratio and

$$2(H_0 - H_2)/(H_0 + H_2) \approx 0.477, \qquad h/l \ll 1, \tag{30a}$$

and

$$2(H_0 - H_2)/(H_0 + H_2) \approx 1.064(h/l)^{-\frac{1}{2}}, \qquad h/l \gg 1, \tag{30b}$$

for the ratio of the difference of the fields to their average. Equations (29) and (30) are easily obtainable from equations (27) and (28).

Inspection of these limiting applied field plots and asymptotes yields the following: The field magnitudes decrease exponentially with decreasing plate thickness, rapidly becoming unmanageably small for very thin plates. The fields increase monotonically with increasing plate thickness toward the common value, $4\pi M_s$. The relative variation of applied fields, allowed within the region of stable circular domains, is smaller than the allowed relative variation of diameters, approaching a constant for thin plates and zero for thick plates.

Note that $4\pi M_s$ determinations are most accurately carried out in moderately thick plates (within the limits of the cylindrical wall approximation and the coercivity limits described in the next section) since here $H_0/4\pi M_s$ (and $H_2/4\pi M_s$) is a weak function of l and is asymptotic to one. Measurements of the characteristic length, on the other hand, are best carried out in moderately thin plates (within coercivity limits) where the diameter is a strong function of l. The appreciation of the validity of these statements will be greatly enhanced if the reader will try a few constructions on a plot of the F and S_n functions.

The implications of the foregoing for device design are summarized as follows: For a given value of h/l there is only an approximately three-to-one variation in the domain diameter permitted within the stable region so that given material and plate thickness, domain size is closely determined. The minimum domain diameter occurs for $h/l \approx 4$. For thinner and thinner plates, both the plate thickness margins and the applied field magnitude decrease exponentially. While field margins are reasonable for $h/l \approx 4$ ($H_0/H_2 = 1.41$) [or $2(H_0 - H_2)/(H_0 + H_2) \approx 0.34$], they become vanishingly small in very thick plates. These considerations thus demonstrate that from stability considerations alone, for a given value of the characteristic length, there is both a preferred range of thicknesses and a preferred range of domain diameters (or applied fields) for device operation. The preferred range of each is centered about the preferred values

$$h_p = 4l \tag{31}$$

and

$$d_p = 8l \tag{32}$$

for which the bias field is

$$H_p = 0.279(4\pi M_s). \tag{33}$$

These values will be shown to be preferred in another sense in the next section.

2.4 *Domain Diameter as a Function of Applied Field*

Since domain diameter is in practice only measurable for diameters between d_0 and d_2 (applied fields between H_0 and H_2), these points form natural endpoints for plotting the diameter as a function of applied field. Figure 8 is a plot of $(d - d_0)/(d_2 - d_0)$, as a function of $(H - H_2)/(H_0 - H_2)$, for various values of the normal thickness in the infinite disk. The plots for finite nonzero values of the thickness were obtained numerically using the force equation (14), the defining relation for the d_n [equation (21)] and the expressions for F [equation (11)], S_0 [equation (12)] and S_2 [equation (13)]. In the limit of very thick plates, diameter and applied field are related by

$$\frac{H_0 - H}{H_0 - H_2} = \frac{3}{4}\frac{(d - d_0)^2}{dd_0}, \qquad h \to \infty \tag{34a}$$

using the small d/h expansions of F, S_0 and S_2 from Ref. 1, Appendix B, to order $(d/h)^2$. In the limit of very thin plates, diameter and applied field are related by

$$\frac{H_0 - H}{H_0 - H_2} = \frac{1}{1 - \frac{7}{3}\exp\left(-\frac{4}{3}\right)}\left[1 - \frac{d_0}{d}\left(1 + \ln\left|\frac{d}{d_0}\right|\right)\right], \qquad h \to 0 \tag{34b}$$

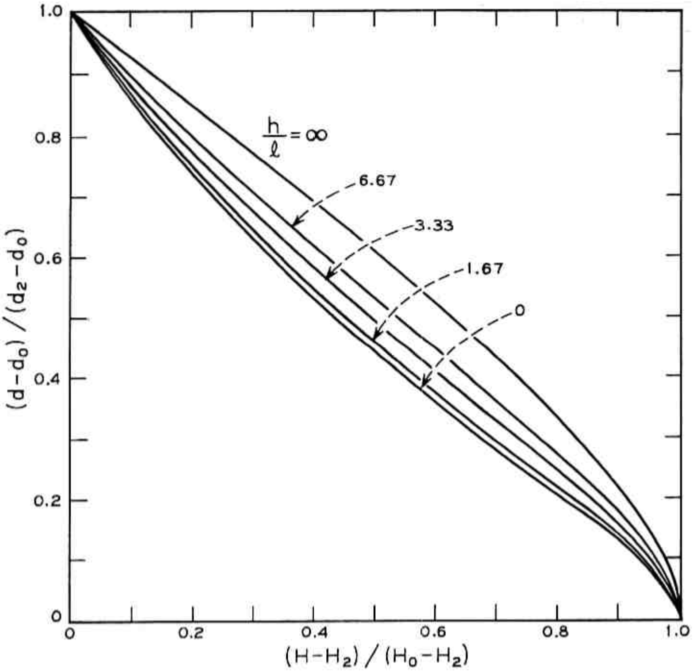


Fig. 8—Relative diameter, $(d - d_0)/(d_2 - d_0)$, as a function of relative applied field, $(H - H_2)/(H_0 - H_2)$, in the infinite plate for various values of the plate thickness, h, measured in units of the characteristic length, l.

using the constant and simple log terms of the expansions of F, S_0 and S_2 from Appendix B of Ref. 1. The curves for intermediate thicknesses are seen to lie, in order, between the limiting curves and are all nearly linear except in the neighborhood of the collapse diameter. In the neighborhood of the collapse diameter, the slopes of all the curves are infinite. This is easily appreciated by expressing the derivative of the domain diameter with respect to applied field in terms of the radial stability coefficient.

The derivative of the domain diameter with respect to the applied field is obtained by considering equation (14) to be continuously solved and then differentiating with respect to d to obtain

$$\frac{1}{4\pi M_s}\left(\frac{d}{h}\frac{\partial H}{\partial d} + \frac{1}{h}H\right) - \frac{\partial}{\partial d}F(d/h) = 0. \tag{35}$$

Eliminating H with the force equation, using the equation for S_0 (12)

and rearranging yields

$$\frac{4\pi M_s}{d}\frac{\partial d}{\partial H} = \frac{d}{h}\frac{1}{l/h - S_0\left(\frac{d}{h}\right)}. \tag{36}$$

So that indeed the derivative of diameter with respect to applied field approaches infinity as radial metastability is approached.

2.5 *The Effect of Wall Energy Anisotropy on the Size, Shape and Stability of Cylindrical Domains*

The domains observed in orthoferrites are never precisely circular but always have some degree of ellipticity. The present subsection relates this anisotropy in domain shape to more fundamental domain parameters.

In the orthoferrites $K_u \gg 2\pi M_s^2$ so that the magnetization lies rigidly along the plate normal and the wall width is narrow as compared to the domain diameter (1). The applied field energy, E_H and the internal magnetostatic energy, E_M , terms of the total energy expression (4) thus make no contribution to producing the domain anisotropy. Therefore the anisotropy results from an anisotropy in the wall energy density σ_w . Considering again the wall energy density to be independent of wall curvature and the wall to be oriented with its normal perpendicular to the plate normal, i.e., a cylindrical wall, the wall energy density may be expanded as

$$\sigma_w = \bar{\sigma}_w + \tfrac{1}{2}\sum_{n=1}^{\infty} \sigma_{2n} \cos\left[2n(\nu - \nu_{2n})\right] \tag{37}$$

where ν is the angle between the wall normal, **N**, and the x axis (see Fig. 9) and the expansion coefficients, $\bar{\sigma}_w$, σ_{2n} and ν_{2n} are taken to be positive. The odd angular periodicity expansion coefficients are deleted from equation (37) since the energy of the system is invariant under time reversal while the direction of the wall normal (referred to the magnetization direction) changes sign.

To describe the anisotropy observed in orthoferrites, it has proven sufficient to include only the average and two-fold terms in equation (37). In addition, if the plate is assumed oriented so that the wall energy is maximized when the wall normal lies along the x axis, the wall energy density is

$$\sigma_w = \bar{\sigma}_w + \tfrac{1}{2}\,\Delta\sigma_w \cos 2\nu. \tag{38}$$

Although the method which now will be employed to calculate the

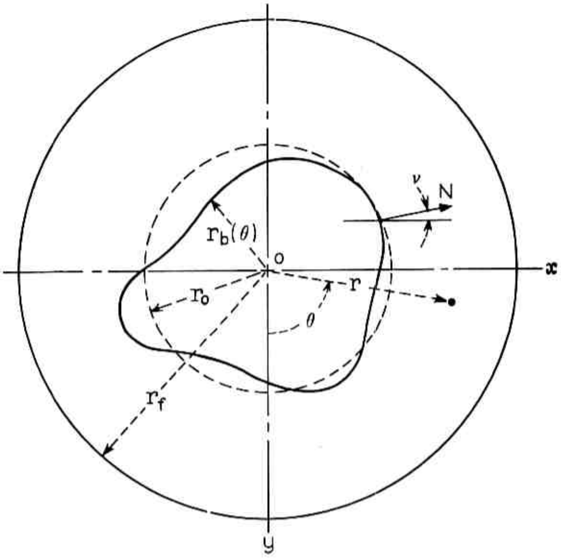


Fig. 9—Coordinate system used in consideration of anisotropic wall energy.

implications of equation (38) is clearly applicable to the more general energy expression (37), attention will be restricted to equation (38).

The total wall energy is

$$E_W = h \oint \sigma_w \, ds \qquad (39)$$

where s is the arc length along the curve describing the domain shape in the plane. When the wall energy density (38) is substituted into equation (39) and the differential arc length, ds, and the wall normal orientation angle, ν (see Fig. 9), are expressed in terms of θ, $r_b(\theta)$ and $\partial r_b(\theta)/\partial\theta$, the total wall energy expression becomes

$$\begin{aligned} E_W = {} & h\bar{\sigma}_w \int_0^{2\pi} \left[r_b^2 + \left(\frac{\partial r_b}{\partial \theta} \right)^2 \right]^{\frac{1}{2}} d\theta \\ & + \tfrac{1}{2} h \, \Delta\sigma_w \int_0^{2\pi} \left\{ \left[r_b^2 + \left(\frac{\partial r_b}{\partial \theta} \right)^2 \right]^{\frac{1}{2}} \cos 2\theta \right. \\ & \left. - 2 \frac{\partial r_b}{\partial \theta} \left(\frac{\partial r_b}{\partial \theta} \cos 2\theta - r_b \sin 2\theta \right) \left[r_b^2 + \left(\frac{\partial r_b}{\partial \theta} \right)^2 \right]^{-\frac{1}{2}} \right\} d\theta. \qquad (40) \end{aligned}$$

In the isotropic wall energy density case, $\sigma_w = \bar{\sigma}_w$, the first term in equation (40) is identical to the isotropic wall energy expression [equation (8) of Ref. 1], the second term in equation (40) representing entirely the effect of the wall energy anisotropy.

To evaluate the effect of the anisotropy term even to lowest order, it is necessary to obtain all the first and second derivatives with respect to the expansion coefficients, Δr_n and $\Delta\theta_n$ of the expansion of $r_b(\theta)$ (2) for the case of a strictly circular domain $r_b(\theta) = r_0$. In the expressions for the derivatives, $\partial\Delta r_n$ and $\partial\Delta\theta_n$ are again abbreviated to ∂r_n and $\partial\theta_n$ and evaluation at $r_b(\theta) = r_0$ is denoted by a zero subscript. The first derivative with respect to Δr_n is

$$\begin{aligned}\frac{\partial E_W}{\partial r_n} = {} & h\bar{\sigma}_w \int_0^{2\pi} \left\{ r_b \frac{\partial r_b}{\partial r_n} + \frac{\partial r_b}{\partial \theta} \frac{\partial\left(\dfrac{\partial r_b}{\partial\theta}\right)}{\partial r_n} \right\} \left[r_b^2 + \left(\frac{\partial r_b}{\partial\theta}\right)^2 \right]^{-\frac{1}{2}} d\theta \\ & + \tfrac{1}{2} h\, \Delta\sigma_w \int_0^{2\pi} \biggl\{ \left[r_b \frac{\partial r_b}{\partial r_n} + \frac{\partial r_b}{\partial \theta} \frac{\partial\left(\dfrac{\partial r_b}{\partial\theta}\right)}{\partial r_n} \right] \left[r_b^2 + \left(\frac{\partial r_b}{\partial\theta}\right)^2 \right]^{-\frac{1}{2}} \cos 2\theta \\ & - 2 \frac{\partial\left(\dfrac{\partial r_b}{\partial\theta}\right)}{\partial r_n} \left(\frac{\partial r_b}{\partial\theta} \cos 2\theta - r_b \sin 2\theta \right) \left[r_b^2 + \left(\frac{\partial r_b}{\partial\theta}\right)^2 \right]^{-\frac{1}{2}} \\ & - 2 \frac{\partial r_b}{\partial\theta} \frac{\partial}{\partial r_n} \left[\left(\frac{\partial r_b}{\partial\theta} \cos 2\theta - r_b \sin 2\theta \right) \left(r_b^2 + \left(\frac{\partial r_b}{\partial\theta}\right)^2 \right)^{-\frac{1}{2}} \right] \biggr\} d\theta. \qquad (41)\end{aligned}$$

Setting $r_b = r_0$ and $\partial r_b / \partial\theta = 0$ and carrying out the integration yields

$$\left(\frac{\partial E_W}{\partial r_n}\right)_0 = 2\pi h \bar{\sigma}_w\, \delta_{n0} - \frac{3\pi}{2} h\, \Delta\sigma_w\, \delta_{n2} \cos 2\theta_2 \qquad (42a)$$

where δ_{mn} is the Kronecker delta function. The remaining derivatives which are evaluated in an entirely similar manner are:

$$\left(\frac{\partial E_W}{\partial \theta_n}\right)_0 = 0, \qquad (42b)$$

$$\begin{aligned}\left(\frac{\partial^2 E_W}{\partial r_m\, \partial r_n}\right)_0 = {} & \frac{\pi}{r_0} h \bar{\sigma}_w n^2\, \delta_{mn} \\ & + \frac{3\pi}{4r_0} h\, \Delta\sigma_w mn \{ \delta_{n1}\, \delta_{m1} \cos 2\theta_1 - \delta_{n,m\pm2} \cos(n\theta_n - m\theta_m) \},\end{aligned}$$
$$(42c)$$

$$\left(\frac{\partial^2 E_W}{\partial\theta_m\,\partial\theta_n}\right)_0 = 0, \tag{42d}$$

and

$$\left(\frac{\partial^2 E_W}{\partial\theta_m\,\partial r_n}\right)_0 = 3\pi h\ \Delta\sigma_w\ \delta_{m2}\ \delta_{n2}\sin 2\theta_2\ . \tag{42e}$$

The terms of equation (42) in $\bar{\sigma}_w$ are all identical to the corresponding terms in the isotropic case [equation (14) of Ref. 1] so that adding the terms of equation (42) in $\Delta\sigma_w$ to equation (7) yields the total energy expansion for the anisotropic case. The domain is therefore in equilibrium when the force equation for the isotropic case (14) with $\sigma_w = \bar{\sigma}_w$ is solved and the force tending to make the domain elliptical is somehow balanced. Inclusion of the effect of second order energy variations terms solves the latter part of the equilibrium problem to lowest order.

Before proceeding with this solution, it is appropriate to comment upon the significance of the various energy terms. As required by translational symmetry [Ref. 1, Sec. 4.2.2] $(\partial E_{\Delta W}/\partial r_2)_0 = -2r_0 \cdot (\partial^2 E_{\Delta W}/\partial r_1^2)_0$ when $\theta_1 = \theta_2$ where $E_{\Delta W}$ is the $\Delta\sigma_W$ contribution to the energy. The $(\partial^2 E_W/\partial r_1^2)_0$ term in the stiffness matrix thus has only kinematic significance. On the other hand, $-\partial[\Delta r_2(\partial E_W/\partial r_2)_0]/\partial\theta_2 = -\Delta r_2(\partial E_W/\partial\theta_2\ \partial r_2)_0$ is a torque tending to turn an elliptical domain into the direction in which the force tending to make the domain elliptical is most positive.

To solve the elliptical equilibrium problem, it is convenient to write the energy variation expression in the form

$$\Delta E_T = -\mathbf{F}\cdot\mathbf{X} + \tfrac{1}{2}\mathbf{X}\cdot\mathbf{S}\cdot\mathbf{X} + 0_3(\mathbf{X}) \tag{43a}$$

where

$$\mathbf{X} \equiv [x_1\ ,\ x_2\ ,\ x_3\ ,\ x_4\ ,\ x_5\ ,\ x_6\ ,\ x_7\ ,\ \cdots] \tag{43b}$$

$$\equiv [\Delta r_0\ ,\ \Delta r_1\ ,\ \Delta\theta_1\ ,\ \Delta r_2\ ,\ \Delta\theta_2\ ,\ \Delta r_3\ ,\ \Delta\theta_3\ ,\ \cdots] \tag{43c}$$

and where the elements of the force vector, F, and the stiffness matrix, S, are

$$F_i = -\left(\frac{\partial E_T}{\partial x_i}\right)_0 \tag{43d}$$

and

$$S_{ij} = \left(\frac{\partial E_T}{\partial x_i\,\partial x_j}\right)_0\ . \tag{43e}$$

Setting the gradient of the total energy in the space of the x_i equal to zero and solving for the equilibrium yields displacements when third order energy terms are neglected

$$\mathbf{X} = \mathbf{S}^{-1}\mathbf{F} \tag{44}$$

where $\mathbf{S}^{-1}$ is the inverse stiffness or compliance matrix. From the comments in the preceding two paragraphs and inspection of equations (43) and (44), it can be seen that stable equilibrium is obtained when: r_0 is the stable solution to the isotropic force equation (14) with $\sigma_w = \bar{\sigma}_w$, $\theta_n = 0$ and $\Delta\theta_n = 0$ for all n (allowing now negative values of Δr_n) and $\Delta r_n = 0$ for $n = 0$ and n odd. The remaining even n Δr_n are determined from equation (44) written with respect to these variables only. Because of the special form of the resulting F and S, inverting only a finite submatrix of S yields X to any finite order in $\Delta\sigma_w$. When this is done there results

$$\frac{\Delta r_2}{r_0} = \frac{1}{2}\left[\frac{l}{h} - S_2\left(\frac{d}{h}\right)\right]^{-1}\left(\frac{\Delta l}{h}\right) + O_3\left(\frac{\Delta l}{h}\right), \tag{45a}$$

$$\frac{\Delta r_4}{r_0} = \frac{1}{5}\left[\frac{l}{h} - S_2\left(\frac{d}{h}\right)\right]^{-1}\left[\frac{l}{h} - S_4\left(\frac{d}{h}\right)\right]^{-1}\left(\frac{\Delta l}{h}\right)^2 + O_4\left(\frac{\Delta l}{h}\right), \tag{45b}$$

$$\frac{\Delta r_n}{r_0} = O_3\left(\frac{\Delta l}{h}\right), \qquad n \geqq 6, \tag{45c}$$

where

$$\Delta l \equiv \Delta\sigma_w/4\pi M_s^2 . \tag{46}$$

Equation (45a) has also been obtained by E. Della Torre and M. Dimyan.[9]

In the small Δr_n approximation of this calculation the wall anisotropy term has no effect on r_0 or any of the odd n Δr_n. The collapse diameter, d_0, and collapse field, H_0, are thus unchanged with respect to the isotropic case. At all diameters, however, the anisotropy pulls the initially circular domain out into an elliptical shape, the ellipticity being smallest at the collapse diameter and blowing up to infinity at the isotropic elliptical runout diameter (45a). At the isotropic runout diameter this calculation clearly cannot be applied to obtaining the reduction in the range of stability which wall anisotropy produces. It is also clear from the blow up of Δr_2 that any statement of such a reduction in the region of stability should be in terms of applied fields rather than domain diameters. In calculations of $\bar{\sigma}_w$ and $\Delta\sigma_w$ from measurements of the major and minor axes of elliptical domains, it should be noted that $\Delta r_4/r_0$ [equation (45b)] makes a small contribution to the length of each axis which is not removed by averaging.

In platelets having the preferred thickness, $h = 4l$, and biased to the collapse diameter, $\Delta r_2/r_0 = 2.61(\Delta l/h)$ and $\Delta r_4/r_0 = 4.65(\Delta l/h)^2$, while in the neighborhood of the isotropic elliptical runout diameter where $\Delta r_2 \rightarrow \infty$, $\Delta r_4/r_0 = 3.59(\Delta l/h)\ \Delta r_2/r_0$.

In rare earth orthoferrites, the anisotropy in wall energy is typically (within a factor of two) $\Delta\sigma_w/\bar{\sigma}_w = 3\%$ so that $\Delta l/h$ for $h = 4l$ is only $3/4\%$. The ellipticity at collapse is thus small. The contribution of Δr_4 is very small at all diameters ($h = 4l$) and further, near collapse, the cylindrical wall and infinitesimal wall width approximations quite possibly produce larger errors.

The implication of the foregoing for device applications is that wall energy anisotropy has no effect upon the collapse diameter and collapse field. As far as the elliptical runout conditions are concerned, careful measurements carried out on a low coercivity platelet of $TmFeO_3$ in which $\Delta\sigma_w/\bar{\sigma}_w = 3\%$ show that the reduction in the ratio of the collapse to the elliptical runout field is only of the order of 1%.[10]

III. COERCIVITY AND MOBILITY

The preceding sections treated only forces arising from reversible processes. The present section considers forces arising from the irreversible processes describable by the wall motion coercivity and wall mobility. Several relations between wall parameters and domain parameters are obtained. In particular, this section computes the domain coercivity and mobility in terms of the wall coercivity and obtains the maximum allowable coercivity which permits device operation in the hard bubble mode. Inversely, it is shown how the wall mobility and coercivity may be obtained by domain measurements.

3.1 *Domain Dissipation*

The method used to take dissipative effects into account is to compute the power dissipation produced by a general variation in domain shape using the wall dissipation equation and then to set this equal to the power produced by the variation. By this procedure the equations for the various modes (translation, size change and deformation) are obtained. The dissipation equation approach is taken as a best first guess to the solution of coercivity problems. The reader is cautioned not to take any of the coercivity results too literally, since coercivity is never uniform and in many cases depends on the direction of wall motion.[11]

Wall motion in many materials is describable in terms of a wall motion coercivity and a wall motion mobility.[12–14] Such a description should be valid in any material for sufficiently low velocities and coercivities. Presently known uniaxial materials such as $BaFe_{12}O_{19}$ for which the velocity-drive field relation is not describable by coercivity and mobility at high drive fields[15–18] (they show roughly a limiting velocity) have low mobilities which reduce their usefulness in device applications. The present work therefore restricts attention to 180° domain walls in materials having the velocity-drive field relation

$$|v_n| = \begin{cases} \mu_w(|H_L| - H_c), & |H_L| > H_c\,, \\ 0, & |H_L| \leqq H_c\,, \end{cases} \tag{47}$$

where: v_n is the local wall velocity in a direction normal to the wall, H_L is the total local field component parallel to the magnetization, and where equation (47) serves as the defining relation for the wall motion coercive field, H_c, and the wall mobility μ_w. The field H_L includes the effect of all magnetic fields as well as any effective fields such as the "wall energy field," and may vary from point to point. The symbol, H_L, is used to distinguish it from H, used elsewhere to denote the spatially uniform z component of the applied field. For a planar wall in the absence of internal magnetostatic forces, H_L is the applied field. Equation (47) is assumed to hold whether or not the wall is accelerating since for presently known materials wall inertia effects are negligible.

Consider now a segment of 180° domain wall such as a portion of the domain wall shown in Fig. 1 and assume that H_L is positive when it lies in the positive z direction. Under these conditions, the power input per unit area from the local field to an inward moving (decreasing r_b) segment of the domain wall is independent of the details of the magnetic configuration within the wall and is

$$\rho_{\text{in}} = 2M_s\,|H_L v_n|. \tag{48}$$

Since inertial effects have been assumed to be negligible, the power input must be equal to the power dissipated per unit wall area, ρ_{diss}, so that eliminating H_L from equations (47) and (48) results in

$$\rho_{\text{diss}} = 2M_s\left[H_c\,|v_n| + \frac{1}{\mu_w}v_n^2\right]\cdot \tag{49}$$

Integrating the dissipation density over the domain wall area yields the total dissipation

$$P_{\text{diss}} = \int_{\text{wall}} 2M_s\left[H_c\ |v_n| + \frac{1}{\mu_w} v_n^2\right] d\alpha \tag{50}$$

where $d\alpha$ is the differential wall area. For a circular domain, where $r_b(\theta) = r_0$, this expression becomes

$$P_{\text{diss}} = 2M_s h r_0 \int_0^{2\pi} \left[H_c\ |\mathbf{v}\cdot\mathbf{i}_r| + \frac{1}{\mu_w} |\mathbf{v}\cdot\mathbf{i}_r|^2\right] d\theta \tag{51}$$

where $\mathbf{i}_r$ is the unit vector in the radial direction. Now for a circular domain

$$\begin{aligned} \mathbf{v}\cdot\mathbf{i}_r &= \frac{dr_b(\theta)}{dt} \\ &= \sum_{n=0}^{\infty}\left[\left(\frac{\partial r_b(\theta)}{\partial r_n}\right)_0 \frac{dr_n}{dt} + \left(\frac{\partial r_b}{\partial \theta_n}\right)\frac{d\theta_n}{dt}\right] \\ &= \sum_{n=0}^{\infty} \cos\left[n(\theta - \theta_n)\right]\frac{dr_n}{dt} \end{aligned} \tag{52}$$

where again $d\Delta r_n$ and $d\Delta\theta_n$ have been abbreviated to dr_n and $d\theta_n$. Substituting equation (52) into equation (51) and carrying out the integration yields

$$\begin{aligned} P_{\text{diss}} = 4\pi M_s h r_0 \Bigg\{ & H_c\left[\left|\frac{dr_0}{dt}\right| + \frac{2}{\pi}\sum_{n=1}^{\infty}\left|\frac{dr_n}{dt}\right| + NO_1\left(\frac{dr_n}{dt}\right)\right] \\ & + \frac{1}{\mu_w}\left[\left(\frac{dr_0}{dt}\right)^2 + \frac{1}{2}\sum_{n=1}^{\infty}\left(\frac{dr_n}{dt}\right)^2 + O_3\left(\frac{dr_n}{dt}\right)\right]\Bigg\} \end{aligned} \tag{53}$$

where $NO_1\ (dr_n/dt)$ indicates the nonlinear coercivity coupling terms which appear even in lowest order. The nonlinear coupling terms tend in general to couple in additional modes even when only one dr_n/dt is initially nonzero. An exception to this is the uniform radial mode of motion dr_0/dt which, because of its symmetry, may take place without coupling in the other modes. If the coercivity is negligible $[H_c \ll 1/\mu_w(dr_n/dt)]$ then the Δr_n and $\Delta\theta_n$ remain uncoupled normal modes of the system. The consequences of the damping of the various individual modes will now be considered in the order $n = 1$, $n = 0$, $n \geqq 2$.

3.2 *Domain Coercivity and Mobility* ($n = 1$)

Consider an initially circular domain in which only $d\Delta r_1/dt$ is nonzero. In this case the component of the domain wall velocity normal to the wall is

$$v_n = dr_b(\theta)/dt = (d\Delta r_1/dt)\cos(\theta - \theta_1). \tag{54}$$

Since the distribution of the wall velocity component normal to the domain wall for a circular domain propagating with velocity, v_d, in the θ_d direction is

$$v_n = \mathbf{i}_r \cdot \mathbf{v}_d = |v_d| \cos(\theta - \theta_d), \tag{55}$$

a variation in Δr_1 may be identified with translation in the θ_d direction with

$$|\mathbf{v}_d| = |d\Delta r_1/dt|. \tag{56}$$

Using equation (53), the power dissipated by a uniformly moving circular domain is

$$P_{\text{diss}} = \pi M_s h r_0 \left[\frac{8}{\pi} H_c\, |\mathbf{v}_d| + \frac{2}{\mu_w} \mathbf{v}_d^2 \right] \tag{57}$$

and, since $P_{\text{diss}} = -\mathbf{v}_d \cdot \mathbf{F}_d$ where $\mathbf{F}_d$ is the drag force, the equivalent force for a domain moving with a nonzero velocity is

$$\mathbf{F}_d = -\pi M_s h r_0 \left[\frac{8}{\pi} H_c + \frac{2}{\mu_w} |\mathbf{v}_d| \right] \mathbf{i}_v \tag{58}$$

where $\mathbf{i}_v = \mathbf{v}_d/|\mathbf{v}_d|$ is the unit vector in the direction of the motion. Notice that $\mathbf{F}_d$ is an ordinary linear force which could be measured mechanically with the aid of a magnetic probe.

Translational forces are most easily produced by gradients in the applied field. The power input to the domain from such a force is obtained by integrating the power input density, equation (48), over the domain wall area. In the case of a uniform gradient the local field at the domain wall is

$$H_L = \bar{H}_L - \tfrac{1}{2}\, |\Delta\mathbf{H}| \cos(\theta - \theta_g) \tag{59}$$

where $\Delta\mathbf{H}$ is a vector orientated in the direction in which the bias field decreases most rapidly, θ_g, and has a magnitude equal to the maximum difference in field across the domain (a gradient of magnitude $|\Delta H|/2r_0$ and direction $\theta_g + \pi$). In equation (59) $\bar{H}_L$ includes: the bias field at the center of the domain, the "wall field" and the demagnetizing field, all of which are independent of angle. All of the field components in equation (59), H_L, $\bar{H}_L$ and $\Delta\mathbf{H}$ are to be understood as the z-averaged z components of the actual field since this is the quantity which interacts with a 180° cylindrical domain wall. With this understanding, integration over the wall area may be replaced by multiplication by h and integration over θ. Assuming the domain to be exactly circular, the total power input to the domain is

$$\begin{aligned} P_{\text{in}} &= 2M_s r_0 h \\ &\quad \cdot \int_0^{2\pi} [\bar{H}_L - \tfrac{1}{2}\, |\Delta\mathbf{H}| \cos(\theta - \theta_g)] \cdot [-|v_d| \cos(\theta - \theta_d)]\, d\theta \\ &= \pi r_0 h M_s\, \Delta\mathbf{H} \cdot \mathbf{v}_d \end{aligned} \tag{60}$$

which is just the expression for the power input to a dipole of strength $(\pi r_0^{\,2} h)(2M_s)$ propagating in a gradient of magnitude $|\Delta\mathbf{H}|/2r_0$.

The domain will propagate in the direction in which the input power is greatest, ($\theta_g = \theta_d$ the direction in which the bias field decreases most rapidly), and the magnitude of the velocity will be such that the dissipated power (57) balances the input power (60). Such a balance always occurs for $v_d = 0$. When $|\Delta\mathbf{H}| \leqq (8/\pi)H_c$ this is the only condition in which the balance is maintained. The domain velocity is therefore

$$|\mathbf{v}_d| = \frac{1}{2}\, \mu_w\left(|\Delta H| - \frac{8}{\pi}\, \mathbf{H}_c\right), \qquad |\Delta\mathbf{H}| > \frac{8}{\pi}\, H_c\ , \tag{61a}$$

$$|\mathbf{v}_d| = \qquad 0, \qquad\qquad |\Delta\mathbf{H}| \leqq \frac{8}{\pi}\, H_c\ . \tag{61b}$$

Comparison of equations (47) and (61) shows that it is possible to define a domain mobility and coercivity by taking $\Delta\mathbf{H}$ as the driving field in terms of the wall mobility and coercivity as

$$\mu_d = \tfrac{1}{2}\mu_w \tag{62}$$

and

$$H_{cd} = \frac{8}{\pi}\, H_c \tag{63}$$

if $\Delta\mathbf{H}$ is taken to be the drive field.

Note that in the integral for the power input (60), the angle independent field term, $\bar{H}_L$, does not contribute so that the effective power input density with respect to angle for a domain propagating down a bias field gradient is $r_0 h M_s\, |\Delta\mathbf{H}|\, |\mathbf{v}_d| \cos^2(\theta - \theta_d)$ which has the same angular factor as the mobility associated dissipation (50) and (55). The power input and power dissipated thus cancel (after z averaging) at each point on the perimeter of the domain for domains propagating under the influence of a uniform field gradient and viscous damping. In this case therefore there are no forces tending to distort the domain shape from circular. When coercivity is present (even perfectly uniform coercivity), the power input and dissipation do not cancel locally and thus internal stresses tending to distort the domain from circular are present.

When a bias field is applied to a circular domain which is not simply a uniform gradient, the affect of the additional nonuniformities may be taken into account by Fourier decomposition of the z-averaged z component of the applied field at the domain wall with respect to angle. When this is done the constant term determines domain size, the θ term translates the domain, and the $n\theta$ terms, for $n \geqq 2$, deform the domain. The procedure is, of course, only applicable if the domain shape remains near circular.

Coercivity and mobility may be measured either by applying fields to walls[19] and using equation (47) or by applying field gradients to entire cylindrical domains[20] and using equation (61). In the platelets of material which have a coercivity which is so low as to make them useful for device applications it has been found convenient to use a second domain to provide the field gradient for the coercivity measurements. In this measurement two domains are brought together and then released, the coercivity being given by the formula[5]

$$H_c = (4\pi M_s) \frac{3\pi}{8} \frac{r_0^3 h}{s^4} \tag{64}$$

where s is the center-to-center distance of the two circular domains. Equation (64) was obtained using equation (61) and the z component of the dipole field from the second domain at the center of the plate, $2M_s\pi r_0^2 h/s^3$.

3.3 *Domain Size in the Presence of Dissipative Processes* ($n = 0$)

The domain size variation mode Δr_0 has two properties which are not common to the other modes: First, large variations from equilibrium may be considered using the force function, $F(d/h)$, as well as small variations from equilibrium using the radial stability function, $S_0(d/h)$. Second, arbitrary dissipation functions may be considered because the rotational symmetry of the motion removes any tendency for an initially circular domain to couple in other modes.

3.3.1 *Large Variations in Domain Size*

Setting all the $d\Delta r_n/dt$ except $d\Delta r_0/dt$ equal to zero in equation (53) yields the dissipation produced by a domain size change

$$P_{\text{diss}} = 4\pi M_s h r_0 \left[\pm H_c + \frac{1}{\mu_w} \frac{d\Delta r_0}{dt} \right] \frac{d\Delta r_0}{dt} \tag{65}$$

where the upper sign is for an expanding domain. Setting the sum of the power dissipation and the rate of energy change from equation (7)

$$\frac{dE_T}{dt} = \frac{dE_T}{d\Delta r_0}\frac{d\Delta r_0}{dt} = 4(2\pi M_s^2)(\pi h^2)\left[\frac{l}{h} + \frac{d}{h}\frac{H}{4\pi M_s} - F\left(\frac{d}{h}\right)\right]\frac{d\Delta r_0}{dt} \tag{66}$$

equal to zero yields the differential equation for the domain diameter $[d\Delta r_0 = dr_0 = \frac{1}{2}d(d)]$

$$-\frac{h}{2\mu_w(4\pi M_s)}\frac{d}{h}\frac{d\left(\frac{d}{h}\right)}{dt} = \frac{l}{h} + \frac{d}{h}\frac{H \pm H_c}{4\pi M_s} - F\left(\frac{d}{h}\right). \tag{67}$$

The domain diameter thus relaxes toward a value which is a solution to the force equation (14) in which the bias field H has been replaced by a composite bias field, $H \pm H_c$. After this substitution is carried out, the equilibrium diameters are obtained as described in Section 2.1 except that the sign of $\pm H_c$ must now be determined in each case and now there is a small continuous range of stable solutions about both the stable and unstable zero coercivity solutions. There are again two solution diameters for each $H \pm H_c$ when $0 < H \pm H_c < H_0$. In the present case, however, coercivity produces two stable ranges of solutions rather than one stable solution point and one unstable solution point. The large diameter solutions to the force equation for $H + H_c$ and $H - H_c$ bound the solution range which brackets the zero coercivity stable solution and similarly the small diameter solutions to the force equation for $H + H_c$ and $H - H_c$ bound the solution range which brackets the zero coercivity unstable solution.

Figure 10 shows the graphical construction for the case $l/h = 0.300$, $H/4\pi M_s = 0.2544$, $H_c/4\pi M_s = 0.020$. The zero coercivity stable and unstable solutions at $d/h = 2.000$ and $d/h = 0.767$ are marked $\mathcal{S}$ and $\mathcal{U}$ respectively; the large diameter solutions for $H + H_c$ and $H - H_c$ at $d/h = 1.527$ and $d/h = 2.468$ are marked $\mathcal{S}_+$ and $\mathcal{S}_-$ respectively and the small diameter solutions for $H + H_c$ and $H - H_c$ at $d/h = 0.919$ and $d/h = 0.685$ are marked $\mathcal{U}_+$ and $\mathcal{U}_-$ respectively. A domain having a diameter greater than that at $\mathcal{S}_-$ will relax toward the diameter at $\mathcal{S}_-$ as indicated by the arrow. Coercivity stabilizes domains having diameters between $\mathcal{S}_+$ and $\mathcal{S}_-$. Domains having diameters between the diameter at $\mathcal{S}_+$ and $\mathcal{U}_+$ will relax toward $\mathcal{S}_+$ as indicated by the arrow. Coercivity stabilizes domains having diameters between $\mathcal{U}_+$ and $\mathcal{U}_-$. Small diameter coercivity stabilized solutions have been observed in the process of carrying out the mobility measurement described in the second paragraph which follows.[7] Domains having diameters smaller

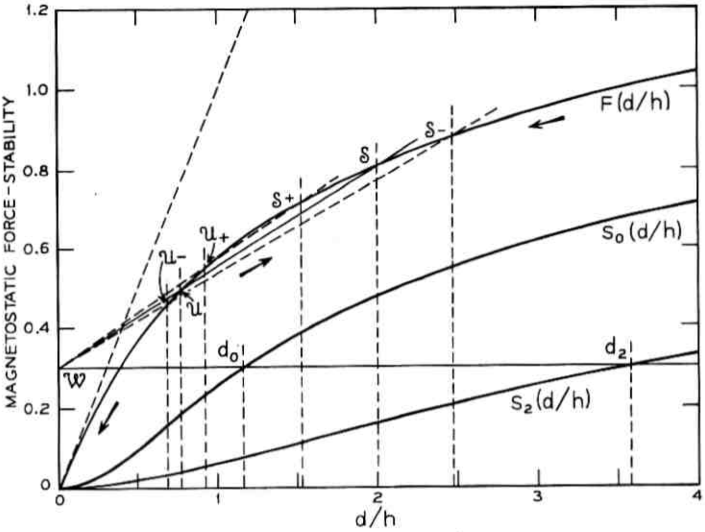


Fig. 10—Construction of solutions to the force equation in the presence of coercivity, $l/h = 0.3$, $H/4\pi M_s = 0.2544$, $H_c/4\pi M_s = 0.020$.

than $\mathcal{U}_-$ will collapse as indicated by the arrow. The rate of change of domain diameter in any of the dynamic processes described is, of course, given by equation (67).

It can be seen that the collapse diameter is independent of coercivity to the extent that the coercivity is uniform (this is only approximately true at best). A measurement of the collapse diameter, denoted by d_m here, yields the characteristic length

$$l = hS_0(d_m/h) \tag{68}$$

independent of the coercivity. If H_c is then measured from the diameter hysteresis and this value is subtracted from the measured collapse field, the value of $4\pi M_s$ is obtained.

Bobeck[17,18] has developed a method for measuring mobilities which requires only the observation of static domain states. In this method the domain is biased to a stable diameter and then a short duration field pulse having a magnitude, H_P , such that the total field amplitude exceeds the collapse field, $H + H_P > H_0$, is applied. The pulse duration is at first kept so short that at the end of the pulse, the domain has a diameter between the zero coercivity stable and unstable solution so that it recovers its initial size. The pulse length is then gradually increased until the diameter at the end of the pulse is less than the unstable solution diameter so that the domain collapses.

Figure 11 is a construction illustrating the mobility method for $l/h = 0.30$, $H/4\pi M_s = 0.254$, $H_P/4\pi M_s = 0.246$ and zero coercivity. Under these conditions, $(H + H_P)/4\pi M_s = 0.500 > H_0/4\pi M_s = 0.283$, and the solution diameters are $d_s/h = 2.00 > d_0/h = 1.16 > d_u/h = 0.77$ where d_s is the stable solution diameter and d_u is the unstable solution diameter. In the figure the static collapse point is denoted by $\mathcal{C}$. The domain is initially at the stable equilibrium point $\mathcal{S}$. Application of a fast rise bias field pulse of amplitude H_P takes the domain to $\mathcal{S}'$ where it collapses towards $\mathcal{W}$. If the pulse is turned off at $\mathcal{R}'$ (above the unstable solution point); the diameter increases until the point $\mathcal{S}$ is again reached. If, on the other hand, the pulse has sufficient duration to reach $\mathcal{D}'$ so that the domain state reaches $\mathcal{D}$, the domain collapses to $\mathcal{W}$. In the insert, the field pulse shape used in making the measurement is shown as a function of time.

When the domain velocity is proportional to the applied field and

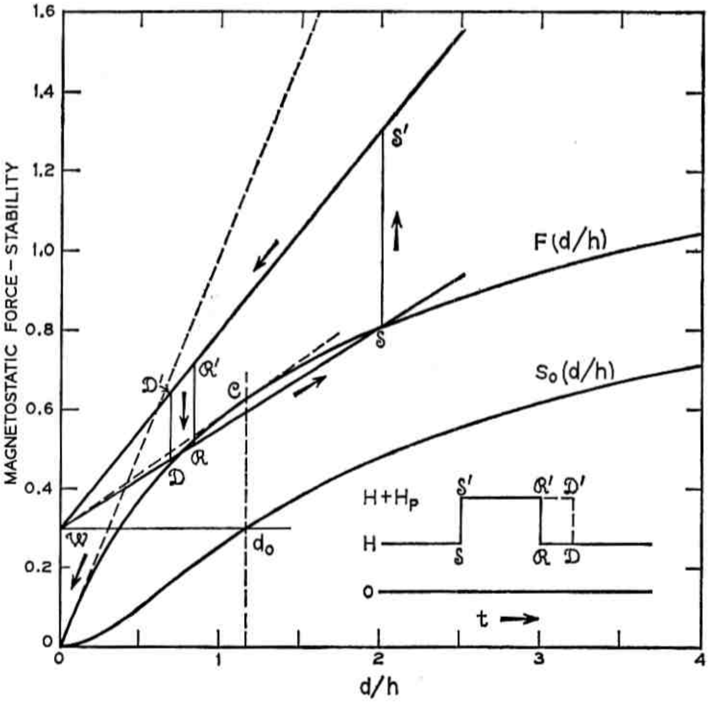


Fig. 11—Construction used in the pulsed bias field mobility method, $l/h = 0.3$, $H/4\pi M_s = 0.254$, $(H + H_P)/4\pi M_s = 0.500$.

pulse rise time effects may be neglected, the mobility is

$$\mu_w = \frac{h}{2T4\pi M_s} \int_{d_u/h}^{d_s/h} \frac{\dfrac{d}{h}\, d\left(\dfrac{d}{h}\right)}{\dfrac{l}{h} + \dfrac{d}{h}\dfrac{H}{4\pi M_s} - F\left(\dfrac{d}{h}\right)} \tag{69}$$

where T is the minimum pulse duration which results in the collapse of the domain. When coercivity must be taken into account the limits of the integral change as described previously. (At the beginning of the pulse the domain is at the point $\mathcal{S}_+$ of Fig. 10.) If additionally nonlinear velocity-drive field relations and pulse rise times must be considered, it is not possible to obtain the mobility as a simple integral over the diameter and the integration is better carried out with respect to time.

3.3.2 *Small Variations in Domain Size from Equilibrium*

This section considers the effect of dissipative processes on domain size for the case where the energy variation expression (7) may be considered an expansion of the energy about the domain diameter which is a solution to the force equation (14). In this case the rate of change of energy with respect to time using (7) is

$$\begin{aligned} \frac{dE_T}{dt} &= \frac{\partial E_T}{\partial \Delta r_0}\frac{d\Delta r_0}{dt} \\ &= -8\pi h(2\pi M_s^2)\,\frac{h}{d}\left[\frac{l}{h} - \mathcal{S}_0\left(\frac{d}{h}\right)\right]\Delta r_0\,\frac{d\Delta r_0}{dt}. \end{aligned} \tag{70}$$

Again equating the rate of energy decrease to the power dissipation (65) yields for the linearized differential equation for radial motion

$$\frac{1}{\mu_w(4\pi M_s)}\frac{d\Delta r_0}{dt} = \left(\frac{h}{d}\right)\left[\frac{l}{h} - \mathcal{S}_0\left(\frac{d}{h}\right)\right]\frac{\Delta r_0}{r_0} \mp \frac{H_c}{4\pi M_s}. \tag{71}$$

The domain radius thus relaxes towards

$$\frac{|\Delta r_0|}{r_0} = \frac{H_c}{4\pi M_s}\left|\frac{d/h}{l/h - \mathcal{S}_0(d/h)}\right| \tag{72}$$

if the domain is stable or, in the case of an unstable domain, coercivity stabilizes the domain for departures in radius from equilibrium up to this same value. In either case, stable or unstable, the relaxation time [defined by the time factor $\exp(-t/\tau)$] is

$$\tau = -\frac{r_0}{\mu_w 4\pi M_s}\,\frac{d/h}{l/h - \mathcal{S}_0(d/h)}. \tag{73}$$

The factor $d/h[l/h - S_0(d/h)]^{-1}$ in equations (71), (72) and (73) is a measure of the radial compliance of the domain relative to coercivity or mobility and will be called the radial relative compliance function. The value of this function may be obtained as the inverse slope of a line constructed on a plot of the force and stability functions. Such a construction is shown in Fig. 12 for $l/h = 0.30$, $d/h = 2.00$, $H/4\pi M_s = 0.2544$. The numerical slope of the line $\mathcal{W}\mathcal{R}$ drawn from l/h on the vertical axis to the point on S_0 at $d/h = 2.00$ is 0.0883 so that $d/h[l/h - S_0]^{-1} = 11.33$. Thus, in this case, for a domain to have a diameter defined to within ten percent, the coercivity must be $H_c < 0.01(4\pi M_s)$.

At the collapse diameter, d_0, the relative radial compliance is, of course, infinite while at the other end of the range of stability (the elliptical runout diameter) where in the present example $d_2/h = 0.358$, the relative radial compliance has the value 9.57. The minimum value of the relative radial compliance for $l/h = 0.30$ is 9.52 occurring at a diameter of $d/h = 3.22$. This behavior of the radial relative compliance at $l/h = 0.30$ is typical for thicknesses near the preferred value of $h/l = 4$ as can be seen from Fig. 12. In all such cases, the minimum compliance is achieved at some diameter less than d_2, the compliance being nearly constant from $d = d_2$ down to $(d_0d_2)^{\frac{1}{2}}$ [for

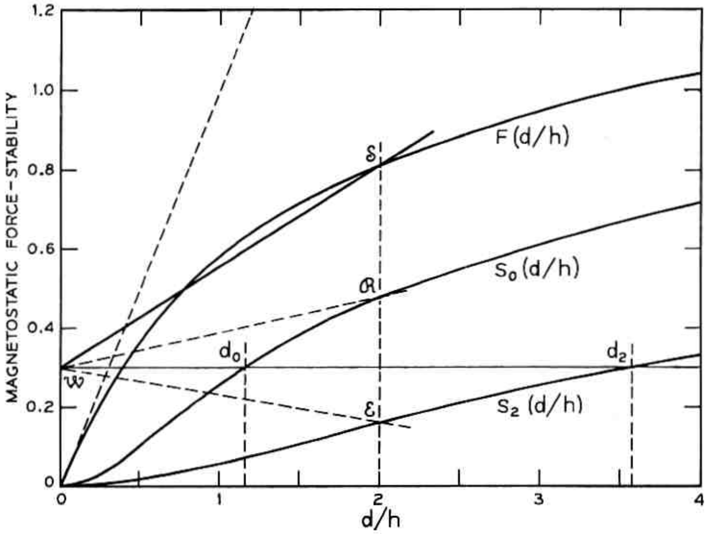


Fig. 12—Construction of the radial and elliptical relative compliance functions for $l/h = 0.3$, $d/h = 2$.

$l/h = 0.30$, $(d_0 d_2)^{\frac{1}{2}}/h = 2.04]$ then increasing rapidly to infinity at d_0. The constancy of the radial compliance over a large part of the stable range is related to the linearity of the diameter-field curves away from the collapse diameter [see (36) and Fig. 8].

In Section 3.3.1, the diameters bounding the coercivity stabilized solutions to the equilibrium problem were computed for the case $l/h = 0.30$, $H/4\pi M_s = 0.2544$ and $(d_u/h = 0.767$ $d_s/h = 2.00$ and $H_c/4\pi M_s = 0.020$ using the force equation in the presence of coercivity. As an illustrative consistency check, these bounding diameters will now be computed using the linearized equation (72). The diameters bounding the solution region bracketing the zero coercivity unstable solution obtained from this equation are

$$\frac{d}{h} = \frac{d_u}{h}\left\{1 \pm \frac{H_c}{4\pi M_s}\frac{d_u}{h}\left[\frac{l}{h} - S_0\left(\frac{d_u}{h}\right)\right]^{-1}\right\} = 0.664 \quad \text{and} \quad 0.869,$$

and the diameters bounding the solution region bracketing the zero coercivity stable solution region are

$$\frac{d}{h} = \frac{d_s}{h}\left\{1 \pm \frac{H_c}{4\pi M_s}\frac{d_s}{h}\left[\frac{l}{h} - S_0\left(\frac{d_s}{h}\right)\right]^{-1}\right\} = 1.547 \quad \text{and} \quad 2.453.$$

The corresponding diameters obtained directly from the force equation were $d/h = 0.685$, 0.919, 1.527, and 2.468. The excellent agreement of the diameters bounding the zero coercivity stable solution computed by the two methods is again related to the linearity of the diameter-field curves. Note that the agreement of splittings of the diameters bounding the zero coercivity unstable solution computed the two ways is also quite good.

3.4 *Domain Shape in the Presence of Dissipative Processes* $(n \geqq 2)$

The power dissipation produced by the motion of a single mode is from equation (53)

$$P_{\text{diss}} = 4\pi M_s h r_0\left[\pm\frac{2}{\pi}H_c + \frac{1}{2}\frac{1}{\mu_w}\frac{d\Delta r_n}{dt}\right]\frac{d\Delta r_n}{dt}, \qquad n \geqq 2, \tag{74}$$

where the upper sign is for positive $d\Delta r_n/dt$. The rate of change of domain energy under these conditions is from (7)

$$\begin{aligned}\frac{dE_T}{dt} &= \frac{\partial E_T}{\partial \Delta r_n}\frac{\partial \Delta r_n}{dt} \\ &= 4\pi h(2\pi M_s^2)(n^2 - 1)\left(\frac{h}{d}\right)\left[\frac{l}{h} - S_n\left(\frac{d}{h}\right)\right]\Delta r_n\frac{d\Delta r_n}{dt}, \qquad n \geqq 2. \quad (75)\end{aligned}$$

Setting the sum of the rate of change of the energy and the power dissipation equal to zero yields the differential equation for the relaxation of the variation of a single Δr_n,

$$\frac{1}{\mu_w 4\pi M_s}\frac{d\Delta r_n}{dt} = -(n^2 - 1)\left(\frac{h}{d}\right)\left[\frac{l}{h} - S_n\left(\frac{d}{h}\right)\right]\frac{\Delta r_n}{r_0} \pm \frac{4}{\pi}\frac{H_c}{4\pi M_s}. \tag{76}$$

The domain shape variation amplitude thus relaxes towards

$$\frac{|\Delta r_n|}{r_0} = \frac{H_c}{4\pi M_s}\left|\frac{4}{\pi}\frac{1}{(n^2 - 1)}\frac{d/h}{[l/h - S_n(d/h)]}\right|, \qquad n \geqq 2, \tag{77}$$

if the variation is stable or, in the case of an unstable domain, coercivity stabilizes the domain for variations in amplitude up to this same value. In either case, stable or unstable, the relaxation time (defined by the time factor $\exp(-t/\tau)$) is

$$\tau = \frac{r_0}{\mu_w 4\pi M_s}\frac{1}{(n^2 - 1)}\frac{d/h}{[l/h - S_n(d/h)]}, \qquad n \geqq 2. \tag{78}$$

Whenever the coercivity is effectively zero so that the mobility characterizes all dissipative processes, the normal modes of the domain remain decoupled (within the small amplitude approximation). Each of these modes relaxes according to equations (71) or (76). When coercivity must be considered, these equations become rather crude approximations because even perfectly uniform coercivity introduces nonlinear mode coupling. The nonlinear mode coupling is especially noticeable at the end of the relaxation of a single mode. The reason for phrasing the discussion of the effects of dissipative processes in terms of dissipation equations was to account correctly for the effects of coercivity to lowest order without being required to examine the coupling of the modes or the origins of coercivity. The results obtained do provide a general picture of the dependence of the effect of coercivity on the various domain parameters and in particular they provide a measure of the dependence of the stiffness of the domains on these parameters.

Equations (77) and (78) show that the entire n dependence of the residual distortion of a domain recovering from a fluctuation of a single mode and the relaxation time with which it recovers is contained in the stiffness factor $(n^2 - 1)[l/h - S_n(d/h)]$. The domain stability condition, $S_0 \geqq l/h \geqq S_2$, and the inequality (18) imply that $[l/h - S_n]$ is a monotonic increasing function of n. Therefore the residual Δr_n and relaxation time both decrease slightly faster than $1/n^2$ (see Ref. 1).

Thus, to characterize the effect of coercivity in limiting the attainment of stable movable cylindrical domains it is necessary only to consider the elliptical shape variation mode ($n = 2$) in addition to the size ($n = 0$) and translation ($n = 1$) modes discussed previously.

The factor $(4/3\pi)(d/h)[l/h - S_2(d/h)]^{-1}$ in equations (76), and (77) and (78) for $n = 2$ is a measure of the elliptical compliance of the domain relative to coercivity of mobility and will be called the elliptical relative compliance function. The value of this function is proportional to the inverse slope of a line constructed on a plot of the force and stability functions. Figure 12 shows such a construction for $l/h = 0.30$, $d/h = 2.00$, $H/4\pi M_s = 0.2544$. The numerical slope of the line $\mathcal{W}\mathcal{E}$ drawn from l/h on the vertical axis to the point on S_2 at $d/h = 2.00$ is -0.068 so that $(4/3\pi)(d/h)[l/h - S_2]^{-1} = 6.25$. Thus a coercivity less than that value which defines the domain diameter to within ten percent ($H_c < 0.01(4\pi M_s)$) defines the ratio of the difference in the ellipse semiaxes to their average ($2\Delta r_2/r_0$) to 12 percent.

At the elliptical runout diameter, d_2, the relative elliptical compliance is, of course, infinite while at the other end of the range of stability (the collapse diameter) where in the present case $d_0/h = 1.16$ the relative elliptical compliance has the value 2.22. The figure shows that this behavior is true for any value of l/h. In general the minimum value of the relative elliptical compliance occurs at the collapse diameter, the compliance increasing from the minimum in a regular fashion to infinity at the elliptical runout diameter. This regular behavior is in contrast to the behavior of the relative radial compliance.

3.5 *Restrictions Placed on the Region of Device Operation by Consideration of Dissipative Effects.*

The relative compliance functions appearing as factors in equations (72), (73), (77) and (78) contain the dependence of both the normalized residual distortion $\Delta r_n/r_0$, and the normalized relaxation time $\tau(\mu_w 4\pi M_s)/r_0$, upon n, l, h, and d, or H. Although the discussion which follows is phrased in terms of the residual $\Delta r_n/r_0$, it should be kept in mind that the same remarks apply to the relaxation times scaled to $r_0/(\mu_w 4\pi M_s)$, the time required to propagate the domain one radius when the maximum field difference across the domain, ΔH, is $8\pi M_s$ [equation (61)].

In preceding subsections, the following has been demonstrated: Except in the immediate neighborhood of the collapse diameter, the domain diameter and bias field are, within the range of stability, approximately linearly related (Fig. 8). In the neighborhood of the plate thickness previously termed preferred, $h_p = 4l$ [equation (31)], the radial compliance, considered as a function of the bias field, is roughly constant from $d = (d_0d_2)^{\frac{1}{2}}$ to $d = d_2$, and at $d = (d_0d_2)^{\frac{1}{2}}$, the elliptical compliance is one half the radial compliance so that $d = (d_0d_2)^{\frac{1}{2}}$ represents a reasonable bias condition. Since for $h/l \approx 4$ or in general for any plate thickness, the diameter and bias field ranges are relatively narrow (Fig. 5) the radial and elliptical compliance of domains suitable for device application may be completely characterized with respect to plate thickness by plotting the relative compliance functions with respect to h/l for the bias condition $d = (d_0d_2)^{\frac{1}{2}}$.

The relative radial compliance,

$$\frac{\Delta r_0/r_0}{H_c/4\pi M_s} = \frac{d/h}{S_0(d/h) - l/h}, \tag{79a}$$

$$\approx 1.085(h/l)^{\frac{1}{2}}, \qquad h/l \gg 1, \tag{79b}$$

$$\approx 3.783 \exp{(\pi l/h)}, \qquad h/l \ll 1, \tag{79c}$$

and the relative elliptical compliance,

$$\frac{\Delta r_2/r_0}{H_c/4\pi M_s} = -\frac{4}{3\pi} \cdot \frac{d/h}{S_2(d/h) - l/h}, \tag{80a}$$

$$\approx 1.382(h/l)^{\frac{1}{2}}, \qquad h/l \gg 1, \tag{80b}$$

$$\approx 1.606 \exp{(\pi l/h)}, \qquad h/l \ll 1, \tag{80c}$$

are plotted in Fig. 13 as functions of h/l for the bias condition $d = (d_0d_2)^{\frac{1}{2}}$. The function values and asymptotic forms were obtained by methods which were used in Section 2.3 to obtain the diameter and field functions. The feature which distinguishes the relative compliance functions from the diameter and field functions is that the diameter and field functions bound the region of domain stability whereas the relative compliance functions provide a measure of the magnitude of the stability of the domains within the stable region.

The constructions of the radial and elliptical relative compliance functions shown in Fig. 12 and described in Sections 3.3.2 and 3.4 for $h/l = 3.33$, $d/h = 2.00$ may be taken as approximate construction for the values of these functions at $h/l = 3.33$ since $(d_0d_2)^{\frac{1}{2}}/h = 2.04$.

The minimum value of the radial compliance $[d = (d_0d_2)^{\frac{1}{2}}]$ is ≈ 7.9 occurring at a thickness of $h/l \approx 10.3$ and the minimum value of the elliptical compliance $[d = (d_0d_2)^{\frac{1}{2}}]$ is ≈ 5.9 occurring at a thickness

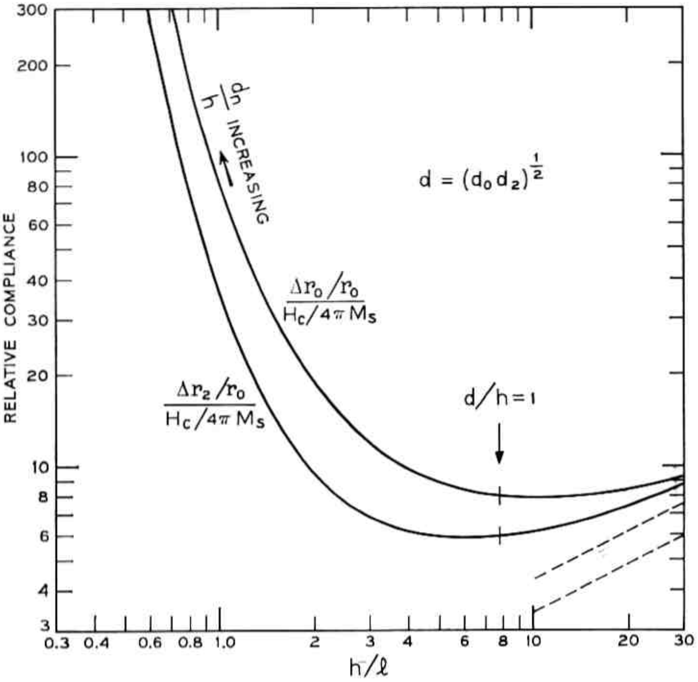


Fig. 13—Radial and elliptical relative compliance functions as a function of the thickness, h, measured in units of the characteristic length, l, for the bias condition $d = (d_0 d_2)^{\frac{1}{2}}$.

of $h/l \approx 6.1$.The compliance minima occur at somewhat greater plate thicknesses than the diameter minima. However, since the functions are quite flat bottomed, they increase only slightly in value from their minima to the thickness value previously termed preferred, $h/l = 4.0$ [equation (31)]. {At $h/l = 4.0$, $(d_0 d_2)^{\frac{1}{2}}/l \approx 6.8$. However, the preferred values were taken to be $h_p = 4l$ [equation (31)], $d_p = 8l$ [equation (32)] so that $d_p/h_p = 2.0$. The exact preferred values depend on the device structure in which the domain is located.}

As in the example of Sections 3.3.2 and 3.4, the coercivity requirement

$$H_c < 0.01(4\pi M_s) \tag{81}$$

will insure that a domain having the preferred thickness and diameter values will show radius function variations from the equilibrium radius of no more than ten percent. Just below $h/l = 4.0$, the increase in the compliance functions is seen to become exponential so that the observation of stable moveable cylindrical domains in reasonably thin plates is experimentally unlikely. Since the relative compliance functions increase slowly with thickness for thick plates, cylindrical domains may be observed well into the region in which the assumption of cylindrical walls becomes dubious (see Ref. 1, Section 6.1).

Although domains may be observed in thick plates, other considerations cause the realizable bit rate to decrease at least inversely with increasing thickness. These effects are as follows: The ΔH appearing in the domain velocity expression (61) is twice the $n = 1$ Fourier component of the z-averaged z component of the applied field. Since the field gradient is applied from the surface of the plate and obeys Laplace's equation, the variation in the applied field intensity decreases exponentially into the plate from its value at the surface. Thus for a given field variation intensity at the plate surface, the z-averaged z component of the intensity variation decreases inversely with increasing h/d for $h/d > 1$. Now for $d = (d_0 \, d_2)^{\frac{1}{2}}$, $d/h = 1$ occurs at $h/l \approx 8$ (see Figs. 2 or 13) and above this value $d/h \approx 2.17(h/l)^{-\frac{1}{2}}$ [see Fig. 4 and equations (23b) and (24b)] so that for thicknesses greater than $h/l \approx 8$, the mobility with respect to the field gradient at the surface of the plate decreases according to $(h/l)^{-\frac{1}{2}}$. Since in a given material the absolute domain diameter increases with increasing plate thickness in this region according to $d/l \approx 2.17(h/l)^{\frac{1}{2}}$, the bit-rate decreases according to $(h/l)^{-1}$. If additionally the maximum field difference at the surface of the plate is assumed to be some fraction of the difference of the collapse and runout fields, $H_0 - H_2$, then it is seen from equations (27b), (28b) and (30b) that the bit-rate decreases according to $(h/l)^{-\frac{3}{2}}$.

Thus in summary, consideration of the effects of dissipative processes even more strongly defines the neighborhood of $h_p/l = 4$, $d_p/l = 8$ as the preferred region than did considerations of stability only and additionally yields the requirement $H_c/4\pi M_s < 0.01$ for the attainment of stable movable domains.

IV. DETERMINATION OF MATERIAL PARAMETERS FROM PREFERRED DEVICE PARAMETERS*

The preceding sections have provided preferred values of the plate thickness (31) and domain diameter (32) or bias field, and the least

* Reference 21 includes part of the material of this section in a discussion of the relation of the M_s and K_u values of materials (available at that time) to the preferred values of these parameters.

permissible value of the anisotropy constant (1). Additionally it was shown that the wall mobility acts to form the scale factor for time. For device construction it is desirable to specify the domain diameter, d_p, from considerations of bit density and the resolution of mask-making and etching procedures while maximizing the bit rate. Once the domain diameter is specified, the desired characteristic material length is determined by equation (32) as $l \approx d_p/8$, the thickness is determined by equation (31) as $h_p \approx \frac{1}{2}d_p$, and the applied field is specified by equation (14) as $H \approx 0.28(4\pi M_s)$. It will now be shown that by adding assumptions about mobility and room temperature operation to conditions (1) and (31), it is possible to specify uniquely the three parameters A, K_u and M_s appearing in the energy density expression for the simplest uniaxial material[22]

$$\rho_E = A\left[\left(\frac{\partial\theta}{\partial s}\right)^2 + \sin^2\theta\left(\frac{\partial\phi}{\partial s}\right)^2\right] + K_u \sin^2\theta - \mathbf{H}_L\cdot\mathbf{M}. \tag{82}$$

In equation (82), A is the isotropic exchange constant, θ is the polar angle (the angle between the magnetization and the z axis), ϕ is the azimuthal angle, s is the distance through the wall, K_u is the uniaxial anisotropy constant, $\mathbf{M}$ is the magnetization vector ($|\mathbf{M}| = M_s$) and $\mathbf{H}_L$ is the sum of the applied and demagnetizing fields. Only Bloch walls in this simplest uniaxial material will be considered. Achievement of the coercivity condition (81) is a function of both intrinsic material properties and processing and will not be considered here.

In Section VI of Ref. 1, it was shown that from several standpoints

$$q \equiv K_u/2\pi M_s^2\,, \tag{83}$$

the dimensionless ratio of the uniaxial anisotropy constant to the energy density of a volume containing a magnetic field of strength $4\pi M_s$ must at least be greater than one, equation (1). From the definitions of q, l, and d_p and the expressions for the Bloch wall width and energy in the simplest uniaxial material (82), $l_w = \pi(A/K_u)^{\frac{1}{2}}$ and $\sigma_w = 4(AK_u)^{\frac{1}{2}}$,[22] the ratio of preferred domain diameter to the wall width is

$$\frac{d_p}{l_w} = \frac{16}{\pi}\,q. \tag{84}$$

If q were much less than one then the domain wall width would be larger than the domain diameter and clearly no domain of the type which has been considered here could exist. On the other hand, if q is very large the domain wall is very narrow with respect to the domain diameter. For a given rate of flipping over of the spin systems comprising the wall, the bit-rate will thus be inversely proportional to q. It may thus be expected that everything else being equal (it is not clear at the present time just exactly what is to be held constant) that materials with high q values will have low bit-rates in devices whose speed is drive field limited. A q value of approximately three may thus be termed preferred.

The range of values of the exchange constant, A, occurring in materials which may be considered for room temperature device applications is quite limited as can be seen from the following argument: For a given structure and density of spin systems, the exchange constant is expected to be proportional to the Curie or Néel temperature, T_n.[23] For room temperature device applications, as a practical matter, T_n must be above approximately 400°K. Since the highest observed T_n are approximately 1000°K, the range of allowable T_n and therefore the range of A in acceptable materials is nearly determined.

F. B. Hagedorn, D. H. Smith, and F. C. Rossol have combined domain measurements of l with magnetometer measurements of K_u and $4\pi M_s$ to obtain the exchange constant in two materials.[24] They find for $Sm_xTb_{1-x}FeO_3$ ($x \approx 0.55$, $T_n = 661°\text{K}$), $A = 0.4 \times 10^{-6}$ ergs per cm and for $PbFe_{12-x}Al_xO_{19}$ ($x \approx 4.0$, $T_n = 508°\text{K}$) $A = 0.1 \times 10^{-6}$ ergs per cm, values which are apparently typical for high T_n iron oxides.

Since the exchange constant, A, is to be considered fixed, q has the preferred value three and d has the preferred value d_p, it is appropriate to solve for the magnetization and anisotropy constant in terms of these quantities,

$$4\pi M_s = 32(2\pi q A)^{\frac{1}{2}}/d_p \tag{85}$$

and

$$K_u = 256 A q^2/d_p^2 \, . \tag{86}$$

If $A = 4.0 \times 10^{-7}$ ergs per cm, $q = 3$ and $d_p = 10^{-3}$ cm (approximately one mil bit spacing) then $4\pi M_s \approx 80$ Gauss and $K_u \approx 900$ ergs per *cc* which are both numerically small.

Maintaining the values of A and d_p but considering q as variable (85) becomes $4\pi M_s \approx 50\sqrt{q}$. Thus q any within two orders of magnitude of the preferred value produces a value of $4\pi M_s$ of one kilogauss or smaller. Since the magnetic moment per spin system and the volume of the individual spin systems are approximately constants and since the exchange interaction is of short range, low saturation magnetizations cannot be achieved by dilution of the spin systems with magnetically unordered materials. Spin system density thus must be high while the net magnetic moment per unit volume remains low. Some sort of antiferromagnetic character is thus required in magnetic materials which are candidates for use in cylindrical domain devices. Typical systems are ferrimagnets, and canted antiferromagnets.

The K_u value of 900 ergs per cc is quite low for a symmetry allowed intrinsic uniaxial anisotropy constant.[21] When low anisotropies are obtained by operating near the reorientation temperature in canted systems[25] or near the Neel temperature, the material parameters tend to be undesirably temperature dependent. Since uniaxial anisotropy energy densities of the required value may be induced, it appears that the use of materials having induced anisotropy, such as the recently announced garnets,[17,18,26] appears quite promising.

Having determined that A is to be considered fixed and that K_u and M_s have preferred values, this section concludes by showing the dependence of several overall device parameters on these parameters and the mobility. The device parameters considered are the bit density, bit rate, the domain flux which is important in Hall effect detectors[27] and the induced voltage which is important in wire pickup loop detectors.

In typical domain devices, the bit positions form a square array with the bit spacing being three to four domain diameters (see figures of References 4 and 5). If the bit spacing is assumed to be $3d_p$, then the number of bit locations per square centimeter is

$$\rho_b = (3d_p)^{-2} = 6.9 \times 10^{-7}(4\pi M_s)^4/AK_u . \tag{87}$$

Since the bit density is such a strong function of $4\pi M_s$ the magnetization is nearly determined once a bit density is specified.

The difference of the collapse and runout bias field for a plate of the preferred thickness $h = 4l$ is $H_0 - H_2 \approx 0.1\ (4\pi M_s)$. Retaining the assumption of a bit spacing of $3d_p$ and assuming that the device structure continuously maintains a field difference across the domain of $0.1(4\pi M_s)$ the bit rate in bits per second is

$$f_b = v_b/3d_p = \mu_w 4\pi M_s/60d_p = 4.1 \times 10^{-5}\mu_w(4\pi M_s)^3(AK_u)^{-\frac{1}{2}}, \tag{88}$$

linear in the mobility and again dominated by the $4\pi M_s$ dependence of d_p. If it is assumed that a pickup loop intercepts one half of the flux emerging from the magnetic charges forming the upper surface of the cylindrical domain, then the flux change produced by moving a cylindrical domain under the loop in gauss square centimeters is

$$\Phi = \frac{\pi}{8} d_p^2(4\pi M_s) = 6.4 \times 10^4 AK_u(4\pi M_s)^{-3}. \tag{89}$$

If it is assumed that this flux change takes place in the time required for a domain driven by a field difference of $0.1(4\pi M_s)$ to propagate a distance d_p, the induced (MKS practical) voltage is

$$V = \frac{\pi}{160} \mu_w(4\pi M_s)^2 d_p \times 10^{-8} = 7.9 \times 10^{-8}\mu_w(AK_u)^{\frac{1}{2}}. \tag{90}$$

V. ACKNOWLEDGMENTS

The author wishes to acknowledge H. E. D. Scovil for his encouragement to complete this work, A. W. Anderson and F. B. Hagedorn for reading the manuscript and making comments, P. I. Bonyhard for initially demonstrating the usefulness of graphical methods for obtaining the properties of the solutions to the force equation and A. H. Bobeck, whose experiments provided much of the motivational basis for this work.

REFERENCES

1. Thiele, A. A., "The Theory of Cylindrical Magnetic Domains," B.S.T.J., *48,* No. 10 (December, 1969), pp. 3287–3335.
2. Thiele, A. A., "Theory of the Static Stability of Cylindrical Domains in Uniaxial Platelets," J. Appl. Phys., *41,* No. 3 (March, 1970), pp. 1139–1145.
3. Kooy, C., and Enz, U., "Experimental and Theoretical Study of the Domain Configuration in Thin Layers of $BaFe_{12}O_{19}$," Philips Res. Rep., *15,* No. 1 (February, 1960), pp. 7–29.
4. Bobeck, A. H., "Properties and Device Applications of Magnetic Domains in Orthoferrites," B.S.T.J., *46,* No. 8 (October, 1967), pp. 1901–1925.
5. Bobeck, A. H., Fischer, R. F., Perneski, A. J., Remeika, J. P., and Van Uitert, L. G., "Application of Orthoferrites to Domain-Wall Devices," IEEE Trans. Magnetics, *5,* No. 3 (September, 1969), pp. 544–553.
6. Perneski, A. J., "Propagation of Cylindrical Magnetic Domains in Orthoferrites," IEEE Trans. Magnetics, *5,* No. 3 (September, 1969), pp. 554–557.
7. Bobeck, A. H., unpublished work.
8. Bobeck, A. H., and Sherwood, R. C., unpublished work.
9. Della Torre, E., and Dimyan, M. Y., "Anisotropy of Wall Energy in Orthoferrites," IEEE Trans. Magnetics *MAG-6,* No. 3 (September 1970), pp. 489–492.
10. Hagedorn, F. B., Tabor, W. J., and Thiele, A. A., unpublished work.
11. Shumate, P. W., unpublished work.
12. Galt, J. K., "Motion of a Ferromagnetic Domain Wall in Fe_3O_4," Phys. Rev., *85,* No. 4 (February, 1952), pp. 664–669.
13. Galt, J. K., "Motion of Individual Domain Walls in a Nickel-Iron Ferrite," B.S.T.J., *33,* No. 5 (September, 1954), pp. 1023–1054.
14. Dillon, J. F., Jr., and Earl, H. E., Jr., "Domain Wall Motion and Ferromag-

netic Resonance in a Maganese Ferrite," J. Appl. Phys., *30,* No. 2 (February, 1959), pp. 202–213.

15. Asti, G., Colombo, M., Giudici, M., and Levialdi, A., "Magnetization Dynamics in $BaO \cdot 6Fe_2O_3$ Single Crystals Using Pulsed Magnetic Fields," J. Appl. Phys., *36,* No. 11 (November, 1965), pp. 3581–3585.
16. Asti, G., Colombo, M., Giudici, M., and Levialdi, A., "Domain-Wall Motion in Barium Ferrite Single Crystals," J. Appl. Phys., *38,* No. 5 (April, 1967), pp. 2195–2198.
17. Bobeck, A. H., "A Second Look at Magnetic Bubbles," talk given at the Int. Magnetics Conf., Washington, D. C., April 21–24, 1970.
18. Bobeck, A. H., Danylchuk, I., Remeika, J. P., Van Uitert, L. G., and Walters, E. M., "Dynamic Properties of Bubble Domains," talk given at the Int. Conf. Ferrites, Kyoto, Japan, July 6–9, 1970.
19. Rossol, F. C., "Domain-Wall Mobility in Rare-Earth Orthoferrites by Direct Stroboscopic Observation of Moving Domain Walls," J. Appl. Phys., *40,* No. 3 (March, 1969), pp. 1082–1083.
20. Rossol, F. C., and Thiele, A. A., "Domain Wall Dynamics Measured Using Cylindrical Domains," J. Appl. Phys., *41,* No. 3 (March, 1970), pp. 1163–1164.
21. Gianola, U. F., Smith, D. H., Thiele, A. A., and Van Uitert, L. G., "Material Requirements for Circular Magnetic Domain Devices," IEEE Trans. Magnetics, *5,* No. 3 (September, 1969), pp. 558–561.
22. Chikazumi, S., *Physics of Magnetism,* New York: John Wiley and Sons, Inc., 1964, Chap. 9.
23. Weiss, P. R., "The Application of the Bethe-Peierls Method to Ferromagnetism," Phys. Rev., *74,* No. 10 (November, 1948), pp. 1493–1504.
24. Hagedorn, F. B., Rossol, F. C., and Smith, D. H., unpublished work.
25. Rossol, F. C., "Temperature Dependence of Rare-Earth Orthoferrite Properties Relevant to Propagating Domain Device Applications," IEEE Trans. Magnetics, *5,* No. 3 (September, 1969), pp. 562–565.
26. Bobeck, A. H., Spencer, E. G., Van Uitert, L. G., Abrahams, S., Barns, R., Grodkiewicz, W. H., Sherwood, Schmidt, P., Smith, D. H., and Walters, E. M., "Uniaxial Magnetic Garnets for Domain Wall Bubble Devices," Appl. Phys. Letters, *17,* No. 3 (August 1970), pp. 131–134.
27. Strauss, W., and Smith, G. E., "Hall-Effect Domain Detector," J. Appl. Phys., *41,* No. 3 (March, 1970), pp. 1169–1170.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

Spreading Resistance Between Constant Potential Surfaces

By R. D. BROOKS and H. G. MATTES

(Manuscript received October 16, 1970)

We determine the spreading resistance for a sheet of homogeneous material of uniform thickness with a disk contact source on one side and with current collected (i) over the entire back plane, and (ii) at a corresponding disk on the back plane. A constant driving potential is assumed over the source resulting in mixed boundary conditions in the plane of the source. For each case, a closed form integral solution is derived and then numerically integrated for a range of geometric ratios resulting in a universal spreading resistance curve. The results are used to evaluate the spreading resistances encountered in a typical (1 Ω *cm*) *semiconductor material.*

I. INTRODUCTION

The resistance associated with the nonparallel current flow between a spacially separated source and sink is referred to as spreading resistance. Calculation of spreading resistance is often required in the analytical treatment of semiconductor devices. In particular, electrical current flow in a slice of silicon between a surface contact and a back-plane contact involves the calculation of the ohmic spreading resistance. The heat flow between an active transistor or integrated circuit and an external heat sink involves a calculation of the thermal spreading resistance in the device carrier. Also, for a given structure, the capacitance including fringing is directly related to the conductance including spreading.

Two cases are considered in this paper. The spreading resistance is determined for an infinite sheet of homogeneous material of uniform thickness with a disk contact source on one side and with current collected (*i*) at a completely metallized back plane, and (*ii*) at a corresponding disk on the back plane. D. P. Kennedy analyzed these cases for finite cylindrical volumes but with nonmixed boundary conditions.[1] He assumed a constant flux over the source region and defined spreading resistance in terms of the maximum temperature (or potential) on the disk. In this paper, we assume a constant driving potential over the source. This assumption results in a mixed boundary condition in the source plane because the potential is specified over part of the surface, and the normal derivative is specified over the rest of the surface. In the limiting case of a cylinder of infinite radius, there appears to be some difference between Kennedy's results and those presented herein. This appears to be associated with Kennedy's definition of spreading resistance in terms of the maximum disk potential. However, the comparison is based on the extrapolation of curves given by Kennedy.

A. Gray, et al., considered certain special cases of the infinite region problem with mixed boundary conditions in the source plane.[2] However, they imposed one of two constraints. Either (*i*) there was sufficient separation between source and sink so the flux distribution at the disk could be taken equal to the limiting half-space case or (*ii*) the disk was small enough to be considered a point source. In essence, these constraints again reduce the problem to one with a nonmixed boundary condition in the source plane. Neither constraint is imposed in this paper.

In this paper, for each case, a closed form integral solution is derived which is numerically integrated for a range of geometric ratios resulting in a universal spreading resistance curve. For both cases, the curves approach the half-space limit for thick sheets, and the curves asymptotically approach the nonfringing limit for very thin sheets. Finally, the results are used to evaluate the spreading resistances typical of those encountered in semiconductor technology.

II. ALGEBRAIC SOLUTION

The geometry and system of coordinates are given in Fig. 1. Solving La Place's equation in cylindrical coordinates when the potential ϕ is independent of θ gives[3]

$$\phi(\rho, z) = \int_0^\infty f(k)(C_1 \cosh kz + C_2 \sinh kz)J_0(k\rho)\, dk, \qquad (1)$$

where $f(k)$ must be determined by the boundary conditions at $z = 0$. For nonmixed boundary conditions at $z = 0$, $f(k)$ can be found by inverting the Hankel transform.[1] In the present case, however, the mixed boundary condition must be imposed at the $z = 0$ surface. Thus,

$$\begin{aligned} V &= \int_0^\infty f(k)C_1J_0(k\rho)\,dk, \qquad 0 \leqq \rho \leqq a; \\ \left.\frac{\partial\phi}{\partial z}\right|_{z=0} &= \int_0^\infty f(k)k(C_1\sinh kz + C_2\cosh kz)J_0(k\rho)\,dk\bigg|_{z=0} = 0. \end{aligned} \tag{2}$$

This results in a dual set of integral equations.

$$\begin{aligned} V &= \int_0^\infty f(k)C_1J_0(k\rho)\,dk, \qquad 0 \leqq \rho \leqq a; \\ 0 &= \int_0^\infty f(k)C_2J_0(k\rho)\,dk, \qquad \rho \geqq a. \end{aligned} \tag{3}$$

J. D. Jackson observes that the solution of this set of equations is[3]

$$f(k) = \frac{2}{\pi}\frac{Va}{C_1}\frac{\sin ka}{ka}. \tag{4}$$

Thus

$$\phi(\rho, z) = \frac{2Va}{\pi}\int_0^\infty \frac{\sin ka}{ka}\left(\cosh kz + \frac{C_2}{C_1}\sinh kz\right)J_0(k\rho)\,dk. \tag{5}$$

2.1 *Case 1—Back Plane Grounded*

In the first case considered, the second electrode is a completely metallized or grounded back plane. The boundary conditions are given in Fig. 2. This situation is representative of heat flow from an integrated circuit through an insulated header to a can acting as

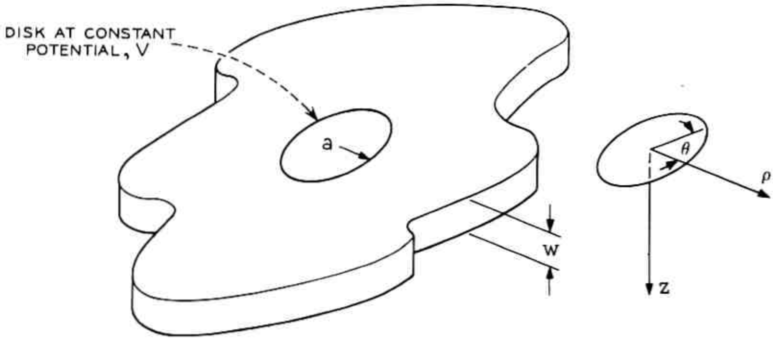


Fig. 1—Physical geometry and coordinate system for calculating spreading resistance.

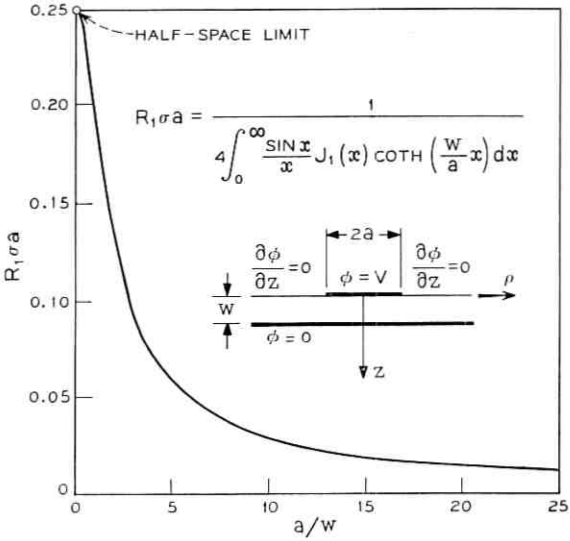


Fig. 2—Spreading resistance R_1 between disk and grounded back plane.

a thermal radiator.

$$\text{At } z = w,\ \phi(\rho, z)\,\big|_{z=w} = 0,$$

and thus

$$\int_0^\infty \frac{\sin ka}{ka} J_0(k\rho)\left(\cosh kw + \frac{C_2}{C_1} \sinh kw\right) dk = 0. \tag{6}$$

Since this must hold for all w,

$$\frac{C_2}{C_1} \sinh kw = -\cosh kw. \tag{7}$$

Thus,

$$\phi(\rho, z) = \frac{2Va}{\pi} \int_0^\infty \frac{\sin ka}{ka} J_0(k\rho)(\cosh kz - \coth kw \sinh kz)\, dk. \tag{8}$$

Now,

$$\bar{J} = \sigma \bar{E} = -\sigma \nabla \phi \qquad \text{at} \qquad z = 0, \tag{9}$$

$$I = \int_0^{2\pi} \int_0^a J \rho \, d\rho \, d\theta, \tag{10}$$

and

$$\nabla\phi \bigg|_{z=0} = \frac{\partial\phi}{\partial z}\bigg|_{z=0} = -\frac{2V}{\pi}\int_0^\infty \sin kaJ_0(k\rho)\coth kw\, dk. \tag{11}$$

Combining the last three equations and carrying out the indicated integrations, yields

$$I = 4V\sigma a \int_0^\infty \frac{\coth kw \sin kaJ_1(ka)}{k}\, dk. \tag{12}$$

Thus

$$\frac{1}{R_1} \equiv \frac{I}{V} = 4\sigma a \int_0^\infty \frac{\coth kw \sin kaJ_1(ka)}{k}\, dk \tag{13}$$

and finally,

$$R_1\sigma a = \frac{1}{4\displaystyle\int_0^\infty \frac{\sin x}{x} J_1(x) \coth\left(\frac{wx}{a}\right) dx}. \tag{14}$$

2.2 *Case 2—Disk on BackPlane Grounded*

The case where the sink consists of a disk of radius a coaxial with the source can be derived by image theory from Case 1. A drawing of this configuration is shown in Fig. 3. The values of spreading resistance in Case **2** can be derived from Case 1 by setting

$$R_1\sigma a = \frac{R_2\sigma a}{2} \tag{15}$$

and

$$\frac{a}{w}\bigg|_{\text{case 1}} = \frac{2a}{w}\bigg|_{\text{case 2}}. \tag{16}$$

The resulting spreading resistance equation is

$$R_2\sigma a = \frac{1}{2\displaystyle\int_0^\infty \frac{\sin x}{x} J_1(x) \coth\left(\frac{wx}{2a}\right) dx}. \tag{17}$$

III. NUMERICAL EVALUATION

3.1 *Numerical Integration—Universal Curves*

The integrals in equations (14) and (17) can be evaluated numerically for various a/w ratios if the upper limit is finite and the

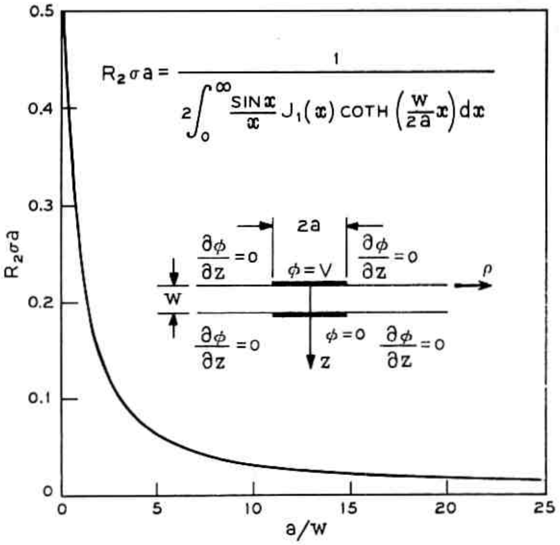


Fig. 3—Spreading resistance R_2 between disk and grounded disk on back plane.

integrands behave well at the lower limit. Investigation of the behavior of the integrands over the entire range of integration, indicated the range could be split into three ranges: 0 to 0.1, 0.1 to 300, and 300 to ∞. For the 300 to ∞ range, and $a/w \geq 1$, the integrals for both cases reduce to

$$\int_{300}^{\infty} \frac{\sin x}{x} J_1(x)\, dx \equiv I_3 \, . \tag{18}$$

But

$$I_3 = \int_0^{\infty} \frac{\sin x}{x} J_1(x)\, dx - \int_0^{300} \frac{\sin x}{x} J_1(x)\, dx \tag{19}$$

or[4]

$$I_3 = 1 - \int_0^{300} \frac{\sin x}{x} J_1(x)\, dx \equiv 1 - I_4 \, . \tag{20}$$

I_4 was evaluated numerically on a digital computer and found to be 0.96439. Thus $I_3 = 0.03561$ and is independent of a/w for $a/w \geqq 1$.

For the 0 to 0.1 range, the integrands were replaced by their small argument approximations. For $x \ll 1$,

$$J_1(x) \to \frac{x/2}{\Gamma(2)} = \frac{x}{2}, \tag{21}$$

$$\frac{\sin x}{x} \to 1, \tag{22}$$

$$\coth \frac{wx}{a} \to \frac{a}{wx}, \tag{23}$$

and,

$$\coth \frac{wx}{2a} \to \frac{2a}{wx}. \tag{24}$$

It follows that the small argument approximations for the integrands are $a/2w$, and a/w for Cases 1 and 2, respectively, and the values of the integrals over the range 0 to 0.1 are $a/20w$ and $a/10w$, respectively. This is a good approximation because the integrands are very flat functions of x near $x = 0$.

Thus for numerical purposes,

$$R_1\sigma a = \frac{1}{4\left[\dfrac{a}{20w} + \displaystyle\int_{0.1}^{300} \frac{\sin x}{x} J_1(x) \coth\left(\frac{wx}{a}\right) dx + 0.03561\right]} \tag{25}$$

and

$$R_2\sigma a = \frac{1}{2\left[\dfrac{a}{10w} + \displaystyle\int_{0.1}^{300} \frac{\sin x}{x} J_1(x) \coth\left(\frac{wx}{2a}\right) dx + 0.03561\right]}. \tag{26}$$

These results hold for $a/w \geqq 1$. For $a/w = 0$, the half-space limiting case for finite a and infinite w can be used to find values for R_1 and R_2. The resultant values are $\frac{1}{4}$ and $\frac{1}{2}$ for cases 1 and 2, respectively. This result follows directly from Jackson's work.[3]

Using the half-space limits and carrying out the numerical integrations for various values of a/w results in the universal spreading resistance curves given in Figs. 2 and 3.

3.2 *Asymptotic Limits for Cases 1 and 2*

As the ratio a/w becomes larger, the components of R_1 and R_2 due to fringing become progressively smaller. Neglecting fringing, the resistance for either case is

$$\bar{R} = \frac{w}{\sigma\pi a^2} \tag{27}$$

or

$$\bar{R}\sigma a = \frac{w}{\pi a}. \tag{28}$$

The results given in this section can be presented more clearly in terms of conductances because the fringing and nonfringing components are in parallel and $\bar{R}^{-1} \equiv \bar{G}$ is a linear function of a/w. Let the conductances corresponding to R_1 and R_2 be designated G_1 and G_2, respectively. For a given a and w, there should be more fringing in Case 1 than in Case 2, and no fringing for $\bar{R}$. Thus for any a/w,

$$G_1 > G_2 > \bar{G}. \tag{29}$$

Also G_1 and G_2 should asymptotically approach $\bar{G}$ as a function of a/w. These observations are supported by the curves of G_1, G_2, and $\bar{G}$ versus a/w given in Fig. 4. The differences of fringing components for Cases 1 and 2 are given in Fig. 5. G_1, G_2 and $\bar{G}$ are within one percent for $a/w \geqq 10$.

3.3 *Typical Example*

The results given above have been used to evaluate spreading resistances typical of those encountered in semiconductor technology. A

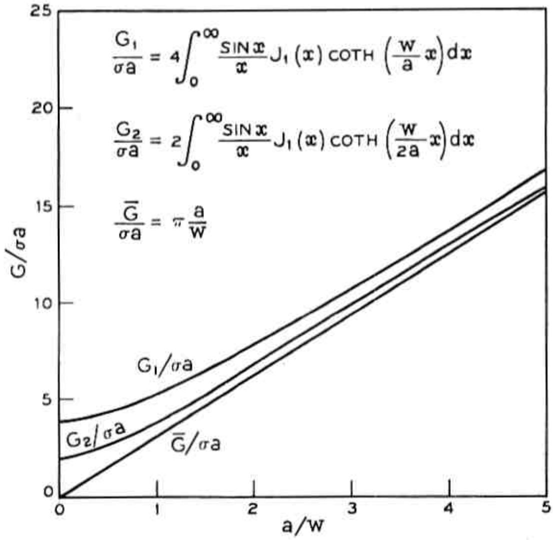


Fig. 4—Spreading conductances with and without fringing between disk and grounded back plane and between disk and grounded disk on back plane.

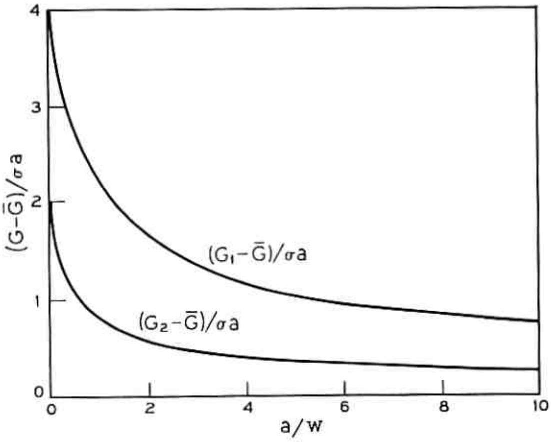


Fig. 5—Spreading conductance fringing components.

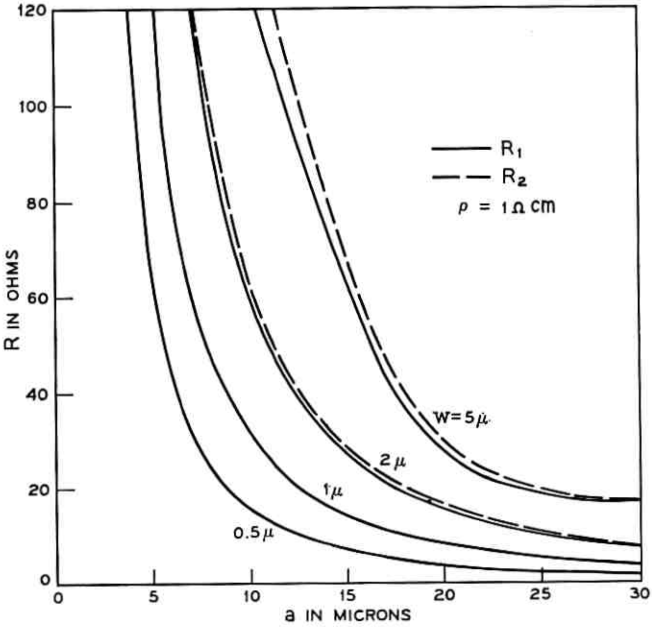


Fig. 6—Spreading resistances R_1 and R_2 for 1 Ω cm material.

1 Ω cm slice of silicon with circular contacts was considered. Figure 6 gives the calculated R_1 and R_2 for such a configuration.

IV. SUMMARY AND CONCLUSIONS

An investigation was made to determine the spreading resistance in a sheet of homogeneous material of uniform thickness with a disk contact source on one side and with current collected (*i*) at the entire back plane, and (*ii*) at a corresponding disk on the back plane. These cases were analyzed and evaluated exactly by solution of a dual set of integral equations. The method represented the boundary conditions as they physically exist. In contrast to previous work, a single, universal, spreading resistance curve was presented for each case. These curves should be useful in designing devices and analyzing materials. In particular, the curves could be used to determine conductivity of a sheet of semiconductor material if its thickness is known. Also, the curves could be used directly in the calculation of total capacitance and fringing capacitance by relabeling the ordinates with $\epsilon a/C$ instead of $R\sigma a$ and $C/\epsilon a$ instead of $G/\sigma a$.

