THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLII SEPTEMBER 1963 NUMBER 5

Phase-Lock Loop Design for Coherent Angle-Error Detection in the *Telstar* Satellite Tracking System

By W. L. NELSON

(Manuscript received January 14, 1963)

The function of the angle-error detector is to provide pointing-error signals to the ground antenna control system, which allows operation in the autotrack mode once the satellite beacon has been acquired. The limitations on the accuracy of this system imposed by noise, phase jitter and Doppler effects are evaluated and the optimum design in terms of minimum mean-square error is developed. Design examples are given for both the horn-reflector antenna autotrack system and the precision tracker antenna system.

I. INTRODUCTION AND SUMMARY

To insure the acquisition and accurate tracking of the Telstar communication satellites, a sequence of tracking modes is provided at the ground stations in Andover, Maine and Pleumeur-Bodou, France.[1]

Initial pointing directions to both the precision tracker[2] and the horn-reflector antennas are provided from orbital data appropriately processed and up-dated for each satellite pass. Once the precision tracker acquires and tracks the satellite, the horn-reflector antenna can use the pointing directions received from the precision tracker control system to acquire the satellite beacon signal in its narrow beamwidth. Finally, the autotrack system[3] provides closed-loop automatic control of the horn-reflector antenna using error signals derived from the satellite beacon.*

* After orbital data becomes sufficiently accurate, it is possible for the horn antenna to acquire the satellite from initial pointing directions and then go directly into the autotrack mode without using the precision tracker.

The detection of these error signals is accomplished in the system described in this paper. The inputs to this system are obtained by means of a mode separation technique[3] in the waveguide of the horn antenna, one mode having a peak amplitude on target, the other having a null, similar to the sum and difference signals in conventional monopulse tracking systems. The characteristics of these input signals and noise are discussed in Section II. An analysis of the phase-lock detection scheme which converts these inputs into the desired antenna pointing-error signals is given in Section III. The accuracy of these pointing-error signals is shown to be critically dependent upon the degree of phase coherence achieved by the phase-lock loop, which is discussed in Section IV. The design of the phase-lock loop to minimize the mean-square phase error in the output signals is considered in Section V.

In Section VI a numerical example is given for the optimum design of the phase-lock detector in the vernier autotrack system. Since the precision tracker also uses essentially the same angle-error detection scheme, a parallel design of this system is included for comparison.

II. INPUT SIGNAL AND NOISE CHARACTERISTICS

The function of the angle error detector is to develop electrical error signals proportional to the pointing angle error, β, between the antenna boresight and the actual satellite position. The expressions for the desired output error signals are

$$\begin{aligned} \epsilon_x &= \beta \cos \varphi \\ \epsilon_y &= \beta \sin \varphi \end{aligned} \tag{1}$$

where x and y are Cartesian coordinates in the plane normal to the antenna boresight (electrical) and φ is the angle which the projection on the x-y plane of the radius vector, R, to the satellite makes with the x-axis, (see Fig. 1).

The information on the parameters β and φ necessary for the error signals (1) is contained in the amplitude and phase, respectively, of the difference channel received signal relative to the sum channel* received signal. For a pointing error, β, which is within the beamwidth

* The designations "difference" and "sum" channels are carried over from conventional monopulse usage. For the autotrack system, these terms should be "TM_{10}" and "TE_{11}" channels, respectively. This analysis is applicable to a single plane in conventional (linearly polarized) monopulse if the angle φ is taken to be 0 or 180 degrees.