REFERENCES

1. Kennedy, D. P., "Spreading Resistance in Cylindrical Semiconductor Devices," J. Applied Physics, *31,* No. 8 (August 1960), p. 1490.
2. Gray, A., Mathews, G. B., and Macrobert, T. M., *Bessel Functions, Second Edition,* London: Macmillan and Co., 1952, Chapter XII.
3. Jackson, J. D., *Classical Electrodynamics,* New York: John Wiley and Sons, 1962, pp. 89–93.
4. Watson, G. N., *Theory of Bessel Functions,* London: Cambridge University Press, 1952, p. 405.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

Performance of an Adaptive Echo Canceller Operating in a Noisy, Linear, Time-Invariant Environment

By J. R. ROSENBERGER and E. J. THOMAS

(Manuscript received October 8, 1970)

In this paper we derive equations describing the performance of various adaptive echo canceller configurations operating in a linear, time-invariant environment. We relate the parameters in these equations to measurable environmental factors, discuss their effect on performance, and verify the results empirically.

In general, the performance of an echo canceller cannot be exactly predicted for speech inputs. Therefore, the derived equations assume a stationary constant power random input. However, it is shown that the results obtained in this manner give useful estimates of the performance to be expected with speech inputs. The similarities and differences of the results for a constant power random input and speech input are discussed in detail.

I. INTRODUCTION

The echo problem in the telephone network is caused by the interaction of the following three factors: (*i*) The impedance mismatches that exist at hybrid junctions cause reflections of incident electrical waves. (*ii*) The existence of a bi-directional transmission medium permits the reflected signal to reach the talker as echo. (*iii*) Time delay due to the finite propagation time of a signal makes the echo annoying. Historically the problem has been alleviated by increasing trunk loss, balancing hybrids, applying four-wire circuits where practical, and providing echo suppressors.[1]

Echo suppressors are used when the echo delay exceed about 45 ms. An echo suppressor is a voice-operated device which switches a large loss in the echo path, as shown in Fig. 1a. This loss blocks the echo effectively but also tends to block speech from the near-end customer

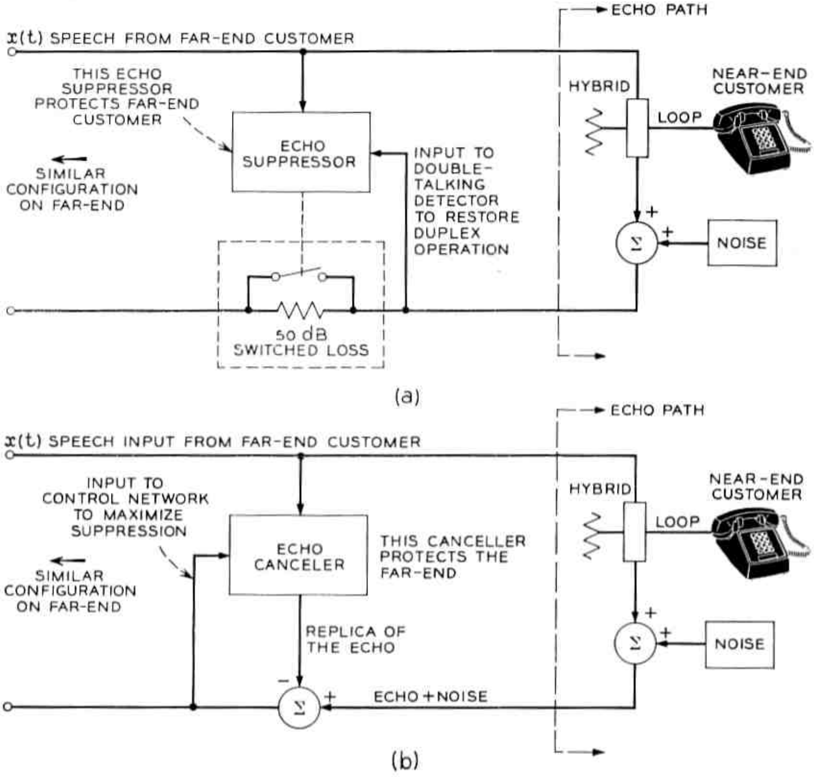

ECHO PATH
x(t) SPEECH FROM FAR-END CUSTOMER
THIS ECHO SUPPRESSOR PROTECTS FAR-END CUSTOMER
HYBRID
NEAR-END CUSTOMER
LOOP
ECHO SUPPRESSOR
SIMILAR CONFIGURATION ON FAR-END
INPUT TO DOUBLE-TALKING DETECTOR TO RESTORE DUPLEX OPERATION
+
Σ
+
NOISE
50 dB SWITCHED LOSS
(a)
ECHO PATH
x(t) SPEECH INPUT FROM FAR-END CUSTOMER
INPUT TO CONTROL NETWORK TO MAXIMIZE SUPPRESSION
HYBRID
NEAR-END CUSTOMER
LOOP
ECHO CANCELER
THIS CANCELLER PROTECTS THE FAR-END
SIMILAR CONFIGURATION ON FAR-END
REPLICA OF THE ECHO
+
Σ
+
NOISE
−
Σ
+
ECHO + NOISE
(b)

Fig. 1—Block diagrams showing how (a) an echo suppressor is applied in a telephone connection and (b) an echo canceller could be applied in a telephone connection.

when he wishes to interrupt the far-end customer. This situation is known as double-talking. During double talking it is necessary to restore the connection to full duplex. Some speech mutilation (called chopping) and echo occur during these double-talking periods. It has been shown that these degradations become increasingly disturbing as the echo delay increases.[2,3]

The performance of echo suppressors on synchronous satellite circuits is less than satisfying due to the very long delay of such circuits.[4] A new approach to the echo problem, called adaptive echo cancellation,[5-7] has been suggested as a possible alternative. In an echo canceller an approximation of the echo signal is automatically constructed and subtracted from the actual echo with no impairment to duplex operation.

In this paper, our aim is to analyze the operation of an adaptive echo canceller in a linear time-invariant environment. Our hope is to give the reader insight into the parameters which affect performance and their interrelatedness. Our method of presentation is as follows. In Section II we discuss the environmental factors affecting performance and we derive equations describing the operation of various echo canceller configurations in a linear time-invariant environment. Since we were unable to characterize an echo canceller for speech inputs, the derived equations assume a stationary random input. However, in Section III we interpret the equations and show empirically that the results obtained give useful estimates of the performance to be expected with speech inputs.

II. MATHEMATICAL DESCRIPTION OF THE ENVIRONMENT AND THE ADAPTIVE ECHO CANCELLER

The echo paths that we will consider are assumed to be linear, time-invariant channels, not necessarily band-limited and otherwise general. This is not to imply that all real echo paths can be so characterized. In fact, time-varying echo paths have been observed and others are suspected of being significantly nonlinear. These deviations from the conditions assumed above may result in serious performance limitations. References 8 and 9 describe the effect on performance when the environment is either nonlinear or time-variant.

A digital echo canceller, having filters with bandwidth, B, determined by the sampling interval, T, can be used with all echo paths so long as the filter bandwidths are at least as wide as that of the input signal, $x(t)$, i.e., $T \leqq 1/2B$. The same can be said for the bandwidths of the filters of an analog canceller. We will assume that these are also bandlimited to B Hz.

Assuming $x(t)$ and the echo signal, $y(t)$, are bandlimited to B Hz, we can equivalently represent them as sequences of the sampled values at times $t = nT$ where $n = 0, 1, 2, \cdots$. Similarly other signals pertinent to the echo canceller are discrete or continuous and have the independent variable nT or t, respectively. For the sake of brevity we will adopt a common notation, letting ζ denote t or nT. Also the convolution operation will be denoted as

$$\alpha(\zeta) * \beta(\zeta)$$

which corresponds to

$$\int_{-\infty}^{\infty} \alpha(\tau)\beta(\zeta - \tau)\, d\tau$$

in the analog case and to

$$\sum_{k=-\infty}^{\infty} \alpha(kT)\beta(\zeta - kT)$$

in the digital case. Where other differences occur the specific variable will be used.

Whether the echo canceller is digital or analog, it would be inserted into the connection as shown in Fig. 1b. The customer on the left (far end) is being protected from echo by the canceller shown. The customer on the right (near end) is being protected by a similar echo canceller on the far end of the connection.

Three factors that affect the performance of an echo canceller are the types of signals used, echo paths encountered, and echo canceller configuration. We will consider these three points separately. There are three different signals present in the echo canceller environment:

(*i*) The speech of the far-end customer, called the input signal $x(t)$.
(*ii*) The speech of the near-end customer. When the near-end customer and far-end customer speak simultaneously, we have double-talking. This constitutes an interference to the echo canceller.
(*iii*) Interfering circuit noise which is inherent to the echo path.

The echo canceller must perform satisfactorily when these signals are present in all possible combinations. Circuit noise, denoted as $\rho(t)$, is assumed to be a zero mean random process with variance σ_ρ^2 band-limited to B Hz.

A block diagram of the canceller circuit used is shown in Fig. 2. The basic components of this canceller are:

(*i*) A set of M filters having orthonormal impulse responses which are the first M members of a complete basis set.
(*ii*) A control network which automatically weights and sums the outputs of the M filters to generate an approximation of the echo.
(*iii*) Devices to couple the canceller to the telephone plant. The A-D and D-A converters are required for an analog canceller operating in a digital plant or vice versa. The set of M filters have impulse responses $\lambda_1(\zeta)$, $\lambda_2(\zeta)$, $\cdots$, $\lambda_M(\zeta)$. The output of each filter, denoted as $w_m(\zeta)$ and given by

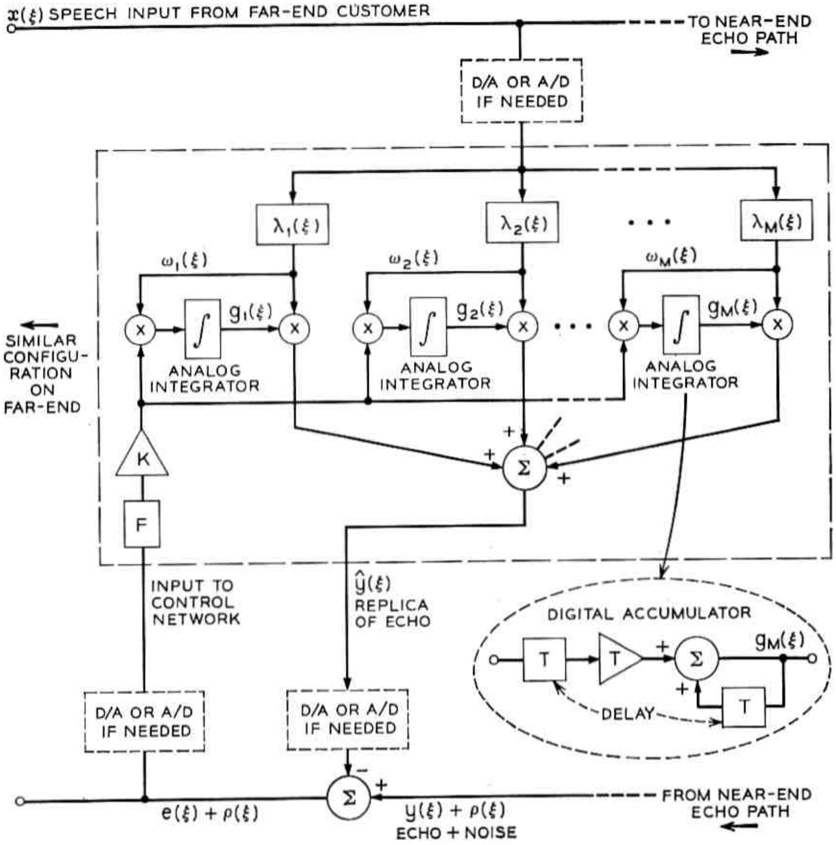


Fig. 2—Block diagram of the structure of an echo canceller.

$$w_m(\zeta) = \lambda_m(\zeta) * x(\zeta), \tag{1}$$

is weighted by the value of the tap gain $g_m(\zeta)$. At $\zeta = 0$ the tap gain is set to some initial value (usually assumed to be zero). The sum of these weighted outputs is the approximation of the echo and is denoted as $\hat{y}(\zeta)$. Thus we have

$$\hat{y}(\zeta) = \sum_{m=1}^{M} g_m(\zeta) w_m(\zeta) \tag{2}$$

which is subtracted from $y(\zeta)$ to give the cancelled echo denoted as $e(\zeta)$. The cancelled echo plus noise $\rho(\zeta)$ is operated on by a function F and then multiplied by a positive factor K. F may be any odd nondecreasing function.

We will consider the two cases:

$$(i)\quad F[\cdot] = [\cdot] \quad \text{and}$$

$$(ii)\quad F[\cdot] = \operatorname{sgn}[\cdot] = \begin{cases} +1 & \text{if} \quad [\cdot] > 0, \\ -1 & \text{if} \quad [\cdot] < 0. \end{cases}$$

The resulting signal is multiplied by $w_m(\zeta)$ and integrated to yield the value of each $g_m(\zeta)$.

The analog network is governed by the set of differential equations

$$\frac{d}{dt} g_m(t) = KF[e(t) + \rho(t)]w_m(t), \qquad m = 1, 2, \cdots, M. \tag{3}$$

The digital network is governed by the set of difference equations

$$\begin{aligned} g_m(nT) = g_m(nT - T) &+ KTF[e(nT - T) \\ &+ \rho(nT - T)]w_m(nT - T), \qquad m = 1, 2, \cdots, M. \end{aligned} \tag{4}$$

We can write the error, $e(\zeta)$, as

$$\begin{aligned} e(\zeta) &\equiv y(\zeta) - \hat{y}(\zeta) \\ &= \sum_{m=1}^{M} [c_m - g_m(\zeta)]w_m(\zeta) + q(\zeta), \end{aligned} \tag{5}$$

where

$$q(\zeta) \equiv \sum_{m=M+1}^{\infty} c_m \lambda_m(\zeta) * x(\zeta). \tag{6}$$

The coefficients c_m, $m = 1, 2, \cdots$ are the generalized Fourier coefficients of the echo path transfer function $H(f)$ over the bandwidth $|f| \leqq B$ relative to the complete basis set. They are given by the equation

$$c_m = \int_{-B}^{B} H(f)\overline{\Lambda_m}(f),^{\dagger} \qquad m = 1, 2, \cdots \tag{7}$$

where $\Lambda_m(f)$ is the Fourier transform of $\lambda_m(\zeta)$. The term $q(\zeta)$ is called the uncancellable part of $e(\zeta)$.

Echo suppression achieved ζ seconds after the start of canceller operation is defined as

$$S(\zeta) \equiv -10 \log \frac{E[e^2(\zeta)]^{\ddagger}}{E[y^2(\zeta)]}. \tag{8}$$

[†] The overbar denotes complex conjugation.

[‡] E denotes ensemble average.

Maximum achievable suppression is denoted as $S_{\max}$ and equals $\lim_{\zeta\to\infty}[S(\zeta)]$. We define the average settling time t_s to be the time in seconds required for the suppression $S(\zeta)$ to each 98 percent of $S_{\max}$ in decibels.

We will now derive equations for maximum achievable suppression and average settling time for the two cases $F[\cdot] = [\cdot]$ and $F[\cdot] = \operatorname{sgn}[\cdot]$. To facilitate the derivations we define the column matrices

$$W(\zeta) \equiv \begin{bmatrix} w_1(\zeta) \\ \vdots \\ w_M(\zeta) \end{bmatrix}$$

and

$$R(\zeta) \equiv C - G(\zeta) \equiv \begin{bmatrix} c_1 \\ \vdots \\ c_M \end{bmatrix} - \begin{bmatrix} g_1(\zeta) \\ \vdots \\ g_M(\zeta) \end{bmatrix}.$$

Using these matrices, we may write $e(\zeta)$ as

$$e(\zeta) = R'(\zeta)\cdot W(\zeta) + q(\zeta)^{\dagger} \tag{9}$$

and $y(\zeta)$ as

$$y(\zeta) = C'\cdot W(\zeta) + q(\zeta). \tag{10}$$

In the derivations which follow we will assume:

(*i*) The input signal, $x(t)$, is a stationary random process having a rectangular power density spectrum

$$P_x(f) = \begin{cases} \sigma_x^2/2B; & |f| \leq B; \\ 0 \;\;; & |f| > B; \end{cases}$$

(*ii*) The circuit noise, $\rho(t)$, is a stationary zero mean random process bandlimitted to B Hz and independent of $x(t)$;

(*iii*) For the case $F[\cdot] = \operatorname{sgn}[\cdot]$ we will further assume that $x(t)$ and $\rho(t)$ are gaussian with zero mean;

(*iv*) $R(\zeta)$ is independent of both $x(\zeta)$ and $\rho(\zeta)$.

With regard to the last assumption, it is clear that, since $R(\zeta)$ is a function of $x(\zeta)$ and $\rho(\zeta)$, it cannot truly be independent of $x(\zeta)$ and $\rho(\zeta)$. For reasonable values of the feedback factor K, the rate of change

[†] An apostrophe denotes matrix transpose and a dot denotes scalar multiplication.

of $R(\varsigma)$ will be much slower than that of $x(\varsigma)$ or $\rho(\varsigma)$, so that the assumption is justified for a wide range of operating conditions.

Substituting equations (9) and (10) into equation (8), it can be shown that

$$S(\varsigma) = -10 \log \left[\frac{\int_{-B}^{B} |H(f)|^2 \, df - C' \cdot C + E[R'(\varsigma) \cdot R(\varsigma)]}{\int_{-B}^{B} |H(f)|^2 \, df} \right]. \tag{11}$$

The integral in the denominator of equation (11) is defined as the echo path energy which will be denoted as ψ. We see from equation (11) that $S(\zeta)$ is maximized for a given echo path if $E[R'(\zeta) \cdot R(\zeta)] = 0$. The maximum value is

$$S(\varsigma) = -10 \log \left[\frac{\psi - C' \cdot C}{\psi} \right]. \tag{12}$$

Actually this suppression may not be achieved because $E[R'(\zeta) \cdot R(\zeta)]$ may not vanish. However, equation (12) gives a theoretical limit on suppression as defined—this limit being a function of the basis set used and the filter set truncation. We will define an incompleteness (truncation) factor I as

$$I \equiv \frac{\psi - C' \cdot C}{\psi}. \tag{13}$$

Note that I is a nonseparable function of the environment and the echo canceller. That is to say, to calculate the incompleteness factor one must know the echo-path transfer function over the bandwidth of the input signal, the filter set used in the echo canceller and the number of taps employed in the canceller.

To find maximum achievable suppression and average settling time, we must evaluate the term $E[R'(\zeta) \cdot R(\zeta)]$ for the digital and analog case under the two conditions of the function F.

2.1 *Evaluation of $E[R'(t) \cdot R(t)]$ for the Analog Case to Yield $S_{\max}$ and t_s*

Using the definition of $R(\varsigma)$ and equation (3), we may write

$$\frac{d}{dt}[R'(t) \cdot R(t)] = -2KF[R'(t) \cdot W(t) + q(t) + \rho(t)]R'(t) \cdot W(t). \tag{14}$$

2.1.1 *The Case $F[\cdot] = [\cdot]$*

For this case, we can write equation (14) as

$$\frac{d}{dt}\,[R'(t)\cdot R(t)] = -2K\{[R'(t)\cdot W(t)]^2 + [R'(t)\cdot W(t)][q(t) + \rho(t)]\}. \quad (15)$$

Solving equation (15) for the expectation of $R'(t)\cdot R(t)$ gives

$$E[R'(t)\cdot R(t)] = R'(0)\cdot R(0)\,\exp\left(-\frac{K\sigma_x^2 t}{B}\right) \qquad t \geqq 0. \qquad (16)$$

Substituting equation (16) into equation (11) yields

$$S(t) = -10\,\log\left[I + \frac{R'(0)\cdot R(0)}{\psi}\,\exp\left(-\frac{K\sigma_x^2}{B}\,t\right)\right]. \qquad (17)$$

As $t \to \infty$, we have

$$S_{\max} = -10\,\log I. \qquad (18)$$

For the given assumptions, the maximum achievable suppression is not a function of circuit noise and is limited only by the incompleteness of the filter set. Of course, strictly speaking, $S_{\max}$ would be less than this limit by an amount depending on the correlation existing between $R(t)$ and $\rho(t)$.

Defining the term

$$s(t_s) \equiv \log^{-1}\left(-\frac{0.98S_{\max}}{10}\right), \qquad (19)$$

the antilog of the suppression at the settling time t_s, substituting this into equation (17) and rearranging we get

$$t_s = \frac{B\,\log\left[\dfrac{R'(0)\cdot R(0)}{\psi(s(t_s) - I)}\right]}{0.434K\sigma_x^2}. \qquad (20)$$

2.1.2 *The Case* $F[\cdot] = \operatorname{sgn}[\cdot]$ *(hard limiter)*

As above, $S_{\max}$ and t_s may be derived yielding the following two equations:

$$S_{\max} = -10\,\log I \qquad (21)$$

and

$$t_s = \frac{B}{K\sigma_x^2\sqrt{\dfrac{2}{\pi}}}\left[2(\gamma - \mu) + 2.3\theta\,\log\left(\left[\frac{\gamma - \theta}{\gamma + \theta}\right]\left[\frac{\mu + \theta}{\mu - \theta}\right]\right)\right] \qquad (22)$$

where

$$\mu = \left(\frac{\sigma_\rho^2 + \psi\sigma_x^2 s(t_s)}{2B}\right)^{\frac{1}{2}}.$$

$$\theta = \left(\sigma_\rho^2 + \frac{\sigma_x^2(\psi - C'\cdot C)}{2B}\right)^{\frac{1}{2}},$$

$$\gamma = \left(\sigma_\rho^2 + \frac{\sigma_x^2[\psi - C'\cdot C + R'(0)\cdot R(0)]}{2B}\right)^{\frac{1}{2}}.$$

If the circuit noise $\rho(t)$ can be neglected and if $I \approx 0$ then equation (22) can be written as

$$t = \frac{1}{\sigma_x K}\,[\pi B R'(0)\cdot R(0)]^{\frac{1}{2}}[1 - \sqrt{s(t)}] \tag{23}$$

where $s(t)$ is the antilog of the suppression at time t.

2.2 *Evaluation of* $E[R'(nT)\cdot R(nT)]$ *for the Digital Case to Yield* $S_{\max}$ *and* t_s

Using the definition of $R(\zeta)$ and equation (4), we can write

$$\begin{aligned} R'(nT)\cdot R(nT) &- R'(nT - T)\cdot R(nT - T) \\ &= -2KTF[R'(nT - T)\cdot W(nT - T) + q(nT - T) \\ &\quad + \rho(nT - T)]R'(nT - T)\cdot W(nT - T) \\ &\quad + K^2T^2F^2[R'(nT - T)\cdot W(nT - T) + q(nT - T) \\ &\quad + \rho(nT - T)]W'(nT - T)\cdot W(nT - T). \end{aligned} \tag{24}$$

2.2.1 *The Case* $F[\cdot] = [\cdot]$

Solving equation (24) for the expectation of $R'(t)\cdot R(t)$, it may be shown that $S(nT)$ is given by

$$\begin{aligned} S(nT) = -10 \log \Bigg[I\left(1 + MK^2T^2\sigma_x^4\left[\frac{\alpha^n - 1}{\alpha - 1}\right]\right) \\ + \frac{R'(0)\cdot R(0)}{\psi}\,\alpha^n + \frac{MK^2T^2\sigma_x^4}{\nu}\left(\frac{\alpha^n - 1}{\alpha - 1}\right)\Bigg] \end{aligned} \tag{25}$$

for $n = 0, 1, 2, \cdots$ and where

$$\begin{aligned} \alpha &= 1 - KT\sigma_x^2[2 - KT(M + 2)\sigma_x^2], \qquad 0 < \alpha < 1; \\ \beta &= MK^2T^2\sigma_x^4(\psi - C'\cdot C) + MK^2T^2\sigma_x^2\sigma_\rho^2. \end{aligned} \tag{26}$$

Note that the limits on α are necessary to yield a convergent system. We define the signal to noise ratio ν at the output of the echo path as

$$\nu = \frac{\psi\sigma_x^2}{\sigma_\rho^2}, \tag{27}$$

and

$$\frac{\mathrm{S}}{\mathrm{N}} = 10 \log \nu \text{ dB}.$$

The maximum suppression is given from the limiting value of $S(nT)$ as $n \to \infty$ as

$$S_{\max} = -10 \log \left[I\left(1 - \frac{MK^2T^2\sigma_x^4}{\alpha - 1}\right) - \frac{MK^2T^2\sigma_x^4}{\nu(\alpha - 1)} \right], \tag{28}$$

and t_s is given by

$$t_s = \frac{T \log \left[\dfrac{s(t_s) - s_{\max}}{\dfrac{R(0)\cdot R(0)}{\psi} + \dfrac{IM(KT\sigma_x^2)^2}{\alpha - 1} + \dfrac{M(KT\sigma_x^2)^2}{\nu(\alpha - 1)}} \right]}{\log [\alpha]}. \tag{29}$$

2.2.2 *The Case* $F[\cdot] = \operatorname{sgn}[\cdot]$

With the hard limiter we can write equation (24) as

$$\begin{aligned} R'(nT)\cdot R(nT) - R'(nT - T)\cdot R(nT - T) \\ = -2KT \operatorname{sgn} [R'(nT - T)\cdot W(nT - T) + q(nT - T) \\ + \rho(nT - T)]W'(nT - T)\cdot W(nT - T) \\ + K^2T^2W'(nT - T)\cdot W(nT - T). \end{aligned} \tag{30}$$

Using the same assumptions and analysis technique as used for the analog case, we can derive the average value of equation (30) obtaining

$$\begin{aligned} E[R'(nT)\cdot R(nT)] - E[R'(nT - T)\cdot R(nT - T)] \\ = -2KT\sigma_x\sqrt{\frac{2}{\pi\psi}} \frac{E[R'(nT - T)\cdot R(nT - T)]}{\sqrt{\dfrac{E[R'(nT - T)\cdot R(nT - T)]}{\psi} + I + \dfrac{1}{\nu}}} \\ + MK^2T^2\sigma_x^2. \end{aligned} \tag{31}$$

Rather than attempt a solution of this nonlinear difference equation, we will find only the limiting value of $E[R'(nT)\cdot R(nT)]$. For

$E[R'(nT) \cdot R(nT)]$ to converge, we require that the right hand side of equation (31) be nonpositive. Using this inequality and solving for $E[R'(nT) \cdot R(nT)]$ gives

$$S_{\max} \leqq -10 \log \left\{ I + \frac{\pi (MKT\sigma_x)^2}{16\psi} + \left[\frac{\pi^2 (MKT\sigma_x)^4}{256\psi^2} + \frac{2\pi (MKT\sigma_x)^2}{16\psi} \left(I + \frac{1}{\nu} \right) \right]^{\frac{1}{2}} \right\}. \tag{32}$$

An explicit expression for the average settling time t_s for this case is not available. As an alternative, the analog equation for settling time t_s, equation (22), with K replaced by KT may be used to predict settling time for this digital case. When using equation (22), we should use equation (32) to calculate $s(t_s)$.

III. EMPIRICAL RESULTS AND INTERPRETATIONS OF THE THEORETICAL RESULTS

In this section we will compare the theoretical results with empirical findings. We will then discuss the performance predicted by the equations as several of the parameters are varied. Finally we will demonstrate that although the equations were derived for a noise input, they yield useful information about the operation of the canceller for speech inputs provided that the echo path is linear and time-invariant. The empirical results tabulated below pertain only to digital implementations, since an analog system was not available for testing.

The echo canceller shown in Fig. 2 was simulated on a digital computer. Also an echo path, chosen to have characteristics which are similar to those of real echo paths which have been observed, was simulated on the computer. Experiments have also been performed incorporating various analog echo paths with the results in general agreement with predicted performance. For the sake of brevity the latter results are omitted.

The measure of echo canceller suppression which we use to monitor canceller performance is defined by the equation

$$S(\zeta) = -10 \log \frac{\displaystyle\int_0^B \left| H(f) - \sum_{m=1}^{M} g_m(\zeta) \Lambda_m(f) \right|^2 df}{\displaystyle\int_0^B |H(f)|^2 \, df} . \tag{33}$$

Equation (33) yields a measure of the goodness of fit across the entire band at time ζ. It is not necessarily equivalent to the subjective echo reduction which a listener would perceive. In fact subjective testing on a very limited basis indicates that perceived loss may be greater than is indicated by this measure. It is important that this subjective factor be considered when comparing the predicted suppression of an echo canceller with what is considered to be necessary for adequate echo reduction. It is clear that equation (33) would give values of suppression identical to the values given by the previously derived equations if the input to the canceller is noise with a rectangular power density spectrum.

3.1 *Random Noise Input Results*

3.1.1 $F[\cdot] = [\cdot]$

In Fig. 3 we plot suppression for a typical simulation. The choice of parameters for the simulation is listed on the figure. The crosses indicate the results of the computer simulation [equation (33)] which is compared against the values of settling time, t_s, and maximum suppression, $S_{\max}$, as predicted by equations (29) and (28) respectively. We also compare the results of the simulation against the suppression as a function of time predicted by equation (25). We see from the figure that a high value of suppression is obtainable.

In Fig. 4, we allowed the echo canceller to reach its maximum suppression with no circuit noise present. Then we introduced a high noise level, S/N = −18 dB, for approximately 1.5 seconds and then removed it, simulating doubletalking. It is clear from the figure that the results are in very good agreement with the equations.

From these results and numerous others using different echo paths and basis sets we draw the following conclusions:

> The assumption of the independence of $R(\zeta)$ from $x(\zeta)$ and $p(\zeta)$ is quite reasonable for suppressions of up to 40 dB, S/N ratios as low as −20 dB, and settling times of 0.3 second and greater. Therefore for random noise inputs we conclude that the derived equations are very accurate predictors of the performance of an echo canceller.

We now focus our attention on the nature of the equations, and discuss the effects of various environmental factors upon them.

In Figs. 5 through 7, we plot maximum suppression (28) versus $KT\sigma_x^2$ for 100 taps and the incompleteness factors $I = 0.01, 0.001, 0.0001$. In all three figures $S_{\max}$ decreases as the S/N decreases. Note that for small values of $KT\sigma_x^2$, as the incompleteness factor I decreases,

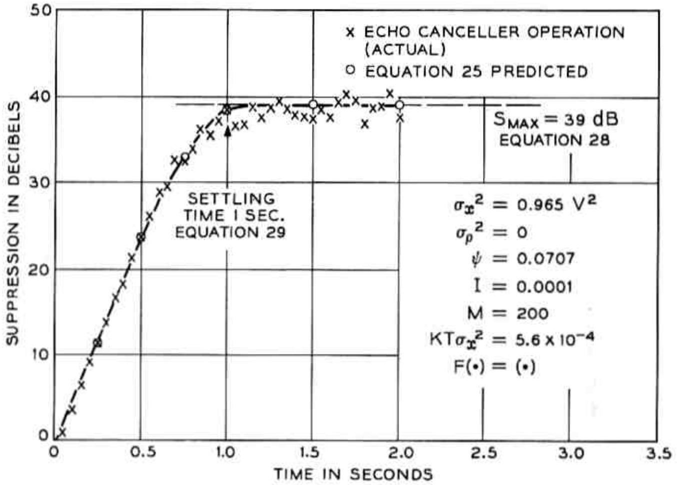


Fig. 3—Comparison of the results of a computer simulation of an echo canceller with the results predicted by the equations. The crosses indicate the simulation results and the circles indicate the predicted results.

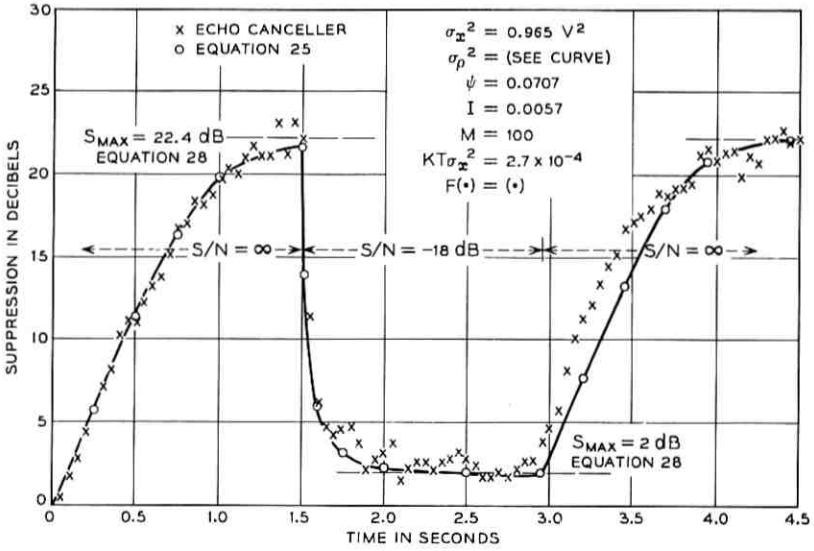


Fig. 4—The same as Fig. 3 except for a different choice of parameters.

the increase in suppression becomes larger for equal increments in S/N. That is, circuit noise (and double-talking) becomes more troublesome as the echo canceller is designed to give higher and higher values of $S_{\max}$.

Note that for fixed S/N, K, and T, an increase in the input signal power reduces the maximum achievable suppression. On the other hand, for fixed σ_p^2 and $KT\sigma_x^2 < 10^{-3}$, a change in signal has no appreciable affect upon maximum suppression. However from equation (25), we see that convergence is assured if and only if $0 < \alpha < 1$. This in turn, implies that $KT\sigma_x^2 < 2/(M + 2)$. Therefore, we cannot make $KT\sigma_x^2$ arbitrarily large.

Figures 5 through 7 were calculated for $M = 100$. In order to investigate the sensitivity of $S_{\max}$ to M we have plotted $S_{\max}$ versus M for $KT\sigma_x^2 = 0.0001$ and $I = 0.001$ and 0.01 in Fig. 8. Observe that $S_{\max}$ is a weak function of M. Thus, Figs. 5 through 7 can be used to predict $S_{\max}$ for given I and $KT\sigma_x^2$ with little regard for M.

In Figs. 9 through 11, we plot settling time versus $KT\sigma_x^2$ for various choices of I and S/N. Note that a decreasing S/N results in a decreasing settling time. This could lead to the erroneous conclusion that high noise levels help convergence. We find, however, that as the noise level increases, $S_{\max}$ decreases. In some cases of very high noise levels, the echo canceller could even provide a net gain. Intuitively it is clear that, starting with zero suppression, it should take less time to settle to the lower level of suppression. For example, consider Fig. 9, with $KT\sigma_x^2 =$

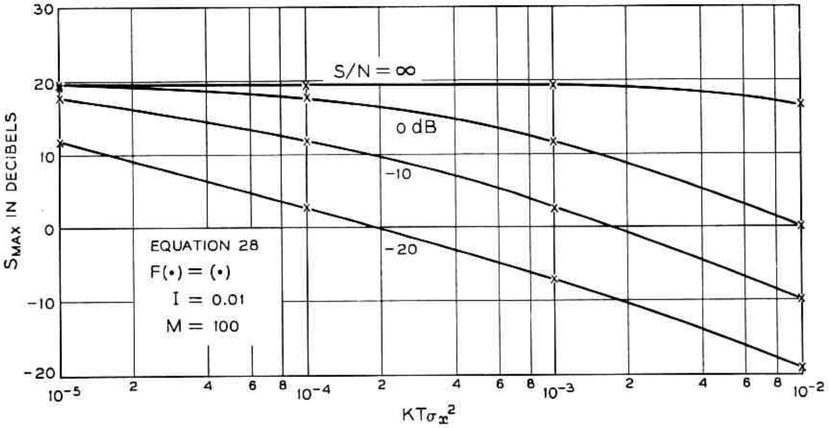


Fig. 5—Theoretically maximum suppression versus $KT\sigma_x^2$ for an incompleteness factor of 0.01 and various echo-to-noise ratios.

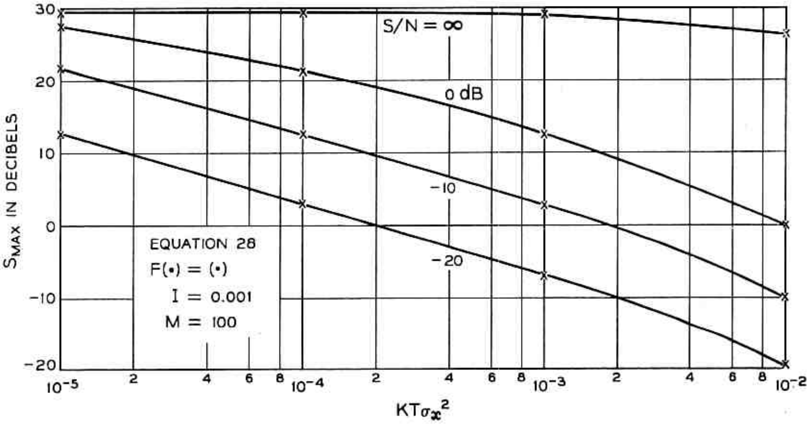


Fig. 6—The same as Fig. 5 but with an incompleteness factor of 0.001.

10^{-3}. It shows that for S/N = ∞ the settling time is approximately 3.6×10^3 iterations versus 1.6×10^3 for S/N = −20 dB. However, Fig. 5 shows that $S_{\max}$ is 19 dB and −7 dB respectively. This situation also illustrates what may happen when a strong interference such as double-talking occurs. The interference will cause divergence to a reduced suppression and may even cause a net gain. We also conclude

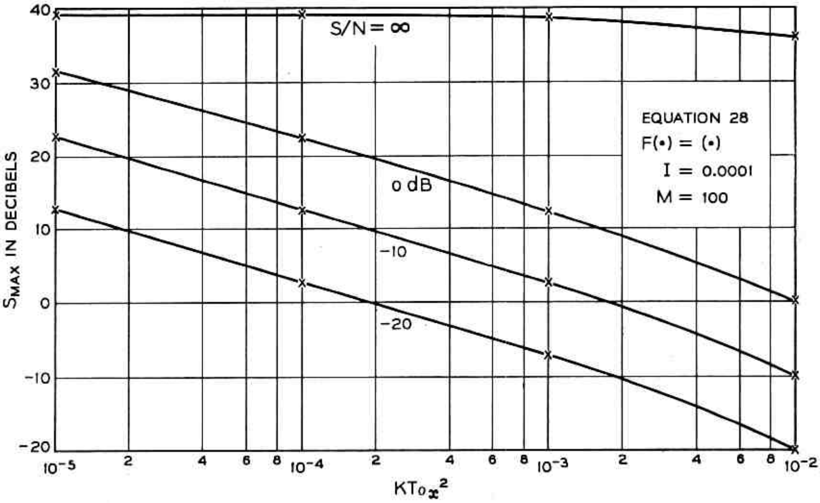


Fig. 7—The same as Figs. 5 and 6 but with an incompleteness factor of 0.0001.

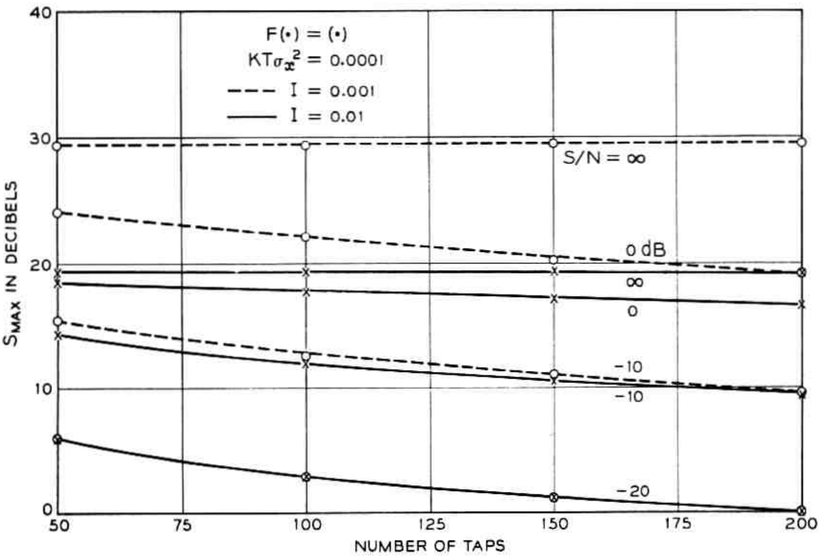


Fig. 8—Theoretically maximum suppression versus number of taps for various echo-to-noise ratios and incompleteness factors.

that for fixed S/N and KT, the larger the input signal power, the faster the settling time will be. However, for the reasons explained previously, the signal power cannot be made arbitrarily large. For constant noise power and fixed KT, it can be seen that settling time is decreased and the rate of increase of suppression is made larger with increased input signal power.

In Fig. 12 we plot settling time as a function of the number of filters M for several values of S/N and I. We see that settling time is relatively insensitive to M and that Figs. 9 through 11 may be used to estimate settling time irrespective of M.

3.1.2 $F[\cdot] = \text{sgn}\,[\cdot]$

We now consider the echo canceller with a hard limiter in the feedback loop. We cannot predict the exact temporal performance of this canceller configuration because we have no solution to the governing difference equation (31). However, we may estimate it by using the solution of the analog differential equation and replacing K with KT. Since this imposes no limit on maximum suppression, we must combine this with the limiting value of $S_{\max}$ given by equation (32). This technique yields a reasonable prediction of the operation. For this case

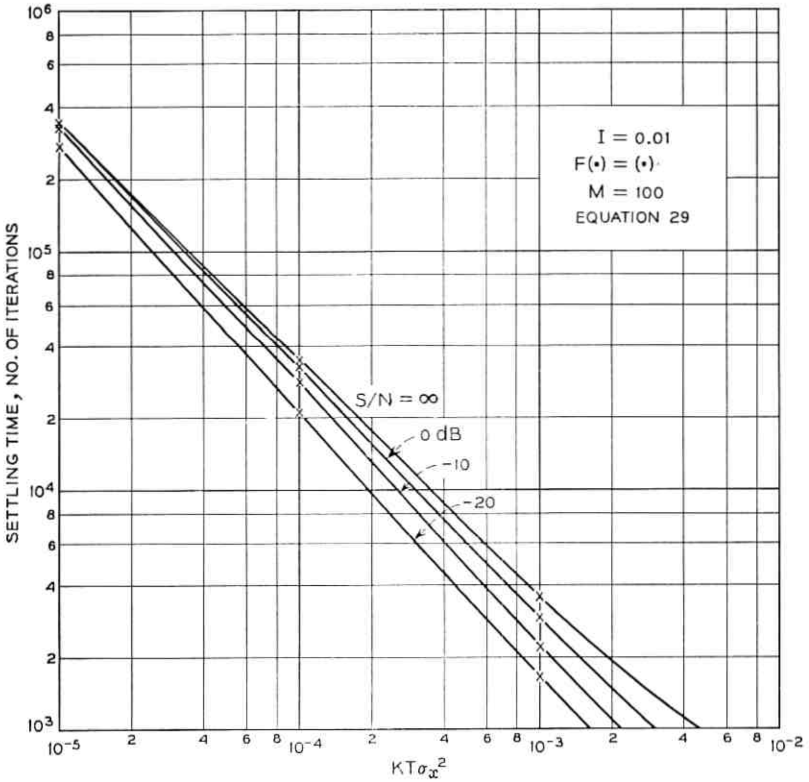


Fig. 9—Settling time versus $KT\sigma_x^2$ for an incompleteness factor of 0.01 and various echo-to-noise ratios.

there are too many variables to present easily a set of curves which describes the operation quantitatively. Therefore we will make some qualitative observations which are generally true. We will use Figs. 13 and 14 as typical examples but we emphasize that these curves are only quantitatively valid for the particular choice of parameters given.

We observe that $S_{\max}$ and settling time are inversely proportional to σ_x. For fixed $KT\sigma_x$, $S_{\max}$ decreases as S/N decreases. For constant signal to noise ratio, $S_{\max}$ decreases with increasing $KT\sigma_x$. However, for constant noise level, $S_{\max}$ is relatively insensitive to changes in $KT\sigma_x$ over a wide range. Note too that for $KT\sigma_x$ sufficiently large the canceller may introduce a net gain. For fixed S/N, settling time is a decreasing function of $KT\sigma_x$. For fixed $KT\sigma_x$ a decrease in S/N produces an increase in settling time, while for the case $F[\cdot] = [\cdot]$ the opposite was true. We will now turn our attention to the operation of the echo canceller with speech; we will attempt to interpret the equations in this new light. We will also attempt to show empirically that the results we obtain give useful estimates of performance.

3.2 *Operation With Speech*

The fundamental differences between noise and speech are:

(*i*) The short time (50 ms) average power of speech is erratic from time interval to time interval whereas by comparison it is relatively constant for the random noise.

(*ii*) The spectral density of speech is nonuniform, and depends on

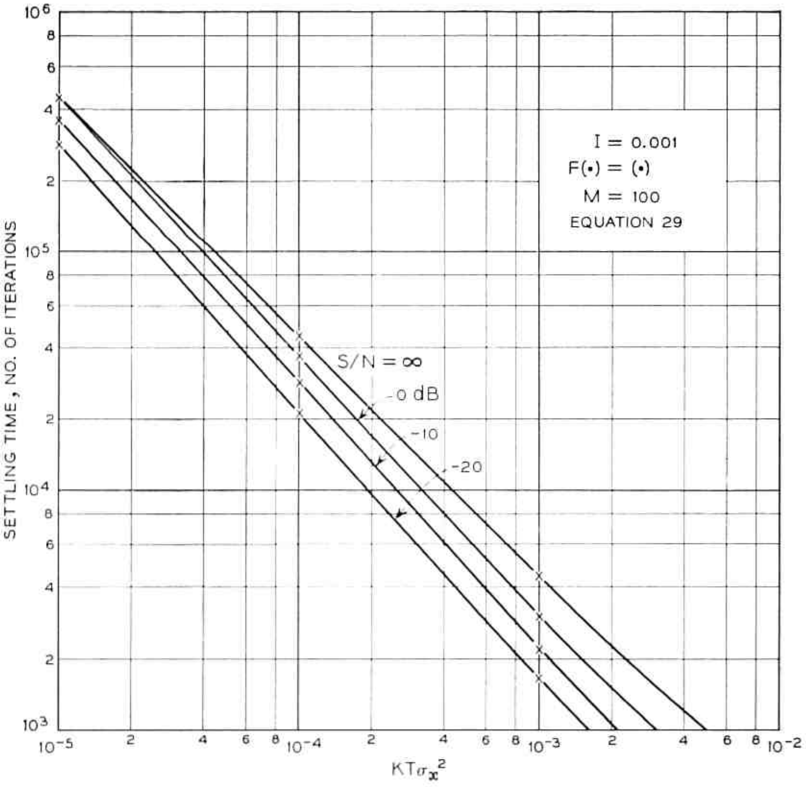


Fig. 10—The same as Fig. 9 but with an incompleteness factor of 0.001.

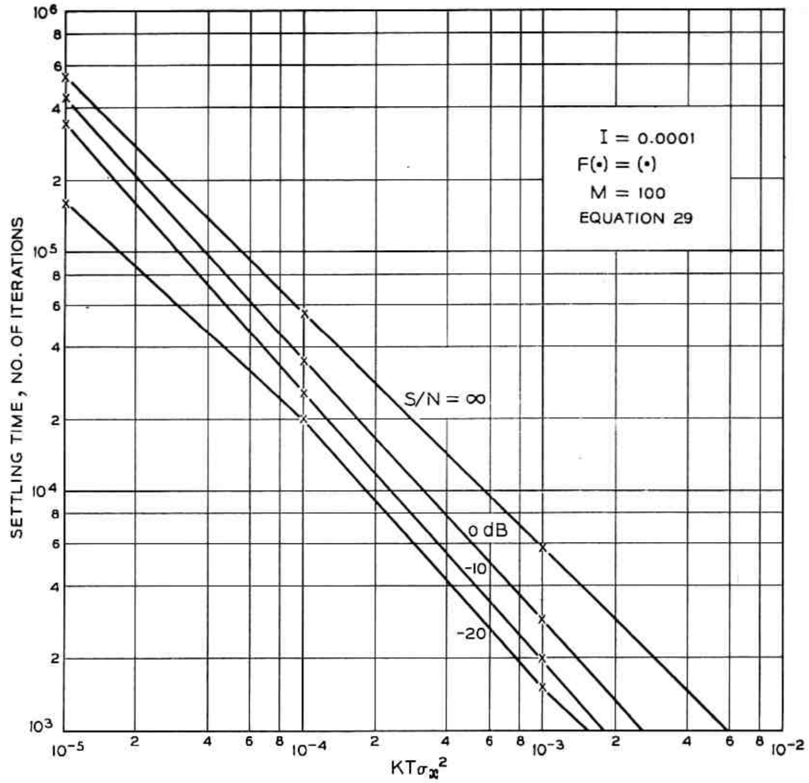


Fig. 11—The same as Figs. 9 and 10 but with an incompleteness factor of 0.0001.

which phoneme is spoken and who speaks it. In fact, the speech power is usually concentrated only a few narrow frequency bands at a time. However, if enough time is allowed to elapse, it is reasonable to assume that the speech power* will eventually scan the entire available bandwidth.

At present, no adequate statistical description of a speech signal accounting for the above properties is available. Using the long-time (several seconds) estimate of average speech power, we have found that the results derived in this paper for random noise may be used as an estimate of the performance which can be expected with speech

* Strictly speaking, this is also true for the random-noise case. However, for noise, the power density spectrum may be considered uniform for a shorter period of elapsed time.

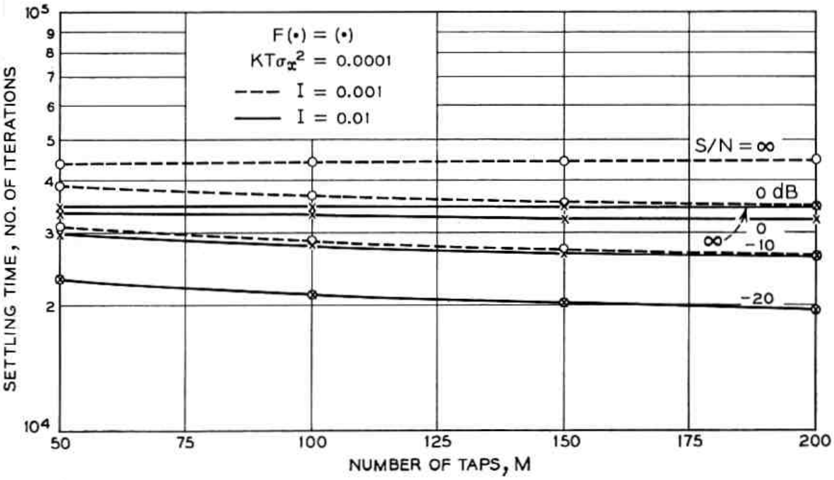


Fig. 12—Settling time versus number of taps for incompleteness factors of 0.01 and 0.001.

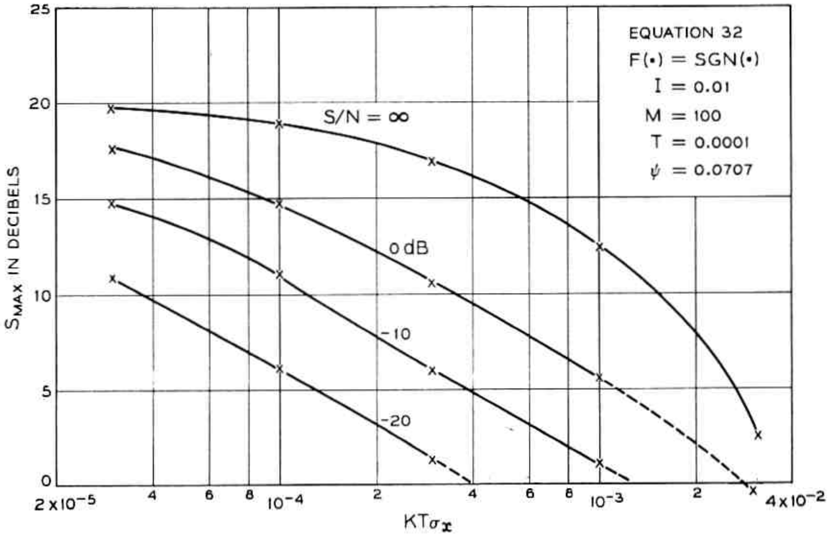


Fig. 13—Theoretically maximum suppression versus $KT\sigma_x$ for an incompleteness factor of 0.01 and various echo-to-noise ratios.

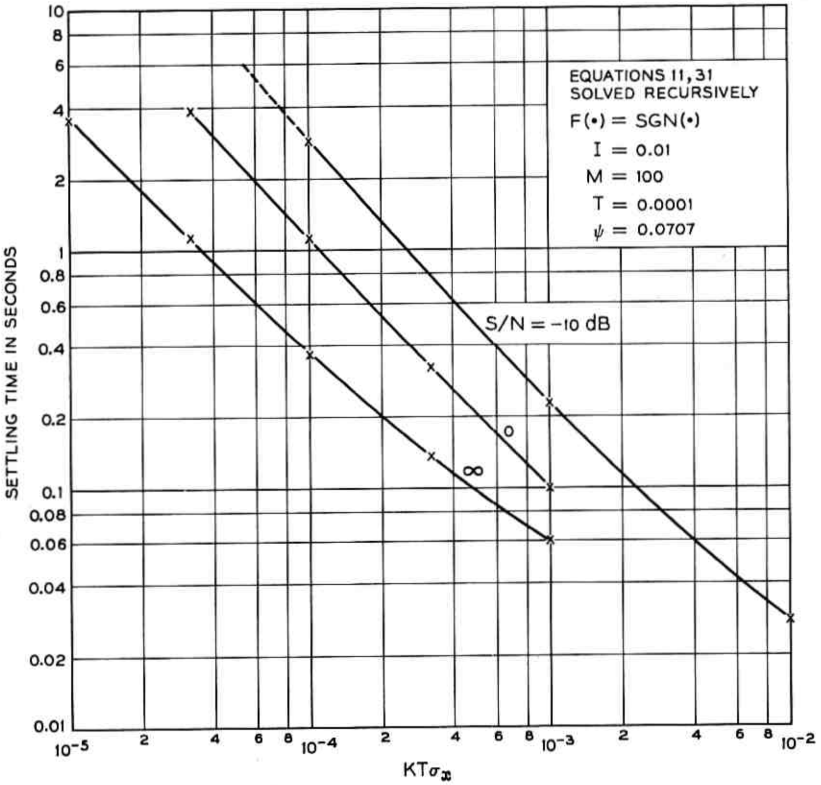


Fig. 14—Settling time versus $KT\sigma_x$ for an incompleteness factor of 0.01 and various echo-to-noise ratios.

inputs. However, the blind use of the equations may give erroneous and optimistically deceiving estimates. We show below how the equations should be used to predict the maximum achievable suppression obtained from the echo canceller with a speech input, and discuss the significance of the settling time estimates.

Because of the variation in speech power level on a short-time (50 ms) basis, we find that the rate of convergence of an echo canceller is erratic. To illustrate this, assume for the moment that we have available speech with a uniform spectral density but with short-time power level variations. For such an input signal, we would find that the operation would be as predicted by the equations with σ_x^2 (short-time power estimate) considered a function of time.

The time variation of the speech spectrum causes individual tap settings of the echo canceller to become correlated. This effect is not taken into account in the equations. This correlation causes the settling-time to be longer than the equations predict. Because of the variations in the spectral density of speech over time, we find that the echo canceller converges on a frequency selective basis. Figure 15 shows the plot of suppression versus time for the echo canceller with a random noise input and $F[\cdot] = \text{sgn}[\cdot]$. The suppression was measured in 20 adjacent frequency bands approximately 200 Hz wide from 0 to 4000 Hz. The suppression in each band was computed by integrating only over that band using equation (33). Two of these bands are shown in Fig. 15. Note that each one converges at approximately the same rate to a limiting value where it then begins to oscillate. Note also that for each band the limiting suppression is reached very close to the predicted settling time of 0.3 second. The other 18 bands behaved similarly. Similar results were obtained with $F[\cdot] = [\cdot]$.

Figure 16 demonstrates what happens when speech is used instead

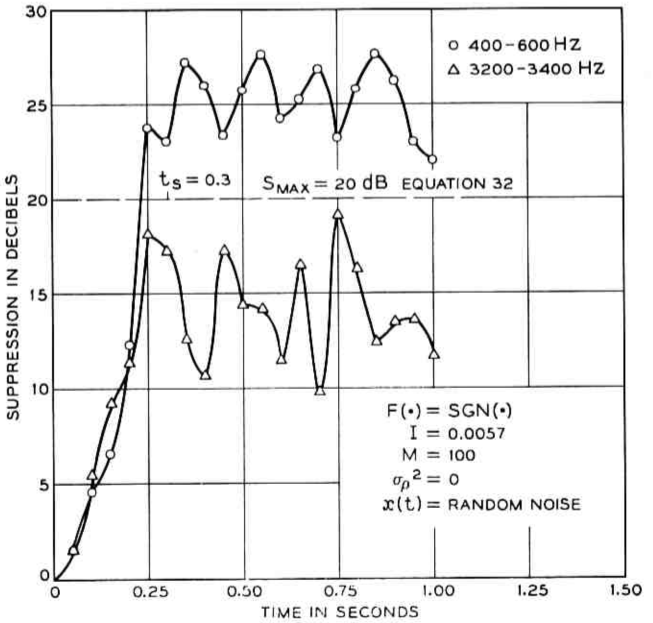


Fig. 15—Suppression as a function of time in the 400- to 500-Hz and 3200- to 3400-Hz frequency bands for a random noise input signal.

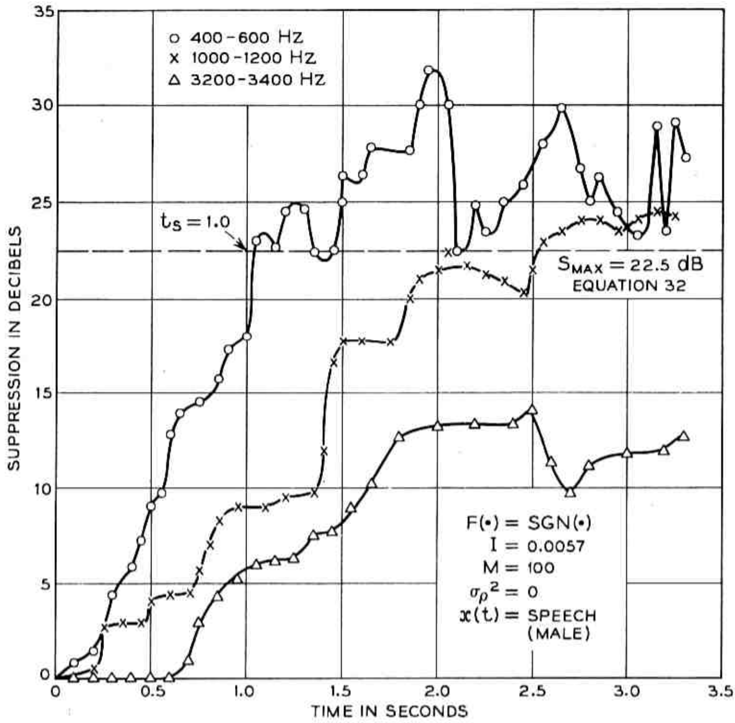


Fig. 16—Suppression as a function of time in the 400- to 600-Hz, 1000- to 1200-Hz and 3200- to 3400-Hz bands with a speech input signal.

of random noise. The designated value of settling time in this experiment was $t_S = 1$ second, using the long-time average (5 seconds) speech power as σ_x^2 . Note that the suppression increases at different rates. For example, between $t = 0.75$ second and $t = 1.25$ seconds, the suppression in the 400- to 600-Hz band increases 9 dB while the suppression in the other 2 bands increases 4 dB at most. Between $t = 1.25$ seconds and $t = 1.5$ seconds, however, the rate of convergence becomes most rapid in the 1000–1200 Hz band. This frequency selective convergence is undoubtedly due to the variation in the spectral distribution of speech power. One result of this is a longer overall settling time based on our measure of suppression. The experiment indicates that although the overall settling time may be longer, the echo canceller converges in some frequency bands more rapidly than average. The bands where this speedy convergence takes place are those where the speech power is greatest at a given time. Because of this, we believe that the perceived settling time would be shorter than is indicated by our measure of suppression, equation (33).* Over a long time (several seconds), this selective convergence results in a fit almost equivalent to that of the random noise across the bandwidth of the speech input. We find that the maximum suppression $S_{\max}$ achieved with a speech input is very nearly equal to that given by the equations when the long-time average speech power is used for σ_x^2.

Figure 17 shows some typical results which were obtained for a simulation with a speech input. Curve (A) shows the results of simulations where $S(nT)$ is calculated every 50 ms using equation (33). Curve (B) shows the performance as predicted using equation (25) and a long-time average (5 seconds) speech power for σ_x^2. The settling time was 2.5 seconds. Curve (C) shows the resulting prediction when

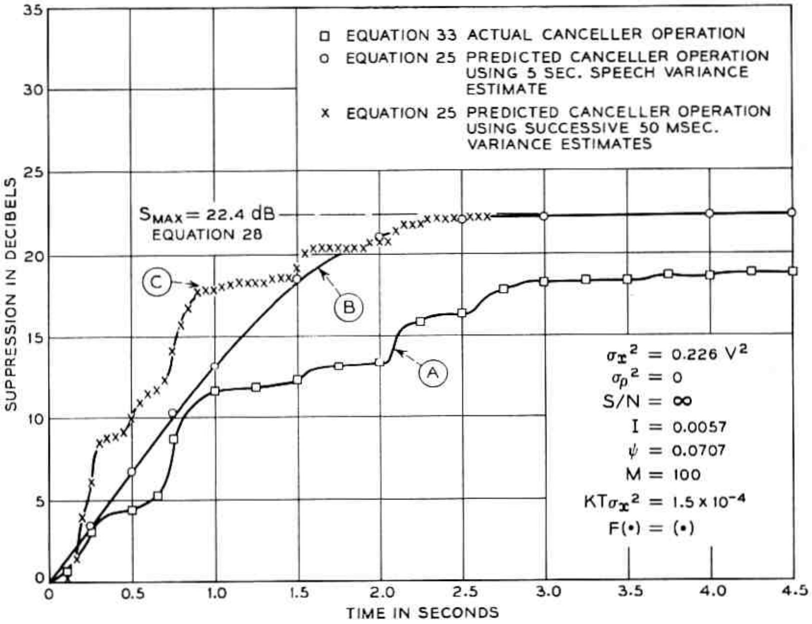


Fig. 17—Comparison ot the results of a computer simulation of an echo canceller for a speech input with those predicted by the equations.

[*] A need for subjective tests which relate suppression (Equation 44) to perceived suppression exists.

σ_x^2 is replaced by a function of time $\sigma_x^2(\zeta)$, which in this case is the successive short-time (50 ms) average speech power. Note that Curves (A) and (C) are almost identical in shape. However, (C) settles more rapidly to its final value as expected. As explained, the nonuniformity of the speech spectrum causes the longer settling time. Had the simulation been plotted for time longer than 4.5 seconds, we would observe that (A) would converge to its limiting value near $S_{\max} = 22.4$ dB.

Another typical case is shown in Fig. 18. A hard limiter was used in the feedback loop of the canceller. The same segment of speech was used here as used in Fig. 17. The value of K was chosen to give a settling time for random noise of 1 second. Note that the echo canceller converged to within 1 dB of $S_{\max}$ in 3 to 4 seconds.

The effects of high noise levels are shown in Fig. 19. With the S/N ratio computed to be -10 dB, we see that the echo canceller converged in approximately 1.5 seconds to $S_{\max} = 11$ dB, where the suppression then tended to vary around this value. This demonstrates that the canceller converges to $S_{\max}$ as predicted in the presence of high levels of noise. Such a strong noise simulates the effect of double-talking. Had the

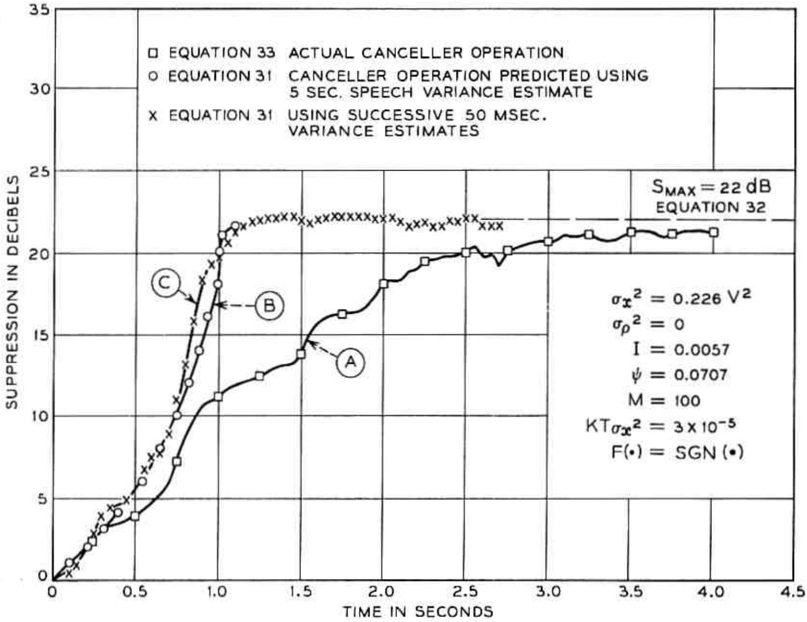


Fig. 18—The same as Fig. 17 with $KT\sigma_x^2 = 3.5 \times 10^{-5}$.

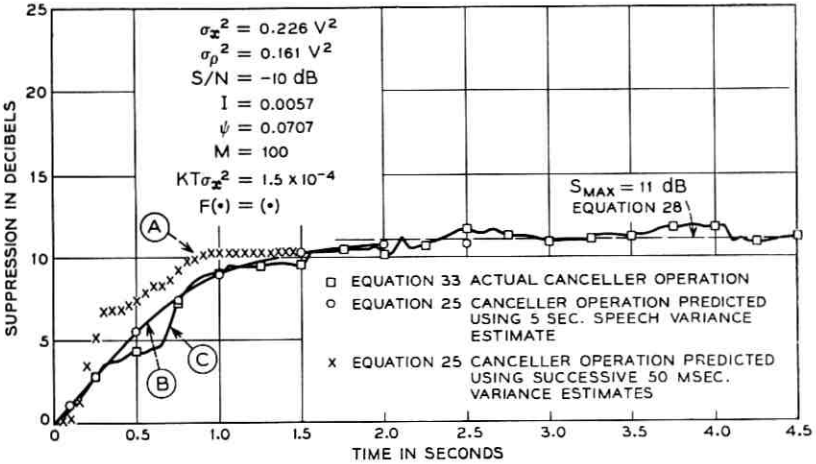


Fig. 19—The same as Fig. 17 with S/N = −10 dB.

noise been introduced after the canceller had converged to some higher $S_{\max}$, the canceller would begin to diverge to the limiting value of $S = 11$ dB. The rate of divergence with speech double-talking is slower than that with random noise. Thus we would find that the effect of double-talking as predicted by the equations would be more severe than it actually is. Also, such an interfering signal causes frequency selective divergence of the echo canceller's transfer function. The divergence is greatest where the interfering signal spectrum is largest. This is not necessarily where the input signal spectrum is largest.

In summary, we see that a long time estimate of speech power can be used in the equations to give a good estimate of the limiting value of suppression $S_{\max}$. Also we see that the canceller performs generally as predicted with a speech input, and in the presence of strong interfering noise. In general, the settling time of the canceller is longer than that predicted by the equations. The settling time may be reduced by increasing K but this must be weighed against the resulting decrease in $S_{\max}$. Also if K is made too large convergence may not take place at all.

IV. SUMMARY

We have described the performance of an adaptive echo canceller operating in a linear, time-invariant, noisy environment. Both digital and analog implementations were considered. In both cases the echo cancellers were assumed to consist of a set of M filters having orthonormal impulse responses which were selected as the first M member of a complete basis set. A weighted sum of the filter outputs approximates the echo signal. The approximation is substracted from the real echo and the difference signal is used to continually improve the approximation so that the cancelled echo power tends toward a minimum. We have used the mean-squared value of the difference between the transfer functions of the echo canceller and the echo path uniformly weighted over the bandwidth of the input signal as a measure of suppression. This measure is not necessarily equivalent to the subjective echo loss (apparent loss perceived by listeners) other than on a relative basis. Sets of equations were derived giving maximum achievable suppression and settling time of the echo canceller. We have shown that despite certain simplifying assumptions made in their derivations, the equations accurately describe the performance for a random noise input.

Families of curves—Figs. 5 through 7, 9 through 11, 13, and 14 show maximum suppression and average settling time for a range of incompleteness factors I, S/N, and a factor related to the input power. The results of simulations are shown to be in close agreement with the predictions.

For a speech input we have found that the equations for maximum suppression can be used to predict performance. The long-time (several seconds) average speech power is used in the equations. The short-time variability of speech power and spectral variations of the speech signal cause the settling time of the echo canceller to be longer than that given by the equations. We have found that during convergence a speech input causes the transfer function of an echo canceller to converge on a frequency selective basis—the fit being best where the power spectrum of the input is greatest. We find that, given enough time, the transfer function of the echo canceller converges to essentially a uniform fit of the echo path transfer function over the bandwidth of the input signal. We have also found that an interfering speech signal (such as exists during double-talking) will cause the echo canceller to diverge on a frequency selective basis. The rate of divergence with speech interference is less than that for the random noise interference.

Before concluding, two final points should be reemphasized. All the previous analysis is only valid when the environment is linear and time-invariant. At present, we suspect that certain systems (compandored systems for example) exhibit non-negligible nonlinearities.

For such systems, the previous analysis may not suffice depending on the type and magnitude of the nonlinearities. In any case it becomes extremely dangerous for compandor type nonlinearities to attempt to relate the performance of an echo canceller to a speech input from the white noise equations given previously.

Also, it should be stressed, that the measure of performances we have defined are objective in nature. These measures are not necessarily equivalent to the subjective echo reduction which a listener will preceive. A need exists to relate the objective and subjective.

REFERENCES

1. Unrue, J. E., "Controlling Echo in the Bell System," Bell Labs Record, *47,* No. 7 (August 1969), pp. 233–238.
2. Emling, J. W., and Mitchell, D., "Effects of Time Delay and Echos on Telephone Conversations," B.S.T.J., *42,* No. 6 (November 1963), pp. 2869–2891.
3. Klemmer, E. T., "Subjective Evaluation of Transmission Delay in Telephone Conversation," B.S.T.J., *46,* No. 6 (July–August 1967), pp. 1141–1147.
4. Helder, G. K., "Customer Evaluation of Telephone Circuits with Delay," B.S.T.J., *45,* No. 7 (September 1966), pp. 1157–1191.
5. Sondhi, M. M., "An Adaptive Echo Canceller," B.S.T.J., *46,* No. 3 (March 1967), pp. 497–511.
6. Sondhi, M. M., and Presti, A. J., "A Self Adaptive Echo Canceller," B.S.T.J., *45,* No. 10 (December 1966), pp. 1851–1854.
7. Becker, F. K., and Rudin, H. R., "Application of Automatic Transversal Filters to the Problem of Echo Suppression," B.S.T.J., *45,* No. 10 (December 1966), pp. 1847–1850.
8. Thomas, E. J., "An Adaptive Echo Canceller in a Non-Ideal Environment," unpublished work.
9. Thomas, E. J., "Some Considerations of the Application of the Volterra Representation of Nonlinear Networks to Adaptive Echo Cancellers and Control Systems," unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

Number and Duration of Fades at 6 and 4 GHz

By ARVIDS VIGANTS

(Manuscript received November 9, 1970)

We present in this paper experimental and theoretical results on the number of fades, average durations of fades, and the probability distributions of fade durations for line-of-sight microwave transmission in the 6- and 4-GHz bands. Both fading of single signals and simultaneous fading of pairs of signals within each of the two bands are treated. The experimental results are based on data obtained in 1966 in Ohio. Mathematical description of the number and average duration of fades is based on a theory in which pairs of signals are treated as correlated random variables that are jointly Rayleigh distributed. The durations of deep fades of single signals tend to be lognormally distributed. A probability distribution of the duration of simultaneous fades, which agrees with experimental data, is obtained from the lognormal distribution using a heuristic model.

I. INTRODUCTION

Microwave transmission on line-of-sight radio-relay links is affected by the lower atmosphere. When atmospheric conditions permit multipath propagation, the output from a receiving antenna can be practically zero for seconds at a time. Such deep fades are rare events. Still, they are sufficiently numerous to cause problems in high-performance communication systems.

There is a certainty to fading that other causes of outages and performance degradation do not possess. For example, catastrophic equipment failure may or may not occur during the projected life of the equipment on a communications link. On the other hand, come summer and fall, one can state with some assurance that fading will occur on certain links in a particular microwave radio-relay system.

The number of fades and the durations of the fades have a direct bearing on system performance. Previous investigations of frequency diversity at Bell Laboratories[1,2] and reports on extensive British[3] and Japanese[4] experiments dwell mostly on fade-depth distributions. The available information on fade duration distributions is rudimentary.[5–7]

As a part of a current and continuing effort, both experimental and theoretical, to describe the fundamental properties of line-of-sight microwave channels,[2,8,9] we will present results on the number of fades, the average durations of fades, and the probability distributions of fade durations. Both fading of single signals and simultaneous fading of pairs of signals within a frequency band will be treated. Theoretical description of fading being far from complete, we will put major emphasis on experimental data.

Deep fades rarely occur simultaneously at two frequencies when the frequency separation becomes larger than a few tens of MHz. We will show that this experimental result can be described in terms of a model in which the signals at the two frequencies are treated as correlated random variables that are jointly Rayleigh distributed. We will also show that the observed average fade durations follow from this model. The advantage of the Rayleigh model is that it contains only a small number of parameters.

Durations of deep fades tend to be lognormally distributed. When the durations are normalized to their means, the distribution becomes independent of fade depth. We will show that a heuristic model can be used to describe how the durations of simultaneous fades of two signals are related to fade durations of single signals. The model permits transformation of the lognormal distribution into a distribution for the durations of simultaneous fades.

The final section of this work contains a comparison of the effectiveness of frequency-diversity and space-diversity reception. The results summarized in this section may be of particular interest to readers who need numerical values for diversity, reliability, or interference calculations.

II. FORM OF THE EXPERIMENTAL DATA

The results presented here are based on experimental data obtained at West Unity, Ohio.[2] Briefly, the basic data consist of measurements of received power for signals at various frequencies on a 28.5-mile path. The transmitted power for each signal, which was angle modulated, was constant. The center frequencies of the signals in the 6-GHz and 4-GHz bands are listed in Table I. The received power for each signal was sampled five times a second, converted to a decibel scale,

TABLE I—LIST OF FREQUENCIES

Frequencies			
6-GHz Band		4-GHz Band	
i	GHz	i	GHz
1	5.9452	1	3.750
2	6.0045	2	3.770
3	6.0342	3	3.830
4	6.0638	4	3.850
5	6.1231	5	3.910
6	6.1528	6	4.070
		7	4.170

and recorded in digital form for subsequent computer processing (in the absence of fading the recording rate was less than the sampling rate). The data were obtained during a 68-day period in 1966 (July 22 to September 28). The time covered by the data is 5.26×10^6 seconds. We refer to the 5.26×10^6 seconds as the test period.

During computer processing, the received power for each signal was normalized to its value in the absence of fading. The normalized dB values are denoted by 20 log R_i, where the subscript i identifies a frequency in Table I (the data in the two frequency bands will be discussed separately; there should be no confusion about which of the two bands a subscript refers to). As a consequence of the normalization, the voltage envelopes R_i are unity in the absence of fading.

To investigate simultaneous fading of signals at two frequencies, an envelope consisting of the larger of the two envelopes,

$$R_{ij} = \max (R_i \, , R_j) \tag{1}$$

was constructed in the computer during the processing of the data. Fifteen such envelopes were constructed in each of the two bands. These are listed in Tables II and III. The frequency separations Δf in Table II have been rounded to the nearest multiple of 30 MHz. The parameter q, discussed later, will be used in the description of simultaneous fading.

An idealized picture of fading is shown in Fig. 1. A signal is said to be in a fade of depth 20 log L when its envelope becomes less than L. We refer to fades of R_{ij} as simultaneous fades. We limit our discussion to fades deeper than -20 dB; that is, to values of L that are less than a tenth.

TABLE II—LIST OF FREQUENCY PAIRS IN THE 6-GHz BAND

Δf in MHz	ij	q	Figure
30	23	0.001	6
30	34	0.001	7
30	56	0.001	8
60	12	0.002	9
60	24	0.002	10
60	45	0.002	11
90	13	0.004	12
90	35	0.004	13
90	46	0.004	14
120	14	0.005	15
120	25	0.005	16
120	36	0.005	17
150	26	0.007	18
180	15	0.007	19
210	16	0.010	20

TABLE III—LIST OF FREQUENCY PAIRS IN THE 4-GHz BAND

Δf in MHz	ij	q	Figure
20	12	0.002	21
20	34	0.002	22
60	23	0.007	23
60	45	0.007	24
80	24	0.010	25
80	35	0.010	26
100	14	0.013	27
160	15	0.020	28
160	56	0.020	29
220	46	0.030	30
260	57	0.030	31
300	26	0.035	32
340	37	0.035	33
400	27	0.050	34
420	17	0.050	35

III. NUMBER AND AVERAGE DURATION OF FADES OF SINGLE SIGNALS

The number of fades of depth 20 log L dB is equal to, by definition, the number of times an envelope crosses L in an upward direction (see Fig. 1). The data on this for the six signals in the 6-GHz band and the seven signals in the 4-GHz band are summarized in Figs. 2 and 3 respectively. The data scatter, but the overall impression is that the number of fades is proportional to L. Least-squares fitting of lines

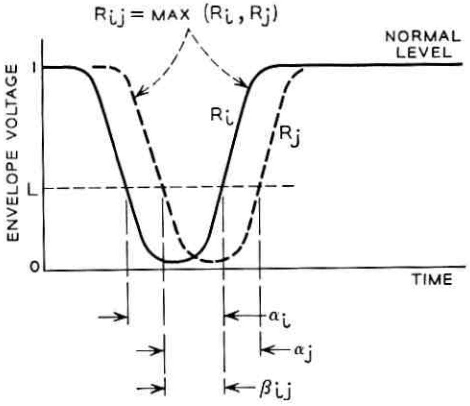


Fig. 1—Definition of terms.

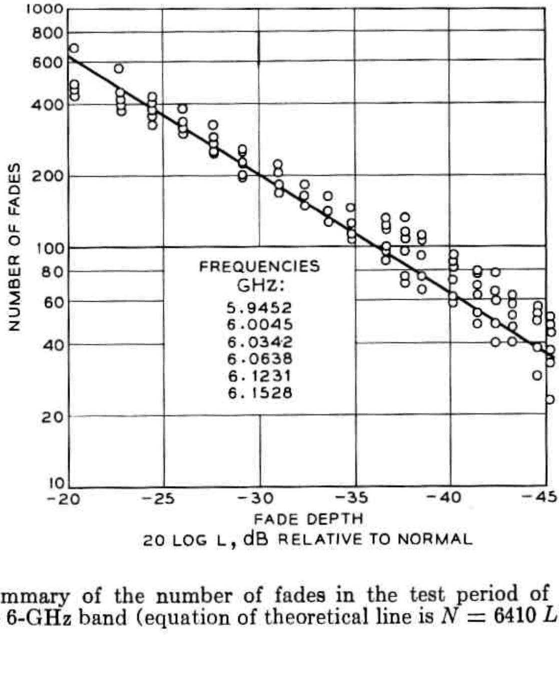


Fig. 2—Summary of the number of fades in the test period of the six single signals in the 6-GHz band (equation of theoretical line is $N = 6410\ L$).

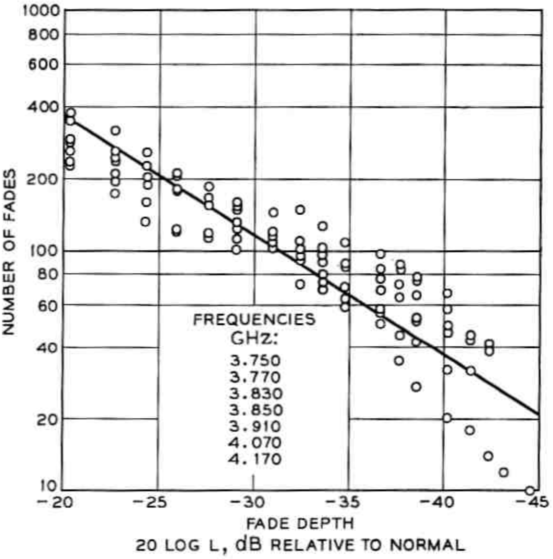


Fig. 3—Summary of the number of fades in the test period of the seven single signals in the 4-GHz band (equation of theoretical line is $N = 3670\ L$).

with this dependence on L to the data gives the following for the number of fades in the test period

$$N = 6410\ L, \text{ in the 6-GHz band,} \tag{2}$$

$$N = 3670\ L, \text{ in the 4-GHz band.} \tag{3}$$

As an example, we note that N in the 6-GHz band averages to roughly one -40 dB fade per day during the test period.

The average durations of fades were obtained by taking the total time spent in fades of a given depth[2] and dividing this by the number of fades of that depth. The resulting data are shown in Fig. 4 for the six signals in the 6-GHz band and in Fig. 5 for the seven signals in the 4-GHz band. The average durations are roughly proportional to L. Least-squares lines with this dependence on L fitted to the data are

$$\langle \alpha \rangle = 490\ L \text{ seconds, in the 6-GHz band,} \tag{4}$$

$$\langle \alpha \rangle = 408\ L \text{ seconds, in the 4-GHz band,} \tag{5}$$

where α denotes fade duration (see Fig. 1), and where the brackets denote averages. The fit of equation (4) to the points in Fig. 4 is better than that of equation (5) to the points in Fig. 5. The points in Fig. 4 are based on a larger number of observations than those in

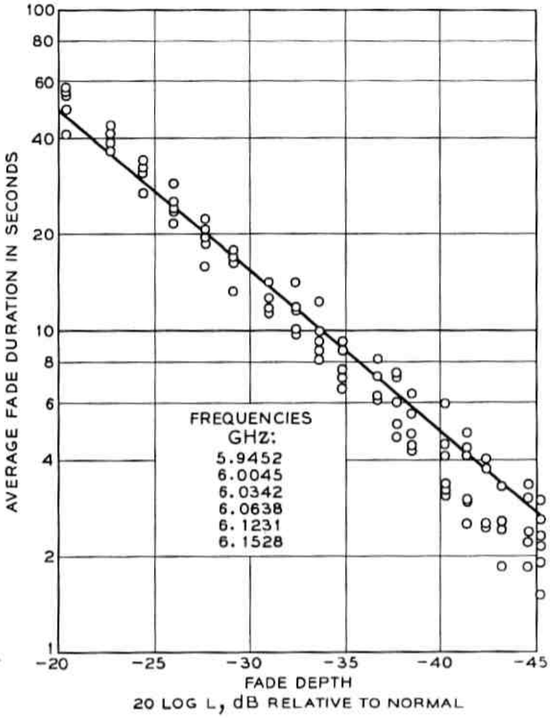


Fig. 4—Summary of the average fade durations of the six single signals in the 6-GHz band (equation of theoretical line is $\langle\alpha\rangle = 490\ L$).

Fig. 5. It is therefore possible that in Fig. 5 we are beginning to see the effects of small sample size. We will assume that this is indeed so and put greater emphasis on the 6-GHz data in the subsequent analysis.

The data show that average fade durations in the -40 dB range are of the order of seconds. This has direct bearing on system performance calculations, where it has been assumed, at times, that deep fades have durations of the order of minutes. (Section 5.2.1 in Ref. 10.)

The total time spent in fades is the product $N\ \langle\alpha\rangle$, which is proportional to L^2. The observed behavior of fade-depth distributions for deep fades is indeed this,[2] which suggests comparison of the experimental results to theoretical results for Rayleigh distributed variables. The theoretical expressions for the number of fades and the average fade durations are, in the case of deep fades,[9,11]

$$N \approx (rT_oc)\ L,\ L < 0.1\ \ ; \tag{6}$$

$$\langle\alpha\rangle \approx c^{-1}L, \quad L < 0.1; \tag{7}$$

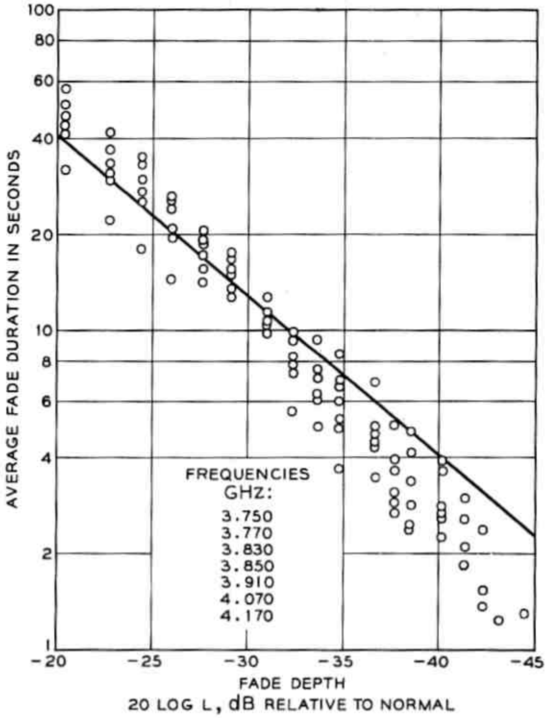
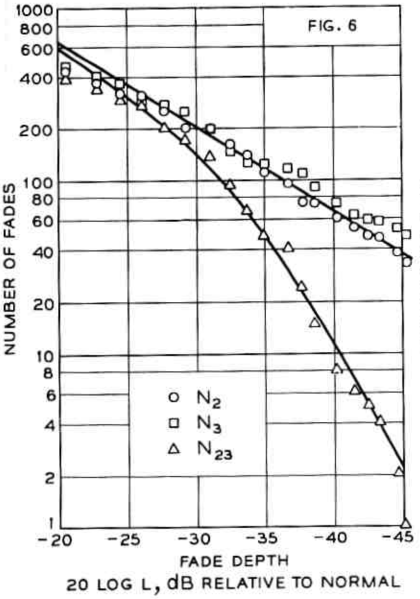


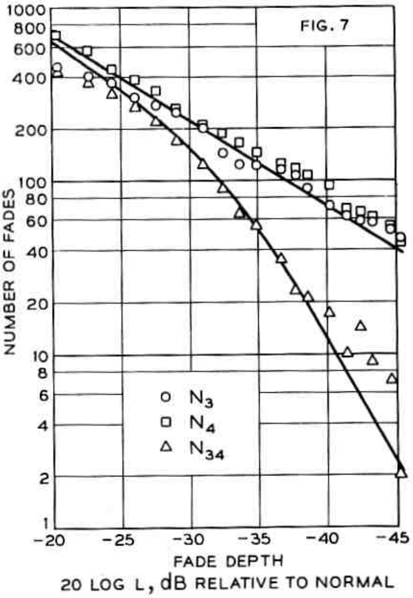
Fig. 5—Summary of the average fade durations of the seven single signals in the 4-GHz band (equation of theoretical line is $\langle \alpha \rangle = 408\ L$).

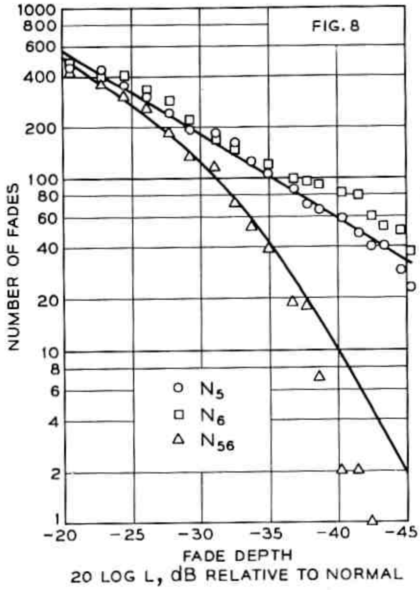
where the dimensionless factor r is proportional to the amount of time during which atmospheric conditions allow multipath propagation; T_0 is the test period (5.26×10^6 seconds in our case); the quantity c, which has the dimensions of inverse time, is proportional to the spectral width of the fluctuations of R_i. The theoretical equations (6) and (7) clarify the meaning of the coefficients in equations (2) through (5). Numerical values for the parameters r and c can be obtained readily from a comparison of the coefficients. However, r varies with path length and other parameters.[12] We will not discuss this here, because the variability of fading from path to path is a topic in itself.

IV. NUMBER OF SIMULTANEOUS FADES

Simultaneous fades are fades of R_{ij} (see Fig. 1). The experimental results on the number of simultaneous fades are shown in Figs. 6 to 20







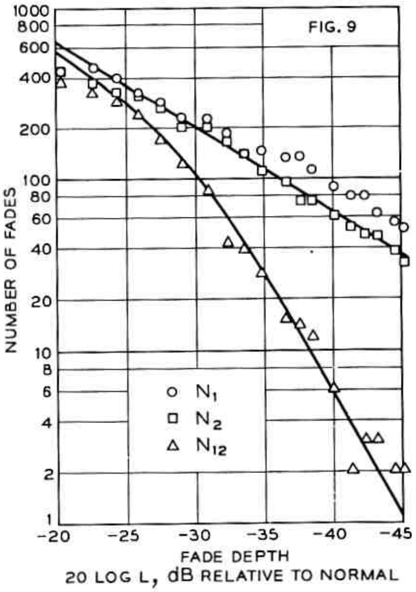
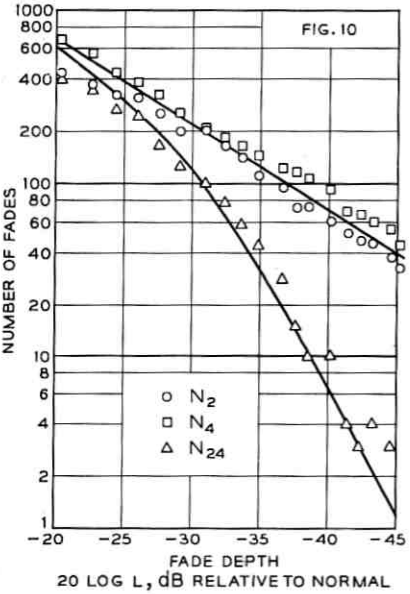
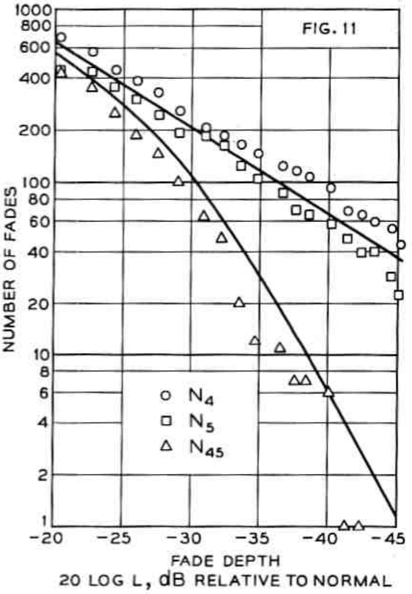
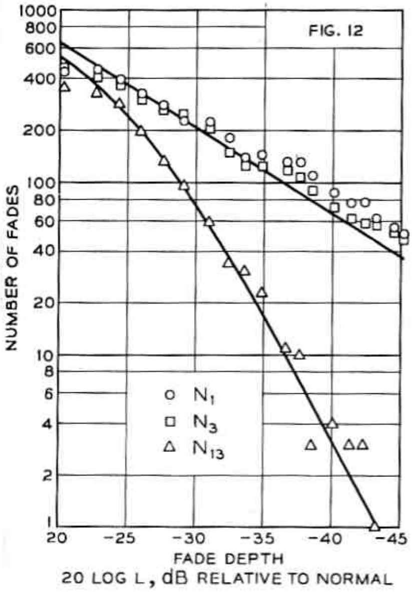


Fig. 6—Number of fades in test period (6-GHz band, $\Delta f = 30$ MHz).
Fig. 7—Number of fades in test period (6-GHz band, $\Delta f = 30$ MHz).
Fig. 8—Number of fades in test period (6-GHz band, $\Delta f = 30$ MHz).
Fig. 9—Number of fades in test period (6-GHz band, $\Delta f = 60$ MHz).







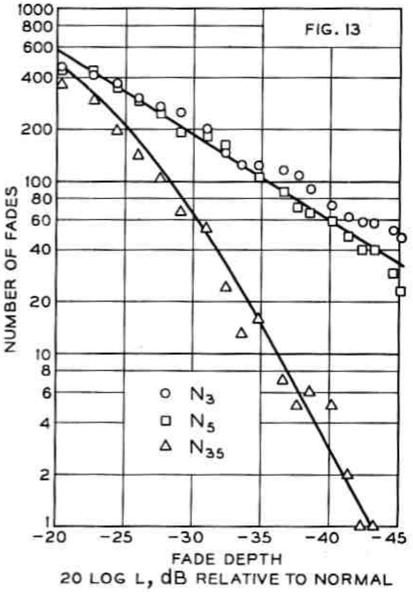
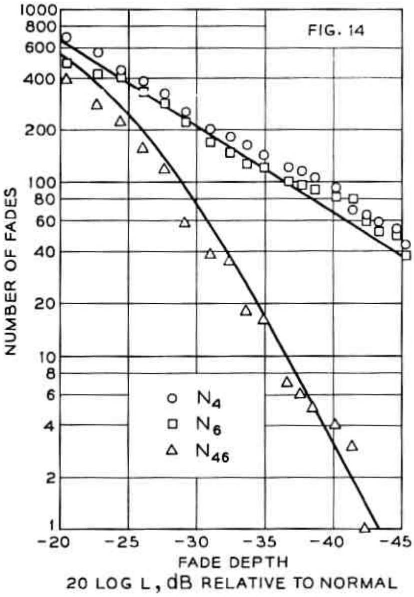
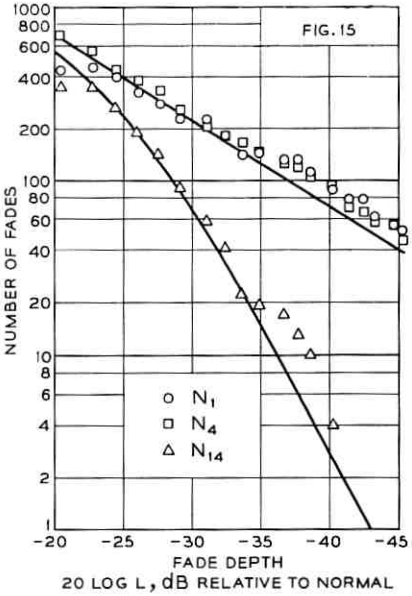
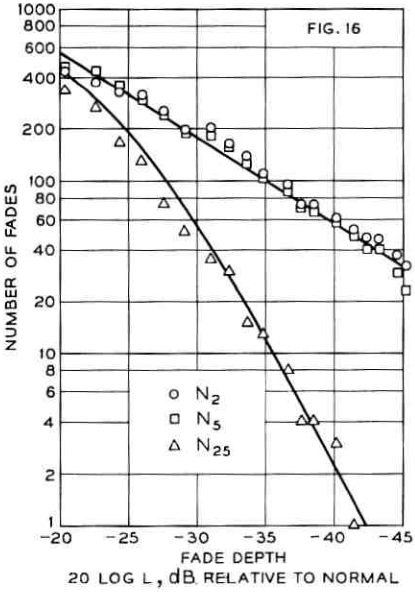


Fig. 10—Number of fades in test period (6-GHz band, $\Delta f = 60$ MHz).
Fig. 11—Number of fades in test period (6-GHz band, $\Delta f = 60$ MHz).
Fig. 12—Number of fades in test period (6-GHz band, $\Delta f = 90$ MHz).
Fig. 13—Number of fades in test period (6-GHz band, $\Delta f = 90$ MHz).







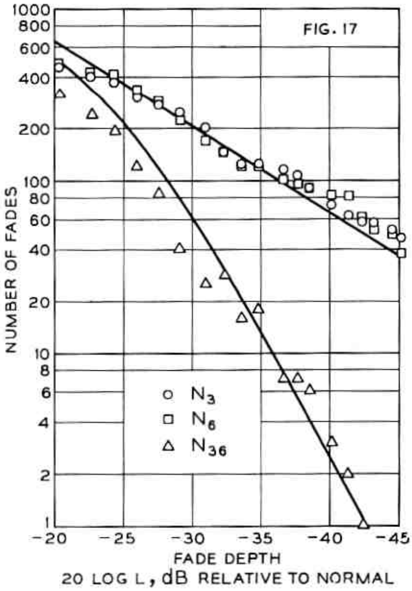


Fig. 14—Number of fades in test period (6-GHz band, $\Delta f = 90$ MHz).
Fig. 15—Number of fades in test period (6-GHz band, $\Delta f = 120$ MHz).
Fig. 16—Number of fades in test period (6-GHz band, $\Delta f = 120$ MHz).
Fig. 17—Number of fades in test period (6-GHz band, $\Delta f = 120$ MHz).

for the 6-GHz band and Figs. 21 to 35 for the 4-GHz band. The figures are arranged in order of increasing frequency separation in each of the two bands. The number of fades of R_i in the test period is denoted by N_i. The number of simultaneous fades in the test period is denoted by N_{ij}. The points on the drawings are the experimental data; the lines are theoretical and will be discussed shortly.

The data show, for example, that for a one percent frequency separation in the 6-GHz band (60 MHz, Figs. 9 to 11) N_{ij} averages to about one −40 dB fade per ten days. The ratio of N_i to N_{ij} at this fade depth is about ten.

The number of deep simultaneous fades is roughly proportional to L^3, which again suggests comparison to theoretical results for Rayleigh distributed variables. When R_i and R_j are treated as correlated Rayleigh distributed variables, the theoretical expression for the number of deep simultaneous fades is[9]

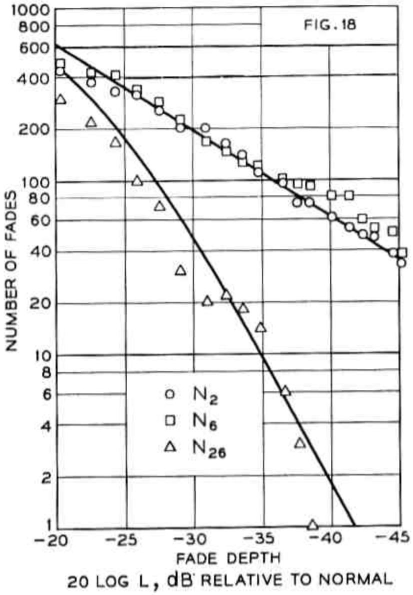
$$N_{ij} \approx (rT_0c)2q^{-1}L^3, \quad L < 0.1, \quad q^{-1}L^2 < 0.1 \tag{8}$$

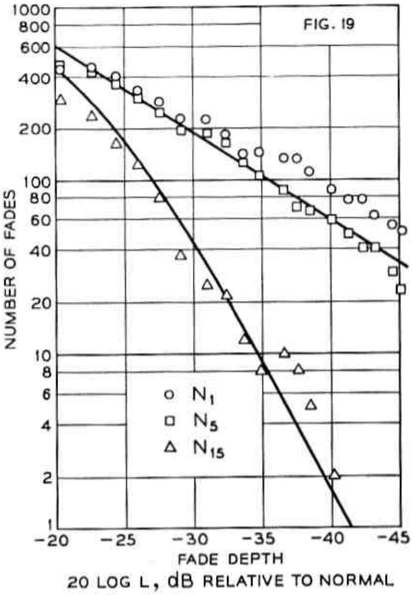
where the parameter q is a function of the frequency separation of the signals. Values of q for the various signal pairs are listed in Tables II and III.*

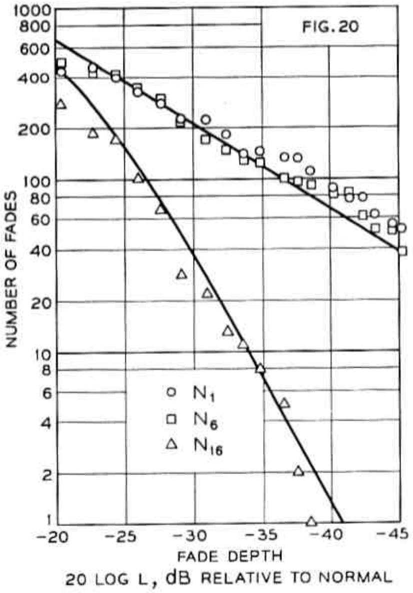
The comparison of the theoretical expressions to the experimental data was carried out as follows. Values of q for the various signal pairs (R_i, R_j) were obtained from Table II or Table III. On each drawing, equation (6) was used to describe the average of N_i and N_j, and equation (8) was used to describe N_{ij}. Actually, equation (8) describes only the part of N_{ij} appearing as a straight line on a drawing; the curved portions of N_{ij} were computed from more complex expressions in previous work.[9] The value of q determines the position of equation (8) relative to equation (6), and curve fitting becomes a matter of fitting two curves with a fixed relative position to the three sets of data on a drawing. The result of a fitting is a value for the product rT_0c. The values of rT_0c vary somewhat from drawing to drawing, reflecting the apparently random variations in N_i from signal to signal. The average values of rT_0c in the two frequency bands appear as coefficients in equations (2) and (3). The overall impression is that the data agree with the theoretically predicted dependence on the fade depth L.

The behavior of the data on the number of fades is typically that of

* These are values of q obtained by Barnett[2] for each signal pair from experimental data on fade depth distributions. Equations describing q are shown in Section X.







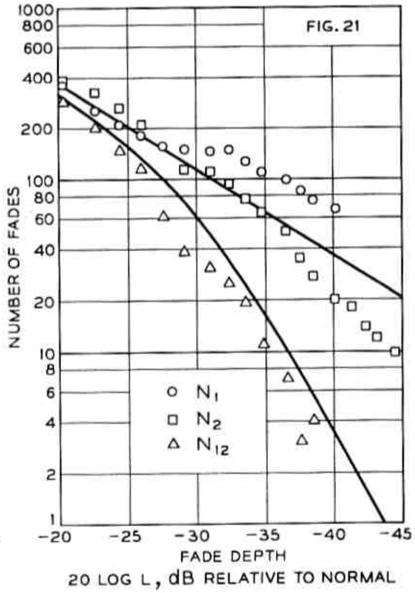
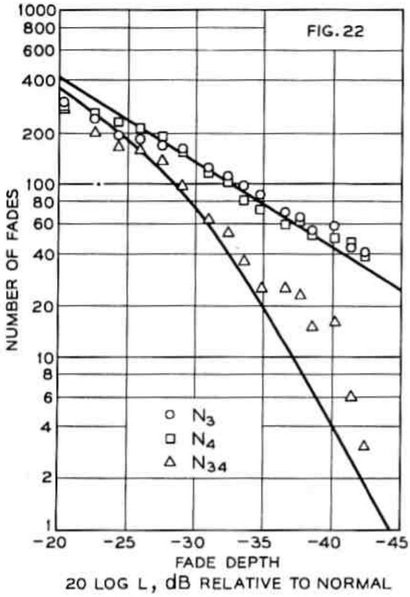
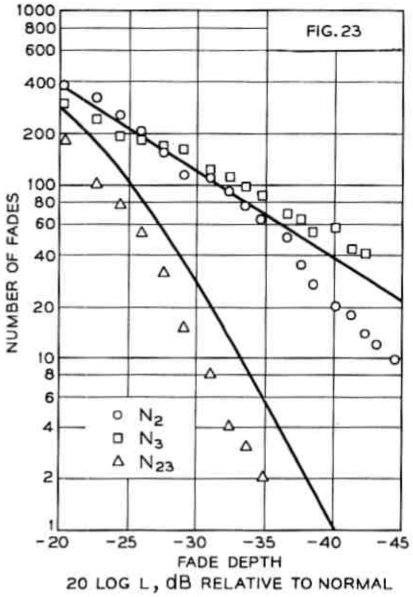
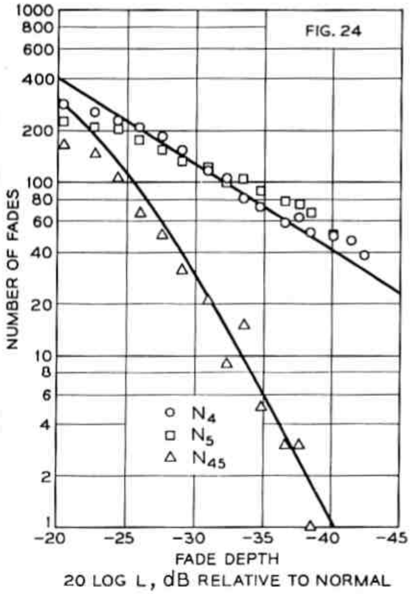


Fig. 18—Number of fades in test period (6-GHz band, $\Delta f = 150$ MHz).
Fig. 19—Number of fades in test period (6-GHz band, $\Delta f = 180$ MHz).
Fig. 20—Number of fades in test period (6-GHz band, $\Delta f = 210$ MHz).
Fig. 21—Number of fades in test period (4-GHz band, $\Delta f = 20$ MHz).







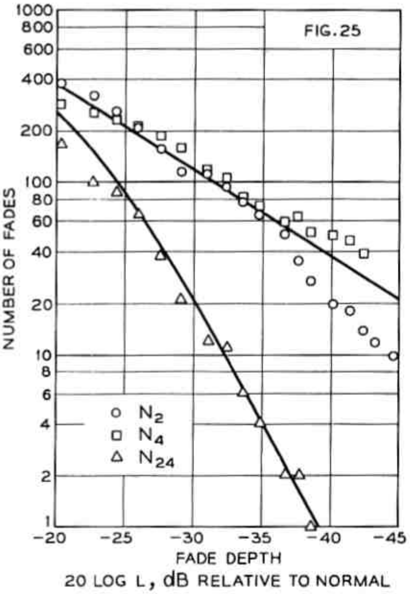


Fig. 22—Number of fades in test period (4-GHz band, $\Delta f = 20$ MHz).
Fig. 23—Number of fades in test period (4-GHz band, $\Delta f = 60$ MHz).
Fig. 24—Number of fades in test period (4-GHz band, $\Delta f = 60$ MHz).
Fig. 25—Number of fades in test period (4-GHz band, $\Delta f = 80$ MHz).

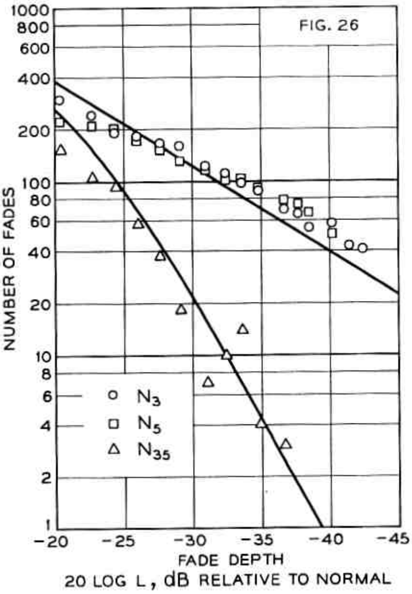


FIG. 26

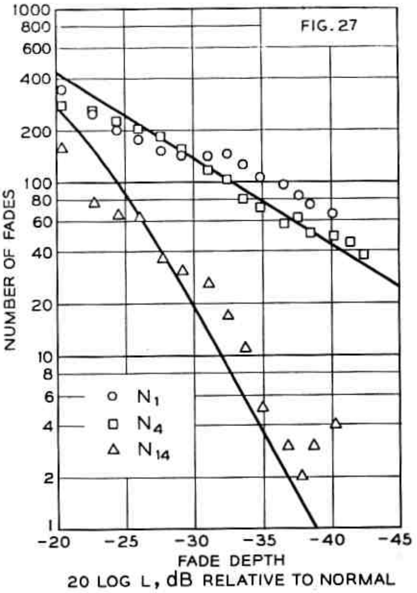


FIG. 27

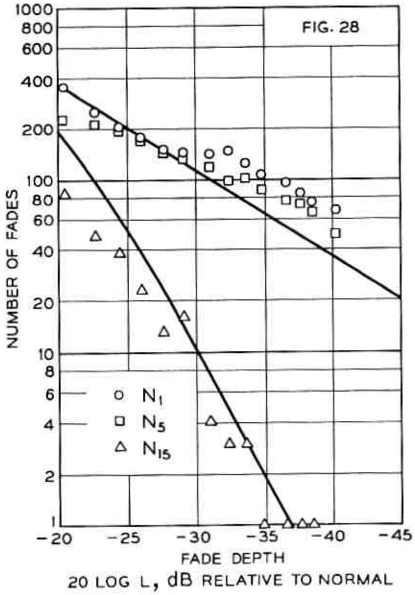


FIG. 28

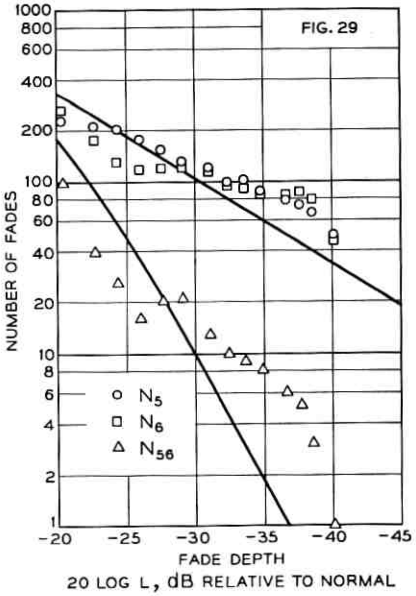
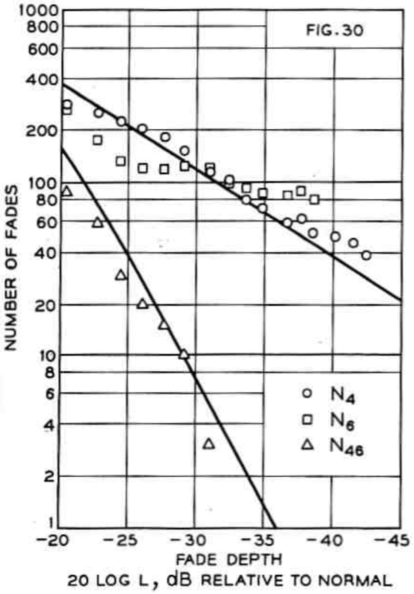
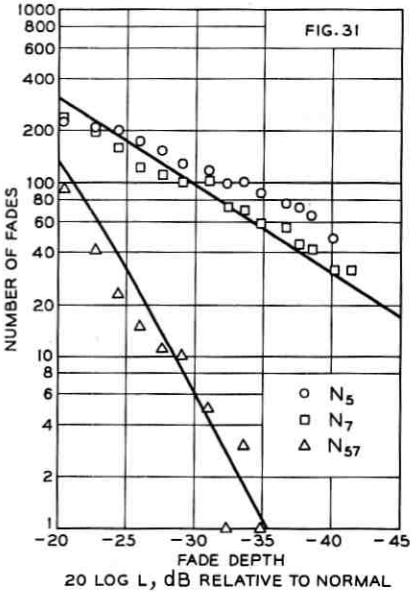
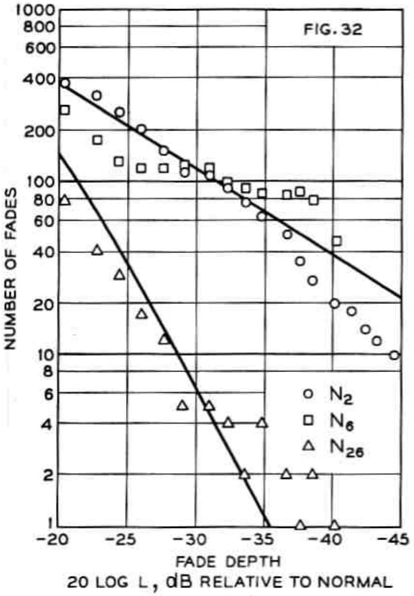


FIG. 29

Fig. 26—Number of fades in test period (4-GHz band, $\Delta f = 80$ MHz).
Fig. 27—Number of fades in test period (4-GHz band, $\Delta f = 100$ MHz).
Fig. 28—Number of fades in test period (4-GHz band, $\Delta f = 160$ MHz).
Fig. 29—Number of fades in test period (4-GHz band, $\Delta f = 160$ MHz).







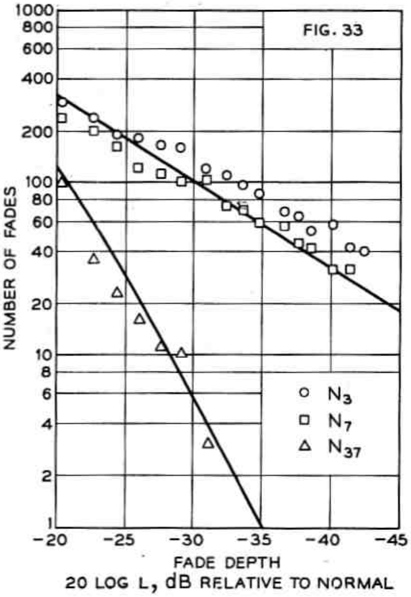


Fig. 30—Number of fades in test period (4-GHz band, $\Delta f = 220$ MHz).
Fig. 31—Number of fades in test period (4-GHz band, $\Delta f = 260$ MHz).
Fig. 32—Number of fades in test period (4-GHz band, $\Delta f = 300$ MHz).
Fig. 33—Number of fades in test period (4-GHz band, $\Delta f = 340$ MHz).

propagation data in general. The processes in the atmosphere causing fading are complex, and this shows up as scatter and deviations of the experimental points from the theoretical lines. Some of this is magnified by the logarithmic scales used in the drawings. The deviations of the experimental points from the theoretical lines in the 4-GHz band are larger than those in the 6-GHz. This may be because the number of observations in the 4-GHz band is smaller, or because the theoretical description becomes less applicable as the path length, measured in wavelengths, becomes shorter. The deviations of the points from the lines around −20 dB reinforce the observation that the theory based on the joint Rayleigh distribution applies only to fades deeper than about −20 dB. A possible explanation is that the physical processes generating deep fades differ from those generating shallow fades.

V. AVERAGE DURATION OF SIMULTANEOUS FADES

The determination of the average durations of simultaneous fades is difficult, because there are too few simultaneous fades, particularly at the larger frequency separations, to obtain meaningful averages. The data for the first entry in Table II are shown in Fig. 36. The curves are theoretical. The top curve is equation (7), with c^{-1} adjusted to describe the average of α_2 and α_3 (the notation is defined in Fig. 1). Given the top curve and the value of q in Table II, the bottom curve describing β_{23} is calculated under the assumption that R_2 and R_3 are jointly Rayleigh distributed.[9] Note from the curves that as the fades become deeper, the average durations of simultaneous fades tend to one-half of those of single signals. The theory predicts this for all frequency separations, provided the fades are sufficiently deep,[9]

$$\langle \beta_{ij} \rangle \approx \tfrac{1}{2} \langle \alpha_i \rangle, \quad L < 0.1, \quad q^{-1}L^2 < 0.1. \tag{9}$$

It is interesting that the same result can also be inferred from a heuristic argument, given in Section VIII.

The data and the theory are in agreement in Fig. 36. In general, the data on the average durations of simultaneous fades show a large amount of scatter, much larger than in the example in Fig. 36, because of the small number of observations.

VI. THE SCALE OF SCATTER IN THE OBSERVATIONS

Experimental data can scatter simply because the number of observations is small. In the case of average fade durations we can estimate the amount of scatter possible, and it is instructive to digress and do so. As an example, the average durations at the highest frequency in the 6-GHz band are shown in Fig. 37 (the points denoted by black squares are at fade depths for which probability distributions of fade durations will be obtained later). Each point on the drawing is the average of a sample of size N_i. This average, the so called sample mean, is a random variable. For example, if the test period contains, by chance, an excess of small α_i, the sample mean will be less than the true average. As N_i increases, the sample mean tends to deviate less from the true average. The standard deviations of the sample means in Fig. 37 are estimated in the Appendix. The straight line in Fig. 37 is the result of a least squares fitting of equation (7) to the data. The curves are at three standard deviations of the sample mean above and below the straight line.

The tolerance band defined by the curves in Fig. 37 is asymmetric, especially so for the deeper fades, because of the logarithmic scale on which the average fade durations are plotted. The band is quite wide in the -40 dB region, and the conclusion is that the deviations of the points from the line are well within those predicted possible by elementary sampling theory.

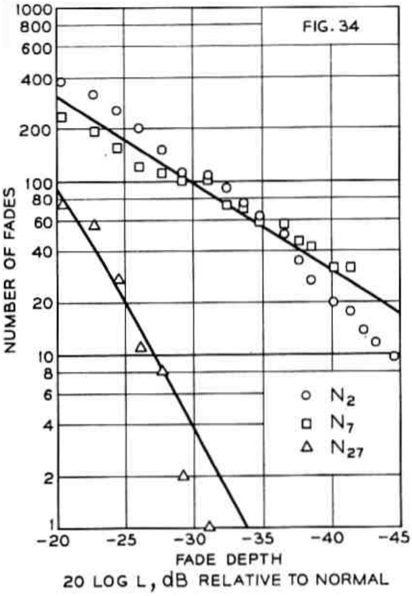
A corresponding tolerance band for the 4-GHz data would be wider because of the smaller number of fades observed. It would be much wider for the simultaneous fades because their number is much smaller. Because of the potential for large scatter in the 4-GHz data, we will discuss, in the next section, duration distributions for the 6-GHz band only.

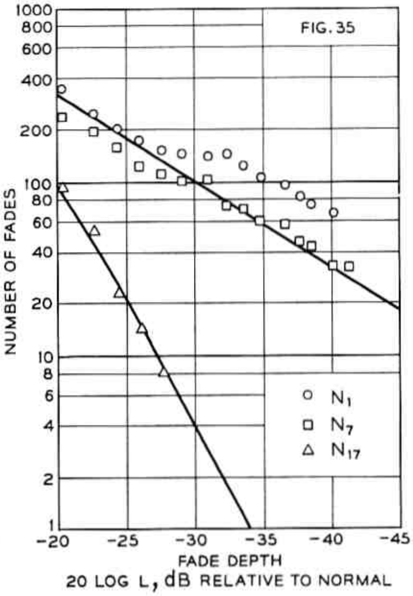
VII. OBSERVED DISTRIBUTIONS OF FADE DURATIONS

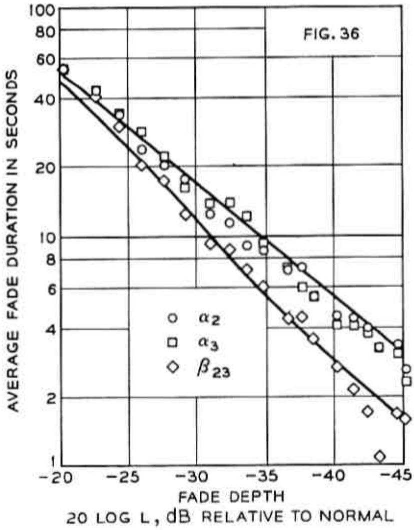
Estimates of the probability that a fade of depth 20 log L dB (see Fig. 1) lasts longer than t seconds were obtained by taking the number of fades longer than t seconds and dividing this by the total number of fades of depth 20 log L dB. The results of this for the highest frequency in the 6-GHz band are shown in Fig. 38 for five values of fade depth (the average durations at these depths are denoted by the black squares in Fig. 37). The probability scale is normal and the duration scale is logarithmic.

The behavior of the probabilities shows a more readily discernible pattern if the durations are measured not in seconds but in units of average fade durations; that is, if we look at the probability that x_i is larger than a number u, where

$$x_i = \alpha_i / \langle \alpha_i \rangle. \tag{10}$$







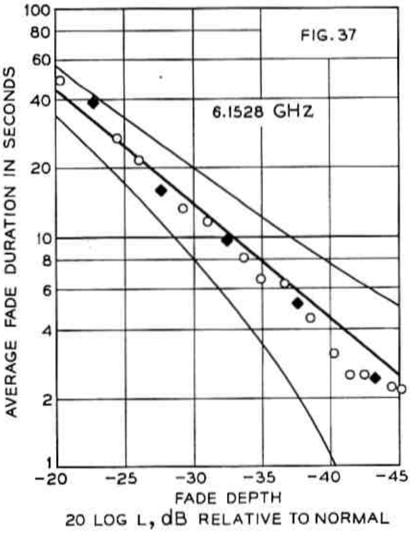


Fig. 34—Number of fades in test period (4-GHz band, $\Delta f = 400$ MHz).

Fig. 35—Number of fades in test period (4-GHz band, $\Delta f = 420$ MHz).

Fig. 36—Average durations of simultaneous fades of signals at 6.0045 and 6.0342 GHz (the lines are theoretical)

Fig. 37—Average durations of fades of the signal at 6.1528 GHz, with superimposed control curves at three standard deviations above and below the straight line, estimated from elementary sampling theory.

When this is done, using values of $\langle \alpha_i \rangle$ denoted by solid squares in Fig. 37, the fairly tightly grouped set of points in Fig. 39 is obtained. From the behavior of the points for the various fade depths one can argue that the data approach a lognormal distribution, denoted by the straight line, as fades become deeper and longer. When points from the other five frequencies are superimposed on this, Fig. 40 is obtained. Because we have pooled data from various frequencies, we have dropped the subscripts on the variables.

The pattern formed by the points in Fig. 40 suggests that the distribution of durations longer than average can be described by a straight line (the points for the long durations begin to scatter because the number of observations is small). The equation of the line shown is

$$\Pr(x > u) = \tfrac{1}{2} \operatorname{erfc}[(\ln u - \mu)/\sqrt{2}\,\sigma] \tag{11}$$

with

$$\mu = -0.673, \quad \sigma = 1.27 \tag{12}$$

where ln denotes the natural logarithm, and erfc denotes the complementary error function. The lognormal distribution (11) specifies that one percent of the durations are longer than ten times average, and that thirty percent of the durations are longer than average.

The results for the durations of the simultaneous fades also fall into a pattern when the durations are normalized to their averages. Letting

$$y_{ij} = \beta_{ij}/\langle \beta_{ij} \rangle \tag{13}$$

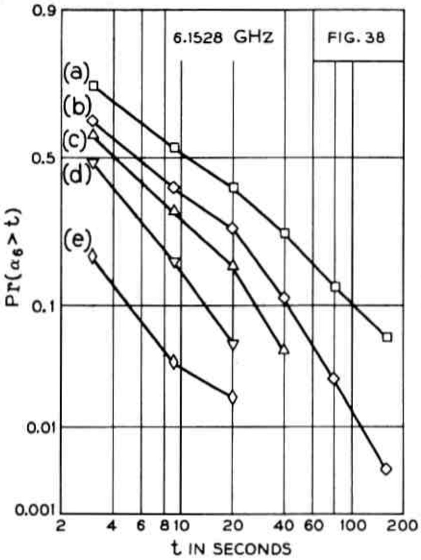
and pooling the data for all fifteen entries in Table II, we obtain Fig. 41. Again, subscripts are dropped for the pooled data.

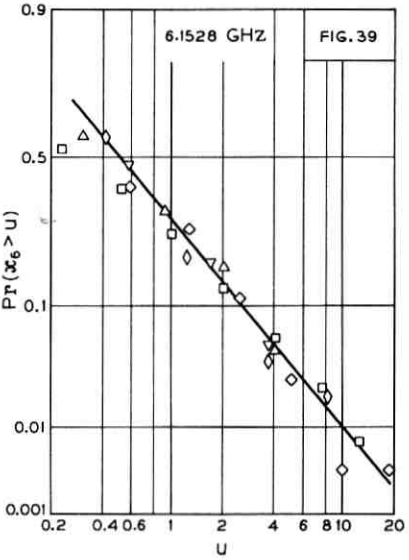
The behavior of the data for the simultaneous fades in Fig. 41 is very similar to that of the data in Fig. 40. The line from Fig. 40 is shown dotted on Fig. 41 (the solid line on Fig. 41 will be discussed shortly).

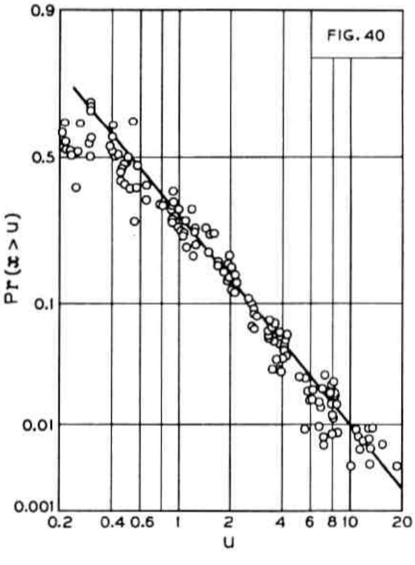
Comparison to previous data on duration distributions[5–7] is difficult, beyond noting the similar lognormal behavior, since the previous data do not cover fades in the −30 to −40 dB range of interest to us. Furthermore, in one of the cases the emphasis is on a grazing path[5, 6] and in the other on overwater paths,[7] whereas our interest is in "standard" overland paths.

VIII. THEORETICAL DESCRIPTION OF DURATION DISTRIBUTIONS

The probability distribution of α_i is related to the power spectrum of the fluctuations of R_i.[11] Calculation of the probability distribution of







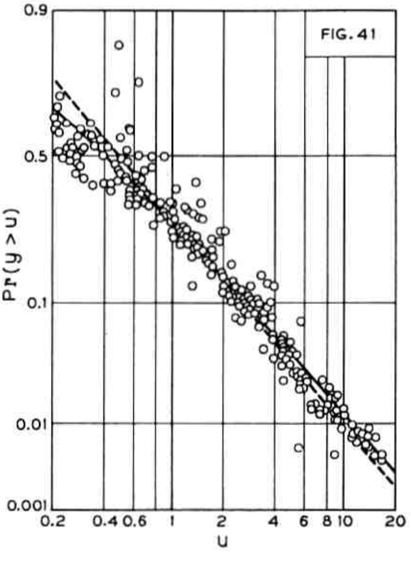


Fig. 38—Probability that duration of fade is longer than a given number of seconds for five fade depths [20 log L is: (a) −22.7 dB, (b) −27.6 dB, (c) −32.4 dB, (d) −37.6 dB, (e) −43.2 dB].

Fig. 39—Probability that duration of fade, normalized to its mean, is longer than a given number (same data as on Fig. 38—line fitted to show trend for deep fades).

Fig. 40—Probability that duration of fade, normalized to its mean, is longer than a given number (summary for the six signals in the 6-GHz band at the fade depths listed for Fig. 38—the line is from Fig. 39).

Fig. 41—Probability that duration of a simultaneous fade, normalized to its mean, is longer than a given number (summary for the 15 pairs of signals in the 6-GHz band listed in Table II—dotted line from Fig. 39—solid line is theoretical).

α_i is difficult theoretically and also because we do not have experimental information on the power spectra. Experimental determination of the spectra is far from straightforward, because fading occurs in spurts and because, presumably, the fluctuations of R_i are generated by a mix of random processes. Various theoretical spectra have been treated by S. O. Rice[11] (we obtained the idea of normalizing to the average durations from this work), but our observed durations are longer than those that follow from the spectra considered. For example, the spectrum that gives the largest percentage of long durations provides one percent of durations longer than 5.6 times the average (from Fig. 4 in Ref. 11). The data in Fig. 40 show that one percent of the durations are longer than ten times average.

To circumvent these difficulties we use the empirically determined expression (11) for $\Pr(x > u)$, and determine the duration distribution of simultaneous fades from this using a heuristic model. The model is based on the observation that when signals separated in frequency undergo a deep fade, the fade durations are often roughly the same, but the fades are offset in time. Furthermore, as fades become deeper they become shorter, and the offsets soon become larger than the fade durations, which explains why there is a large reduction in the number of deep simultaneous fades. A simultaneous fade occurs when the fades at two frequencies overlap. There is apparently no preferred overlap position, and the relationship between α and β (see Fig. 1) can be expressed mathematically as

$$\beta = g\alpha \tag{14}$$

where g is a random variable, distributed uniformly between zero and unity. The relationship (9) for the averages follows from this immediately.

The duration distribution of the simultaneous fades also follows immediately when the transformation (14) is applied to $\Pr(x > u)$. Specifically,

$$\begin{aligned}\Pr(y > u) &= \Pr(2gx > u),\\ &= \int_0^1 \Pr(x > \tfrac{1}{2}g^{-1}u)\, dg,\end{aligned} \tag{15}$$

where the integrand is (11), with appropriate change of variable. Fortunately, this can be transformed into an expression that can be integrated.[13] The result, after a fair amount of algebra, is

$$\Pr(y > u) = \tfrac{1}{2}\operatorname{erfc}\left(\frac{\ln(\tfrac{1}{2}u) - \mu}{\sqrt{2}\,\sigma}\right) - \tfrac{1}{4}u\exp(-\mu + \tfrac{1}{2}\sigma^2)\operatorname{erfc}\left(\frac{\ln(\tfrac{1}{2}u) - \mu}{\sqrt{2}\,\sigma} + \frac{\sigma}{\sqrt{2}}\right). \qquad (16)$$

The solid line in Fig. 41 is equation (16) with numerical values for μ and σ as in equation (12). The line provides adequate description of durations longer than average.

The generality imparted to the distribution functions $\Pr(x > u)$ and $\Pr(y > u)$ by the normalization of the durations to their means is a subject for further work. We would venture a guess, for example, that the distributions of x and y in the 4-GHz band are also given by equations (11) and (16). A somewhat different thought is that the idea of uniformly distributed overlaps of deep fades, which lead to equation (14), can perhaps be extended to durations of simultaneous fades of more than two signals. This approach might provide results for a problem that otherwise seems to be mathematically intractable.

IX. SIMILARITY TO DURATION DISTRIBUTION IN BEYOND-HORIZON PROPAGATION

It is interesting to compare the line-of-sight data for $\Pr(x > u)$ to similar data for tropospheric beyond-horizon propagation,[14] where fade durations have also been observed to be lognormally distributed. Table IV shows that the numerical values for line-of-sight (from our Fig. 40) are quite close to the numerical values for long term tropospheric data (from Table 2 of Ref. 14).

In line-of-sight transmission, the receiving antenna is illuminated directly by the transmitting antenna. In beyond-horizon tropospheric transmission the receiving antenna is illuminated by reflected (scattered) energy. There is insufficient knowledge to decide whether the

TABLE IV—COMPARISON OF FADE-DURATION PROBABILITIES IN LINE-OF-SIGHT AND BEYOND-HORIZON PROPAGATION

$\Pr(x > u)$	u	
	Line-of-Sight	Beyond-Horizon
0.1	2.6	2.2
0.01	10	9.1

structures of the electromagnetic waves at the receiving antennas in the two cases are similar in some sense, or whether the numerical closeness of the distributions of the durations normalized to their means is accidental.

X. COMPARISON OF FREQUENCY AND SPACE DIVERSITY

The number of simultaneous fades of the two signals must be small to obtain good diversity reception. More specifically, the ratio

$$F_N = N_i/N_{ij} \tag{17}$$

must be large. For signal envelopes that are jointly Rayleigh distributed, a theoretical result for deep fades is[9]

$$F_N \approx \tfrac{1}{2}qL^{-2}, \quad L < 0.1, \quad q^{-1}L^2 < 0.1. \tag{18}$$

This shows how F_N varies with fade depth. Information about the separation of the signals is contained in the parameter q. A theory from which q can be calculated for line-of-sight microwave links still remains to be established. However, empirical expressions for q, based on experimental data, are available,[2] and these provide a means for comparing the effectiveness of separations in frequency and space.

For separations in frequency, expressions for q determined from probability distributions of R_{ij} are[2]

$$q = \tfrac{1}{4}(\Delta f/f), \text{ in the 6-GHz band;} \tag{19}$$

$$q = \tfrac{1}{2}(\Delta f/f), \text{ in the 4-GHz band;} \tag{20}$$

where Δf is the frequency separation of R_i and R_j, and where the values of f used in the determination of the coefficients in equations (19) and (20) were 6.175 and 3.950 GHz respectively. The above expressions for q are based on data obtained on a 28.5-mile path in Ohio. We do not know, for example, how the expressions are affected by changes in path length.

For two signals received at the same frequency on two vertically separated receiving antennas, an empirical expression for q is[8, 9]

$$q = (2.75)^{-1}(s^2/\lambda d) \tag{21}$$

where s is the vertical center-to-center separation of the receiving antennas, λ is the wavelength, and d is the path length—all measured in the same units. Equation (21) is also based on data from the test path in Ohio, which was picked because fading on it was thought to be typical of many paths in the United States. Some generality is lent to equation (21) by the fact that one of the experimental points on which it is based comes from data obtained in Texas.

The equations for q can be used to compare the effects of separations in frequency and space. For example, separations that provide equal values of F_N are shown in Fig. 42. The curves apply for deep fades that satisfy the conditions $L < 0.1$ and $q^{-1}L^2 < 0.1$. As an illustration, values of F_N at $20 \log L = -40$ are indicated along the curves. Since the expressions for q are based on data for fairly small separations, equations (19) through (21) should be viewed as first terms in power series, and extrapolation of the results to separations larger than those in Fig. 42 may not be advisable.

The term improvement has been used to describe the ratio of the total time spent in fades to the total time spent in simultaneous fades.[2, 8] In terms of F_N, the deep fade approximation for the improvement F is[9]

$$F \approx 2F_N, \quad L < 0.1, \quad q^{-1}L^2 < 0.1 \tag{22}$$

and Fig. 42 therefore describes separations that provide equal improvement.

XI. CONCLUSIONS

We have presented data on aspects of fading that are important but have not been covered in previous investigations of fading on line-of-

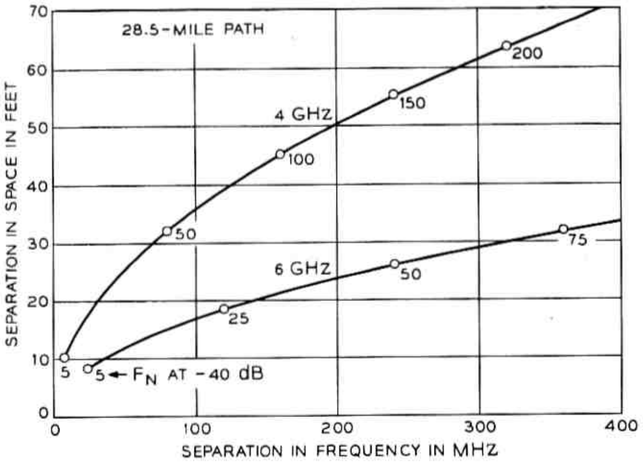


Fig. 42—Separations in space and in frequency that provide equal values of the ratio of the number of fades to the number of simultaneous fades.

sight microwave links. The theoretical framework we have provided for the data will be, we anticipate, of practical value to designers of communication systems.

The results presented here also demonstrate that further work, both experimental and theoretical, is needed. For example, the parameters in our equations are empirical and are based on data from one propagation path only.

XII. ACKNOWLEDGMENTS

The data presented come from an experiment to which many people at Bell Laboratories have contributed. From an overall point of view, there are three equally important parts to the work associated with the experiment. First, the design of the Multiple Input Data Acquisition System (MIDAS) by G. A. Zimmerman. Second, data processing, based on the work and ideas of C. H. Menzel. Third, data analysis and theory, which is our own area of work. In all parts of the work, discussions with W. T. Barnett have been invaluable.

APPENDIX

Control Curves

The equation for the control curves in Fig. 37 is, denoting the standard deviation of the sample means by σ_s ,

$$f(L) = \langle \alpha_i \rangle \pm 3\sigma_s \, . \tag{23}$$

From elementary sampling theory[15]

$$\sigma_s^2 = N_i^{-1} \sigma_\alpha^2 \tag{24}$$

where σ_α is the standard deviation of the fade durations α_i . Further

$$\sigma_\alpha = \langle \alpha_i \rangle \sigma_x \tag{25}$$

where σ_x is the standard deviation of x, and x denotes the fade durations normalized to their means. The probability distribution of the natural logarithm of x is given by equation (11). In terms of the variance of that distribution

$$\sigma_x^2 = \langle x \rangle^2 \left(\exp\left(\sigma^2\right) - 1\right). \tag{26}$$

Since $\langle x \rangle$ is unity by definition, use of σ from equation (12) gives

$$\sigma_x^2 \approx 4. \tag{27}$$

Using equation (2) to describe N_i, we obtain

$$f(L) \approx \langle \alpha_i \rangle \left\{ 1 \pm \frac{6}{\sqrt{6410L}} \right\}. \qquad (28)$$

The straight line in Fig. 37 is $\langle \alpha_i \rangle$, and equation (28) therefore provides numerical values for drawing of the control curves.

REFERENCES

1. Kaylor, R. L., "A Statistical Study of Selective Fading of Super-High Frequency Radio Signals," B.S.T.J., *32,* No. 5 (September 1953), pp. 1187–1202.
2. Barnett, W. T., "Microwave Line-of-Sight Propagation with and without Frequency Diversity," B.S.T.J., *49,* No. 8 (October 1970), pp. 1827–1871.
3. Baker, A. E., and Brice, P. J., "Radio-Wave Propagation as a Factor in the Design of Microwave Relay Links," Post Office Elec. Eng. J., *62,* No. 4 (January 1970), pp. 239–246.
4. Ugai, S., "Characteristics of Fading due to Ducts and Quantitative Estimation of Fading," Rev. Elec. Commun. Lab. (Japan), *9,* No. 5–6 (May–June 1961), pp. 319–360.
5. Wheeler, B. F., and Mathwich, H. R., "Use of Distribution Curves in Evaluating Microwave Path Clearance," IRE Trans. Instrumentation, *PGI-4,* (October 1955), p. 31.
6. Mathwich, H. R., Nuttall, E. D., Pitman, J. E., and Randolph, A. M. "Propagation Test on 955.5 Mc, 1965 Mc and 6730 Mc," AIEE Trans. Commun. Elec., Part I, *75* (January 1957), pp. 685–691.
7. Gudmandsen, P., and Larsen, B. F., "Statistical Data for Microwave Propagation Measurements on Two Oversea Paths in Denmark," Trans. IRE on Antennas and Propagation, *AP-5,* No. 3 (July 1957), pp. 255–259.
8. Vigants, A., "Space-Diversity Performance as a Function of Antenna Separation," IEEE Trans. Commun. Tech., *COM-16,* No. 6 (December 1968), pp. 831–836.
9. Vigants, A., "The Number of Fades in Space-Diversity Reception," B.S.T.J., *49,* No. 7 (September 1970), pp. 1513–1530.
10. Pearson, K. W., "Method for the Prediction of the Fading Performance of a Multisection Microwave Link," Proc. IEE (London), *112,* No. 7 (July 1965), pp. 1291–1300.
11. Rice, S. O., "Distribution of the Duration of Fades in Radio Transmission: Gaussian Noise Model," B.S.T.J., *37,* No. 3 (May 1958), pp. 581–635.
12. Barnett, W. T., "Occurrence of Selective Fading as a Function of Path Length, Frequency, and Geography," 1969 USNC/URSI Spring Meeting, April 21–24, 1969, Washington, D. C.
13. Abramowitz, M., and Stegun, I. A., editors, *Handbook of Mathematical Functions,* New York: Dover Publications, 1965, p. 304.
14. Grosskopf, J., and Fehlhaber, L., "Rate and Duration of Single Deep Fades in Tropospheric Scatter Links," NTZ-Commun. J., *1,* No. 3 (1962), pp. 125–132.
15. Hoel, P. G., *Introduction to Mathematical Statistics,* New York: John Wiley & Sons, 1954, pp. 105–109.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

Time Dispersion in Dielectric Waveguides

By S. D. PERSONICK

(Manuscript received October 16, 1970)

In dielectric waveguides operating at optical frequencies, the primary cause of time dispersion of narrow pulses can be mode conversion. In this paper we argue that under certain assumptions a dielectric waveguide acts as a linear system in intensity. That is, given the intensity input, the intensity output is equal to the input convolved with an intensity impulse response. We show that contrary to intuition, the width of the impulse response gets narrower when coupling between guided modes increases. Using the perturbation results of D. Marcuse, we obtain an interesting model of energy propagation down imperfect guides. We conclude that the intensity response width increases as the square root of the guide length for sufficiently long guides and approaches a gaussian shape for sufficiently long guides.

We conclude from the theory that the dispersion in dielectric waveguides may be orders of magnitude below that which was previously expected in guides of sufficiently long length having properly controlled large amounts of mode conversion. These theoretical results have not yet been verified experimentally.

I. INTRODUCTION

In multimode dielectric waveguides operating at optical frequencies, the primary cause of time dispersion of narrow pulses can be mode conversion. In a geometrically perfect guide with more than a single mode, energy initially launched in a given mode remains in that mode as it propagates down the guide. Physical guides have imperfections from perfect geometric shape (e.g., roughness at the core-cladding interface of a nominally right circular cylindrical guide) which allows energy to couple between modes during propagation down the guide. Since group velocities differ in general amongst the modes, a pulse of energy initially launched in a single mode or combination of modes will be broadened due to the spread of propagation times of different parts of the energy.

In this paper we argue that under certain assumptions, a dielectric waveguide acts as a linear system in intensity as well as in voltage. That is, we show that the relationship between the input intensity and output intensity of the guide is defined in terms of an intensity impulse response. We argue that this intensity impulse response, for sufficiently long guides, has a mean-square width about its mean which increases only linearly with length. Further, in the limit of very long guides, we argue that the response shape is gaussian. We also show that the greater the coupling between modes, the less the time dispersion—a result which at first contradicts intuition. Finally, we obtain quantitative results and an interesting model of an optical guide under certain assumptions.

We conclude from the theory that the dispersion in dielectric waveguides may be orders of magnitude below that which was previously expected in guides of sufficiently long length having properly controlled *large* amounts of mode conversion. These theoretical results have not yet been verified experimentally.

II. AN OUTLINE OF THE ARGUMENTS

We next outline the steps of the derivations to follow, so that the reader can follow the train of thought.

We start with the fact that the optical guide is a linear system in voltage. That is, if we expand the input signal in spatial modes and expand the output signal in the same modes, then the time varying coefficients of the modes at the output are related to the coefficients at the input by a set of voltage impulse responses. We then make an assumption about the associated set of transfer functions (Fourier transforms of the impulse responses) which allows us to argue that the set of average output intensities and the set of input intensities are also related by a set of impulse responses. Thus the guide is also linear in intensity under the assumptions.

We next argue that for sufficiently long guides, these intensity impulse responses coupling a chosen input mode coefficient and a chosen output mode coefficient are indifferent to the modes chosen except perhaps for a magnitude scale factor.

Finally, this allows us to show that for guides longer than the above scale, the intensity impulse response which is now in common for all input-output pairs has a mean-square deviation about its mean which grows linearly in length, and which approaches the gaussian shape in the limit of long guides.

III. THE OPTICAL GUIDE AS A LINEAR SYSTEM IN INTENSITY

3.1 *The Random Channel*

We now argue that under a simple assumption, the expected value of the intensity at the output of a random channel is related to the intensity of the input to the channel by a simple convolution with an intensity impulse response. We start with input baseband signal $a(t)$. We use $a(t)$ to linearly modulate a carrier $m(t)$ which may be coherent or a stationary (wide sense) random process centered at frequency f_0. We assume that the result $x(t) = a(t)m(t)$, has bandwidth B, i.e., its spectrum extends from $-B/2 + f_0$ to $B/2 + f_0$.

We pass $x(t)$ through a time invariant filter with a random impulse response $h(t)$ representing the channel, resulting in the final output $y(t)$. We define a Fourier transform relationship between the function $\Lambda(f)$ and the function $\lambda(t)$

$$\Lambda(f) = \int_{-\infty}^{\infty} \exp\,[i2\pi f(t)]\lambda(t)\,dt, \qquad (1)$$

$$\Lambda(f) \Leftrightarrow \lambda(t).$$

By simple linear system theory if

$$\begin{aligned} X(f) &\Leftrightarrow x(t),\\ H(f) &\Leftrightarrow h(t),\\ Y(f) &\Leftrightarrow y(t), \end{aligned} \qquad (2)$$

then

$$Y(f) = X(f)H(f).$$

Define the envelopes of $x(t)$ and $y(t)$ by

$$\begin{aligned} x(t) &= \sqrt{2}\ \mathrm{Re}\ \{x_e(t)\exp\,(i2\pi f_0 t)\},\\ y(t) &= \sqrt{2}\ \mathrm{Re}\ \{y_e(t)\exp\,(i2\pi f_0 t)\}. \end{aligned} \qquad (3)$$

The intensity of the input and output signals are defined as

$$\begin{aligned} I_{\mathrm{in}}(t) &= |x_e(t)|^2 = a^2(t)\ |m_e(t)|^2,\\ I_{\mathrm{out}}(t) &= |y_e(t)|^2, \end{aligned} \qquad (4)$$

where $m_e(t)$ = carrier envelope.

Assumption: Stationarity of channel transfer function

$$\langle H^*(\alpha)H(f + \alpha)\rangle = \Gamma(f) \qquad (5)$$

provided $f_0 - B/2 < \alpha, f + \alpha < f_0 + B/2$.

The assumption, while apparently arbitrary, is essential to the results which follow. Perturbation results of Marcuse,[1] to be discussed in Section 4.1, indicate that equation (5) may be satisfied for the input-output temporal transfer function of a given spatial eigenmode of an optical dielectric waveguide with mechanical imperfections, provided the mechanical imperfections satisfy constraints also to be discussed.

Define for any function $U(\alpha)$

$$\begin{aligned} U_+(\alpha) &= U(\alpha) \qquad \alpha \geqq 0, \\ &= 0 \qquad\qquad \alpha < 0. \end{aligned} \tag{6}$$

Then it has been shown that (See Appendix C)

$$|y_e(t)|^2 \Leftrightarrow 2 \int_{-\infty}^{\infty} Y_+(f + \alpha) Y_+^*(\alpha) \, d\alpha. \tag{7}$$

Then clearly

$$|y_e(t)|^2 \Leftrightarrow 2 \int_{-\infty}^{\infty} X_+(f + \alpha) H_+(f + \alpha) H_+^*(\alpha) X_+^*(\alpha) \, d\alpha. \tag{8}$$

Using equation (5) we obtain

$$\begin{aligned} \langle |y_e(t)|^2 \rangle &\Leftrightarrow 2\Gamma(f) \int_{-\infty}^{\infty} \langle X_+(f + \alpha) X_+^*(\alpha) \rangle \, d\alpha, \\ \langle |y_e(t)|^2 \rangle &\Leftrightarrow \Gamma(f) I_{\text{in}}(f), \end{aligned} \tag{9}$$

where

$$I_{\text{in}}(f) \Leftrightarrow \langle I_{\text{in}}(t) \rangle.$$

Thus*

$$\langle I_{\text{out}}(t) \rangle = \langle I_{\text{in}}(t) \rangle * \gamma(t) \tag{10}$$

where

$$\gamma(t) \Leftrightarrow \Gamma(f).$$

Thus we have a linear system relationship between the channel input and output intensities.

3.2 *Extension to Vector Channels*

Suppose we have a vector channel (corresponding to multimode guide) consisting of a vector of L input functions

* The notation $x(t) * y(t)$ signifies convolution:

$$x(t) * y(t) \triangleq \int_{-\infty}^{\infty} x(t - u) y(u) du.$$

$$\mathbf{X}(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_L(t) \end{bmatrix}$$

and L output functions

$$\mathbf{Y}(t) = \begin{bmatrix} y_1(t) \\ \vdots \\ y_L(t) \end{bmatrix}.$$

The input and output functions are related by an $L \times L$ matrix of impulse responses

$$\boldsymbol{\theta}(t) = [h_{ij}(t)] \tag{11}$$

where we have

$$y_l(t) = \sum_{j=1}^{L} h_{lj}(t) * x_j(t), \tag{12}$$

i.e.,

$$\mathbf{Y}(t) = \boldsymbol{\theta}(t) * \mathbf{X}(t).$$

Now consider a cascade of two vector channels having impulse response matrices ${}^1\boldsymbol{\theta}(t)$ and ${}^2\boldsymbol{\theta}(t)$. The input passes first through channel 1 and then through channel 2. The output of channel 2 is given by

$$\mathbf{Y}(t) = {}^2\boldsymbol{\theta}(t) * {}^1\boldsymbol{\theta}(t) * \mathbf{X}(t). \tag{13}$$

Define the envelope of $y_k(t)$, $y_{ke}(t)$: we know that

$$|y_{ke}(t)|^2 \Leftrightarrow 2 \int_{-\infty}^{\infty} Y_{k+}(\alpha + f) Y_{k+}^*(\alpha)\, d\alpha,$$

$$|y_{ke}(t)|^2 \Leftrightarrow 2 \int_{-\infty}^{\infty} \sum_j \sum_l \sum_m \sum_n {}^2H_{kl+}(f + \alpha){}^1H_{lj+}(f + \alpha)X_{j+}(f + \alpha) \tag{14}$$

$$\cdot {}^2H_{kn+}^*(\alpha){}^1H_{nm+}^*(\alpha)X_{m+}^*(\alpha)\, d\alpha.$$

Assumption a. Stationarity of Mode Transfer function.
b. Mode Transfer functions uncorrelated.

$$\langle {}^1H_{lj}(f + \alpha){}^2H_{kl}(f + \alpha){}^1H_{nm}^*(\alpha){}^2H_{rn}^*(\alpha)\rangle$$

$$= {}^1\Gamma_{lj}(f){}^2\Gamma_{kl}(f)\ \delta_{l,n}\ \delta_{j,m}\ \delta_{k,r} \tag{15}$$

where

$$\delta_{X,Y} \triangleq \text{Kronecka delta}$$

for $\{f + \alpha, \alpha\} \epsilon$ input signal bands.

We are implying that randomness, especially in phase, erases correlation between the transfer functions of different modes. The validity of this assumption for optical guides with more than one mode will be discussed in Section 4.1.

It then follows that

$$\begin{aligned} \langle |y_{ke}(t)|^2 \rangle &\Leftrightarrow 2 \int_{-\infty}^{\infty} \sum_{j=1}^{L} \sum_{l=1}^{L} {}^1\Gamma_{lj}(f)^2\Gamma_{kl}(f) \langle X_{j+}(f + \alpha) X_{j+}^*(\alpha) \rangle \, d\alpha, \\ \langle |y_{ke}(t)|^2 \rangle &\Leftrightarrow \sum_{j=1}^{L} \sum_{l=1}^{L} {}^2\Gamma_{kl}(f)^1\Gamma_{lj}(f) I_{\mathrm{inj}}(f) \end{aligned} \tag{16}$$

where

$$I_{\mathrm{inj}}(f) \Leftrightarrow \langle | x_{je}(t) |^2 \rangle.$$

Thus under assumption (15) we have

$$\langle |y_{ke}(t)|^2 \rangle = \sum_{j=1}^{L} \sum_{l=1}^{L} {}^2\gamma_{kl}(t) * {}^1\gamma_{lj}(t) * \langle |x_{je}(t)|^2 \rangle \tag{17}$$

where

$$\gamma_{kl}(t) \Leftrightarrow \Gamma_{kl}(f).$$

Forming the matrix ${}^1G(t)$ with elements ${}^1\gamma_{kl}(t)$ and similarly ${}^2G(t)$; the vectors of input and output *intensities* are related by

$$| Y_e |^2 = {}^2G(t) * {}^1G(t) * |X_e |^2. \tag{18}$$

Thus the vector channel is a vector linear system in intensity as well as voltage [compare equation (18) to equation (13)].

3.3 *A Limit Theorem for a Cascade of Vector Channels*

Now consider a cascade of a large number M of vector channels, each behaving as described in Sections 3.1 and 3.2, i.e., if $|Y_e(t)|^2$ is the vector of the average intensity responses at the output, $|X_e(t)|^2$ the input intensity vector; we have

$$|Y_e(t)|^2 = \left(\mathop{*}_{j=1}^{M} \mathbf{G}_j(t) \right) * |\mathbf{X}_e(t)|^2 = \mathbf{G}_T * |\mathbf{X}_e(t)|^2$$

where

$$\mathbf{G}_T = \mathop{*}_{j=1}^{M} G_j(t) = \mathbf{G}_M * \mathbf{G}_{M-1} \cdots * \mathbf{G}_1 . \tag{20}$$

We would like to argue now that for sufficiently large M, all the elements of G_T are identical in waveform, differing at most by a constant. In other words, we would like to argue that the shape of the intensity response between any input mode and any output mode is indifferent to the choices of input and output modes, for sufficiently long guides, except perhaps for the magnitude of the responses.

We shall prove our results for a lossless two-mode guide. Define $\gamma_{ij}(t,\ b)$ as the $j \to i$ intensity impulse response for b sections of guide. We obtain (see Appendix A for derivation)

$$\gamma_{21}(t,\ b) = \sum_{R=1}^{b} \gamma_{11}(t,\ b - R) * \gamma_{21}(t) * \gamma_{22}^{*R-1}(t) \tag{21}$$

where

$$\gamma_{22}^{*R-1}(t) \triangleq \gamma_{22}(t) * \gamma_{22}(t) \cdots R - 1 \text{ times}$$

and

$$\gamma_{11}(t,\ 0) = \delta(t).$$

Now assume that the guide is lossless, i.e.,

$$\int_{-\infty}^{\infty} \{\gamma_{11}(t) + \gamma_{21}(t)\}\ dt = 1. \tag{22}$$

Further define

$$\gamma_{ij}(t) = a_{ij}\rho_{ij}(t),$$

$$\int_{-\infty}^{\infty} \rho_{ij}(t)\ dt = 1, \qquad 0 < a_{ij} < 1. \tag{23}$$

Thus

$$\gamma_{21}(t,\ b) = \sum_{R=1}^{b} \gamma_{11}(t,\ b - R) * \rho_{21}(t) * \rho_{22}^{*R-1}(t)(a_{21}a_{22}^{R-1}). \tag{24}$$

For the lossless guide, and b sufficiently large $\gamma_{11}(b - R) \approx \gamma_{11}(t,\ b)$ for $R \ll b$. Furthermore a convolution of $\gamma_{11}(t,\ b - R)$ with $\rho_{21}(t) * \rho_{22}^{*R-1}(t)$ is approximately equal to $\gamma_{11}(t,\ b - R)$ for $R \ll b$ since the response $\gamma_{11}(t,\ b - R)$, which is a convolution of $b - R$ terms, has a narrow spectrum compared to the other R term convolution for $b \gg R$. Furthermore $a_{21}a_{22}^{R-1} \to 0$ for R large. Thus for b sufficiently large

$$\gamma_{21}(t,\ b) \approx \gamma_{11}(t,\ b)(a_{21}/(1 - a_{22})). \tag{25}$$

Similarly we have

$$\gamma_{11}(t,\,b) = \sum_{R=1}^{b-1} \gamma_{12}(t,\,b-R) * \rho_{21}(t) * \rho_{11}^{*R-1}(t)a_{21}a_{11}^{R-1},$$

$$\approx \gamma_{12}(t,\,b)\,\frac{a_{21}}{1-a_{11}} = \gamma_{12}(t,\,b), \tag{26}$$

(since $a_{21} + a_{11} = 1$). Thus

$$\mathbf{G}(t,\,b) \cong \rho(t,\,b)\begin{bmatrix} \dfrac{1}{1+\eta} & \dfrac{1}{1+\eta} \\ \dfrac{1}{1+1/\eta} & \dfrac{1}{1+1/\eta} \end{bmatrix} = \rho(t,\,b)\mathbf{A}, \tag{27}$$

where $\eta = a_{21}/a_{12}$, and $\rho(t,\,b) \triangleq \gamma_{11}(t,\,b)/\int_{-\infty}^{\infty} \gamma_{11}(t,\,b)\,dt$. Finally we obtain the response of a guide of kb sections

$$\mathbf{G}(t,\,kb) = \rho^{*k}(t,\,b)\mathbf{A}.$$

Note that $\mathbf{A}$ is idempotent, i.e., $\mathbf{A}^2 = \mathbf{A}$ and $\rho(t,\,b)$ is a positive unit area function.

3.4 *Application to Long Optical Guide*

For a multimode lossless of guide sufficiently long length, $l = kL$, with finite coupling between all modes, we can generalize equation (27) to conclude that the intensity impulse response between an input and output mode is a constant times some positive unit area function $\rho(t,\,L)$ convolved with itself l/L times when L is a scale on which equation (27) holds in the generalized case (more than two modes). Since the central limit theorem states that the convolution of a large number of unit area positive functions approaches a gaussian shape,* we conclude that the impulse response should approach a gaussian shape in the limit of long guides. Further, the impulse response's second moment about its mean increases linearly with increasing guide length for guides longer than L.† That is, the second moment about the mean of the response is

$$M_2(l) = M_2(L)l/L. \tag{28}$$

If τ_1 is the "differential delay" (time/meter) of propagation in the

* Provided that $\int_{-\infty}^{\infty} \rho(t,\,L)t^2dt < \infty$, when we add similar independent random variables with finite second moments, the probability density of the sum, which is the *convolution* of the individual densities, approaches a gaussian shape.

† The second central moment of a convolution is the sum of the individual second central moments.

slowest modes and τ_2 is the differential delay in the fastest mode, then*

$$M_2(L) \leqq \frac{(\tau_1 - \tau_2)^2 L^2}{4} = \frac{(\Delta\tau L)^2}{4}. \tag{29}$$

Therefore,

$$M_2(l) \leqq \frac{(\Delta\tau)^2}{4} Ll$$

for any L where equation (27) holds [for an N mode case we have an $N \times N$ matrix multiplying $\rho(t, l)$].

IV. QUANTITATIVE RESULTS

4.1 *Perturbation Theory*

We shall now apply the above results to the case of a lossless slab dielectric waveguide previously studied by Marcuse.[1] We expand the input field to the guide as

$$\epsilon(t, x, 0) = \sum \epsilon_k(t, 0)\psi_k(x) \tag{30}$$

where x is the cross-sectional position parameter and the $\psi_k(x)$ are the eigenmodes of the guide. The field a distance l down the guide is written as

$$\epsilon(t, x, l) = \sum \epsilon_k(t, l)\psi_k(x). \tag{31}$$

We have a linear voltage impulse response relationship between the vector of input voltages $[\epsilon_k(t, 0)]$ and the vector of output voltages $[\epsilon_k(t, l)]$. Defining $E_k(\omega, l)$ as the Fourier transform of $\epsilon_k(t, l)$ we have

$$E_k(\omega, l) = \sum_j C_{kj}(\omega, l)E_j(\omega, 0). \tag{32}$$

Marcuse has shown that a *perturbation* theory solution for the $C_{kj}(\omega, l)$ is given by

$$C_{kj}(\omega, l) = \lambda_{kj} \exp [i\beta_j(\omega)l] \int_0^l g(z) \exp \{i[\beta_k(\omega) - \beta_j(\omega)]z\} \, dz \tag{33}$$

where λ_{kj} is a constant weakly dependent upon ω and $g(z)$ is the wall perturbation from straightness. It is assumed $k \neq j$ [For $k = j$, $C_{kj}(\omega, l) = 1$]. We have therefore

* The right side of equation (29) is the mean-square intensity impulse response width if the response consists of an impulse of area ½ at the shortest delay and an impulse of area ½ at the longest delay.

$$C_{kj}(\omega, l)C^*_{np}(\omega + \sigma, l)$$
$$= \lambda_{kj}\lambda^*_{np} \exp\{i(\beta_j(\omega) - \beta_p(\omega + \sigma))l\} \int_0^l \int g(z)g(z')$$
$$\cdot \exp\{i[(\beta_k(\omega) - \beta_j(\omega))z - (\beta_n(\omega + \sigma) - \beta_p(\omega + \sigma))z']\}\, dz\, dz'. \qquad (34)$$

Defining the correlation function

$$\langle g(z)g(z')\rangle = R_g(z - z') \qquad (35)$$

(we assume $g(z)$ is a wide sense stationary process) we obtain

$$\langle C_{kj}(\omega, l)C^*_{np}(\omega + \sigma, l)\rangle = \lambda_{k1}\lambda^*_{np} \exp(i\,\Delta\beta_2 l) \int^l \int R_g(z - z')$$
$$\cdot \exp[i(\beta_k(\omega) - \beta_j\omega)(z - z')] \exp[i\,\Delta\beta(\omega, \sigma)z']\, dz\, dz' \qquad (36)$$

where

$$\Delta\beta = (\beta_k(\omega) - \beta_n(\omega + \sigma)) - (\beta_j(\omega) - \beta_p(\omega + \sigma)),$$
$$\Delta\beta_2 = \beta_j(\omega) - \beta_p(\omega + \sigma).$$

If $R_g(z - z')$ drops off quickly for $(z - z')$ in an interval of length l, then we have the approximate result

$$\langle C_{kj}(\omega, l)C^*_{np}(\omega + \sigma, l)\rangle \simeq \lambda_{kj}\lambda^*_{np} l S_g(\beta_k(\omega) - \beta_j(\omega))$$
$$\cdot \exp(i\,\Delta\beta_2 l)\, \frac{1 - \exp[i\,\Delta\beta(\omega, \sigma)l]}{i\,\Delta\beta(\omega, \sigma)l} \qquad (37)$$

where $S_g(\cdot)$ = Fourier Transform of $R_g(\cdot)$, and is assumed to be constant as a function of $\beta_k(\omega) - \beta_i(\omega)$ for ω within the excitation bandwidth. For $k = n$, $j = p$

$$\Delta\beta \simeq \left[\frac{\partial\beta_k(\omega)}{\partial\omega} - \frac{\partial\beta_j(\omega)}{\partial\omega}\right]\sigma \triangleq \Delta\tau_{kj}\sigma,$$
$$\Delta\beta_2 = \left(\frac{\partial\beta_j}{\partial\omega}\right)\sigma. \qquad (38)$$

That is we assume σ is small enough so that there is negligible dispersion of energy travelling in a single mode. Thus, the intensity impulse response between input j and output k is (see Fig. 1)

$$\gamma_{kj}(t, l) \simeq \lambda_{mj}\lambda^*_{mj} S_g(\beta_m(\omega) - \beta_j(\omega)) f(t - \tau_j l) \qquad (39)$$

where

$$f(t) = 1, \qquad t \epsilon [0, \Delta\tau_{mj} l];$$
$$= 0, \qquad \text{otherwise};$$

for l small enough for the perturbation theory to hold.

From equations (36) through (38), it is clear that for cases where we do not have $k = n$, $j = p$ the correlation function $C_{kj}(\omega)C_{np}^{*}(\omega + \sigma)$ will be negligible provided $\Delta\beta(\omega, \sigma)$ is sufficiently large in the band of input frequencies. Thus we can use the perturbation length impulse response to find a long-guide response by means of equation (20).

From equation (37) we see that we increase coupling between modes by making the mechanical perturbation spectrum large at frequencies which correspond to the difference of the inverses of the phase velocities of the modes at excitation frequencies. It should be emphasized that while making the perturbation spectrum high at frequencies that couple guided modes, we will wish to avoid making it too high at frequencies that couple guided to unguided radiation modes since such coupling results in loss.

4.2 *A Hydraulic Model of Dispersion*

We shall now show that the perturbation results imply a model which is an interesting interpretation of the propagation process, and which allows easy computation of the response of a long guide.

Suppose energy traveled down the guide as follows. We start with a large number of indivisible bundles of energy at the guide input. Each bundle begins propagating down the guide randomly jumping from mode to mode. At any point down the guide, a bundle travels at the group velocity associated with the mode it is currently in. At any position, the probability that a bundle will jump to mode k, given that it is in mode j, in the next increment of distance dl is $\lambda_{kj}\lambda_{kj}^{*}S_g(\beta_k(\omega) - \beta_j(\omega))\, dl$.

Since we have a very large number of bundles, the output response of the guide in intensity should have the same shape as the probability distribution of the arrival time at the output of an individual bundle. For a short guide of length L, the probability that a bundle is in mode k given that it started in mode j is $| \lambda_{kj} |^2 S_g(\beta_k(\omega) - \beta_j(\omega))L$ and its arrival time distribution is given exactly by Fig. 1. Since this distribution is the perturbation solution for the intensity impulse response of a short guide, we see that the hydraulic model gives the same result as the perturbation theory. Further, a little thought will show (see Appendix B) that the extrapolation from a short guide to a long guide in the hydraulic model is analytically the same as equation (20). Thus any technique which can be used to determine the intensity impulse response characteristics using the hydraulic model will be valid for the solution of equation (20) using the perturbation results.

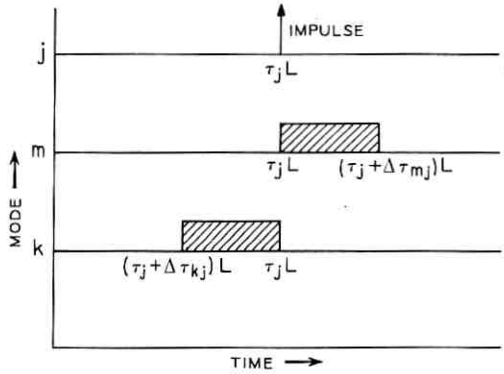


Fig. 1—Output intensity response for modes *j*, *m* and *k*, given mode *j* is excited.

4.3 *The Solution of Some Hydraulic Model Response*

4.3.1 *Characteristics*

We now wish to determine some probability density moments of the arrival time of a bundle of energy at the output of a long guide recalling that this has the shape of the guide impulse response.

Let $H(l')$ be the mode a bundle is in at distance l' down the guide. In that mode the bundle travels with differential delay (time/meter) $\tau_H = \tau(l')$. The total propagation time down the guide is

$$T = \int_0^l \tau(l')\, dl'. \tag{40}$$

The expected propagation time down the guide is

$$\langle T \rangle_{\text{av}} = \int_0^l \langle \tau(l')\, dl' \rangle_{\text{av}} \,. \tag{41}$$

The variance about the mean is

$$\begin{aligned}\langle (T - \langle T \rangle_{\text{av}})^2 \rangle_{\text{av}} &= \left\langle \int_0^l \int_0^l (\tau(l') - \langle \tau \rangle_{\text{av}}(l'))(\tau(l'') - \langle \tau \rangle_{\text{av}}(l''))\, dl'\, dl'' \right\rangle \\ &= \left[\int_0^l \int R_\tau(l', l'')\, dl'\, dl'' \right] - \langle T \rangle_{\text{av}}^2 \end{aligned} \tag{42}$$

where $R_\tau(l', l'') = E(\tau(l')\tau(l''))$. We need the correlation function $R_\tau(l, l')$ and the mean $\langle \tau \rangle_{\text{av}}(l')$.

4.4 *Calculation for a Lossless Two-Mode Guide*

For the lossless two-mode guide we have an energy bundle making a Poisson number of mode changes in any length L with mean $|\lambda_{12}|^2 S_g(\beta_1(\omega) - \beta_2(\omega))L$. The correlation function $R_\tau(l, l')$ (assuming we start off randomly in one of the modes) is that of a random telegraph wave[2] and is given by

$$R_\tau(l, l') = \tfrac{1}{4} |\Delta\tau_{12}|^2 \exp(-2|l - l'|/l_c) + \left(\frac{\tau_1 + \tau_2}{2}\right)^2 (l - l')^2 \qquad (43)$$

where

$$1/l_c = |\lambda_{12}|^2 S_g(\beta_1(\omega) - \beta_2(\omega)),$$

and

$$\Delta\tau_{12} = \left(\frac{\partial\beta_1}{\partial\omega} - \frac{\partial\beta_2}{\partial\omega}\right) = \tau_1 - \tau_2 .$$

We obtain the mean and second moment about the mean of the intensity response of a guide of length L.

$$\langle T\rangle_{\text{av}} = \left(\frac{\tau_1 + \tau_2}{2}\right)L,$$

$$\langle (T - \langle T\rangle_{\text{av}})^2\rangle_{\text{av}} = \frac{(\Delta\tau_{12})^2 l_c L}{4}\left[1 - \frac{l_c}{2L}(1 - \exp(-2L/l_c))\right], \qquad (44)$$

$$\lim_{[L/l_c]\to\infty} \langle (T - \langle T\rangle_{\text{av}})^2\rangle_{\text{av}} = \frac{(\Delta\tau_{12})^2 L l_c}{4},$$

where

$$l_c = [|\lambda_{12}|^2 S_g(\beta_1(\omega) - \beta_2(\omega))]^{-1},$$

$$\lim_{[L/l_c]\to 0} \langle (T - \langle T\rangle_{\text{av}})^2\rangle_{\text{av}} = \frac{(\Delta\tau_{12})^2 L^2}{4},$$

(compare equation (44) to equation (29)).

4.5 *Extension to the Two-Mode Guide With Loss*

We can use the hydraulic model to extend the above results to a two-mode guide with loss and differential loss.

Assume that when travelling at distance dl in mode j, a bundle of light has probability $\alpha_j dl$ of being absorbed.

If in travelling down a guide of length L, the bundle spends a distance L_1 in mode 1 and L_2 in mode 2, then the probability that it is not absorbed is

$$P = \exp\left[-(\alpha_1 L_1 + \alpha_2 L_2)\right] = \exp\left[-((\alpha_1 - \alpha_2)L_1 + \alpha_2 L)\right]. \quad (45)$$

The number of bundles entering the guide at time zero and arriving at the output end at time t in the presence of loss equals the number that would arrive at time t in the absence of loss times the probability that a bundle with total travel time t is not absorbed. But we have

$$\begin{aligned} &L_1 + L_2 = L, \\ &\qquad t = \tau_1 L_1 + \tau_2 L_2 = (\tau_1 - \tau_2)L_1 + \tau_2 L, \\ &\qquad P = \exp\left[-(\alpha_1 L_1 + \alpha_2 L_2)\right] = \exp\left\{-\left(\Delta\alpha \frac{(t - \tau_2 L)}{\Delta t_{12}} + \alpha_2 L\right)\right\} \end{aligned} \quad (46)$$

where

$$\Delta\tau_{12} = \tau_1 - \tau_2 \quad \Delta\alpha = \alpha_1 - \alpha_2.$$

With a little algebra we obtain

$$P = \exp\left[-\left\{\frac{\Delta\alpha}{\Delta\tau_{12}}(t - \langle\tau\rangle_{\text{av}}L) + \langle\alpha\rangle_{\text{av}}L\right\}\right] \quad (47)$$

where

$$\begin{aligned} \langle\alpha\rangle_{\text{av}} &= (\alpha_1 + \alpha_2)/2, \\ \langle\tau\rangle_{\text{av}} &= (\tau_1 + \tau_2)/2. \end{aligned}$$

Thus the intensity impulse response for a two-mode guide with loss is equal to the lossless response multiplied by P of equation (47).

Note that the gaussian shape for long guides still holds because the product of a gaussian and an exponential envelope is a shifted gaussian.

V. CONCLUSIONS

We can conclude at least one important result. Long optical fiber waveguides need not have large dispersion due to random imperfections if properly controlled mode coupling exists. From equation (44) we see that a mechanical perturbation spectrum which is peaked at frequencies that couple guided modes will lower dispersion. However, to avoid loss, we must not make the mechanical perturbation spectrum too high at frequencies that couple guided and radiating modes.

The above conclusions have been obtained by D. T. Young and H. E. Rowe,[3] for the two-mode guide by solving the coupled line equations directly under the assumption of white noise coupling.

APPENDIX A

We wish to derive equation (21) from the following relationship

$$G(t, b) = [\gamma_{ij}(t, b)] = [\gamma_{ij}(t)]^*[\gamma_{ij}(t)]^* \cdots (b \text{ Times}), \qquad i, j = 1, 2. \qquad (48)$$

Equation (48) implies the following model shown in Fig. 2. The transfer function $\gamma_{21}(t, b)$ is the overall transmission response between input 1 and output 2. This can be obtained by adding up the transmission responses over all different paths between input 1 and output 2 using any desired bookkeeping scheme. Every path between input 1 and output 2 must pass through the $\gamma_{21}(t)$ function *for the last time* in some section. If a path passes through the $\gamma_{21}(t)$ function for the last time in the fifth section from the end, then it must pass through four $\gamma_{22}(t)$ functions on its way to output 2. The sum of the path transfer functions between input 1 and the input to the $\gamma_{21}(t)$ function in the fifth section from the end is $\gamma_{11}(t, b - 5)$. Thus the contribution to the overall transfer function between input 1 and output 2 due to all paths which pass through a $\gamma_{21}(t)$ function for the last time in the fifth section from the end is $\gamma_{11}(t, b - 5) * \gamma_{21}(t) * \gamma_{22}^{*4}(t)$. Equation (21) merely expresses the sum of the contributions over all positions of last passage through a γ_{21} (t) function.

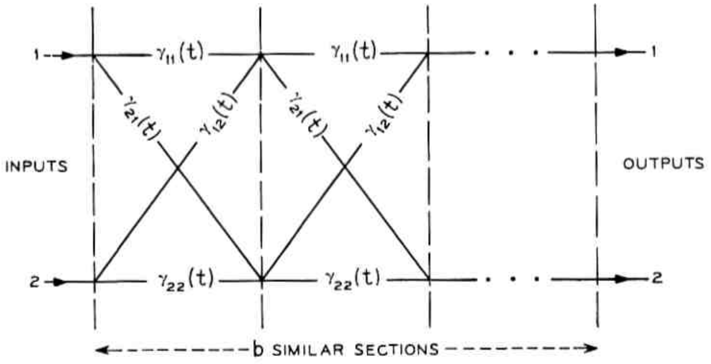


Fig. 2—*b*-section guide.

$$\gamma_{21}(t, b) = \sum_{R=1}^{b} \gamma_{11}(t, b - R) * \gamma_{21}(t) * \gamma_{22}^{*R-1}(t). \tag{21}$$

APPENDIX B

We wish to show that the equations for obtaining the intensity response of a long guide, given the intensity response of a short guide, for the hydraulic model are identical to the extrapolation equations for the intensity response given by equation (20).

Suppose we have the probability density, for a short guide, that a bundle of energy starting off in mode j of the guide (at time zero) arrives at the output of the guide at time t in mode i. Thus we have the matrix of densities $P(t, L)$ where L is the guide length and the elements $p_{ij}(t, L)$ are the previously described densities. Let $I_j(t, 0)$ be the probability density that a bundle of light arrives at input j at time t. Let $I_j(t, L)$ be the probability density that a bundle arrives at the output position L in mode j at time t. Let $I(t, \cdot)$ be the corresponding vectors. Using the laws of addition of random variables we obtain

$$I_i(t, L) = \sum_s p_{is}(t, L) * I_s(t, 0)$$

of in matrix notation

$$I(t, L) = P(t, L) * I(t, 0)$$

therefore

$$I(t, kL) = \left(\overset{k}{\underset{1}{*}} P(t, L)\right) * I(t, 0).$$

We see that the probability density of the output arrival mode and time of a bundle of energy for a long guide, which corresponds to the intensity response, is extrapolated from the short-guide response exactly as in equation (20). Thus since the perturbation results of Marcuse correspond to the hydraulic model in the limit of short guides and satisfy the conditions for extrapolation using equation (20), it follows that properties of the hydaulic model solution for long guides will correspond to the solution of equation (20) starting with these perturbation results. This is true no matter what techniques we use to find these hydaulic model properties.

APPENDIX C

We wish to establish that for a narrowband high frequency signal

$$y(t) = (y_e(t) \exp (i\omega_0 t) + y_e^*(t) \exp (-i\omega_0 t))/\sqrt{2}$$

of carrier frequency $f_0 = \omega_0/2\pi$ and envelope $y_e(t)$, the intensity is given by

$$|y_e(t)|^2 = 2 \int \left[\int_{-\infty}^{\infty} Y_+^*(f) Y_+(f + \alpha)\, df \right] \exp (-i2\pi\alpha t)\, d\alpha,$$

$$\Leftrightarrow 2 \int_{-\infty}^{\infty} Y_+^*(f) Y_+(f + \alpha)\, df.$$

Define

$$Y_e(f) = \int_{-\infty}^{\infty} y_e(t) \exp (i2\pi f t)\, dt, = \sqrt{2}\, Y_+(f + f_0),$$

[provided $y_e(t)$ is narrowband compared with f_0];

$$|y_e(t)|^2 = \int_{-\infty}^{\infty} Y_e(f) \exp (-2i\pi f t) Y_e^*(f') \exp (i2\pi f' t)\, df\, df',$$

$$= \int_{-\infty}^{\infty} \int \exp [-i2\pi(f - f')t][Y_e(f) Y_e^*(f')]\, df\, df',$$

$$= \int_{-\infty}^{\infty} \int^{\infty} \exp (-i2\pi\gamma t)[Y_e^*(f') Y_e(f' + \gamma)]\, d\gamma\, df',$$

where $\gamma = f - f'$;

$$= 2 \int_{-\infty}^{\infty} \exp (-i2\pi\gamma t)[Y_+^*(g) Y_+(g + \gamma)\, d\gamma\, dg]$$

where $g = (f' + f_0)$.

Q.E.D.

REFERENCES

1. Marcuse, D., "Mode Conversion Caused By Surface Imperfections In A Dielectric Slab Waveguide," B.S.T.J., *48*, No. 10 (December 1969), pp. 3187–3215.
2. Davenport, W. L., and Root, W. B., *Random Signals and Noise*, New York: McGraw-Hill, 1958, pp. 61–62.
3. Young, D. T., and Rowe, H. E., "Impulse Response Of A Two Mode Random Guide," Unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

Anisotropic Scattering Due to Rain at Radio-Relay Frequencies

By DAVID E. SETZER

(Manuscript received October 30, 1970)

L. T. Gusler and D. C. Hogg have estimated that interference coupling due to rain is a significant factor in coordinating the shared use of frequencies between satellite-communications and terrestrial microwave radio-relay systems. Their calculations are based on the assumption that rain scatters isotropically. Calculations given here, based on the exact anisotropic angular scattering functions, do not alter their conclusions. The exact scattering patterns at selected frequencies in the range 1.4 to 300 GHz for rains whose drop sizes obey the Laws-Parsons distribution are presented for the range of rain rates 1 to 150 mm/hr.

I. INTRODUCTION

Recently L. T. Gusler and D. C. Hogg[1] calculated the degree of coupling between satellite-communications and terrestrial radio-relay systems due to scattering by rain at frequencies of 4, 6, 11, 18.5 and 30 GHz. In their work they took the scattering by raindrops to be isotropic and to be based on the Laws and Parsons drop-size distribution.[2] They have pointed out correctly that it is known that raindrops do not scatter isotropically. Since the isotropic assumption is incorrect, it is of interest to examine the magnitude of the resulting error. Also, for the purposes of documentation, we present the scattering patterns at selected frequencies in the range 1.4 to 300 GHz at rain rates in the range 1 to 150 mm/hr.

It has been shown that the Mie solution to the problem of scattering from a single sphere can be used to generate the Stokes scattering matrix for atmospheric type aerosols, such as rain, and in particular for Laws and Parsons type rains.[3,4] One of the elements of this matirx is the angular scattering function $[P_1(\theta) + P_2(\theta)]$, sometimes referred to as the normalized Mie intensity function. Here, θ is the scattering angle (0° to 180°) with zero representing the forward direction. Since the aerosol is assumed to be made up of spherical particles, the scattering is symmetric about the direction of propagation; therefore, the total scattering field can be described in terms of the single variable θ. Essentially, the function $[P_1(\theta) + P_2(\theta)]$ is a prescription for the relative amount of energy scattered into a differential solid angle in the direction θ. For isotropic scattering,

$$P_1(\theta) + P_2(\theta) = 2, \tag{1}$$

and for Rayleigh scattering,

$$P_1(\theta) + P_2(\theta) = \tfrac{3}{2}(1 + \cos^2 \theta). \tag{2}$$

Generally, as the ratio of particle circumference to the wavelength of the scattered energy becomes small, the scattering function $[P_1(\theta) + P_2(\theta)]$ approaches the Rayleigh formula.

II. COMPUTER PROGRAM

We devised a computer program which was used to generate the angular scattering functions for Laws and Parsons rains of 1, 50, 100 and 150 mm/hr at 1.4, 2, 3, 6, 16, 30, 60, 100, 150 and 300 GHz.[4] The results are plotted on Figs. 1 through 4 along with plots for isotropic and Rayleigh scattering. The Laws and Parsons rains are described in Table I.

Figures 1 through 4 show that the scattering at 1.4, 2 and 3 GHz is described by the Rayleigh function. Also, for many applications it appears that the assumption of Rayleigh scattering at 6, 16 and 30 GHz will be in order; however, the scattering patterns at 60, 100, 150 and 300 GHz deviate greatly from the Rayleigh function. Note that the scattering diagrams are only mildly dependent on rain rate, the most noticeable feature being the increase in the forward peak with increase in rain rate. It is to be expected that, in general, for a Laws and Parsons rain, the angular scattering function will not be greatly influenced by rain rate. This is because the angular scattering function of an aerosol is determined by the shape of the particle size distribution and the mean particle size, rather than the density of particles. As can be seen by examining Table I, the mean particle size and the general distribution shape are not very different over the range of rain rates presented.

Finally, note that none of the scattering patterns can be described as isotropic. However, in the range of frequency considered by Gusler

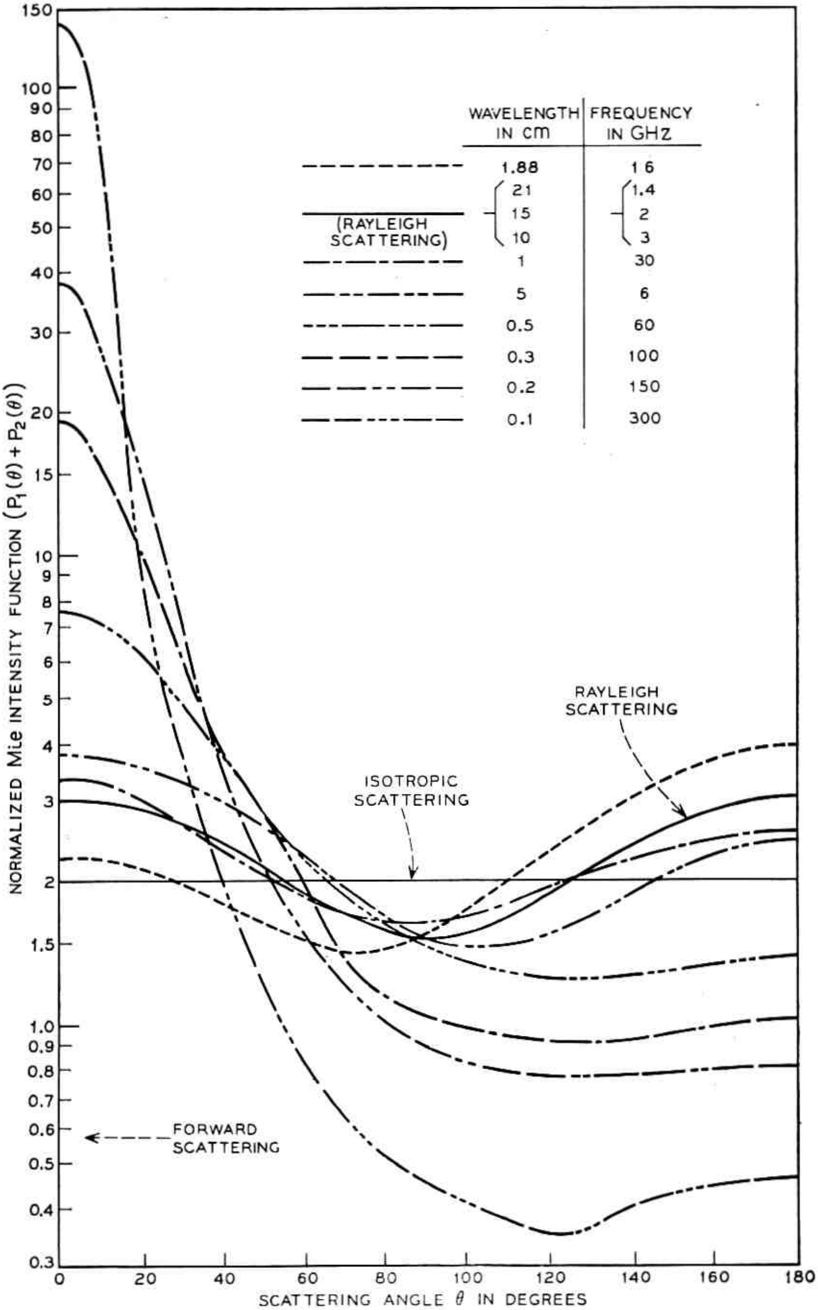


Fig. 1—Intensity versus scattering angle, rain rate = 150 mm/hr.

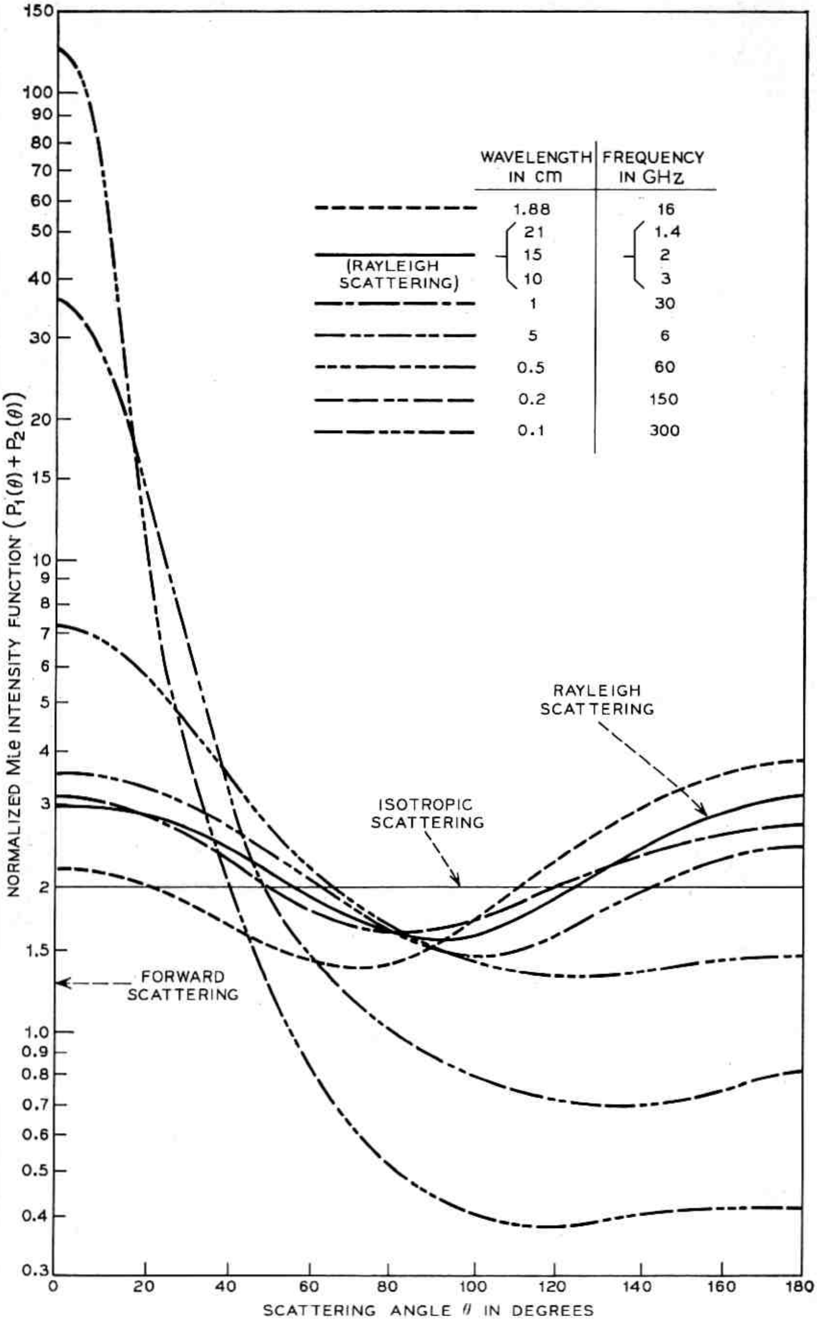


Fig. 2—Intensity versus scattering angle, rain rate = 100 mm/hr.

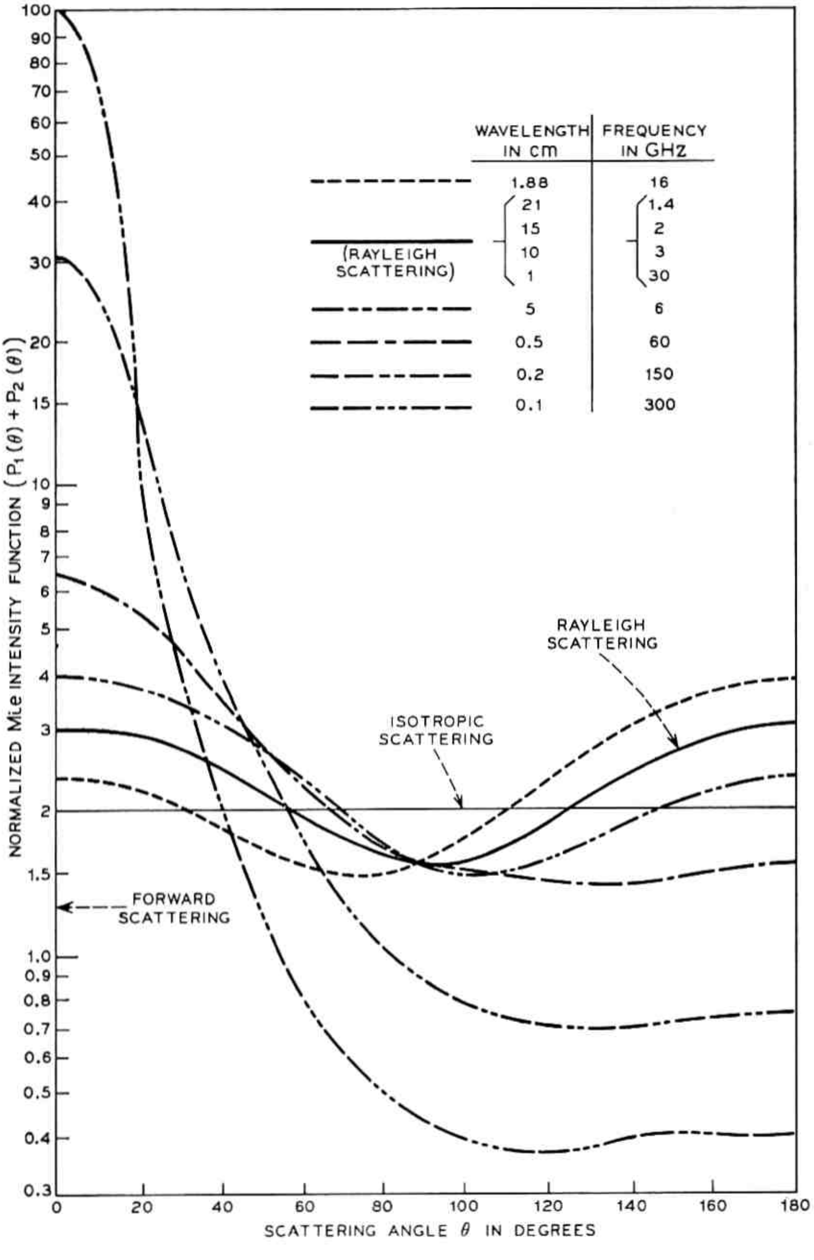


Fig. 3—Intensity versus scattering angle, rain rate = 50 mm/hr.