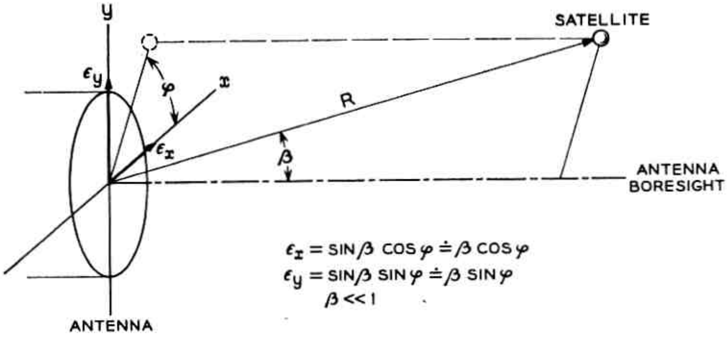

y
SATELLITE
ϵ_y
φ
x
R
ϵ_x
β
ANTENNA
BORESIGHT
$\epsilon_x = \mathrm{SIN}\,\beta\ \mathrm{COS}\,\varphi \doteq \beta\ \mathrm{COS}\,\varphi$
$\epsilon_y = \mathrm{SIN}\,\beta\ \mathrm{SIN}\,\varphi \doteq \beta\ \mathrm{SIN}\,\varphi$
$\beta \ll 1$
ANTENNA

Fig. 1 — Pointing error, β, and the angle φ in the x-y plane which determine the orthogonal error signals: $\epsilon_x = \beta \cos \varphi$; $\epsilon_y = \beta \sin \varphi$.

of the sum pattern (the TE_{11} mode pattern in the horn-reflector antenna), the received signals in the sum and difference channels can be expressed as[3]

$$\begin{aligned} e_s(t) &= E_s(R,\beta) \cos (\omega_b t + \theta_i(t)) + N_s(t) \\ e_d(t) &= \eta\beta E_s(R,\beta) \cos (\omega_b t + \theta_i(t) + \varphi) + N_d(t) \end{aligned} \tag{2}$$

where

$$\begin{aligned} \omega_b &= 2\pi \times \text{frequency of satellite beacon transmitter,} \\ R &= \text{range of the satellite,} \\ E_s(R,\beta) &= \text{sum channel signal amplitude,} \\ \eta &= \text{difference channel relative sensitivity,} \\ &= \frac{1}{E_s} \left| \frac{\Delta E_d}{\Delta \beta} \right|_{\beta=0} \end{aligned}$$

and

$\theta_i = \theta_s(t) + \theta_n(t)$, is the signal phase relative to a reference phase, $\theta_r = 0$, plus a random phase fluctuation, $\theta_n(t)$, discussed below.

$N_s(t)$ and $N_d(t)$ are the thermal noise components at the inputs of the sum and difference channels, respectively, whose one-sided power-spectral densities are assumed identical and equal to

$$\Phi_N = kT_{eq} \text{ watts/cps} \tag{3}$$

where

$$\begin{aligned} k &= 1.38 \times 10^{-23} \text{ watt-sec/°K} \\ T_{eq} &= \text{equivalent receiver noise temperature, °K.} \end{aligned}$$

The random phase fluctuation, $\theta_n(t)$, results from the frequency instability of the various oscillators in the system, principally the beacon oscillator in the satellite, since elaborate frequency stabilizing techniques are not feasible from weight and space considerations. The one-sided spectral density of the resultant phase fluctuation can be expressed as

$$\Phi_\theta = \frac{2}{\tau_{ce}\omega^2} \text{ rad}^2/\text{cps} \tag{4}$$

where

τ_{ce} = equivalent coherence time of the system oscillators.[4]

For the purpose of this analysis, the thermal noise terms in (2) will be represented by the usual in-phase and quadrature notation[4,5]

$$N(t) = X(t) \cos (\omega_b t + \theta_i) + Y(t) \sin (\omega_b t + \theta_i)$$

where $X(t)$, $Y(t)$ = independent Gaussian random voltages with one-sided power spectral density, $2\Phi_N$, with Φ_N given in (3).*