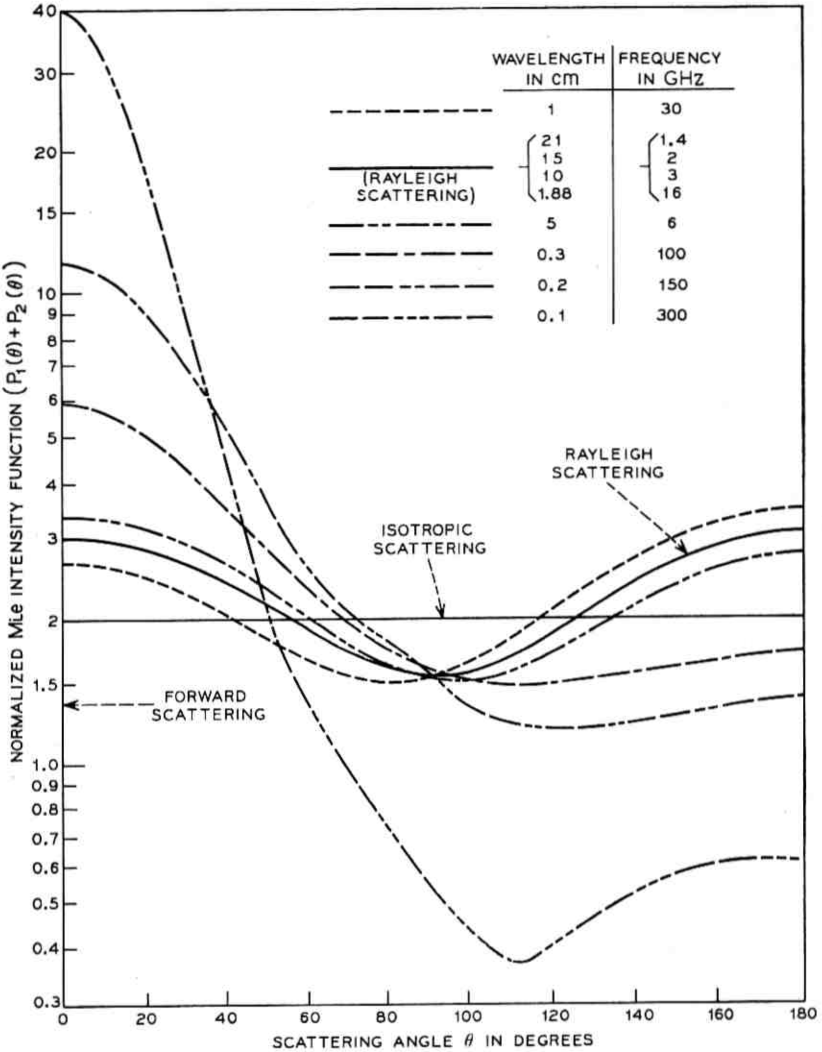


Fig. 4—Intensity versus scattering angle, rain rate = 1 mm/hr.

and Hogg, 4 to 30 GHz, the actual scattering functions are everywhere within a factor of two (3 dB) of the isotropic function. Since the Gusler-Hogg scattering model predicts an interference level between radio relay and satellite systems proportional to the scattering function, the solutions can be in error by at most 3 dB. It is important

TABLE I—LAWS AND PARSONS DROP-SIZE DISTRIBUTIONS FOR VARIOUS PRECIPITATION RATES

Drop Diameter (cm)	Rain Rate (mm/hour)								
	0.25	1.25	2.5	5	12.5	25	50	100	150
	Percent of Total Volume								
0.05	28.0	10.9	7.3	4.7	2.6	1.7	1.2	1.0	1.0
0.1	50.1	37.1	27.8	20.3	11.5	7.6	5.4	4.6	4.1
0.15	18.2	31.3	32.8	31.0	24.5	18.4	12.5	8.8	7.6
0.2	3.0	13.5	19.0	22.2	25.4	23.9	19.9	13.9	11.7
0.25	0.7	4.9	7.9	11.8	17.3	19.9	20.9	17.1	13.9
0.3		1.5	3.3	5.7	10.1	12.8	15.6	18.4	17.7
0.35		0.6	1.1	2.5	4.3	8.2	10.9	15.0	16.1
0.4		0.2	0.6	1.0	2.3	3.5	6.7	9.0	11.9
0.45			0.2	0.5	1.2	2.1	3.3	5.8	7.7
0.5				0.3	0.6	1.1	1.8	3.0	3.6
0.55					0.2	0.5	1.1	1.7	2.2
0.6						0.3	0.5	1.0	1.2
0.65							0.2	0.7	1.0
0.7									0.3

to remember that here we are speaking only of the error due to invoking the isotropic scattering assumption. Nothing is intended to be said about any other aspect of the model and, most importantly, this work does not imply anything about the magnitude of the scattering cross section other than that the Rayleigh cross section is probably a reasonable assumption at those wavelengths where the angular scattering function is Rayleigh. For a detailed look at the Mie cross sections for Laws and Parsons rains see Ref. 4.

III. SUMMARY

Gusler and Hogg have estimated maximum coupling between satellite communications and terrestrial radio-relay systems due to scattering by rain to be of the order of −150 dB. A 3-dB uncertainty in those results is not really significant, considering all of the other uncertain aspects of modeling such a complicated physical problem. In summary, the results of Gusler and Hogg are not significantly altered if the exact angular scattering functions rather than isotropic scattering functions are used in their calculations.

REFERENCES

1. Gusler, L. T., and Hogg, D. C., "Some Calculations on Coupling Between Satellite Communications and Terrestrial Radio-Relay Systems Due to Scattering by Rain," B.S.T.J., *49,* No. 7 (September 1970), pp. 1491–1511.

2. Medhurst, R. G., "Rainfall Attenuation of Centimeter Waves: Comparison of Theory and Measurement," IEEE Trans. Antennas and Propagation, *13,* No. 4 (July 1965), pp. 550–564.
3. Setzer, D. E., "Comparison of Measured and Predicted Aerosol Scattering Functions," Applied Optics, *8,* No. 5 (May 1969), pp. 905–911.
4. Setzer, D. E., "Computed Transmission Through Rain at Microwave and Visible Frequencies," B.S.T.J., *49,* No. 8 (October 1970), pp. 1873–1892.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

Short-Term Frequency Stability of Precision Oscillators and Frequency Generators

By Raymond E. Barber

(Manuscript received August 3, 1970)

We present in this paper two definitions of short-term frequency stability: (i) time domain, the expected value of the variance of the fractional frequency fluctuations from nominal frequency, $\langle \sigma_\nu^2(N, T, \tau) \rangle$, *in which* N *is the number of samples,* T *is the averaging time plus the dead time between samples, and* τ *is the averaging time; and (ii) frequency domain, the power spectral density of the fractional frequency departure from nominal frequency,* $S_\nu(f)$. *We discuss the topics of conversion from the frequency domain to the time domain and conversion among time domain measures.*

All measurements were made in the time domain, using period counting techniques. An oscillator was offset in frequency by using a specially built quartz crystal unit plated for 10 kHz below the frequency of the other sources. This oscillator was used to obtain the beat frequency required by the period counting approach.

Since the use of $\langle \sigma_\nu^2(2, T, \tau) \rangle$ *as a measure of short-term stability has significant advantages over* $\langle \sigma_\nu^2(N, T, \tau) \rangle$, *the relationship between the two quantities was investigated. For averaging times of one second or less,* $\langle \sigma_\nu^2(2, T, \tau) \rangle$ *and* $\langle \sigma_\nu^2(N, T, \tau) \rangle$ *were almost equal.*

The short-term stability of several quartz crystal oscillators and precision frequency generators was measured. The stability over the shorter averaging times was nearly equal for most of the oscillators. At longer times, the stability of each oscillator was unique. The frequency generators demonstrated similar stability over averaging times of 10 and 100 milliseconds, but were unique elsewhere. The accuracy of all measurements was limited by systematic effects from the environment and the measurement equipment itself.

I. INTRODUCTION

When the phrase "short-term frequency stability" is mentioned, one can immediately infer that frequency varies with time and that a short time interval is involved. Many questions are, at this point, unanswered: How short is the interval? How is stability defined? What system malfunctions can result from excessive instability? What definitions are appropriate for the specific application and why?

Some answers to these questions are developed here; others must be determined by the user in terms of the specific system involved. If a measurement of short-term frequency stability is to be of significant use, each of these questions must be resolved.

Although ever-increasing interest in short-term stability has existed for 25 or 30 years, no universally accepted definition of short-term stability exists. Primarily, early research in this area was extremely application-oriented. Little general work was performed with the result that many definitions were developed. Individuals and organizations defined and used stability in terms of their own applications. The resultant confusion showed that a more general definition applicable to the majority of cases was desirable in the topic of short-term frequency stability.

Today, many users talk about stable oscillators without understanding why or even if such precision is necessary. Manufacturers add to the present confusion through lack of rigor in their performance specifications. That is, they often merely quote a number without defining what, how, or why they are measuring the stability of their equipment.

Sophisticated systems exist today which require more precision than frequency standards of 10 to 15 years ago. Only recently has the precision oscillator been liberated from its position in secluded portions of carefully controlled laboratories.[1] Spacecraft applications suffice as good examples of the strides made in this area; wide temperature variations, severe mechanical vibrations, and varying oscillator voltage are often encountered. In the immediate future, these requirements will tighten even further. A proposed collision avoidance system for aircraft will require stable oscillators. With higher and higher data rates in communication systems, even the dependable telephone office will contain precision oscillators. As precision oscillators come into general use, accepted measurement theory and techniques are absolutely necessary.

In view of the above, two definitions of frequency stability are presented in this paper. One definition is in the time domain; the other is in the frequency domain. Conversion between the domains, an extremely important subject, is also treated. Every attempt has been made to be consistent with the IEEE Subcommittee on Frequency Stability.* It is anticipated that the subcommittee will publish its formal definition sometime in 1971.

II. GENERAL DISCUSSION

2.1 *General Definitions*

The general definition of instantaneous angular frequency is the time-rate of change of phase. That is

$$\omega(t) = \frac{d\Phi}{dt} = \dot{\Phi}. \tag{1}$$

An oscillator output signal may be described as

$$g(t) = [A + \epsilon(t)] \cos [\omega t + \phi(t)], \tag{2a}$$

$$g(t) = [A + \epsilon(t)] \cos [2\pi F t + \phi(t)], \tag{2b}$$

where

ω = nominal angular frequency of the oscillator,
$F = \omega/2\pi$ = nominal frequency in hertz of the oscillator,
A = nominal amplitude of the oscillator,
$\epsilon(t)$ = small, slowly time-varying amplitude fluctuations, and
$\phi(t)$ = slowly time-varying real function (phase).

Here, $g(t)$ may be considered as either a voltage or a current. If the oscillator is to qualify as a *precision* oscillator, it is required that

$$\left|\frac{\epsilon(t)}{A}\right| \ll 1, \tag{3}$$

$$\left|\frac{\dot{\phi}(t)}{2\pi F}\right| \ll 1. \tag{4}$$

2.2 *Statistical Processes*

For the present, assume that the variables of equation (2) are random processes. It is generally assumed in the study of frequency stability that amplitude deviations, $\epsilon(t)$, do not directly affect frequency or

* This subcommittee, formed in 1964 as a result of the IEEE-NASA Symposium on short-term frequency stability, is formally named the Subcommittee on Frequency Stability of the Technical Committee on Frequency and Time of the Group on Instrumentation and Measurement of the IEEE.

phase.[2] It must be stressed that $\phi(t)$ presumably contains all frequency and phase fluctuations and is the main quantity considered.

If a random process is stationary in the strict sense, it is unaffected by translations of the origin for time.[3] This implies that the probability distribution of values in the ensemble will be the same at any two instants. Hence, examination of the ensemble yields no data as to which instant of time the examination occurred.

Texts on random noise and stochastic processes assume that most noise functions may be represented or approximated as stationary gaussian random processes with zero averages.[4] The justification for assuming a gaussian distribution lies in the central limit theorem. This well-known theorem states that the distribution of the sum of a large number of independent random variables will approach a gaussian distribution. S. O. Rice showed that noise does, in general, conform to a gaussian distribution, provided that a sufficiently large sample is taken.[5] Zero averages of each random variable are assumed for convenience.

A process can be defined as ergodic when the statistics of one system over an infinite period of time are equivalent to the statistics of an infinite ensemble of systems at any given instant. (Note that stationarity is a necessary but not sufficient condition for ergodicity.) Since the processes are stationary and gaussian, the noise functions considered in this paper are assumed ergodic.

In the real world, no process can be stationary in the strict sense. To be stationary, the probability distributions across the ensemble must be the same for any selected instant of time. This stipulation includes time as it approaches infinity. If the process is terminated at some future time, the concept of stationarity in the strict sense is violated. The random process, X_t, is called stationary in the wide sense, if the first and second order moments of its random variables exist and satisfy

$$E(X_t) = \text{constant},$$

$$\sigma^2(X_t) = \text{constant},$$

$$E\{[X_{t+\tau} - E(X_{t+\tau})][X_t - E(X_t)]\} = R(\tau),$$

where

$R(\tau)$ = autocorrelation function, and
τ = some arbitrary finite time interval.

The actual verification of stationarity is not feasible. The main requirement is that the physical process be *consistent* with the concept of stationarity. That is, if measurements are to be of any use, they must reasonably describe the process at all times in the future, prior to termination of the process.

2.3 *Systematics*

When a study is made of a precision oscillator or frequency generator, great pains are often taken to isolate the circuits from electrical disturbances. This is done to minimize the systematic effects so that only the truly random noise can be observed. For evaluation of the generator itself, this is a sound practice. However, in the real world, electrical disturbances do exist. For example, in a data processing system there are card readers, magnetic tape transports, and other devices. Each of these devices has numerous electro-mechanical relays which are continually chattering. Thus, it is desirable to insure that a frequency generator has the necessary stability. This obviously means it must demonstrate adequate stability in the operating environment with the contribution of the systematic effects.

In equation (2), it was assumed that the main contribution of $\phi(t)$ came from the oscillator. Now it is necessary to loosen this assumption to include the contributions to $\phi(t)$ from the operating environment. While these added uncertainties are not caused by the oscillator, any attempt to observe $\phi(t)$ will include these effects. Therefore, $\phi(t)$ in equation (2) can be considered to include

$$\phi(t) = c(t) + s(t) + n(t), \tag{5}$$

where

$c(t)$ = phase due to aging (drift),
$s(t)$ = phase due to environmental systematic effects, and
$n(t)$ = phase due to oscillator random noise.

The contributions of $c(t)$ are strictly long term, affecting time intervals of a thousand seconds or longer. Hence, for time periods of less than a thousand seconds, $c(t)$ becomes almost constant and thus is not significant. The exact composition of $s(t)$ is unknown, although it has both long- and short-term effects. Again, the long-term component of the systematic effects becomes insignificant for short time periods. For these short intervals, $\phi(t)$ becomes

$$\phi(t) = s_n(t) + n(t), \tag{6}$$

where

$s_n(t)$ = short-term components of the systematic effects.

For the present, it will be assumed that the short-term component of the systematic effects is a random process which is stationary at least in the wide sense: gaussian, with zero averages, and ergodic. This stipulation may later fall as more is learned about the nature of systematic effects. Certainly, even if not truly random, the effects will appear as short-term instabilities.

Any environment will contribute a term $s(t)$ to (5). Obviously, if the systematics, $s(t)$, can be minimized to the point where their contribution is several orders of magnitude less than the oscillator noise, $n(t)$ becomes the only significant measurable. On the other hand, if the environment is extremely noise contaminated, the actual oscillator instabilities may be "swamped" by the systematic effects. Conditions can exist that are so contaminated that coherent operation of many electronic systems is not possible. The systematic effects will contribute to the short-term instabilities in an additive manner on a power basis. If an estimate of the short-term stability of an oscillator or frequency generator can be obtained in a carefully controlled, low-noise environment and if a subsequent estimate can be made in an operating atmosphere, the resulting instabilities may be compared. The contribution of the noise due to the systematic effects of the environment may then be described.

III. DEFINITIONS

3.1 *General*

Many applications exist that necessitate the use of highly stable oscillators or frequency generators. Due to the accuracy requirements of these various applications, some measure of oscillator instability is imperative. Although many persons view such a quantity merely as a vehicle for oscillator selection and comparison between oscillators, these uses are only academic. Strictly speaking, a measure of instability enables prediction of oscillator performance or, if not sufficiently stable, nonperformance in a given system. More precisely, *any* deviation from the intended frequency will degrade the performance of any system. Specific cases of the effects of frequency stability on an operation are discussed by J. A. Barnes.[6, 7]

Presented below are definitions of short-term frequency stability both in the time and frequency domains. Translation between the domains is also discussed.

3.2 *Time Domain*

In this paper, short-term frequency stability in the time domain is defined as the ensemble average of the variance of fractional frequency fluctuations from nominal frequency. This definition is based on the work of D. W. Allan.[8]

It is convenient to make the following definition

$$y(t) = \frac{\dot{\phi}(t)}{2\pi F_1}, \tag{7}$$

where

$\dot{\phi}(t) = d\phi/dt = \omega(t)$ {from equations (1) and (2)}, and
F_1 = long-term average frequency, hertz, of the oscillator {from equation (2)}.

Taking the short-term average value of equation (7)

$$\bar{y}_n = \frac{1}{\tau}\int_{t_n}^{t_n+\tau} y(t)\,dt = \frac{\phi(t_n+\tau) - \phi(t_n)}{2\pi F_1\tau}, \tag{8}$$

$$\bar{y}_k = \frac{1}{\tau}\int_{t_k}^{t_k+\tau} y(t)\,dt = \frac{\phi(t_k+\tau) - \phi(t_k)}{2\pi F_1\tau}, \tag{9}$$

where $n, k = 1, 2, \cdots, m, \cdots$.

Using equations (8) and (9), an expression in terms of the sample variance is obtained

$$\sigma_y^2(N, T, \tau) = \frac{1}{N-1}\sum_{n=1}^{N}\left(\bar{y}_n - \frac{1}{N}\sum_{k=1}^{N}\bar{y}_k\right)^2, \tag{10}$$

where

N = number of samples,
T = time between the beginning of successive sample intervals, and
τ = sample time.

Short-term frequency stability in the time domain is defined as the expected value or ensemble average of equation (10), the variance of fractional frequency fluctuations from nominal frequency. That is

$$\langle\sigma_y^2(N, T, \tau)\rangle = E\left[\frac{1}{N-1}\sum_{n=1}^{N}\left(\bar{y}_n - \frac{1}{N}\sum_{k=1}^{N}\bar{y}_k\right)^2\right], \tag{11}$$

where $E[X]$ denotes the expected value or ensemble average of X. Note that this definition of frequency stability is dimensionless.

The justification of the use of the term $1/(N - 1)$, located directly to the left of the first summation in equations (10) and (11), follows from the theory of estimation (see Ref. 9). Thus, in order to obtain an unbiased estimate of the population variance, the sample variance is multiplied by $N/(N - 1)$, resulting in equations (10) and (11). Clearly, for large samples, the population and sample variances become very nearly equal.

A popular expression of short-term stability in the time domain is the expected value of the standard deviation (rms value) of the fractional frequency fluctuations from nominal frequency. Expressed functionally, this becomes

$$\langle \sigma_y(N, T, \tau) \rangle = E\left\{ \left[\frac{1}{N - 1} \sum_{n=1}^{N} \left(\bar{y}_n - \frac{1}{N} \sum_{k=1}^{N} \bar{y}_k \right)^2 \right]^{\frac{1}{2}} \right\}. \tag{12}$$

It should be noted that the square root of equation (11), $\langle \sigma_y^2(N, T, \tau) \rangle$, is equal to equation (12), $\langle \sigma_y(N, T, \tau) \rangle$, only when the successive samples are truly stationary in the wide sense. If the samples appear only slightly nonstationary, approximate equality may be assumed as

$$(\langle \sigma_y^2(N, T, \tau) \rangle)^{\frac{1}{2}} \cong \langle \sigma_y(N, T, \tau) \rangle. \tag{13}$$

3.3 *Frequency Domain*

L. S. Cutler and C. L. Searle showed that a practical definition of short-term frequency stability can be given in terms of autocorrelation functions and power spectral densities.[10]

The autocorrelation function of phase is defined as

$$R_\phi(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_0^T \phi(t + \tau)\phi(t)\, dt. \tag{14}$$

$R_\phi(\tau)$ can also be determined using numerical methods on a digital computer if enough samples can be taken, such as

$$R_\phi(\tau) = \frac{1}{N} \sum_{k=1}^{N} \phi_k(t + \tau)\phi_k(t), \tag{15}$$

where N is some large positive integer.

Let $R_\phi(\tau)$ be the autocorrelation function of a process which is stationary in the wide sense and continuous. From the Wiener-Khintchine theorem, it is known that autocorrelation functions and power spectral densities are Fourier transform pairs. Hence, the two-sided spectral density of phase becomes[10]*

$$S_\phi(f) = \int_{-\infty}^{\infty} R_\phi(\tau) e^{-j2\pi f\tau}\, d\tau. \tag{16}$$

From this, the one-sided spectral density of phase, $S_\phi(f)$, becomes[6]

$$S_\phi(f) = 2\int_0^\infty R_\phi(\tau)\cos(2\pi f\tau)\, d\tau, \tag{17}$$

$$R_\phi(\tau) = 2\int_0^\infty S_\phi(f)\cos(2\pi f\tau)\, df. \tag{18}$$

In normal Fourier analysis,[4] differentiation in the time domain corresponds to multiplication by $j2\pi f$ in the frequency domain. In terms of power spectral densities, this becomes $(2\pi f)^2$. Thus

$$S_{\dot\phi}(f) = (2\pi f)^2 S_\phi(f). \tag{19}$$

But $S_{\dot\phi}(f)$ is the power spectral density of frequency fluctuations. To obtain the power spectral density of instantaneous fractional frequency departure from nominal frequency, equation (19) must be normalized

$$S_y(f) = \frac{S_{\dot\phi}(f)}{(2\pi F_1)^2} = \frac{f^2}{F_1^2} S_\phi(f). \tag{20}$$

Thus, $S_y(f)$, equation (20), is the definition of short-term frequency stability in the frequency domain. Note that $S_y(f)$ has the units "per hertz."

3.4 *Translations Between Domains*

3.4.1 *Frequency Domain to Time Domain*

L. S. Cutler has shown that the following equation allows calculation of the time-domain stability from the frequency domain[6]

$$\langle \sigma_y^2(N, T, \tau)\rangle = \frac{N}{(N-1)}\int_0^\infty S_y(f)\frac{[\sin^2(\pi f\tau)]}{(\pi f\tau)^2}\left[1 - \frac{\sin^2(\pi r f N\tau)}{N^2\sin^2(\pi r f\tau)}\right] df, \tag{21}$$

where

$$r = T/\tau.$$

* In this paper, the convention is used that the term F_1, as in equations (2) and (7), is the cycle frequency, hertz, of the oscillator. The term f is a Fourier frequency, hertz, and is not a function of time.

3.4.2 *Translations Among Time-Domain Measures*

In certain applications, translation among time-domain measures is of interest. The following method, presented by Barnes,[11] allows calculation of $\langle \sigma_y^2(N_2, T_2, \tau_2) \rangle$, given an estimate of $\langle \sigma_y^2(N, T, \tau) \rangle$, for functions which have a power law spectral density.

Since most precision oscillators and frequency generators have power law spectral densities, it is possible to define two bias functions, $B_1(N, r, \mu)$ and $B_2(r, \mu)$, as follows

$$B_1(N, r, \mu) \equiv \frac{\langle \sigma_y^2(N, T, \tau) \rangle}{\langle \sigma_y^2(2, T, \tau) \rangle}, \tag{22}$$

$$B_2(r, \mu) \equiv \frac{\langle \sigma_y^2(2, T, \tau) \rangle}{\langle \sigma_y^2(2, \tau, \tau) \rangle}, \tag{23}$$

where

μ = spectral type (related to the shape of the spectrum. In Ref. 11, Barnes shows how μ may be found, given B_1).

Now an estimate of $\langle \sigma_y^2(N_2, T_2, \tau_2) \rangle$ can be made

$$\langle \sigma_y^2(N_2, T_2, \tau_2) \rangle = \left(\frac{\tau_2}{\tau_1} \right)^{\mu} \frac{B_1(N_2, r_2, \mu) B_2(r_2, \mu)}{B_1(N_1, r_1, \mu) B_2(r_1, \mu)} \langle \sigma_y^2(N_1, T_1, \tau_1) \rangle. \tag{24}$$

See Ref. 11 for a listing of the bias functions for various spectral types, number of samples, and r.

IV. MEASUREMENT TECHNIQUES

4.1 *Period Counting Technique*

The basic functional description of this method is shown in Fig. 1. This arrangement is similar to that presented in Ref. 2. Two similar oscillators are offset in frequency. This offset can be produced by adjustment of the tuning control on the oscillator or use of a crystal which produces a slightly different average frequency. It is assumed that these procedures change only the long-term average frequency of the oscillators and that the short-term instabilities are not affected.

As in equation (2), the output signals from these similar oscillators are fed to a mixer or product detector. The resulting output is fed to a low pass filter. This yields an expression:

$$g_0(t) \cong \frac{A_1 A_2}{2} \{ \cos [2\pi (F_1 - F_2)t + (\phi_1(t) - \phi_2(t))] \}. \tag{25}$$

This signal is the nominal frequency difference between the sources plus the instabilities of each. (For a more detailed discussion, see Refs. 8, 10 or 12.) Define

$$F_0 = F_1 - F_2 ,$$

$$\Phi(t) = [\phi_1(t) - \phi_2(t)].$$

Substituting these values into equation (25)

$$g_0(t) \cong \frac{A_1 A_2}{2} \cos [2\pi F_0 t + \Phi(t)]. \tag{26}$$

This quantity is fed directly to a digital counter.

The theoretical time, τ_0, between the first and Mth positive-going zero crossings, if the signals were ideal, is

$$\tau_0 = \frac{M}{F_0}.$$

The actual elapsed time, τ, between the first and Mth zero crossing is

$$\tau = \frac{M}{F_0} - \frac{[\Phi(t_0 + \tau) - \Phi(t_0)]}{2\pi F_0}. \tag{27}$$

The uncertainties are characterized as

$$\delta\tau = \frac{\Phi(t_0 + \tau) - \Phi(t_0)}{2\pi F_0}.$$

Therefore, equation (27) becomes

$$\tau = \tau_0 - \delta\tau. \tag{28}$$

Because of equation (4), $\delta\tau$ is small. Cutler and Searle show that if

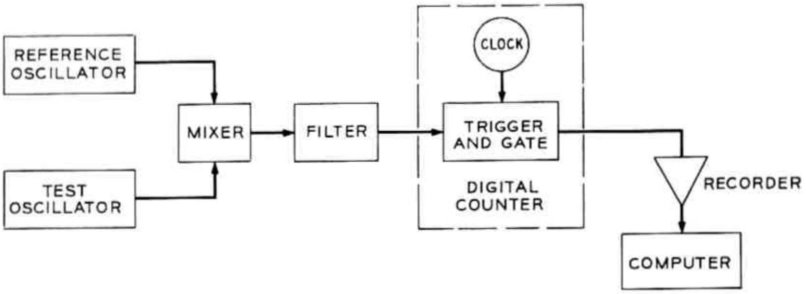


Fig. 1—Fundamental description, period counting technique.

variations of $\delta\tau$ are small, the following may be assumed

$$\Phi(t_0 + \tau) \cong \Phi(t_0 + \tau_0), \tag{29}$$

substituting

$$\delta\tau \cong \frac{\Phi(t_0 + \tau_0) - \Phi(t_0)}{2\pi F_0}. \tag{30}$$

Substituting equation (30) into equations (8) and (9) gives

$$\bar{y}_n \cong \frac{F_0}{F_1\tau_0}\left[\frac{\Phi(t_{0n} + \tau_0) - \Phi(t_{0n})}{2\pi F_0}\right] = \frac{F_0}{F_1\tau_0}\,\delta\tau_n\,, \tag{31}$$

$$\bar{y}_k \cong \frac{F_0}{F_1\tau_0}\left[\frac{\Phi(t_{0k} + \tau_0) - \Phi(t_{0k})}{2\pi F_0}\right] = \frac{F_0}{F_1\tau_0}\,\delta\tau_k\,. \tag{32}$$

For each sampling sequence, N estimates of $\bar{y}_n$ and $\bar{y}_k$ are computed. Substituting these values into equation (11)

$$\langle\sigma_v^2(N,\,T,\,\tau)\rangle = E\left[\frac{1}{N-1}\sum_{n=1}^{N}\left(\frac{F_0}{F_1\tau_0}\,\delta\tau_n - \frac{1}{N}\cdot\sum_{k=1}^{N}\frac{F_0}{F_1\tau_0}\,\delta\tau_k\right)^2\right], \tag{33}$$

$$\langle\sigma_v^2(N,\,T,\,\tau)\rangle = E\left[\left(\frac{1}{N-1}\right)\frac{F_0^2}{F_1^2\tau_0^2}\sum_{n=1}^{N}\left(\delta\tau_n - \frac{1}{N}\sum_{k=1}^{N}\delta\tau_k\right)^2\right]. \tag{34}$$

4.2 *A Special Case*

When $N = 2$, equation (11) is simplified

$$\langle\sigma_v^2(2,\,T,\,\tau)\rangle = E\left[\sum_{n=1}^{2}\left(\bar{y}_n - \frac{1}{2}\sum_{k=1}^{2}\bar{y}_k\right)^2\right]. \tag{35}$$

Likewise, as in equation (34),

$$\langle\sigma_v^2(2,\,T,\,\tau)\rangle = E\left[\frac{F_0^2}{F_1^2\tau_0^2}\sum_{n=1}^{2}\left(\delta\tau_n - \frac{1}{2}\sum_{k=1}^{2}\delta\tau_k\right)^2\right]. \tag{36}$$

Recall that $\tau = \tau_0 - \delta\tau$

$$\langle\sigma_v^2(2,\,T,\,\tau)\rangle = E\left[\frac{F_0^2}{F_1^2\tau_0^2}\sum_{n=1}^{2}\left\{(\tau_0 - \tau_n) - \frac{1}{2}\sum_{k=1}^{2}(\tau_0 - \tau_k)\right\}^2\right]. \tag{37}$$

Expanding the right side of this equation and reducing

$$\langle\sigma_v^2(2,\,T,\,\tau)\rangle = E\left[\frac{F_0^2}{F_1^2\tau_0^2}\,\frac{(\tau_2 - \tau_1)^2}{2}\right]. \tag{38}$$

Let $\tau_2 - \tau_1 = \Delta\tau_0$

$$\langle\sigma_v^2(2,\,T,\,\tau)\rangle = E\left[\frac{F_0^2}{2F_1^2}\left(\frac{\Delta\tau_0}{\tau_0}\right)^2\right]. \tag{39}$$

The following term often appears in the literature on frequency stability[13]

$$\frac{\Delta F}{F} = \frac{\Delta t}{T}, \tag{40}$$

where

ΔF = frequency difference between received and local standards (hertz),
F = nominal frequency,
Δt = accumulated time error corresponding to change in phase, and
T = averaging time.

The expression in equation (40) is frequently used in calculating the long-term stability or drift rate of an oscillator. The application of equation (40) has been well documented in many sources. The difference between two readings of equation (40) is recognized as the definition of long-term stability, yielding a peak-to-peak value of frequency drift over a specified time period.

As the time interval is compressed, the oscillator instabilities demonstrate more randomness while being less subject to drift. In the limit, equation (40) yields an instantaneous peak-to-peak value of the fractional frequency fluctuations from nominal frequency. Letting $t = \tau$ and substituting equation (40) into equation (39),

$$\langle \sigma_y^2(2, T, \tau) \rangle = E\left[\frac{F_0^2}{2F_1^2}\left(\frac{\Delta F_0}{F_0}\right)^2\right], \tag{41a}$$

$$\langle \sigma_y^2(2, T, \tau) \rangle = E\left[\frac{1}{2F_1^2}(\Delta F_0)^2\right], \tag{41b}$$

$$\langle \sigma_y^2(2, T, \tau) \rangle = E\left[\frac{1}{2F_1^2}(F_{02} - F_{01})^2\right], \tag{41c}$$

where

F_{01}, F_{02} = two successive samples of the beat frequency, F_0, over an averaging time, τ.

In a similar manner, the standard deviation becomes

$$\langle \sigma_y(2, T, \tau) \rangle = E\left\{\frac{1}{\sqrt{2}\,F_1}[(F_{02} - F_{01})^2]^{1/2}\right\}. \tag{42}$$

In Section 3.4.2, it was shown that, given the spectral type, μ, translations between time domain measures can be made. Recalling equation (22)

$$B_1(N, r, \mu) = \frac{\langle \sigma_v^2(N, T, \tau)\rangle}{\langle \sigma_v^2(2, T, \tau)\rangle}.$$

It is obvious that if

$$\langle \sigma_v^2(N, T, \tau)\rangle = \langle \sigma_v^2(2, T, \tau)\rangle, \tag{43}$$

then

$$B_1(N, r, \mu) = 1. \tag{44}$$

Barnes showed that for $r > 1$, equation (44) holds in two cases, when $\mu = -1.00$ and -2.00.[11] For $r = 1$, equation (44) holds for $\mu = -1.00$ only. Allan showed that when the spectral type, μ, is equal to minus one, white noise frequency modulation is present.[8] When μ equals minus two, flicker noise phase modulation is indicated.

As shown by E. A. Gerber and R. A. Sykes,[14] and Cutler and Searle,[10] there are three main sources of noise in oscillators. These are:

(*i*) additive noise (added to signal),

(*ii*) thermal and shot noise (perturbs oscillation), and

(*iii*) flicker ($1/f$) noise frequency modulation. (See Ref. 15 for a discussion of flicker noise.)

For very short time intervals, the additive (white) noise predominates. For longer intervals, flicker noise frequency modulation prevails.[16] Thermal and shot noise are overpowered by the other two types.[10]

For oscillators, μ normally assumes two values

$$\mu \cong 0,$$

$$\mu \cong -1.$$

When the latter is true, $B_1(N, r, \mu) \cong 1$ and $\langle \sigma_v^2(N, T, \tau)\rangle \cong \langle \sigma_v^2(2, T, \tau)\rangle$. When $\mu \cong 0$, however, this does not hold.

There are many indications that the most basic parameter of frequency stability in the time domain is $\langle \sigma_v^2(2, \tau, \tau)\rangle$ (no dead time between samples). Initially, there is no guarantee that $\langle \sigma_v^2(\infty, T, \tau)\rangle$ will converge. Secondly, even if it were possible to assume convergence, it is not practical to take enough samples at the longer intervals to assure meaningful results. Next, it can be shown[6] that $\langle \sigma_v^2(2, \tau, \tau,)\rangle$ converges even for divergent $\langle \sigma_v^2(N, T, \tau)\rangle$. Therefore, the greatest significance of $\langle \sigma_\nu^2(2, T, \tau) \rangle$ as a measure of short-term frequency stability is that it eliminates the embarrassing divergence of flicker noise contributions as $N \rightarrow \infty$ in $\langle \sigma_\nu^2(N, T, \tau) \rangle$.

In addition, some noise functions have extremely long periods, even beyond one cycle per year. These low frequency components affect estimates of $\langle \sigma_\nu^2(N, T, \tau) \rangle$ even though their periods are too long to have any influence on an actual system. As a result, $\langle \sigma_\nu^2(2, \tau, \tau) \rangle$ is more constant in time than $\langle \sigma_\nu^2(N, T, \tau) \rangle$.

In physical applications, $\langle \sigma_\nu^2(2, \tau, \tau) \rangle$ is often more directly applicable than other expressions. For example, in radar it can be shown[6] that the expression for calculating doppler range errors is directly proportional to $\langle \sigma_\nu^2(2, \tau, \tau) \rangle$. In timing, the mean-square second difference of phase is often useful.[7,15] It can be shown[6] that $\langle \sigma_\nu^2(2, \tau, \tau) \rangle$ is directly proportional to the mean-square second difference of phase.

In a practical sense, measurement of $\langle \sigma_\nu^2(2, T, \tau) \rangle$ is easier to achieve than $\langle \sigma_\nu^2(N, T, \tau) \rangle$. Storage requirements for the latter become excessive as N becomes large. Usually, the most practical method of making estimates of $\langle \sigma_\nu^2(N, T, \tau) \rangle$ is the use of a magnetic tape unit for storage of data. The data may then be processed off-line on a special purpose or commercial computer.

For $\langle \sigma_\nu^2(2, T, \tau) \rangle$, storage requirements are small. Neglecting processing requirements, only one summation register is necessary to obtain the ensemble average. Resulting equipment is small and portable. Computations can be made on-line, in real time. Using this approach, reliable measurements of short-term stability in the time domain are practical at field locations.

V. MEASUREMENTS

5.1 *General*

Of prime interest to this paper is the measurement of the short-term frequency stability of a frequency and time of year generator to be used in a large, special purpose computer. This generator consists of a rack of equipment which produces a total of 17 different square wave signals ranging from 9.5367 Hz to 20 MHz. These signals are used as clock frequencies for the computer.

The primary frequency source consists of three 5-MHz quartz crystal oscillators.[1] Two of these oscillators, designated as slave and standby, are phase locked to the other oscillator, called the master. The outputs of the master-slave pair are combined in three digital mixers and fed to three counter chains. Each counter chain output is compared to the corresponding outputs of the other two counters. The two signals that very nearly coincide are combined and fed to external users via cables. The output from the standby oscillator is not used. The standby oscillator is present only because of the five day initial warm-up period. It is phase locked to the master oscillator to speed switching into the system should a failure of one of the other two oscillators occur.

The time of year generator simply uses a 1-MHz combined output and generates a 42-bit BCD parallel time of year code.

A VLF receiver is used to check the 5-MHz oscillators against precision VLF stations. Daily frequency checks reduce the actual frequency error. The specified long-term stability of each oscillator is 1 pp 10^{10} per day. [This is a measure of $\Delta F/F$ as shown in equation (40). The expression 1 pp 10^{10} is a commonly used abbreviation for 1×10^{-10}.] Thus, in theory at least, long-term error (drift) never exceeds that which can be accumulated in a single day.

Provisions have been made to indicate to operating personnel if errors have occurred. These include sensors to detect missing pulses, the loss of an output, counter synchronization errors, or a phase lock circuit approaching its limits. These techniques are present merely to add additional reliability to the system.

The phase-lock circuit permits two oscillators to be phase locked over a frequency change of ± 1 pp 10^8 with a maximum phase error between the two oscillators of less than ± 60 milliradians. The circuit can "capture" over a range of ± 2.5 pp 10^8 with a phase error of ± 160 milliradians.

Whether instabilities are caused directly by an oscillator or by associated circuitry, various problems can result from excessive instabilities. These include (*i*) loss of phase lock, (*ii*) loss of counter chain synchronization resulting in one or more false outputs, (*iii*) degradation of stability delivered to users, and (*iv*) failure of the frequency generator. Loss of phase lock will occur when linear drift or phase errors caused by excessive instability of an oscillator exceed the above-given limits. When phase lock is lost, short-term stability is degraded even further. When a counter chain loses synchronization, the resulting false outputs could have serious consequences to the data processing system.

Once the stability is degraded to some critical level, the data processing system will begin to lose accuracy and resolution; time-of-year errors will accumulate. Pulses propagating through delay lines will not coincide with intended clock pulses. Loss of data and parity errors will result. When the parity count is sufficiently large, a software program will attempt to regain coherence of the data processing system, which will cause destruction of data. Since the generator is the system clock, failure of the generator will cause failure of the system. Possibly the most serious problems occur when the stability is marginal, causing the system to respond with false data, without any indication to operating personnel.

Since the generator is located in the operating environment, all electrical disturbances caused by peripheral equipment add to the instabilities (see Section 2.3). The signals are distributed to users via cables. Each cable, up to 300 feet in length, acts as an "antenna" to the systematic noise, resulting in greater degradation of stability.

It was desirable to determine the short-term frequency stability in several configurations:

(*i*) Oscillator output in a "quiet" environment (sinusoid).

(*ii*) Output at terminals of generator in a "quiet" environment (square wave).

(*iii*) Oscillator output in an operating environment (sinusoid).

(*iv*) Output at terminals of generator in an operating environment (square wave).

(*v*) Generator output at end of a 200-foot cable in an operating environment (square wave).

(*vi*) Generator output at end of a 200-foot cable in a "quiet" environment (square wave).

If a comparison between oscillator stability in a "quiet" and an operating environment is made, an estimate of the noise contributed by the environment is possible. Comparison between the results for the oscillator and the generator yields an estimate of the instabilities contributed by the generator itself.

Comparison of the stabilities of the generator outputs over short and long cables yields an estimate of the degradation due to the cable in the environment. Finally, comparison between the stability at the end of a long cable in both operating and "quiet" environments yields an estimate of the absolute contribution of the cable and another estimate of environmental noise.

Almost all subsystems using outputs from this frequency generator are defined and operated in terms of real time. Therefore, short-term frequency stability is more appropriately defined and measured in the time domain.

5.2 *Measurement Systems*

In Fig. 2, a measurement system similar to that of Cutler and Searle is shown. The optional multipliers are used for two reasons. First, it is often desirable to use similar oscillators. If the oscillators are truly of high precision, they are tunable only over a small range (one hertz or less). One oscillator is tuned to nominal frequency. The other is offset by some predetermined amount. Using frequency multipliers, both oscillator outputs are multiplied up until the frequency difference between the two is 10 kHz. This permits measurement down to intervals of 100 microseconds. Secondly, in addition to the basic frequencies, all instabilities are multiplied as well. The use of multipliers enables the use of a counter that is less accurate than otherwise required. For example, assume that it is necessary to measure the stability of an oscillator to 1 pp 10^7 in one millisecond, using a beat frequency of 10 kHz. To be able to resolve this quantity, the counter must be accurate to one millihertz in a millisecond. If multipliers are used, however, the instabilities are increased by the multiplication factor. If the factor used is 100, the counter need resolve only 0.1 Hz. For higher multiplication factors, counter requirements are proportionately reduced.

Up to this point, it has been assumed that the multipliers are perfect. Unfortunately, this is not the case. Highly accurate multipliers are extremely difficult to construct. Multipliers are susceptible to temperature variations and, if not properly designed, are sources of phase instability. When using multipliers, there is no guarantee that the measured instabilities are due to the oscillators. They may be due to instabilities within the multipliers.

The problem of obtaining a large enough frequency offset is easily solved. Assuming two similar precision oscillators are present, it is possible to replace the quartz crystal in one oscillator with a similar type plated for either 10 kHz above or below 5 MHz.

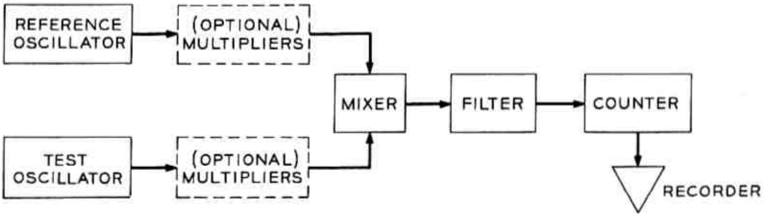


Fig. 2—Measurement system of L. S. Cutler and C. L. Searle.[10]

The problem of obtaining the necessary resolution can be solved by using commercially available counters. There are at least two types available that have sufficient accuracy to enable precise measurement at short averaging times, without multipliers. The first, a Hewlett-Packard 5248L counter, has a 100-MHz clock, allowing measurements down to one millisecond with an offset of 1000 hertz. The second, a Hewlett-Packard 5360A Computing Counter, allows measurements down to 100 microseconds with an offset of 10 kHz.

In an effort to verify the assumptions made in Section 4.2, the arrangement shown in Fig. 3 was used. The HP5248L was modified to have a one-millsecond recycle delay time. It was not possible to use the HP5360A in this configuration due to slower data transfer capabilities to the computer and interface incompatibility. A large number of samples were taken at various averaging times and the data recorded on magnetic tape. The quantities $\langle \sigma_y(N, T, \tau) \rangle$ and $\langle \sigma_y(2, T, \tau) \rangle$ were computed from the same data and compared.

Using the configuration shown in Fig. 4, measurements were made with the HP5360A Computing Counter and associated keyboard. This unit can be programmed to calculate $\langle \sigma_y(2, T, \tau) \rangle$ from its measured data. It is portable and was transported by automobile to test oscillators and frequency generators in use at various locations.

5.3 *Precision Oscillators*

The measurement systems used by the author are shown in Figs. 3 and 4. Note that a 5-MHz square wave signal from the frequency generator was used as an external time base for the counter in Fig. 4. A slight improvement was noted due to the superior stability of the frequency generator over the internal time base of the counter. It was originally planned to use the 5-MHz oscillator being tested to drive the external time base. Unfortunately, the additional load degraded measurable stability.

In the system shown in Fig 3, the oscillators, synthesizer, mixer, and video amplifier were located inside a shielded room. The counter, computer, and magnetic tape unit were located immediately outside. No frequency standard was available at the counter for use as an external time base.

In Section 2.2., the assumption was made that instabilities in oscillators are independent random processes. It can be shown[9] that the variance of the sum of two independent or uncorrelated random variables is equal to the sum of the variances, provided that the variances exist.

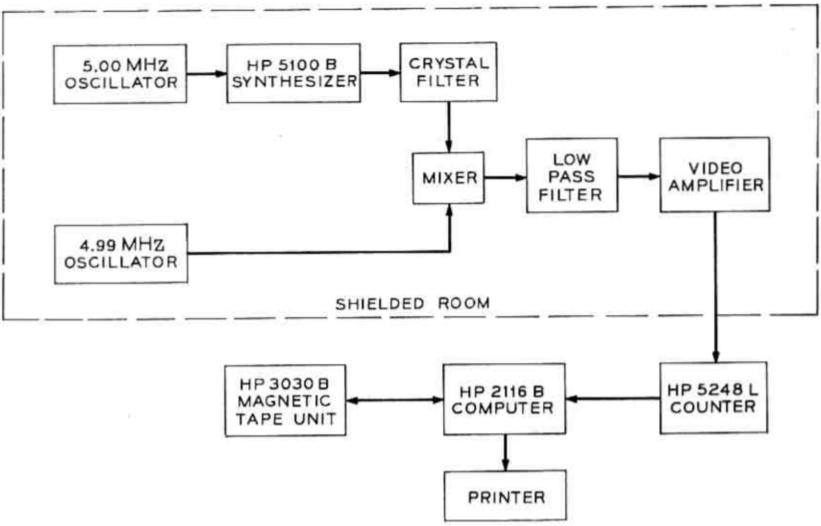


Fig. 3—Measurement system used for comparison of $\langle \sigma_y(N, T, \tau) \rangle$ and $\langle \sigma_y(2, T, \tau) \rangle$.

It was previously stated that it is often desirable to use similar oscillators. The process of selecting two oscillators at random from a large population of oscillators can be approximated by procuring two oscillators from a large supplier. Since the manufacturing processes are identical, it is assumed that the probability density and distribution functions of the oscillators are the same. The total measurable instabilities of the pair are then equal to twice the instabilities of either oscillator. Therefore, the variance measured is divided by two. Likewise, the standard deviation can be divided by $\sqrt{2}$. In the case of the

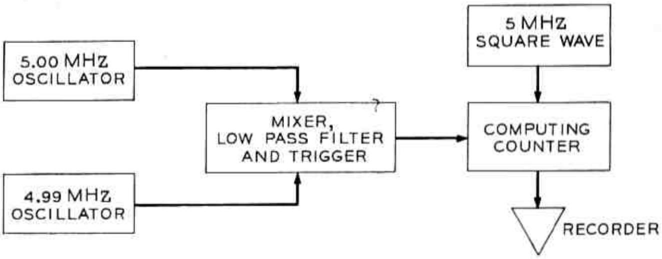


Fig. 4—Measurement system used by the author.

offset oscillator, it is assumed that modification of the crystal did not change the statistics by a significant amount.

Using the arrangement shown in Fig. 3, a comparison between $\langle \sigma_y^2(N, T, \tau) \rangle$ and $\langle \sigma_y^2(2, T, \tau) \rangle$ can be made. These quantities represent the sum of the variance contributed by the oscillator plus the variance contributed by the synthesizer, since the processes are independent.

When the arrangement shown in Fig. 4 is used, a very good estimate of the short-term stability in the time domain will result.

5.4 *Frequency Generators*

It was desirable to investigate the short-term stability of the 5-MHz square wave output of the frequency generator, described briefly in Section 5.1. The measurement system used was the same as that shown in Fig. 4 except that the 5-MHz square wave was substituted for the 5-MHz oscillator.

It seems logical that the stability of the generator would be worse than that of an oscillator, since the original sine wave produced by an oscillator has gone through shapers, counters, and other circuits used to generate the square wave signal. The exact magnitude of the generator instabilities can be easily determined. First, the total instabilities of the oscillator-generator combination are determined. The contribution of the oscillator, discussed in Section 5.3, may be subtracted out. The remainder is a good estimate of the variance of fractional frequency fluctuations from nominal frequency of the 5-MHz square wave.

5.5 *Synthesizers*

Short-term stability must often be measured when no auxiliary precision oscillators are available. By using a synthesizer, it is possible to obtain the necessary offset without the use of multipliers. Subsequently, it is possibile to arrive at an estimate of the short-term stability, provided a sufficiently accurate counter is available (see Figs. 3 and 5). Several problems are introduced, however. In the preceding examples, the stability of each oscillator or frequency generator could be determined. Such is not possible using a synthesizer since the contributions of the oscillator and synthesizer to equation (37) are unknown.

If two similar synthesizers were present, these could be driven by an external precision oscillator and the stability determined as in the case of two similar oscillators. With only one synthesizer, an estimate of the stability of the oscillator-synthesizer combination can be deter-

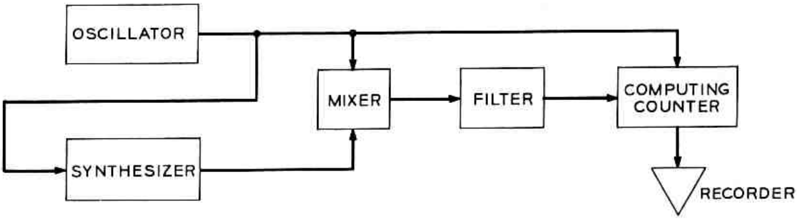


Fig. 5—Measurement system using a frequency synthesizer.

mined. Then, it can be stated that the actual stability of the oscillator is at least as good as that measured for the combination. If the synthesizer was much better than the oscillator, the majority of the instabilities measured would be contributed by the oscillator. Unfortunately, the reverse is usually true; the instabilities of the oscillator are small compared with those of the synthesizer. In some cases, however, the measured stability using such a configuration may be useful in evaluating system performance.

VI. RESULTS

6.1 *Precision Oscillators*

In order to compare the actual relationship between $\langle \sigma_v^2(N, T, \tau) \rangle$ and $\langle \sigma_v^2(2, T, \tau) \rangle$, the arrangement shown in Fig. 3 was used. The two oscillators shown are similar except for the modified crystal in the 4.99-MHz oscillator. It was assumed that the instabilities of the oscillators were similar. Therefore, the channel containing the synthesizer was inherently more noisy due to the mere presence of the synthesizer.

In an attempt to minimize the measurable stability, a narrowband crystal filter was inserted between the synthesizer output and the mixer input. At averaging times shorter than the reciprocal of the bandwidth of the filter, significant improvement in measurable stability of the system was noted. At times longer than the reciprocal of the bandwidth, the presence of the filter caused a slight degradation of measurable stability. As a result, for these longer averaging times, the filter was removed from the synthesizer output.

The total recycle delay time of the counter is about one millisecond. Therefore, $T = \tau + 1$ millisecond. For longer averaging times, this delay time is not significant. It was assumed for the longer intervals that $T \cong \tau$.

The quantities observed are listed in Table I. Note the close corre-

TABLE I—COMPARISON BETWEEN $\langle\sigma_y(N, T, \tau)\rangle$ AND $\langle\sigma_y(2, T, \tau)\rangle$

Time Interval	No. of Groups	Samples per Group (N)	$\langle\sigma_y(N, T, \tau)\rangle$	$\langle\sigma_y(2, T, \tau)\rangle$	$B_1(N, r, \mu)$	Remarks
1 millisecond	10	250	8.27 pp 10^9	8.26 pp 10^9	1.0016	See Fig. 3 for Equipment Arrangement
2 milliseconds	10	250	4.92 pp 10^9	4.88 pp 10^9	1.0164	↓
4 milliseconds	10	250	3.08 pp 10^9	3.03 pp 10^9	1.0336	
10 milliseconds	10	250	1.31 pp 10^9	1.28 pp 10^9	1.046	
20 milliseconds	10	250	6.34 pp 10^{10}	6.52 pp 10^{10}	0.9434	
40 milliseconds	10	250	3.00 pp 10^{10}	3.03 pp 10^{10}	0.98	
100 milliseconds	10	250	1.34 pp 10^{10}	1.31 pp 10^{10}	1.0429	Crystal Filter Removed From Synthesizer Output
200 milliseconds	10	250	6.88 pp 10^{11}	7.14 pp 10^{11}	0.9287	↓
400 milliseconds	10	250	3.73 pp 10^{11}	3.78 pp 10^{11}	0.9713	
1 second	10	250	1.72 pp 10^{11}	1.73 pp 10^{11}	0.9843	
10 seconds	8	100	1.57 pp 10^{12}	6.47 pp 10^{13}	5.9	Synthesizer Removed From Circuit, 4999.95 kHz Oscillator Substituted for 4990.0 kHz Oscillator
108 seconds	10	54	1.61 pp 10^{12}	6.13 pp 10^{13}	6.8765	
200 seconds	20	11	8.19 pp 10^{12}	5.56 pp 10^{12}	2.1682	Standard Oscillator Versus an Experimental, Oscillator
1080 seconds	31	11	1.85 pp 10^{12}	9.94 pp 10^{13}	3.4482	Same Configuration as 10 and 108 seconds

spondence of $\langle\sigma_y(N, T, \tau)\rangle$ and $\langle\sigma_y(2, T, \tau)\rangle$ for averaging times of one second and less. Since neither quantity is consistently larger than the other at all sampling times, it appears that correspondence would become even better if more samples were taken.

At averaging times of ten or more seconds, $\langle\sigma_y(2, T, \tau)\rangle$ becomes smaller than $\langle\sigma_y(N, T, \tau)\rangle$. It can be seen, in these longer intervals, that $B_1(N, r, \mu)$ is in the range predicted by Barnes[11] for flicker noise frequency modulation and superpositions of white and flicker noise frequency modulation.

More precise measurements of $\langle\sigma_y(2, T, \tau)\rangle$ over all time intervals were possible by using the configuration shown in Fig. 4. The improved precision was achieved by the elimination of the synthesizer and the use of a more accurate counter. The recycle delay time of the HP5360A is on the order of 1.5 milliseconds. As before, for the longer averaging times, it was assumed that $T \cong \tau$.

First, measurements were made using several standard oscillators (averaging times from 0.1 millisecond to 10 seconds). Although the instabilities contributed by similar oscillators are theoretically the same, differences in stability were observed. Among the six standard oscillators tested, using the system shown in Fig. 4, one oscillator demonstrated generally better characteristics than the other five at the short averaging times. At longer time intervals, the measured stability of each oscillator was unique. For example, at a one second averaging time, the best oscillator demonstrated a stability of better than 6.3 pp 10^{12}. [Here, the quantity measured was $\langle\sigma_y(2, T, \tau)\rangle$.] The worst was about 1.4 pp 10^{11}. The only differences in the oscillators were that oscillators one, two and three used solder-mounted crystals. The offset oscillator and oscillators four, five and six all used thermally bonded crystal units. It appears that the thermally bonded units exhibit more similarity between crystals. It is interesting to note, however, that the best stability at an averaging time of 10 seconds was observed using an oscillator containing a solder-mounted crystal unit. Likewise, the oscillator which demonstrated the worst stability also used a solder-mounted unit. Although the differences in the measurable stability of the oscillators using the thermally bonded crystal units were significant, these differences were still somewhat small. The stabilities measured are summarized in Table II and Fig. 6.

In an attempt to determine the contribution of the environment to oscillator instabilities, all equipment in the system was de-energized. The only equipment which remained under power was the frequency

Averaging Time

Configuration	0.1 ms	0.4 ms	1 ms	10 ms	100 ms	1 s	10 s
Oscillator No. 1 Operating Environment	3.3 pp 10^8	1.3 pp 10^8	6.1 pp 10^9	6.6 pp 10^{10}	6.8 pp 10^{11}	1.41 pp 10^{11}	1.5 pp 10^{11}
Oscillator No. 1 Quiet Environment	3.3 pp 10^8	1.3 pp 10^8	6.0 pp 10^9	6.46 pp 10^{10}	6.8 pp 10^{11}	1.1 pp 10^{11}	1.2 pp 10^{11}
Oscillator No. 2 Operating Environment	3.2 pp 10^8	1.3 pp 10^8	6.0 pp 10^9	7.4 pp 10^{10}	6.7 pp 10^{11}	7.0 pp 10^{12}	3.0 pp 10^{12}
Oscillator No. 2 Quiet Environment	3.2 pp 10^8	1.3 pp 10^8	6.0 pp 10^9	7.5 pp 10^{10}	6.7 pp 10^{11}	7.0 pp 10^{12}	3.0 pp 10^{12}
Oscillator No. 3. Operating Environment	3.1 pp 10^8	1.2 pp 10^8	5.7 pp 10^9	6.3 pp 10^{10}	6.1 pp 10^{11}	6.4 pp 10^{12}	2.2 pp 10^{12}
Oscillator No. 3 Quiet Environment	3.1 pp 10^8	1.2 pp 10^8	5.7 pp 10^9	6.2 pp 10^{10}	5.8 pp 10^{11}	6.4 pp 10^{12}	2.2 pp 10^{12}
Oscillator No. 4 Operating Environment	2.9 pp 10^8	1.14 pp 10^8	5.2 pp 10^9	5.8 pp 10^{10}	5.86 pp 10^{11}	7.0 pp 10^{12}	3.93 pp 10^{12}
Oscillator No. 5 Operating Environment	3.05 pp 10^8	1.23 pp 10^8	5.7 pp 10^9	6.17 pp 10^{10}	6.2 pp 10^{11}	6.25 pp 10^{12}	4.0 pp 10^{12}
Oscillator No. 6 Operating Environment	3.26 pp 10^8	1.25 pp 10^8	5.95 pp 10^9	6.33 pp 10^{10}	6.17 pp 10^{11}	6.53 pp 10^{12}	3.44 pp 10^{12}

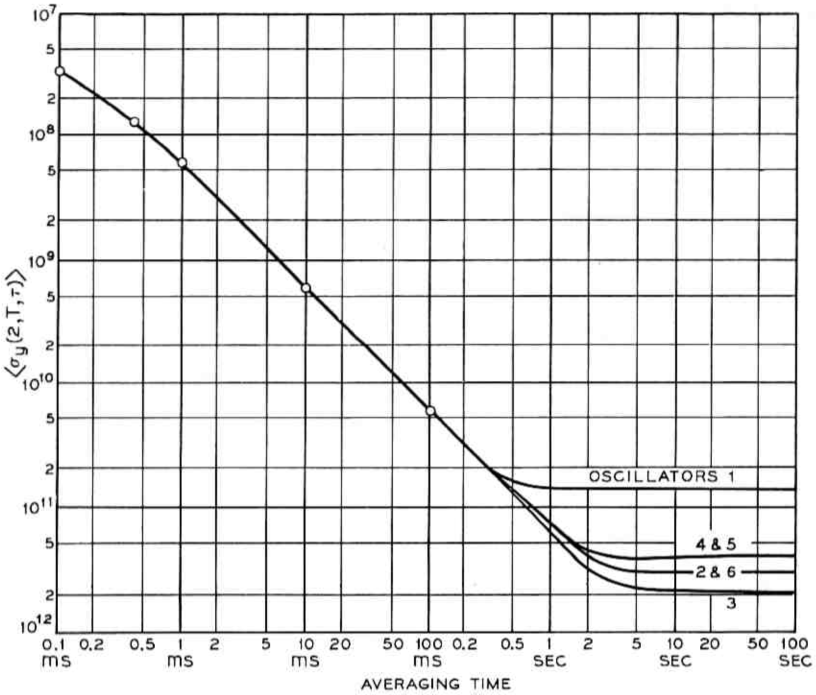


Fig. 6—Short-term stability of six precision oscillators.

generator, the offset oscillator, measurement equipment, and the air conditioning equipment required for forced air cooling of the frequency generator. Measurements were repeated for all time intervals. At the shorter intervals, no differences were apparent. In some instances, stability was slightly improved at intervals of 100 milliseconds or longer but only by an amount of questionable significance.

Since these results were not as anticipated, the causes for the disparity were investigated. The only obvious reason was inherent in the design of the mixer circuit. The mixer contains four small ferrite cores. About twenty feet from the frequency generator is a large power transformer which is used in the power distribution network for the entire data processing laboratory. Even with the remainder of the data processing system idle, this transformer must be energized, as it supplies power to the generator. Ferrites have been found to adversely affect short-term stability when used in the presence of electromagnetic fields.[17] The effect of the field created by the power transformer on mixer performance is unknown.

Other than the ferrites, no reason for the discrepancy was apparent. The most logical reason, however, would be that the measurable stability is limited by the noise characteristics of the measurement system itself. If the noise characteristics of the system were much worse than that of the oscillators, no differences in measurable stability would occur. Since differences between oscillators can be observed, the oscillators must contribute a measurable amount of the noise. As a result, if the noise characteristics of the measurement system were improved, measurable stability may improve, but probably by less than an order of magnitude.

6.2 *Frequency Generators*

6.2.1 *General*

Each frequency generator contains three precision oscillators. Any of the three oscillators may assume any of the three functions. The functions are switched by means of control panel and associated relays. The normally closed positions of the relays constitute the usual arrangement and is referred to as the "normal" mode. These are

Oscillator #1 = MASTER,
Oscillator #2 = SLAVE, and
Oscillator #3 = STANDBY.

The outputs from the phase-locked master-slave pair are digitally mixed and sent through counter chains and frequency multipliers to arrive at the various output frequencies. As mentioned above, it follows that the stability of the generator should, in general, follow the stability of the oscillators used as the frequency sources.

The observed stability of the six oscillators differed for time intervals of one second or longer (see Section 6.1). Ironically, the oscillator installed in the number one position in frequency generator #1 demonstrated the worst stability at the longer intervals. Oscillator number three demonstrated the best.

The oscillator functions were redesignated

Oscillator #3 = MASTER,
Oscillator #2 = SLAVE, and
Oscillator #1 = STANDBY.

As anticipated, the observed stability improved. Most measurements were taken using the "normal" mode, as this is the usual configuration employed during operation of the data processing system.

6.2.2 *5-MHz Square Wave, Ten-Foot Cable*

On frequency generator #1, measurements were taken in both operating and idle environments. As was the case of precision oscillators, little difference was apparent. On frequency generator #2, measurements were taken in an operating environment only.

To make estimates of the short-term stability of the generator, it was necessary to obtain the variance of the fractional frequency fluctuations of the offset oscillator-frequency generator pair. When these were determined, the variance contributed by the offset oscillator was subtracted out of the data.

The question at this point was, which oscillator-data should be subtracted out to arrive at a reasonable estimate of the square wave stability? The data from the best oscillator pair from each frequency generator was used. If the data from the worst pair was used, the stability of the generator may be too optimistic. Since the offset oscillator used a thermally bonded crystal unit, the actual stability is probably similar to the other oscillators using thermally bonded units.

In extremely critical applications, the measured stability between the oscillator-generator pair can be considered to be the stability of the generator itself. It may be safely assumed that the actual stability of the generator is no worse than the measured stability of the oscillator-generator pair.

See Table III for a listing of the quantities observed. Figure 7 shows the same data in graphical form.

6.2.3 *5-MHz Square Wave, 200-Foot Cable*

On frequency generator #2, measurements were made only in an operating environment. Measurements were taken on frequency generator #1 in both operating and idle environments. Little, if any, difference was noted between the two environments except as noted below.

At intervals of 100 and 400 microseconds, little difference between the short and long cables was apparent. At averaging times of 1 millisecond, the measured instabilities began to increase. At 10 milliseconds, the additional instabilities reached a peak about forty percent above the short cable. For 100 milliseconds, the figure had declined, although not to the level present using the short cable. At intervals

TABLE III—SHORT-TERM STABILITY OF TWO FREQUENCY GENERATORS

	Averaging Time						
Configuration	0.1 ms	0.4 ms	1 ms	10 ms	100 ms	1 s	10 s
Generator No. 1 10-Foot Cable Operating	5.6 pp 10^8	2.3 pp 10^8	1.1 pp 10^8	1.4 pp 10^9	1.4 pp 10^{10}	2.2 pp 10^{11}	2.54 pp 10^{11}
Generator No. 1 10-Foot Cable Quiet	5.6 pp 10^8	2.26 pp 10^8	1.1 pp 10^8	1.3 pp 10^9	1.3 pp 10^{10}	2.2 pp 10^{11}	2.0 pp 10^{11}
Generator No. 1 200-Foot Cable Operating	5.5 pp 10^8	2.3 pp 10^8	1.13 pp 10^8	1.91 pp 10^9	1.51 pp 10^{10}	2.2 pp 10^{11}	2.4 pp 10^{11}
Generator No. 1 200-Foot Cable Quiet	5.5 pp 10^8	2.25 pp 10^8	1.12 pp 10^8	1.81 pp 10^9	1.4 pp 10^{10}	2.1 pp 10^{11}	2.3 pp 10^{11}
Generator No. 2 10-Foot Cable Operating	7.0 pp 10^8	2.6 pp 10^8	1.24 pp 10^8	1.46 pp 10^9	1.36 pp 10^{10}	1.65 pp 10^{11}	7.7 pp 10^{12}
Generator No. 2 200-Foot Cable Operating	6.8 pp 10^8	2.57 pp 10^8	1.25 pp 10^8	1.84 pp 10^9	1.65 pp 10^{10}	2.2 pp 10^{11}	6.64 pp 10^{12}

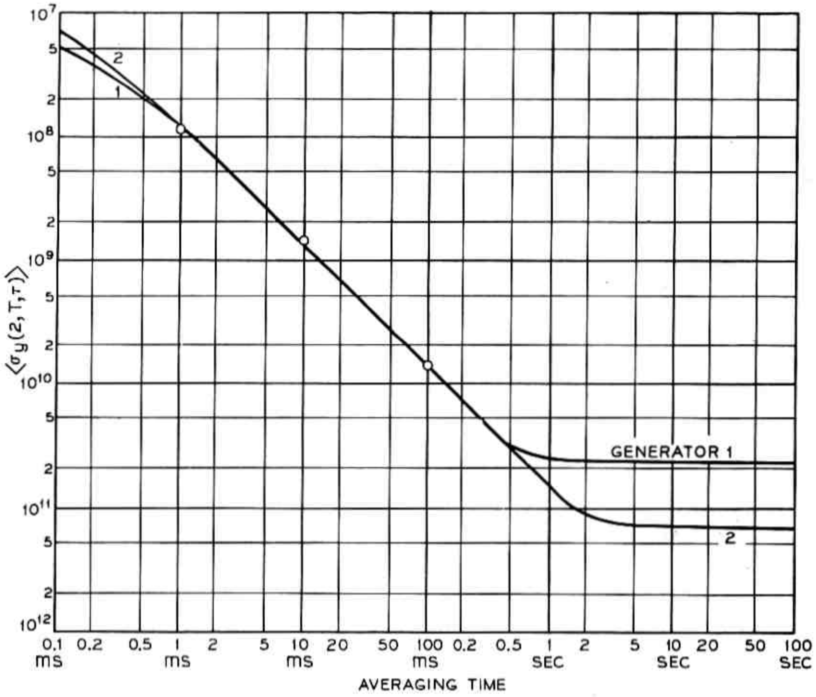


Fig. 7—Stability of two frequency generators, ten-foot cable.

of one and ten seconds, instabilities returned to the level observed using the short cable.

Some of the causes of the additional instabilities centered at ten milliseconds are identifiable in this instance. In many data processing systems, pulse rates of 100 bits per second are used. Since this rate has a duty cycle of ten milliseconds, it is not surprising that averaging times of ten milliseconds would be affected. Comparing the data between quiet and operating environments, it became apparent that averaging times of ten milliseconds were indeed affected by the operation of the data processing system. In addition, the period of the second harmonic of the power line frequency occurs near this rate. Much longer time intervals are not affected by either of these noise sources.

The stabilities observed for both frequency generators are plotted in Fig. 8. The same data is presented in tabular form in Table III.

6.3 *Synthesizers*

A Hewlett-Packard model 5105A Frequency Synthesizer/5110B Synthesizer Driver was arranged as shown in Fig. 5. A model 5100A Frequency Synthesizer/5110B Synthesizer Driver was used in Fig. 3. A 5-MHz precision oscillator was used as an external frequency standard input to the driver in both applications. In the neighborhood of 5 MHz, the short-term stability of the 5100A is much better at all time intervals than that of the 5105A. As a result, the 5100A was more useful in evaluating performance of oscillators and frequency generators. The 5105A has a higher frequency capability than the 5100A, and produces the best short-term stability at these frequencies.

Using the 5105A, the synthesizer output frequency was adjusted for 4.99 MHz. The output was mixed with the 5-MHz oscillator which was being used as the external reference. Here, the stability of the pair was about 2 pp 10^{10} over one second. A 5-MHz square wave was then

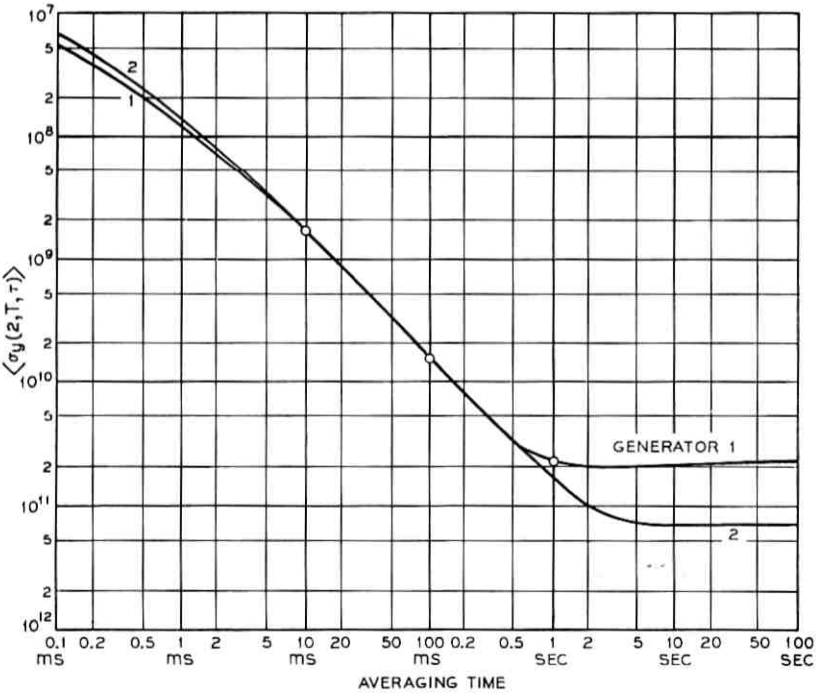


Fig. 8—Stability of two frequency generators, 200-foot cable.

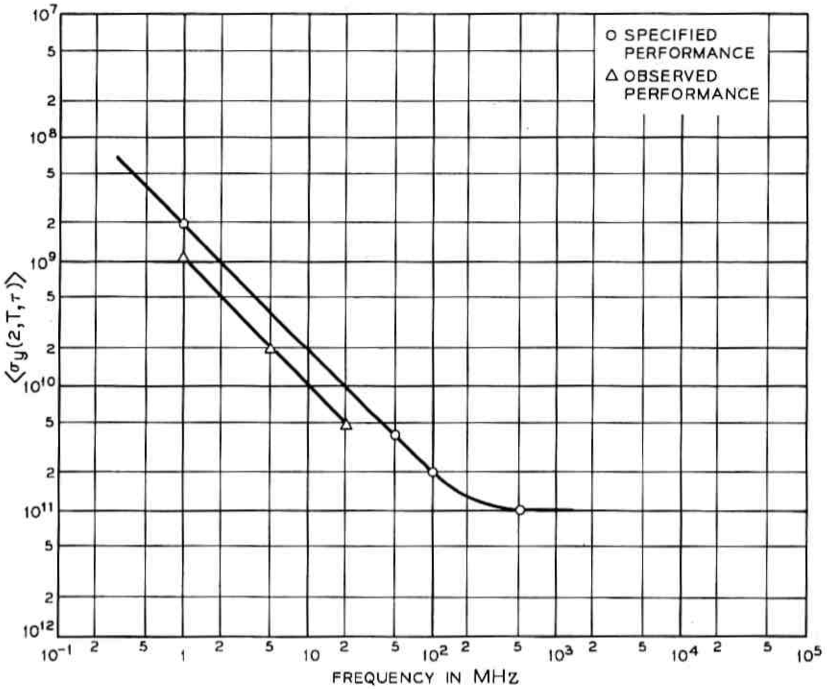


Fig. 9—Performance of frequency synthesizer, one-second averaging time.

mixed with the synthesizer output with no significant change in results.

The frequency of the synthesizer was then readjusted for 990 kHz and the output was mixed with a 1-MHz square wave signal from the generator. The stability was now reduced to 1.5 pp 10^{9} over one second.

The frequency was then increased to 19.99 MHz and the output was mixed with a 20-MHz square wave signal from the generator. In this configuration, measurable stability improved to 5 pp 10^{11} over one second.

Figure 9 shows a plot of the frequency versus the standard deviation of fractional frequency fluctuations for a one-second averaging time. The specifications from the manufacturer's catalog are plotted on the same graph. Since the slopes appear the same, the synthesizer is the most likely source of the majority of noise. It may be assumed that the stability of all the frequencies measured are of the same order as the stability observed for the 20-MHz square wave. In reality, all may be somewhat better. Similar data for the 5100A Synthesizer was not taken.

6.4 *Systematic Effects*

From the experimental data, it seems that systematics from the data processing system do not contribute as large a portion of the instabilities as previously suspected. It must be noted, however, that the data processing environments used for testing purposes were only small, developmental varieties of the larger system (to be constructed at a later date). It is anticipated that the larger system will contribute more systematic noise to all frequency sources within the system.

Several discussions took place with individuals involved in short- and long-term stability measurements. Some of these individuals had made such measurements in areas where machine tools or other heavy electrical equipment were operated. There seemed to be a general consensus that stability is degraded at 8:00 a.m. and 4:30 p.m. or whenever the equipment is being started or stopped. Since no heavy machinery was located in the near vicinity of the data processing systems, these effects were not observed.

Each day, somewhere between 4:00 and 4:15 p.m., it was noted that observable stability became two to three times worse than that normally observed. About 4:30 p.m., the instabilities returned to their former level. Occasionally, the same effects were observed at other times during system operation.

Investigation yielded two causes for these observations. First, many individuals are dumping data on the various input-output devices to prepare for the change in shifts. Second, a concept of "rotation" is present to make the system available to individuals wishing to run software tests. Time intervals of five minutes are scheduled in advance. Individuals are generally prepared with magnetic and paper tapes and subject the system to heavy use during their five-minute allotment. It appeared that the increased instabilities observed during mid-day operation occurred during rotation periods.

Accuracy of measurements was probably reduced due to systematic noise introduced by the measurement system itself. For a good characterization of the noise due strictly to the oscillator or frequency generator plus induced noise from the environment, the system should have a noise level capability of at least two orders of magnitude better than the anticipated stability of the unit being tested.

VII. RECOMMENDATIONS AND CONCLUSIONS

The aspects of short-term frequency stability applicable to most situations have been discussed. Although definitions were made in both the frequency and time domains, the time domain definitions were used for measurement.

With the vast increase in the use of precision oscillators, measurement techniques must be fast, accurate, and easy to perform. Frequency domain measurements are extremely difficult to make accurately. In addition, such measurements take a long time to gather sufficient data.

The actual observable relationship between $\langle\sigma_y(N,\ T,\ \tau)\rangle$ and $\langle\sigma_y(2,\ T,\ \tau)\rangle$ was discussed. It was shown that the two quantities are very nearly equal for averaging times of one second or less. For longer times, the results are in agreement with those predicted by Barnes.[11]

Time domain measurements, using a device such as the HP5360A computing counter, are fast and consistent. The accuracy limitations of such measurements depend mainly on mixing and filtering equipment. A good method for determination of the short-term stability of precision oscillators and frequency generators is the use of an offset oscillator to obtain a frequency difference.

Further investigation of systematic effects is in order. Initially, systematic noise due to the measurement system must be minimized. Then, estimates of the stability of the oscillators in a quiet environment should be performed. Next, estimates should be made in an operating environment under controlled activity levels. Using this data, it may be possible to generate a mathematical model of the contributions of the systematic effects of the data processing system to the short-term frequency stability.

Investigation of the distribution of the instabilities would be extremely useful in evaluating performance of the measurement system. Such computations can be made by appropriate programming of the HP2116B computer shown in Fig. 3.

VIII. ACKNOWLEDGMENTS

Many persons have contributed to this study. The author expresses thanks to them all for their assistance. Special thanks must go to J. R. Hahn, Jr., J. Leighton, and M. E. Fabian for their advice and assistance while this study was being conducted. Gratitude is expressed to the IEEE Subcommittee on Frequency Stability for their assistance, particularly J. A. Barnes, Chairman. W. L. Smith, a member of the subcommittee, provided invaluable assistance in discussion, criticism, suggestions, and measurements.

REFERENCES

1. Pustarfi, H. S., "An Improved 5-MHz Reference Oscillator for Time and Frequency Standard Applications," IEEE Trans. for Time and Frequency Standard Applications, *IM-15,* No. 4 (December 1966), pp. 196–202.
2. Baghdady, E. J., Lincoln, R. N., and Nelin, B. D., "Short-Term Frequency Stability: Characterization, Theory, and Measurement," Proc. IEEE-NASA Symp. on Short-Term Frequency Stability, U. S. Government Printing Office, Washington, D. C., NASA SP-80, pp. 65–87. Also published in Proc. IEEE, *53,* No. 7 (July 1965), pp. 704–722.
3. Blackman, R. B., and Tukey, J. W., *The Measurement of Power Spectra,* New York: Dover Publications, Inc., 1959.
4. Davenport, W. B., and Root, W. L., *Random Signals and Noise,* New York: McGraw-Hill, Inc., 1958.
5. Rice, S. O., "Mathematical Analysis of Random Noise," B.S.T.J., *23,* No. 3 (July 1944), pp. 282–332. Also published in *Selected Papers on Noise and Stochastic Processes,* Wax, N., editor, New York: Dover Publications, Inc., 1958, pp. 133–294.
6. IEEE Subcommittee on Frequency Stability, private communication.
7. Barnes, J. A., "Atomic Timekeeping and the Statistics of Precision Signal Generators," Proc. IEEE, *54,* No. 2 (February 1966), pp. 207–220.
8. Allan, D. W., "Statistics of Atomic Frequency Standards," Proc. IEEE, *54,* No. 2 (February 1966), pp. 221–230.
9. Fisz, M., *Probability Theory and Mathematical Statistics,* New York: John Wiley and Sons, Inc., 1963.
10. Cutler, L. S., and Searle, C. L., "Some Aspects of the Theory and Measurement of Frequency Fluctuations in Frequency Standards," Proc. IEEE, *54,* No. 2 (February 1966), pp. 136–154.
11. Barnes, J. A., "Tables of Bias Functions, B_1 and B_2, for Variances Based on Finite Samples of Processes with Power Law Spectral Densities," NBS Technical Note No. 275, U. S. Government Printing Office, Washington, D. C., 1969.
12. Barber, R. E., "Studies of the Short-Term Frequency Stability of Precision Oscillators and Frequency Generators," North Carolina State University, Raleigh, North Carolina, 1970.
13. Hewlett-Packard Company, "Frequency and Time Standards," Application Note No. 52, Palo Alto, California, 1965.
14. Gerber, E. A., and Sykes, R. A., "State of the Art—Quartz Crystal Units and Oscillators," Proc. IEEE, *54,* No. 2 (February 1966), pp. 103–116.
15. Barnes, J. A., and Allan, D. W., "A Statistical Model of Flicker Noise," Proc. IEEE, *54,* No. 2 (February 1966), pp. 176–178.
16. Paradysz, R. E., and Smith, W. L., "Measurement of FM Noise Spectra of Low-Noise VHF Crystal Controlled Oscillators," IEEE Trans. on Instrumentation and Measurement, *IM-15,* No. 4 (December 1966), pp. 202–211.
17. Smith, W. L., private communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

On the Design and Analysis of a Class of PCM Systems

By H. HEFFES, S. HORING, D. L. JAGERMAN

(Manuscript received July 20, 1970)

This paper considers the problem of transmitting bandlimited signals using binary signaling over a noise-free channel. An analytical framework is presented for the design and analysis of a class of PCM systems where peak error is of primary interest. For a specific class of input signals which includes deterministic amplitude-constrained bandlimited functions as well as bandlimited wide-sense stationary, second-order, random processes, results are obtained which provide trade-offs between the sampling rate, quantizer and reconstruction filter.

I. INTRODUCTION

This paper considers the problem of transmitting bandlimited signals using binary signaling over a noise-free channel. A functional block diagram of the type of PCM system under consideration is shown in Fig. 1.*

The two major differences between the problem considered here and previous work are the measure of system performance and signal classes considered. The measure of system performance usually considered is related to the integral mean-squared error.[1-3] This type of performance measure lends itself to a frequency-domain analysis. While this criterion is widely used, it doesn't provide direct information regarding the size of the error as a function of time. To investigate this time behavior, we use a time-domain approach and use the maximum error over time as a measure of system performance.

Because of the desire to consider input signals that engineers would normally call bandlimited [i.e., trigonometric polynomials, sinusoids as well as functions in $\mathcal{L}_2(-\infty, \infty)$ with finite bandwidth] we treat a

* For special classes of signals, other types of PCM systems such as DPCM and Delta Modulation are sometimes used. These systems will not be considered here.

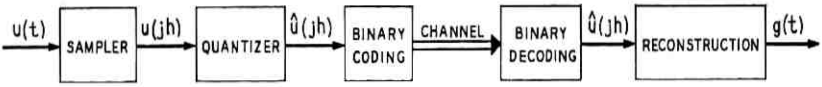


Fig. 1—UPCM system.

somewhat broader class of signals than normally considered in the literature. (The input class is usually considered to be a wide-sense stationary random process.)

In this paper, a technique is presented which simultaneously selects the sampling rate, quantizer and reconstruction filter in such a way as to minimize a bound on the peak error between the reconstructed and transmitted-signals. Since our interest centers on studying the trade-offs between the various system parameters, we desire precise mathematical results which are valid over the entire range of possible system parameters. By using an upper bound on system performance this aim was achieved for several classes of input signals. Using a criterion which evaluates a given encoding-decoding scheme in terms of its performance for the worst signal in the input class, a specific encoding-decoding algorithm, which will be called Uniform PCM or UPCM, is suggested and evaluated. The results are presented in the form of a set of normalized curves which plot an upper bound on the percentage error associated with the proposed UPCM system as a function of a normalized parameter, ρ, which represents the ratio of the bit rate to the bandwidth of the input class. The optimized values of the various parameters which define the system can also be determined from the plots (these include the sampling rate, the number of quantizing levels and the delay associated with the decoder).

II. AN ENCODING-DECODING ALGORITHM FOR THE TRANSMISSION OF BANDLIMITED FUNCTIONS

The class B_σ of functions consists of entire functions of exponential order one and type σ which are bounded on the real axis.[4] It includes all functions in $\mathfrak{L}_2(-\infty, \infty)$ having finite radian bandwidth, σ, and all trigonometric polynomials of degree $[\sigma]$.*

The performance of a given system is defined by

* [ ] denotes integral part. A trigonometric polynomial of degree $[\sigma]$ has the form $\frac{1}{2}a_0 + \sum_{k=1}^{[\sigma]} (a_k \cos kt + b_k \sin kt)$.

$$\epsilon = \sup_{t \varepsilon [0, T]} \sup_{u \varepsilon B_\sigma(M)} \| u(t) - g(t; u) \| \tag{1}$$

where $u(t)$ is the value of the input signal at time t and $g(t; u)$ represents the value of the output signal at time t corresponding to the input function u. The input function u is an element of B_σ and the values of the input and output signals at any time $t \varepsilon [0, T]$ belong to a normed linear space, Ω. By proper choice of Ω and the norm on Ω, a variety of different input classes and performance measures can be treated. Specifically, define the set $B_\sigma(M)$ by

$$B_\sigma(M) = \{x : x \varepsilon B_\sigma , x(t) \varepsilon \Omega, \| x(t) \| \leqq M \ \forall \ t\}.$$

In arriving at the proposed algorithm, the following representation is used for elements of B_σ (and hence $B_\sigma(M)$:*

$$u(t) = \sum_{j=-\infty}^{\infty} u(jh)\theta_j(t) \tag{2}$$

where

$$\theta(t) = \frac{\sin \dfrac{\delta\sigma}{1 - \delta} t \sin \dfrac{\sigma}{1 - \delta} t}{\dfrac{\delta\sigma}{1 - \delta} t \quad \dfrac{\sigma}{1 - \delta} t} ; \tag{3}$$

$$\theta_j(t) = \theta(t - jh), \qquad h = \frac{\pi(1 - \delta)}{\sigma}. \tag{4}$$

This representation is valid for any $\delta \varepsilon (0, 1)$, hence δ may be chosen appropriately for each application. The parameter δ will be called the fractional guardband since the time between samples is $\pi(1 - \delta)/\sigma$ which is less than the time between samples π/σ which corresponds to the Nyquist rate.

In essence the proposed algorithm is a scheme for approximating any element of $B_\sigma(M)$† by one of a finite set of appropriately chosen functions. These functions are determined by first truncating the infinite series, (2). This truncation process produces an approximation to $u(t)$ (over the time interval $t \varepsilon [0, T]$) of the form

$$\bar{u}(t) = \sum_{j=-L}^{[T/h]+L} u(jh)\theta_j(t). \tag{5}$$

* This is a special case of a broader class of representations for elements of B_σ[5]. It was chosen because it provided the smallest bound on quantization error.

† The results presented in this paper are applicable to the class of signals having representation (2) with $\|u(jh)\| \leqq M$. This class is larger than $B_\sigma(M)$.

Thus we have replaced the requirement of transmitting an infinite number of sample values by the problem of transmitting a finite number. To achieve our ultimate objective, these sample values are then quantized and the quantized samples are used to reconstruct the signal

$$g(t) = \sum_{j=-L}^{[T/h]+L} \hat{u}(jh)\theta_j(t) \tag{6}$$

where $\hat{u}(jh)$ represents the quantized value of $u(jh)$.

The encoder will therefore consist of appropriate quantization of the input samples while the decoder simply represents the reconstruction of the truncated series using the quantized samples [i.e., the signal $g(t)$ given in equation (6)]. An analog interpretation of the decoding process is discussed in Section V. It should be noted that in order to construct $g(t)$ according to equation (6), a delay of $T + Lh$ seconds is required. The trade-offs between this delay and the accuracy of reconstruction will become clear as the results are presented.

Having constrained the general form of the proposed system, we now seek to determine the various parameters which define it (e.g., sampling rate and quantizer) in such a way as to minimize an upper bound on the value of ϵ. Because of the binary nature of the signaling which is being considered, the number of quantization levels is constrained to be 2^ν, where ν is an integer.* Thus ν represents the number of bits per sample. The bit rate is then given by

$$B = \frac{\nu}{h}\,\text{b/s}. \tag{7}$$

It is convenient to define a normalized parameter ρ, which is given by

$$\rho = \frac{2\pi B}{\sigma} = \frac{B}{f} \tag{8}$$

where σ is the radian bandwidth of the input class and f is the bandwidth in Hz. In terms of these parameters, from equation (4), we have

$$\delta = 1 - \frac{2\nu}{\rho}. \tag{9}$$

Since $\delta \epsilon (0, 1)$, the number of bits per sample must satisfy

$$1 \leqq \nu \leqq \left[\frac{\rho}{2}\right]. \tag{10}$$

* This is a practical constraint and not a theoretical one.

In terms of these parameters a simple bound on the reconstruction error results.

The analysis of the reconstruction error, given in the Appendix, shows that ϵ, defined by equation (1), satisfies

$$\epsilon \leqq \bar{\epsilon} = \frac{MK}{\pi^2 \, \delta L} + \frac{\sigma_q(\nu)}{\sqrt{\delta}} \tag{11}$$

where

$$K = \frac{\dfrac{T}{Lh} + 2}{\dfrac{T}{Lh} + 1}, \tag{12}$$

$$\sigma_q(\nu) = \sup_{j} \sup_{u \varepsilon B_\sigma(M)} \| u(jh) - \hat{u}(jh) \| \tag{13}$$

and δ is given by equation (9). Thus the actual value of ϵ which is achieved by the proposed algorithm may be less than the value $\bar{\epsilon}$ given by equation (11). In fact, since the reconstruction formula (6) is interpolatory, the error at sample points can never exceed the raw quantization error, $\sigma_q(\nu)$. The first term on the right-hand side of equation (11) may be viewed as the error due to truncating $u(t)$ as given by equation (2) while the second term represents the quantization error. The effect of the delay on the error can be easily seen from equations (11) and (12).

As a design procedure, one might first choose the quantizer which minimizes $\sigma_q(\nu)$ and then determine the value of ν [subject to (10)] which minimizes the error bound given by $\bar{\epsilon}$ [see equation (11)].

In the next sections, several important special cases are considered and explicit design curves are presented for these cases.

III. DETERMINISTIC, AMPLITUDE CONSTRAINED BANDLIMITED FUNCTIONS

In this case, $\Omega = R$ (the real line), and $||x|| = |x|$. Thus

$$B_\sigma(M) = \{u : u \varepsilon B_\sigma , u(t) \varepsilon R, \mid u(t) \mid \leqq M \ \forall \ t\}^*.$$

The quantizer which minimizes $\sigma_q(\nu)$ for this case is shown in Fig. 2. The optimal value of $\sigma_q(\nu)$ is given by

$$\sigma_q(\nu) = \frac{M}{2^\nu}. \tag{14}$$

* Insofar as information content is concerned, if $M_1 \leqq u(t) \leqq M_2$, then it is equivalent to consider $|u(t)| \leqq M$, where $M = (M_2 - M_1)/2$.

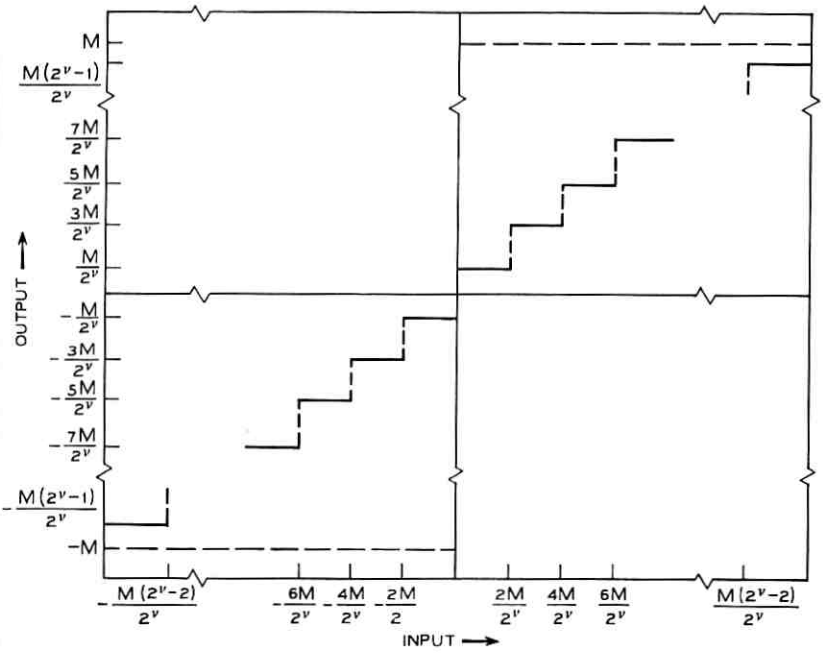


Fig. 2—Uniform quantizer.

Using equation (14), equation (11) becomes

$$\epsilon \leqq \bar{\epsilon} = \frac{MK}{\pi^2\, \delta L} + \frac{M}{\sqrt{\delta}\, 2^\nu}. \tag{15}$$

For later use, we define

$$\bar{\epsilon} = \frac{MK}{\pi^2\, \delta L} + \frac{K_0 M}{\sqrt{\delta}\, 2^\nu}. \qquad (16)^*$$

For the present case $K_0 = 1$ and equation (15) thus becomes $\epsilon \leqq \bar{\epsilon} = \bar{\epsilon}$.

The bound represented by equation (15) is valid for any allowable set of system parameters. For fixed values of ρ, T, and L, the value of ν which minimizes $\bar{\epsilon}$ can easily be found. If we define the normalized percentage error as $100(\bar{\epsilon}/2MK_0)$, then Fig. 3 plots this as a function of ρ, using the optimized values of ν. These values of ν (denoted ν^*) are indicated on the curves. Curves are plotted for $K_0L/K = 5, 10, 20, 40,$

* K_0 is introduced here for the purpose of unifying the results of this section and the next. For the class of signals in this section, K_0 can be replaced by unity.

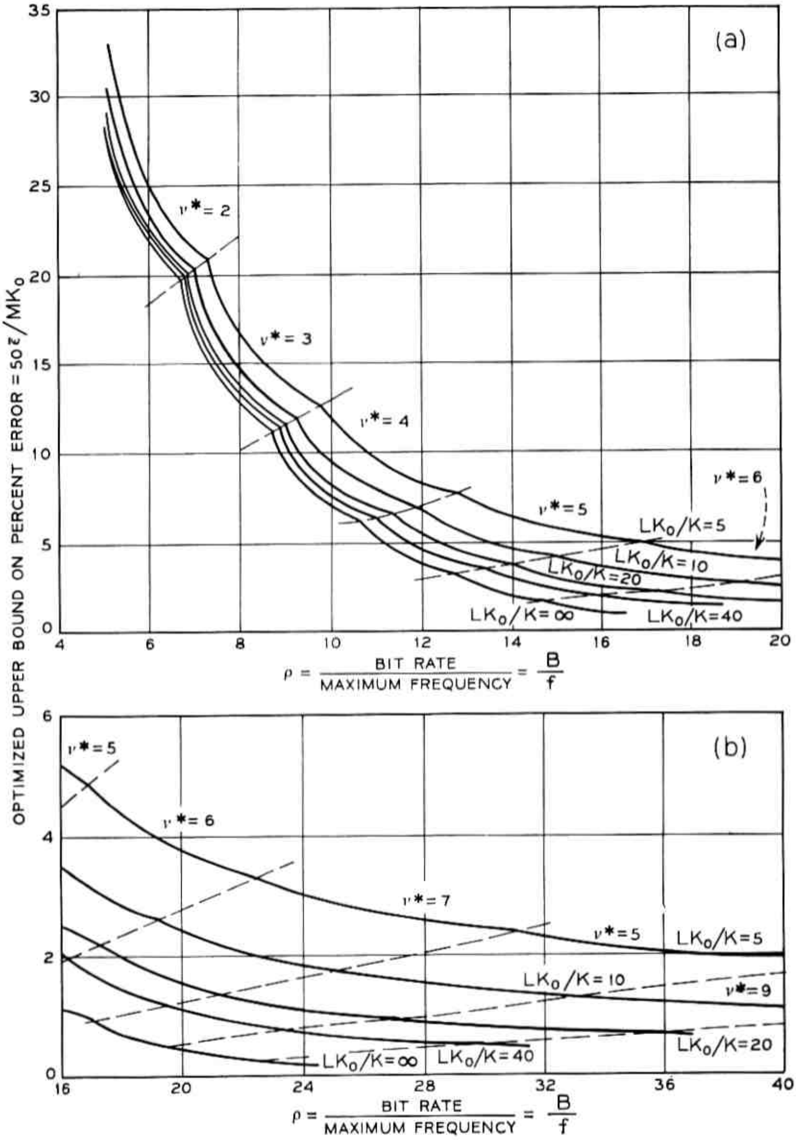


Fig. 3a and b—Performance curves. (Notes: 1. For analog implementation, use $K = 1$; delay $= (L + 1)\ h$. 2. For digital implementation, $K = (T/Lh + 2)/(T/Lh + 1)$; delay $= T + Lh$. 3. ν^* is the optimized number of bits per sample. 4. The optimum sampling rate is $1/h = B/\nu^*$.)

and ∞. Since $1 \leqq K \leqq 2$, for a given value of the ratio L/K, one can associate any value of L satisfying $L/K \leqq L \leqq 2L/K$. Using equation (12), the corresponding value of T can then be computed. The delay associated with reconstruction in accordance with equation (6) is then given by $T_d = T + Lh$. As discussed in Section V, when an analog implementation is used for the reconstruction, a value of $K = 1$ should be used and the delay associated with the filter which accomplishes the reconstruction is given by $T_d = (L + 1)h$.

To illustrate the use of these curves, consider the problem defined by the following parameter values:

$$\sigma = \pi \times 10^6,$$

$$B = 6.3 \times 10^6 \text{ b/s}.$$

Using these values, we compute $\rho = 12.6$. From the curves, for $L/K = 20$, we have $50\bar{\epsilon}/M = 4.67$ percent, $\nu^* = 5$. Using $M = \frac{3}{8}$, we have $\bar{\epsilon} = 0.035$. Since $B = \nu/h$, we have $h^* = 0.79 \times 10^{-6}$ seconds or $1/h^* = 1.26 \times 10^6$ samples/second. For $L = 40$, the delay associated with the error [using equation (6) to accomplish the reconstruction] is about 32 μs. For the corresponding analog implementation (See Section V) these results correspond to a value of $L = 20$ (and $K = 1$) with a corresponding delay of 16.6 μs.

It is clear that the normalized nature of the curves in Fig. 3 facilitates their use in a wide variety of ways. For example, one could easily answer questions such as:

(*i*) For a given class of signals, what is the smallest bit rate that can be used to guarantee an error not exceeding a prescribed level?

(*ii*) For a given bit rate, what is the largest class of signals that can be handled with a maximum error not exceeding a prescribed level?

In the next section, analogous results are developed for some important classes of random input signals.

IV. SOME IMPORTANT CLASSES OF RANDOM INPUTS

In this section we consider the case where Ω is the space of zero mean second-order random variables* with norm given by

$$\| x \| = (Ex^2)^{\frac{1}{2}} \equiv \sigma_x , \qquad x \in \Omega. \tag{17}$$

* There is no loss of generality in the zero-mean assumption.

Then

$$B_\sigma(M) = \{u : u \varepsilon B_\sigma\,,\ u(t) \varepsilon \Omega,\ \sigma_{u(t)} \leqq M\}. \tag{18}*$$

Thus $B_\sigma(M)$ consists of second-order random processes (with $\sigma_{u(t)} \leqq M$) with sample functions which are all in B_σ. For the important special case of wide-sense stationary random processes, $\sigma_{u(t)} \equiv \sigma_u$. For this case, the quantizer will be characterized by the property that it minimizes the mean-squared error at the sample times. The problem of designing such a quantizer for a given probability distribution of the input amplitude has been considered by B. Smith.[6] An approximate upper bound on the optimized quantizing error is given by

$$\sigma_q \leqq \bar{\sigma}_q \cong \bar{\sigma}_q = \frac{1}{2^\nu}\sqrt{\frac{2}{3}}\left[\int_0^\infty p^{\frac{1}{3}}(u)\,du\right]^{\frac{3}{2}} \tag{19}$$

where $p(u)$ is the probability density function of $u(t)$. The corresponding quantizer is described in Ref. 6.

For the important cases where $p(u)$ is uniform, gaussian, or exponential, $\bar{\sigma}_q$ can be written in the form

$$\bar{\sigma}_q = \frac{K_0\sigma_u}{2^\nu} \tag{20}$$

where K_0 is a constant depending on the particular distribution function. Table I gives the value of K_0 for each of the input classes of interest. The bound on the system performance which is given by equation (11) can thus be written as

$$\epsilon \leqq \bar{\epsilon} \cong \bar{\epsilon} = \frac{MK}{\pi^2\,\delta L} + \frac{K_0 M}{\sqrt{\delta}\,2^\nu} \tag{21}$$

which is identical in form to the corresponding bound for the deterministic case [see equation (16)]. The use of the design-analysis curves of Fig. 3 for these cases is thus identical to the previously described deterministic case.

It is interesting to note that for the case of a uniform amplitude distribution as well as the amplitude constrained input class, $\sigma_q = \bar{\sigma}_q$ and the optimal quantizer is a uniform quantizer. In each of these cases $\epsilon \leqq \bar{\epsilon}$.

In the next section, the interpretation and implementation of the proposed algorithm will be discussed. It will be shown that the analog implementation takes the form of a PCM system where the sampling

* More precisely, $B_\sigma(M) = \{u(\cdot\,,\ \cdot): u(\cdot\,,\ \omega) \varepsilon B_\sigma,\ u(t,\ \cdot) \varepsilon \Omega,\ \|u(t,\ \cdot)\| \leqq M\}$.

TABLE I—INPUT SIGNAL DESCRIPTION*

Deterministic	Random		
	Uniform	Gaussian	Exponential
$\lvert u(t)\rvert \leqq M$	$p(u) = \begin{cases} \frac{1}{2a} & \lvert u\rvert \leqq a \\ 0 & \lvert u\rvert > a \end{cases}$	$p(u) = \frac{1}{\sqrt{2\pi}\,\sigma_u}$ $\exp\,(-u^2/2\sigma_u^2)$	$p(u) = \frac{1}{\sqrt{2}\,\sigma_u}$ $\exp\,(-\sqrt{2}u/\sigma_u)$
$K_0 = 1$	$K_0 = 1$	$K_0 = 1.65$	$K_0 = 2.12$
	$\sigma_u = \frac{a}{\sqrt{3}} \equiv M$	$M \equiv \sigma_u$	$M \equiv \sigma_u$

* All signals are in B_σ.

rate, quantizer and low-pass filter which is used to accomplish the decoding are carefully chosen.

V. ANALOG RECONSTRUCTION IN THE UPCM SYSTEM

In this section we discuss an analog reconstruction in the UPCM system. For the analog reconstruction, the system takes the form shown in Fig. 4. The identification of the low-pass filter results from the following analysis.

If we first consider the response, $\bar{g}(t)$, of a causal, stationary filter [with impulse response $h(t)$] to the input

$$u^*(t) = \sum_{j=-\infty}^{\infty} \hat{u}(jh)\ \delta(t - jh) \tag{22}$$

then

$$\bar{g}(t) = \sum_{-\infty}^{[t/h]} \hat{u}(jh)h(t - jh). \tag{23}$$

We now consider $h(t)$ to be a delayed, truncated (for negative times)

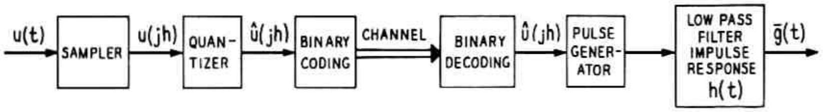


Fig. 4—Analog UPCM system.

version of $\theta(t)$ [see equation (3)]

$$h(t) = \theta(t - (L + 1)h)S(t) = \theta_{L+1}(t)S(t) \tag{24}$$

where $\theta(t)$ is given in equation (3) and $S(t)$ is the unit step. The response $\bar{g}(t)$ can thus be written

$$\bar{g}(t) = \sum_{-\infty}^{[t/h]} \hat{u}(jh)\theta_j(t - (L + 1)h). \tag{25}$$

If $\bar{u}(t)$, given in equation (5), were more accurately represented by

$$\tilde{u}(t) = \sum_{j=-\infty}^{[T/h]+L} u(jh)\theta_j(t), \qquad t\,\varepsilon\,[0, T] \tag{5'}$$

then $g(t)$, given by equation (6), would more accurately be represented (for $T = h$) by

$$\tilde{g}(t) = \sum_{j=-\infty}^{L+1} u(jh)\theta_j(t), \qquad t\,\varepsilon\,[0, T]. \tag{6'}$$

Equation (6′) can be made valid for all t with

$$\tilde{g}(t) = \sum_{j=-\infty}^{[t/h]+L+1} \hat{u}(jh)\theta_j(t). \tag{26}$$

Thus

$$\tilde{g}(t - (L + 1)h) = \sum_{-\infty}^{[t/h]} \hat{u}(jh)\theta_j(t - (L + 1)h). \tag{27}$$

We thus see that the output $\bar{g}(t)$ of the filter, with impulse response $h(t)$ given by equation (24), is a delayed version [delay of $(L + 1)h$ seconds] of the more accurate representation of $g(t)$ given by equation (6′). It should be noted that the more accurate representation of $g(t)$ was obtained by truncating only future values in the infinite sum as opposed to truncating both past and future values. The corresponding error bound is reduced to

$$|| u(t - (L + 1)h) - \bar{g}(t) || \leqq \frac{MK_0}{\sqrt{\delta}\, 2^\nu} + \frac{M}{\pi^2\, \delta L}. \tag{28}^*$$

This corresponds to the case $K = 1$ for the curves shown in Fig. 3.

Figure 5 is a plot of $\theta_{L+1}(t)$ and its transform

$$\Theta_{L+1}(\omega) = \int_{-\infty}^{\infty} \theta_{L+1}(t) \exp(-j\omega t)\, dt.$$

* This can easily be obtained by eliminating the second term in equation (38) which corresponds to truncation of past values.

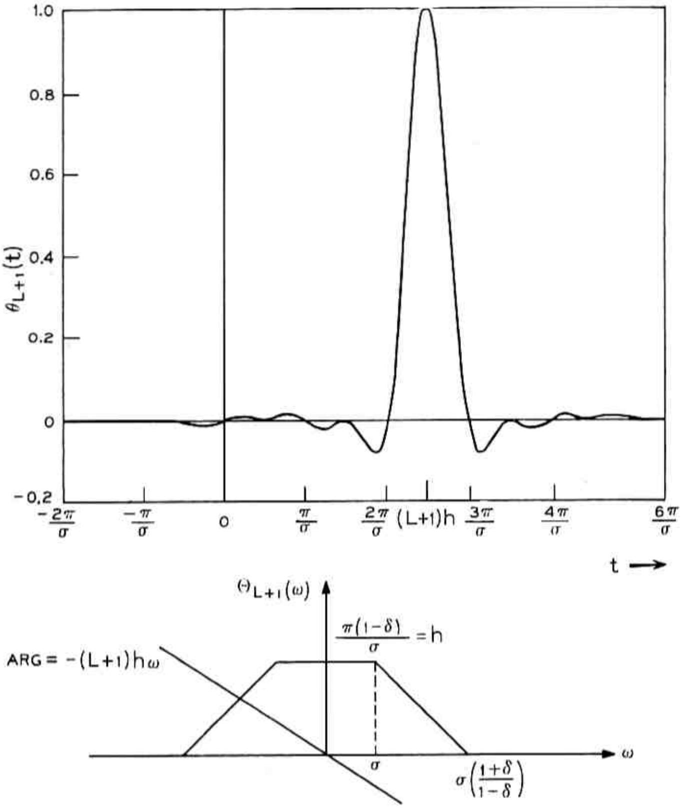


Fig. 5—Curve shown for $L = 4$, $\delta = 1/2$.

The causal impulse response $h(t)$ can be written as

$$h(t) = \theta_{L+1}(t) - \theta_{L+1}(t)S(-t)$$

where the second term represents the negative time tail of $\theta_{L+1}(t)$. The corresponding transform is

$$H(\omega) = \Theta_{L+1}(\omega) - \tilde{H}(\omega) \tag{29}$$

where $\tilde{H}(\omega)$ is the transform of $\theta_{L+1}(t)S(-t)$. The transform of the causal filter is thus seen to be a slight perturbation (for L sufficiently large) of the low-pass characteristic of $\Theta_{L+1}(\omega)$.

VI. DISCUSSION

In this paper, the design and analysis of a class of PCM systems has been considered. In arriving at these results, advantage has been taken of the fact that the sampling rate is higher than the Nyquist rate. The approach taken in this paper takes advantage of the alternate representations of the input class which are made possible by the higher sampling rate and chooses one for which a good bound on the effects of quantization errors can be derived.

In their present form the results which have been presented may be viewed as a step towards the study of peak error behavior of PCM systems. The extension of these results to input signals about which more information is available (e.g., power spectral information) is the next step towards providing a more generally applicable framework for the design and analysis of PCM systems where peak error is the natural criterion.

VII. ACKNOWLEDGMENTS

The authors would like to thank L. J. Forys for many helpful discussions. They are also indebted to Mrs. Sharon Miller for the excellent programming jobs she did in connection with this project.

APPENDIX

Derivation of Error Bound for UPCM System

The class B_σ of functions consists of entire functions of exponential order one and type σ which are bounded on the real axis.[4] It includes all functions in $\mathfrak{L}_2(-\infty, \infty)$ having finite radian bandwidth, σ, and all trigonometric polynomials of degree $[\sigma]$.*

Any element in B_σ has the representation[5]

$$u(t) = \sum_{j=-\infty}^{\infty} u(jh)\theta_j(t) \tag{30}$$

where

$$\theta(t) = \frac{\sin \dfrac{\delta\sigma}{1-\delta} t}{\dfrac{\delta\sigma}{1-\delta} t} \frac{\sin \dfrac{\sigma}{1-\delta} t}{\dfrac{\sigma}{1-\delta} t}, \tag{31}$$

$$\theta_j(t) = \theta(t - jh), \qquad h = \frac{\pi(1-\delta)}{\sigma}. \tag{32}$$

* [ ] denotes integral part. A trigonometric polynomial of degree $[\sigma]$ has the form $\frac{1}{2}a_0 + \sum_{k=1}^{[\sigma]} (a_k \cos kt + b_k \sin kt)$.

This representation is valid for any $\delta \in (0, 1)$, hence δ may be chosen appropriately for each application. The parameter δ will be called the fractional guardband since the time between samples is $\pi(1 - \delta)/\sigma$ which is less than the time between samples π/σ which corresponds to the Nyquist rate.

Let $\bar{u}(t)$ represent an approximation to $u(t)$ in the interval $[-T/2, T/2]$ obtained by truncating equation (30). Thus

$$\bar{u}(t) = \sum_{j=-N}^{N} u(jh)\theta_j(t) \tag{33}$$

and

$$u(t) - \bar{u}(t) = \sum_{|j|>N} u(jh)\theta_j(t). \tag{34}$$

Since $u \in B_\sigma(M)$, we have

$$\sup_{u \in B_\sigma(M)} \| u(t) - \bar{u}(t) \| \leqq M \sum_{|j|>N} | \theta_j(t) |. \tag{35}$$

Let $2N + 1 = T/h + 2L$, T/h odd, and $| j | > N$, then for $| t | \leqq T/2$,

$$| \theta_j(t) | \leqq \frac{1}{\pi^2 \delta} \frac{1}{\left(j - \frac{t}{h}\right)^2}. \tag{36}$$

Thus

$$\sup_{u \in B_\sigma(M)} \| u(t) - \bar{u}(t) \| \leqq \frac{M}{\pi^2 \delta} \sum_{j>N} \left\{ \left(j - \frac{t}{h}\right)^{-2} + \left(j + \frac{t}{h}\right)^{-2} \right\}. \tag{37}$$

Using the Sonin formula[7] to sum the remainder series in equation (37) yields

$$\sup_{u \in B_\sigma(M)} \| u(t) - \bar{u}(t) \|$$
$$\leqq \frac{M}{\pi^2 \delta} \left\{ \left(N + \frac{1}{2} - \frac{t}{h}\right)^{-1} + \left(N + \frac{1}{2} + \frac{t}{h}\right)^{-1} \right\}. \tag{38}$$

To obtain a uniform bound for $t \in [-T/2, T/2]$, we observe that the right side of equation (38) is maximized for $t = T/2$. Using this and $N + \frac{1}{2} = T/2h + L$ yields

$$\sup_{t \in [-T/2, T/2]} \sup_{u \in B_\sigma(M)} \| u(t) - \bar{u}(t) \| \leqq \frac{MK}{\pi^2 \delta L}, \tag{39}$$

where

$$K = \frac{\frac{T}{Lh} + 2}{\frac{T}{Lh} + 1}, \qquad 1 \leqq K \leqq 2.$$

Thus equation (39) represents a uniform bound on the truncation error associated with the finite sum approximation (33) to an arbitrary $u \varepsilon B_\sigma(M)$ over the time interval $-T/2 \leqq t \leqq T/2$. It is clear that this bound can be made arbitrarily small by appropriate choice of L, however the quantity $T + Lh$ represents the delay associated with the decoder and hence some compromise will generally be called for.

To evaluate the error due to quantizing the sample values, it is convenient to define the quantity $A(\delta)$ by

$$A(\delta) = \sup_{-\infty < t < \infty} \sum_{j=-\infty}^{\infty} | \theta_j(t) |. \tag{40}$$

$A(\delta)$ plays a crucial role in relating the effects of quantization error at sample times to the error between samples. In fact, one of the significant characteristics of the representation (30), which is not shared by the Cardinal series representation (for functions in $W_\sigma \subset B_\sigma$) is the ability to establish a useful bound on $A(\delta)$. To do this, we first apply the Cauchy-Schwarz inequality and obtain

$$A(\delta)^2 \leqq \sup_{-\infty < t < \infty} \sum_{j=-\infty}^{\infty} \left[\frac{\sin \frac{\delta\sigma}{1-\delta}(t - jh)}{\frac{\delta\sigma}{1-\delta}(t - jh)} \right]^2 \cdot \sup_{-\infty < t < \infty} \sum_{j=-\infty} \left[\frac{\sin \frac{\sigma}{}(t - jh)}{\frac{\sigma}{1-\delta}(t - jh)} \right]^2. \tag{41}$$

The Parseval theorem for functions in B_σ which are also in $\mathfrak{L}_2(-\infty, \infty)$ provides the following explicit evaluations

$$\sum_{j=-\infty}^{\infty} \left[\frac{\sin \frac{\sigma}{1-\delta}(t - jh)}{\frac{\sigma}{1-\delta}(t - jh)} \right]^2 = 1, \tag{42}$$

$$\sum_{j=-\infty}^{\infty} \left[\frac{\sin \frac{\delta\sigma}{1-\delta}(t - jh)}{\frac{\delta\sigma}{1-\delta}(t - jh)} \right]^2 = \frac{1}{\delta}. \tag{43}$$

Hence

$$A(\delta) \leqq \frac{1}{\sqrt{\delta}}. \tag{44}$$

Consider a quantizer with 2^ν output levels. If we denote by $g(t)$ the signal

$$g(t) = \sum_{j=-N}^{N} \hat{u}(jh)\theta_j(t), \tag{45}$$

where $\hat{u}(jh)$ represents the quantized value of $u(jh)$, then

$$u(t) - g(t) = \sum_{j=-N}^{N} (u(jh) - \hat{u}(jh))\theta_j(t) + \sum_{|j|>N} u(jh)\theta_j(t). \tag{46}$$

Since the maximum quantization error *at sample times* is $\sigma_q(\nu)$, using equations (44) and (39) we obtain

$$\sup_{t\epsilon[-T/2,T/2]} \sup_{u\epsilon B_\sigma(M)} || u(t) - g(t) || \leqq \frac{\sigma_q(\nu)}{\sqrt{\delta}} + \frac{MK}{\pi^2 \delta L} \triangleq \tilde{\epsilon}. \tag{47}$$

REFERENCES

1. Ruchkin, D. S., "Linear Reconstruction of Quantized and Sampled Random Signals," IRE Trans. Commun. Syst. *9*, No. 4 (December 1961), pp. 350–355.
2. Bennett, W. R., "Spectra of Quantized Signals," B.S.T.J., *27*, No. 3 (July 1948), pp. 446–472.
3. Totty, R. E., and Clark, G. C., "Reconstruction Error in Waveform Transmission," IEEE Trans. Infor. Theory, *13*, No. 2 (April 1967), pp. 336–338.
4. Achieser, N. I., *Theory of Approximation,* New York: Ungar, 1956, pp. 137–140.
5. Jagerman, D., "Information Theory and Approximation of Bandlimited Functions," B.S.T.J., *49,* No. 8 (October 1970), pp. 1915–1945.
6. Smith, B., "Instantaneous Companding of Quantized Signals," B.S.T.J., *36,* No. 3 (May 1967), pp. 653–709.
7. Vinogradov, I. M., *Elements of Number Theory,* New York: Dover, 1954, p. 31.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

Digital Phase Demodulator

By BERNARD GLANCE

(Manuscript received October 2, 1970)

Injection-locked oscillators are shown to act as narrow-band tunable filters for FM signals if the modulation rate is much larger than the locking bandwidth. The filtering action of the injection-locked oscillator for FM signals is found analogous to that of a high Q passive cavity. The effective Q of the injection-locked oscillator can be as high as 10^5 if the stability of the injected signal carrier and oscillator frequencies is better than 10^6.

These filtering properties can be applied to a digital demodulator for coherent phase detection of a coded FM signal. The local source which is required for coherent phase detection is provided by using a fraction of the received signal to lock an oscillator. Sideband suppression and carrier amplification of the injected signal are achieved simultaneously by using the filtering action of the injection-locked oscillator.

The simplicity of this digital demodulator makes it appear useful for repeaters in microwave radio relays.

I. INTRODUCTION

Injection-locked oscillators can perform a wide variety of functions required in microwave radio relays, such as amplification, amplitude limitation,[1] frequency modulation[2] and demodulation[3] to mention only the most important applications.

It is shown in this paper that injection-locked oscillators can also be used as narrow band tunable filters for angle modulated signals. These filtering properties can be used in a digital demodulator for coherent phase detection and such a demodulator is described here. Its configuration, shown in Fig. 1, is similar to the injection-locked oscillator FM receiver proposed by C. L. Ruthroff.[3] The principles of operation, however, are different. For proper operation of the injection-locked oscillator FM receiver the output signal of the oscillator contains all of the frequency modulation on the input signal. This signal is multiplied by a fraction of the input signal to

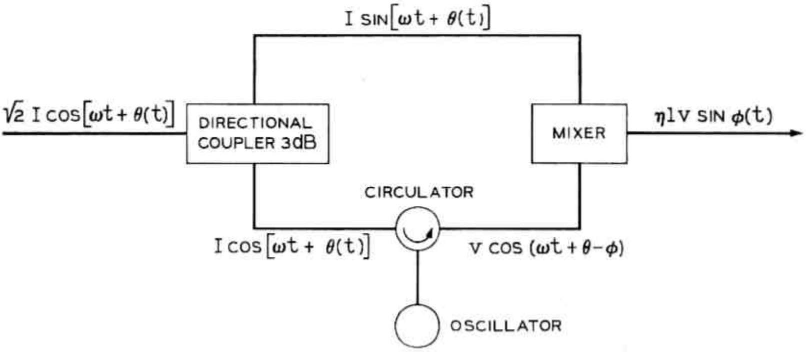


Fig. 1—Scheme of the digital demodulator for coherent phase detection.

produce the demodulated output. In the digital demodulator, the filtering properties of an injection-locked oscillator are used to remove the frequency modulation from the input signal and to deliver a sinusoidal output at the frequency and phase of the unmodulated carrier; this carrier is used as the local reference clock signal in a synchronous detector. The mixing of the received signal with the local source gives a current dependent on the phase of the received signal provided the time average of the phase modulation stays small over periods shorter than the time constant of the oscillator.

The first part of this paper includes an analysis of the filtering properties of the injection-locked oscillator for phase modulated signals. An approximate analytical solution for the filtering action is derived from the locking equation. This solution is then compared with exact numerical calculations and with experimental results.

The second part of this paper describes the properties and the limitations of a digital demodulator which uses the filtering action of the injection-locked oscillator analyzed in the first section.

II. ANALYSIS OF THE INJECTION-LOCKED OSCILLATOR FILTER

2.1 *Locking Equation Analysis*

The injection-locked oscillator performs two functions in the digital demodulator: it removes the phase modulation of the input signal and it amplifies the carrier signal up to the free-running oscillator output level. The oscillator output signal which is obtained can be used as a reference signal for synchronous detection.

The filtering properties can be derived from the locking equation. These properties can also be obtained from a frequency domain analysis which provides a clearer physical picture of the phenomena.

Let us consider first the locking equation analysis. Assume that the injection-locked oscillator is driven by an injected signal phase modulated by a function $\theta(t)$, thus $i(t) = I \cos [\omega t + \theta(t)]$. Locking occurs if ω is sufficiently close to the natural oscillator frequency ω_0. Within the locking range, the oscillator output voltage is given by $v(t) = V(t) \cos [\omega t + \theta(t) - \varphi(t)]$.[4] $\varphi(t)$, called the tracking angle, is the difference in phase between the input and output signals of the oscillator. The phases of the input and output signals are related by the well-known locking equation[5,6]

$$\frac{d\varphi}{dt} = \frac{d\theta}{dt} + \omega - \omega_0 - \Delta\omega_L \sin \varphi(t), \tag{1}$$

where

$$\Delta\omega_L = I/(2V(t)G_L) \times \omega_0/Q$$

is one-half the locking bandwidth for an unmodulated injected signal of amplitude I, G_L is the load and Q is the external loaded circuit Q. The output signal amplitude $V(t)$ is usually nearly constant[1] and therefore $\Delta\omega_L$ may be assumed to be time independent.

Removing the phase modulation from the injected signal requires that the phase of the oscillator output signal, $\theta(t) - \varphi(t)$, becomes time independent. It will be shown that this condition is approximately fulfilled if the rate of phase modulation is much larger than the locking bandwidth.

Let us consider an input signal with a sinusoidal phase modulation given by

$$\theta(t) = \theta_0 \sin \Omega t. \tag{2}$$

Substitution of equation (2) into equation (1) gives

$$\frac{1}{\Omega}\frac{d\varphi}{dt} - \theta_0 \cos \Omega t = \frac{\Delta\omega_L}{\Omega}\left[\frac{\omega - \omega_0}{\Delta\omega_L} - \sin \varphi(t)\right]. \tag{3}$$

With a rate of phase modulation $\Omega \gg \Delta\omega_L$, the right side of equation (3) remains much smaller than unity as long as $|\omega - \omega_0| \leqq \Delta\omega_L$. Equation (3) can therefore be approximated by

$$\frac{1}{\Omega}\frac{d}{dt}[\varphi(t) - \theta(t)] \approx 0 \tag{4}$$

which gives for the output signal a phase approximately time independent, thus

$$\varphi(t) - \theta(t) \approx \varphi_0 . \tag{5}$$

The constant of integration, φ_0, can be obtained from the locking condition which is obtained by taking the time average of equation (1) yielding

$$\left\langle \frac{d\varphi}{dt} \right\rangle = \left\langle \frac{d\theta}{dt} \right\rangle + \omega - \omega_0 - \Delta\omega_L \langle \sin \varphi(t) \rangle . \tag{6}$$

From equations (2) and (4),

$$\begin{aligned} \left\langle \frac{d\theta}{dt} \right\rangle &= 0, \\ \left\langle \frac{d\varphi}{dt} \right\rangle &= 0, \end{aligned} \tag{6a}$$

and

$$\begin{aligned} \langle \sin \varphi(t) \rangle &= \langle \sin (\varphi_0 + \theta_0 \sin \Omega t) \rangle \\ &= J_0(\theta_0) \sin \varphi_0 , \end{aligned}$$

where J_0 is the Bessel function of order zero. Substitution of equation (6a) into equation (6) gives the locking condition

$$\omega - \omega_0 = \Delta\omega_L J_0(\theta_0) \sin \varphi_0 . \tag{7}$$

The frequency range of locking is determined by the condition $| \sin \varphi_0 | \leqq 1$. Therefore the locking bandwidth is, from equation (7),

$$2(\omega - \omega_0)_{\max} = 2 \, \Delta\omega_L J_0(\theta_0). \tag{8}$$

The maximum locking bandwidth is reduced by the factor $J_0(\theta_0)$ compared with the unmodulated case; in particular, it becomes equal to zero for $\theta_0 = 2.405$ radians.

These results have been obtained from a first-order solution of equation (3). The magnitude of the filtering effect, resulting from the residual phase modulation $\theta(t) - \varphi(t)$ of the oscillator output signal, can be calculated by solving equation (3) to the second order.

Before calculating the magnitude of the filtering effect, it is interesting to give a physical picture of this phenomenon through a frequency domain analysis. Experimental results observed with a spectrum analyzer will be compared with the results of this analysis.

2.2 *Frequency Domain Discussion*

Let us consider as before an oscillator locked by an injected signal with a sinusoidal phase modulation, thus

$$i(t) = I \cos [\omega t + \theta_0 \sin \Omega t]. \tag{9}$$

The current expression, expanded in Bessel series, can be written as[7]

$$i(t) = I\left\{J_0(\theta_0) \cos \omega t + \sum_{\substack{-\infty \\ n\neq 0}}^{+\infty} J_n(\theta_0) \cos (\omega + n\Omega)t\right\}. \tag{10}$$

If one assumes that the carrier frequency ω is within the locking bandwidth $\Delta\omega_L$ and that $\Omega \gg \Delta\omega_L$, the spectral components at $\omega + n\Omega$ ($n = \pm 1, \pm 2, \cdots$ etc) have a small effect on the oscillator. It can be expected in a first approximation that the oscillator is locked by the injected current component $IJ_0(\theta_0) \cos \omega t$ corresponding to the carrier frequency. The locking bandwidth, which is proportional to the effective driving current amplitude $IJ_0(\theta_0)$, is reduced by the factor $J_0(\theta_0)$ as found previously in Section 2.1. The locking bandwidth decreases with increasing index of modulation and becomes equal to zero for $\theta_0 = 2.405$. This particular case corresponds to an injected FM signal with a suppressed carrier.

The sideband suppression effect, of the injection-locked oscillator, can be seen clearly in this analysis as an increasing function of the rate of phase modulation. Furthermore it can be expected from this analysis that locking can also occur for any of the spectral components at $\omega + n\Omega$ ($n = \pm 1, \pm 2, \cdots$). Locking can be obtained by tuning the oscillator natural frequency to $\omega + n\Omega$. The locking range for each frequency $\omega + n\Omega$ is proportional to the spectral line amplitude, $IJ_n(\theta_0)$.

2.3 *Experimental Verification*

The validity of the assumptions made in this analysis has been checked by locking a 35-MHz oscillator with an FM injected signal. The index of modulation, the rate of modulation and the locking bandwidth were adjusted to be about $\pi/2$, 100 kHz, and 10 kHz, respectively.

Stable locking was obtained by tuning the injected signal frequency such that the main spectral lines, $J_0(\pi/2) \cos \omega t$, $J_{\pm 1}(\pi/2) \cos (\omega \pm \Omega)t$ and $J_{\pm 2}(\pi/2 \cos (\omega \pm 2\Omega)t$ lie consecutively in the locking range. Figure 2 shows the locking obtained with the three first spectral components and also shows the characteristic beat modulation which occurs just before locking. The largest locking ranges correspond as expected to the spectral components $J_{\pm 1}(\pi/2) \cos (\omega \pm \Omega)t$ which have the largest amplitudes.

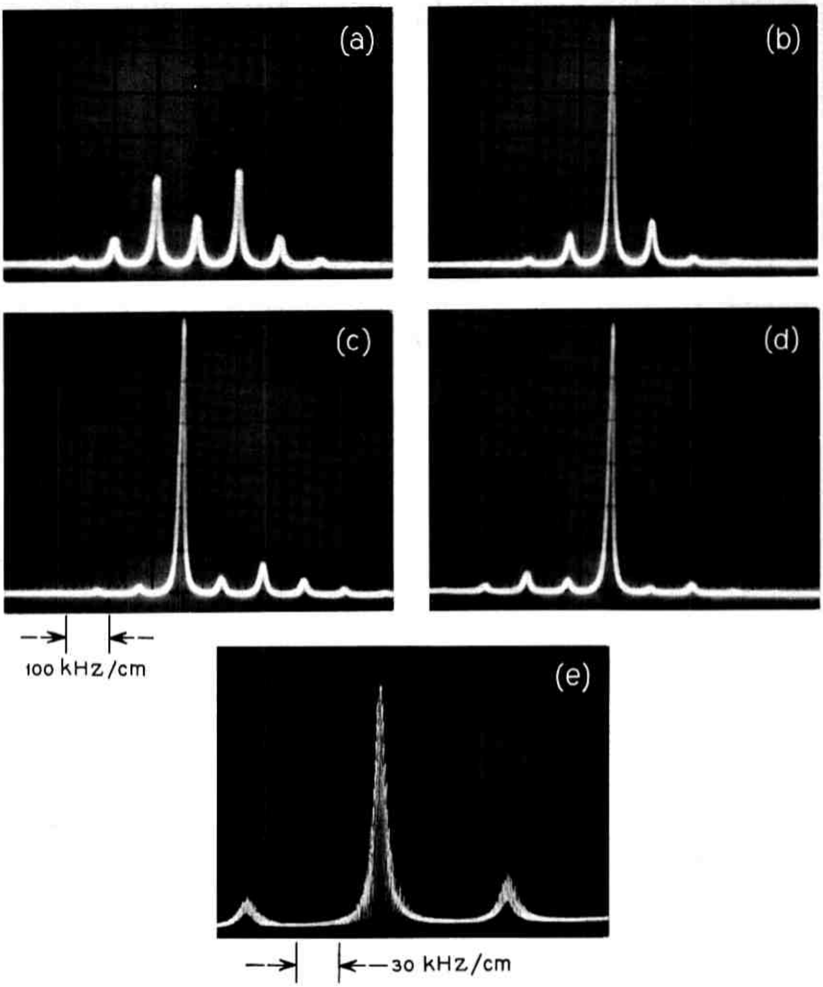


Fig. 2—Spectra of the injected signal and the oscillator output signal: (a) injected signal spectrum, (b) output signal with locking at ω, (c) output signal with locking at $\omega - \Omega$, (d) output signal with locking at $\omega + \Omega$, and (e) output signal just before locking. ($\theta \approx \pi/2 \sin \Omega t$, $\Omega = 2\pi \times 100$ kHz, $\Delta\omega_L \approx 2\pi \times 10$ kHz, $F_0 = 35$ MHz)

2.4 *Sideband Attenuation*

The filtering effect of the injection-locked oscillator suppresses the sidebands of the output signal. This effect is shown in Fig. 2 where the spectra of the input and output signals of the oscillator can be compared.

Sideband attenuation can be defined by

$$\left(\frac{\text{Sideband Power}}{\text{Carrier Power}}\right)_{\text{output}} \Big/ \left(\frac{\text{Sideband Power}}{\text{Carrier Power}}\right)_{\text{input}}.$$

This ratio is related to the residual phase modulation $\theta(t) - \varphi(t)$ which can be calculated by solving equation (3) to the second order. The calculations are given in Appendix A for $\theta_0 = \pi/2$. The steady-state solution is

$$\varphi(t) \simeq \frac{\pi}{2}\sin\Omega t + \varphi_0 + 2\frac{\Delta\omega_L}{\Omega}\left\{J_1\left(\frac{\pi}{2}\right)\cos\varphi_0\cos\Omega t - \frac{J_2\left(\frac{\pi}{2}\right)}{2}\sin\varphi_0\sin 2\Omega t\right\}, \tag{11}$$

where from equation (7)

$$\sin\varphi_0 = \frac{\omega - \omega_0}{J_0\left(\frac{\pi}{2}\right)\Delta\omega_L}. \tag{12}$$

For simplicity let us assume that $\omega = \omega_0$, thus the injected current is

$$i(t) = I\cos\left[\omega_0 t + \frac{\pi}{2}\sin\Omega t\right]. \tag{13}$$

The tracking angle becomes

$$\varphi(t) \simeq \frac{\pi}{2}\sin\Omega t + 2\frac{\Delta\omega_L}{\Omega}J_1\left(\frac{\pi}{2}\right)\cos\Omega t, \tag{14}$$

and the oscillator output voltage is

$$v(t) = V\cos\left[\omega_0 t - 2\frac{\Delta\omega_L}{\Omega}J_1\left(\frac{\pi}{2}\right)\cos\Omega t\right]. \tag{15}$$

The expressions of the input and output power, expanded in Bessel series, yield for the ratio of sideband power to carrier power:

a) *Input signal*

$$\left(\frac{\text{Sideband Power}}{\text{Carrier Power}}\right)_{\text{input}} = \frac{1 - J_0^2\left(\frac{\pi}{2}\right)}{J_0^2\left(\frac{\pi}{2}\right)} = 3.484. \tag{16}$$

b) *Output signal*

$$\left(\frac{\text{Sideband Power}}{\text{Carrier Power}}\right)_{\text{output}} = \frac{1 - J_0^2\left[2\,\frac{\Delta\omega_L}{\Omega}\,J_1\left(\frac{\pi}{2}\right)\right]}{J_0^2\left[2\,\frac{\Delta\omega_L}{\Omega}\,J_1\left(\frac{\pi}{2}\right)\right]} = 0.0064. \tag{17}$$

The calculated sideband attenuation, in the present case, is about 27.4 dB.

In order to verify these results, equation (3) has been solved numerically for the following parameter values: $\theta_0 = \pi/2$, $(\omega - \omega_0)/\Omega = -0.01$ and $(\Delta\omega_L)/\Omega = 0.1$. Results of this computation are shown in Fig. 3 and are compared in Table I with the second-order approximation given by equation (11).

2.5 *Comparison Between the Active and Passive Resonators*

In the case of an injected current phase-modulated by a sinusoidal phase excursion, the filtering properties of the injection-locked oscillator characterized by a single-pole resonator with a negative resistance can be compared to the filtering effect of the same passive circuit.

Assuming an injected current equal to $I \cos(\omega_0 t + \pi/2 \sin \Omega t)$ the power ratio between the spectral component at $\omega_0 + \Omega$ and the carrier at ω_0 is, for the passive resonator,

$$\frac{1}{1 + 4Q^2\left(\frac{\Omega}{\omega_0}\right)^2}\left[\frac{J_1\left(\frac{\pi}{2}\right)}{J_0\left(\frac{\pi}{2}\right)}\right]^2 \approx \frac{1}{4Q^2}\left(\frac{\omega_0}{\Omega}\right)^2\left[\frac{J_1\left(\frac{\pi}{2}\right)}{J_0\left(\frac{\pi}{2}\right)}\right]^2. \tag{18}$$

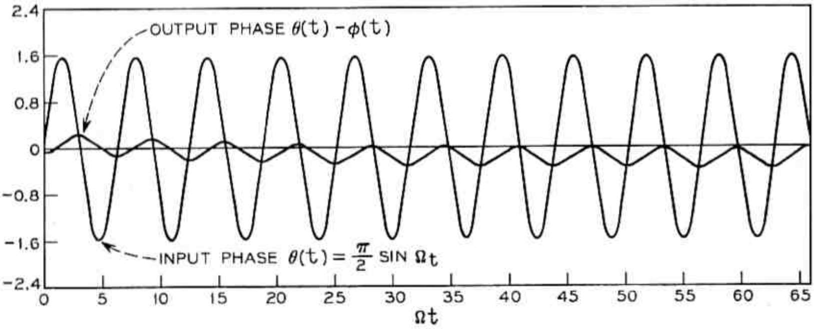


Fig. 3—Phase of the oscillator output signal for an injected signal phase modulated by $\theta(t) = \pi/2 \sin \Omega t$. $[(\Delta\omega_L)/\Omega = 0.1,\ (\omega - \omega_0)/\Omega = -0.01,\ \theta_0 = \pi/2.]$

Table I—Comparison Between the Analytic Approximation Solution and the Computed Solution

Analytic Second-Order Approximation Solution	Numerically Computed Solution
$\sin \varphi_0 \simeq \dfrac{\omega - \omega_0}{J_0(\pi/2)\Delta\omega_L} = -0.212$	$\sin \varphi_0 = -0.199$
$\lvert\varphi - \varphi_0 - \theta\rvert_{\max} = 0.112$	$\lvert\varphi - \varphi_0 - \theta\rvert_{\max} = 0.160$
Sideband attenuation = 27.4 dB	Sideband attenuation = 23.7 dB

Taking into account that $\Delta\omega_L = \omega_0/Q(P_i/P_0)^{1/2}$, the same ratio for the locked oscillator gives, for $\Omega \gg \Delta\omega_L$,

$$\frac{1}{Q^2}\left(\frac{\omega_0}{\Omega}\right)^2 J_1^2\left(\frac{\pi}{2}\right)\left[\frac{P_i}{P_0}\right] \tag{19}$$

where P_0/P_i is the locking gain. For an FM injected signal having a sinusoidal phase excursion of amplitude $\pi/2$, the locked oscillator acts as a singly resonant filter with an effective Q given by

$$Q_e = Q\left(\frac{P_0}{P_i}\right)^{\frac{1}{2}} \frac{1}{2J_0\left(\frac{\pi}{2}\right)} \simeq Q\left(\frac{P_0}{P_i}\right)^{\frac{1}{2}} = \left[\frac{\Delta\omega_L}{\omega_0}\right]^{-1}. \tag{20}$$

The maximum effective Q which can be obtained depends on the minimum-locking bandwidth achievable. The minimum-locking bandwidth is limited by

(*i*) minimum frequency offset $\omega - \omega_0$,
(*ii*) oscillator free running frequency stability $\delta\omega_0/\omega_0$,
(*iii*) injected signal frequency stability $\delta\omega/\omega$.

The slow variation of the frequency offset $\omega - \omega_0$, between the injected signal frequency and the oscillator free-running frequency, can be made approximately equal to zero with a low-frequency feedback loop as shown in Ref. 3. In that scheme, the mixer output signal contains a current proportional to $\omega - \omega_0$; the oscillator natural frequency ω_0 can be kept tuned to ω by using this current to control a suitable oscillator parameter.

If one assumes a frequency stability of 10^{-6} for ω and ω_0 and takes a safety margin $\Delta\omega_L = 10\ \delta\omega$, one obtains a maximum effective Q given by

$$Q_{e\max} \simeq \frac{1}{10}\left[\frac{\Delta\omega_L}{\omega}\right]^{-1} = 10^5.$$

III. DIGITAL DEMODULATOR

3.1 *Coherent Phase Detection*

Coherent phase detection involves multiplication of the PM received signal by a local source which has a constant phase and the frequency of the carrier associated with the received signal. The local source, which is needed to achieve coherent phase detection, can be obtained by using the filtering properties of the injection-locked oscillator for phase-modulated signals.

A block diagram of the digital demodulator is shown in Fig. 1. The oscillator, in this configuration, is locked by a fraction of the received signal. The filtering properties of the injection-locked oscillator are used to remove the phase modulation from the input signal and to deliver a sinusoidal output at the frequency and phase of the unmodulated carrier; this carrier is used as the local reference signal for synchronous detection. The mixing of the received signal with the local source gives a current dependent on the phase of the received signal. Correct demodulation is obtained if the time average of the phase modulation stays small over periods shorter than $\Delta\omega_L^{-1}$. This restriction results from the impossibility of maintaining the phase of the reference signal constant if the phase of the injected signal has an average value different from zero. This problem arises for instance with a long pulse sequence of plus ones made by binary digital encoding using polar pulses.

3.2 *Demodulation of a Binary Polar Signal*

Let us consider an unmodulated carrier given by cos ωt. Starting at $t = 0$, the phase is modulated by a pulse train of plus ones made from raised cosines of maximum amplitude equal to $\pi/2$. The phase of the received signal can be written for $t \geqq 0$

$$\begin{aligned}\theta(t) &= \frac{\pi}{2}\sin^2 \Omega t\\ &= \frac{\pi}{4}[1 - \cos 2\Omega t]\end{aligned} \tag{21}$$

where 2Ω is the signaling frequency equal here to the rate of phase modulation. The phase modulation of the received signal has an average value equal to $\pi/4$. The phase of the local source $\theta(t) - \varphi(t)$ at $t > 0$ is obtained by the transient solution of the equation

$$\frac{d\varphi}{d(\Omega t)} = \frac{\pi}{2}\sin 2\Omega t - \frac{\Delta\omega_L}{\Omega}\sin\varphi(t), \tag{22}$$

where $\omega - \omega_0$ has been set equal to zero in order to simplify the analysis.

The filtering condition $\Delta\omega_L/\Omega \ll 1$ suggests that the time constant associated with the transient solution of equation (22) is large compared to $2\pi/\Omega$. An equation giving approximately the transient effect can be obtained by taking the time average of equation (22) over a large number of periods of $\sin 2\Omega t$. Equation (22) then becomes

$$\frac{d\langle\varphi\rangle}{dt} = -\Delta\omega_L\langle\sin\varphi\rangle. \tag{23}$$

Equation (23) can be solved approximately by replacing $\langle\sin\varphi\rangle$ by $\langle\varphi\rangle$ which yields

$$\langle\varphi\rangle \simeq \frac{\pi}{4}\exp(-\Delta\omega_L t). \tag{24}$$

The transient solution of equation (22) is therefore approximated by

$$\varphi(t) \simeq \frac{\pi}{4}[\exp(-\Delta\omega_L t) - \cos 2\Omega t], \tag{25}$$

where $-\pi/4 \cos 2\Omega t$ is the first order steady state solution of equation (22). The phase of the local source is given for $t \geqq 0$ by

$$\theta(t) - \varphi(t) \simeq \frac{\pi}{4}[1 - \exp(-\Delta\omega_L t)]. \tag{26}$$

An exact numerical solution, shown in Fig. 4, agrees well with the analytical solution given by equation (26). This result shows that the phase of the filtered signal used as a local source increases exponentially from zero to $\pi/4$ with a time constant equal to about $\Delta\omega_L^{-1}$.

Correct demodulation requires that the magnitude of the phase of the reference signal remains small compared to $\pi/4$. The maximum number of consecutive pulses of the same polarity, which can be decoded without error, increases with the ratio $\Omega/\Delta\omega_L$. It is important to note that the sideband suppression effect improves by the same factor.

In general, the pulse polarity varies in a nearly random fashion from pulse to pulse. The phase of the reference signal is then a function of the random processes which give the pulse polarity distribution. Its maximum magnitude variation can be equal to $\pm\pi/4$. It will remain smaller than $|\pi/4|$ if the probability of having positive or negative pulses is about the same over periods shorter than $\Delta\omega_L^{-1}$. This condition introduces some restriction on the coding.

3.3 *Demodulation of a Signal Symmetrically Phase Modulated*

The problem of the reference phase discussed in the previous section disappears in case the average phase deviation is equal to zero for each

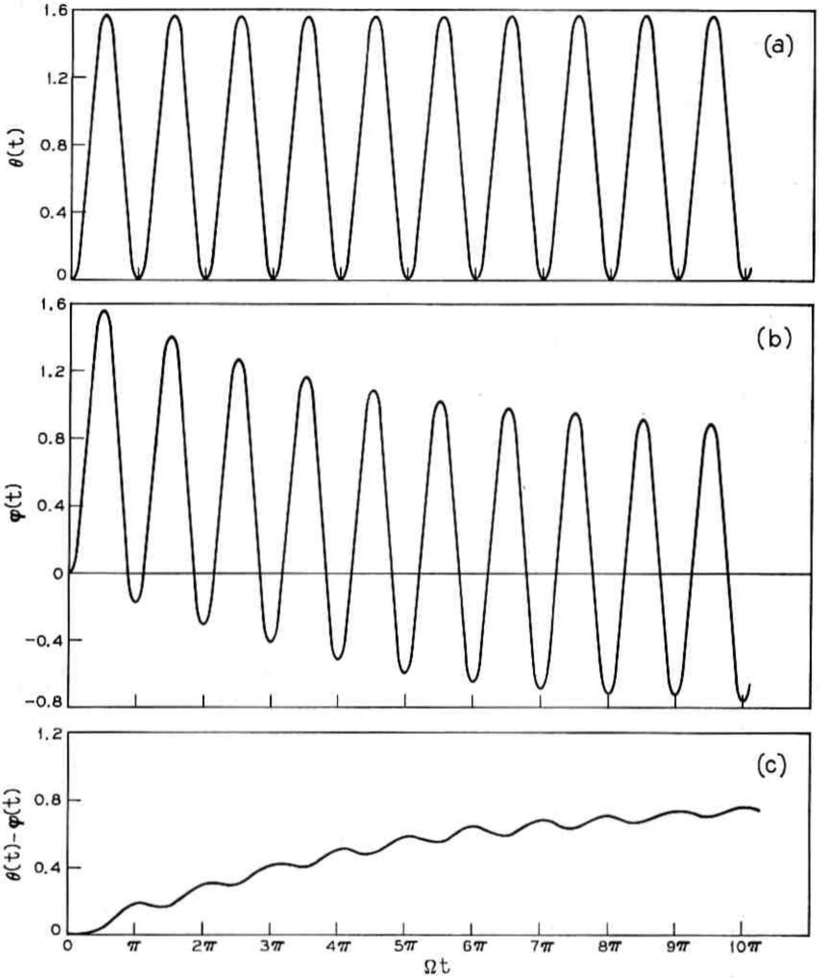


Fig. 4—Tracking angle and phase of the oscillator output signal for an injected signal phase modulated by $\theta(t) = (\pi/2) \sin^2 \Omega t$ with $(\Delta\omega_L)/\Omega = 0.1$. (a) Phase of the input signal, (b) tracking angle, (c) phase of the oscillator output signal.

pulse. A simple example of such a pulse shape is a sine wave starting from zero and limited to one period. The phase of the received signal can be written in this case[7]

$$\theta(t) = \sum_{k=-\infty}^{+\infty} a(k) \frac{\pi}{2} \sin \Omega(t - kT), \tag{27}$$

where $a(k)$ takes, for example, the value 0 or +1 according to a binary code. The phase of the received signal becomes, for a pulse train of +1s

$$\theta(t) = \frac{\pi}{2} \sin \Omega t. \tag{28}$$

This case has been solved in Section II, equation (11); it gives, if $(\omega - \omega_0)/\Delta\omega_L \ll 1$, a local source with a phase equal to

$$\theta(t) - \varphi(t) \approx \frac{\omega - \omega_0}{J_0\left(\frac{\pi}{2}\right)\Delta\omega_L} + 2\,\frac{\Delta\omega_L}{\Omega}\, J_1\left(\frac{\pi}{2}\right) \cos \Omega t. \tag{29}$$

This result is shown in Fig. 3.

The phase of the local source can be made nearly constant and adjusted close to zero by making $(\omega - \omega_0)/\Delta\omega_L$ and $(\Delta\omega_L)/\Omega$ very small.

This phase shift is also calculated for other types of pulse distributions. Its effect is shown in Fig. 5 which gives the tracking angle $\varphi(t)$ and the demodulated signal $\sin[\varphi(t)]$ for an injected signal phase modulated by a pulse train of alternate ones and zero. These curves are calculated for the parameters $(\Delta\omega_L)/\Omega = 0.1$ and $\omega - \omega_0/\Delta\omega_L = \pm 0.1$. Figure 6 shows the same functions for a random phase modulation with the same parameter values. In these two cases, the distortion due to the phase shift of the local source is minimum for $(\omega - \omega_0)/\Delta\omega = -0.1$. Partial compensation is then obtained between the phase shift resulting from the frequency offset and the phase shift due to the residual of the filtering effect. In all cases the distortion can be minimized by setting $(\Delta\omega_L)/\Omega$ and $(\omega - \omega_0)/\Delta\omega_L$ small compared to unity. Correct demodulation can then be obtained without an encoding restriction.

IV. CONCLUSION

Injection-locked oscillators can be used as filters for sideband suppression of FM signals if the modulation rate is much larger than the locking range. These filtering properties can be summarized as

(*i*) high effective Q,
(*ii*) power amplification for the carrier, and
(*iii*) circuit simplicity.

The filtering properties of an injection-locked oscillator can be used in a digital demodulator to provide a local source for coherent phase detection of a particular class of digitally modulated signals. In particular, correct demodulation is obtained for pulse shapes which give an average phase deviation equal to zero for each pulse.

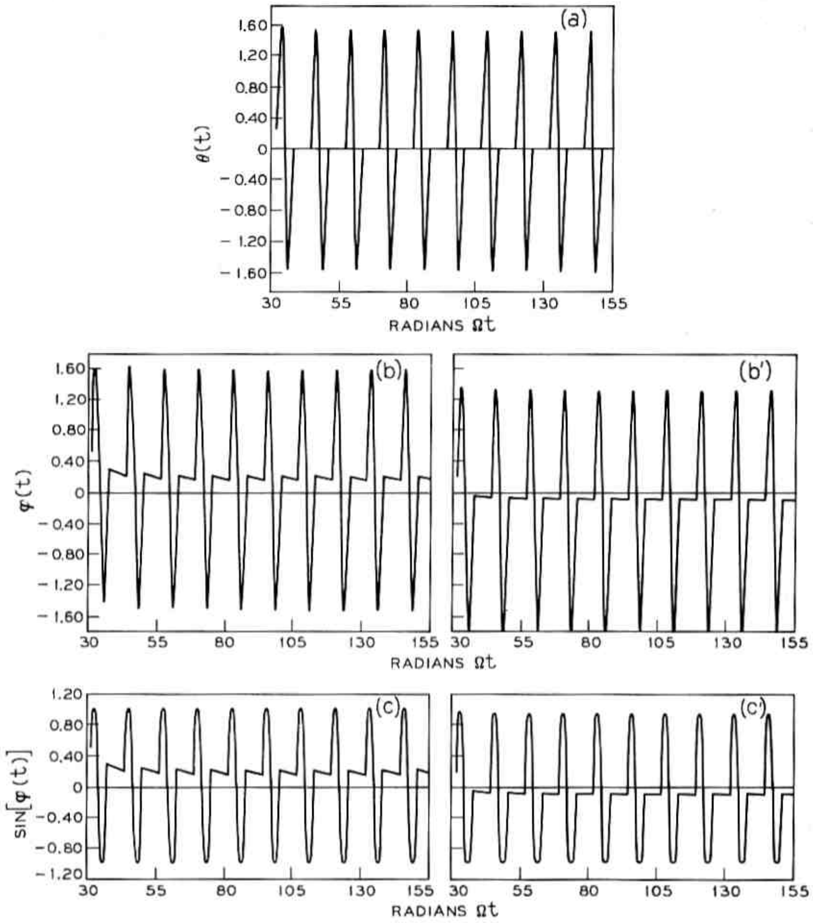


Fig. 5—Tracking angle and demodulated signal for an injected signal phase modulated by

$$\theta(t) = \frac{\pi}{2} \sin \Omega t \left[\frac{1}{2} + \frac{2}{\pi} \sum_{0}^{\infty} \frac{\sin (2p + 1)\Omega t/2}{2p + 1} \right]$$

which corresponds to a pulse train of 1, 0, 1, 0, $\cdots$. (a) Input modulation, $\theta(t) = (\pi/2) \sin \Omega t \{1/2 + 2/\pi \sum_{0}^{\infty} [\sin (2p + 1)(\Omega t/2)]/(2p + 1)\}$. (b) Tracking angle $\phi(t)$ and (c) output mixer $\sin \phi(t)$, $(\Delta\omega_L/\Omega) = 0.1$ and $(\omega - \omega_0/\Omega) = 0.01$. (b′) Tracking angle $\phi(t)$ and (c′) output $\sin \phi(t)$, $(\Delta\omega_L/\Omega) = 0.1$ and $(\omega - \omega_0/\Omega) = -0.01$.

A binary digital coding using polar pulses requires a coding which gives a small average phase deviation over periods shorter than $\Delta\omega_L^{-1}$

V. ACKNOWLEDGMENT

I am grateful to C. L. Ruthroff for helpful comments, and to Mrs. C. L. Beattie for making the numerical calculations.

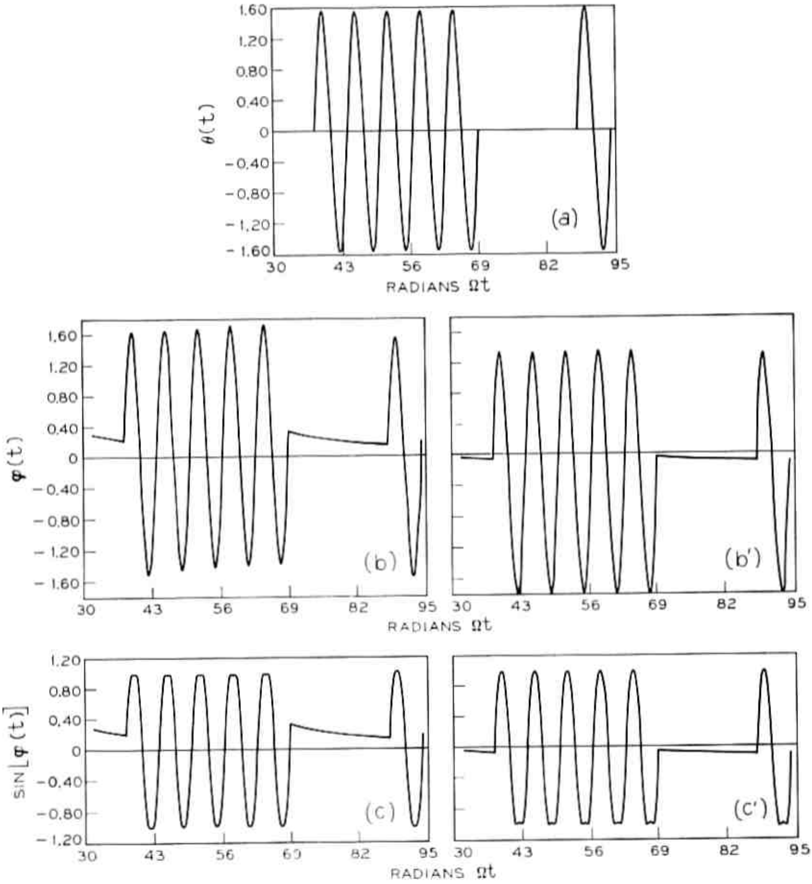


Fig. 6—Tracking angle and demodulated signal for an injected signal phase modulated by $\theta(t) = \Gamma(t)\pi/2 \sin \Omega t$ where $\Gamma(t)$ is a random telegraphic signal equal to +1 or zero with signaling frequency equal to Ω. (a) Random input modulation $\theta(t)$. (b) Tracking angle $\phi(t)$ and (c) mixer output $\sin \phi(t)$, $(\Delta\omega_L/\Omega) = 0.1$ and $(\omega - \omega_0/\Omega) = 0.01$. (b′) Tracking angle $\phi(t)$ and (c′) mixer output $\sin \phi(t)$, $(\Delta\omega_L/\Omega) = 0.1$ and $(\omega - \omega_0/\Omega) = -0.01$.

APPENDIX

Second-Order Solution of the Locking Equation

The locking equation (3), obtained for an injected signal phase modulated by $\theta(t) = \pi/2 \sin \Omega t$, can be written as:

$$\frac{d\varphi}{dx} = \frac{\pi}{2} \cos x + \beta - \alpha \sin \varphi \tag{30}$$

where $n = \Omega t$, $\beta = (\omega - \omega_0)/\Omega$ and $\alpha = (\Delta\omega_L)/\Omega$.

Equation (30) integrated for $\beta < \alpha \ll 1$ has a solution in the first approximation given by

$$\varphi_1 = \frac{\pi}{2} \sin x + \varphi_0 \tag{31}$$

with

$$\beta = \alpha J_0\left(\frac{\pi}{2}\right) \sin \varphi_0 \, . \tag{32}$$

The second-order approximation is calculated by putting $\varphi_2 = \varphi_1 + \eta$, with $\eta \sim \alpha \ll 1$. Substitution of these results into equation (30) gives, keeping the first-order terms in α,

$$\frac{d\eta}{dx} + \alpha\eta\left\{\left[J_0\left(\frac{\pi}{2}\right) + 2J_2\left(\frac{\pi}{2}\right) \cos 2x\right] \cos \varphi_0 - 2J_1\left(\frac{\pi}{2}\right) \sin x \sin \varphi_0\right\}$$
$$= -2\alpha\left\{J_1\left(\frac{\pi}{2}\right) \cos \varphi_0 \sin x + J_2\left(\frac{\pi}{2}\right) \sin \varphi_0 \cos 2x\right\}. \tag{33}$$

The approximate solution of equation (33) is

$$\eta = 2\alpha\left\{J_1\left(\frac{\pi}{2}\right) \cos \varphi_0 \cos x - \frac{J_2\left(\frac{\pi}{2}\right)}{2} \sin \varphi_0 \sin 2x\right\}$$
$$+ C \exp\left[-\alpha\left\{J_0\left(\frac{\pi}{2}\right)x + J_2\left(\frac{\pi}{2}\right) \cos \varphi_0\right.\right.$$
$$\left.\left. \cdot \sin 2x + 2J_1\left(\frac{\pi}{2}\right) \sin \varphi_0 \sin x\right\}\right], \tag{34}$$

which gives for φ_2 after the initial transient

$$\varphi_2 = \frac{\pi}{2}\sin x + \varphi_0 + 2\alpha\left\{J_1\left(\frac{\pi}{2}\right)\cos\varphi_0\cos x - \frac{J_2\left(\frac{\pi}{2}\right)}{2}\sin\varphi_0\sin 2x\right\}. \tag{35}$$

REFERENCES

1. Osborne, T. L., "Amplitude Behavior of Injection-Locked Oscillators," unpublished work.
2. Bodtmann, W. F., and Ruthroff, C. L., "A Linear Phase Modulator for Large Baseband Bandwidths," B.S.T.J., *49*, No. 8 (October 1970), pp. 1893–1903.
3. Ruthroff, C. L., "Injection-Locked Oscillator—FM Receiver Analysis," B.S.T.J., *47*, No. 8 (October 1968), pp. 1653–1661.
4. Van der Pol, B., "The Nonlinear Theory of Electric Oscillators," Proc. IRE, *22*, No. 9 (September 1934), pp. 1051–1081.
5. Adler, R., "A Study of Locking Phenomena in Oscillators," Proc. IRE, *34*, No. 6 (June 1964), pp. 351–357.
6. Kurokawa, K., "Noise in Synchronized Oscillators," IEEE Trans. Microwave Theory and Techniques, *16*, No. 4 (April 1968).
7. Rowe, H. E., *Signals and Noise in Communications Systems,* D. Van Nostrand Company, 1965.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

Self-Synchronizing Sequential Coding with Low Redundancy

By PETER G. NEUMANN

(Manuscript received August 31, 1970)

In this paper, we present sequential codes which have interesting properties in three respects. First, these codes may be used to achieve low redundancy (e.g., bandwidth compression through coding) by employing a multiplicity of variable-length codes to encode transitions between successive source symbols. Second, the coding complexity is surprisingly low. Third, many of these codes have exceedingly good intrinsic recoverability properties following errors. These codes compare favorably with a difference code environment in which the differences between successive source symbols are encoded. The scope of the sequential codes presented here includes, but is much wider than, difference code schemes. Where comparable, the sequential codes have slightly greater complexity and may have lower redundancy. They normally have vastly superior error recovery. These codes are applicable in situations such as video transmission in which the message source is highly correlated and where errors can be tolerated for a short period of time.

I. INTRODUCTION

In several previous papers, the author has pursued two apparently separate paths of development. The first path involves classes of slightly suboptimal variable-length prefix codes[1,2] whose self-synchronizing abilities are vastly superior to the optimal (Huffman) codes which minimize redundancy. The second path involves self-synchronizing sequential codes using information-lossless sequential machines as encoders and decoders.[3,4] In this paper, these two paths of development are joined. The result produces highly efficient sequential codes (with low redundancy) which have good self-synchronizing abilities and surprisingly low decoding complexity. These codes are applicable in situations in which the message source is highly correlated.

1.1 *Difference Codes*

Given an environment in which successive source signals are likely to be similar (or at least strongly correlated), considerable compression can be obtained by encoding the level difference between successive quantized signal levels (temporally or spatially). Such is the situation in a video picture environment, for example, with respect to adjacent points horizontally or vertically, or even to the same point in successive frames. By encoding differences, however, any error in quantizing, encoding, transmitting, decoding or reconstructing the image tends to persist. That is, once an error has occurred in a signal level, subsequent levels will continue to be in error by the same offset, unless terminated by boundary effects or by compensating errors. In an encoding of level differences between successive points in a line, for example, errors tend to propagate until the end of the line; in an encoding of level differences between the same point in successive frames, on the other hand, errors may continue forever. In order to prevent such error effects from propagating indefinitely, it may be necessary to terminate the propagation forcibly, for example by transmitting periodically the set of signal levels for the entire frame ("replenishment") rather than their frame-to-frame differences. Thus the use of difference coding for compression may be compromised by the need to resynchronize. This is true in general of frame-to-frame difference codes. An example of the use of such codes in a differential pulse-code modulation environment is given in Ref. 5.

II. SELF-SYNCHRONIZATION

The main purpose of this paper is to present codes which have compression capabilities at least as good as difference codes, along with roughly comparable decoding complexity, as well as having rather remarkable intrinsic self-synchronization properties. (These codes are in fact much more general than difference codes in terms of compression capabilities.) These codes recover quickly from the effects of errors in that arbitrary errors give incorrect results for a period of time, after which the entire system resumes correct operation without any explicit effort. (Note that self-synchronization is a property of the coding scheme, and should not be confused with video picture frame synchronization.) It is important to note that in such a scheme the errors are not corrected (in the sense of error-correcting codes); instead errors are *tolerated*, with the expectation that their effect will cease quickly. Video coding is an example where such an approach is reasonable since loss of information for a short period of time can often be tolerated.

2.1 *Self-Synchronization in Variable-Length Codes*

The use of variable-length codes for reducing redundancy is well understood. For example, D. A. Huffman[6] shows how to obtain a code which minimizes the transmitted information for a given independent distribution of source symbols. However, there has been relatively little quantitative concern for the effects of errors on these codes. In earlier papers[1,2] the author has shown that a slight sacrifice in efficiency (i.e., a slight increase in transmitted information) can be rewarded with tremendous gains in self-synchronizing capability. Some of this work is required here and is reviewed briefly, although from a different viewpoint.

A *code* is a collection of sequences of digits (code digits), each sequence being called a *code word*. *Code text* is obtained by concatenating code words. An *encoding* is a mapping of source symbols $S(i)$ onto code words $W(i)$. A code is a prefix code if and only if no code word occurs as the beginning (prefix) of any other code word. Thus in prefix code text, a code word can be decoded as soon as it is received, even though there are no explicit interword markers. A code is *exhaustive* if and only if every sequence of code digits is the prefix of some code text (i.e., of some sequence of code words). (A uniquely decodable code must be a prefix code if it is exhaustive.[7]) A sequence of code digits is a *synchronizing sequence* for a given code if the occurrence of the end of that sequence in (correct) code text must correspond to the end of a code word (although not necessarily to a particular code word), irrespective of what preceded that sequence. M. P. Schützenberger[8] and E. N. Gilbert and E. F. Moore[7] have shown that most exhaustive prefix codes tend to resynchronize themselves following loss of synchronization (e.g., after arbitrary errors, or at start-up). If an exhaustive code has at least one synchronizing sequence, then the code tends to resynchronize itself following errors with a finite average delay (assuming a suitable randomness). All codes considered here are exhaustive unless explicitly stated otherwise. Note that resynchronization is an intrinsic property of the code, and no externally implied synchronization is required. Synchronization following ambiguity occurs as a result of any synchronizing sequence occurring naturally in code text.

As an example, consider the code of Fig. 1, consisting of the five code words 00, 01, 10, 110, 111. The tree of Fig. 1 may be interpreted

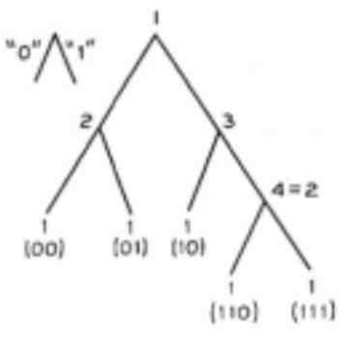


Fig. 1—A simple prefix code.

as the state diagram for a sequential machine[9] which detects the end of a code word whenever the initial state 1 recurs. In this figure (and throughout this paper) a "0" code digit corresponds to left-downward motion, a "1" to right-downward motion. These digits are the inputs to the sequential machine. In Fig. 1, state 4 is equivalent to state 2 (in the usual sequential machine sense[9]), since the respective next states are equivalent, for each input digit. Thus Fig. 1 represents a 3-state machine whose recurrence of state 1 indicates the end of a code word in text.

The synchronizing diagram[3,10] for the code of Fig. 1 is shown in Fig. 2. It is obtained from Fig. 1 by examining the set of next states resulting from each input digit, beginning with the set of all states, i.e., total ambiguity. For example, a "0" digit can lead only to state 1 or 2, and a "1" can lead only to state 1, 2 (formerly 4), or 3. Given the set of states 1, 2, a "1" can lead only to state 1 or 3. In this way it is seen that the sequence 0110 always culminates in the occurrence of state 1, irrespective of the actual state at the beginning. (This is indicated in Fig. 2 by the dark path.) Thus 0110 is a synchronizing sequence for the code. (Note that its occurrence in code text following ambiguity does not imply the conclusion of a particular code word; either 10 or 110 could be involved.) There is an infinite set of synchronizing sequences described by Fig. 2, including as other examples 0111110 and 10110.

If a sequential machine (or a code) has at least one synchronizing sequence, then it tends to resynchronize itself with probability one,[11] assuming all input sequences are possible. In order to obtain a measure of how well text for a given code is self-synchronizing following arbitrary errors, the code digits "0" and "1" are assumed to occur independently and equiprobably. This is in fact a meaningful and useful assumption under various real-life circumstances, even when it is only

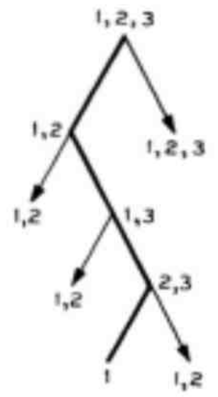


Fig. 2—Synchronizing diagram for the code of Fig. 1, showing unresolved ambiguity.

approximately valid. (This assumption holds exactly whenever each code word occurs independently with its *characteristic probability* 2^{-d}, where d is the length of the code word in digits.) Assuming this randomness of 0 and 1 in code text, the synchronizing diagram of Fig. 2 is redrawn in Fig. 3 to show that on the average $I = 16$ digits of code text are required for code text to resynchronize itself following arbitrary errors. (This computation may be done in several ways[1] which are not relevant to the present discussion. Note that the randomness assumption implies that at each node in the synchronizing diagram the residual lag is one more than the average between the residual lags of the two nodes below.) In general, the *synchronization lag* (or, simply, the lag) I of a (prefix) code is defined as the average number of code digits until synchronization can be guaranteed to the end of some (not necessarily known) code word following total ambiguity, assuming

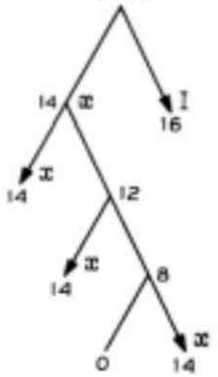


Fig. 3—Copy of Fig. 2, showing lag at each node.

randomness of "0" and "1" as above. Thus the lag is the average length of the synchronizing sequences. (It is also convenient to speak of the *actual lag*, the same average but based on the probabilities of the actual source distribution rather than on random text. If the randomness assumption is not roughly applicable in a given case, then it is necessary to investigate the actual lag. If the likelihood of the various synchronizing sequences actually occurring in text is smaller than under the randomness assumption [e.g., because of constraints on the source symbols], the actual lag is larger than I. However, in such cases it is often possible to change the encoding slightly to assure that the actual lag is less than I, without adversely affecting compression. It may also be possible to use a different code with the same set of code-word lengths whose lag is less. For example, the code 00, 01, 11, 100, 101 has $I = 5$, a marked improvement over the code of Fig. 1, with $I = 16$; for purposes of compression, these two codes are equivalent. In general, the randomness assumption and the resulting lag are quite useful.)

[Synchronizing sequences also exist for uniquely decodable non-prefix codes. For example, consider the code 00, 01, 11, 001, 011.[3] The sequence 10 in code text guarantees that a code-word end occurred between the "1" and the "0"; the lag is 4. Since such codes may require decoding delays (in some cases infinite) beyond the end of their code words, they are of little practical significance. (This paragraph and others delimited by square brackets may be omitted on casual reading.)]

Codes vary widely in their ability to resynchronize code text. For each number n of code words, there is a code with $I = 2$ (the "best"), and a related code with $I = 2^{(n-1)} - 2$ (the "bad" code). These codes are shown in Fig. 4 for each $n \leqq 9$, along with a few other examples. (Note that the "best" code and the "bad" code for each n have the same set of code-word lengths, and are thus equivalent with respect to compression considerations.) The only finite exhaustive codes with $I = \infty$ known to Schützenberger[8] (and to the author) are the block codes (e.g., the fixed-length codes (a), (b) in Fig. 4) and three classes of non-block codes: the codes with greatest common divisor of their code-word lengths greater than one (e.g., code (c) in Fig. 4), the uniformly composed codes C^f obtained by concatenating f times the code words of some code in all combinations (e.g., code (d) in Fig. 4, composed from 0, 10, 11 with $f = 2$), and Schützenberger's "anagrammatic" codes[8] which when scanned backwards are also prefix codes (e.g., code (e) in Fig. 4). Note that block codes have the properties of all of these three classes. (By sacrificing exhaustivity [and optimality], i.e., by

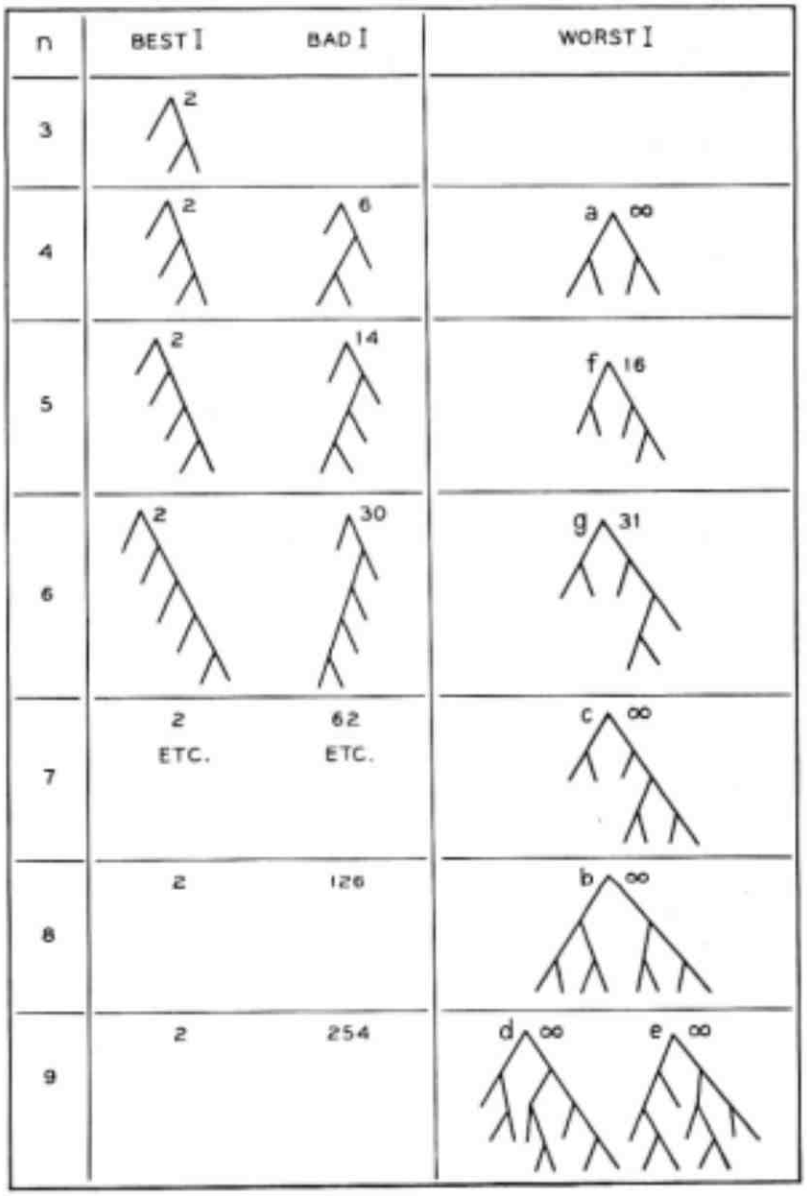


Fig. 4—Prefix codes with external lags for $n \leqq 9$.

eliminating at least one code word from a code, the lag is never increased; in some cases I is reduced substantially. [A slight extension of the definition of the lag is necessary for non-exhaustive prefix codes to handle sequences which cannot arise.] Thus there are synchronization advantages of nonexhaustive codes.)

The only finite exhaustive codes known to the author which have $2^{(n-1)} - 2 < I < \infty$ for any n are the two codes (f) and (g) in Fig. 4. Excluding these two codes and the "bad" codes, all remaining finite lag codes seem to have $I < 2^{(n-1)} - 2$, for all n; the worst of the remaining codes seem to be far from the "bad" code bound ($I = 5.5$ for $n = 5$, $I = 12$ for $n = 6$). Nevertheless, as n increases, the lag of an arbitrary code may be quite bad. Since prefix codes offer compressions which become more spectacular as n increases, this would be distressing were it not for the small-lag infinite systematic codes of Refs. 1 and 2. Since these are helpful in the construction of sequential prefix codes, they are summarized next.

2.2 *Infinite Codes with Good Synchronization Properties*

The "systematic" prefix codes of Ref. 1 are infinite codes generated by sequential machines whose synchronizing properties guarantee that the codes will themselves have good synchronizing properties. (These codes have properties suggested by B. Mandelbrot.[12]) When truncated, these codes give highly efficient (but not optimal because they are no longer exhaustive) encodings with self-synchronizability far exceeding that of the optimal (Huffman[8]) encodings. A typical reduction for a code for English words with $n = 5537$ is from I $\gg$ 1,000,000 for an optimal code down to $I = 10.7$ for a systematic code which is within 3 percent of the optimal code in terms of compression. (The resulting compression is about a factor of three better than a block coding of the English letters, due to the redundancy of the language.) High-efficiency compression using these codes is discussed in Ref. 1.

Several examples are given in Fig. 5. The convention of this and succeeding figures is that a terminal node without an arrow indicates the end of a code word. A terminal node with an arrow indicates the occurrence of a state shown elsewhere in the diagram. If there is no label on the arrow, the corresponding state is that occurring at the top (the root) of the diagram. The first code (a) is a "definite" code,[2] with any occurrence of the sequence "10" in code text indicating the end

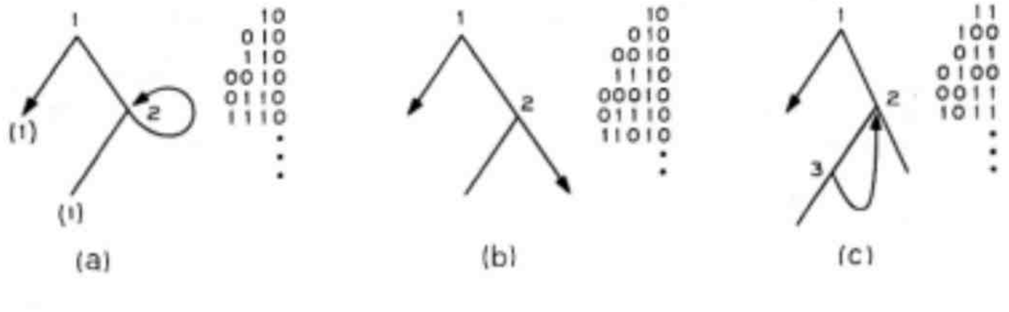


Fig. 5—Examples of systematic prefix codes with small lags I. (a) Definite code, $I = 4$, $L = 4$; $N(d) = 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7$ . . . . (b) Tree-based code, $I = 8$, $L = 6$; $N(d) = 0\ 1\ 1\ 2\ 3\ 5\ 8\ 13$ . . . . (c) General systematic, $I = 6$, $L = 4$; $N(d) = 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7$ . . . .

of a code word. A *definite* code is one in which a code-word end occurs in code text if and only if one of a finite set of finite sequences occurs; in the example, the sequence "10" forms the set. The other codes are not definite, but nevertheless have small lags. The number $N(d)$ of code words of each length d is given for each code. The function $N(d)$ is usually called the *structure function* of the code. The average code-word length (assuming randomness of "0" and "1" in code text) is L, the sum of $d\ N(d)\ 2^{-d}$ over all d. (If each code word occurs independently with its characteristic probability, then L is equal to the entropy H of the source distribution.) Note that $I \geqq L$ for all prefix codes, with equality if and only if the code is definite. That is, on the average, one code word is sufficient to resynchronize a definite code, longer being required for a non-definite code.

It appears that asymptotically almost all [completely specified, strongly connected, deterministic] sequential machines have synchronizing sequences; for those that do, the resulting systematic prefix codes are self-synchronizing (i.e., $I < \infty$). The spectrum of values of I, however, is wide. As is the case with exhaustive finite prefix codes, the infinite codes with $I = \infty$ may be of several classes. There exist codes with a non-unit greatest common divisor of their code-word lengths (e.g., Fig. 6a), with uniform composition (e.g., Fig. 6b), and with the anagrammatic property (e.g., Fig. 6c). Unlike the case for finite exhaustive codes, infinite exhaustive codes can easily be constructed which belong to none of these three classes (e.g., Fig. 6d). (Uniformly composed codes are studied in Ref. 13.)

[The worst value of I for a code derived from a synchronizable r-state machine appears to be about $I \leqq (r - \frac{1}{2})\ (2^r - 2) + 1$; that is, just about a factor of r worse than the "bad" code of Fig. 4 with r states, $r = n - 1$. A. E. Laemmel and B. Rudner[14] exhibit for each r a machine whose shortest synchronizing sequence is $(r - 1)^2$. For all r-state machines with synchronizing sequences, the best bound known to the author for the longest synchronizing sequence is that given by M. A. Fischler and M. Tannenbaum[15]: $(r - 1)^2$, exact for $r \leqq 4$; $(r^3 - r)/6$ for small $r \geqq 5$; $11r^3/48$ for large r. On the other hand, there are many machines with extremely short synchronizing sequences and small values of I.]

III. SYNCHRONIZATION IN SEQUENTIAL PREFIX CODES

For purposes of this paper, sequential coding implies the use of a sequential machine for the encoder, and a corresponding sequential

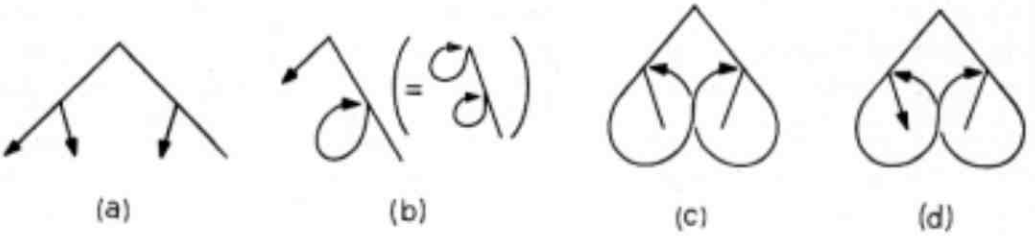


Fig. 6—Examples of systematic codes with infinite lag ($I = \infty$). (a) Greatest common divisor 2; $N(d) = 0\ 1\ 0\ 3\ 0\ 9\ 0\ 27\ 0\ 81\ \ldots$. (b) Uniform composition; $f = 2$, $N(d) = 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ \ldots$. (c) Anagrammatic; $N(d) = 0\ 2\ 2\ 2\ 2\ 2\ 2\ \ldots$. (d) None of the three classes; $N(d) = 0\ 1\ 1\ 2\ 3\ 5\ 8\ 13\ 21\ \ldots$.

machine (the inverse or "quasi-inverse") for the decoder.[16] In an earlier paper,[3] the author explored the use of particular input sequences to the encoder (synchronizing input sequences) each of which synchronizes the encoder to a particular known state. Also involved are particular output sequences from the encoder (synchronizing output sequences) whose presence in code text (in the absence of errors) guarantees the occurrence of a particular state of the encoder at a particular point in the code text. Synchronizing output sequences correspond to the synchronizing sequences for prefix codes.

Given an encoder with both synchronizing input sequences and synchronizing output sequences, the entire system is self-synchronizing on the average. That is, following some arbitrary errors, two things happen. First the encoder resynchronizes itself and begins to encode correct text again. Then the decoder tends to resynchronize itself with the encoder (the synchronizing output sequences for the encoder act as synchronizing input sequences for the decoder), and correct decoding resumes. This occurs spontaneously as an intrinsic property of the coding system, with no externally imposed resynchronization required.

A (first-order) *sequential encoding* is a mapping of symbols $S(i)$ onto code words $w(i \mid j)$ where the code word selected depends on the previously encoded symbol $S(j)$ as well as on $S(i)$. If the set of code words $\{w(i \mid j)\}$ for each j is a prefix code, then the set of code words $\{w(i \mid j)\}$ for all i, j is a *sequential prefix code*. For such codes a *synchronizing sequence* is a sequence of code digits the end of which must correspond to the end of a code word (possibly unknown) resulting from a known symbol $S(i)$, irrespective of what preceded that sequence. Thereafter subsequent decoding is correct, irrespective of the initial ambiguity. The remainder of this paper is concerned with sequential prefix codes, and investigates their compression, decoding and self-synchronizing properties. (The development is also applicable to higher-order sequential encodings, with the code word depending on the present message $S(i)$ and some finite number of previous messages.)

As an example, consider the sequential encoding given by Table I. A, B, C and D represent four source symbols $S(i)$, $i = 0, 1, 2, 3$, where i is the *level* of the symbol. This is an example of an encoding in which the code word $w(i \mid j)$ to be transmitted is a function of the cyclic difference between the level of the symbol $S(i)$ to be encoded (column headings) and the level of the symbol $S(j)$ just previously encoded (row headings): $w(i \mid j) = W(k)$, where $k \equiv i - j \pmod{4}$. This encoding is thus a difference encoding. Note that, irrespective of the choice of the code $\{W(k)\}$, there is always ambiguity in decoding as soon as an error is made. If for example $S(2)$ is decoded instead of $S(1)$ as a result of a transmission error, subsequent decoding will consistently produce $S(i + 1)$ instead of $S(i)$ where $i + 1$ is modulo 4, as long as further errors do not compensate for the original errors. (Throughout the paper, all additive operations involving i and j are modulo n.)

As a second example, consider the sequential prefix code of Table II. In this example four different prefix encodings $w(i \mid j)$ are used for $j = 0, 1, 2, 3$, depending upon the symbol $S(j)$ previously encoded. For example, if "A" was just encoded, then A, B, C, D are encoded as 0, 11, 100, 101, respectively. Thus any symbol $S(i)$ as input to the encoder acts as a synchronizing input sequence. (The encoder is a 1-definite machine,[1] with its output being a function of the present input symbol and the previous input symbol.)

A state diagram for the decoder is given in Fig. 7. The states A, B, C, D represent the successful decoding of these four symbols, while the states a, b, c, d, a′, b′, c′, d′ are intermediate states. The state diagram is shown in four pieces, which fit together as the upper-case letters indicate. The synchronization diagram for this state diagram is shown in Fig. 8. Its construction follows the usual technique[3,10] and is similar to the synchronization diagram for prefix codes (cf. Fig. 3). The top

TABLE I—FOUR-LEVEL DIFFERENCE CODE

	$i = 0$	1	2	3
$S(j)$	$S(i) = $ A	B	C	D
$j = 0$ A	W(0)	W(1)	W(2)	W(3)
$j = 1$ B	W(3)	W(0)	W(1)	W(2)
$j = 2$ C	W(2)	W(3)	W(0)	W(1)
$j = 3$ D	W(1)	W(2)	W(3)	W(0)

TABLE II—A GOOD SEQUENTIAL PREFIX CODE

	$i = 0$	1	2	3
$S(j)$	$S(i) =$ A	B	C	D
$j = 0$ A	0	11	100	101
$j = 1$ B	010	1	00	011
$j = 2$ C	100	11	0	101
$j = 3$ D	010	011	00	1

node corresponds to the set of all states. Each terminal node corresponds to the occurrence of the end of a code word resulting from a particular symbol $S(i)$. Given a "0" input, the next state must be one of A, C, b, d, a′, c′, for example, beginning with the set of all states. The synchronizing diagram of Fig. 8 results after noting several equivalences (e.g., Abda′c′ with ACbda′c′).

From the synchronizing diagram of Fig. 8 it is seen that the sequence 0011 can arise in code text only if the corresponding source sequence ends in a "B". Thus 0011 synchronizes the decoder to the end of a code word corresponding to the symbol "B" irrespective of what preceded it; similarly 00101 synchronizes to "D", 1100 to "C", and 11010 to "A". Assuming 0 and 1 are random in the above sense, it is easily shown that synchronization results from total ambiguity after an average of $J = 7.67$ digits. The *sequential synchronization lag J* is the average number of code digits until the end of a code word is achieved corresponding to a known symbol; that is, J is the average length of the synchronizing sequences. In this example, the occurrences in code text of the encoded versions of CB, CD, BC and BA imply synchronizing sequences for the decoder. (For example, note that CB is encoded as 0011 following a "B" or "D", as 10011 following an "A",

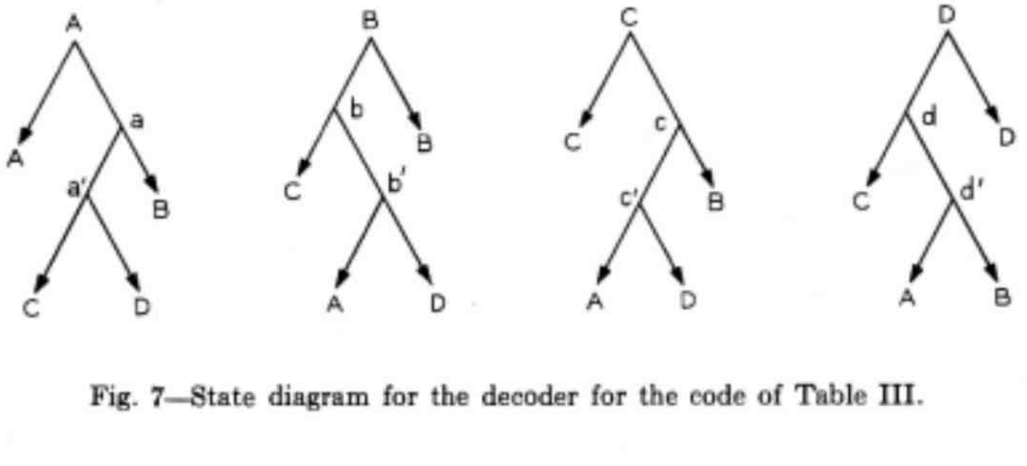


Fig. 7—State diagram for the decoder for the code of Table III.

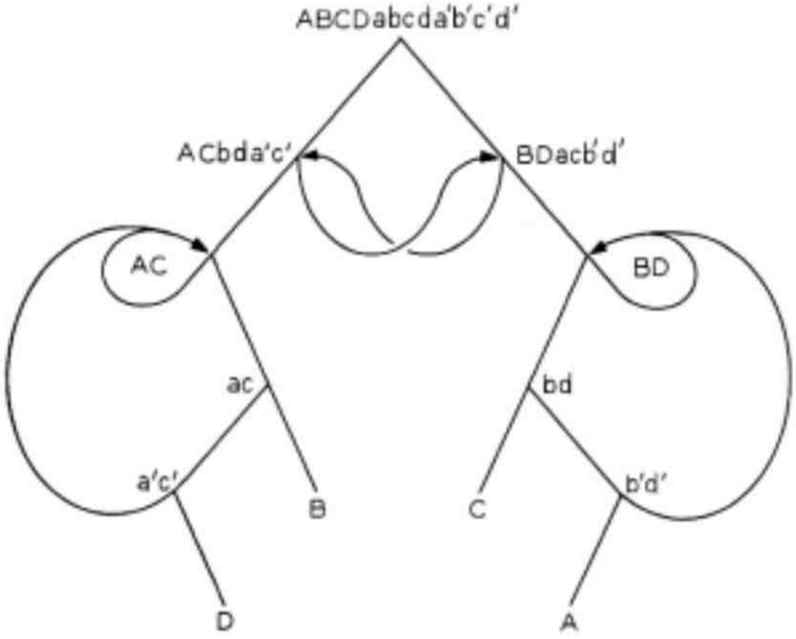


Fig. 8—Synchronizing diagram for the decoder of Fig. 7, $J = 7.67$, $J' = 3$, $J'' = 4.67$.

and as 011 following a "C"; in the last case, 011 must have been preceded by a "0".)

In the example, any sequence ending in 00 or 11 (cf. Fig. 8) guarantees the end of a code word, although not the code word for a known symbol. For purposes of this paper, synchronization to the end of some (unspecified) code word is called *first-stage* synchronization. Synchronization to the end of a code word corresponding to a particular symbol is called *second-stage* synchronization. The former is of concern in prefix codes, and both are of concern in sequential codes. In rare cases (particularly asymmetric ones), the second stage is achieved simultaneously with the first stage. In many useful cases, however, they may be treated independently. (In all but one example given in this paper, second-stage synchronization implies a particular known code word as well as a particular known symbol.)

[A word of caution is again needed regarding the randomness assumption. Again, the actual lag may be defined in terms of the actual probabilities. If synchronizing sequences occur naturally in code text, then the lag J is a fair estimate of the actual lag. If the sequences occur only as a result of unlikely sequences of symbols, then the lag J is smaller than the actual lag. However, in such cases the encoding can often be altered so that J is realistic. In general, it is desirable to have codes with widely distributed likely synchronizing sequences, as in Fig. 8, rather than being totally dependent on a few obscure sequences.]

IV. CODING COMPLEXITY

It is clear that the choice of the code of Table II (with $J = 7.67$) is (infinitely) superior to the code of Table I (with $J = \infty$) in terms of resynchronizability. If $W(k) = 0, 11, 100, 101$ in Table I for $k = 0, 1, 2, 3$, then the codes of Tables I and II are identical with respect to compression capabilities. The significant concern remaining is the computational complexity of the encoder and the decoder for the code of Table II. Intuitively, one might expect that an n-state sequential prefix code would be almost n times as complex as a difference code in terms of its encoding and decoding circuitry. Somewhat surprisingly, codes are developed below for which the complexity is essentially the same as the difference codes, while attaining good synchronizability and compression.

Examination of the code $w(i \mid j)$ of Table II shows that the prefix code $w(\mathrm{i} \mid 1)$ for state B $(j = 1)$ is the binary complement of the code $w(i \mid 0)$ for state A $(j = 0)$, cyclically shifted by one word: $w(i \mid 1) = w'(i - 1 \mid 0)$. Further, the code $w(i \mid 2)$ for state C $(j = 2)$ is related to $w(i \mid 0)$, having the same code words but with a different mapping. In particular, $w(i \mid 2) = w(2 - i \mid 0)$. Finally, the code for state D $(j = 3)$ is the shift of the complement of $w(i \mid 2)$, or the complement of the shift, and thus $w(i \mid 3) = w'(3 - i \mid 0)$. Thus all four of the prefix codes are closely related to any one of them.

The code $w(i \mid j)$ as a function of $w(i \mid 0)$ is summarized in Table III for each j. The structure of the code is somewhat more transparent when related to the difference code of Table I, with $W(k) = 0, 11, 100, 101$ for $k = 0, 1, 2, 3$ and $k \equiv i - j \pmod 4$. In this case, $w(i \mid j) = W(k), W'(k), W(-k), W'(-k)$ for $j = 0, 1, 2, 3$, respectively. The encoder and decoder for Table II are thus easily specified in terms of the difference code. If (p, q) is the binary representation of j as shown in Table III, then $W(k)$ is replaced by $W(-k)$ if $p = 1$, and the result is complemented if $q = 1$. The encoder for the difference code is shown in Fig. 9, while the encoder for the sequential prefix code of Table II is given in Fig. 10. ("$\triangle$" represents a one-digit delay.) It is seen that an AND gate $(\cdot)$ and two EXCLUSIVE OR gates $(\oplus)$ represent the marginal cost of encoding the latter code, compared to the difference code. The same is true of the decoder, which employs precisely the same set of gates.