With identical receivers in the sum and difference channels which amplify the signals (2) by a factor K_0 and reduce the center frequency from ω_b to an intermediate frequency, $\omega_i = 2\pi \times 60$ mc, the input to the sum and difference channels of the coherent angle-error detector can be represented by

$$\begin{aligned} e_{s_i}(t) &= K_0[E_s(R,\beta) \cos (\omega_i t + \theta_i(t)) \\ &\quad + X_s(t) \cos (\omega_i t + \theta_i(t)) \\ &\quad + Y_s(t) \sin (\omega_i t + \theta_i(t))] \\ e_{d_i}(t) &= K_0[\eta\beta E_s(R,\beta) \cos (\omega_i t + \varphi + \theta_i(t)) \\ &\quad + X_d(t) \cos (\omega_i t + \theta_i(t)) \\ &\quad + Y_d(t) \sin (\omega_i t + \theta_i(t))] \end{aligned} \tag{5}$$

where $X_s(t)$, $X_d(t)$, $Y_s(t)$, $Y_d(t)$ have identical one-sided power density spectra $2\Phi_N$ band-limited by the IF bandwidth, B_{IF}, hence have mean-square expected values, $\overline{X^2} = \overline{Y^2} = \Phi_N B_{\text{IF}}$.

III. LINEAR ANALYSIS OF COHERENT ANGLE-ERROR DETECTION SYSTEM

A block diagram of the coherent angle-error detection system is given in Fig. 2. The coherence between the input error signal and the

* The fluctuation of θ_i due to Doppler and random phase effects is assumed to have negligible effect on the power spectral densities of $X(t)$ and $Y(t)$.

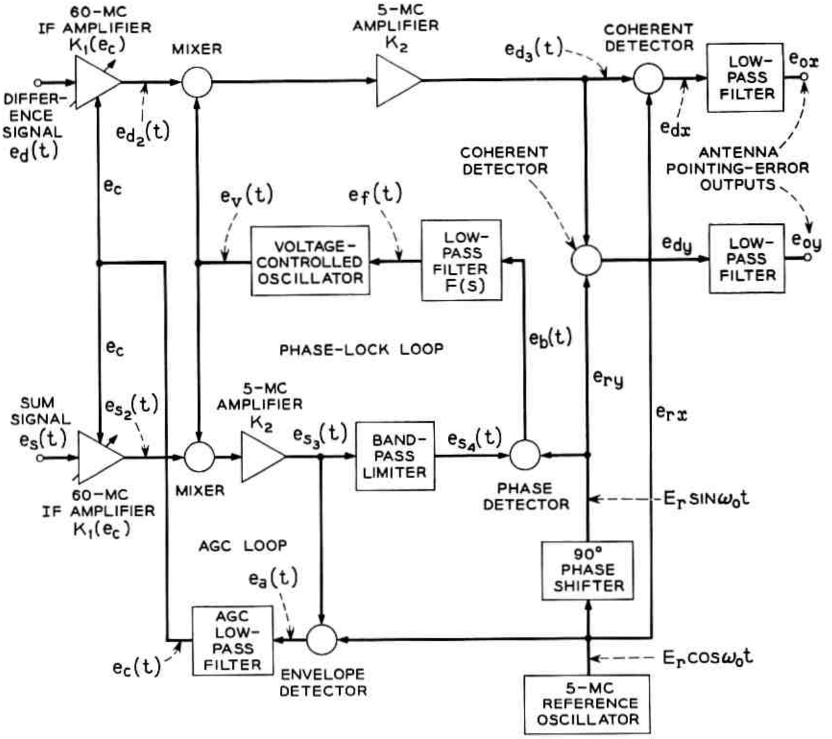


Fig. 2 — Coherent angle-error detection system block diagram.

local reference signal is achieved through the action of the phase-lock loop. The mixers and detectors indicated by circles in Fig. 2 are assumed to be ideal multipliers with unity gain. The AGC action is assumed to respond perfectly to variations in the sum channel signal level, so the gain of the 60-mc IF amplifiers in both channels can be expressed as

$$K_1(e_c) = \frac{E}{E_s(R,\beta)}, \qquad E = \text{constant}. \tag{6}$$

Using (5) and (6), the input to the mixers in each channel in Fig. 2 can be written

$$\begin{aligned} e_{s_2}(t) = K_0 E \biggl[\cos(\omega_i t + \theta_i) &+ \frac{X_s}{E_s} \cos(\omega_i t + \theta_i) \\ &+ \frac{Y_s}{E_s} \sin(\omega_i t + \theta_i) \biggr] \end{aligned} \tag{7a}$$

$$\begin{aligned} e_{d_2}(t) = K_0 E \bigg[\eta\beta \cos(\omega_i t + \varphi + \theta_i) + \frac{X_d}{E_s} \cos(\omega_i t + \theta_i) \\ + \frac{Y_d}{E_s} \sin(\omega_i t + \theta_i) \bigg] \end{aligned} \tag{7b}$$

where the dependence of E_s on R and β and the time dependence of θ_i, X, and Y is understood, but not indicated explicitly in (7) and the subsequent analysis, to simplify the equations.