TABLE III—SUMMARY OF ENCODER AND DECODER FOR THE CODE OF TABLE II

j	p	q	Code $w(i\|j)$
0	0	0	$w(i\|0) = w(i\|0) = W(k)$
1	0	1	$w(i\|1) = w'(i-1\|0) = W'(k)$
2	1	0	$w(i\|2) = w(2-i\|0) = W(-k)$
3	1	1	$w(i\|3) = w'(3-i\|0) = W'(-k)$

V. CONSTRUCTION OF GOOD SEQUENTIAL PREFIX CODES (WITH SMALL LAGS)

The synchronizing properties of sequential prefix codes fall into two classes, those which depend on the individual choice of the encoding $w(i \mid j)$ for each i, j, and those (in varying degrees) which are independent of the actual choice of $w(i \mid j)$. [The encoding of Table I, for example, has $J = \infty$ irrespective of the $w(i \mid j)$.] Such properties are said to be *choice-dependent* and *choice-independent* for these two classes. An example of how choice-independent properties may be treated separately is given by the following theorem.

5.1 *Very Bad Codes* ($J = \infty$)

Theorem 1: A sequential prefix code has $J = \infty$ *in each of the following cases: (a) if there are precisely* n *distinct code words among the* $n \times n$ $\{w(i \mid j)\}$, *each of which occurs exactly once for each* i *(a "Latin square" code); (b) if for some value of* s $(0 < s < n)$ *the relation* $w(i + s \mid j + s) = w(i \mid j)$ *holds for all* i *and* j.

Proof: Case (a). Consider an ambiguity between any two symbols which could have led to a given code word. Then any subsequent code word could have resulted from either of two distinct symbols. Thus ambiguity is never reduced, and $J = \infty$ irrespective of the choice of the n independent $w(i \mid j)$. [Note that if the prefix code for any j (the same code words, but a different encoding for each j) is self-synchronizing $(I < \infty)$, then it is possible to reduce ambiguity to the end of some (unknown) code word, but no further. Recall that in the definitions, a distinction is made between the code (the set of code words) and

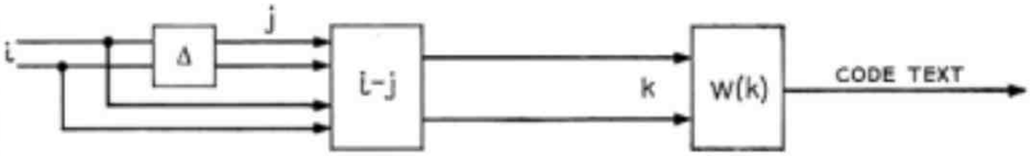


Fig. 9—Encoder for the difference code of Table I.

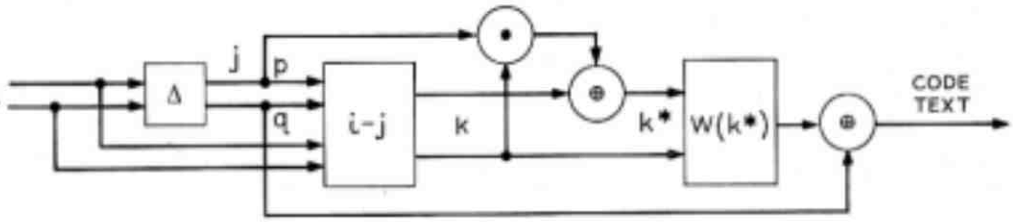


Fig. 10—Encoder for the sequential prefix code of Table II.

the encoding (the mapping between code words and source messages).]

Case (b). Without loss of generality, suppose that s is the smallest value of s for which $w(i+s \mid j+s) = w(i \mid j)$ for all i and j. (If s is 1, the code is a difference code.) Then $w(i+gs \mid j+gs) = w(i \mid j)$ for all i, j, for every integer g. Let u be the smallest value of $g > 0$ for which $gs \equiv 0 \pmod{n}$. Then since s is as small as possible, it follows that $su = n$, i.e., s divides n: Thus there are s prefix codes $w(i \mid j)$, e.g., for $j = 0, 1, \cdots, s-1$, which completely determine the prefix codes for $j = s, s+1, \cdots, n-1$. That is, the same prefix code (shifted) occurs for each $j = gs + h \pmod{n}$ for $g = 0, 1, \cdots, (n-s)/s$, for any fixed value of h, $0 \leqq h \leqq s-1$. Thus there must always remain an ambiguity among all the symbols with $i = gs + h \pmod{n}$ for $g = 0, 1, \cdots, (n-s)/s$, for any particular value of h, $0 \leqq h \leqq s-1$. Therefore J is infinite again, irrespective of the choice of the sn independent $w(i \mid j)$. QED.

[Case (b) of Theorem 1 can easily be extended to include cases for which $w(i+s \mid j+t) = w(i \mid j)$ for all i, j: if s and t are each relatively prime to n, or more generally if s and t have the same order in the field of integers {i.e., if $u = v$, where u and v are the smallest non-zero values for which $us \equiv 0 \pmod{n}$ and $vt \equiv 0 \pmod{n}$}. As these cases are less natural to the present environment, they are mentioned parenthetically.]

The difference codes satisfy both cases (a) and (b). Another example of case (b) is given in Table IV. If e, f, g, h are replaced by b, c, d, a, respectively; this example also satisfies case (a), and is a "sum" code rather than a difference code.

5.2 *Properties of Good Codes*

It is highly desirable that sequential codes $\{w(i \mid j)\}$ have considerable structure, in order to simplify encoding and decoding, to simplify the analysis of synchronizing properties, and to facilitate the construction of good large codes. Several intuitively evolved properties have been examined which are found to contribute considerably to this desired structure. *Firstly*, to simplify encoding and decoding, the number m of distinct prefix codes (not counting complements) used to form a sequential code should be small (one or at most two). If a prefix code and its complement are both used, the sequential code is *complemented*. In the complemented code of Tables II and III, m is one since the prefix code for each j is $W(k)$ or its complement. Since the circuitry required is up to m times as complex as when $m = 1$, large values of m are to be diligently avoided. The m distinct prefix codes (ignoring complements) are called *kernels*. The property of keeping m small is called the *symmetrization* property; it is choice-independent. *Secondly*, if all code words $w(i \mid j)$ for a given i end in the same digit, irrespective of j, then first-stage synchronization is greatly enhanced. This property of making code-word endings uniform by column (as in Table II) is called the *columnization* property. (It is more or less choice-independent.) A third property involves the number of different symbols to which a given code word can correspond. (In the example of Table II, half of the code words occur only for one value of i each.) Second-stage synchronization is greatly aided by having almost all code words occur only for a relatively small number of different symbols, avoiding having each occurrence correspond to a different symbol $S(i)$. This is called the *association* property, and is also choice-independent.

It should be noted that none of the above properties is necessary for obtaining finite lag codes; however, these properties are found to be helpful in achieving low lags. Although the example of Table V is a (complemented) one-kernel code, it violates both the columnization property and the association property. (Note that each code word occurs only twice, but always for different symbols.) Its lag J, though finite, is quite horrendous (around 200), with the shortest synchronizing sequences being of length ten (e.g., 0101000100). It thus compares badly with the code of Table II with respect to its lag. (It even compares badly with the worst columnized one-kernel finite-lag code using the given prefix code, for which $J = 23.1$.)

TABLE IV—EXAMPLE OF A CODE WITH $J = \infty$ BY THEOREM *1*(b)
$n = 4, s = 2$

a	b	c	d
e	f	g	h
c	d	a	b
g	h	e	f

TABLE V—A BAD FINITE LAG CODE

0	11	100	101
1	00	011	010
101	100	11	0
010	011	00	1

An example of a two-kernel code which satisfies the columnization property and the association property is shown in Table VI. The code words which occur for only one value of i are indicated by asterisks. This code has $J = 8.9$, with synchronizing sequences including 0011 (for B) and 11100 (for C).

In a columnized code, the set of symbols $S(i)$ for which all code words end in "0" ("1") is called the *0-set (1-set)*. In the example of Table VI, the sequence 00 (among others) guarantees the end of a code word corresponding to the 0-set A, C or E (i even), while a 111 guarantees the end of a code word corresponding to the 1-set B or D (i odd). These are two of the first-stage synchronizing sequences. Having reduced the ambiguity to a 0-set symbol or to a 1-set symbol, the association property within these sets is of great aid to second-stage synchronization. For example, the code words which optimally satisfy the association property (e.g., those with asterisks in Table VI) themselves act as second-stage synchronizing sequences in this example. Association is especially helpful to second-stage synchronization if the prefix code is the same for each symbol $S(j)$ in the 0-set, and similarly for the 1-set. This is the *bifurcation* property of columnized codes, that the 0-set and the 1-set use one prefix code each (although the two codes may be identical). Note that this property is found in the code of Table VI, in which two distinct prefix codes are used. If present, the bifurcation property implies that the symmetrization property is met with $m = 1$ or 2. (By definition, bifurcated codes must be columnized.)

An example of a one-kernel code satisfying all four of the above properties is given in Table VII. Apart from 00 and 01, no code word

TABLE VI—A TWO-KERNEL EXAMPLE, $J = 8.9$, $I' = 4$, $J'' = 5.1$

$S(j)$	$S(i)$ = A	B	C	D	E
A	0	11*	100	1011*	1010
B	010*	1	00*	0111	0110*
C	1010	11*	0	1011*	100
D	010*	0111	00*	1	0110*
E	1010	11*	100	1011*	0

TABLE VII—A ONE-KERNEL EXAMPLE, $J = 20.2$, $I = 8$, $J'' = 14.7$

$S(j)$	$S(i) = $A	B	C	D	E	F	G	H
A, B	00	01	100	101	1100	1101	1110	1111
C, D	1100	1101	00	01	100	101	1110	1111
E, F	1100	1101	100	101	00	01	1110	1111
G, H	1100	1101	1110	1111	100	101	00	01

occurs for more than two symbols $S(i)$ (i.e., in more than two columns). The lag is seen to be $J = 20.2$.

5.3 *Balanced Codes*

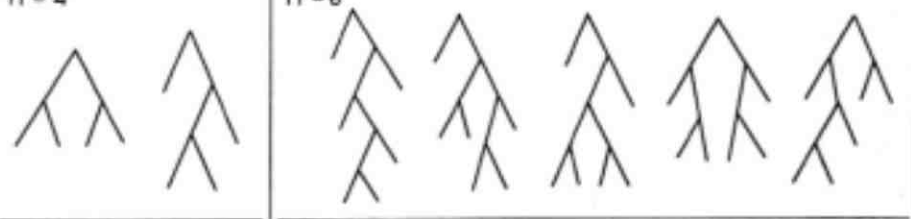
If the columnization property is to be achieved in a complemented sequential code, then any kernel prefix code and its complement must each have the same number of code words ending in zero (or one). Consequently, the prefix code (and its complement) must have half of its code words ending in each digit. Such a prefix code is called a *balanced* code, and of course must have an even number of code words. The codes of Tables II and VII are examples of balanced prefix codes used in complemented and uncomplemented one-kernel codes, respectively.

Theorem 2: For every set of code-word lengths with n even for which there exists an exhaustive prefix code, there exists at least one balanced prefix code.

Proof: A simple proof involves a construction procedure during which the number of code words ending in "0" differs by at most one from the number of code words ending in "1". Thus since n is even, the resulting code is balanced. Such a procedure is easy to construct, but is omitted here since it does not contribute to a basic understanding of the paper. [It can in fact be shown that the ratio of balanced exhaustive codes to all exhaustive codes is asymptotic for large n to $1/(n\pi)^{\frac{1}{2}}$.]

A few balanced exhaustive codes are shown in Fig. 11, including one for each structure function $N(d)$ with $n = 4$ and 6. In each of these cases, the code shown has the smallest possible lag I among all balanced codes with the given $N(d)$.

If the sequential code is not complemented, the kernels need not be balanced. However, the use of balanced prefix codes as kernels is highly beneficial. It greatly enhances flexibility in the assignment of the $w(i \mid j)$ according to the needs of synchronization (e.g., via columnization and association), coding complexity and compression.



CODE	n = 4		n = 6				
N(d)	04	112	11112	1032	1104	024	0312
I	∞	6	5.2	5.3	7	10	12
I'	∞	3	3	6.1	5.5	10	11
$I^{`}$	∞	5	31	6.3	5.9	10	13
I^{*}	∞	2	2	5.1	4.5	9	10

Fig. 11—Some balanced codes for $n \leqq 6$.

5.4 *Analysis of First-Stage Synchronization*

Another property emerges from considering the relation between the lag J of a bifurcated sequential code and the lags of the kernel codes. For an uncomplemented one-kernel code, first-stage synchronization is guaranteed by the synchronizing sequences of the kernel, i.e., with the lag I of the kernel. Thus this lag should clearly be small. On the other hand, for a bifurcated complemented one-kernel code, the lage I is not relevant to first-stage synchronization. Instead the mutual lag I' of a prefix code and its complement is needed. The *mutual lag* of a 0-set code and a 1-set code is obtained by considering the state diagrams of both codes. Imposing the restriction that a code word ending in "0" ("1") is followed by a 0-set (1-set) code word, these two state diagrams become one just as the four diagrams in Fig. 7 become one. The mutual lag I' is then the lag of this combined state diagram, obtained from the *mutual synchronizing diagram*, i.e., the synchronizing diagram for the combined state diagram. Terminal nodes consist of first-stage synchronization, i.e., solely of 0-set or 1-set codeword ends. (Note that the mutual lag is partially choice-independent, depending on the choice of the 0-set and 1-set prefix codes, but not on the actual $w(i \mid j)$.) Mutual lags are relevant primarily for balanced prefix codes and their complements, as used in (one-kernel) bifurcated complemented codes; this case is assumed unless otherwise specified. They are also meaningful for two-kernel bifurcated codes, for which the mutual synchronization diagram also guarantees first-state synchronization. (For example, the two prefix codes used in Table VI have $I' = 4$.) Note that the mutual lag of a code with itself is $I' = I$ (since the 0-set code and the 1-set code are identical).

Consider as an example the prefix code {0, 11, 100, 101} used in

Table II. Suppose that it is used for the 0-set symbols A and C (Fig. 12a), while its complement is used for the 1-set symbols B and D (Fig. 12b). The state 0 (1) corresponds to a code word ending in "0" ("1"), and implies that the next code word is taken from the 0-set (1-set) code. (Note that irrespective of the actual $w(i \mid j)$, the occurrence of a 00 or a 11 suffices to guarantee the end of a code word, as in Fig. 8.) The mutual synchronization diagram is given in Fig. 12c, and has $I' = 3$.

Values of the mutual lag I' are given in Fig. 11 for each code shown there and its complement. In each case, the code shown has the smallest value of I' for any balanced code with the given set of code-word lengths $N(d)$. A *numerological curiosity* is provided by Fig. 11: for each $N(d)$, the indicated best value of I' is precisely one greater than the best value I^* of I attainable by any code (unbalanced, in fact) with the given $N(d)$. Since this curiosity is true of all $N(d)$ for exhaustive codes with $n \leqq 6$, including those for odd n, it seems highly likely for all n, for all $N(d)$ for which exhaustive codes exist.

[It should be noted that the value of I' for a given prefix code and its complement is obtained with the given code as the 0-set code. If it is used instead as the 1-set code, the resulting value of I' is the mutual lag of the complemented prefix code, and is designated by $I^{\backprime}$. Values of $I^{\backprime}$ are also shown in Fig. 11. It is seen that I' and $I^{\backprime}$ are not usually the same.]

The *kernel lag* J' of a bifurcated code is then defined as the synchronization lag of the first stage. For a one-kernel uncomplemented code, $J' = I$; for a two-kernel (uncomplemented) code, or a one-kernel complemented code, $J' = I'$. The *kernel lag property* then states that the kernel lag J' of a bifurcated code should be small, in order to help minimize the lag J.

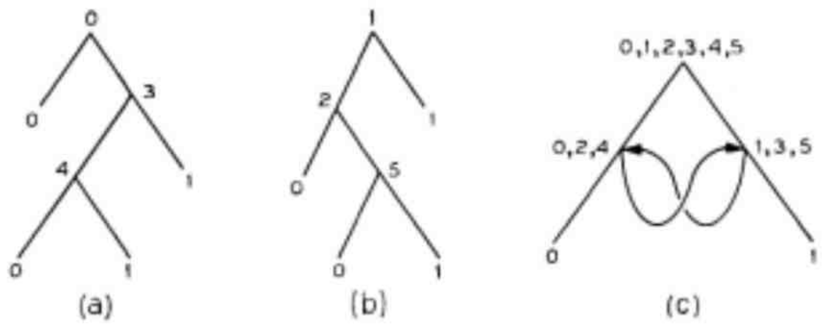


Fig. 12—Example of first-stage synchronization in a complemented code with $I' = 3$.

5.5 *Further Properties of Bifurcated Codes*

The second-stage lag may be approximated by the *structural lag* J'' which assumes first-stage synchronization, and results in a known symbol (at the end of a code word). It is obtained by averaging the second-stage lags, K beginning with the 0-set and K' beginning with the 1-set; K and K' are weighted according to the probabilities of reaching the 0-set and 1-set, respectively, in the mutual synchronization diagram. (In the example of Table VI, for example, it is seen that these weights are $\frac{2}{3}$ and $\frac{1}{3}$.) The *structural lag property* states that this lag should be as small as possible. A rough measure of J'' may be made independent of the prefix code by assuming all code words of equal length, and the 0-set and 1-set equiprobable.

Theorem 3: For a bifurcated code,

$$J \leqq J' + J''.$$

Proof: The theorem follows from the definitions. Inequality occurs when, in the sequential synchronization diagram, first-stage synchronization occurs in at least one case as the result of an advantageous subset of either the 0-set or the 1-set. By "advantageous" is meant a subset which accelerates second-stage synchronization. The value of J' gives precisely the first-stage synchronization lag in any event. When there are no such subsets, equality holds, as in Fig. 8. (For example, the code of Table VI has $J' = 4$, $J'' = 5.1$ and $J = 8.9$; the sequence 1100 guarantees not just the 0-set, but specifically a "C" or an "E" at its end. The code of Table VII also has such subsets. These subsets are obtainable from the first-stage synchronization diagram, and may be used to calculate the exact second-stage lag $J^{\sim}$ based on the subsets with their appropriate weights, rather than just on the 0-set and the 1-set. Then $J = J' + J^{\sim}$. In all cases $J'' \geqq J^{\sim}$. Also, if J'' is infinite, then so is $J^{\sim}$.)

Theorem 4: A bifurcated code has $J = \infty$ if and only if either J' or J'' is infinite.

Proof: The "only if" follows as a corollary of Theorem 3. The "if" requires two cases. If $J'' = \infty$, then $J = \infty$ since the 0-set and 1-set are subsets of the set of all states, on which J is based. If $J' = \infty$, first-stage synchronization is never reached. Since a bifurcated code must go through this stage in order to reach second-stage synchronization, and since J' is the true value of first-stage lag even when there is inequality in Theorem 3, it follows that $J = \infty$. QED. (Note that in Theorem 4 J'' may be replaced by $J^{\sim}$.)

In many symmetric cases, the 0-set and the 1-set are equivalent structurally, and $K = K' = J''$. This occurs in the example of Table II, since the code obtained under the transformation $A \Leftrightarrow D$, $B \Leftrightarrow C$ is the complement of the Table II code (a complementary reflective symmetry). It also occurs in Table VII, since the 0-set and 1-set symbols are paired with identical encodings. Such symmetric cases are advantageous for encoding/decoding and analytic simplicity.

In certain cases it is desirable to use complemented bifurcated codes. If the above numerological curiosity is true in general, there is essentially no sacrifice in the best I' of a balanced code compared to the best $I = I^*$ for the given $N(d)$. That is, first-stage synchronization is just as rapid for these complemented codes, while avoiding the difficulties arising from the use of an unbalanced code. Further, considering only balanced codes for a given $N(d)$, the best value of I' is often better than the best value of I (cf. Fig. 11). Thus the overall lag is frequently better. Furthermore, the flexibility of compression available with the complemented codes is often greater. For example, consider again the code of Table II, this time only in terms of its code-word lengths. Using an exhaustive prefix code, it is impossible to have length one on the $i = j$ diagonal with an uncomplemented one-kernel code unless the columnization property is violated; this in turn may greatly increase the lag of the code.

[Surprisingly, prefix codes with $I = \infty$ are not altogether useless. If a one-kernel code uses a block code or a code the greatest common divisor (gcd) of whose lengths is greater than one, then it follows that $J' = J = \infty$. It is interesting to note, however, that many uniformly composed codes and anagrammatic codes (with $I = \infty$) have $I' < \infty$, and thus may give rise to one-kernel codes with $J < \infty$, assuming $J' < \infty$. The smallest possible finite exhaustive anagrammatic code with $I' = I'' = \infty$ has $n = 18$, and is its own complement. An infinite example with the same properties is provided by the self-complementing anagrammatic code of Fig. 6c. Thus the use of such prefix codes in a one-kernel sequential code must result in $J = \infty$, by Theorem 4.]

VI. GOOD STRUCTURE FOR LARGE SEQUENTIAL CODES

The codes of Tables II and VII are rather small examples, albeit good ones, of one-kernel codes. Since great compressions are found primarily in large codes, the next question is whether low lags can be achieved for large codes without sacrificing coding simplicity and compression.

Assume for now that it is possible to approximate the second-order probabilities associated with the code words $w(i \mid j)$ as a function of $k = i - j \pmod{n}$, and hence that from a compression point of view a difference code is meaningful. (This assumption will be generalized below.) Now consider the general n-level one-kernel encoding framework shown in Fig. 13. The difference k is modified as a function of j, and the result k^* is encoded by an ordinary prefix encoding $W(k^*)$. If the code is a complemented code, the result is complemented (e.g., if j is odd). (Otherwise the complementing circuitry is not needed, as in Table VII.) The corresponding decoder is shown in Fig. 14. (Again the complementing circuitry may not be required.) In many cases the demodification circuitry of Fig. 14 is identical to the modification circuitry of Fig. 13, with the input and output interchanged. Note that the codes of Tables II (see Figs. 9 and 10) and VII are examples of codes amendable to this framework. The incremental cost of encoding and decoding compared to difference codes is embodied in the modification logic and the trivial complementing logic; as long as the modification logic can be kept simple, the incremental cost is low. The compression can in general be made at least as good as the difference code.

If a balanced code is used, the framework defined by Fig. 13 makes it easy to satisfy the columnization property, the symmetrization property (with $m = 1$), and the bifurcation property. The choice of a prefix code is dictated by the kernel lag property, and by the conditional probability distribution of the symbols $S(i)$, given $S(j)$. The suitable structure is dictated by those probabilities, mitigated by the structural lag property, and by the complexity of the modification logic desired. Considerable experimentation has shown that the properties discussed here are central to the construction of good codes. Columnization and the choice of prefix codes with small J' greatly facilitate first-stage synchronization. Association greatly influences second-stage synchronization. Bifurcation greatly simplifies coding complexity and improves second stage synchronization. The use of balanced codes is

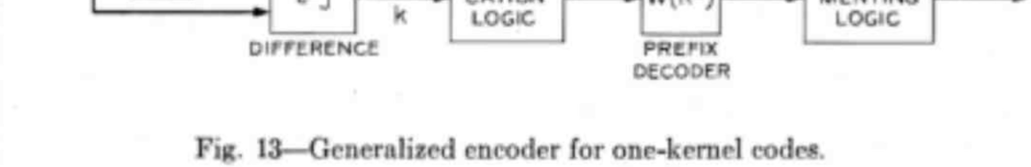


Fig. 13—Generalized encoder for one-kernel codes.

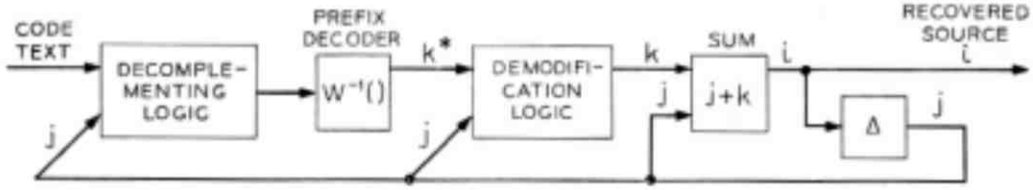


Fig. 14—Generalized decoder for one-kernel codes.

very helpful. If the number n of messages is odd, however, it may be desirable to use a two-kernel code as in Table VI to achieve columnization.

If the conditional probabilities are strongly asymmetric and/or not approximately representable by difference probabilities, it may be advantageous from a compression standpoint to use a multi-kernel code. Alternatively, or additionally, it may be desirable to eliminate the difference/sum circuitry in Figs. 13 and 14, and to deal directly with the i and j. The properties described in this paper are equally relevant in such cases. No restrictive assumptions have been made regarding the first-order probabilities. It is also unnecessary for 0-set and 1-set code words to match even and odd i, respectively; this is merely a descriptive convenience.

6.1 *Large Balanced Codes*

In order to enhance the kernel lag property for codes with large n, it may be desirable to use truncated systematic prefix codes rather than optimal prefix codes as kernels. In this way it is possible for I to remain small as n increases. Several examples of such codes which may easily be truncated to give balanced codes are summarized in Fig. 15, along with their structure functions $N(d)$ and their lags I, I' and Γ. The best lag I^* for any code with the specified $N(d)$ is also given, along with the random average code-word length L for $N(d)$. The selection of efficient codes for compression purposes is considered in Ref. 1.

There exist many classes of codes for which J remains finite (and in fact quite small) as n increases without limit. A simple (rather extreme) example is indicated in Table VIII for $n = 8$. The code is columnized, complemented, bifurcated, and maximally associated without being trivial. (It assumes very high probability of $i = j$ for compression purposes.) To avoid confusion, the symbols are given as A to H for i and j from 0 to 7; the integer k is given to indicate the occurrence of the code word $W(k)$. If j is odd, the complementary code word

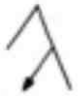
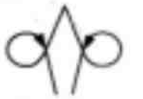
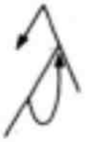
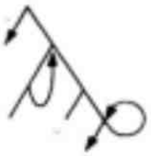
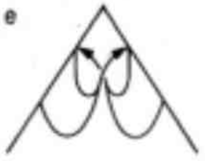
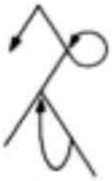
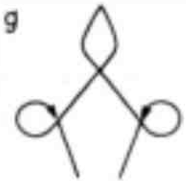
CODE	I	I′	I′	I*	L	N(d) FOR d =									
						1	2	3	4	5	6	7	8	9	⋯
a 	4	3	∞	2	2	1	1	1	1	1	1	1	1	1	⋯
b 	5	5	5	4	3	0	2	2	2	2	2	2	2	2	⋯
c 	6	5	8	4	4	0	1	2	3	4	5	6	7	8	⋯
d 	8	–	–	8	8	0	0	2	2	4	4	8	12	24	⋯
e 	9	9	9	8	7	0	0	2	2	4	6	10	16	26	⋯
f 	8	9	10.5	≤8	6	0	0	1	3	6	10	15	21	28	⋯
g 	8.5	8.5	8.5	≤7.5	4	0	0	4	4	4	4	4	4	4	⋯

Fig. 15—Some systematic prefix codes easily truncated to balanced codes.

$W'(k)$ is used, and the table is read from the bottom up. The code thus has complementary reflective symmetry. If $W(k) = 0, 11, 100, 1011, 10100, 101011, 1010100, 1010101$ for $k = 0, 1, \cdots, 7$, then $J = 8.62$. This code is readily extended to arbitrary even n with a similar pattern of k's: for each even j other than 0, $k = 0$ for $i = j$, $k = j$ for $i = 0$; otherwise $k = i$; code words for odd j are then specified by the complementary reflective symmetry. The modification logic of Figs. 13 and 14 is thus exceedingly simple, and the complementing logic is again a single EXCLUSIVE OR gate as in Fig. 10. The limiting prefix code is code (a) of Fig. 15. Analysis shows that a sequential code with $J' = I' = 3$, $J'' = 4[1 + \frac{1}{3} + \frac{1}{15} + \frac{1}{63} + \frac{1}{255} + \cdots] < 5.7$, $J < 8.7$ exists for every even n. (For suitably skewed distributions, the compression factor arising from this code approaches the base two logarithm of n.) This is an example of a class of codes for which the differential circuitry is not relevant; it is easier to encode directly as $k = i$, with $(n - 1)/2$ pairs of exceptions. Variants for which the 1-set is treated in this fashion while the 0-set is treated differentially as before are also easily implemented. These variants provide a very simple structure which results in low J and simple circuitry.

[A word of caution is in order when considering sequences of exhaustive prefix codes approaching systematic codes in the limit: the limit of a sequence of values of I (or I') thus obtained is normally somewhat larger than the value of I (or I') for the limiting systematic code. (The previous example is an exception with respect to I'.) Code (b) of Fig. 15 is an example in point, with $I = I' = 5$. The limit of the value of I (or I') for the codes 01, 10, 001, 110, 000, 111 ($I = I' = 10$), 01, 10, 001, 110, 0001, 1110, 0000, 1111, ($I = I' = 8$), etc., is seven, although this sequence approaches the systematic code (with $I = I' = 5$) in the limit. In general, truncated systematic codes have better synchronization properties than their finite exhaustive approximations, since the values of I and I' are essentially unaffected by truncation. For large codes, there is little if any noticeable degradation in compression caused by using truncated infinite codes.]

The use of definite codes is suggested since their lags are minimal ($I = L$), while $I > L$ for non-definite codes. However, non-definite codes can be quite good. Codes (a) and (c) in Fig. 15, for example,

Table VIII—Example of a Good Code from an Infinite Class of Codes for Which $J < 8.7$ for all n

j	k for desired $W(k)$								
even	A	B	C	D	E	F	G	H = $S(i)$	
j: A	0	1	2	3	4	5	6	7	H
C	2	1	0	3	4	5	6	7	F
E	4	1	2	3	0	5	6	7	D
G	6	1	2	3	4	5	0	7	B: j
$S(i)$ =	H	G	F	E	D	C	B	A	j odd
	k for desired $W'(k)$								

have $N(d)$ for which definite codes exist (whence $I^* = L$). In these cases, $I' = L + 1$, recalling the above numerological curiosity. No counterexample to this curiosity is known among $N(d)$ for which definite codes exist, or even among all systematic codes. In general, there are many non-definite cases for which I or I' is quite close to $L + 1$. Thus in these cases the non-definite systematic codes do almost as well as comparable definite codes in achieving first-stage synchronization. Subsequently the non-definite codes may be much better in reaching the second stage. Definite codes are necessarily completely unbalanced,[1,2] with all code words ending in the same digit; therefore a complemented columnized code is impossible. Although a non-complemented code is thus automatically columnized if it uses a definite prefix code as its kernel, second-stage synchronization often must begin with the set of symbols, not just half of them as in the case of a balanced code. However, some definite codes are balanced on two distinct terminal sequences, e.g., on the next to last digit of each code word or on some earlier digit position. In these cases, it is possible to columnize on the two distinct terminal sequences. As an example, consider code (d) of Fig. 15. This is a definite code defined by the set of sequences {0100,0110}, with $N(d) = 0\ 0\ 2\ 2\ 4\ 4\ 8\ 12\ 24\ \cdots$. For each length there are exactly as many code words ending in 00 as in 10. If the sequential code is columnized according to the last two digits of each code word into a 00-set and a 10-set, first-stage synchronization results in one of these two sets, as in the balanced code situation. Since $I = L = 8$, this code has synchronization properties excelling any non-definite code with the given $N(d)$.

6.2 *Guidelines for Choosing Good Encodings*

Because of the perversity of the three-dimensional tradeoffs among compression, complexity and synchronizability, it is pointless to try to give a specific algorithm for choosing the best encoding for a given application. Optimality in any one dimension is of little concern, for slight sacrifices in any of these dimensions often result in great savings in the other two. Besides, no sensible cost metric is known. Nevertheless, the techniques of this paper provide a set of guidelines for the construction of good codes and good encodings.

The first step in selecting a sequential encoding is to establish for each j the code-word lengths which are optimal for the code word $w(i \mid j)$ in the prefix code for the given j, based on the conditional probabilities of $S(i)$, given $S(j)$. This may be done simply using any variant of the Huffman algorithm[6] which derives the code-word lengths. Inspection of the matrix of lengths thus obtained (i.e., of the set of n structure functions, one for each j) indicates what kind of symmetries the code might reasonably have, e.g., whether a reflective symmetry is in order, and whether the code can be a one-kernel code. A complemented bifurcated code may be required for compression reasons, as in Table II. The next step is to choose the basis code(s), considering the set of lengths and the first-stage lag J'. For large codes, techniques of Ref. 1 may be required to aid code selection. The encodings should then be arranged to have the columnization and association properties to help minimize the first- and second-stage lags, respectively. The desired code-word lengths should be taken as suggestive rather than as mandatory; slight departures from these lengths are generally not harmful to compression (especially among large lengths), and may help greatly in decreasing $J^{\sim}$. Care should be taken to avoid having short synchronizing sequences occur only as the result of unlikely sequences of source messages.

VII. CONCLUSIONS

The object of this paper is to present codes with low redundancy, reasonable complexity and intrinsic error tolerance, within a single class of codes designed for that purpose. The approach taken is by no means the only one, although the codes exhibited here seem quite powerful in view of their capabilities. In combination with some additional redundancy (e.g., as in Refs. 1, 4, 17 and 18) to be used for error-detection and/or correction and/or forced framing, the intrinsic properties described here may be used to great advantage.

The sequential prefix codes of this paper can be useful for a wide range of conditional probability distributions. Quantized video picture information provides an example of a class of distributions[19-21] to which these codes seem suited, with respect to compression and synchronizability, as well as to the meaningfulness of tolerating errors for a short period of time. Such distributions are strongly geometric,[20] for which the codes presented here are naturally applicable.

The techniques described here are also applicable to other compression situations. Examples include run-length coding (which has the same essential synchronization problem as differential coding) and predictive, averaging or smoothing schemes, or combinations thereof. In addition, the same techniques are applicable when code words $w(i \mid j)$ need not be provided for many of the transitions, for example, when only relatively small differences are possible.

In conclusion, a wide variety of sequential prefix codes has been presented, with a considerable range of tradeoffs among coding complexity, compression and synchronizability. By not insisting on optimality in any of these, it is possible to obtain codes which are highly satisfactory in all of these respects at the same time.

VIII. ACKNOWLEDGMENT

The author is indebted to E. N. Gilbert and J. O. Limb for helpful comments on the manuscript.

REFERENCES

1. Neumann, P. G., "Efficient Error-Limiting Variable-Length Codes," IRE Trans. Inform. Theory, *IT-8,* (July 1962), pp. 292–304.
2. Neumann, P. G., "On a Class of Efficient Error-Limiting Variable-Length Codes," IRE Trans. Inform. Theory, *IT-8,* (September 1962), pp. S260–266.
3. Neumann, P. G., "Error-Limiting Coding Using Information-Lossless Sequential Machines," IEEE Trans. Inform. Theory, *IT-10,* (April 1964), pp. 108–115.
4. Neumann, P. G., "On Error-Limiting Variable-Length Codes," IEEE Trans. Inform. Theory, *IT-9,* (July 1963), p. 209.
5. Chow, M. C., "Shannon-Fano Encoding for Differentially Coded Video Telephone Signals," Mervin J. Kelly Communication Conference, University of Missouri, Rolla, October 5–7, 1970.
6. Huffman, D. A., "A Method for the Construction of Minimum Redundancy Codes," PROC IRE, *40,* (September 1952), pp. 1098–1101.
7. Gilbert, E. N., and Moore, E. F., "Variable-Length Binary Encodings," B.S.T.J., *38,* No. 4 (July 1959), pp. 933–967.
8. Schützenberger, M. P., "On an Application of Semi-Group Methods to Some Problems in Coding," IRE Trans. Inform. Theory, *IT-2,* (September 1956), pp. 47–60.
9. Gill, A., *Introduction to the Theory of Finite-State Machines,* New York: McGraw-Hill Book Co., (1962).
10. Hennie, F., *Finite State Models for Sequential Machines,* John Wiley and Sons, (1968).
11. Winograd, S., "Input-Error-Limiting Automata," J. Assn. Computing Machinery, *XI,* (1964), pp. 338–351.
12. Mandelbrot, B., "On Recurrent Noise Limiting Coding," Proc. Symp. on Information Networks, Polytechnic Inst. of Brooklyn, (April 1954), pp. 205–221.
13. Schützenberger, M. P., "On Synchronizing Prefix Codes," Information and Control, *7,* (1964), pp. 23–26.
14. Laemmel, A. E. and Rudner, B., "Study of the Application of Coding Theory," Internal Report PIBEP-69-034 (AD 691 764), Polytechnic Inst. of Brooklyn (June 1969).
15. Fischler, M. A., and Tannenbaum, M., "Synchronizing and Representation Problems for Sequential Machines with Masked Outputs," Eleventh Annual Symposium on Switching and Automata Theory, Santa Monica, California, October 1970.
16. Huffman, D. A., "Canonical Forms for Information-Lossless Finite-State Logical Machines," IRE Trans., *CT-6* or *IT-5* (Supplement) (May 1959), pp. 41–59. Also appears in E. F. Moore, *Sequential Machine Papers,* New York: Addison-Wesley (1964).
17. Hatcher, T. R., "On a Family of Error-Correcting and Synchronizable

Codes," IEEE Trans. Inform. Theory, *IT-15*, (September 1969), pp. 620–624.
18. Scholtz, R. A., and Welch, L. R., "Mechanization of Codes with Bounded Synchronization Delays," IEEE Trans. Inform. Theory, *IT-16*, (July 1970), pp. 438–446.
19. Oliver, B. M., "Efficient Coding," B.S.T.J., *31*, No. 4 (July 1952), pp. 724–750.
20. Schreiber, W. F., "The Measurement of Third Order Probability Distributions of Television Signals," IRE Trans. Inform. Theory, *IT-2*, (September 1956), pp. 94–105.
21. Graham, R. E., "Predictive Quantizing of Television Signals," IRE WESCON, Part 4, (August 1958), pp. 147–157.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

Theory for Some Asynchronous Time-Division Switches

By J. E. MAZO

(Manuscript received September 24, 1970)

A two-wire active asynchronous time-division switch has recently been proposed by J. O. Dimmick, T. G. Lewis, and J. F. O'Neill. The interrupted energy transfer from one filter to another is accomplished asynchronously in order that more efficient use may be made of processing time of talker pairs on the switching bus. After first formulating, in a concise mathematical fashion, the effect of passing a signal through such a randomly time-varying circuit, we focus attention on optimizing an important filter response function. Typically, two percent rms jitter in the transfer times yields an output S/N of 30 dB independent of signal spectrum. The timing stabilization required to obtain small jitter is also discussed and an exact solution for exponential processing time is obtained. This latter result may be put to good use in studying the efficiency of the switch. Conservatively, asynchronous operation should increase traffic capacity threefold.

Finally, a speech wave which passes through a sample-and-hold circuit with random sampling times is considered upon being reconstructed with a fixed filter. This is a model for a four-wire active asynchronous switch, and results are compared with the two-wire situation.

I. INTRODUCTION

Voice switching systems have most commonly been based on controlling electromechanical switches which select and hold a spatially distinct path for each conversation. The technology used to implement such a space-division network (crossbar or ferreed switches) usually results, in practice, of an individual path having much larger bandwidth than is required for faithful transmission of the signal. The space and cost of these switches makes other solutions desirable for many applications. One line of attack has been to keep the space-division concept, but replace the electromechanical relays with semiconductor switches. These techniques, however, still suffer from the hard-to-grow nature of multistage space division networks. A more promising solution seems to be to place all the conversations on a wideband bus using time-division techniques. In fact, the 101 Electronic Switching System (ESS) is such a time-division switch. This system uses resonant energy transfer to "move" periodically measured samples of a speech wave from an incoming line to an outgoing one. In the same vein, an asynchronous time-division switch has recently been proposed by J. O. Dimmick, T. G. Lewis, and J. F. O'Neill.[1] This switching arrangement makes use of active energy transfer between filters rather than resonant energy transfer and allows a variable time slot for transferring each speech sample through the switch. The asynchronous nature of this switch allows a more efficient use of processing time than is possible to achieve synchronously. However, a consequence of this virtue is a periodic sampling of the input waves. While the synchronous switch can make use of the usual sampling theory to guarantee faithful reproduction, the asynchronous switch cannot. Further, the modifications of the theory which are required to discuss the proposed asynchronous switch are not simple, for the random sampling causes some feedback energy which is later retransmitted, further adding to the output noise. Our immediate purpose will be, then, to present theoretical work relevant to this problem. We discuss the quality of transmission [measured by the output signal-to-noise ratio (S/N)] as a function of jitter in the sample values and as a functional of a certain filter response function upon which the feedback energy depends. The optimum function is found. Also a technique suggested in Ref. 1 for keeping the jitter small is discussed theoretically, and the combined question of how jitter, quality of transmission, and increased efficiency are related is answered.

Section X summarizes our conclusions and, barring some terminology introduced in the text, may be read next.

II. MATHEMATICAL MODEL

Consider the diagram in Fig. 1 which represents a talker and listener on a switching bus. There will be many such pairs on a particular bus, but we need now concentrate on only one. The way the switch works is that, at approximately periodic instants of time, the two identical

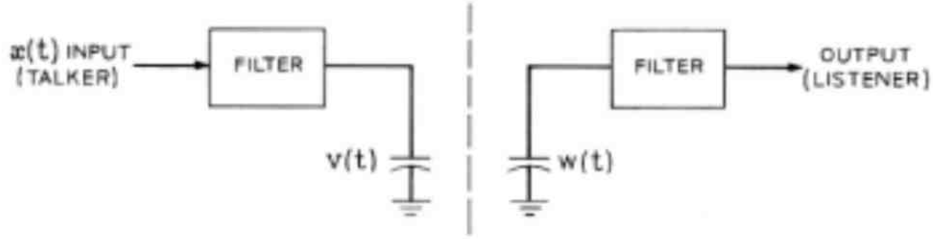


Fig. 1—Model for a "talker"—"listener" pair on an asynchronous switching bus.

capacitors which are shown in Fig. 1 are interchanged* and energy transfer is effected. The following assumptions are made:

(*i*) The input signal $x(t)$ is bandlimited to W Hz and the switching occurs at times $t_n = nT + \epsilon_n$ where $T \leqq 2W$ and ϵ_n is small compared to T.

(*ii*) The filtering aspect of the filter is neglected in the sense that if *no* switching is performed, then $v(t) = x(t)$. In particular this means that if current impulses of area $x(nT)$ are applied at times nT to the listeners capacitor, $x(t)$ will occur at the output.

(*iii*) Suppose that one volt is placed on the listening capacitor when no energy is stored in the filter. The voltage across the capacitor $(t \geqq 0)$ under these conditions will be called $z(t)$, and we assume $z(0+) = 1$.

Our goal will be to determine, for given spectrum of the input, and for given second-order statistics of the ϵ_n sequence, the output S/N. The proper design of $z(t)$ will be of major concern in the analysis.

We shall call the problem just described two-wire switching.

III. FUNDAMENTAL EQUATIONS

In this section an exact equation which describes the two-wire talker-listener switching situation will be derived. As in Fig. 1, let $v(t)$ be the actual voltage on the talker's capacitor and $w(t)$ the actual voltage on the listener's capacitor. In the absence of switching, $v(t) = x(t)$. In addition, at times $nT + \epsilon_n$ voltages $[w(nT + \epsilon_n-) - v(nT + \epsilon_n-)]$ are placed on the capacitor, where $w(t-)$ means the limiting

* Of course in reality they are not interchanged. What happens is that at the given instant of time at which the "switch" is to occur, the instantaneous voltages are measured, certain current sources (not shown) are activated and a fixed value of current flows for a duration just sufficient to effect the interchange of charges, and hence voltages, of the equal capacitors. All this occurs in a negligible period of time compared with the response time of the filters.

value of $w(t')$ as t' approaches t from below. Since switching occurs at times $nT + \epsilon_n$, continuity cannot be assumed at these instants and left and right limits must be distinguished. Using the definition of $z(t)$ we have by the superposition principle

$$v(t) = x(t) + \sum_{n=0}^{l(t)} [w(nT + \epsilon_n -) - v(nT + \epsilon_n -)]z(t - nT - \epsilon_n) \tag{1}$$

where $l(t)$ is the integer which satisfies

$$l(t)T + \epsilon_l \leqq t < (l(t) + 1)T + \epsilon_{l+1}.$$

Likewise for $w(t)$ we may write

$$w(t) = \sum_{n=0}^{l(t)} [v(nT + \epsilon_n -) - w(nT + \epsilon_n -)]z(t - nT - \epsilon_n). \tag{2}$$

If we let $t \rightarrow (kT + \epsilon_k -)$ in equations (1) and (2), we have the pair of equations

$$v(kT + \epsilon_k -) = x(kT + \epsilon_k) - \sum_{n=0}^{l(t)} [v(nT + \epsilon_n -) - w(nT + \epsilon_n -)] \cdot z[(k - n)T + \epsilon_k - \epsilon_n -] \tag{3}$$

and

$$w(kT + \epsilon_k -) = \sum_{n=0}^{l(t)} [v(nT + \epsilon_n -) - w(nT + \epsilon_n -)] \cdot z[(k - n)T + \epsilon_k - \epsilon_n -]. \tag{4}$$

It is very useful to now introduce the vectors Y, V, W and a matrix Z defined by

$$\begin{aligned} y_k &= x(kT + \epsilon_k), \\ v_k &= v(kT + \epsilon_k -), \\ w_k &= w(kT + \epsilon_k -), \end{aligned} \tag{5}$$

and

$$Z_{kn} = z[(k - n)T + \epsilon_k - \epsilon_n -]. \tag{6}$$

Note $Z_{kn} = 0$ if $k \leqq n$ since $z(t) = 0$ for $t < 0$.
Using equations (5) and (6), (3) and (4) become

$$V = Y - Z(V - W), \tag{7}$$

$$W = Z(V - W). \tag{8}$$

Solving the above pair for $(V - W)$ gives*

$$(V - W) = [I + 2Z]^{-1}Y. \tag{9}$$

Equation (9) determines exactly the behavior of the switch, for the output is determined by applying the sequence $\{V_k - W_k\}$ at times $kT + \epsilon_k$ to the filter (here assumed ideal). We emphasize that the jitter enters not only through the time at which the δ-functions are applied, but also in the quantities Z and Y.

Consider now a $z(t)$ such as that shown in Fig. 2 which passes through zero at all times kT, $k > 0$. If further there is no jitter, then $Z = 0$, and equations (5) and (9) show that impulses of area $x(nT)$ are applied at times nT, thus giving $x(t)$ at the output. In short, in the absence of jitter, any $z(t)$ which passes through zero at all positive integer multiples of the sampling interval will be optimum. One might now conclude that a $z(t)$ which is zero in a sufficiently large neighborhood of each such crossing will not see the jitter and would therefore be optimum in the presence of jitter. However the following argument (by J. F. O'Neill) suggests that such is not the case. Consider sampling a dc signal at random times. We must get enough power through the switch to reproduce the signal. Surely if we sample late, we are lagging in power and we would like to increase the area of the impulse; likewise if we sample early, we seem to be supplying extra power and so should decrease the impulse area. There should be a design of $z(t)$ which would conspire with the jitter to reduce jitter noise even below that noise obtained for the class of $z(t)$ which do not see the jitter.

IV. WHAT IS THE OUTPUT NOISE?

Define the vector

$$\gamma \equiv V - W \tag{10}$$

and the functions

$$\psi_n(t) = \frac{\sin \frac{\pi}{T}(t - nT)}{\frac{\pi}{T}(t - nT)} \tag{11}$$

for all integer n. The properties of $\psi_n(t)$ that we use are

$$\psi_n(t) = \psi_{-n}(-t), \tag{12}$$

* The inverse of $(I + 2Z)$ always exists since it is triangular and all its diagonal elements are unity.

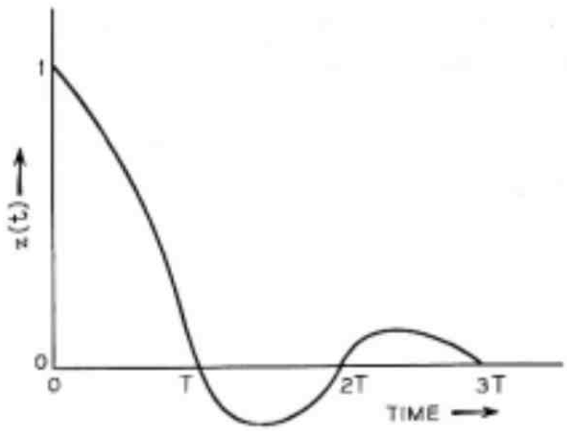


Fig. 2—An illustrative curve for $z(t)$.

$$\int_{-\infty}^{\infty} \psi_n(t)\psi_m(t)\, dt = T\delta_{nm}\,, \tag{13}$$

and

$$\psi_n(t_1 - t_2) = \sum_{m=-\infty}^{\infty} \psi_{n+m}(t_1)\psi_m(t_2). \tag{14}$$

We assume that the reconstructing filter impulse response is given by $\psi_0(t)$ and thus the output error signal is

$$e(t) = \sum_{n=0}^{\infty} \gamma_n \psi_n(t - \epsilon_n) - \sum_{n=0}^{\infty} x_n \psi_n(t) \tag{15}$$

and has average power[*] N where

$$N = \lim_{k\to\infty} \frac{1}{KT} \int_{-KT}^{KT} e^2(t)\, dt = \lim_{k\to\infty} \frac{1}{KT} \int_{-\infty}^{\infty} e_K^2(t)\, dt. \tag{16}$$

In the right side of equation (16), we have introduced $e_k(t)$ which is the error signal truncated to K pulses; that is, the upper limit of the sums in equation (15) is K instead of infinity. If we use equation (14) to expand $\psi_n(t - \epsilon_n)$ which occurs in equation (15), find $e^2(t)$ and do the time integral, we obtain

$$\int_{-\infty}^{\infty} e_K^2(t) = T \sum_{n,k=0}^{K} \gamma_n\gamma_k \sum_{m=-\infty}^{\infty} \psi_m(\epsilon_n)\psi_{n+m-k}(\epsilon_k) + T \sum_{0}^{K} x_n^2 - 2T \sum_{n,k=0}^{K} x_n\gamma_k\psi_{n-k}(\epsilon_k). \tag{17}$$

[*] This corresponds to the noise in a band up to $1/2T$ Hz. If T is less than the Nyquist rate, then some out-of-band noise is included here. In practice T will be about half the Nyquist interval in order to make filter design problems easier, but then the inclusion of the out-of-band noise seems fair at this stage. Figure 3 shows a picture of the relevant bandwidths.

Equation (14) may be used again to recombine the first terms of the right side of equation (17) and so

$$\frac{1}{KT}\int_{-\infty}^{\infty} e_K^2(t) = \frac{1}{K}\sum_{n,k=0}^{K} \gamma_n\gamma_k\psi_{n-k}(\epsilon_k - \epsilon_n) + \frac{1}{K}\sum_{0}^{K} x_n^2 - \frac{2}{K}\sum_{n,k=0}^{K} x_n\gamma_k\psi_{n-k}(\epsilon_k). \tag{18}$$

Recall again that

$$\begin{aligned} Z_{kn} &= z[(k-n)T + \epsilon_k - \epsilon_n], \\ \gamma_k &= (1 + 2Z)_{kl}^{-1} y_l\ , \\ y_l &= x(lT + \epsilon_l). \end{aligned} \tag{19}$$

It is useful to define a new matrix θ by

$$\theta = \frac{2Z}{1 + 2Z} \tag{20}$$

so that

$$\gamma = Y - \theta Y. \tag{21}$$

Substitution of equation (21) in equation (18) splits the expressions into two types of terms. The first type, called collectively A_K, do not involve the filter $z(t)$ while the remaining ones, called B_K, do. We thus have

$$N = \lim_{K\to\infty} E[A_K] + \lim_{K\to\infty} E[B_K] \equiv A + B, \tag{22}$$

where in equation (22), the expectation is taken over both the signal and jitter statistics. The quantities A_k and B_k are given by

$$A_K = \frac{1}{K}\sum_{n,k=0}^{K} y_n y_k\psi_{n-k}(\epsilon_k - \epsilon_n) + \frac{1}{K}\sum_{0}^{K} x_n^2 - \frac{2}{K}\sum_{n,k=0}^{K} x_n y_k\psi_{n-k}(\epsilon_k), \tag{23}$$

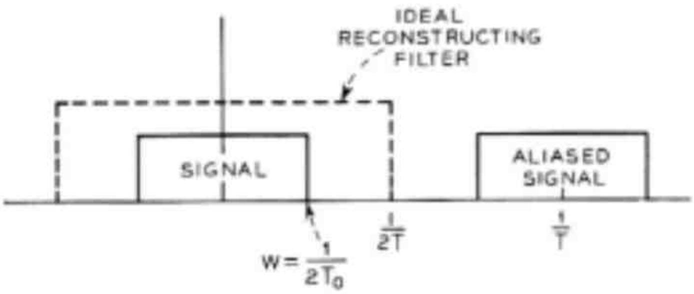


Fig. 3—Relevant bandwidths in terms of the sampling interval T.

$$B_K = -\frac{2}{K}\sum_{n,k=0}^{K} y_n(\theta y)_k[\psi_{n-k}(\epsilon_k - \epsilon_n) - \psi_{n-k}(\epsilon_k)] - \frac{2}{K}\sum_{n,k=0}^{K} (\theta y)_k[y_n - x_n]\psi_{n-k}(\epsilon_k) + \frac{1}{K}\sum_{n,k=0}^{K} (\theta y)_n(\theta y)_k\psi_{n-k}(\epsilon_k - \epsilon_n). \tag{24}$$

One may note that if $z(t)$ is such that when the filter is designed so as not to feel the jitter (as described in Section III), then $B = 0$ and the noise would be given by the filter independent term A alone. The function of the filter design is to make B as negative as possible.

To proceed further we make essential use of the fact that the ϵ_k are small. In addition, since $z(kT) = 0$ is an optimum solution in the absence of jitter we shall keep the requirement

$$z(kT) = 0, \qquad k = 1, 2, \cdots \tag{25}$$

and see what further optimization can now be made. This will amount to designing the slopes of the function $z(t)$ when it goes through zero. The evaluation of A and B for small ϵ is carried out in Appendix B. Introducing the correlation function of the signal

$$R(\tau) = E[x(t)x(t + \tau)] \tag{26}$$

and the correlation function of the jitter

$$J(k - n) = E[\epsilon_k\epsilon_n] \tag{27}$$

we find that when the jitter is independent from sample to sample,

$$A = \lim E[A_K] = \sigma_\epsilon^2\left[-\ddot{R}(0) + \frac{1}{3}\frac{\pi^2}{T^2}R(0)\right]. \tag{28}$$

In equation (28), σ_ϵ^2 is the variance of the jitter and is given, for independent jitter, by

$$\sigma_\epsilon^2 = \overline{\epsilon^2} - \bar{\epsilon}^2 = J(0) - J(n \neq 0). \tag{29}$$

To write the corresponding expression for B_k we introduce, as is done in Appendix B, the derivatives $\dot{z}_s$ of the function $z(t)$ at the zeros, that is

$$\dot{z}_s \equiv \frac{d}{dt}z(t)\bigg|_{t=sT} \qquad s = 1, 2, 3, \cdots \tag{30}$$

and also the constants*

$$\Gamma_s = \dot{\psi}_s(0) = \dot{\psi}_0(-s) = \frac{(-1)^{s+1}}{Ts}, \qquad s \neq 0. \tag{31}$$

Then, for jitter independent from sample to sample,

$$\begin{aligned} B &= \lim E[B_K] \\ &= 4R(0)\sigma_\epsilon^2\left[\sum_{s=1}^{\infty}\Gamma_s\dot{z}_s + 2\sum_{s=1}^{\infty}\dot{z}_s^2 - \sum_{s=1}^{\infty}\frac{\dot{R}(s)}{R(0)}\dot{z}_s + 2\sum_{\substack{s,t=1\\ s>t}}^{\infty}\dot{z}_s\dot{z}_t\frac{R(s-t)}{R(0)}\right]. \end{aligned} \tag{32}$$

Expressions for A_k and B_k which include the effects of correlation in the jitter are also derived in Appendix B and will be discussed in Section V. At the moment we merely state that positive correlation between jitter values will tend to reduce the noise power given by the sum of equations (28) and (32). Equation (32) gives, incidentally, more of a physical interpretation as to the breakup of the noise into A and B terms. The filter mentioned at the end of Section III which does not see the jitter certainly has $\dot{z}(t) = 0$ at each crossing, and thus from equation (32), $B = 0$ for that filter. Thus the A term is the noise power for the "blind" filter; it will be improved upon whenever B is negative.

If we now define the functional $F[z]$ of $z(t)$ by

$$F[z] = \sum_{s=1}^{\infty}\Gamma_s\dot{z}_s + 2\sum_{s=1}^{\infty}\dot{z}_s^2 - \sum_{s=1}^{\infty}\frac{\dot{R}(s)}{R(0)}\dot{z}_s + 2\sum_{\substack{s,t=1\\ s>t}}^{\infty}\dot{z}_s\dot{z}_t\frac{R(s-t)}{R(0)}, \tag{33}$$

then optimum choice of any $\dot{z}_k$ is found by simple differentiation of equation (33):

$$\frac{\partial F[z]}{\partial \dot{z}_k} = 0 = \Gamma_k + 2\dot{z}_k - \frac{R(k)}{R(0)} + 2\sum_{t=1}^{\infty}\dot{z}_t\frac{R(k-t)}{R(0)}, \qquad k = 1, 2, \cdots. \tag{34}$$

Thus for given signal statistics, equation (34) must be solved for the optimum set of $\{z_k\}$. In reality, equation (34) does not have to be taken seriously for all positive integer k, since the response of a realistic filter will die off rapidly with time. Thus in equation (33), most of the $\dot{z}_k$ can be set to zero and only a few retained. If the first Z are thus retained we shall refer to this as designing the first Z zero

* Recall the definition of $\psi_s(t)$ given in equation (11). Also in equation (31), dots denote differentiation with respect to the time variable.

crossings. Of course equation (34) will hold only for the k such that $\dot{z}_k \neq 0$ *a priori*.

As our first example we consider the case where the signal spectrum is flat and extends to $\Omega_{\max} = \pi/T$ rad/sec. We refer to this as the full spectrum case. The interval T in this case is also the Nyquist interval, and no oversampling is done as one would expect to do in practice (the latter case will be treated shortly). We have

Full Spectrum Case:

$$\begin{aligned} R(\tau) &= R(0)\psi_0(\tau), \\ R(s) &= 0, \qquad s \neq 0, \\ \dot{R}(0) &= 0, \\ \dot{R}(s) &= R(0)\,\frac{(-1)^s}{Ts}\,, \qquad s \neq 0, \\ -\ddot{R}(0) &= R(0)\,\frac{1}{3}\,\frac{\pi^2}{T^2}. \end{aligned} \tag{35}$$

Thus equation (34) yields as the optimum solution for designing all zeros in the full spectrum case

$$\dot{z}_k = \frac{(-1)^k}{2Tk} = -\frac{\Gamma_k}{2}. \tag{36}$$

Proceeding to evaluate A and B, we have

$$\begin{aligned} A &= R(0)\cdot\frac{2}{3}\,\frac{\pi^2}{T^2}\,\sigma_\epsilon^2\,, \\ B &= -R(0)\cdot\frac{1}{3}\,\frac{\pi^2}{T^2}\,\sigma_\epsilon^2\,. \end{aligned} \tag{37}$$

Since $R(0)$ is the signal power, we have

$$\frac{S}{N} = \frac{1}{3}\,\frac{\pi^2}{T^2}\,\sigma_\epsilon^2\,. \tag{38}$$

An important fact can be gleaned from equation (37): one should not generally expect the optimum filter to be very much better than the "blind" filter.* In the reasonable example which we have just worked, only a gain of 3 dB is achieved.

* Of course the blind filter is itself a highly designed filter.

Table I—Full Spectrum

Output S/N	Jitter Standard Deviation σ_s
15 dB	$0.1\ T$
21 dB	$0.05\ T$
29 dB	$0.02\ T$
35 dB	$0.01\ T$
55 dB	$0.001\ T$

Table I calculates equation (38) for several values of jitter. For an output S/N of 30 dB, two percent jitter is required.

V. EFFECTS OF SIGNAL CORRELATION

Equation (34), the basic design equation for the filter response $z(t)$, can be conveniently rewritten in matrix notation

$$(I + M)\dot{z} = A \tag{39}$$

where the vectors $\dot{z}$ and A have components

$$(\dot{z})_k = \dot{z}_k \,, \tag{40a}$$

$$A_k = \frac{1}{2}\left[\frac{\dot{R}(k)}{R(0)} - \Gamma_k\right] \qquad k = 1, 2, \cdots,$$

and the matrix M is defined by

$$M_{kl} = \frac{R(k - l)}{R(0)}. \tag{40b}$$

The matrix I is of course the identity. The dimension of all the above quantities is Z, where Z is the number of zero crossings that one wishes to design for.

To solve equation (39) one must in general invert the matrix $(I + M)$. This was possible in the full spectrum case because M was a diagonal matrix; in general it is hard to do exactly, unless the dimension Z is small. A case of special interest for applications is a qualitative understanding of the situation for a flat signal bandwidth up to $\Omega_{\max} = \pi/2T$. This corresponds to sampling at twice the Nyquist rate and better approximates the situation to be encountered in practice. We call it the half spectrum case. We have

Half Spectrum Case:

$$\begin{aligned} R(sT) &= R(0)\,\frac{2}{\pi s}\sin\frac{\pi}{2}s, \qquad s \neq 0; \\ \dot{R}(sT) &= R(0)\left[\frac{\cos\frac{\pi}{2}s}{sT} - \frac{2}{\pi}\,\frac{\sin\frac{\pi}{2}s}{s^2T}\right]; \\ -\ddot{R}(0) &= \frac{1}{3}\left(\frac{\pi}{2T}\right)^2\cdot R(0). \end{aligned} \tag{41}$$

The A term yields, from equation (28)

$$\frac{A}{R(0)} = \frac{5}{12}\,\frac{\pi^2}{T^2}\,\sigma_\epsilon^2 = \frac{\sigma_\epsilon^2}{T^2}\,(4.16). \tag{42}$$

In the B term we must solve equation (39). We do so by inverting the matrix by hand fot the cases of designing either the first zero, the first two zeros, or the first three zeros ($Z = 1, 2, 3$ respectively). We obtain

$$\begin{aligned} \dot{z}_{\text{opt}} &\underset{Z=1}{=} \frac{1}{T}\,[-0.409], \\ \dot{z}_{\text{opt}} &\underset{Z=2}{=} \frac{1}{T}\begin{bmatrix} -0.454 \\ +0.145 \end{bmatrix}, \\ \dot{z}_{\text{opt}} &\underset{Z=3}{=} \frac{1}{T}\begin{bmatrix} -0.442 \\ +0.179 \\ -0.118 \end{bmatrix}. \end{aligned} \tag{43}$$

These numbers give for B

$$\begin{aligned} \frac{B}{R(0)} &\underset{Z=1}{=} \frac{\sigma_\epsilon^2}{T^2}\,(-1.34), \\ \frac{B}{R(0)} &\underset{Z=2}{=} \frac{\sigma_\epsilon^2}{T^2}\,(-1.49), \\ \frac{B}{R(0)} &\underset{Z=3}{=} \frac{\sigma_\epsilon^2}{T^2}\,(-1.50). \end{aligned} \tag{44}$$

From equations (44) and (42) we note that $Z = 1$ design improves the "blind" filter by only 1.7 dB and $Z = 2, 3$ by 1.8 dB. Output S/N are given in Table II. For a given fraction of jitter essentially the same output S/N is obtained as for the full spectrum case shown in Fig. 1. Let

Table II—Half Spectrum*

Output S/N			σ_s/T
Z			
1	2	3	
15.5 dB			0.1
21.5 dB	S	S	0.05
29.5 dB	A	A	0.02
35.5 dB	M	M	0.01
55.5 dB	E	E	0.001

* The noise is here measured in a bandwidth twice the signal bandwidth.

us emphasize here that the jitter is listed as a fraction of the nominal sampling time and not in absolute units. Thus for a *fixed* signal spectrum and *fixed* jitter in *seconds,* one percent jitter for the full spectrum case would correspond to two percent for half spectrum situation.

The effect of positive correlation can be seen by comparing equation (43) with equation (36). Positive correlation tends to flatten the slopes at the zero crossings somewhat. Pursuing positive signal correlation to the utmost, we consider one more case, the case of a dc signal. In this case $\dot{R}(\tau) = 0$, and we want to solve

$$(I + M)\dot{z} = -\tfrac{1}{2}\Gamma \tag{45}$$

where

$$I + M = \begin{bmatrix} 2 & 1 & 1 & 1 & \cdots \\ 1 & 2 & 1 & 1 & \cdots \\ 1 & 1 & 2 & 1 & \cdots \\ 1 & 1 & 1 & 2 & \cdots \end{bmatrix}. \tag{46}$$

Let us first consider the $(Z \times Z)$ version of equation (45). One may verify that for dc

$$(I + M)^{-1} = I - \frac{1}{Z + 1} U \tag{47}$$

where U is the $Z \times Z$ matrix that has all its elements equal to unity. If we let V equal the Z dimensional vector which has all its components equal to unity, then we get

$$\dot{z} = -\frac{\Gamma}{2} + \frac{1}{2(Z + 1)T}\left[\sum_{n=1}^{Z} \frac{(-1)^{n+1}}{n}\right]V. \tag{48}$$

Finally for dc signal we have

$$\frac{A}{R(0)} = \frac{\sigma_\epsilon^2}{T^2} \cdot \frac{\pi^2}{3} \tag{49}$$

and

$$\frac{B}{R(0)} = \sigma_\epsilon^2(-4)\dot{z}'\left(-\frac{\Gamma}{2}\right)$$

which has the limiting value $(Z \to \infty)$

$$\frac{B}{R(0)} = \frac{-\sigma_\epsilon^2}{T^2} \sum_{s=1}^{\infty} \frac{1}{s^2} = \frac{-\sigma_\epsilon^2}{T^2} \frac{\pi^2}{6}. \tag{50}$$

Again we see for an exact solution that only 3 dB is gained by pursuing optimum design past the "blind" filter design. The output (S/N) is 3 dB better than equation (38); one would expect dc to be influenced less by jitter.

We note the remarkable fact that as $Z \to \infty$, equation (48) yields exactly the same solution $\dot{z}_k = -\Gamma_k/2$, as one gets for the full spectrum case. This is not true for a fixed finite Z.

Judging from our two exact and one approximate solution, *designing*

$$\dot{z}_k = -\frac{\Gamma_k}{2} \tag{51}$$

is an optimum design independent of signal spectrum. Also the "blind" filter is a very good design independent of signal spectrum, being about only 3 dB worse than optimum. This all assumes that the jitter is independent from sample to sample.

VI. EFFECTS OF JITTER CORRELATION

Until now, any correlation between jitter samples has been ignored. In fact, some positive correlation is to be expected in the jitter statistics. We do not feel it is large for the way we have described the jitter in the previous sections* (call it ϵ-jitter), so that our previous design is not affected. Nevertheless, we should note that any positive correlation which exists between the jitter values will improve the output S/N above that obtained by assuming that the jitter from sample to sample is independent. The physical argument as to why this should be true is quite simple. Assume ϵ_i is a nonzero constant, the same con-

* See Section VII.

stant for all i. Then we are sampling at time $t_n = nT + \epsilon$. From equation (19), $Z = 0$ and we are supplying current impulses of area $x(n\tau + \epsilon)$ at times $n\tau + \epsilon$ to the output filter and the signal is reconstructed perfectly. Positive correlation helps.

An important check on our work thus far can be made by assuming $\epsilon = \text{const.}$, and checking that the A and B terms, which for jitter are given in Appendix B, vanish. Since for this case

$$J(0) = J(s), \qquad \text{for all } s, \tag{52}$$

the B term given in equation (112) obviously vanishes. The explicit demonstration of the vanishing of the A term in equation (107) is not at all so obvious. Term (107), the A term, for constant jitter, is seen to be proportional to [note $\dot{\psi}_0(0) = 0$]

$$\begin{aligned} -\ddot{R}(0) - \sum_s R(s)\ddot{\psi}_s(0) + 2 \sum_s R(s)\dot{\psi}_s(0) \\ = -\ddot{R}(0) - \sum_s R(s)\ddot{\psi}_0(sT) - 2 \sum_s \dot{R}(s)\dot{\psi}_0(sT). \end{aligned} \tag{53}$$

Define a function of time $g(t)$ by

$$g(t) = \frac{d^2}{dt^2}\,[R(t)\psi_0(t)].$$

Then the right side of equation (53) is

$$\sum_{s=-\infty}^{\infty} g(sT)$$

and by Poisson's sum formula

$$\sum_{s=-\infty}^{\infty} g(sT) = \frac{1}{T} \sum_{m=-\infty}^{\infty} G\left(\frac{2\pi}{T}\, m\right) \tag{54}$$

where $G(\omega)$ is the Fourier Transform of $g(t)$. Now $\psi_0(t)$ is band-limited to $|\omega| < \pi/T$ and so is $R(t)$. The spectrum of $R(t)\psi_0(t)$ extends only to $2\pi/T$ and in fact vanishes at $\omega = 2\pi/T$ since it is a convolution. Thus only the $m = 0$ term of equation (54) could possibly contribute, but this contribution also vanishes because the second time derivative in the definition of $g(t)$ introduces a double zero in the spectrum of $G(\omega)$ at $\omega = 0$.

Further use will be made of the expression for output noise with correlated jitter when the ϵ-model for jitter is compared with another model discussed later in Section IX.

VII. TIMING STABILIZATION

At this point we return to consideration of many talker-listener pairs. Let L be the number of possible such pairs. This number depends on how rapidly each talker can be processed (switched). For example, if each talker is sampled at a nominal rate of $1/T$ and t_i is the processing time for the ith speaker, a constraint which reads something like

$$\sum_{i=1}^{L} t_i \approx T \tag{55}$$

must be imposed. Further, the timing errors ϵ_i are hopefully small. A scheme for stabilizing the sampling rate has, in fact, been proposed but T. G. Lewis.[1,2] An interpretation of this scheme due to Saltzberg and Pasternak is shown in Fig. 4. For immediate convenience let us not worry about normalization and let the dashed line have unity slope. The large dots in Fig. 4 represent a talker starting to be sampled, and let the length of time required for the ith sample be t_i. The time t_i represents the length of time that the current source pumps current to effect the interchange of the talker-listener capacitors and, in accordance with our earlier assumptions about measurement times being short compared to filter bandwidths, we have $t_i \ll T$. After many counts (measurements), say L, we return to resample a given talker. We particularly note that a talker is never sampled early. We see that in Fig. 4 after the time t_3, our path hits the 45° line early, and we

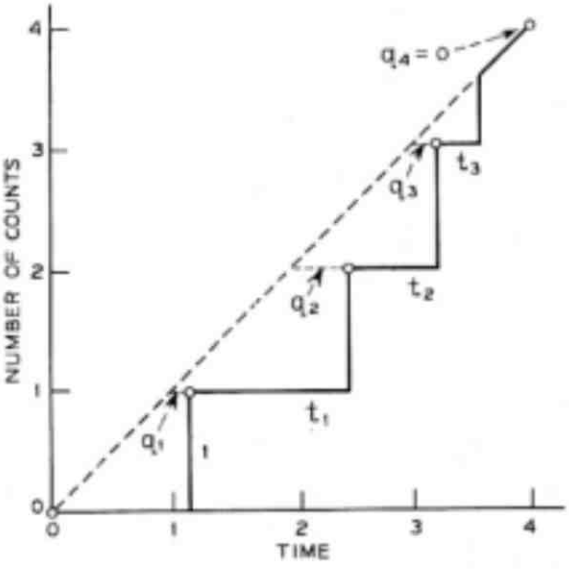


Fig. 4—Timing stabilization.

"pad" time until the fourth count occurs. Also, as shown, the variables q_n represent the horizontal distance from the 45° line to the dot representing the count. They are positive random variables and are given by the equation*

$$q_n = \max \begin{cases} 0 \\ q_{n-1} + t_{n-1} - 1. \end{cases} \tag{56}$$

The random length of times t_{n-1} are assumed to be identically distributed and independent of q_{n-1}. We shall proceed to derive an integral equation for the probability distribution function for q_n.† We note the possibility of a δ-function at $q_n = 0$ and thus write for the density function $q_n(x)$

$$q_n(x) = \alpha_n \, \delta(x) + (1 - \alpha_n) p_n(x). \tag{57}$$

The number α_n is the probability of having $q_n = 0$, and $p_n(x)$ describes the continuous portion of the density. If we let $u(t)$ denote the density of t_n and $*$ denote convolution, we have, using equation (56),

$$\alpha_n = \alpha_{n-1} \int_0^1 u(t) \, dt + (1 - \alpha_{n-1}) \int_0^1 [p_{n-1} * u] \, dt', \tag{58a}$$

$$d_n(x) = \alpha_{n-1} \frac{u(t + 1)}{\displaystyle\int_1^\infty u(t') \, dt'} + (1 - \alpha_{n-1}) \frac{\displaystyle\int_0^\infty p_{n-1}(t') u(x + 1 - t') \, dt'}{\displaystyle\int_1^\infty dt \int_0^\infty p_{n-1}(t') u(t - t') \, dt'} . \tag{58b}$$

A steady-state α and $p(x)$ would obey equation (58) with all indices removed. We shall write the steady-state equation. Let K denote the known constant,

$$K \equiv \int_1^\infty u(t') \, dt', \tag{59}$$

let, for $x > 0$, $v(t)$ denote the known density function

$$v(t) \equiv \frac{u(t + 1)}{K}, \tag{60}$$

* Except for a time scale normalization, the sequence $\{q_{nL}\}$ $n = 1, 2 \cdots$ corresponds to the sequence $\{\epsilon_n\}$ of previous sections when there are L talkers on the switch.

† B R. Saltzberg and G. P. Pasternak[3] had also derived an integral equation for q_n, and no doubt our equation is substantially equivalent to their equation. An entirely different approach (i.e., an equation for a different set of variables to describe the same problem) has also been discussed by J. Salz and R. D. Gitlin.[4]

and introduce the *unknown* constant

$$\rho = \int_0^1 [p * u(t')] \, dt'. \tag{61}$$

Then the steady-state solution ($\lim n \to \infty$) satisfies, if it exists,

$$\alpha = (1 - K)\alpha + (1 - \alpha)\rho, \tag{62a}$$

$$p(x) = \alpha v(x) + (1 - \alpha) \int_0^\infty p(t')v(x - t') \, dt'. \tag{62b}$$

If we let $C(\omega)$ denote the characteristic function of $p(x)$ then equation (62) yields

$$C(\omega) = \frac{\alpha V(\omega)}{1 - (1 - \alpha) V(\omega)} \tag{63}$$

where

$$\alpha = \frac{\rho}{K + \rho}. \tag{64}$$

One may now imagine determining the unknown constant ρ in the following manner. Equation (63) determines $C(\omega)$ and therefore the density $p(x)$ *in terms of* ρ. Form the convolution of p with u and set the finite integral of this convolution equal to ρ in accordance with equation (61). This then is an equation which can be used for the numerical determination of ρ. For the case

$$u(t) = \beta \exp(-\beta t) \tag{65}$$

the procedure may be carried through exactly. We find that the density $p(x)$ of the q variable is also exponential and is

$$p(x) = \alpha\beta \exp(-\alpha\beta x), \tag{66}$$

where the probability α of having q exactly zero is related to β by

$$1 - \alpha = \exp(-\beta\alpha). \tag{67}$$

Using the inequality

$$1 - x < e^{-x}, \qquad x \neq 0,$$

we see that equation (67) has a nonvanishing solution for α if and only if $\beta > 1$. Realizing that the average of t_i is $1/\beta$, this says the system will be stable if the average duration of t_i is less than one.

From the above considerations we calculate that

$$\sigma_q^2 = \frac{1 - \alpha^2}{\alpha^2 \beta^2}. \tag{68}$$

VIII. EFFICIENCY OF SWITCH

We propose here to give an idea of how much is gained by asynchronous (but stabilized) versus synchronous sampling. We have shown in the previous section that (introducing unnormalized quantities now) for a speech processing duration distributed exponentially, i.e.,

$$u(t) = \beta e^{-\beta t}, \tag{69}$$

we have

$$\sigma_s^2 = \frac{1 - \alpha^2}{\alpha^2 \beta^2} \tag{70}$$

where α and β are related by

$$1 - \alpha = \exp\left[-\frac{\alpha \beta T}{L} \right], \tag{71}$$

which requires

$$\beta > \frac{L}{T}. \tag{72}$$

Again, L is the number of talkers. To effect the comparison with the synchronous version of the above scheme a new parameter has to be introduced which represents the peak-to-average *voltage* (not power) ratio of the signal. Recall $1/\beta$ is the average processing time, and therefore represents the average voltage. In the synchronous case a maximum time, $t_{\max}$, is allowed for each talker. Clearly $t_{\max}$ is a representative of the peak voltage. The ratio A is then

$$A \equiv \beta t_{\max} . \tag{73}$$

If L_s is the number of synchronous talkers, we also have

$$L_s t_{\max} = T. \tag{74}$$

Now use equation (70) to solve for α, substitute the expression for α into equation (71), eliminate β via equations (73) and (74) and obtain

$$1 - \frac{1}{\sqrt{1 + Z^2}} = \exp\left[\frac{-A\left(\frac{L_S}{L}\right)}{\sqrt{1 + Z^2}}\right] \tag{75}$$

where

$$Z^2 = A^2 L_S^2 \left(\frac{\sigma_t}{T}\right)^2. \tag{76}$$

An approximate solution to equation (75), which is accurate for large L_S, is that

$$\frac{L}{L_S} \approx A, \tag{77}$$

i.e., in the large L_S limit, the ratio of possible asynchronous talkers to synchronous ones is approximately the peak-to-average voltage ratio of the speech signal. One would expect this number to be at least four. Actually for L_S of interest (77) is not an accurate enough approximation. The solution of equations (75) and (76) is shown if Figs. 5 and 6 for $A = 4$ and 8 respectively. Taking $A = 4$ (which is conservative), we see from Fig. 5 that if the technology would permit 50 talkers to be put on the switch synchronously, the switch could accommodate 150 simultaneous speakers, sampled on a stabilized asynchronous manner, with an output S/N of 35 dB.* To repeat, factor of 3 increase in efficiency seems like a conservative estimate.

IX. FOUR-WIRE CONSIDERATIONS

The last topic we consider is what we call a four-wire treatment of the problem. Here we treat the details of a model proposed by F. K. Becker[5] which should deal with active asynchronous energy transfer when four-wire facilities are available. The model is this. A speech waveform is sampled and held for a variable time Δ_i before the next sample is taken (see Fig. 7). The holding times Δ_i are independent, identically distributed, and have average value

$$\bar{\Delta}_i = T. \tag{78}$$

The initial speech waveform is approximately reconstructed by passing the jittered box car through a filter having an inband characteristic

$$h(\omega) = \frac{\frac{\omega T}{2}}{\sin \frac{\omega T}{2}} \exp\left(\frac{i\omega T}{2}\right). \tag{79}$$

* The S/N is read from Table II for $\sigma_t/T = 0.01$.

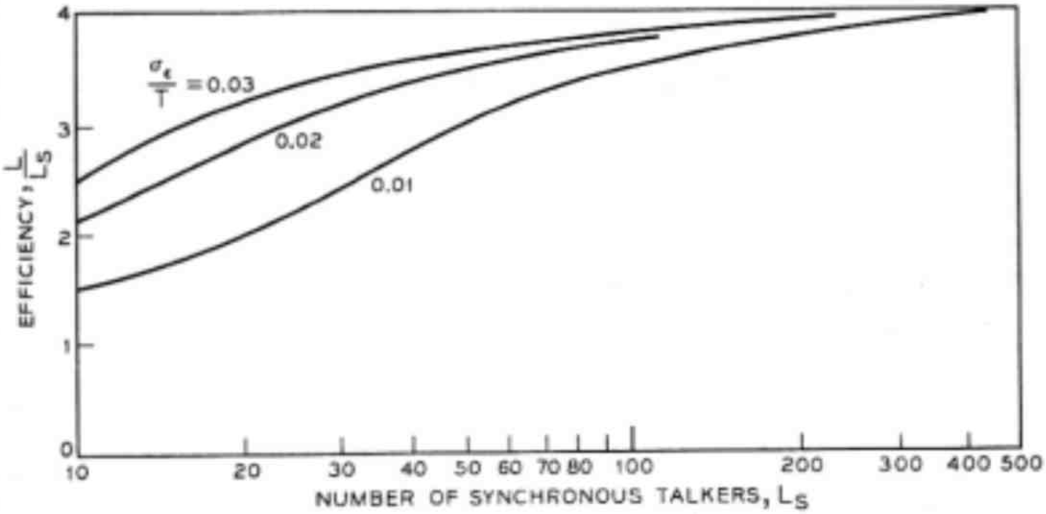


Fig. 5—Efficiency versus number of synchronous talkers for peak-to-average voltage ratio $A = 4$.

The Δ_i in the above model will not necessarily be assumed to be sharply distributed about T; later a stabilized version using an ϵ-jitter model will be used.

Let $x(t)$ denote the speech wave and $x_J(t)$ denote the jittered box car version. Further denote the spectra of the two processes by $S(\omega)$ and $S_J(\omega)$ respectively. Our major interest shall focus on determining the spectrum $E(\omega)$ of the error

$$\text{error} = h * x_j(t) - x(t), \tag{80}$$

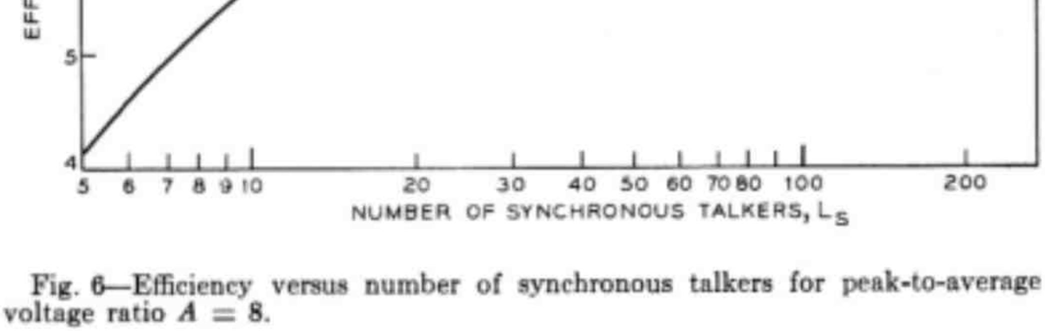


Fig. 6—Efficiency versus number of synchronous talkers for peak-to-average voltage ratio $A = 8$.

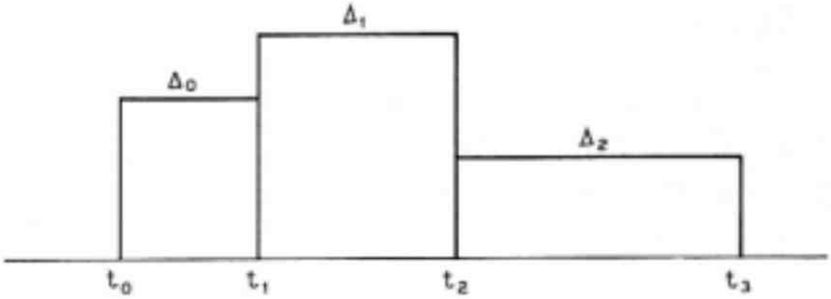


Fig. 7—"Jittered-boxcar" sampled speech wave.

assuming $h(t)$ is the reconstructing filter. The output noise power can then be obtained by integrating the spectrum. It may be shown that for the Δ-model of jitter described above the *exact* error spectrum is

$$E(\omega) = S(\omega)\left|1 - \frac{i}{\omega T}h^*(\omega)[1 - C(\omega)]\right|^2 + \frac{2\,|h(\omega)|^2}{T\omega^2}\,\mathrm{Re}\left\{[1 - C^*(\omega)]P\int_{-\infty}^{\infty} S(\omega')\,\frac{1 - C(\omega')}{1 - C(\omega' - \omega)}\,\frac{d\omega'}{2\pi}\right\}. \quad (81)$$

In equation (81), the function $C(\omega)$ is the characteristic function of the variable Δ, that is,

$$C(\omega) = \int_{-\infty}^{\infty} \exp\,(i\omega\Delta)p(\Delta)\,d\Delta, \quad (82)$$

$p(\Delta)$ being the probability density of Δ. Also

$$h(\omega) = \int_{-\infty}^{\infty} \exp\,(-i\omega t)h(t)\,dt.$$

The symbol P in front of the integral in equation (81) denotes that the principal value is to be taken for the simple pole $1/[1 - C(\omega' - \omega)]$. If the term in the braces in equation (81) is rewritten as

$$\{\ \} = \int_{-\infty}^{\infty} S(\omega')\,\mathrm{Re}\left[\frac{(1 - C^*(\omega))(1 - C(\omega'))}{1 - C(\omega' - \omega)}\right]\frac{d\omega'}{2\pi}, \quad (83)$$

the reader should be assured that now no singularity will arise in the integrand, and the principal value distinction need not be made.

For numerical purposes, we plot the error spectrum given in equations (81) and (83) for the case when Δ is gaussian distributed about mean value T, and for standard deviations $(\sigma_\Delta/T) = 0.316$, 0.1.*

* The random variable Δ is always positive, while the gaussian assumption allows it to become negative with some probability. We have verified that this effect is not significant here.

These larger variances are chosen here because we are mainly interested in the unstabilized version. Figures 8 and 9 show, respectively, these error spectra, assuming the input spectrum is flat up to a maximum frequency $\Omega = \pi$ rad/s, and assuming a sampling interval $T = 0.5$ s. This corresponds to sampling at twice the Nyquist rate. Also in calculating the out-of-band noise we have taken the reconstructing filter (79) to extend to *twice* the baseband spectrum. This emphasizes the higher frequencies more than desirable, perhaps. Thus in Fig. 9, the S/N for noise measured in twice the baseband interval is 24 dB while inband we have a S/N of 29 dB. Thus for independent Δ_i a 10 percent jitter about the mean value seems tolerable for sampling at about twice the Nyquist rate.*

In the event that the unstabilized four-wire version is unsatisfactory, we recalculate the error spectrum when timing is stabilized according to ϵ-model discussed earlier. In this case the Fourier transform $e_L(\omega)$ of the truncated error signal is found to be given by

$$\begin{aligned} e_L(\omega) = \frac{i}{\omega} \sum_{n=0}^{L} x(nT + \epsilon_n) \exp\left(-i\omega nT - i\omega \frac{T}{2}\right) \\ \times \left[\exp\left(-\frac{i\omega T}{2} - i\omega\epsilon_{n+1}\right) - \exp\left(\frac{i\omega T}{2} - i\omega\epsilon_n\right)\right] \\ \times \left[\frac{\omega T}{2} \frac{1}{\sin\dfrac{\omega T}{2}} \exp\left(\frac{i\omega T}{2}\right)\right] - \begin{pmatrix}\text{same terms}\\ \text{with } \epsilon = 0\end{pmatrix}. \qquad (84) \end{aligned}$$

Again, the reconstructing filter (79) has been assumed. Further details of the calculation will not be recorded, but we simply state that the error spectrum $E(\omega)$ is, retaining only second-order terms as in previous work involving ϵ-jitter,

$$\begin{aligned} E(\omega) &= \lim_{L\to\infty} \frac{E\,|e^L(\omega)|^2}{L}, \\ &= \frac{T}{\sin^2\dfrac{\omega T}{2}} \sum_{s=-\infty}^{\infty} \exp(-i\omega sT)\left\{\frac{\omega^2}{2} J(s)\left[R(s) - \frac{R(s+1)+R(s-1)}{2}\right]\right. \\ &\quad - \sin^2\frac{\omega T}{2} \bar{R}(s)J(s) \end{aligned}$$

[*] Let us remark that the Δ model for jitter described here is significantly different from the ϵ-model used previously. The Δ model implies positive correlation between adjacent ϵ variables and for a given variance produces less noise. We shall return to these questions shortly.

$$+ \frac{\omega}{2} \sin \frac{\omega T}{2} J(s) \Big[(-\dot{R}(s) + \ddot{R}(s+1)) \exp\left(-\frac{i\omega T}{2}\right)$$

$$- (-\dot{R}(s) + \dot{R}(s-1)) \exp\left(\frac{i\omega T}{2}\right) \Big] \Big\}. \tag{85}$$

One may verify that the above expression is indeed real, that is, it vanishes when $\epsilon_i = \epsilon$ and when the correlation function corresponds to a dc input signal. The noise power N in any bandwidth Ω is gotten from equation (85) by

$$N = \frac{1}{2\pi} \int_{-\Omega}^{\Omega} E(\omega)\, d\omega. \tag{86}$$

Choosing $\Omega = \pi/T$ we obtain from equation (86)

$$N = -\ddot{R}(0)J(0) + \frac{1}{T^2} \sum_{s=-\infty}^{\infty} \alpha_s J(s) \left[R(s) - \frac{R(s+1) + R(s-1)}{2} \right]$$

$$+ \frac{1}{T} \sum_{s=-\infty}^{\infty} \beta_s J(s) \frac{\dot{R}(s+1) - \dot{R}(s-1)}{2}$$

$$- \frac{1}{T} \sum_{s=-\infty}^{\infty} \delta_s J(s) \left[-\dot{R}(s) + \frac{\dot{R}(s+1) + \dot{R}(s-1)}{2} \right], \tag{87}$$

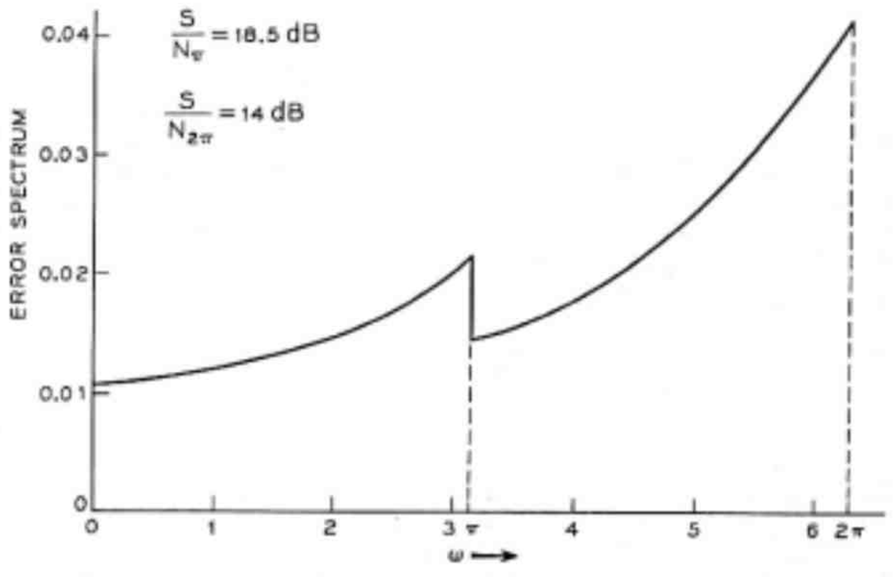


Fig. 8—Error spectrum for $T = 1/2$, $\sigma_\Delta/T = 0.316$. Undistorted spectrum is one for $-\pi \leqq \omega \leqq \pi$, and zero otherwise. $S/N_\pi = 18.5$ dB, $S/N_{2\pi} = 14$ dB.

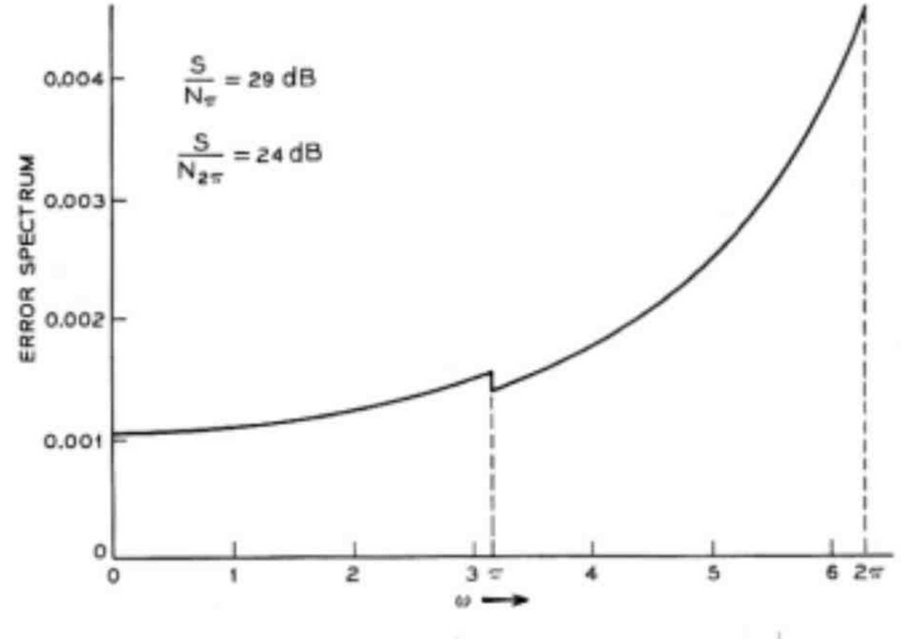


Fig. 9—Error spectrum for $T = 1/2$, $\sigma_\Delta/T = 0.1$. Undistorted spectrum is one for $-\pi \leqq \omega \leqq \pi$, and zero otherwise. S/N$_\pi$ = 29 dB, S/N$_{2\pi}$ = 24 dB.

where

$$\alpha_s = \frac{1}{2\pi}\int_0^\pi \frac{\cos sx}{\sin^2 \frac{x}{2}}\, x^2\, dx, \tag{88a}$$

$$\beta_s = \frac{1}{\pi}\int_0^\pi x \cos sx \frac{\cos \frac{x}{2}}{\sin \frac{x}{2}}\, dx, \tag{88b}$$

$$\delta_s = \frac{1}{\pi}\int_0^\pi x \sin sx\, dx = \frac{(-1)^{s+1}}{s}, \qquad s \neq 0. \tag{88c}$$

Numerically

$$\begin{aligned} \alpha_0 &= 2\beta_0 \approx 2.78, \\ \alpha_1 &= -0.469, \\ \beta_0 &= 1.39, \\ \beta_1 &= 0.386. \end{aligned} \tag{89}$$

For the case of independent ϵ-jitter, equation (87) reduces to

$$\begin{matrix}\text{noise power;}\\ \text{independent } \epsilon\end{matrix} = R(0)\sigma_\epsilon^2\left[-\frac{\ddot{R}(0)}{R(0)} + \frac{1}{T^2}\,\alpha_0\left(1 - \frac{R(1)}{R(0)}\right) + \frac{1}{T}\,\beta_0\,\frac{\dot{R}(1)}{R(0)}\right]. \tag{90}$$

The results for the stabilized full spectrum and half spectrum cases are obtained by using equations (35) and (41) in equation (90).

Full Spectrum:

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right) = \frac{\sigma_\epsilon^2}{T^2}\left[\frac{\pi^2}{3} + \beta_0\right] \approx \frac{\sigma_\epsilon^2}{T^2}\,(4.67). \tag{91}$$

Half Spectrum:

$$\left(\frac{\mathrm{S}}{\mathrm{N}}\right) = \frac{\sigma_\epsilon^2}{T^2}\left[\frac{\pi^2}{12} + \alpha_0\left(1 - \frac{2}{\pi}\right) - \beta_0\,\frac{2}{\pi}\right] \approx \frac{\sigma_\epsilon^2}{T^2}\,(0.946). \tag{92}$$

By comparison with equation (38), the four-wire full spectrum answer (91) is about 1.5 dB worse than the optimum two-wire result, but the four-wire result has about a 4.5-dB advantage for the half-spectrum case [compare equation (92) with equations (42) and (44)]. The latter case corresponds more to the case of practical interest. We note further that in this comparison the four-wire might be penalized by two dB or so due to the $(x/\sin x)$ characteristic being used out of band also. Secondly, the more concentrated toward dc the signal spectrum is, the better the four-wire version will become since it has vanishing distortion for dc, while the optimum two-wire version does not.