The other input to the mixers is the output of the voltage-controlled oscillator (VCO) in the phase-lock loop, which has the form

$$e_v(t) = E_v \cos(\omega_v t + \theta_v) \tag{8}$$

where $\omega_v = 2\pi \times 65$ mc and $\theta_v = \theta_v(t)$ is the instantaneous phase of the VCO output, determined by the operation of the feedback loop in the sum channel, which is discussed in the next section. Multiplying (7) and (8) and taking only the low-frequency (5-mc) components gives at the outputs of the 5-mc IF amplifier in the sum and difference channels (see Fig. 2)

$$\begin{aligned} e_{s_3}(t) &= E_3 \bigg[\cos(\omega_o t - \theta_i + \theta_v) + \frac{X_s}{E_s} \cos(\omega_o t - \theta_i + \theta_v) \\ &\quad - \frac{Y_s}{E_s} \sin(\omega_o t - \theta_i + \theta_v) \bigg] \\ e_{d_3}(t) &= E_3 \bigg[\eta\beta \cos(\omega_o t - \varphi - \theta_i + \theta_v) \\ &\quad + \frac{X_d}{E_s} \cos(\omega_o t - \theta_i + \theta_v) - \frac{Y_d}{E_s} \sin(\omega_o t - \theta_i + \theta_v) \bigg] \end{aligned} \tag{9}$$

where

$$E_3 = \tfrac{1}{2} K_0 K_2 K_m E E_v \qquad K_m = \text{mixer gain, (volts)}^{-1}$$

$$\omega_0 = \omega_v - \omega_i = 2\pi \times 5 \text{ mc.}$$

The difference channel voltage, $e_{d_3}(t)$, is applied to the coherent detectors. The other input to these detectors comes from the 5-mc reference oscillator which produces signals

$$\begin{aligned} e_{r_x} &= E_r \cos \omega_0 t \\ e_{r_y} &= E_r \sin \omega_0 t \end{aligned} \tag{10}$$

for the detection of the desired x and y error components given in (1). The phase of these signals is the reference phase, $\theta_r = 0$.

The low-pass filters following the coherent detectors pass only the baseband components of the products $(e_{d_3} \cdot e_{r_x})$ and $(e_{d_3} \cdot e_{r_y})$. Using (9) and (10) these baseband components are

$$\begin{aligned} e_{dx} &= A\left[\eta\beta \cos(\varphi + \theta_i - \theta_v) + \frac{X_d}{E_s}\cos(\theta_i - \theta_v)\right. \\ &\quad \left. + \frac{Y_d}{E_s}\sin(\theta_i - \theta_v)\right] \\ e_{dy} &= A\left[\eta\beta \sin(\varphi + \theta_i - \theta_v) + \frac{X_d}{E_s}\sin(\theta_i - \theta_v)\right. \\ &\quad \left. - \frac{Y_d}{E_s}\cos(\theta_i - \theta_v)\right] \end{aligned} \tag{11}$$

where

$$A = \tfrac{1}{2}K_d E_3 E_r = \text{channel amplification factor}$$

$$K_d = \text{detector gain, (volts)}^{-1}.$$

If the phase-lock loop is tracking properly, the phase of the VCO output, θ_v, will follow closely the phase of the input, θ_i. Assuming that the rms value of $(\theta_i - \theta_v)$ is small compared to 1 radian, then the following approximations hold with high probability*

$$\begin{aligned} \sin(\theta_i - \theta_v) &\doteq (\theta_i - \theta_v) \ll 1 \\ \cos(\theta_i - \theta_v) &\doteq 1 \end{aligned}$$