Before leaving this topic, a comment on the relation between Δ-jitter and ϵ-jitter should be made. We begin by looking at two successive Δ-variables Δ_n and Δ_{n+1} in terms of ϵ-variables.

$$\begin{aligned} \Delta_n &= T + \epsilon_n - \epsilon_{n-1}, \\ \Delta_{n+1} &= T + \epsilon_{n+1} - \epsilon_n. \end{aligned} \tag{93}$$

Clearly

$$E(\Delta) = T, \tag{94a}$$

$$\sigma_\Delta^2 = 2\overline{\epsilon^2}(1 - \rho), \tag{94b}$$

$$E[(\Delta_{n+1} - T)(\Delta_n - T)] = 2\overline{\epsilon^2}\rho - \bar{\epsilon}^2 - \overline{\epsilon^2}. \tag{94c}$$

In writing equation (94 a through c), we define

$$E[\epsilon_n\epsilon_{n+1}] = \overline{\epsilon^2}\rho \tag{95}$$

and assume no ϵ-correlation after one displacement, i.e.,

$$E(\epsilon_0\epsilon_2) = \bar{\epsilon}^2.$$

To make two successive Δ variables uncorrelated, we set equation

(94c) equal to zero and obtain

$$\rho = \frac{1}{2} + \frac{1}{2}\frac{\bar{\epsilon}^2}{\overline{\epsilon^2}} \tag{96}$$

and

$$\sigma_\Delta^2 = \sigma_\epsilon^2 . \tag{97}$$

Thus to compare independent Δ-jitter with ϵ-jitter, a good way to do it is to choose the same variance according to equation (97) and introduce a correlation between adjacent ϵ's by equation (96). This simple way of comparing the two schemes is not exact, but should be good if signal correlations do not persist for extended periods. In any event, it illustrates why for $\sigma_\Delta^2 = \sigma_\epsilon^2$ the independent Δ-model will yield appreciably better results than the independent ϵ model. The first implies positive correlation in the second.

X. CONCLUSIONS

We began our study with consideration of the two-wire switching problem (Fig. 1). After obtaining a concise mathematical description of the operation of this switch, [equation (9)], attention focused on the optimum capacitor discharge $z(t)$ when one volt is placed on the capacitor (Fig. 2). We have seen that a good design for the slopes $\dot{z}_k$ of the function $z(t)$ as it passes through its zeros (see Fig. 2) are the values $\dot{z}_k = (-1)^k/(2kT)$. Exact solutions for full spectrum and for dc indicate that this result is not sensitive to the signal spectrum. This optimum design can be expected to yield only a 3 dB improvement over the "blind" filter which is defined to be flat at the zero crossings and consequently does not "see" the jitter. Typically two percent jitter yields an output S/N of 30 dB. Any positive correlation in the jitter will tend to improve this figure.

When the processing time of an individual speaker is exponentially distributed, an exact distribution of the timing jitter with a stabilized clock is obtained. This result is used to study the efficiency of the asynchronous switch. This increase in capacity over a switch using the same technology but employing a synchronous strategy is, roughly, the peak-to-average voltage ratio of the signal. More accurate descriptions are presented in Figs. 5 and 6. Conservatively, the asynchronous switch should handle three times the traffic.

Finally, a four-wire version modeled as a jittered boxcar reconstructed with a filter has been considered. Under the independent

Δ-model for jitter, Fig. 7, 10 percent jitter can yield 30 dB S/N, largely because of the correlations this implies for ϵ-jitter. Exact error spectra have been plotted for this case also (Figs. 8 and 9). With timing stabilization, and consequently the ϵ-jitter model, the reconstructed boxcar should yield several dB improvement over the two-wire results for cases of practical interest (spectrum concentrated at dc).

XI. ACKNOWLEDGMENT

We are happy to acknowledge several discussions with J. F. O'Neill and F. K. Becker during which several of the problems dealt with here were outlined.

APPENDIX A

Some Lemmas on Limits of Sums

We list here certain lemmas which will be needed in Appendix B.

Lemma 1: If either $\sum_{s=0}^{\infty} | f(s) |$ *or* $\sum_{s=0}^{\infty} sf(s)$ *converges,* then*

$$\lim_{K\to\infty} \frac{1}{K} \sum_{\substack{n,k=0\\ n\geq k}}^{K} f(n-k) = \sum_{s=0}^{\infty} f(s). \tag{98}$$

Lemma 2: If $\sum_{s=0}^{\infty} f(s)$ *converges, then*

$$\lim_{K\to\infty} \frac{1}{K} \sum_{n=1}^{K} \sum_{s=1}^{n} f(s) = \sum_{s=1}^{\infty} f(s). \tag{99}$$

Likewise

$$\lim_{K\to\infty} \frac{1}{K} \sum_{n=1}^{K} \sum_{s,t=1}^{n} f(s, t) = \sum_{s,t=1}^{\infty} f(s, t). \tag{100}$$

Lemma 3: If $\sum_{\substack{s=1\\ j=-\infty}}^{\infty} f(j, s)$ *converges, then*

$$\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \sum_{s=1}^{k} \sum_{j=-k}^{K-k} f(j, s) = \sum_{\substack{s=1\\ j=-\infty}}^{\infty} f(j, s). \tag{101}$$

APPENDIX B

Derivation of Output Noise Power

Our first step is to derive the filter independent noise, i.e.

* Neither condition implies the other; each implies the convergence of $\Sigma f(s)$.

$$A = \lim_{K\to\infty} A_K$$

where A_K is given by equation (23). Our expressions are valid to second order in the $\{\epsilon_l\}$. We note the following relations

$$y_n = x(nT + \epsilon_n) \approx x(nT) + \epsilon_n\dot{x}(nT) + \tfrac{1}{2}\epsilon_n^2\ddot{x}(nT); \qquad (102)$$

$$\begin{aligned} \psi_m(\epsilon_l) &\approx 1 + \ddot{\psi}_0\frac{\epsilon_l^2}{2}, \qquad \text{if} \quad m = 0; \\ &\approx \psi_m(0)\epsilon_l + \tfrac{1}{2}\ddot{\psi}_m(0)\epsilon_l^2/2, \qquad \text{if} \quad m \neq 0; \qquad (103) \\ \psi_k(0) &= -\psi_{-k}(0); \end{aligned}$$

making use of this in equation (23), we have, to second order in $\{\epsilon_l\}$

$$\begin{aligned} A_K = \frac{1}{K}\sum_{n=0}^{K}(\epsilon_n^2\dot{x}_n^2 - x_n^2\ddot{\psi}_0(0)\epsilon_n^2) \\ - \frac{1}{K}\sum_{\substack{n,k=0\\n\neq k}}^{\infty}[x_nx_k\epsilon_n\epsilon_k - 2x_k\dot{x}_n\epsilon_n\epsilon_k\psi_{n-k}(0)]. \end{aligned} \qquad (104)$$

Averaging over the signal statistics by using equation (26) and the further relations

$$\begin{aligned} E[\dot{x}(t+\tau)x(t)] &= \dot{R}(\tau) = -E[\dot{x}(t)x(t+\tau)], \\ E[\dot{x}(t+\tau)\dot{x}(t)] &= -\ddot{R}(\tau), \end{aligned} \qquad (105)$$

and also averaging over the jitter variables using equation (27) gives:

$$\begin{aligned} EA_K = \frac{1}{K}\sum_{n=0}^{K}[-\ddot{R}(0)J(0) - \ddot{\psi}_0(0)J(0)R(0)] \\ - \frac{1}{K}\sum_{\substack{n,k=0\\n\neq k}}^{K}[R(n-k)J(n-k)\ddot{\psi}_{n-k}(0) - 2\Gamma_{n-k}J(n-k)\dot{R}(n-k)]. \end{aligned} \qquad (106)$$

Observing that the second sum is a symmetric function of $(n - k)$ and using Lemma 1 of Appendix A, the required limiting operation yields

$$A = -\ddot{R}(0)J(0) - \sum_{s=-\infty}^{\infty} R(s)J(s)\ddot{\psi}_s(0) + 2\sum_{\substack{s=-\infty\\s\neq 0}}^{\infty}\Gamma_sJ(s)\dot{R}(s). \qquad (107)$$

We list for further possible use

$$\begin{aligned} \ddot{\psi}_s(0) &= 2\,\frac{(-1)^{s+1}}{s^2T^2}\,, \qquad s \neq 0; \\ \ddot{\psi}_0(0) &= -\frac{1}{3}\,\frac{\pi^2}{T^2}\,; \end{aligned} \tag{108}$$

and also

$$\sum_{s=-\infty}^{\infty} \ddot{\psi}_s(0) = 0. \tag{109}$$

The latter relation may be derived by direct evaluation of the sum, of course, or by using the Poisson sum formula and the bandlimited nature of $\psi(t)$.

We now proceed to derive an expression for the filter dependent terms B, starting from equation (24). Gathering the second-order terms proceeds much as it did for the A_k terms, with the following additional complication. Recall from equation (20) that θ is given by

$$\theta = \frac{2Z}{1 + 2Z} = 2Z - 4Z^2 + 8Z^3 - \cdots$$

From equation (6),

$$\begin{aligned} Z_{kn} &= (\epsilon_k - \epsilon_n)\dot{z}_{k-n} + \begin{pmatrix}\text{higher}\\ \text{terms}\end{pmatrix}, &\quad \text{if} \quad k > n; \\ &= 0, &\quad \text{if} \quad k \leq n. \end{aligned}$$

Thus the matrix θ is at least linear in the $\{\epsilon_i\}$. Proceeding now to collect terms and averaging, we have to second order

$$\begin{aligned} E[B_K] = {} & \frac{4}{K} \sum_{n=1}^{K} \sum_{m,l=0}^{n-1} \dot{z}_{n-m}\dot{z}_{n-l}R(l - m) \\ & \cdot[J(0) + J(m - l) - J(n - m) - J(n - l)] \\ & - \frac{4}{K} \sum_{n=1}^{K} \sum_{m=0}^{n-1} \dot{z}_{n-m}\dot{R}(n - m)[J(0) - J(n - m)] \\ & + \frac{4}{K} \sum_{\substack{n,k=1 \\ n\neq k}}^{K} \sum_{m=0}^{k-1} \Gamma_{n-k}\dot{z}_{k-m}R(n - m)[J(n - k) - J(n - m)]. \end{aligned} \tag{110}$$

In the second term of the above, we change summation variables by replacing the sum over m with a sum over s, where

$$s = n - m, \qquad s = 1, 2, \cdots, n.$$

We make the same change in the first term, and also introduce

$$t = n - l, \qquad t = 1, 2, \cdots, n.$$

In the last term let

$$\begin{aligned} s &= k - m, \qquad s = 1, 2, \cdots, k, \\ j &= n - k, \qquad j = -(k-1), \cdots, K - k, \\ &\qquad\qquad j \neq 0. \end{aligned}$$

Then

$$\begin{aligned} E[B_K] = \frac{4}{K} \sum_{n=1}^{K} \sum_{s,t=1}^{n} \dot{z}_s \dot{z}_t R(s-t)[J(0) + J(s-t) - J(s) - J(t)] \\ - \frac{4}{K} \sum_{n=1}^{K} \sum_{s=1}^{n} \dot{z}_s \dot{R}(s)[J(0) - J(s)] \\ + \frac{4}{K} \sum_{k=1}^{K} \sum_{s=1}^{k} \sum_{\substack{j=-k+1 \\ j \neq 0}}^{K-k} \Gamma_j \dot{z}_s R(j+s)[J(j) - J(j+s)]. \end{aligned} \tag{111}$$

Making use of Lemmas 2 and 3 of Appendix A yields

$$\begin{aligned} B = \lim_{K \to \infty} E[B_K] = -4 \sum_{s=1}^{\infty} \dot{z}_s R(s)[J(0) - J(s)] \\ + 4 \sum_{s,t=1}^{\infty} \dot{z}_s \dot{z}_t R(s-t)[J(0) + J(s-t) - J(s) - J(t)] \\ + 4 \sum_{s=1}^{\infty} \sum_{\substack{j=-\infty \\ j \neq 0}}^{\infty} \Gamma_j \dot{z}_s R(j+s)[J(j) - J(j+s)]. \end{aligned} \tag{112}$$

The simpler equations (28) and (32) which hold for independent jitter may be gotten from the above by using the result of Section VI. Namely for $J(s) = \text{const.}$, all s, the A and B terms vanish. When we have independent jitter $J(s) = \text{const.}$, $s \neq 0$. So we add and subtract a term with $J(0)$ replaced by $J(s \neq 0)$ and use equation (29) to obtain the independent jitter results.

APPENDIX C

Exact Error Spectrum for Independent Δ-Jitter

Consider a long time L, and denote by $[L]$ the best integer so that

$$L \cong [L]T. \tag{113}$$

The jittered box car (truncated to the long interval L) can then be represented as

$$x_J(t) = \sum_{n=0}^{[L]} x(t_n)B[t_{n+1},\, t_n], \tag{114}$$

where B is a unit box extending from t_n to t_{n+1}. The Fourier transform of this truncated version is then

$$s_L(\omega) = \frac{i}{\omega} \sum_{n=0}^{[L]} x(t_n) \exp{(-i\omega t_n)}[\exp{(-i\omega\, \Delta_n)} - 1]. \tag{115}$$

The power spectra for a process $y(t)$ is generally calculated according to

$$G(\omega) = \lim_{L\to\infty} \frac{1}{L} E\, |y^L(\omega)|^2,$$

where $y^L(\omega)$ is the Fourier transform of $y(t)$ truncated to an interval L, and E denotes an ensemble average with respect to all random parameters. If we apply the above to spectrum of the error signal (80), we obtain immediately, using notation of the text,

$$E(\omega) = |\, h(\omega)\, |^2\, S_J(\omega) + S(\omega) + D(\omega) \tag{116}$$

where

$$D(\omega) = -\lim_{[L]\to\infty} \frac{1}{[L]T} \cdot 2E \text{ Re } h^*(\omega) x_J^{L*}(\omega) x^L(\omega). \tag{117}$$

Expressing $x^L(\omega)$ as a time integral and performing the signal averaging we get

$$D(\omega) = -\lim \frac{2}{[L]T} \text{ Re } h^*(\omega) \left\{ \frac{-i}{\omega} \sum_{n=0}^{[L]} \exp{(i\omega t_n)}[\exp{(i\omega\, \Delta_n)} - 1] \right. \tag{118}$$
$$\left. \times \int_{-\infty}^{\infty} \exp{(-i\omega t')} R(t' - t_n)\, dt \right\}.$$

Expressing the signal correlation function in terms of the spectrum, doing the integrals and remaining averages, gives

$$D(\omega) = \frac{-2}{T} \text{ Re } h^*(\omega) \left(\frac{i}{\omega}\right) S(\omega)[1 - C(\omega)]. \tag{119}$$

There remains the calculation of $S_J(\omega)$. Using equation (115), we have (performing averages over signals)

$$E\,|s^L(\omega)|^2 = \frac{1}{\omega^2} E \sum_{n,m=0}^{[L]} R\left[\sum_{k=0}^{n-1} \Delta_k - \sum_{k=0}^{m-1} \Delta_k\right]$$
$$\times\,[\exp(-i\omega\,\Delta_n) - 1](\exp(i\omega\,\Delta_m) - 1]\left[\exp - i\omega\left(\sum_{k=0}^{n-1} \Delta_k - \sum_{k=0}^{m-1} \Delta_k\right)\right] \tag{120}$$

where we note the fact

$$t_{n+1} = \sum_{k=0}^{n} \Delta_n. \tag{121}$$

Performing the Δ-averaging yields

$$\frac{E\,|s^L(\omega)|^2}{[L]T} = \frac{R(0)}{[L]T\omega^2}\, 2[1 - \operatorname{Re} C(\omega)] \sum_{n=0}^{[L]} 1$$
$$+ \frac{2}{[L]T\omega^2} \operatorname{Re}\left\{[1 - C^*(\omega)] \sum_{n=m+1}^{[L]} \sum_{m=0}^{[L]}\right.$$
$$\left.\cdot \int_{-\infty}^{\infty} \frac{d\omega'}{2\pi}\, S(\omega')C(\omega' - \omega)\cdot[C(\omega' - \omega) - C(\omega')]\right\}\cdot \tag{122}$$

To perform the limits of the sums we first replace $C^{n-m-1}(\omega' - \omega)$ by $[\exp(-\epsilon)C^{-n-m-1}(\omega' - \omega)]$ to insure convergence and then let $\epsilon \rightarrow 0$. We then have

$$S_J(\omega) = \lim_{[L]\to\infty} \frac{1}{[L]T} E\,|s^L(\omega)|^2$$
$$= \frac{2}{\omega^2 T} \operatorname{Re}\left\{[1 - C^*(\omega)] \int_{-\infty}^{\infty} \frac{d\omega}{2\pi}\, S(\omega')\, \frac{1 - C(\omega')}{1 - C(\omega' - \omega)e^{-\epsilon}}\right\}\cdot \tag{123}$$

Making use of the general relation

$$\frac{1}{x \pm i\epsilon} = P\,\frac{1}{x} \mp i\pi\,\delta(x),$$

where P denotes principal value, finally yields

$$S_J(\omega) = \frac{|1 - C(\omega)|^2}{\omega^2}\,\frac{S(\omega)}{T^2}$$
$$+ \frac{2}{T\omega^2} \operatorname{Re}\left\{[1 - C^*(\omega)]P \int_{-\infty}^{\infty} \frac{d\omega'}{2\pi}\, S(\omega')\, \frac{1 - C(\omega)}{1 - C(\omega' - \omega)}\right\}\cdot \tag{124}$$

The interested reader may indeed check that for

$$C(\omega) = \exp(i\omega T),$$

the above expression reproduces the known result for a wave-form impulse modulated at precisely T-second intervals.

REFERENCES

1. Dimmick, J. O., Lewis, T. G., and O'Neill, J. F., "A Time Division Switching Network Using Asynchronous Active Energy Transfer," 1970 IEEE Int. Conf. Commun. Record, p. 43–1.
2. Lewis, T. G., "Asynchronous Time Division Switching Network—Stabilization of Sampling Rate Variation," unpublished work.
3. Saltzberg, B. R., and Pasternak, G. P., unpublished work.
4. Salz, J., and Gitlin, R. D., unpublished work.
5. Becker, F. K., private communication.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

Error Probability of a Multilevel Digital System With Intersymbol Interference and Gaussian Noise

By E. Y. HO and Y. S. YEH

(Manuscript received October 16, 1970)

In a previous paper a series expansion method for calculating the error probability of a binary digital AM system in the presence of intersymbol interference and additive gaussian noise was derived.[1] *In this paper those results are extended to the multilevel case. In the examples calculated for a four-level system, this method is 10^4 times faster than the exhaustive method and is 10^2 times more accurate than the Chernoff bound. The actual computation time with an 11-sample approximation to the real system impulse response is only 1.3 seconds with the GE Mark II time-sharing system.*

I. INTRODUCTION

In a recent paper[1] we have developed a new method to calculate the error probability of a binary digital data system in the presence of intersymbol interference and additive gaussian noise. A similar method has also been reported by M. I. Celebiler and O. Shimbo.[2] The purpose of this paper is to extend the previous results to multilevel systems.

The existing methods for the estimation of the error probability are the Chernoff bound or the worst case bound,[3,4] the results of which are generally too loose. Another alternative is the time-consuming exhaustive method.[3] For example, it would require $4^{10} (\approx 10^6)$ calculations of the error function to find the error probability of a four-level digital system where intersymbol interference resulted from ten nonzero samples of the channel impulse response.

II. DERIVATION OF THE EXPRESSION FOR ERROR PROBABILITY

For a $2m$-level digital AM system, the corrupted received sequence at the input to the receiver detector is

$$y(t) = \sum_{l=-\infty}^{\infty} a_l r(t - lT) + n(t), \tag{1}$$

where

$n(t)$ is additive gaussian noise,

$a_l = \pm 1, \pm 3, \cdots, \pm(2m - 1)$ with equal probability,

and

$r(t)$ is the given noiseless system impulse response.

At the detector, $y(t)$ is sampled every T seconds to determine the amplitude of the transmitted signal. At sampling time t_0, the sampled signal is,

$$y(t_0) = a_0 r(t_0) + \sum_{\substack{l=-\infty \\ l \neq 0}}^{\infty} a_l r(t_0 - lT) + n(t_0). \tag{2}$$

The first term is the desired signal while the second and the third terms represent the intersymbol interference and gaussian noise respectively.

The set of slicing levels are,[5]

$$0, \pm 2r(t_0), \pm 4r(t_0), \cdots, \pm(2m - 2)r(t_0). \tag{3}$$

Given a particular transmitted signal level, a_0, the conditional error probability is,

$$P_r(e/a_0) = \begin{cases} P\{y(t_0) \geqq -2(m-1)r(t_0)\}, & a_0 = -(2m-1), \\ P\{y(t_0) \leqq 2(m-1)r(t_0)\}, & a_0 = 2m-1, \\ P\{(y(t_0) \geqq (a_0+1)r(t_0))U(y(t_0) \leqq (a_0-1)r(t_0))\}, & \\ & a_0 \neq \pm(2m-1), \end{cases} \tag{4}$$

where AUB is the union of the events A and B.

Substituting equation (2) into (4), we obtain

$$P_r(e/a_0) = \begin{cases} P\{\sum_{l \neq 0} a_l r(t_0 - lT) + n(t_0) \geqq r(t_0)\}, & a_0 = -(2m-1), \\ P\{\sum_{l \neq 0} a_l r(t_0 - lT) + n(t_0) \leqq -r(t_0)\}, & a_0 = 2m-1, \\ P\{(\sum_{l \neq 0} a_l r(t_0 - lT) + n(t_0) \geqq r(t_0))U, & \\ \quad (\sum_{l \neq 0} a_l r(t_0 - lT) + n(t_0) \leqq -r(t_0))\}, & a_0 \neq \pm(2m-1). \end{cases} \tag{5}$$

Since $\sum_{l \neq 0} a_l r(t_0 - lT)$ and $n(t_0)$ are equally likely to be positive or negative, equation (5) reduces to

$$P_r(e/a_0) = \begin{cases} P\{\sum_{l \neq 0} a_l r(t_0 - lT) + n(t_0) \geq r(t_0)\}, & a_0 = \pm(2m-1), \\ 2P\{\sum_{l \neq 0} a_l r(t_0 - lT) + n(t_0) \geqq r(t_0)\}, & a_0 \neq \pm(2m-1). \end{cases} \tag{6}$$

The error probability of the system is,

$$\begin{aligned} P_e &= \sum_{all\ a_0} P_r(e/a_0) P_r(a_0), \\ &= \frac{2m-1}{m} P\{\sum_{l \neq 0} a_l r(t_0 - lT) + n(t_0) \geqq r(t_0)\}. \end{aligned} \tag{7}$$

We note that equation (7) is similar to equation (7) of Ref. 1, with the only exception that the a_l can now assume multiple values. According to equation (9) of Ref. 1, we obtain the following expansion for P_e,

$$\begin{aligned} P_e = {} & \frac{2m-1}{2m} \operatorname{erfc}\left(-\frac{r(t_0)}{\sqrt{2}\,\sigma}\right) \\ & + \frac{2m-1}{m} \sum_{k=1}^{\infty} \frac{1}{(2k)!} \left(\frac{1}{2\sigma^2}\right)^k \frac{1}{\sqrt{\pi}} \exp\left[-\frac{r^2(t_0)}{2\sigma^2}\right] \\ & \cdot H_{2k-1}\left(\frac{r(t_0)}{\sqrt{2}\,\sigma}\right) M_{2k}, \end{aligned} \tag{8}$$

where

σ^2 is the noise power,
H_{2k-1} is the Hermite polynomial,
erfc is the complementary error function,
M_{2k} is the $2k$th moment of the random variable X,

and

$$X = \sum_{l \neq 0} a_l r(t_0 - lT), \qquad a_l = \pm 1, \pm 3, \cdots, \pm(2m-1). \tag{9}$$

The moments can again be obtained through the characteristic function of X without the explicit evaluation of the distribution function. The characteristic function is,

$$\begin{aligned} \Phi(\omega) &= \prod_{l \neq 0} \left\{ \sum_{k=-m+1}^{m} \exp\,[j\omega(2k-1) r(t_0 - lT)]/2m \right\}, \\ &= \prod_{l \neq 0} \left\{ \frac{1}{2m} \sin\,[2m\omega r(t_0 - lT)] \cdot \csc\,[\omega r(t_0 - lT)] \right\}. \end{aligned} \tag{10*}$$

* See Ref. 6, equation (1.342).

Therefore,

$$\Phi'(\omega) = \Phi(\omega)\{2m \sum_{l\neq 0} r(t_0 - lT) \cot [2m\omega(t_0 - lT)] - \sum_{l\neq 0} r(t_0 - lT) \cot [\omega r(t_0 - lT)]\}. \tag{11}$$

Since $M_{2k} = (-1)^k\Phi^{2k}(0)$ and $M_{2k-1} = 0$, we obtain a recurrence formula for M_{2k} by successive differentiation of equation (11),

$$M_{2k} = -\sum_{i=1}^{k} \binom{2k-1}{2i-1} M_{2(k-i)}(-1)^i \frac{2^{2i}[(2m)^{2i} - 1]}{2i} \mid B_{2i} \mid \cdot [\sum_{l\neq 0} r(t_0 - lT)^{2i}], \tag{12}$$

where B_{2i} is the Bernoulli number obtained by series expansion of cotangent function about the origin. Knowing that $M_0 = 1$, all the M_{2k}'s can be calculated successively through equation (12) for an N-sample approximation of the channel pulse response. The N-sample truncation is equivalent to the approximation of $\sum_{l\neq 0} r(t_0 - lT)^{2k}$ by $(N - 1)$ summation terms.

The error probability of a $2m$-level system can thus be obtained by equations (8) and (12). In the special case $m = 1$, equations (8) and (12) agree with the results of the binary system, i.e., equations (9) and (15) of Ref. 1.

III. TRUNCATION ERROR BOUND

The error incurred by truncating the series expansion of equation (8) at a finite term $n - 1$ is,

$$R_{2n} = \frac{2m-1}{m} \sum_{k=n}^{\infty} \frac{1}{(2k)!} \left[\frac{1}{2\sigma^2}\right]^k \frac{1}{\sqrt{\pi}} \exp\left[-\frac{r^2(t_0)}{2\sigma^2}\right] \cdot H_{2k-1}\left(\frac{r(t_0)}{\sqrt{2}\,\sigma}\right) M_{2k} . \tag{13}$$

Let

$$\lambda = \text{maximum } \{\mid \sum_{l\neq 0} a_l r(t_0 - lT) \mid\}, = (2m - 1) \sum_{l\neq 0} \mid r(t_0 - lT) \mid. \tag{14}$$

It can be shown that the moments satisfy the following inequality,

$$M_{2k+2p} \leqq M_{2k}\lambda^{2p}, \qquad p = 0, 1, 2, \cdots \tag{15}$$

For $(2k - 1) \gg x$, the Hermite polynomials are upper bounded by,

$$| H_{2k-1}(x) | \leqq 2^{k-\frac{1}{2}}[(2k - 3)!!]\sqrt{2k - 1} \exp [x^2/2]. \tag{16}$$

Substituting equations (16) and (15) into equation (13) and grouping the terms into geometric series, we obtain an upper bound for R_{2n}.

$$| R_{2n} | \leqq \epsilon_{2n} = \frac{2m - 1}{m} \frac{1}{\sqrt{2\pi}} (\exp [-r^2(t_0)/4\sigma^2]) \cdot \frac{M_{2n}}{(2\sigma^2)^n} \frac{1}{n!\sqrt{2n - 1}} \frac{1}{\left[1 - \dfrac{[(2m - 1) \sum_{l \neq 0} | r(t_0 - lT) |]^2}{2n\sigma^2}\right]}. \tag{17}$$

IV. EXAMPLE

We have calculated the error probability of a four-level digital AM system with the received pulse given by

$$r(t) = \frac{T}{\pi t} \sin (\pi t/T).$$

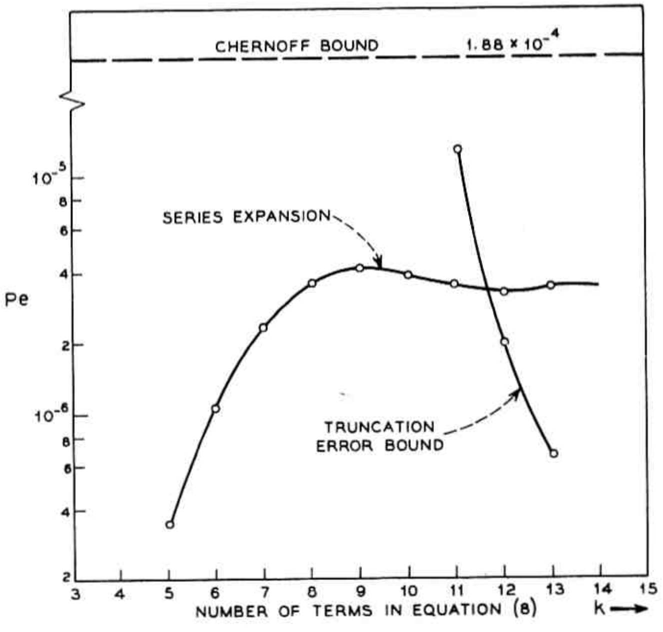


Fig. 1—Comparison of Chernoff Bound and series expansion method. Sin $[\pi t/T]/(\pi/T)t$ pulse, 11-pulse truncation approximation, $t_0 = 0.05T$, $(S/N) =$ 24 dB.

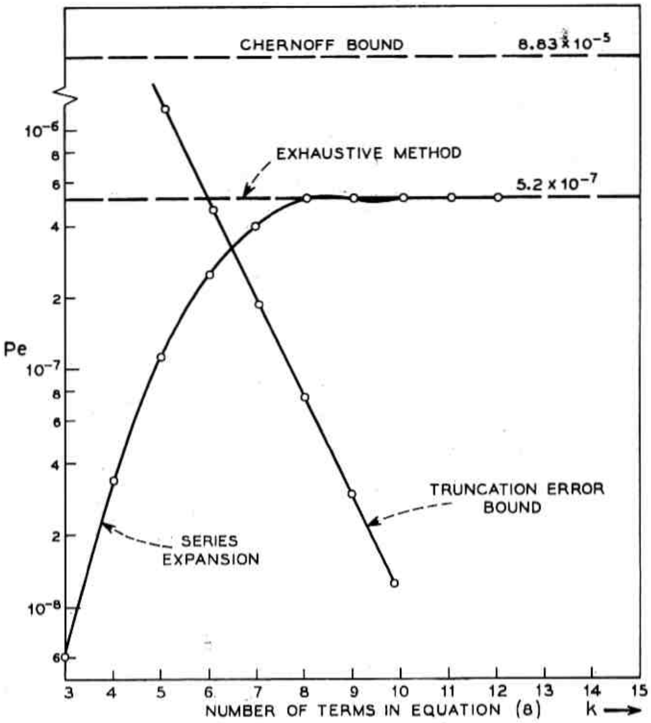


Fig. 2—Comparison of series expansion method with Chernoff Bound and exhaustive method. [SIN $\pi t/T]/(\pi/T)t$ pulse, 5-pulse truncation approximation, $t_o = 0.05T$, $(S/N) = 24$ dB.

With an 11-sample approximation, a S/R of 24 dB and a sampling time of $t_0 = 0.05\ T$, the error probability obtained by the Chernoff[7] * bound is 2×10^{-4}. The result obtained by our method is 3.4×10^{-6} indicating an improvement of two orders of magnitude. The convergence of equation (8) is presented in Fig. 1. Reasonable accuracy is achieved after eight terms of the series are calculated. A check of the accuracy of our method by comparison with the exhaustive method is impossible in this case because the latter requires 10^4 times more computation time as compared to the series expansion method. Instead, we checked our method with the exhaustive method for the case of a five-sample approximation. The results agree well and are presented in

* 11-sample approximation to the channel response is used in calculating the Chernoff bound.

Fig. 2. The computation time spent on GE Mark II time-sharing system is approximately 1.3 seconds for the series expansion method in both the eleven-sample and the five-sample approximations. The truncation error bounds are also presented in both figures.

V. CONCLUSIONS

We have extended the series expansion evaluation of the error probability of a binary digital AM system in the presence of intersymbol interference and additive gaussian noise to the $2m$-level systems. The results are extremely encouraging. For the case examined the series expansion method was calculated several orders of magnitude faster than the exhaustive method, and it was more accurate than the Chernoff bound by two orders of magnitude.

REFERENCES

1. Ho, E. Y., and Yeh, Y. S., "A New Approach for Evaluating the Error Probability in the Presence of Intersymbol Interferences and Additive Gaussian Noise," B.S.T.J., *49,* No. 9 (November 1970), pp. 2249–2266.
2. Celebiler, M. I., and Shimbo, O., "Intersymbol Interference Considerations in Digital Communications," ICC Record, *1* (June 1970), pp. 8.1–8.10.
3. Aein, J. M., and Hancock, J. C., "Reducing the Effects of Intersymbol Interferences with Correlation Receivers," IEEE Trans. Inform. Theory, *IT-9,* No. 3 (July 1963), pp. 167–175.
4. Aaron, M. R., and Tufts, D. W., "Intersymbol Interference and Error Probability," IEEE Trans. Inform. Theory, *IT-12,* No. 1 (January 1966), pp. 26–34.
5. Lucky, R. W., Salz, J., and Welson, E. J., Jr., *Principle of Data Communication,* New York: McGraw-Hill Book Company, 1968, p. 44.
6. Gradshteyn, I. S., and Ryzhik, I. M., *Table of Integrals and Series and Products,* New York and London: Academic Press, 1965, p. 30.
7. Saltzberg, B. R., "Intersymbol Interference Error Bounds with Application to Ideal Band Limited Sampling," IEEE Trans. Inform. Theory, *IT-14,* No. 4 (July 1968), pp. 563–568.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

Spectral Density of a Nonlinear Function of a Gaussian Process

By SCOTTY NEAL

(Manuscript received September 10, 1970)

We derive an expression which can be used in a relatively straightforward manner to obtain either the autocorrelation function or the spectral density of any "reasonable" function of any stationary gaussian process. The expression is used to study the spectral density of a sinusoidal wave which is phase modulated by a hard-limited gaussian process.

I. INTRODUCTION

A band-limited gaussian process having a rectangular power spectrum is assumed for some purposes to be an adequate approximation for several classes of modulating signals encountered in phase modulation (PM) systems.[1] However, in an actual implementation of a PM system, the modulating signal passes through circuitry which saturates when the signal rises above a fixed level. This clipping (i.e., limiting) level is usually adjusted fairly high (nominally at four times the rms of the modulating signal) and then ignored in any subsequent analysis of the system. Consequently, the objective of this study is to determine the qualitative effect of hard-limiting the modulating signal in a PM system.

From a mathematical viewpoint, we can obtain an understanding of the preceding question by investigating the following problem: Find the spectral density of a sinusoidal wave which is phase modulated by a function $g(X_t)$ of the stationary gaussian process X_t. Of course, this version of the problem can also be viewed as finding the spectral density of a (composite) nonlinear function of a gaussian process; a problem originally studied by S. O. Rice,[2] D. Middleton[3] and W. R. Bennett.[4] In fact, we do use their approach (representing the nonlinearity in terms of a transform) to derive an expression which is essentially the starting point of our analysis. However, using our relation avoids some of the complexity associated with the transform method. A derivation of the expression is given in Appendix A.

The notation and general results are presented in Section II along with two examples. In the third section, we obtain specific results for the hard-limiting case.

II. SPECTRAL DENSITY OF A PM WAVE MODULATED BY A NONLINEAR FUNCTION OF A GAUSSIAN PROCESS

Let $W(t)$ denote a constant-amplitude sinusoidal wave which is phase modulated by a real-valued function $g(X_t)$ of the stationary gaussian process X_t. That is,

$$W(t) = A \cos (\omega_c t + g(X_t) + \theta) \tag{1}$$

where A is the wave amplitude, $f_c = \omega_c/2\pi$ is the carrier frequency, and θ is a random variable with probability density function

$$\pi_\theta(\theta) = \begin{cases} \dfrac{1}{2\pi} & \text{for} \quad 0 \leqq \theta < 2\pi, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

To obtain the spectral density for $W(t)$, it is convenient to express the wave in terms of complex variables as[5]

$$W(t) = \text{Re } \{\tilde{W}(t)\} \tag{3}$$

where[†]

$$\tilde{W}(t) = A \exp [j(\omega_c t + \theta)]V(t), \tag{4}$$

and

$$V(t) = \exp [jg(X_t)]. \tag{5}$$

Since X_t is stationary and θ is uniformly distributed, both $\tilde{W}(t)$ and $V(t)$ are wide-sense stationary.[5] Moreover, the spectral density $S_w(f)$ of $W(t)$ is given by

$$S_w(f) = \tfrac{1}{4}[S_v(f - f_c) + S_v(-f - f_c)] \tag{6}$$

where $S_v(f)$ is the spectral density of $V(t)$.[5]

If we define $R_v(\tau) = \langle V(t + \tau)\overline{V(t)}\rangle$, the Wiener–Khintchine theorem implies that

$$S_v(f) = \int_{-\infty}^{\infty} R_v(\tau) \exp [-j2\pi f\tau] \, d\tau. \tag{7}$$

[†] We use Re (z) and Im (z) to denote respectively the real and imaginary parts of the complex variable z. The symbol $\bar{z}$ denotes the complex conjugate of z.

So, assume that $X_{t+\tau}$ and X_t are gaussian with joint probability density function

$$p(x_1\,,\,x_2) = \frac{1}{2\pi\Gamma^2\sqrt{1-r^2}}\exp\left(-\frac{x_1^2 - 2rx_1x_2 + x_2^2}{2\Gamma^2(1-r^2)}\right),$$
$$-\infty < x_i < \infty, \qquad (8)$$

where $R_x(\tau) = \langle X_{t+\tau}X_t\rangle$, $\Gamma^2 = R_x(0)$ and $r = R_x(\tau)/\Gamma^2$. It follows that

$$R_v(\tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\,[j(g(x_1) - g(x_2))]p(x_1\,,\,x_2)\,dx_1\,dx_2\,. \qquad (9)$$

In Appendix A, we show that if $G(x)$ is of exponential order and of bounded variation on bounded intervals,[†] then

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x_1)\overline{G(x_2)}p(x_1\,,\,x_2)\,dx_1\,dx_2 = \sum_{n=0}^{\infty} a_n r^n, \qquad |\,r\,| \leqq 1, \qquad (10)$$

where

$$a_n = \frac{\Gamma^{2n}}{n!}\left|\int_{-\infty}^{\infty} G(x)\left(\frac{d^n}{dx^n}\,p(x)\right)dx\,\right|^2,$$

and

$$p(x) = \frac{1}{\sqrt{2\pi\Gamma^2}}\exp\left(-\frac{x^2}{2\Gamma^2}\right)$$

is the probability density for X_t. Setting $G(x) = \exp[jg(x)]$,

$$R_v(\tau) = \sum_{n=0}^{\infty} C_n r^n, \qquad (11)$$

where[‡]

$$C_n = \frac{\Gamma^{2n}}{n!}\left|\int_{-\infty}^{\infty} \exp\,[jg(x)]\left[\frac{d^n}{dx^n}\,p(x)\right]dx\,\right|^2. \qquad (12)$$

Recall that the Hermite polynomial of degree n, $H_n(x)$, satisfies the relation

$$\frac{d^n}{dx^n}\exp\,(-x^2) = (-1)^n H_n(x)\exp\,(-x^2).$$

[†] We define such functions to be "reasonable."

[‡] Since $|\exp\,[jg(x)]| = 1$ for all finite $g(x)$, it follows that equations (11) and (12) are valid for all functions $g(x)$ which are of bounded variation on finite intervals. This should include most functions which one might encounter in engineering applications.

Consequently,

$$C_n = \frac{1}{2^n n!} \left| \int_{-\infty}^{\infty} \exp\,[jg(x)] H_n\left(\frac{x}{\sqrt{2}\,\Gamma}\right) p(x)\, dx \right|^2 . \tag{13}$$

Since the coefficients $\{C_n\}$ are independent of $R_x(\tau)$, an infinite series representation for the spectral density $S_v(f)$ can be obtained from equations (7) and (11). To obtain the series, let

$$S_1(f) = \int_{-\infty}^{\infty} R_z(\tau) \exp\,(-j2\pi f\tau)\, d\tau \tag{14}$$

and define $S_n(f)$ to be the n-fold convolution of $S_1(f)$ with itself; i.e., $S_n(f) = S_{n-1}(f) * S_1(f)$ for $n \geqq 2$. It follows that

$$S_v(f) = C_0\, \delta(f) + \sum_{n=1}^{\infty} C_n \frac{S_n(f)}{\Gamma^{2n}}, \qquad -\infty < f < \infty . \tag{15}$$

Using equations (13) and (15), one can see that the carrier power is given by

$$C_0 = |\langle \exp\,[jg(X_t)] \rangle|^2 . \tag{16}$$

2.1 *Examples*

To obtain a better understanding of the preceding results, we present two examples. The first example uses $g(x) = x$ (and serves as a partial check on our results since this case is well known[5]). The second example assumes extreme clipping,

$$g(x) = \begin{cases} b & \text{for} \quad x > 0, \\ -b & \text{for} \quad x < 0, \end{cases}$$

and serves as a preview for the hard-limiting case presented in Section III.

2.1.1 *Phase Modulation*

For this example, $g(x) = x$ for all x. Consequently, from equation (13) we have, for $n \geqq 0$,

$$C_n = \frac{1}{2^n n!} \left| \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\,(j\sqrt{2}\,\Gamma x) H_n(x) \exp\,(-x^2)\, dx \right|^2 . \tag{17}$$

Thus,

$$C_{2n} = \frac{1}{2^{2n}(2n)!} \left| \frac{2}{\sqrt{\pi}} \int_0^{\infty} \cos\,(\sqrt{2}\,\Gamma x) H_{2n}(x) \exp\,(-x^2)\, dx \right|^2 , \tag{18}$$

$$C_{2n} = \exp(-\Gamma^2)\frac{(\Gamma^2)^{2n}}{(2n)!}, \tag{19}$$

(see Ref. 6, Section 7.388-3).

In a similar fashion $C_{2n+1} = \exp(-\Gamma^2)(\Gamma^2)^{2n+1}/(2n+1)!$, which implies that

$$C_n = \exp(-\Gamma^2)\frac{\Gamma^{2n}}{n!} \quad \text{for} \quad n \geqq 0. \tag{20}$$

Now, substitute equation (20) into equation (11) to obtain the well-known result[5]

$$R_v(\tau) = \exp[-\Gamma^2(1 - r)]. \tag{21}$$

2.1.2 *Phase Modulation by an Extreme-Limited Gaussian Process*

For this case,

$$g(x) = \begin{cases} b & \text{if} \quad x > 0, \\ -b & \text{if} \quad x < 0. \end{cases} \tag{22}$$

Using equation (22) in equation (13), one obtains

$$C_n = \frac{1}{2^n n!}\left| \exp(jb)\int_0^\infty H_n\left(\frac{x}{\sqrt{2}\,\Gamma}\right)p(x)\,dx + \exp(-jb)\int_{-\infty}^0 H_n\left(\frac{x}{\sqrt{2}\,\Gamma}\right)p(x)\,dx \right|^2, \tag{23}$$

which reduces to

$$C_{2n} = \frac{(\cos(b))^2}{2^{2n}(2n)!}\left(\frac{2}{\sqrt{\pi}}\int_0^\infty H_{2n}(x)\exp(-x^2)\,dx\right)^2 \tag{24}$$

and

$$C_{2n+1} = \frac{(\sin(b))^2}{2^{2n+1}(2n+1)!}\left(\frac{2}{\sqrt{\pi}}\int_0^\infty H_{2n+1}(x)\exp(-x^2)\,dx\right)^2, \tag{25}$$

for $n \geqq 0$. Since

$$\int_0^\infty \exp(-y^2)H_n(y)\,dy = H_{n-1}(0) \quad \text{for} \quad n \geqq 1, \tag{26}$$

we have

$$C_0 = (\cos(b))^2, \tag{27}$$

$$C_{2n} = 0 \quad \text{for} \quad n \geqq 1, \tag{28}$$

and

$$C_{2n+1} = \frac{(\sin (b))^2}{2^{2n+1}(2n+1)!}\left(\frac{2}{\sqrt{\pi}} H_{2n}(0)\right)^2. \tag{29}$$

Using the fact that $H_0(0) = 1$ and $H_{2n}(0) = (-1)^n 2^n [1 \cdot 3 \cdots (2n-1)]$ for $n \geqq 1$, we see that

$$C_1 = \frac{2}{\pi} (\sin (b))^2$$

and

$$C_{2n+1} = \frac{2}{\pi} (\sin (b))^2 \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2^n n!\,(2n+1)} \quad \text{for} \quad n \geqq 1. \tag{30}$$

Substituting equations (27), (28) and (30) into equation (11) and recalling the series expansion for Arcsin x, one obtains

$$R_v(\tau) = (\cos (b))^2 + (\sin (b))^2 \frac{2}{\pi} \text{Arcsin} \left(\frac{R(\tau)}{R(0)}\right), \tag{31}$$

for $-\infty < \tau < \infty$.

Notice that the carrier power, $C_0 = (\cos (b))^2$, vanishes if $b = \pi/2 + k\pi$, $k = 0, 1, \cdots$, while all the power goes into the carrier whenever $b = k\pi$, $k = 0, 1, \cdots$. If $b \neq k\pi$, the continuous part of the spectrum is simply a scaled version of the spectrum for an amplitude-modulated wave when the modulator is an extreme-limited gaussian process. This problem was studied by J. H. VanVleck.[7]

III. PHASE MODULATION BY A HARD-LIMITED GAUSSIAN PROCESS

A problem of interest arises when the function g represents an ideal hard-limiter; i.e.,

$$g(x) = \begin{cases} b & \text{for} \quad x > b, \\ x & \text{for} \quad |x| \leqq b, \\ -b & \text{for} \quad x < -b, \end{cases} \tag{32}$$

for some $b \geqq 0$.

3.1 *Carrier Power*

Noting equations (13) and (16), one can see that the carrier power is given by

$$C_0 = \left(\frac{1}{\sqrt{2\pi\Gamma^2}} \int_{-b}^{b} \cos(x) \exp\left(-\frac{x^2}{2\Gamma^2}\right) dx \right.$$
$$\left. + \cos(b) \frac{2}{\sqrt{2\pi\Gamma^2}} \int_{b}^{\infty} \exp\left(-\frac{x^2}{2\Gamma^2}\right) dx \right)^2 .$$

Defining $\alpha = b/\Gamma$ (the relative clipping level), we have

$$C_0 = \left(\frac{2}{\sqrt{\pi}} \int_0^{\alpha/\sqrt{2}} \cos(\sqrt{2}\,\Gamma x) \exp(-x^2)\, dx + \cos(b)\, \operatorname{erfc}\left(\frac{\alpha}{\sqrt{2}}\right) \right)^2 \tag{33}$$

where

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^{z} \exp(-t^2)\, dt \tag{34}$$

is the error function[8] and erfc $(z) = 1 - \operatorname{erf}(z)$.

A more useful expression for C_0 can be obtained in terms of the error function with complex argument. Apparently,

$$\frac{2}{\sqrt{\pi}} \int_0^{\alpha/\sqrt{2}} \cos(\sqrt{2}\,\Gamma x) \exp(-x^2)\, dx$$

$$= \operatorname{Re}\left\{ \frac{1}{\sqrt{\pi}} \int_{-\alpha/\sqrt{2}}^{\alpha/\sqrt{2}} \exp(-j\sqrt{2}\,\Gamma x - x^2)\, dx \right\}, \tag{35}$$

$$= \exp\left(-\frac{\Gamma^2}{2}\right) \operatorname{Re}\left\{ \frac{1}{\sqrt{\pi}} \int_{-\alpha/\sqrt{2}}^{\alpha/\sqrt{2}} \exp\left(-\left(x + j\frac{\Gamma}{\sqrt{2}}\right)^2\right) dx \right\}, \tag{36}$$

$$= \frac{1}{2} \exp\left(-\frac{\Gamma^2}{2}\right) \operatorname{Re}\left\{ \operatorname{erf}\left(\frac{\alpha}{\sqrt{2}} + j\frac{\Gamma}{\sqrt{2}}\right) - \operatorname{erf}\left(\frac{\alpha}{\sqrt{2}} + j\frac{\Gamma}{\sqrt{2}}\right) \right\}, \tag{37}$$

$$= \exp\left(-\frac{\Gamma^2}{2}\right) \operatorname{Re}\left\{ \operatorname{erf}\left(\frac{\alpha}{\sqrt{2}} + j\frac{\Gamma}{\sqrt{2}}\right) \right\} .^{\dagger} \tag{38}$$

Consequently,

$$C_0 = \left(\exp\left(-\frac{\Gamma^2}{2}\right) \operatorname{Re}\left\{ \operatorname{erf}\left(\frac{\alpha}{\sqrt{2}} + j\frac{\Gamma}{\sqrt{2}}\right) \right\} + \cos(b)\, \operatorname{erfc}\left(\frac{\alpha}{\sqrt{2}}\right) \right)^2 \tag{39}$$

can be calculated numerically.[8] However, we can obtain a better analytical understanding of the carrier power by expressing the integral in equation (36) as

[†] To make the last step, we used the relations[8]

$$\operatorname{erf}(-z) = -\operatorname{erf}(z) \quad \text{and} \quad \operatorname{erf}(\bar{z}) = \overline{\operatorname{erf}(z)}.$$

$$\exp\left(-\frac{\Gamma^2}{2}\right)\left\{\operatorname{Re}\frac{2}{\sqrt{\pi}}\int_0^{\alpha/\sqrt{2}}\exp\left(-\left(x+j\frac{\Gamma}{\sqrt{2}}\right)^2\right)dx\right\}$$
$$=\exp\left(-\frac{\Gamma^2}{2}\right)\operatorname{Re}\left\{\int_{j(\Gamma/\sqrt{2})}^{\alpha/\sqrt{2}+j(\Gamma/\sqrt{2})}\exp(-z^2)\,dz\right\}\cdot \qquad (40)$$

Since $\exp(-z^2)$ is an entire function, the integral in equation (40) is independent of the path taken from $j(\Gamma/\sqrt{2})$ to $\alpha/\sqrt{2}+j(\Gamma/\sqrt{2})$. It is useful to use a contour consisting of three straight line-segments: the first from $j(\Gamma/\sqrt{2})$ to 0 (on the imaginary axis), the second from 0 to $\alpha/\sqrt{2}$ (on the real axis) and the last from $\alpha/\sqrt{2}$ to $\alpha/\sqrt{2}+j(\Gamma/\sqrt{2})$. Integrating along these lines, it is straightforward to show that

$$\exp\left(-\frac{\Gamma^2}{2}\right)\operatorname{Re}\left\{\frac{2}{\sqrt{\pi}}\int_{j(\Gamma/\sqrt{2})}^{\alpha/\sqrt{2}+j(\Gamma/\sqrt{2})}\exp(-z^2)\,dz\right\}$$
$$=\exp\left(-\frac{\Gamma^2}{2}\right)\operatorname{erf}\left(\frac{\alpha}{\sqrt{2}}\right)$$
$$+\exp\left(-\frac{\alpha^2}{2}\right)\frac{2}{\sqrt{\pi}}\int_0^{\Gamma/\sqrt{2}}\sin(\sqrt{2}\,\alpha y)\cdot\exp\left(-\left(\frac{\Gamma^2}{2}-y^2\right)\right)dy \qquad (41)$$

Combining equations (33) and (41) yields

$$C_0=\left(\exp\left(-\frac{\Gamma^2}{2}\right)\operatorname{erf}\left(\frac{\alpha}{\sqrt{2}}\right)+\cos(b)\operatorname{erfc}\left(\frac{\alpha}{\sqrt{2}}\right)\right.$$
$$\left.+\exp\left(-\frac{\alpha^2}{2}\right)\frac{2}{\sqrt{\pi}}\int_0^{\Gamma/\sqrt{2}}\sin(\sqrt{2}\,\alpha y)\exp\left(-\left(\frac{\Gamma^2}{2}-y^2\right)\right)dy\right)^2\cdot \qquad (42)$$

When $\Gamma \ll 1$ (low-index case), the last term inside the brackets is negligible so that

$$C_0\approx\left(\exp\left(-\frac{\Gamma^2}{2}\right)\operatorname{erf}\left(\frac{\alpha}{\sqrt{2}}\right)+\cos(b)\operatorname{erfc}\left(\frac{\alpha}{\sqrt{2}}\right)\right)^2 \text{ when } \Gamma\ll 1. \qquad (43)$$

It is interesting to observe that erf $(\alpha/\sqrt{2})$ is the probability that X_t is inside the clipping levels $[-b, b]$ [and erfc $(\alpha/\sqrt{2})$ is the probability that X_t is outside the clipping levels]. Hence, we can reason that $\exp(-\Gamma^2/2)$ is the average dc voltage associated with $V(t)$ when X_t is inside the clipping levels while cos (b) is the average voltage when X_t is outside the limits. [One can see from equation (16) that the carrier power $C_0 = |\langle V(t)\rangle|^2$.]

Pursuing these thoughts further, one might conclude that the integral in equation (42) represents a high-order correction (to the preceding approximation) associated with transitions of X_t across the clipping thresholds. When Γ is small, these transitions are rapid and the correction (i.e., the integral) is small. However, as Γ increases, the transitions are slower (at least when X_t is bandlimited) and the integral in equation (42) can increase in magnitude for certain values of α. Moreover,

$$\frac{2}{\sqrt{\pi}} \int_0^{\Gamma/\sqrt{2}} \sin(\sqrt{2}\,\alpha y) \exp\left(-\left(\frac{\Gamma^2}{2} - y^2\right)\right) dy$$
$$\approx \sin(b) F\left(\frac{\Gamma}{\sqrt{2}}\right) \quad \text{when} \quad \Gamma \gg \sqrt{2},$$

where $F(x) = 2/\sqrt{\pi} \int_0^x \exp[-(x^2 - t^2)]\, dt$ is Dawson's function.[8] Consequently it follows that when $\Gamma \gg \sqrt{2}$ (i.e., the high-index case),

$$C_0 \approx \left(\exp\left(-\frac{\Gamma^2}{2}\right) \operatorname{erf}\left(\frac{\alpha}{\sqrt{2}}\right) + \cos(b) \operatorname{erfc}\left(\frac{\alpha}{\sqrt{2}}\right)\right.$$
$$\left. + \sin(b) \exp\left(-\frac{\alpha^2}{2}\right) F\left(\frac{\Gamma}{\sqrt{2}}\right)\right)^2. \qquad (44)$$

Thus, when α is not too large,

$$C_0 \approx \left(\cos(b) \operatorname{erfc}\left(\frac{\alpha}{\sqrt{2}}\right) + \sin(b) \exp\left(-\frac{\alpha^2}{2}\right) F\left(\frac{\Gamma}{\sqrt{2}}\right)\right)^2. \qquad (45)$$

The quantity inside the brackets is a sinusoidal function of α. Hence, there are various values of α (for fixed Γ) which will completely null the carrier power C_0 while other clipping levels give rise to a relative maximum carrier power.

Using

$$\langle V(t) \rangle = \exp\left(-\frac{\Gamma^2}{2}\right) \operatorname{Re}\left\{\operatorname{erf}\left(\frac{\alpha}{\sqrt{2}} + j\frac{\Gamma}{\sqrt{2}}\right)\right\} + \cos(b) \operatorname{erfc}\left(\frac{\alpha}{\sqrt{2}}\right),$$

we computed $\exp(\Gamma^2/2)\langle V(t)\rangle$ as a function of α for several values of Γ. Some of the results are displayed in Fig. 1. (Since the dynamic range of this function is so large, we used a *linear scale* between -1 and 1 on the ordinate axis and a logarithmic scale otherwise.)

The general features of the three functions are as indicated above. Since $\langle V(t) \rangle = 1$ whenever $\alpha = 0$, each curve starts at $\exp(\Gamma^2/2)$ for $\alpha = 0$. For $\Gamma \leqq 1$ (not shown in Fig. 1) the curve decreased monotonically to one. For $\Gamma = 2$, this relative average remains positive but oscillates slightly before converging to unity. However, for the high-index case, the relative mean is similar to a damped sinusoid (as dis-

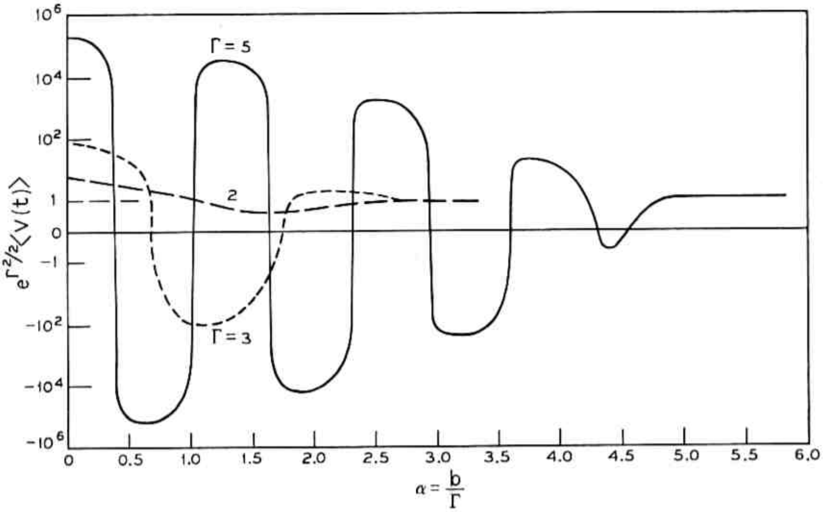


Fig. 1—Relative mean of $V(t)$ as a function of relative clipping level. α is the ratio of the clipping level b to the rms modulation-index Γ.

cussed above). This extreme oscillation about zero apparently occurs for $\Gamma \geqq 3$. Notice that the value of α required for the relative mean (and therefore the carrier power) to stabilize increases as Γ increases. Hence, for high-index modulation, setting $\alpha \approx 4$ does not necessarily guarantee that the carrier power is unaffected by the clipping.

3.2 *Continuous Part of the Spectrum*

Numerical techniques are required to obtain the continuous part of the spectrum. However, one can see from equations (23) and (31) that the coefficients $\{C_n\}$ in the series are independent of $R_x(\tau)$ and need only be computed once for any particular configuration of the hard limiter (and choice of Γ). Consequently, it seems worthwhile to derive an algorithm for efficient computation of $\{C_n\}$.

Noting equations (11), (12) and (13), one can see that

$$R_v(\tau) = \sum_{n=0}^{\infty} \left(\frac{R(\tau)}{\Gamma^2}\right)^n C_n \tag{46}$$

where

$$C_n = \frac{\Gamma^{2n} \mid \nu(n) \mid^2}{n!} \tag{47}$$

and

$$j^n\nu(n) = \int_{-\infty}^{\infty} \exp\left[-jg(x)\right]\left(\frac{d^n}{dx^n}\, p(x)\right) dx. \tag{48}$$

Consequently,

$$\begin{aligned} j\nu(1) & \\ &= \int_{-b}^{b} \exp(-jx)\, dp(x) + \exp(-jb) \int_{b}^{\infty} dp(x) + \exp(jb) \int_{-\infty}^{-b} dp(x), \\ &= \exp\left(-\frac{\Gamma^2}{2}\right) \operatorname{Re}\left(\operatorname{erfc}\left(\frac{\alpha}{\sqrt{2}} + j\,\frac{\Gamma}{\sqrt{2}}\right)\right). \end{aligned} \tag{49}$$

Next, for $n \geqq 2$ we have

$$\begin{aligned} j^n\nu(n) = \int_{-b}^{b} \exp(-jx)\, d\left(\frac{d^{n-1}}{dx^{n-1}}\, p(x)\right) + \exp(-jb) \int_{b}^{\infty} d\left(\frac{d^{n-1}}{dx^{n-1}}\, p(x)\right) \\ + \exp(jb) \int_{-\infty}^{-b} d\left(\frac{d^{n-1}}{dx^{n-1}}\, p(x)\right). \end{aligned}$$

Integrating the above expression by parts, we have

$$\nu(n) = \frac{1}{j^{n-1}} \int_{-b}^{b} \exp(-jx)\left(\frac{d^{n-1}}{dx^{n-1}}\, p(x)\right) dx. \tag{50}$$

Recall that

$$\frac{d^n}{dx^n} \exp\left(-\frac{x^2}{2\Gamma^2}\right) = \left(\frac{-1}{\sqrt{2}\;\Gamma}\right)^n H_n\left(\frac{x}{\sqrt{2}\;\Gamma}\right) \exp\left(-\frac{x^2}{2\Gamma^2}\right)$$

so that integrating equation (50) by parts obtains

$$\nu(n) = \nu(n-1) - \frac{1}{\sqrt{\pi}}\left(\frac{j}{\sqrt{2}\;\Gamma}\right)^{n-1} H_{n-2}\left(\frac{x}{\sqrt{2}\;\Gamma}\right) \exp\left(-jx - \frac{x^2}{2\Gamma^2}\right)\Bigg|_{-b}^{b}. \tag{51}$$

Of course, $H_n(-x) = (-1)^n H_n(x)$, which implies that the exact determination of equation (51) depends on whether n is even or odd. It is straightforward to show that

$$\begin{aligned} \nu(2n) &= \nu(2n-1) \\ &\quad + (-1)^n \sin(b)\, \frac{2}{\sqrt{\pi}} \exp\left(-\frac{\alpha^2}{2}\right)\left(\frac{1}{\sqrt{2}\;\Gamma}\right)^{2n-1} H_{2(n-1)}\left(\frac{\alpha}{\sqrt{2}}\right) \end{aligned} \tag{52}$$

and

$$\nu(2n+1) = \nu(2n) - (-1)^n \cos(b) \frac{2}{\sqrt{\pi}} \exp\left(-\frac{\alpha^2}{2}\right)\left(\frac{1}{\sqrt{2}\,\Gamma}\right)^{2n} H_{2n-1}\left(\frac{\alpha}{\sqrt{2}}\right)$$

for $n \geqq 1$.

It is known that[8]

$$H_0(x) = 1, \qquad H_1(x) = 2x$$

and

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x) \tag{53}$$

for $n \geqq 1$, so that in principle, the sequence $\{C_n\}$ can be computed by using equations (47), (52) and (53). However, in practice the difference-equations (52) and (53) are unstable and care must be taken in their implementation. We found it best to use the following approach:

Let $[x]$ denote the largest integer not exceeding x and let $n!! = n!/2^{[n/2]}[n/2]!$. [So that $(2n-1)!! = 1\cdot 3\cdot 5\cdots(2n-1)$.] Now, one can use equations (52) and (53) to obtain stable algorithms for

$$t_1(n) = \frac{\Gamma^n \nu(n)}{2^{[n/2]}\left[\dfrac{n}{2}\right]!} \quad \text{and} \quad t_2(n) = \frac{\Gamma^n \nu(n)}{n!!}$$

which are used to compute

$$C_n = t_1(n)t_2(n).$$

3.3 *Numerical Results*

To obtain insight into the effect of hard-limiting, we assume that X_t is bandlimited with a rectangular power spectrum. That is to say

$$R_x(\tau) = \Gamma^2 \frac{\sin 2\pi W\tau}{2\pi W\tau} \quad \text{for} \quad |\tau| < \infty. \tag{54}$$

In this case,[9]

$$S_v(f/W) = C_0\,\delta(f) + \frac{1}{2\pi W}\sum_{n=1}^{\infty} C_n F_n(f/W), \tag{55}$$

for $-\infty < f < \infty$. The functions $F_n(\lambda)$ are defined as follows:[9]

$$F_1(\lambda) = \begin{cases} \pi, & -1 < \lambda < 1, \\ 0, & \text{otherwise,} \end{cases} \tag{56}$$

and for $n \geqq 2$,

$$F_n(\lambda) = \begin{cases} n\pi \displaystyle\sum_{k=0}^{M(n,\lambda)} (-1)^k \frac{\left(\dfrac{|\lambda| + n}{2} - k\right)^{n-1}}{k!\,(n-k)!}, & 0 \leqq |\lambda| < n, \\ 0, & \text{otherwise}, \end{cases} \tag{57}$$

where $M(n, \lambda) = [n + |\lambda|/2]$. We use $[x]$ to denote the integer part of x.

As $M(n, \lambda)$ increases, it becomes increasingly difficult to accurately sum the alternating series displayed in equation (57). Since

$$F_n(\lambda) \sim \left(\frac{6\pi}{n}\right)^{\frac{1}{2}} \exp\left(-\frac{3\lambda^2}{n}\right) \tag{58}$$

(see Ref. 9), we used the asymptotic approximation (58) whenever $F_n(\lambda)$ was required for $n \geqq 15$. The threshold 15 was determined empirically. We attempted to keep our numerical estimates of the various spectra accurate to one percent relative error.

We considered three cases: low modulation-index ($\Gamma = 0.1$), moderately high index ($\Gamma = 1$) and high index ($\Gamma = 5$). The results are displayed in Figs. 2 through 4. In each case, we have computed the effect of changes in the relative clipping level $\alpha = b/\Gamma$.

3.3.1 *Low Modulation-Index*

When $\Gamma \ll 1$, it is evident from Fig. 2 and equation (55) that the principal part of the spectrum is well approximated by

$$S_v(f/W) \approx \begin{cases} C_0\,\delta(f) + \dfrac{C_1}{2\pi W} F_1(f/W), & \text{for } |f| < W, \\ 0, & \text{otherwise}. \end{cases} \tag{59}$$

Using results from the preceding section, we can show that when $\Gamma \ll 1$, $C_1 \approx \Gamma^2 \exp(-\Gamma^2)\,(\operatorname{erf}(\alpha/\sqrt{2}))^2$. Hence, for $\Gamma \ll 1$,

$$S_v(f/W) \approx \begin{cases} C_0\,\delta(f) + \left(\operatorname{erf}\left(\dfrac{\alpha}{\sqrt{2}}\right)\right)^2 \dfrac{\Gamma^2 \exp(-\Gamma^2)}{2\pi W} F_1(f/W), & |f| < W, \\ 0, & \text{otherwise}. \end{cases} \tag{60}$$

That is to say, in the low-index case ($\Gamma \ll 1$), hard-limiting causes the principal part of the spectrum to be scaled down by the multiplicative factor $(\operatorname{erf}(\alpha/\sqrt{2}))^2$.

Figure 2 illustrates the actual behavior [as opposed to the approxi-

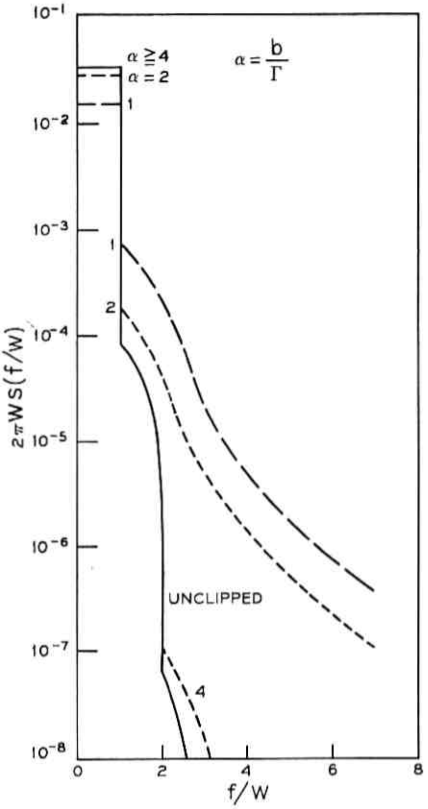


Fig. 2—Spectrum of low-index PM wave ($\Gamma = 0.1$) for various relative clipping levels. α is the ratio of the clipping level b to the rms modulation-index Γ.

mation (60)] of $S_v(f/W)$ for $f > W$. Notice that the tails of the spectrum are raised considerably as the relative clipping level α decreases. This phenomenon is noticeable even for $\alpha = 4$ (i.e., clipping at a four-sigma level). The increase in high-frequency content is apparently caused by the introduction of points at which the derivative of $W(t)$ is discontinuous.

3.3.2 *Moderately High Modulation-Index*

The results for $\Gamma = 1$ are displayed in Fig. 3. The qualitative results are similar to those observed for $\Gamma = 0.1$; i.e., the principal part of the spectrum tends to decrease and the tails increase, as the relative clipping level α decreases. However, notice that for a particular value of α, the frequency f/W at which the tails of the spectrum begin to rise is somewhat larger for $\Gamma = 1$ than for $\Gamma = 0.1$.

3.3.3 *High Modulation-Index*

When $\Gamma = 5$, the behavior of $\langle V(t) \rangle$ as a function of α is quite different from that observed for $\Gamma \leqq 1$ (see Fig. 1). In fact, small changes in α can change the discrete component of the spectrum, $C_0 = \langle V(t) \rangle^2$, from a relative maximum to zero. Consequently, we originally suspected that the continuous part of the spectrum might change significantly as the discrete component changed from a relatively large value to zero. To check this point, four values of α were selected for testing;

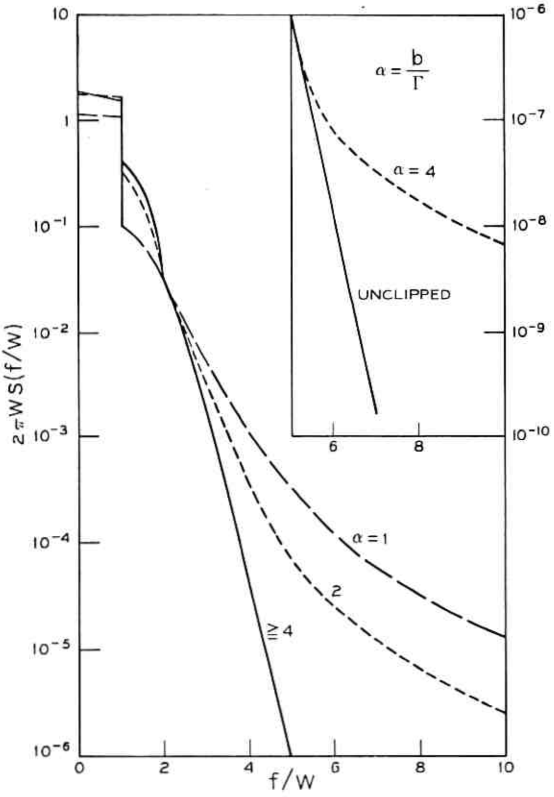
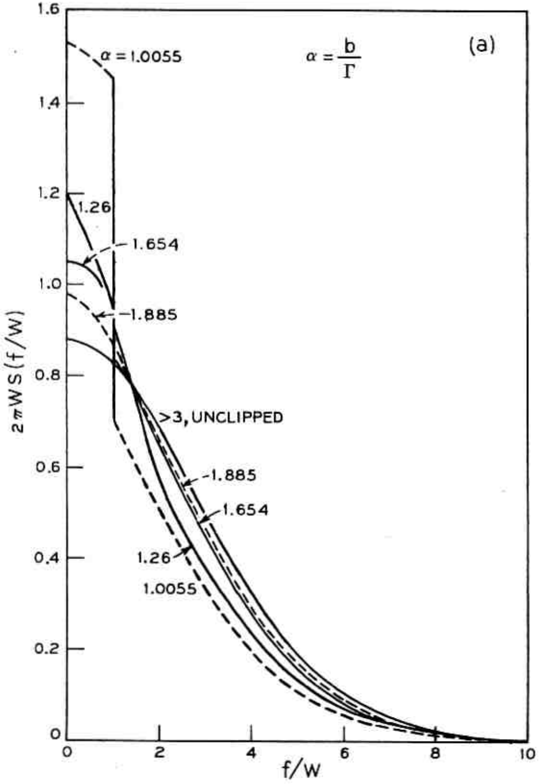


Fig. 3—Spectrum of waves having moderately high modulation-index ($\Gamma = 1$) for various relative clipping-levels. α is the ratio of the clipping level b to the rms modulation-index Γ.



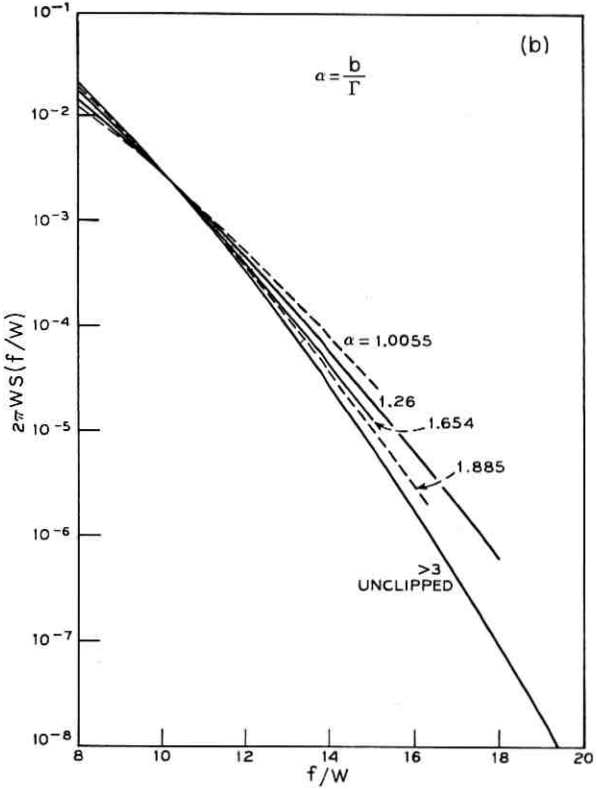


Fig. 4a—Principal part of spectrum for PM waves having high modulation-index ($\Gamma = 5$) for various relative clipping levels. α is the ratio of the clipping level b to the rms modulation-index Γ.

Fig. 4b—Tails of spectrum for PM waves having high modulation-index ($\Gamma = 5$) for various relative clipping levels. α is the ratio of the clipping level b to the rms modulation-index Γ.

two at points of relative maxima for C_0 and two where C_0 was nearly zero. The results are displayed in Figs. 4a and b.

From Figure 4a, we see that the qualitative behavior of $S_v(f/W)$ for high modulation-index ($\Gamma = 5$) differs significantly from that observed for $\Gamma \leqq 1$. However, whether or not C_0 vanishes does not seem to affect the general characteristics of the continuous part of the spectrum. Generally speaking as α decreases, the portion of the spectrum for $0 < f/W < 1$ tends to increase while the part for $1 < f/W < 10$ tends to decrease. Notice that a discontinuity is introduced at $f/W = 1$, a characteristic of lower modulation-index. Of course, limiting reduces the variance of the modulating signal X_t, so that the discontinuity at $f/W = 1$ is not surprising.

To examine the behavior of the tails of the spectra, look at Figure 4b. As in the other cases, decreasing α raises the tails of the spectrum, although the effect is considerably less than that observed for $\Gamma \leqq 1$. In fact, for $\alpha > 3$, we observed very little difference between the clipped and unclipped cases.

3.4 *Comments*

Computation of the spectrum for large values of f/W, using the infinite series approach, is expensive since many terms are required. Moreover, accurate computation of C_n and F_n, for large n, is difficult (if not impossible) because of numerical problems. Consequently, a good estimate of the spectrum for large f/W would be very desirable. Unfortunately, we do not have such an approximation. However, it is possible to estimate the relative frequency f/W at which the tails of the spectrum for a clipped modulator will begin to depart from the unclipped case.

More precisely, consider equations (48) and (49) which constitute an algorithm for the computation of $\nu(n)$. Since the tails of the spectrum of $W(t)$ can rise only when $\nu(n)$ increases as a function of n, we need to examine $\nu(n)$ for large n.

For large n, we know that[6]

$$H_{2n}(x) = (-1)^n 2^n (2n - 1)!! \exp(x^2/2)\left\{\cos(\sqrt{4n + 1}\, x) + O\left(\frac{1}{n^{\frac{1}{2}}}\right)\right\} \tag{61}$$

and

$$H_{2n+1}(x) = (-1)^n 2^{n+\frac{1}{2}}(2n - 1)!!\sqrt{2n + 1} \cdot \exp(x^2/2)\left\{\sin(\sqrt{4n + 3}\, x) + O\left(\frac{1}{n^{\frac{1}{2}}}\right)\right\}. \tag{62}$$

Hence, for large n we have the approximation

$$\nu(2n) \approx \nu(2n-1) - 2\Gamma \sin(b)\sqrt{\frac{2}{\pi}} \cos(\alpha\sqrt{2n - 3/2}) \cdot \exp\left(-\frac{\alpha^2}{4}\right) \prod_{k=1}^{n-1} \left(\frac{2k-1}{\Gamma^2}\right), \tag{63}$$

$$\nu(2n+1) \approx \nu(2n) + \cos(b)\sqrt{\frac{2}{\pi}} \sin(\alpha\sqrt{2n + 3/2}) \cdot \exp\left(-\frac{\alpha^2}{4}\right) \frac{\sqrt{2n+1}}{\Gamma^2} \prod_{k=1}^{n-1} \left(\frac{2k-1}{\Gamma^2}\right). \tag{64}$$

It is evident that the sequence $\mu_n = \prod_{k=1}^{n} [(2k-1))\Gamma]$ is a decreasing function of n for $n < n_0 = [\Gamma^2 + 3/2]$ and is an increasing function for $n > n_0$. For any finite α, μ_n will ultimately exceed $\exp(\alpha^2/4)$ and the forcing function for $\nu(2n)$ will begin to grow without bound. It follows that $\nu(2n)$ will exhibit instability when n is so large that $\mu_n \gg \exp(\alpha^2/4)$. This will certainly be true when $(2n-1)/\Gamma^2 \geqq \exp(\alpha^2/4)$ or for

$$n \geqq \tfrac{1}{2}[\Gamma^2 \exp(\alpha^2/4) + 1]. \tag{65}$$

An inspection of Figs. (2), (3) and (4) shows that equation (65) gives a good indication (at least for those cases considered) of the value of f/W where the tails of the clipped spectra begin a significant departure from the unclipped case. Equations (63) and (64) also indicate that for large n, $\nu(n)$ will oscillate as it increases in magnitude. Consequently, it appears that it will be difficult to get tight bounds on $S_\nu(f/W)$ for large f/W. It is interesting that the carrier power is most significantly altered by the hard-limiting when $\Gamma \gg 1$ while the effect (of hard-limiting) on the tails of the spectrum decreases as Γ increases.

IV. CONCLUSIONS

In principle, equation (68) can be used to obtain either the autocorrelation function or the spectral density of any "reasonable" function of a stationary gaussian process. We made use of the identity to study the spectral density of a sinusoidal wave which is phase-modulated by a hard-limited gaussian process.

The main disadvantage of this approach (i.e., using an infinite series solution) is that numerical techniques are usually required to obtain a solution. Consequently, estimates of the tails of spectra (of interest for PM systems having high modulation-index) are difficult and expensive to compute accurately. However, when no better alternative is available, our approach can be used to obtain numerical results in a relatively straightforward manner.

V. ACKNOWLEDGMENTS

I am grateful to V. K. Prabhu and H. E. Rowe for considerable help and encouragement while working on this problem. Suggestions made by L. Shepp and G. J. Foschini were very helpful in establishing a rigorous proof for equation (68).

APPENDIX

A Relation

Let X_t denote a stationary gaussian process with probability density function

$$p(x) = \frac{1}{2\pi\Gamma^2} \exp\left(-\frac{x^2}{2\Gamma^2}\right), \qquad -\infty < x < \infty. \tag{66}$$

The joint probability density function for $X_{t+\tau}$ and X_t is

$$p(x_1\,,x_2) = \frac{1}{2\pi\Gamma^2\sqrt{1-r^2}} \exp\left[-\frac{x_1^2 - 2rx_1x_2 + x_2^2}{2\Gamma^2(1-r^2)}\right], \qquad -\infty < x_i < \infty, \tag{67}$$

where

$$R_x(\tau) = \langle X_{t+\tau}X_t\rangle, \qquad \Gamma^2 = R_x(0) \quad \text{and} \quad r = \frac{R_x(\tau)}{\Gamma^2}$$

for $-\infty < \tau < \infty$.

If $G(x)$ is an exponentially bounded complex-valued function of the real variable x (i.e., there are real numbers u and v such that $|G(x)| \leqq v \exp(u\,|x|)$ for $-\infty < x < \infty$) and if $G(x)$ is of bounded variation on finite intervals, then we shall show that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x_1)\overline{G(x_2)}p(x_1\,,x_2)\,dx_1\,dx_2 = \sum_{n=0}^{\infty} \frac{(R_x(\tau))^n}{n!}\left|\int_{-\infty}^{\infty} G(x)\left(\frac{d^n}{dx^n}p(x)\right)dx\right|^2. \tag{68}$$

A.1 *Comments*

All the improper integrals displayed in this Appendix are defined in the sense of principal value. For example, equation (68) is more precisely expressed as

$$\lim_{T_1, T_2 \to \infty} \int_{-T_1}^{T_1} \int_{-T_2}^{T_2} G(x_1)\overline{G(x_2)} p(x_1 , x_2) \, dx_1 \, dx_2$$
$$= \sum_{n=0}^{\infty} \frac{(R_x(\tau))^n}{n!} \left| \lim_{T \to \infty} \int_{-T}^{T} G(x)\left(\frac{d^n}{dx^n} p(x)\right) dx \right|^2 \cdot \qquad (69)$$

Since $G(x)$ is exponentially bounded and integrable over $(-T, T)$ for all $T \geqq 0$, it follows that the integrals in equation (69) exist and are bounded. Consequently, the improper integrals in equation (68) also exist. Moreover, one can show that

$$\lim_{T_1 \to \infty} \int_{-T_1}^{T_1} G(x_1)\overline{G(x_2)} p(x_1 , x_2) \, dx_1 = \int_{-\infty}^{\infty} G(x_1)\overline{G(x_2)} p(x_1 , x_2) \, dx_1$$

uniformly for $-\infty < x_2 < \infty$.

It follows that the double limit in equation (69) can be expressed as an iterated limit and further that the improper integral in equation (68) can be evaluated as an iterated improper double integral.

A.2 *Proof of Identity*

We first establish equation (68) for all functions $G(x)$ which are of bounded variation and satisfy

$$\int_{-\infty}^{\infty} | G(x) | \, dx < \infty .$$

In this case (following Rice,[2] Middleton[3] and Bennett[4]), $G(x)$ has a Fourier transform

$$\mathcal{G}(\omega) = \int_{-\infty}^{\infty} G(x) \exp(-j\omega x) \, dx \qquad (70)$$

such that for almost all x, $-\infty < x < \infty$,

$$G(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathcal{G}(\omega) \exp(j\omega x) \, d\omega. \qquad (71)$$

Using equation (71), we can write

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G(x_1)\overline{G(x_2)} p(x_1 , x_2) \, dx_1 \, dx_2$$
$$= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \mathcal{G}(\omega_1) \exp(j\omega_1 x_1) \, d\omega_1 \right) \left(\int_{-\infty}^{\infty} \overline{\mathcal{G}(\omega_2)} \exp(j\omega_2 x_2) \, d\omega_2 \right)$$
$$\cdot p(x_1 , x_2) \, dx_1 \, dx_2 . \qquad (72)$$

Since $G(x)$ is of exponential order, one can show that all combinations of the integrals in the right-hand side of equation (72) converge uniformly so that we are justified in interchanging the order of integration to obtain

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} & G(x_1)G(x_2)p(x_1\,,\,x_2)\,dx_1\,dx_2\\ &= \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathcal{G}(\omega_1)\,\mathcal{G}(\omega_2)\\ &\quad\cdot\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x_1\,,\,x_2)\exp\left[j(\omega_1x_1+\omega_2x_2)\right]dx_1\,dx_2\right)d\omega_1\,d\omega_2\,. \qquad (73)\end{aligned}$$

Using the fact that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x_1\,,\,x_2)\exp\left[j(\omega_1x_1+\omega_2x_2)\right]dx_1\,dx_2 = \exp\left\{-\frac{\Gamma^2}{2}\left(\omega_1^2+2r\omega_1\omega_2+\omega_2^2\right)\right\}$$

(see Ref. 5, pp. 30–35), and recalling that $\sum_{n=0}^{\infty} x^n/n!$ converges uniformly to exp (x) for $-\infty < x < \infty$, we obtain the following from equation (73):

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} & G(x_1)G(x_2)p(x_1\,,\,x_2)\,dx_1\,dx_2\\ &= \sum_{n=0}^{\infty}\frac{(R_x(\tau))^n}{n!}\left\{\frac{1}{2\pi}\int_{-\infty}^{\infty}(j\omega_1)^n\mathcal{G}(\omega_1)\exp\left(-\frac{\Gamma^2}{2}\,\omega_1^2\right)d\omega_1\right\}\\ &\quad\cdot\left\{\frac{1}{2\pi}\int_{-\infty}^{\infty}(j\omega_2)^n\mathcal{G}(\omega_2)\exp\left(-\frac{\Gamma^2}{2}\,\omega_2^2\right)d\omega_2\right\}. \qquad (74)\end{aligned}$$

An application of Parseval's theorem yields

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\mathcal{G}(\omega)(j\omega)^n\exp\left(-\frac{\Gamma^2}{2}\,\omega^2\right)d\omega = \int_{-\infty}^{\infty}G(-x)\left(\frac{d^n}{dx^n}\,p(x)\right)dx \qquad (75)$$

and a similar expression for $\overline{\mathcal{G}}$ and $\bar{G}$. Using equation (75) in equation (74), we obtain equation (68).

Now, we only assume that $G(x)$ is exponentially bounded and of bounded variation on finite intervals. For $k = 1, 2, \cdots$, define

$$G_k(x) = \begin{cases} G(x), & \text{for}\quad -k \leqq x \leqq k,\\ 0, & \text{otherwise.}\end{cases}$$

It follows that $G_k(x)$ is of bounded variation and that $\int_{-\infty}^{\infty} | G_k(x) | \, dx < \infty$ for $k = 1, 2, \cdots$ so that the above results imply that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G_k(x_1)\overline{G_k(x_2)}p(x_1 \, , x_2) \, dx_1 \, dx_2 = \sum_{n=0}^{\infty} \frac{(R_x(\tau))^n}{n!} \left| \int_{-\infty}^{\infty} G_k(x)\left(\frac{d^n}{dx^n} p(x)\right) dx \right|^2 \tag{76}$$

for $k = 1, 2, \cdots$.

We know that

$$\frac{d^n}{dx^n} p(x) = \frac{1}{\sqrt{2\pi\Gamma^2}} \left(-\frac{1}{\sqrt{2}\ \Gamma}\right)^n H_n\left(\frac{x}{\sqrt{2}\ \Gamma}\right) \exp\left(-\frac{x^2}{2\Gamma^2}\right)$$

[$H_n(z)$ is the Hermite polynomial of degree n], so that the series in equation (76) can be expressed as

$$S_k(r) = \sum_{n=0}^{\infty} a_{n,k}\, r^n \tag{77}$$

where

$$a_{n,k} = \frac{1}{2^n n!} \left| \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} G_k(\sqrt{2}\ \Gamma x) H_n(x) \exp(-x^2) \, dx \right|^2 . \tag{78}$$

Making use of the inequality[8]

$$| H_n(x) | < \xi \exp\left(\frac{x^2}{2}\right) 2^{n/2} \sqrt{n!}, \qquad \xi \approx 1.086435,$$

and the assumption

$$| G(x) | \leqq v \exp(u \,| x |), \qquad -\infty < x < \infty,$$

we can see that

$$0 \leqq a_{n,k} < \left(\frac{2\xi v}{\sqrt{\pi}} \int_0^{\infty} \exp(u\sqrt{2}\ \Gamma x - x^2) \, dx\right)^2 < \infty,$$

for all n and all k. Hence, when $| r | < 1$, Weierstrass' M-test implies that the series $S_k(r)$ converges uniformly for $k = 1, 2, \cdots$. Consequently, we have

$$\lim_{k \to \infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G_k(x_1)\overline{G_k(x_2)}p(x_1 \, , x_2) \, dx_1 \, dx_2 = \sum_{n=0}^{\infty} \frac{(R_x(\tau))^n}{n!} \left| \lim_{k \to \infty} \int_{-\infty}^{\infty} G_k(x)\left(\frac{d^n}{dx^n} p(x)\right) dx \right|^2 \tag{79}$$

for all τ such that $R_x(\tau)/\Gamma^2 \neq \pm 1$; i.e. for $| r | < 1$.

One can also show that

$$\lim_{k\to\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G_k(x_1)\overline{G_k(x_2)}p(x_1\ ,\ x_2)\ dx_1\ dx_2$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G(x_1)\overline{G(x_2)}p(x_1\ ,\ x_2)\ dx_1\ dx_2\ , \qquad (80)$$

and that

$$\lim_{k\to\infty} \int_{-\infty}^{\infty} G_k(x)\left(\frac{d^n}{dx^n}\, p(x)\right) dx = \int_{-\infty}^{\infty} G(x)\left(\frac{d^n}{dx^n}\, p(x)\right) dx. \qquad (81)$$

Substituting equations (80) and (81) into equation (79) yields the identity (68) for all τ such that $R_x(\tau)/\Gamma^2 \neq \pm 1$.

We saw above that our identity can also be expressed as

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G(x_1)G(x_2)p(x_1\ ,\ x_2)\ dx_1\ dx_2 = \sum_{n=0}^{\infty} a_n r^n, \qquad |\ r\ | < 1, \qquad (82)$$

where

$$a_n = \frac{1}{2^n n!} \left| \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} G(\sqrt{2}\ \Gamma x) H_n(x) \exp\,(-x^2)\ dx \right|^2 \cdot \qquad (83)$$

Since the left-hand side of equation (82) exists for $r = \pm 1$ and since $a_n \geqq 0$ for all n, it follows that equation (82) is also valid for $r = \pm 1$. (The proof of this result is a Tauberian theorem. For example, see Ref. 10, page 427, exercise 13–34.)

REFERENCES

1. Bennett, W. R., Curtis, H. E., and Rice, S. O., "Interchannel Interference in FM and PM Systems Under Noise Loading Conditions," B.S.T.J., *34*, No. 3 (May 1955), pp. 601–636.
2. Rice, S. O., "Mathematical Analysis of Random Noise," B.S.T.J., *23*, No. 3 (July 1944), pp. 282–332; *24*, No. 1 (January 1945), pp. 46–156.
3. Middleton, D., "Some General Results in the Theory of Noise through Nonlinear Devices," Quart. Applied Math., *5*, No. 4 (January 1948), pp. 445–498.
4. Bennett, W. R., "Methods of Solving Noise Problems," Proc. IRE, *44*, No. 5 (May 1956), pp. 609–638.
5. Rowe, H. E., *Signals and Noise in Communication Systems*, Princeton, N. J.: D. Van Nostrand Co., Inc., 1965, pp. 98–203.
6. Gradshteyn, I. S., and Ryzhik, I. M., *Tables of Integrals, Series and Products*, New York, N. Y.: Academic Press, 1965, pp. 1033–1035 and pp. 836–841.
7. VanVleck, J. H., "The Spectrum of Clipped Noise," Harvard University Radio Res. Laboratory, Report No. 51, Cambridge, Massachusetts, 1943.
8. Abramowitz, M., and Stegun, I. A., *Handbook of Mathematical Functions*, Washington, D. C.: National Bureau of Standards Applied Mathematics Series, *55*, pp. 297–329 and pp. 771–787.
9. Prabhu, V. K., and Rowe, H. E., "Spectral Density Bounds of a PM Wave," B.S.T.J., *48*, No. 3 (March 1969), pp. 769–811.
10. Apostol, T. M., *Mathematical Analysis*, Reading, Mass.: Addison-Wesley, December 1960.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

Television Coding Using Two-Dimensional Spatial Prediction

By D. J. CONNOR, R. F. W. PEASE and W. G. SCHOLES

(Mańuscript received September 30, 1970)

Three classes of differential coders have been built which exploit the two-dimensional spatial correlation present in television pictures. The three classes are distinguished according to the technique used to predict the value of the picture element (pel) to be coded. In the first class, the prediction is the value of the previous element along the scan line averaged with one or more elements in the previous line (of the same field). In the second class, the prediction is the sum of the value of the previous pel with the local element difference in the previous line. In the third class, the transmitter and receiver choose the prediction from either the previous element or some combination of the previous element with the elements in the previous line. The choice is based on differences measured between elements already coded. These are available both at the transmitter and receiver, so no extra information need be transmitted. In the absence of transmission errors, the pictures resulting from all three classes of coder are markedly improved over those in which only the previous element is used as a prediction. In particular, vertical and sloping edges are now well defined. The effect of a transmission error on a single frame of the received picture is scarcely visible in the first class of coder but is much more visible in the second and third classes.

I. INTRODUCTION

Intensity correlation in television pictures extends both horizontally and vertically.[1-3] For a number of good reasons, most practical predictive encoders have exploited only the horizontal correlation:[3]

(*i*) In a noninterlace system, access to the vertically adjacent picture element (pel) requires storage of one line of information (about 8,000 bits for broadcast TV or 2,000 bits for *Picture-*

phone® Service). In a 2 : 1 interlaced system, a field plus a line must be stored (10^6 bits for broadcast or 250,000 for *Picturephone* Service). If access to the vertically adjacent pel in the same field is desired, a line of storage is required.

(*ii*) Most television systems, including *Picturephone* Service, use an interlaced scan. Because of the storage problem, access to the vertically adjacent pel in the same field is most practical. However, this pel is about twice as far away (spatially) from the current pel as is the previous pel in the same scan line. The vertical correlation of intensity within a field is thus less than the horizontal correlation.

(*iii*) Statistical studies of television pictures suggested that for linear predictive encoding very little extra redundancy could be removed by exploiting the vertical correlation as well as the horizontal correlation.[2,3]

Given these reasons, why look at the use of vertically adjacent pels in various predictive encoding schemes? First, in recent years the cost of storage has fallen dramatically and hence the extra memory required for storing one line is not a serious objection. At present, a 2,000-bit line store using MOS shift registers costs about $600.00 and fits on one printed circuit board containing 40 dual-in-line packages. Future costs should be much lower (less than $100.00).

To counter the second and third points above, let us consider for a moment the operation of a differential quantizer. A differential quantizer predicts the value of the current pel by using information that has already been transmitted. The discrepancy between the uncoded value and the prediction is quantized and added to the prediction to form the coded value of the current pel.

Most practical differential quantizers which make use of the horizontal correlation of intensity produce pictures of generally acceptable quality, except for sharp vertical or diagonal edges. These are blurred and have an annoying "busyness." The problem is in the prediction. The differential quantizer has no way of "anticipating" a vertical edge. Consequently, the prediction it makes is grossly in error at these edges.

By making use of information from the previous line, it is possible to anticipate the vertical edges and most of the diagonal ones. To study this possibility, we have built and/or simulated on a computer a number of differential quantizers. These coders can be divided into three classes which are distinguished by the way in which the information from the previous line is used to "anticipate" or predict the value of the current pel. In Fig. 1 we give a labelled diagram of the pels near the current pel.

In the first class of coder, the prediction is the average of the value of the previous pel, A, and the value of one or more of the vertically adjacent pels, B, C, and D. We call this "averaged prediction." In the second class of coder, the prediction is the sum of A and the scaled horizontal element difference or slope between elements on the previous line, e.g., $C - B$ or $\frac{1}{2}$ $(D - B)$. This has been called "planar prediction."[4]

These first two classes of coders use linear prediction since the prediction is a linear sum of previously coded intensities. The third class of coder uses a nonlinear prediction in that the choice of prediction mode is signal-dependent. The prediction can be either A or some linear combination of A, B, C and/or D. The choice is dependent on the magnitude of intensity differences between the pels in the previously coded neighborhood of the current pel, X. This class of coder can be called "optional prediction." (R. E. Graham has proposed and simulated some coders of this type.[5])

In the following sections we describe the particular coders we have built, and give some examples of processed pictures. We also include a discussion of the effect of transmission errors for the various coding strategies.

II. DESCRIPTION

The scan format is similar to that used in *Picturephone* Service: there are 271 scan lines with a 2:1 interlace, 248 elements per line and the frame rate is 30 Hz.

A schematic of a general differential quantizer that uses information from the preceding line is given in Fig. 2. The quantizer characteristics are shown on Table I. The particular coders to be described below differ only in the implementation of the predictor. Note that with the use of sign prediction[6] these are three-bit coders.

For comparison, we chose as a reference an element-coder (i.e., the

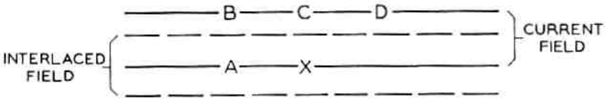


Fig. 1—Diagram of picture elements (pels) near the current pel.

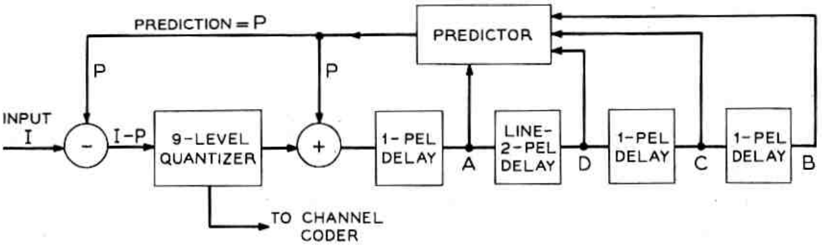


Fig. 2—Diagram of differential quantizer using information in previous line to aid prediction of current pel. The letters *A, B, C* and *D* refer to the pels whose positions are shown in Fig. 1.

value of pel *A* is used as a prediction) having the same quantizer characteristics as above. Although it might be argued that it would make more sense to optimize the quantizing scale for each type of prediction, we found that the subjectively optimum quantizing scale varied as much with picture content as with prediction strategy. For most experiments, therefore, we used one quantizing scale; a few results on varying the quantizing scale are, however, also described in the next section.

In the subjective tests, the $5\frac{1}{2}$ inch by 5 inch raster was viewed through a polarizing screen from a distance of 36 inches in a room having average office illumination (70 foot candles). Two scenes were used: (*i*) a model head, "Penelope," with a draped background (Fig. 3a); and (*ii*) a photographic slide, "Karen," (Fig. 3b).

III. AVERAGED PREDICTION

In this type of coder, the prediction is formed by adding the value of pel *A* to the value of some combination of the pels *B, C* and *D*. Three coders were built and tested; the predictions used were: (*i*) $(A + D)/2$, (*ii*) $(A + C)/2$, (*iii*) $[A + (C + D)/2]/2$. Figure 4a shows a diagram of the arrangement used to test coder (*i*).

In the absence of channel errors, the displayed picture at the trans-

TABLE I—ENCODER CHARACTERISTICS

Input Levels	±0–1	2–7	8–17	18–33	34–255
Output Levels	±0	4	11	25	42

mitter is the same as at the receiver[6] and hence only the transmitter loop was built for the real time tests.

In comparison with the reference coder, all three of the coders gave significantly improved rendition of the vertical and most of the sloping edges. Horizontal edges and those sloping slightly upward appear blurred in comparison with the same edges encoded by the element coder. However, it is a static blurring which is subjectively much less annoying than the "edge busyness" which occurs on some edges encoded by the element coder. The granular noise in pictures encoded with the averaged prediction coders is associated more with high vertical spatial frequencies than with high horizontal spatial frequencies. This effect is to be expected because the relatively large vertical step size will result in more frequent use of the larger (and less accurate) quantized difference signals in regions containing vertical detail. In picture areas with no directional properties, the noise level is about the same as with the element coder. Some observers, however, mentioned that it appeared to be a higher frequency noise.

In comparisons among the "averaged prediction" coders, (*i*) and (*iii*) appeared preferable to (*ii*) since they were able to adequately encode edges with smaller positive slopes than was (*ii*). This is understandable since the presence of pel D in their prediction means that such edges are anticipated. There was very little to choose between coders (*i*) and (*iii*) and so we carried out subjective tests in which eight unskilled and eight skilled observers were asked to state a preference between pictures obtained using coder (*i*) and the reference coder. With "Penelope," all eight of the skilled and seven of the eight unskilled observers preferred the picture given by coder (*i*). With

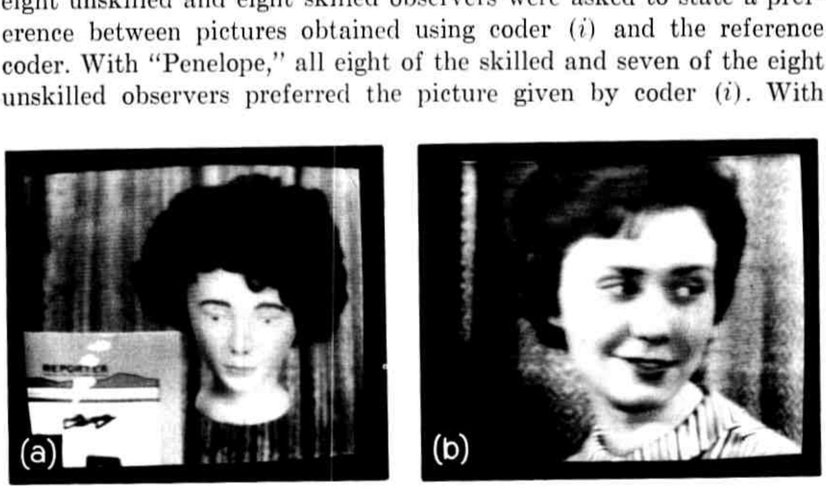
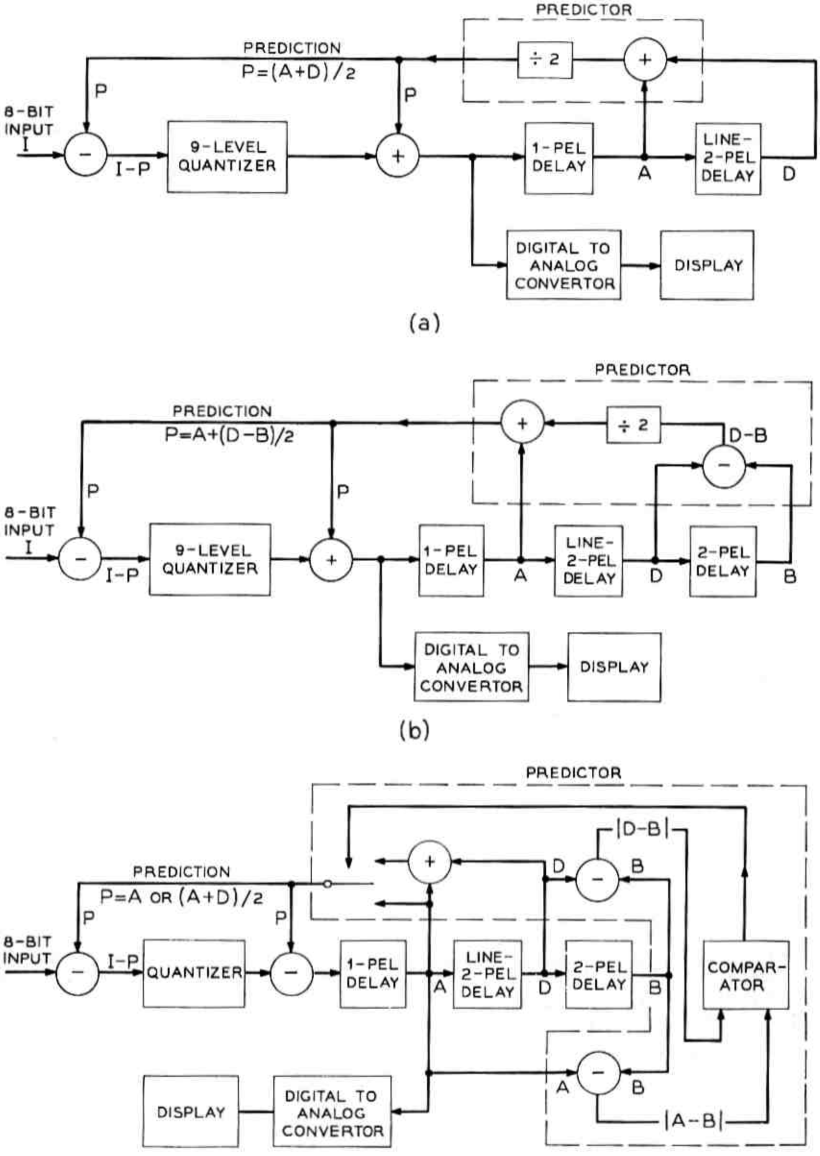


Fig. 3—Scenes viewed for evaluating coders, (a) Penelope and (b) Karen.



(a)

(b)

(c)

Fig. 4a—Diagram of apparatus used to evaluate coder (*i*) in which the value $(A + D)/2$ is used as a prediction of the current picture element. The displayed picture is equivalent to the received picture providing there are no channel errors.

Fig. 4b—Apparatus used to evaluate coder (*v*) in which $A + (D - B)/2$ is used as a prediction of the current pel. The position of pels A, B and D is shown in Fig. 1.

Fig. 4c—Apparatus used to test the optional prediction encoder. If $|A - B| > |D - B|$, A is used as the prediction. If $|A - B| \leqq |D - B|$, $(A + D)/2$ is used. The positions of A, B and D are shown on Fig. 1.

"Karen," all subjects preferred the picture given by coder (*i*). (The skilled observers all remarked on the better edge rendition of the vertical stripes in Karen's blouse.) As the quantizing scale is made coarser, there is improved rendition of vertical edges in the picture from the reference coder but the granular noise also increases. In this situation, the comparison between the two coders is closer, but coder (*i*) was still preferable.

The element coder and coders (*i*), (*ii*) and (*iii*) were simulated on a DDP-224 computer system. Only a single frame was processed so it was not possible to judge frame to frame noise. In the first set of experiments the magnitude of the difference between the original picture (coded as eight-bit PCM) and pictures from the above coders was displayed. Figures 5a and 5b show the discrepancy between the PCM coded picture and pictures coded with the element coder and coder (*i*) respectively. (Figures 5c and d are for coders discussed

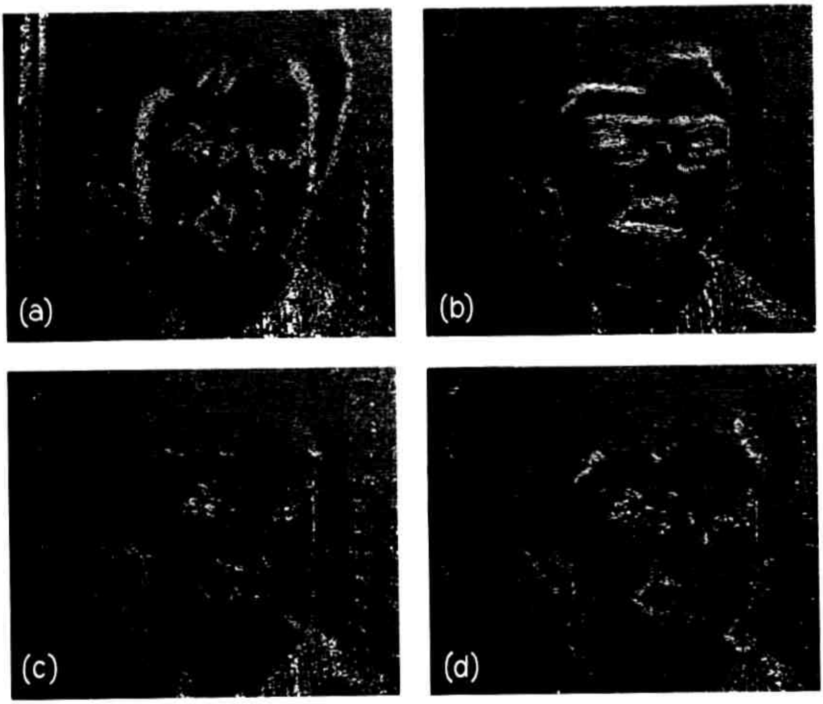


Fig. 5—Magnitude of the difference between a picture coded as 8-bit PCM and: (a) the reference coder; (b) coder (*i*) ($[A + D]/2$ is prediction); (c) coder (*v*) ($[A + (D - B)/2]$ is prediction); (d) optional prediction coder of Fig. 4c.

below.) As expected, the element coder showed differences along vertical and diagonal edges. The average prediction coders, on the other hand, gave differences chiefly along horizontal edges. Coder *(iii)* seemed to show the least discrepancy from the eight-bit coded picture. No subjective improvement was realized by choosing coder *(iii)* over coder *(i)* in the experimental model.

We also investigated on the computer the effect of transmission errors on a single frame of the received picture. It has been suggested that because the average prediction coders made use of information from the previous line as well as from previous pels on the current line, the effect of an error on subsequent lines would be serious. However, because the prediction is an average, the effect of the error decays spatially. For example, consider the effects of an error of 64 quantum levels in coder *(i)*. This error results in the array of intensity errors shown in Table II. Note that the effect of the error decays rapidly both horizontally and diagonally, but more slowly vertically. Also, the spatial derivative of the error pattern is small except near the original error. This fact should help reduce the visibility of the error pattern. In fact, the pictures generated on the computer showed that even serious errors (e.g., a change of sign on the outer level) resulted in a pattern that was scarcely visible.

Contrast the visibility of the errors in Figs. 6a and b. The picture in Fig. 6a was coded with the reference coder. The upper, dark line was caused by a positive-to-negative sign change in an outer quantizer level. The lower light line results from a negative-to-positive sign change in the third quantizer level. The identical type of errors are present in the picture in Fig. 6b which was coded with coder *(i)*. The error in the outer quantizer level is visible as a small black dot at the outside corner of the left-hand eye. The error in the third quantizer level appears as a white streak toward the right end of the upper lip. In any event, the errors are far less visible than the equivalent errors in Fig. 6a.

IV. PLANAR PREDICTION ENCODERS

Two coders were built and tested. The predictions used were *(iv)* $A + (C - B)$, and *(v)* $A + (D - B)/2$. Figure 4(b) shows a schematic for coder *(v)*.

Both coders showed improvement over the reference coder, but not over the coders in the previous section. In particular, the planar prediction encoders gave rise to "edge-busyness" on edges sloping upward

TABLE II—INTENSITY ERRORS FOR AVERAGE PREDICTION ENCODER*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Horizontal Distance in Pels	→x
1								64	32	16	8	4	2	1			
2																	
3							32	32	24	16	10	6	3	1			
4																	
5						16	24	24	20	15	10	7	4	2	1		
6																	
7					8	16	20	20	18	15	11	8	5	3	1		
8																	
9				4	10	15	18	18	16	14	11	8	6	3	2	1	
10																	
11			1	5	10	14	16	16	15	13	11	8	5	3	2	1	
12																	
13			3	7	10	14	15	15	14	12	10	7	5	3	2	1	
14																	
15		2	4	7	11	13	14	14	13	11	9	7	5	4	2	1	
16																	
17	1	2	5	8	10	12	13	13	12	10	9	7	6	4	3	2	1
Vertical 18																	
Distance in Lines 19	1	3	6	8	10	12	12	12	11	10	8	7	6	4	3	2	1

y ↓

* The pattern of intensity errors that results from an error of 64 levels at pel 8,1 ($x = 8$, $y = 1$) when $(A + D)/2$ is used as a prediction of the current pel. The intensity of the error decays spatially and, perhaps more importantly, there is no abrupt large spatial difference in intensity except near pel 8, 1.

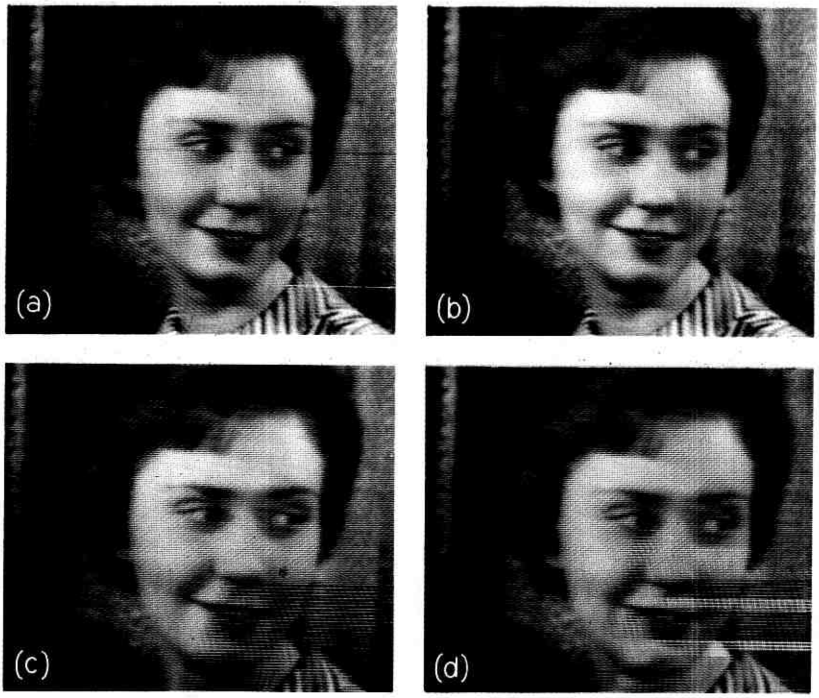


Fig. 6—Decoded pictures showing the effect of intensity errors resulting from channel errors. Figures (a) and (b) each contain two error patterns resulting from errors of 84 and 50 levels (out of 255 levels). (a) Derived using the reference coder (differential quantizer using A as the prediction); (b) results from the use of coder (i) ($[A + D]/2$ is prediction); (c) results from the use of coder (v) ($[A + (D - B)/2]$ is prediction) when there is a single intensity error of 84 levels; (d) results from the use of the optional predictor (see Fig. 4c) when there is a single intensity error of 50 levels. Copies of the original prints may be obtained from the authors.

to the left. In addition, coder (iv) gave rise to the same problem on edges sloping upward to the right.

Consider an edge passing to the left of B in Fig. 1 and between A and X so that $(C - B) = (D - B) = 0$. The values of $A + (C - B)$ and $A + (D - B)/2$ will both be poor predictions of X (the same as the element coder but worse than $(A + D)/2$.) Also, if an edge passes between A and X and between D and C, then $A + (C - B)$ will be a poor prediction of X.

The discrepancy between the picture coded as eight-bit PCM and that coded with coder (v) is shown in Fig. 5c. Although on the average, the picture shows less discrepancies than the equivalent

TABLE III—INTENSITY ERRORS FOR PLANER PREDICTIVE ENCODER*

							Horizontal Distance in Pels →											Start of Blanked Region ↓	
							210	211	212	213	214	215	216	217	218	219	220	221	222	223
1							0	0	64	64	64	64	64	64	64	64	64	64	64	0
3							0	32	64	64	64	64	64	64	64	64	64	64	32	0
5						0	16	48	64	64	64	64	64	64	64	64	64	48	16	0
7					0	8	32	56	64	64	64	64	64	64	64	64	56	32	8	0
9					4	20	44	60	64	64	64	64	64	64	64	60	44	20	4	0
11				2	12	32	52	62	64	64	64	64	64	64	62	52	32	12	2	0
13			1	7	22	42	57	63	64	64	64	64	64	63	57	42	22	7	1	0
15		0	4	14	32	50	61	64	64	64	64	64	64	61	50	32	14	4	0	0
17	0	2	9	23	41	56	63	64	64	64	64	64	63	56	41	23	9	2	0	0
19	1	6	17	33	49	60	64	64	64	64	64	64	60	49	33	17	6	1	0	0
21	3	11	25	4	54	62	64	64	64	64	64	62	54	41	25	11	3	0	0	0

y ↓ Vertical Distance in Lines
(unlike Table II, the interlaced lines are not shown)

* Pattern of intensity errors resulting from an error of 64 levels. At pel 212, 1 when A + (D − B)/2 is used as a prediction of the current pel. Note that the intensity error is constant to the right of pel 212, 1 until the start of the blanked portion of the line (pel 223, 1) when the accumulated value is forced to zero. On successive lines this (correct) reset value is partially used in the prediction near the right-hand side and so the error is reduced.

picture for coder (i), the discrepancies that do occur lie on vertical or diagonal edges sloping upward to the left. It is on just these edges that the "edge busyness" occurs.

The effect of a transmission error is an area of uniformly altered brightness as shown in Fig. 6c. How this pattern arises is illustrated in Table III where we have assumed an error of +64 levels in the element in the top row of numbers.

In view of the increased visibility of error patterns and the absence of a significant improvement in picture quality over the averaged prediction encoder, we saw no overall advantage in the planar prediction encoders for fixed length codes.

V. OPTIONAL PREDICTION ENCODERS

A number of varieties of optional prediction encoders were investigated. The one examined most closely uses either A or $(A + D)/2$ as a prediction. If the difference $|A - B|$ is greater than the difference $|D - B|$, A is used as the prediction, otherwise $(A + D)/2$ is used (Fig. 4c). These differences are made up of pels already coded and available (in the absence of transmission errors) at both the receiver and transmitter. Thus, no extra information need be sent to the receiver to indicate the prediction used by the transmitter.

The pictures obtained showed less granular noise than those from any of the other encoders. They showed little deviation (Fig. 5d) from the pictures coded as eight-bit PCM. Most edges are reproduced well, but those sloping at about 45 degrees upward to the left showed some local slope overload due to a mistaken choice of prediction. Unfortunately, a mistaken choice on one line tended to lead to similar mistakes on the next line and hence such clusters of mistakes tended to be visible. Different strategies for making the choice altered the slope of the edges on which the trouble occurred. So far only simple strategies have been tried; more complicated ones may well eliminate the problem. Obviously, there is a wide variety of techniques to be explored. For instance, there can be more than two options. In view of the low granular noise of the encoded pictures, such exploration is well worthwhile.

The effect of a transmission error on the decoded picture can be catastrophic. The receiver can lose track of the prediction option selected by the transmitter. Indeed, the effect of a large error tends to make the receiver choose incorrectly. This is illustrated by supposing the pel D has an error of +64 levels. This generates a large spatial change; $|D - B|$ is large and hence $(A + D)/2$ will probably be used as the prediction. This propagates the error to the next line. If A should have been used as a prediction instead of $(A + D)/2$, then the differences are added to the wrong prediction at the receiver. The effect on the decoded picture can be catastrophic as is illustrated in Fig. 6d.

VI. DISCUSSION

Because of the improved picture quality and the insensitivity to transmission errors compared with a standard differential quantizer, coders (*i*) and (*iii*) seem at present to be the most practical for intrafield coding of pictures for transmission. This is especially true when fixed length codes are to be used. We have still to determine the best quantizing characteristics and the best weighting of pel A with pels on the previous line to form a prediction.

Except for a few poorly predicted pels, the optional prediction encoders give a better rendering of the scenes described than do the fixed prediction encoders [e.g., coders (*i*) through (*v*)]. Hence we are continuing to study ways of removing the last few flaws in the coded picture. The sensitivity of this type of coder to transmission errors may be a serious problem. We are considering a number of methods for reducing this sensitivity.

REFERENCES

1. Kretzmer, E. R., "Statistics of Television Signals," B.S.T.J., *31,* No. 4 (July 1952), pp. 751–763.
2. O'Neal, J. B., Jr., "Predictive Quantizing Systems (Differential Pulse Code Modulation) for the Transmission of Television Signals," B.S.T.J., *45,* No. 5 (May–June 1966), pp. 689–721.
3. Limb, J. O., "Entropy of Quantized Television Signals," Proc. IEE, *115,* No. 1 (January 1968), pp. 16–20.
4. Harrison, C. W., "Experiments with Linear Prediction in Television," B.S.T.J., *31,* No. 4 (July 1952), pp. 764–784.
5. Graham, R. E., "Predictive Quantizing of Television Signals," IRE WESCON Record, *2,* Part 4, (1958), pp. 147–157.
6. Limb, J. O., and F. W. Mounts, "A Digital Differential Quantizer for Television," B.S.T.J., *48,* No. 7 (September 1969), pp. 2583–2599.

Copyright © 1971 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 50, No. 3, March, 1971
Printed in U.S.A.

On the Control of Linear Multiple Input-Output Systems*

By B. GOPINATH

(Manuscript received July 1, 1970)

The control of linear time-invariant systems is one of the most basic problems of modern automatic control theory. Although "optimal controllers" which minimize certain costs associated with control can be determined, in most applications "simple controllers" suffice, and are often more desirable. The criteria by which these simple controllers are designed are closely related to the problem of assigning the eigenvalues of the fundamental matrix (i.e., the poles of the system) to arbitrary but specified locations. This paper presents an approach to the design of such control systems. Our approach does not involve computing complicated canonical forms, as do some previous methods, and at the same time generalizes easily to multi-input-output systems. A simple solution of the problem of designing feedback control systems with a minimum number of dynamic elements is also presented.

I. INTRODUCTION

In recent years there has been a considerable amount of interest in the problem of designing controllers for linear systems. Although most of the theoretical interest has centered around optimal control approaches, it is generally known that in most standard control systems, simple and usually nonoptimal controllers suffice. One of the oldest problems of control theory is that of stabilizing a linear control system by using feedback (see Fig. 1). Although this problem has been solved in the single input-output case by many people, one of the first clear statements was that by D. G. Luenberger.[1] In the case of multiple input-outputs, elegant solutions are of recent origin (see Ref. 2). Almost all of the published solutions resort to canonical forms, and in the multiple input-output case are not convenient to work

* A talk based on this paper was presented at the Second Assilomar Conference on Circuits and System Theory, 1968.

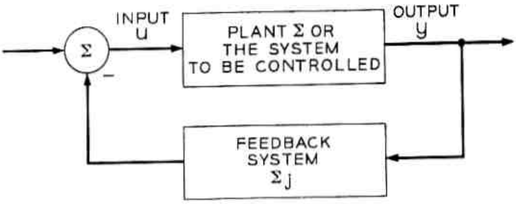


Fig. 1—Model of controller.

with. Also, in almost all cases, since the system is often described in terms of variables that are of direct interest, a transformation to canonical form is inconvenient.

In this memo we present a solution of the problem including the problem of designing controllers of minimal dynamic order (i.e., a controller using the least number of dynamic elements). The present solution does not resort to the use of canonical forms for the design. This approach also helps to systematically exploit the additional freedom that is obtainable due to the multiplicity of the inputs and outputs. In fact there is no previous solution known to the author which solves the problem of designing "minimal order" observers without resorting to complicated canonical forms.

We begin by introducing certain preliminaries and establishing the notation, and then solving the problem of designing controllers and observers in Sections III and IV. In Section V, the problem of designing controllers of low dynamic order is solved.

After this paper was written, the author became aware of the paper by W. M. Wonham.[3] Wonham derives Lemma 1 in the following section. The proof given in Wonham however uses the theory of minimal polynomials, as compared to the proof given in the following section which uses only the concept of linear spaces. Wonham himself has commented in his paper that an abstract proof of his results would be very worthwhile. The author feels that the proof given in the following section is an abstract version. The results of Sections III through V are not to be found in Wonham's paper.

II. LINEAR TIME-INVARIANT SYSTEMS AND THE PROBLEM OF CONTROL

The following definitions contain certain undefined but generally understood concepts such as dynamical system, etc. For a more detailed discussion of these ideas, the reader is referred to Ref. 4.

2.1 *Linear Time-Invariant System*

A linear time-invariant system Σ is a dynamical system governed by the following equation.

$$\dot{x}(t) = Fx(t) + Gu(t), \tag{1}$$

$$y(t) = Hx(t), \tag{2}$$

where $x(t) \in E^n$ is the state of Σ and $u(t) \in E^m$ and $y(t) \in E^p$ are the inputs and outputs of Σ respectively. F, G, and H are $n \times n$, $n \times m$ and $p \times n$ matrices respectively, and are *independent of time t.*

A "system" hereinafter shall denote a linear time-invariant system for brevity.

2.1.1 *Cyclic System*

Σ is cyclic if there exists $x \in E^n$ such that the matrix $[x \; Fx \cdots F^{n-1}x]$ is nonsingular.

2.1.2 *Complete Controllability*

Σ is completely controllable if the rank of $[G \, FG \cdots F^{n-1}G]$ is n. See Ref. 4 for details.

2.1.3 *Complete Observability*

Σ is completely observable if the rank of $[H' \, F'H' \cdots F'^{n-1}H]$ is n. Finally Σ is ordinary if

(*i*) Σ is cyclic; and
(*ii*) Σ is completely controllable and observable.

Most systems ordinarily dealt with are cyclic, because, as will be shown in this section, the condition of not being cyclic is caused by having two identical subsystems embedded in one system and yet completely decoupled from each other. Hence it is a singular situation in the sense that whenever a system is completely reachable and completely observable, a slight amount of feedback can make the system cyclic (see Ref. 5).

It is very interesting to note that most theorems to be given in this paper are dependent on a simple and basic property of linear spaces. This property is stated as the following lemma.

Lemma 1: Let $\mathcal{S}_i$, $i = 1 \cdots n$, be n distinct linear subspaces of a linear space. Let $\mathcal{L}$ be a linear space contained in the set union of $\mathcal{S}_i$,

$i = 1 \cdots n$. Then

$$\mathcal{L} \subseteq \mathcal{S}_j \textit{ for some } j \in \{1, 2, \cdots n\}, \tag{3}$$

where $\subseteq$ denotes "contained in".

Proof: The proof will be based on the principle of finite induction. The lemma is obviously true for $n = 1$. Now suppose the lemma is true for all $n < n_0$; i.e., suppose that, given $\mathcal{S}_i$, $i = 1, 2, \cdots, n < n_0$, then $\mathcal{L} \subseteq \bigcup_{i \leq n} \mathcal{S}_i$ implies that $\mathcal{L} \subseteq \mathcal{S}_j$ for some $j \leqq n$. It will then be proved that the lemma is true for $n = n_0$, which will complete the proof of the lemma by finite induction.

Assume $\mathcal{L}$ is not contained in the set union of any m ($m < n_0$) of the $\mathcal{S}_i$'s, for if it is so contained, then the lemma trivially holds (from previous paragraph). Therefore there exists n_0 vectors x_i such that

$$x_i \in \mathcal{L}, \quad x_i \in \mathcal{S}_i \quad \text{and} \quad x_i \notin \mathcal{S}_j \quad \text{if} \quad i \neq j. \tag{4}$$

Consider now any two of these n_0 vectors x_i, x_j and

$$x_i + \alpha x_j, \qquad i \neq j, \qquad \alpha \in R. \tag{5}$$

(The set of real numbers).

Since $\mathcal{L}$ is linear, $x_i + \alpha x_j \in \mathcal{L}\ \forall\ \alpha$; also, since $\mathcal{L} \subseteq \bigcup_{i \leq n_0} \mathcal{S}_i$

$$x_i + \alpha x_j \in \mathcal{S}_s \quad \text{for some} \quad s(\alpha) \in \{1, 2, \cdots, n_0\}. \tag{6}$$

However, since there are only a finite number of $\mathcal{S}_i$'s while α can assume any values from the uncountably infinite set R, there exists an "s" such that for at least two distinct values of α, namely α_1 and α_2,

$$x_i + \alpha_1 x_j \qquad \text{and} \qquad x_i + \alpha_2 x_j \in \mathcal{S}_s .$$

But this implies that

$$(\alpha_1 - \alpha_2)x_j \in \mathcal{S}_s \text{ since } \mathcal{S}_s \text{ is linear;} \tag{7}$$

i.e.,

$$x_j \in \mathcal{S}_s \text{ since } \alpha_1 \neq \alpha_2 \text{ and } \mathcal{S}_s \text{ are linear.} \tag{8}$$

Therefore $s = j$ by (4), i.e., $x_i + \alpha_1 x_j \in \mathcal{S}_j$, whence $x_i \in \mathcal{S}_j$ since $x_j \in \mathcal{S}_j$ and $\mathcal{S}_j$ is linear. Once again using (4), $x_i \in \mathcal{S}_j \Rightarrow i = j$ which contradicts equation (5). Q.E.D.

Note: As can easily be seen, Lemma 1 does not hold in general for an uncountable union of linear spaces.

Definition: A square matrix F has simple structure if and only if for each eigenvalue λ_i, of F there exists one and only one eigenvector e_i. (In other words, F is simple if no two uncoupled Jorden blocks in the canonical form have the same eigenvalue.)

Note: All the eigenvectors are assumed to be normalized such that the first nonzero component is $+1$.

Lemma 2: The statement that the system Σ is cyclic implies that the square matrix F in equation (1) has simple structure.

Proof: Suppose there exists two eigenvectors e_1 and e_2 of F corresponding to the eigenvalue λ. Then $\exists$ two eigenvectors, d_1 and d_2 of the matrix F' corresponding to $\bar{\lambda}$, where F' is the conjugate transpose of F. Let x be any vector in E^n. Then suppose y is the projection of x in $R(d_1, d_2)$, the subspace spanned by d_1 and d_2. Let

$$z \in R(d_1, d_2) \quad \text{such that} \quad (z, y)^* = 0, \quad z \neq 0. \tag{9}$$

Then since $((x - y), z) = 0$ because $z \in R(d_1, d_2)$, it follows that $(x, z) = 0$ by equation (9) and since $z \in R(d_1, d_2)$, $z = \alpha_1 d_1 + \alpha_2 d_2$.

Therefore

$$\begin{aligned}(z, F^r x) &= ((\alpha_1 d_1 + \alpha_2 d_2), F^r x),\\ &= \bar{\lambda}^r((\alpha_1 d_1 + \alpha_2 d_2), x),\\ &= \bar{\lambda}^r(z, x) = 0 \qquad \forall \quad r.\end{aligned}$$

Therefore the proof is complete. Q.E.D.

Definition: A subspace $\mathfrak{L}$ of E^n is an invariant subspace of F if $x \in \mathfrak{L} \Rightarrow Fx \in \mathfrak{L}$. $\mathcal{I}_s(F)$ denotes an s-dimensional invariant subspace of F.

Lemma 3: The statement that Σ is ordinary implies that $\exists$ an $\alpha \in E^m$ and $\beta \in E^p$, such that

$$\rho([F : G\alpha]) = \rho([F' : H'\beta]) = n^\dagger.$$

Proof: Notice that the number of invariant subspaces of F are finite, since the number of one-dimensional invariant subspaces are finite. [This follows from the familiar structure of invariant subspaces, see equation (6).] Suppose $x \in R(G)$, the space spanned by the columns of G and

$$\rho([F : x]) = s(x) < n.$$

* (z, y) is inner product of z and y.

† The matrix $[F : A]$ will in general denote the matrix $[A \ FA \ \cdots \ F^{n-1}A]$, and $\rho([F : A])$ denotes the rank of $[F : A]$.

Therefore $x \in \mathcal{R}(G)$ belongs to some $\mathcal{I}_s(F)$, $s < n$. Therefore, from Lemma 1 $\mathcal{R}(G) \subseteq \mathcal{I}_s(F)$ for some $s < n$, which contradicts $\rho([F:G]) = n$. Therefore $\exists$ an $\alpha \in E^m$, such that $\rho([F:G\alpha]) = n$. Similarly the other case. Q.E.D.

The above lemma shows that $\exists$ a single input-output system corresponding to every ordinary system, such that the controllability and observability of the new system is implied by that of the old system with multiple inputs and outputs. Lemma 4 shows that the weighting vectors α and β could almost be any vector in E^m and E^p respectively.

*Lemma 4: The statement that Σ is ordinary, that $\alpha \in E^m$ and $\beta \in E^p$, implies that $\rho([F:G\alpha]) = \rho([F':H'\beta]) = n$ almost surely.**

Proof: Notice the determinant of $[F:G\alpha]$ is a polynomial in α, and by Lemma 3 we have shown is nonzero for at least one α. If the distribution of α does not allow nonzero probability to any surface of dimension $< m$ then

$$\text{Probability that } \rho([F:G\alpha]) = n \text{ is } 1. \quad \text{Q.E.D.}$$

The stability of Σ, and the transient response of Σ are generally characterized by the eigenvalues of F, which in turn are given by the characteristic polynomial $\chi(F)$. Hence loosely by dynamics of Σ we mean the characteristic polynomial or the eigenvalues of F.

Given that a system Σ models a "plant" to be controlled the problem that we will consider is that of designing another system Σ_f such that the resultant "closed loop" system has arbitrary dynamics.

III. THE DESIGN OF CONTROLLERS

In order to motivate the nature of the problem in Section IV, we shall first solve the so-called control problem which essentially is a simplified version of the problem postulated in Section II. The H matrix in equation (2) is now assumed to be the "n" dimensional identity denoted by I_n, in other words, the complete state of Σ is available for measurement. In this case, we show that we need only feed back a certain linear function of the state "x" to achieve any given dynamics for the closed loop system.

The problem formally reduces to the following.

Given a plant Σ described by equations (1) and (2) with H in (2)

* α and β are chosen from any joint distribution in E^m and E^p respectively, which does not assign nonzero probability to a surface of dimension less than m and p, respectively.

replaced by I_n, it is required to find an $m \times n$ matrix K referred to in the following as the feedback gain such that the resultant system has the prescribed characteristic polynomial (10)

$$s^n + \sum_{i=1}^{n} \gamma_i s^{n-i}. \tag{10}$$

Let the characteristic polynomial $\chi(F)$ be

$$\chi(F) = s^n + \sum_{i=1}^{n} a_i s^{n-i}. \tag{11}$$

Then from Fig. 2 the problem reduces to finding K such that

$$\chi(F - GK) = s^n + \sum_{i=1}^{n} \gamma_i s^{n-i} \tag{12}$$

since the new differential equation is $\dot{x} = (F - GK)x + Gu$. The' solution is contained in the following theorem.

Theorem 1: If a system Σ *is cyclic and completely controllable then with* $\gamma_1, \gamma_2, \cdots, \gamma_n$ *real constants,*

(1)
$$\chi(F - GK) = s^n + \sum_{i=1}^{n} \gamma_i s^{n-i} \tag{13}$$

for any K *of rank one satisfying*

$$\begin{aligned} \gamma_1 &= a_1 + \operatorname{tr}(GK) \\ &\vdots \\ \gamma_n &= a_n + a_{n+1}\operatorname{tr}(GK) + \cdots + \operatorname{tr}(F^{n-1}GK). \end{aligned} \tag{14}$$

(2) *Moreover there exists at least one* K *satisfying equation* (14).

Proof: The new characteristic polynomial with feedback is

$$\chi(F - GK) = \det(sI - F + GK),$$

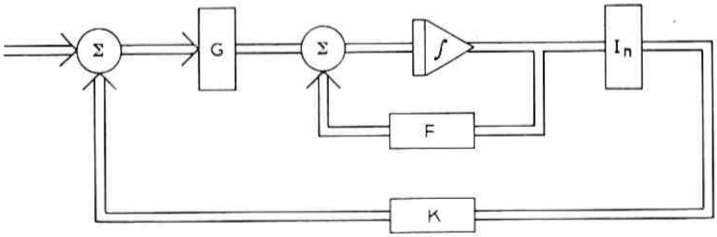


Fig. 2—Feedback control when the state is known.

i.e.,

$$= \det\,[(sI - F)(I + (sI - F)^{-1}GK)], \tag{15}$$

$$= \chi(F)\det\,(I + (sI - F)^{-1}GK).$$

Since K is of rank one, it follows that $(sI - F)^{-1}GK$ has rank one; therefore

$$\det\,(I + (sI - F)^{-1}GK) = 1 + \operatorname{tr}\,(sI - F)^{-1}GK \tag{16}$$

where tr (A) denotes the sum of the diagonal elements of A. Therefore from equation (15)

$$\chi(F = GK) = \chi(F)[1 + \operatorname{tr}\,(sI - F)^{-1}GK], \tag{17}$$

i.e.,

$$\chi(F - GK) = \chi(F) + \operatorname{tr}\,[\chi(F)(sI - F)^{-1}GK] \tag{18}$$

which since

$$(sI - F)^{-1} = \sum_{i=0}^{\infty} F^i s^{-(i+1)}$$

outside an appropriate region in the complex plane (see Ref. 6), becomes

$$\chi(F - GK) = \chi(F) + \operatorname{tr}\left[\chi(F)\sum_{i=0}^{\infty}[F^i s^{-(i+1)}GK]\right]\cdot \quad \cdot \tag{19}$$

Now using the Cayley-Hamilton theorem (see Ref. 6), i.e., using the fact that

$$F^n + \sum_{i=1}^{n} a_i F^{n-i} = 0. \tag{20}$$

and equating coefficients of equation (19), we have that the coefficient of s^{n-j} on R. H. S. of equation (19) is (using $a_0 \triangleq \gamma_0 \triangleq 1$)

$$= 0, \qquad j > n; \tag{21}$$

$$\gamma_j = a_j + a_{j-1}\operatorname{tr} GK + \cdots + \operatorname{tr} F^{j-1}GK, \qquad n \geq j \geq 1; \tag{22}$$

$$= 0, \qquad j < 0. \tag{23}$$

Therefore

$$\begin{aligned} \gamma_1 &= a_1 + \operatorname{tr} GK, \\ \gamma_2 &= a_2 + a_1 \operatorname{tr} GK + \operatorname{tr} FGK, \\ &\vdots \\ \gamma_n &= a_n + \cdots + \operatorname{tr} F^{n-1}GK. \end{aligned} \tag{24}$$

This proves that if there exists K of rank 1 such that equation (14) is satisfied, then K satisfies equation (24), and that if K satisfies equation (24), then equation (12) is satisfied.

Let

$$[\gamma_1, \cdots, \gamma_n] \triangleq \gamma', \qquad (25)$$
$$[a_1, \cdots, a_n] \triangleq a',$$

where $'$ denotes the transpose. We have, rewriting equation (24),

$$\gamma = a + \begin{bmatrix} 1 & 0 & \cdots & 0 \\ a_1 & 1 & \cdots & 0 \\ \vdots & a_1 & \cdots & \cdot \\ a_{n-1} & a_{n-2} & \cdots & a_1 1 \end{bmatrix} \begin{bmatrix} \operatorname{tr} GK \\ \operatorname{tr} FGK \\ \vdots \\ \operatorname{tr} F^{n-1}GK \end{bmatrix}. \qquad (26)$$

Let

$$A \triangleq \begin{bmatrix} 1 & 0 & \cdots 0 & 0 \\ a_1 & 1 & \cdots 0 & 0 \\ \vdots & & \vdots & \vdots \\ a_{n-2} & a_{n-3} & \cdots 1 & \cdot \\ a_{n-1} & a_{n-2} & \cdots a_1 & 1 \end{bmatrix}. \qquad (27)$$

Notice here that A^{-1} always exists and can be evaluated easily.

$$\begin{bmatrix} \operatorname{tr} GK \\ \operatorname{tr} FGK \\ \vdots \\ \operatorname{tr} F^{n-1}GK \end{bmatrix} = A^{-1}(\gamma - a). \qquad (28)$$

Now we assume $K = \alpha k'$ (α and k are $m \times 1$ and $n \times 1$ matrices respectively), such that K is of rank 1 then equation (28) becomes

$$\begin{bmatrix} \operatorname{tr} G\alpha k' \\ \operatorname{tr} FG\alpha k' \\ \vdots \\ \operatorname{tr} F^{n-1}G\alpha k' \end{bmatrix} = A^{-1}(\gamma - a). \qquad (29)$$

Since tr $F^iG\alpha k' = \alpha' G' F'^i k$, equation (29) becomes

$$\begin{bmatrix} \alpha' G' \\ \alpha' G' F' \\ \vdots \\ \alpha' G' F'^{n-1} \end{bmatrix} k = A^{-1}(\gamma = a). \tag{30}$$

But from Lemma 4 it is clear that $\rho([F \; G\alpha]) = n$ for almost all α. Therefore it follows that equation (30) has a unique solution for almost any α, and this completes the proof.

Therefore from the proof of the above theorem, it is easy to see how we can find the required gain matrix. Equation (28) is linear in the elements of K. The freedom in the multi input-output case is essentially one of picking α. Almost any solution of equation (28) which has rank 1 will do the job. Notice that restricting K to be of rank 1 also helps to reduce the number of amplifiers to implement the system, for K then can be realized by $(m + n - 1)$ amplifiers instead of (mn).

IV. DESIGN OF OBSERVERS

In Section III we saw how a gain matrix K could be computed for the system Σ with $H = I_n$. However when $H \neq I_n$, the state 'n' of Σ is not directly observable and an observer to estimate the state has to be designed. It will become clear that Theorem 1 gives the solution to this problem also. The solution consists of designing a linear system Σ_0 which is constructed in such a way that its state $\hat{x}$ can easily be observed, and such that the state of Σ_0 tends to the state of Σ as "rapidly as desired." (The meaning will become clear in the following.)

The system Σ_0 will consist of a model of Σ driven by an input which is equal to the sum of the inputs to a weighted error term which is the difference between the state of Σ and that of Σ_0.

Let Σ_0 be defined by

$$\dot{\hat{x}} = F\hat{x} + LH(x - \hat{x}) + Gu. \tag{31}$$

Let the error $\tilde{x} \triangleq x - \hat{x}$. Then equations (1) and (31) imply

$$\dot{\tilde{x}} = F\tilde{x} - LH\tilde{x}. \tag{32}$$

Now we would like $\tilde{x}$ to decrease to zero according to some dynamics in the sense that $\chi(F - LH)$ should be some prescribed polynomial. It is obvious that once again the problem is to find an L such that

$\chi(F - LH)$ is a prescribed polynomial which by Theorem 1 again is easily done by solving

$$\begin{aligned}\gamma_1 &= a_1 + \operatorname{tr} LH,\\ \gamma_2 &= a_2 + a_1 \operatorname{tr} LH + \cdots + \operatorname{tr} LHF,\\ &\vdots\\ \gamma_n &= a_n + a_{n-1} \operatorname{tr} LH + \cdots + \operatorname{tr} HF^{n-1},\end{aligned} \tag{33}$$

where γ_i are the coefficients of the prescribed polynomial. Hence now the complete solution can be stated as follows.

Step 1: By solving equation (33), construct an observer with dynamics such that it is sufficiently fast compared with the plant.

Step 2: By solving equation (26), construct the controller as in Section III to have the required closed loop dynamics.

Step 3: Cascade the observer and the controller as in Fig. 3.
We can show that the characteristic polynomial of the entire closed loop system of Fig. 3 is actually $\chi(F - GK)\chi(F - LH)$.

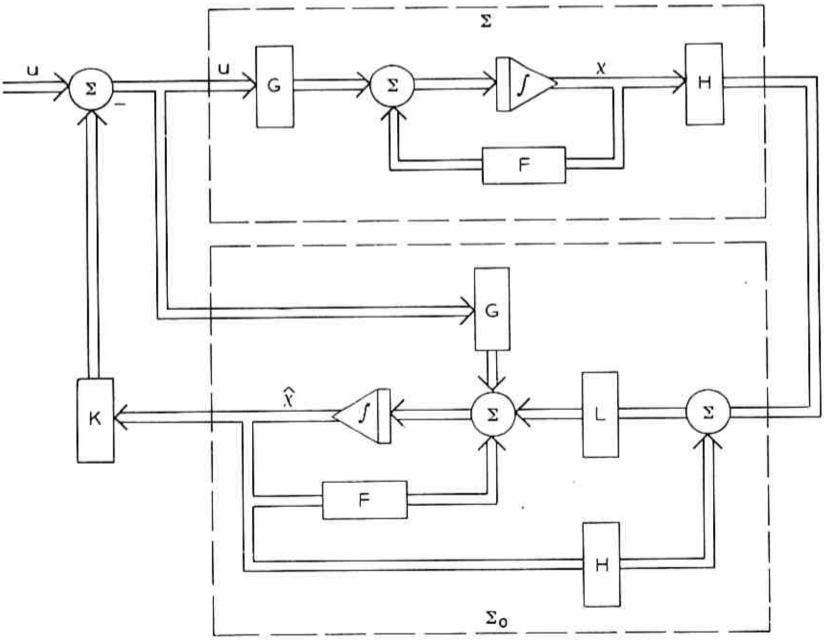


Fig. 3—The structure of a complete "controller."

The differential equation for the complete system of Fig. 3 is as follows.

$$\begin{bmatrix} \dot{x} \\ \dot{\hat{x}} \end{bmatrix} = \begin{bmatrix} F & -GK \\ LH & F - GK + LH \end{bmatrix} \begin{bmatrix} x \\ \hat{x} \end{bmatrix} + \begin{bmatrix} G \\ G \end{bmatrix} u.$$

Since the characteristic equation is unaltered by a nonsingular transformation of the state $[x', \hat{x}']'$, consider the transformation T denoted by

$$\begin{bmatrix} I & 0 \\ I & -I \end{bmatrix}.$$

Obviously T is nonsingular and

$$T \begin{bmatrix} F & -GK \\ LH & F - GK + LH \end{bmatrix} T^{-1} = \begin{bmatrix} F - GK & GK \\ 0 & F - LH \end{bmatrix}$$

whose characteristic polynomial is clearly

$$\chi(F - GK)\chi(F - LH).$$

V. DESIGN OF OBSERVERS OF MINIMAL ORDER

In Section IV, the design of observers was discussed and it was seen that the observer is a system of the same order as Σ, namely n. If it is assumed that H has full rank (this assumption is obviously no loss of generality), it is clear that the knowledge of Hx gives measurements on part of the state immediately; for

$$y = Hx \tag{34}$$

gives the projection of x on the row space of H. Hence it seems that one should be able to find the other component in the complement of the row space of H by making use of a dynamical system of order $(n - p)$ $(H\colon p \times n)$. This was first proved by Luenberger in 1964 in the single input/output case[1] and was proved in the multiple input/output case by the same author in 1966.[2]

In the following section, a simple proof of this proposition is given and a different method for constructing the observer is derived.

Let the given system be, as before:

$$\begin{aligned} \dot{x} &= Fx + Gu, \\ y &= Hx. \end{aligned} \tag{35}$$

Let H be of full rank. It can be assumed that H is of the form

$$H = [I_p\,,\,0], \qquad p < n; \tag{36}$$

since a nonsingular transformation of the state of equation (35) can always be found such that equation (36) holds. Hence, if

$$x = \begin{array}{c} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \end{array}\begin{array}{l} p \\ n-p \end{array}, \tag{37}$$

and

$$F = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}\begin{array}{l} p \\ n-p \end{array}, \tag{38}$$

$$G = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}\begin{array}{l} p \\ n-p. \end{array} \tag{39}$$

Then

$$\begin{aligned} \dot{x}_1 &= F_{11}x_1 + F_{12}x_2 + G_1u, \\ \dot{x}_2 &= F_{21}x_1 + F_{22}x_2 + G_2u. \end{aligned} \tag{40}$$

Now the design procedure will be described in the form of two theorems for clarity.

Theorem 2: If $(H,\, F)$ is completely observable, then $[F_{12}\,,\, F_{22}]$ is completely observable.

Proof: The assertion of the theorem is in some sense intuitively clear, since $y = x_1$ does not give any direct information about x_2. The only information about x_2 is obtained from equation (40). That is,

$$F_{12}x_2 = \dot{x}_1 - F_{11}x_1 - G_1u, \tag{41}$$

which implies that $F_{12}\,,\, F_{22}$ should be completely observable in order that (H, F) be completely observable.

The proof of the theorem is immediate, since $(H,\, F)$ completely observable implies that the rank

$$\rho[(H',\, F'H',\, \cdots\,,\, F'^{n-1}H')] = n,$$

i.e., using the partition of F indicated in equation (38),

$$\rho\left[\begin{bmatrix} I & 0 \\ F_{11} & F_{12} \\ F_{11}^2 + F_{12}F_{21} & F_{11}F_{12} + F_{12}F_{22} \\ \cdots & \cdots \\ \cdots & \cdots \end{bmatrix}\right] = n. \tag{42}$$

[In the following a row of a partitioned matrix will mean the "block" row; for example, the first row of the matrix in equation (42) is $[I\ 0]$, the second row is $[F_{11}\ F_{12}]$, etc.]

The rank of the matrix in equation (42) is unaltered by adding to any row linear combination of other rows. Hence

$$\rho\left(\begin{bmatrix} I & 0 \\ \cdots & F_{12} \\ \cdots & F_{12}F_{22} \\ \cdots & \cdots \\ \cdots & F_{12}F_{22}^{n-1} \end{bmatrix}\right) = n. \tag{43}$$

The third row of the matrix in equation (43) is the third row of the old matrix $-\ F_{11}^2$ (first row) $-\ F_{11}$ (second row), etc. Equation (43) implies that, irrespective of what the first column of the matrix in equation (43) is,

$$\rho\left(\begin{bmatrix} 0 \\ F_{12} \\ \vdots \\ F_{12}F_{22}^{n-1} \end{bmatrix}\right) = (n - p), \tag{44}$$

which by the Cayley–Hamilton theorem implies that one need only include terms up to $F_{12}F_{22}^{n-p-1}$ which gives

$$\rho([F'_{12}\ , F'_{22}F'_{12}\ , \qquad , F'^{n-p-1}_{22}F'_{12}]) = (n - p). \tag{45}$$

Theorem 3: There exists an "observer" of dimension $(n - p)$ *for* Σ *of Theorem 2.*

Proof: Consider the "partitioned" system presented in Theorem 3.

From Section IV it is clear that since $(F_{12}\ , F_{22})$ is completely observable we can find an L such that $(F_{22} - LF_{12})$ has arbitrary eigenvalues. Hence if $\bar{x}_2$ is defined by

$$\dot{\bar{x}}_2 = F_{22}\bar{x}_2 + LF_{12}(x_2 - \bar{x}_2) + G_2u + F_{21}x_1\ , \tag{46}$$

then $\tilde{x} = x_2 - \bar{x}_2$ implies

$$\dot{\tilde{x}} = (\dot{x}_2 - \dot{\bar{x}}_2) = (F_{22} - LF_{12})\tilde{x}_2\ . \tag{47}$$

Therefore by choosing L appropriately, we can make $\tilde{x}_2 \to 0$ as fast as we want. The only problem is observing $(x_2 - \bar{x}_2)$ in equation (46), that can be solved as follows. Using equation (40), equation (46) reduces to

$$\dot{\bar{x}}_2 = (F_{22} - LF_{12})\bar{x}_2 - LF_{11}x_1 - LG_1u + G_2u + F_{21}x_1 + L\dot{x}_1 \,. \tag{48}$$

All the terms on the R. H. S. of equation (48) except $L\dot{x}_1$ can be observed. In order to eliminate the need for getting $L\dot{x}_1$ we replace it by $(F_{22} - LF_{12})Lx_1$, i.e., if

$$\begin{aligned}\dot{\hat{x}}_2 = (F_{22} - LF_{12})\hat{x}_2 - LF_{11}x_1 - LG_1u \\ + G_2u + F_{21}x_1 + (F_{22} - LF_{12})Lx_1 \,.\end{aligned} \tag{49}$$

Then it can be shown by integration by parts that

$$\begin{aligned}\bar{x}_2 = \hat{x}_2 + \exp\,(F_{22} - LF_{12})t \\ \cdot(\bar{x}_2(0) + LF_{11}x_1(0) + LG_1u(0) - G_2u(0)) + Lx_1(t).\end{aligned} \tag{50}$$

Therefore by appropriately choosing initial conditions for the system described by equation (49) we can make

$$\bar{x}_2 = \hat{x}_2 + Lx_1(t) \tag{51}$$

hence the proof is complete. Note that even if the initial conditions for equation (41) were not set, the error term, namely,

$$\bar{x}_1 - \hat{x}_2 - Lx_1(t) \to 0 \quad \text{as fast as} \quad \exp\,(F_{22} - LF_{12})t.$$

The proof of the theorem, though complete above, is easier to see in the form of figures.

Equation (48) says that the observer is of the form presented in Fig. 4.

Theorem 3 implies that Fig. 4 is equivalent to Fig. 5, namely, the input $\dot{x}_1$ shifted over to the right of $(n - p)$ dimensional integrator.

From the above theorems the following method evolves for the construction of minimal-order observers.

Step 1: Construct by the method of §3 an L such that $F_{22} - LF_{12}$ is stable and has the required dynamics.

Step 2: Use the output x_1 in the configuration presented in Fig. 5 to get $[x_1' , \hat{x}_2']$ which gives the required estimate.

VI. EXAMPLE

In order to illustrate the methods presented in this paper, we solve a simple example of a control problem.

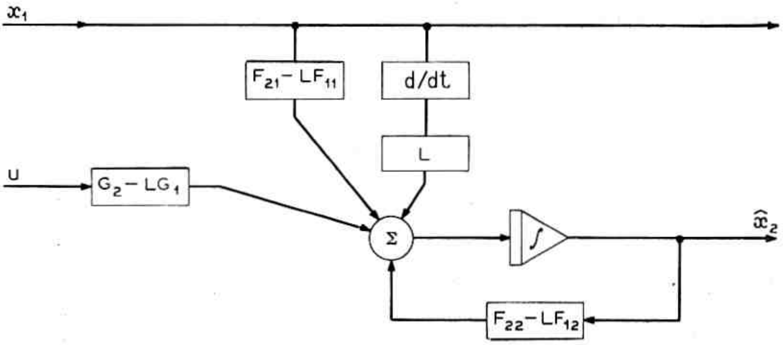


Fig. 4—Minimal order observer with rate information.

The problem is to stabilize the control system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 2 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix},$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \qquad (52)$$

It is readily seen that the characteristic polynomial of this system is

$$s^3 - 3s^2 + 3s - 1, \quad \text{or} \quad (s-1)^3, \qquad (53)$$

which shows that the system is unstable.

It is desired to design a controller which observes the output y and computes a linear feedback law such that the plant (52) behaves as a

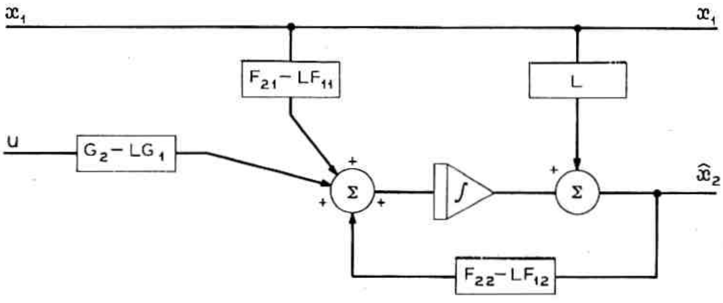


Fig. 5—Minimal order observer without rate information.

system with characteristic polynomial

$$(s + 1)^3 = 0. \tag{54}$$

Proceeding as in Sections III and IV, suppose $(x_1\,, x_2\,, x_3)$ were available; then the feedback control gain matrix say

$$\begin{bmatrix} k_{11} & k_{12}\,, & k_{13} \\ k_{21} & k_{22} & k_{23} \end{bmatrix} \triangleq K \tag{55}$$

can be computed from equation (24) which gives

$$\begin{bmatrix} 1 & 2 & 1 & 0 & 0 & 1 \\ 9 & 4 & 1 & 2 & 2 & 1 \\ 10 & 6 & 1 & 10 & 4 & 1 \end{bmatrix} \begin{bmatrix} k_{11} \\ k_{12} \\ k_{13} \\ k_{21} \\ k_{22} \\ k_{23} \end{bmatrix} = \begin{bmatrix} 6 \\ 18 \\ 38 \end{bmatrix} \cdot \tag{56}$$

Almost any solution (56), subject to the condition that rank of $K = 1$, will do. For the purpose of this discussion a particular solution, namely

$$\begin{bmatrix} 2/3 & 4/3 & 4/3 \\ 2/3 & 4/3 & 4/3 \end{bmatrix} \tag{57}$$

will be considered.

Now since x_3 is not actually available, an estimate will be computed and the error will be required to diminish as fast as $\exp(-2t)$, at the same time using only one integrator in the feedback loop. From equation (52) it follows that

$$\dot{x}_3 = x_3 + u_1 + u_2\,, \tag{58}$$

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 2 \\ 2 \end{bmatrix} x_3 + \begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix} \begin{matrix} u_2 \\ u_2 \end{matrix} . \tag{59}$$

The expressions in equations (58) and (59) can be compared to the more general treatment of Section III by noting that

$$\begin{aligned} F_{11} &= \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}, \qquad & F_{12} &= \begin{bmatrix} 2 \\ 2 \end{bmatrix}; \\ F_{22} &= [1], & F_{21} &= [0 \quad 0]. \end{aligned} \tag{60}$$

Then

$$F_{22} - LF_{12} = 1 - [L_1 \quad L_2]\begin{bmatrix} 2 \\ 2 \end{bmatrix} = -2. \tag{61}$$

One solution for L is

$$L_1 = L_2 = \tfrac{3}{4}. \tag{62}$$

The analogue of equation (23) is

$$\dot{\hat{x}}_3 = \hat{x}_3 + (3/2 \quad 3/2)\begin{bmatrix} 2 \\ 2 \end{bmatrix}(x_3 - \hat{x}_3) + u_1 + u_2 \ ; \tag{63}$$

that is, the solution of equation (46) obviously tends to x_3 as fast as $\exp(-2t)$ for from equation (52)

$$\frac{d}{dt}(x_3 - \hat{x}_3) = -2(x_3 - \hat{x}_3). \tag{64}$$

In order to construct the observer, in equation (63), x_3 has to be replaced by x_1 , x_2 which are directly measurable. Proceeding just as in §3, equation (46) becomes

$$\hat{x}_3 = -2\hat{x}_3 + [3/2 \quad 3/2]\left\{\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} - \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix}\begin{matrix} u_1 \\ u_2 \end{matrix}\right\} + u_1 + u \ . \tag{65}$$

Therefore

$$\dot{\hat{x}} = -2\hat{x}_3 - 3/2\, x_1 - 6x_2 - 9/2\, u_1 + 3/2(\dot{x}_1 + \dot{x}_2) + u_1 + u_2 \ . \tag{66}$$

The only quantity that is not available on the right-hand side of equation (66) is $(\dot{x}_1 + \dot{x}_2)$, but appealing to Theorem 4 it follows that if

$$\dot{\bar{x}}_3 = -2\bar{x}_3 - 3/2\, x_1 - 6x_2 - 9/2\, u_1 - 3(x_1 + x_2) + u_1 + u_2 \tag{67}$$

$$\hat{x}_3 \to \bar{x}_3 + 3/2(x_1 + x_2) \tag{68}$$

and the error tends to zero as exp $(-2t)$ [it can be made zero by setting the initial condition on the integrator simulating equation (62) to $\hat{x}_3(0) - \frac{3}{2}(x_1(0) + x_2(0))$].

Hence $\bar{x}_3 + \frac{3}{2}(x_1 + x_2)$ which can also be gotten as in Fig. 5 tends to x_3 as exp $(-2t)$. Therefore the controller design is complete. The analog computer realization is shown in Fig. 6.

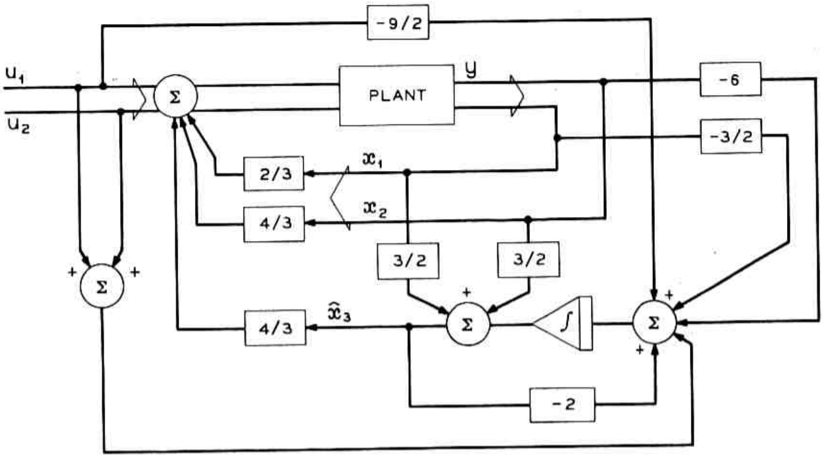


Fig. 6—Example of a controlled system.

REFERENCES

1. Luenberger, D. G., "Observers of Low Dynamic Order," IEEE Trans. Military Elec., *8,* No. 2 (April 1964), pp. 74–80.
2. Luenberger, D. G., "Observers for Multivariable Systems," IEEE Trans. Automatic Control, *AC-11,* No. 2 (April 1966), pp. 190–197.
3. Wonham, W. M., "On Pole Assignment in Multi-Input Controllable Linear Systems," IEEE Trans. Automatic Control, *AC-12,* No. 6 (December 1967), pp. 660–665.
4. Kalman, R. E., "Mathematical Description of Linear Dynamical Systems," J. SIAM Control, *1,* 1963, pp. 152–192.
5. Gopinath, B., "On the Identification and Control of Linear Systems," Ph.D. Dissertation Submitted to the Department of Electrical Engineering, Stanford University, Stanford, California, 1968.
6. Gantmacher, F. R., *Matrix Theory,* Vol. I, New York: Chelsea, 1960.

Contributors to This Issue

Raymond E. Barber, B.S.E.E., 1967, University of Denver; M.S.E.E., 1970, North Carolina State University; Western Electric, 1967–1971. AT&TCo, 1971—. Mr. Barber was assigned to the Western Electric Service Division from 1967 to 1968. In 1968 he was transferred to Defense Activities, Greensboro, North Carolina, where he was "on loan" to Bell Laboratories, involved in precision frequency and time generation. In October 1970, he returned to the Western Electric Service Division, Aurora, Colorado. Currently he is with AT&TCo Long Lines Department, New York City. Member, Eta Kappa Nu, IEEE.

Andrew H. Bobeck, B.S.E.E., 1948, and M.S.E.E., 1949, Purdue University; Bell Telephone Laboratories, 1949—. Mr. Bobeck's early work concerned the design of communication and pulse transformers and, later, the development of one of the first solid-state digital computers. Since 1956, he has specialized in the development of magnetic logic and memory devices. He was responsible for the conception and development of the twistor memory device. Most recently he has been investigating the properties of cylindrical domains found in uniaxial magnetic materials, such as the orthoferrites. Member, Tau Beta Pi, Eta Kappa Nu. Fellow, IEEE.

Richard D. Brooks, B.S.E.E., 1958, Newark College of Engineering; M.S.E.E., 1961, Ohio State University; and Ph.D., 1970, Lehigh University; Bell Telephone Laboratories, 1961—. Since joining Bell Laboratories, Mr. Brooks has been engaged in the development of traveling-wave-tube electron guns, and microwave solid-state harmonic generators. He is currently developing integrated circuits. Member, Eta Kappa Nu, IEEE.

Denis J. Connor, B.A.Sc., 1963, M.A.Sc., 1965, and Ph.D., 1969, University of British Columbia; Bell Telephone Laboratories, 1969—. A member of the Opto-Electronics Research Department, Mr. Connor is currently working on techniques for the efficient coding of television signals by making use of both in-frame and frame-to-frame correlation.

Edward Della Torre, B.E.E., 1954, Brooklyn Polytechnic Institute; M.S. (Electrical Engineering), 1956, Princeton University; M.S. (Physics), 1961, Rutgers University; D.E.Sc. (Electrical Engineering), 1964, Columbia University; Associate Professor, Rutgers University, 1956–1967; Bell Telephone Laboratories, 1967–1968; Associate Professor, McMaster University, 1968—. Mr. Della Torre's early interests were in magnetic recording. Since 1967, he has been interested in bubble devices. He is the author of *The Electromagnetic Field* with C. V. Longo. Member, Sigma Xi, Eta Kappa Nu. Senior Member, IEEE.

U. F. Gianola, B.S. (Physics), 1948, and Ph.D. (Electron Physics), 1951, University of Birmingham, England; Postdoctorate Fellow, Physics Department, University of British Columbia, 1951–1953; Bell Telephone Laboratories, 1953—. Mr. Gianola worked initially in the Communications Research Department where he was concerned with exploration of new techniques for communications devices. Later in the Components and Solid State Device Division, he supervised construction of prototype models of the permanent magnet twistor memory for stored program in ESS No. 1. In 1963 he became Head of the Fundamental Memory Components and was responsible for the exploratory development of advanced memory devices, including magnetic thin film and magnetic "bubble" domain memories. In 1969 he was appointed Head of the Ocean Physics Research Department, responsible for experimental and theoretical studies of ocean acoustics phenomena. Member, American Physical Society, Scientific Research Society of America. Fellow, IEEE.

Bernard Glance, Dipl. Ing., 1958, Ecole Spéciale de Mécanique et Electricité; Dipl. Ing., 1960, Ecole Supérieure d'Electricité, Paris, France; C.S.F., Research Center of Corbeville, Orsay, France, 1960–1966; Dipl. Docteur (Ing.), 1964, Sorbonne, Paris; Bell Telephone Laboratories, 1968—. At C.S.F., Mr. Glance had been engaged in research on microwave tubes. At S.F.D. Laboratories, he had worked on high-power microwave amplifiers. Mr. Glance is presently working on microwave solid-state integrated circuits.

B. Gopinath, M.S. (Mathematical Physics), 1964, University of Bombay, India; M.S.E.E. and Ph.D.(E.E.), 1968, Stanford University; Postdoctoral Research Associate, Stanford University, 1967–1968;

Bell Telephone Laboratories, 1968—. Mr. Gopinath's primary interest, as a member of the Systems Theory Research Group, is in the applications of mathematical methods to physical problems.

HARRY HEFFES, B.E.E., 1962, City College of New York; M.E.E., 1964, and Ph.D., 1968, New York University; Bell Telephone Laboratories, 1962—. Mr. Heffes' interests have been in applications of modern control theory, filtering theory, and system modelling. An Adjunct Associate Professor of Electrical Engineering at New York University, Mr. Heffes is currently Supervisor in the Systems Analysis and Modelling Department. Member, Tau Beta Pi, Eta Kappa Nu.

E. Y. HO, B.S.E.E., 1964, The National Taiwan University; Ph.D., 1969, University of Pennsylvania; Bell Telephone Laboratories, 1969—. Mr. Ho has been engaged in developing and analyzing automatic equalizers for data transmission systems. Member, IEEE.

SHELDON HORING, B.E.E., 1957, City College of New York; M.E.E., 1959, New York University; Ph.D. (E.E.), 1962, Brooklyn Polytechnic Institute; Bell Telephone Laboratories, 1957–1960, 1962—. Mr. Horing completed the communications development training program in 1960. He was first engaged in the design and development of an optical electromechanical control system. After spending two years on the faculty at Brooklyn Polytechnic Institute, he returned to Bell Laboratories where he joined the Mathematical Analysis and Consulting Group. Since that time, he has been engaged in research and consulting in control theory and related areas, as well as in studies of defense systems and air traffic control. He is currently Head of the Systems Analysis and Modelling Department. Member, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

D. L. JAGERMAN, B.E.E., 1949, Cooper Union; M.S., 1954, and Ph.D. (Mathematics), 1962, New York University; Bell Telephone Laboratories, 1964—. Mr. Jagerman has been engaged in mathematical research on numerical quadrature theory, interpolation, mathematical properties of pseudo-random number generators, dynamic programming, and approximation theory. His recent work concerns the theory of widths and entropy with application to the storage and transmission of information. Member, American Mathematical Society, Pi Mu Epsilon.

LEIF LUNDQUIST, Civilingenjör, 1964, Royal Institute of Technology, Stockholm, Sweden; Microwave Department, Royal Institute of Technology, Stockholm, Sweden, 1964–1965; Bell Telephone Laboratories, 1965—. Mr. Lundquist formerly was engaged in interference studies in the Radio Systems Engineering Department. He is currently Supervisor in the Transmission Facilities Engineering Department, involved in application studies. Member, Svenska Teknologföreningen, IEEE.

Hans G. Mattes, B.S.E.E., 1964, California Institute of Technology; M.S.E.E., 1966, and Ph.D., 1968, University of Southern California; Bell Telephone Laboratories, 1968–1970. Mr. Mattes was engaged in the development of substrates for integrated electronics. Presently on a leave of absence, he is teaching at the National Taiwan University. Member, Sigma Xi.

James E. Mazo, B.S., 1958, Massachusetts Institute of Technology; M.S., 1960, and Ph.D., 1963, Syracuse University; Research Associate, University of Indiana, 1963–64; Bell Telephone Laboratories, 1964—. Mr. Mazo was engaged in work on quantum scattering theory at Indiana University. Now he is doing theoretical analysis of data systems. Member, American Physical Society, Sigma Xi, IEEE.

Scotty R. Neal, B.A. (Mathematics), 1961, M.A. (Mathematics), 1963, and Ph.D. (Mathematics), 1965, University of California, Riverside; Research Mathematician, Naval Weapons Center, China Lake, California, 1964–1967; Bell Telephone Laboratories, 1967—. Since coming to Bell Laboratories, Mr. Neal has been primarily concerned with the analysis of various aspects of telephone traffic systems. He has also worked on applications of optimal linear estimation theory and certain aspects of communication theory. Member, American Mathematical Society, SIAM, Sigma Xi.

Peter G. Neumann, A.B., 1954, and S.M., 1955, Harvard University; Dr. rerum naturum 1960, Technische Hochschule Darmstadt, West Germany; Ph.D. (Applied Mathematics), 1961, Harvard University; Bell Telephone Laboratories, 1960—; currently Visiting Mackay Lecturer, University of California at Berkeley (on leave from Bell Laboratories, 1970–71). Mr. Neumann has been engaged in research in computers and communications since 1955, including as special interests coding theory and computer operating systems. Member, ACM, Sigma Xi, Society of Harvard Engineers and Scientists, AAAS, IEEE.

R. F. W. Pease, B.A., 1960, M.A. and Ph.D., 1964, University of Cambridge; Bell Telephone Laboratories, 1967—. Mr. Pease held a faculty appointment at the University of California at Berkeley prior to joining Bell Laboratories and worked on electron microscopy. He is now trying to efficiently encode moving and still pictures.

S. D. Personick, B.E.E., 1967, City College of New York; S.M. in E.E., 1968, E.E., 1969, and Sc.D., 1969, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1967—. Mr. Personick is engaged in work on wire systems engineering.

John R. Rosenberger, B.S.E.E., 1966, University of Tennessee; M.S.E., 1967, University of Pennsylvania; Bell Telephone Laboratories, 1966—. Mr. Rosenberger has studied the problem of echo in the telephone network and the feasibility of employing self-adaptive echo cancellers to control echo. He has also worked in the area of filter design where he was concerned with developing an interactive filter design and analysis program. He is presently engaged in work on a portable computerized testing system for measuring transmission and performance characteristics of telephone circuits. Member, Tau Beta Pi, Eta Kappa Nu.

William G. Scholes, A.A.Sc., 1968, Devry Institute of Technology; Bell Telephone Laboratories, 1968—. Mr. Scholes is in the Opto-Electronics Research Department. He is currently working on frame-to-frame and in-frame coding of television signals.

David E. Setzer, B.S.M.E., 1958, Lehigh University; M.S.E.M., 1960, New York University; Ph.D. (Applied Mechanics), 1964, Lehigh University; Instructor of Mechanics, Lehigh University, 1961–1964; Bell Telephone Laboratories, 1958–1961, 1964—. Mr. Setzer's early work at Bell Laboratories dealt with the design and construction of the TELSTAR® Earth Stations. Later he engaged in experimental and theoretical studies of transmission through natural aerosols. He is presently working in improving electrical and mechanical properties of communication cables. Member, American Association for the Advancement of Science, Pi Tau Sigma, Tau Beta Pi, Pi Mu Epsilon, Sigma Xi, Newtonian Society.

A. A. Thiele, B.S. (Physics), 1960, and Ph.D. (Physics), 1965, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1965—. Mr. Thiele was initially engaged in exploratory device development for various optical memory schemes. He is currently working on the development of the theory of circular domains. Member, Sigma Xi.

Edmond J. Thomas, B.S.E.E., 1964, and M.S., 1965, Rensselaer Polytechnic Institute; Rutgers University, 1968; Bell Telephone Laboratories, 1965—. Mr. Thomas has been engaged in studying the applicability of various adaptive control systems to the echo-control problem. Most recently he has been involved in characterizing the performance of these systems in nonlinear and time-varying environments. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Xi, IEEE.

Arvids Vigants, B.E.E., 1956, City College of New York; M.S. (E.E.), 1957, and Eng. Sc.D. (E.E.), 1962, Columbia University; Bell Telephone Laboratories, 1962—. Mr. Vigants has worked on various electromagnetic wave propagation problems and is currently working on line-of-sight microwave propagation. Member, Eta Kappa Nu, Tau Beta Pi, Sigma Xi, Commission 2 of URSI, IEEE.

Yu S. Yeh, B.S.E.E., 1961, National Taiwan University; M.S.E.E., 1964, and Ph.D., 1966, University of California, Berkeley; Harvard University 1967; Bell Telephone Laboratories, 1967—. Mr. Yeh is a member of the Radio Transmission Research Department and is doing research work concerning Mobile Radio communication.

B.S.T.J BRIEFS

HILO—An Improved Transmission Scheme for Semiconductor Switching Networks

By R. R. LAANE

(Manuscript received July 25, 1969)

I. INTRODUCTION

Foremost among the problems in semiconductor switching network implementations is the difficulty of fabricating a semiconductor crosspoint which duplicates the characteristics of metallic crosspoints of nearly zero "on" impedance. A low crosspoint impedance is necessary for low signal loss through the switching network.

A transmission technique is described in this paper which relieves the requirement for low crosspoint "on" impedance and which allows excellent voice-band transmission characteristics with unbalanced semiconductor switching networks. The technique called HILO converts voltage inputs to the network into current changes using a high input impedance current source and transmits the current changes through the network crosspoints by modulating the network bias current. A low output impedance is provided at the networks receiving terminal, where the modulated currents are decoded to recover normal voltage-current levels.

II. CURRENT TRANSMISSION

The ac equivalent of the HILO current transmission circuit is shown in Fig. 1. Assuming idealized transistor characteristics (zero emitter impedance, infinite collector impedance, and unity α), an input signal e_i controls the current source, Q_1, converting the input into a current change $i = e_i/R_i$. The current change is transmitted through a network path and is supplied as an input to a common base amplifier, Q_2, at the receiving terminal. The recovered output is given by

$$e_0 = iR_0 = R_0/R_ie_i \, . \tag{1}$$

Although equation (1) assumes idealized transistor characteristics to

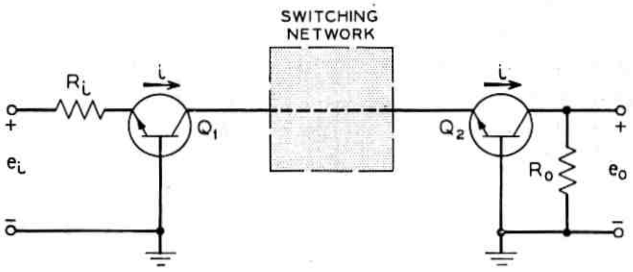


Fig. 1 — Current transmission circuit for switching networks (assume $\alpha \rightarrow 1$).

keep the exposition simple, the equation remains reasonably valid for voice-band applications using Darlington transistor pairs and normal (600-ohm) transmission impedance levels.

Because information is transmitted only as a current change within the switching network, crosspoint resistance and crosspoint nonlinearities do not impair the characteristics of the recovered signal. Recovered transmission level becomes primarily a function of R_i and R_0. As a result, network gain (loss) characteristics can be accurately controlled.

In addition to providing good signal transmission capability, the technique allows nearly ideal isolation between network paths. Capacitive crosstalk is minimized because only a small voltage change is produced within the network transmission path as a result of the transmitted current changes. This is the result of a low input impedance at the common base output stage and the result of relatively low series resistance of the network crosspoints.

Inductive crosstalk is reduced because the high output impedance of the current source driver makes the transmitted currents relatively insensitive to inductive coupling between network paths.

III. BIASING, TRANSMISSION, AND CONTROL

As an example of current transmission, Fig. 2 illustrates the biasing, transmission, and control circuitry necessary for operating an unbalanced thyristor network. Only a single direction of transmission is shown. For a bidirectional connection, a second, identical circuit is required.

To establish a connection through the network, a CONNECT signal is applied at the transmit terminal. The signal sets the control flip-flop, thereby turning off Q_1. As a result current source transistors, Q_2 and Q_3, become forward biased. The CONNECT signal also saturates Q_4 which allows Q_2 and Q_3 to saturate during selection.

Next, appropriate thyristor base selection lines are pulsed simultaneously to form a path through the network stages. Crosspoints are activated when the selected current path appears at the cathode terminal and when a thyristor selection control signal is applied at the base terminal. Thus, thyristors turn on in sequence starting with the stage nearest the network input terminal.

Holding current for the crosspoints is supplied through thyristor base terminals of successive stages until a path has been established from the transmitting to the receiving terminal. At this time, appropriate holding current is supplied through Q_5 and Q_6 and thyristor selection inputs may be terminated. Finally, the CONNECT input is removed, causing Q_4 to turn off and forcing Q_2 and Q_3 to function as a current source.

Supply voltage V_2 is selected to set the level of the current source collector voltage (Q_2 and Q_3) at a higher potential than the thyristor base selection inputs. This prevents faulty connection to already active network paths.

A network path is disconnected by applying a DISCONNECT input to the control flip-flop. This causes Q_1 to saturate, terminating the current flow in the network path and thereby tearing down the connection.

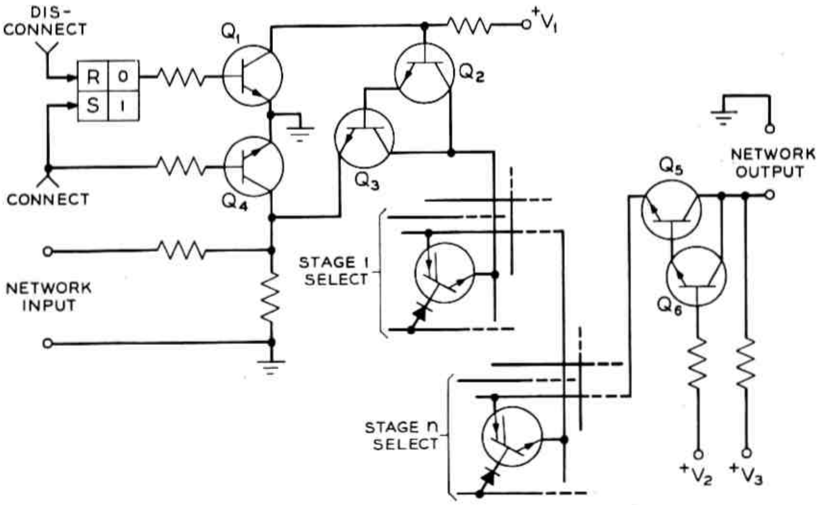


Fig. 2. — Thyristor network using current transmission.

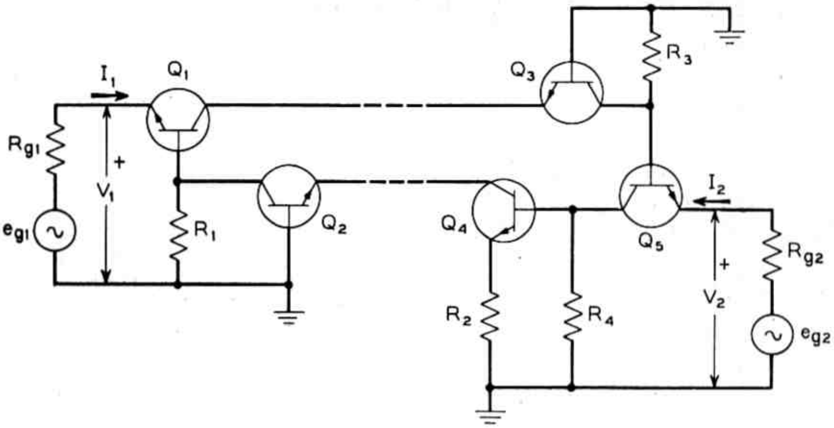


Fig. 3 — Combined hybrid and transmission gate (assume $\alpha \rightarrow 1$).

IV. COMBINED HYBRID AND TRANSMISSION CIRCUIT

A loss-less electronic hybrid can be combined with the transmission and biasing circuits to convert two unidirectional transmission paths in the switching network to bidirectional transmission paths at the network terminals. Figure 3 shows the ac equivalent of the hybrid circuit.

Currents appearing at the emitter of Q_1 are transmitted to the emitter of Q_5 via Q_3; whereas, currents at the emitter of Q_5 are transmitted to Q_1 via Q_4 and Q_2. Inverter stage Q_4 is necessary for proper phase relationship in the transmission loop. The circuit can be described by the following z-parameters

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} 0 & -R_1R_4/R_2 \\ R_3 & 0 \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \end{bmatrix}. \tag{2}$$

Consequently,

$$V_2 = \frac{R_2R_{g2}}{R_1R_4} V_1 \tag{3}$$

and

$$V_1 = -\frac{R_{g1}}{R_3} V_2 . \tag{4}$$

The input impedance at the two input terminals is given by

$$Z_1 = \frac{V_1}{I_1} = \frac{R_1R_3R_4}{R_2R_{g2}} \tag{5}$$

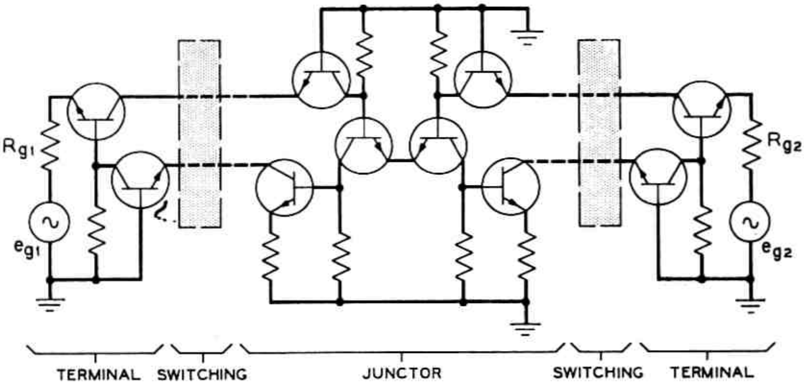


Fig. 4 — Combining two hybrid-transmission circuits in a switching network.

and

$$Z_2 = \frac{V_2}{I_2} = \frac{R_1R_3R_4}{R_2R_{g1}}. \tag{6}$$

To cancel the inverse effect of R_{g1} and R_{g2} on input impedance two identical circuits are cascaded between network terminals. Assuming R_1, R_2, R_3, and R_4 of the circuits are identical, the circuit acts as a gyrator and termination impedances are reflected through the network to the input terminals such that

$$Z_1 = R_{g2}, \qquad Z_2 = R_{g1}. \tag{7}$$

For symmetry, two stages are connected as shown in Fig. 4 to form a switching network with unidirectional network paths and bidirectional network terminals.

The contributions and suggestions of W. B. Gaunt, J. E. Iwersen and D. Vlack on the hybrid circuit are gratefully acknowledged.

A Telephone Traffic Model Based on Randomly Closing Crosspoints, and Its Relationships to Other Models

By V. E. BENEŠ

(Manuscript received December 23, 1970)

I. INTRODUCTION AND SUMMARY

In the theory of traffic in telephone connecting networks it is on one hand a virtual necessity, for practical purposes, to compromise the true complexity of the system under study and to introduce drastic simplifying assumptions that allow some calculation to be done, and on the other, it is perfectly feasible to pursue basic theoretical studies without such compromise and simplification. For this reason, a spectrum of several mathematical models for describing traffic in networks has been developed in recent years.

These models range from "simple" ones that furnish an incomplete description based on strong stochastic independence assumptions, to "complicated" ones that exactly mirror network structure and routing. Each grade of model has its uses: "simple" ones for easy computation and involved ones for general understanding.

An example of a useful "simple" model is the probability linear graph[1] suggested by C. Y. Lee in 1955, an outgrowth of earlier work by L. E. Kittredge and E. C. Molina. At the other end of the scale, an example of a "complicated" model is the Markov process[2] proposed by the author in 1963 as an improvement of the "thermodynamic" model.[3]

We shall describe here another "simple" model, with a basic starting point similar to that of Lee, and then show how a certain natural restriction of this model yields in many cases precisely the thermodynamic model. Our presentation thus clarifies the known relationships between these models, and reveals some unsuspected ones; it strengthens our understanding of them by showing how the apparently realistic "complicated" models can arise through natural and relatively minor modfications of the "simple" ones.

Whereas Lee's model assumes that a *link* ℓ of the network is busy or idle with a probability p_ℓ, all links being independent of each other, we propose instead to assume that each (individual switch or) *crosspoint* c is closed with some probability p_c, again independently. This basis for calculating probabilities has the virtue of assigning probability to every possible way of closing switches, physically meaningful or not. We then modify the model by calculating all probabilities *conditional* on the system's being in a physically meaningful state. This procedure in effect rids our calculations of the irrelevant states by normalizing them out. The resulting new model we call the *crosspoint* model. If, as is usually the case, every call goes through the same number of switches, then conditioning in this way brings in the partition function[3] in a natural way, and the crosspoint model turns out to be formally equivalent to the thermodynamic one; when the latter is in turn modified so as to take realistic account of routing and calling rates, it becomes the Markov process model.

We stress that the suggestion made here of a new model does not really improve our capacity to calculate blocking, load-loss curves, or other practical items. Primarily, it provides a new derivation of the thermodynamic model from simple (and strong) first principles similar to those used for the probability linear graph model of Lee.

II. LEE'S MODEL AND ITS EXTENSION

The probability linear graph model has been extensively discussed[1,4] in the literature, so we include only a resume of the method: to calculate the congestion incurred by traffic between an inlet u and an outlet v, attention is focused on the graph G defined by the permitted paths through the network from u to v; G consists of all nodes and branches through which some path from u to v passes. Given any complete specification of which branches of G are busy and which are idle (at a particular juncture of network operation), it is possible to examine G to see if there is a path from u to v no branch of which is busy. The method now assigns a probability distribution to the possible occupancies by postulating that a link ℓ of G is busy with probability p_ℓ independently of all other links. The congestion for u and v is then calculated as the probability that this distribution assigns to the event "There is no path from u to v composed of idle branches."

We have described the probability linear graph model as assuming something only about certain events having to do with the graph G of paths for a particular call from u to v, and not as providing a probabilistic description of the busy or idle condition of all the links in the network. However, it is entirely possible to extend the probabilistic description, used in Lee's model for links of G, to all the links in the network. This extension is natural because if the description is believable for one inlet-outlet pair, it should be so for all such pairs, and so for all links. It will of course still give only an incomplete stochastic model, since it says rather little about what crosspoints are closed, so that in general it is not possible to tell what inlet is connected to what outlet.

However, the extension does shed some light on the character and accuracy of Lee's model, as we note in the next paragraph, and it also suggests the new model to be proposed.

The fashion in which this extended version of Lee's model works is clear: the states of the network, i.e., all the possible ways of closing crosspoints, whether physically meaningful or not, are partitioned according to the equivalence relation of "having the same links busy," and probabilities are assigned to these equivalence classes. In this situation, it is unfortunately true that physically meaningful and physically irrelevant (microscopic) states occur in the same equivalence class. Were this not so, one could try to remove the effect of the irrelevant states[5] by insisting that all probabilities be taken conditional on being in the set of relevant states. This set, however, does not have probability assigned to it.

III. THE CROSSPOINT MODEL

There is, nevertheless, a basic modification of Lee's approach in which the normalization device for eliminating the "irrelevant" states can be used. The change to be made is this: whereas Lee's model assigns a probability p_ℓ of being busy to each link ℓ, we propose to assign a probability p_c of being closed to each crosspoint c, all independently in both cases. The point is this: if it is known what crosspoints are closed, then it is known what links are busy, but not *vice versa*. This approach has the property of assigning probability to *every* state of the network, physically meaningful or not.

In particular, the set of meaningful states is assigned probability. Once this is true we can restrict attention to these states. We shall eliminate the effect of the irrelevant states by simply normalizing them out of the picture, i.e., by calculating all probabilities of interest *conditional* on being in the set of meaningful states. The distribution of probability (over the set S of physically meaningful states) obtained in this way we shall call the "crosspoint" model, because the basic events to which probability is assigned are closings or openings of crosspoints. In a similar vein, Lee's model might be called the "link" model, because the basic probabilities are assigned to the busy or idle condition of links.

IV. PROPERTIES OF THE CROSSPOINT MODEL

Let S be the set of physically meaningful ways of closing crosspoints, and for $x \in S$ let $c(x)$ be the number of switches or crosspoints closed in x. Let us suppose that all crosspoints have the same chance p of being closed. (This is likely to be true if the network traffic is uniform and if there is neither concentration nor expansion.) The basic *unconditional* probability assigned to x is then

$$p^{c(x)}(1-p)^{C-c(x)} \qquad x \in S$$

where C is the total number of crosspoints in the network. The probability of x conditional on being in the set S of physically meaningful states is then zero if $x \notin S$, and

$$\frac{p^{c(x)}(1-p)^{C-c(x)}}{\sum_{y \in S} p^{c(y)}(1-p)^{C-c(y)}}, \qquad \text{for} \quad x \in S,$$

or, with $\mu = p/(1-p)$,

$$\frac{\mu^{c(x)}}{\sum_{y \in S} \mu^{c(y)}}.$$

This is of the familiar Maxwell–Boltzmann form, with the function $c(\cdot)$ playing the role of the energy. The reader familiar with the thermodynamic model[3] will at once recognize the resemblance of the above expression to the basic (equilibrium) state probabilities in that model, which assigns a meaningful state $x \in S$ a probability

$$\frac{\lambda^{|x|}}{\sum_{y \in S} \lambda^{|y|}},$$

with $|y|$ = number of calls in progress in state y, and λ a positive constant. As we have pointed out, this distribution is obtained by maximizing the entropy functional subject to a fixed mean number of calls in progress. Exactly the same arguments[3] characterize the distribution over S in the crosspoint model, as follows:

Theorem: The distribution $\{q_x, x \in S\}$ *of probability over* S *which maximizes the entropy functional*

$$H(q) = -\sum_{x \in S} q_x \log q_x$$

subject to the constraint

$$\sum_{x \in S} q_x c(x) = c,$$

with c a fixed number >0, is given by

$$q_x = \frac{\mu^{c(x)}}{\sum_{y \epsilon S} \mu^{c(y)}}$$

where $\mu > 0$ is the constant determined uniquely by the equation

$$c = \mu \frac{d}{d\mu} \log \sum_{x \epsilon S} \mu^{c(x)}.$$

Thus the crosspoint model differs from the thermodynamic model only in that the average number of closed crosspoints, rather than that of calls in progress, is fixed while maximizing the entropy. In an important class of cases the two models formally coincide, even. This can be seen from the

Corollary: If every call goes through exactly s switches, then the state probabilities assigned by the crosspoint model with parameter p are exactly the same as those assigned by the thermodynamic model with parameter

$$\lambda = \mu^s = \left(\frac{p}{1-p}\right)^s.$$

For evidently in this case $c(x) = s \mid x \mid$. The property that every call goes through the same number of switches is possessed by virtually all the connecting networks used in practice.

REFERENCES

1. Lee, C. Y., "Analysis of Switching Networks," B.S.T.J., *34,* No. 6 (November 1955), pp. 1287–1315.
2. Beneš, V. E., "Markov Processes Representing Traffic in Connecting Networks," B.S.T.J., *42,* No. 6 (November 1963), pp. 2795–2837.
3. Beneš, V. E., "A 'Thermodynamic' Theory of Traffic in Connecting Networks," B.S.T.J., *42,* No. 3 (May 1963), pp. 567–607.
4. Grantges, R. F., and Sinowitz, N. R., "NEASIM: A General-Purpose Computer Simulation Program for Load-Loss Analysis of Multistage Central Office Switching Networks," B.S.T.J., *43,* No. 3 (May 1964), pp. 965–1004.
5. Beneš, V. E., "On Some Proposed Models for Traffic in Connecting Networks," B.S.T.J., *46,* No. 1 (January 1967), pp. 105–116.