and the coherent detector outputs (11) can be expressed in the approximate form

$$\begin{aligned} e_{dx} &\doteq A\eta\beta\cos\varphi - A\left(\eta\beta\sin\varphi - \frac{Y_d}{E_s}\right)(\theta_i - \theta_v) + A\,\frac{X_d}{E_s}, \\ e_{dy} &\doteq A\eta\beta\sin\varphi + A\left(\eta\beta\cos\varphi + \frac{X_d}{E_s}\right)(\theta_i - \theta_v) - A\,\frac{Y_d}{E_s}. \end{aligned} \tag{12}$$

The first term in each of the expressions in (12) is the desired error component, given in (1), amplified by the total difference channel gain, $A\eta$. The second term represents the perturbation due to the lack of perfect phase coherence, while the third term represents the contribution of thermal noise in the net noise bandwidth of the difference channel.

* The validity of these assumptions is discussed in the next section.

The achievement of good phase coherence in the detector outputs (12) over the expected range of satellite tracking conditions is the objective of the phase-lock loop analysis and design described in the following sections.

IV. PHASE-LOCK LOOP ANALYSIS

The coherent detection of the control error signals depends on the performance of the phase-lock loop in the sum channel (see Fig. 2). The loop must be capable of following the change of phase of the input signal due to frequency instability of the source and Doppler shift and also discriminate against random phase fluctuations caused by thermal noise. These requirements are somewhat contradictory, the former requiring a wide loop bandwidth and the latter requiring a narrow loop bandwidth. Proper design of the phase-lock loop must therefore be based on the best compromise of these requirements consistent with the expected variation of the signal phase and the expected random phase fluctuation.

The sum channel voltage, $e_{s_3}(t)$ at the input to the bandpass limiter is given in (9). The effect of the limiter can be closely approximated as multiplying this voltage by a limiter suppression factor, α, which increases from 0 to 1 as the signal-to-noise ratio at the limiter input increases from 0 to ∞. This limiter action is discussed in Appendix A.

The limiter output voltage, $e_{s_4}(t) = \alpha e_{s_3}(t)$ is applied to the phase detector in the sum channel, together with the reference signal, e_{r_y}, given in (10). The baseband component of the phase detector output is therefore

$$e_b(t) = (e_{r_y} \cdot \alpha e_{s_3})_{\text{baseband}}$$

or, from (9) and (10)

$$e_b(t) = \alpha A\left[\left(1 + \frac{X_s}{E_s}\right) \sin (\theta_i - \theta_v) - \frac{Y_s}{E_s} \cos (\theta_i - \theta_v)\right]. \quad (13)$$

To develop an approximate linear model for the phase-lock loop, the following two assumptions are made:

(*i*) the phase error $(\theta_i - \theta_v)$ is sufficiently small to permit the approximations $\sin (\theta_i - \theta_v) \doteq \theta_i - \theta_v$, $\cos (\theta_i - \theta_v) \doteq 1$, and

(*ii*) the noise component, $X(t)$, is assumed small in the rms sense compared to the signal amplitude, E_s.

Using these assumptions, the phase detector output (13) can be written in the approximate form

$$e_b(t) \doteq \alpha A\left[\theta_i(t) - \theta_v(t) - \frac{Y_s(t)}{E_s}\right]. \quad (14)$$

The low-pass filter passes the baseband component $e_b(t)$, producing a voltage $e_f(t)$ which causes the frequency of the VCO to vary from the center frequency, $\omega_v = 2\pi \times 65$ mc, with a proportionality constant, K_v radians per second per volt. The instantaneous phase of the VCO output is therefore

$$\theta_v(t) = K_v \int e_f(t)\, dt, \text{ radians.}$$

The transfer function of the VCO over the range in which this proportionality holds is then

$$\frac{\theta_v(s)}{e_f(s)} = \frac{K_v}{s}. \tag{15}$$

For a phase-lock loop which is stable over a large range of loop gain variations and which has zero steady-state phase error for a phase ramp input (step frequency change) the low-pass filter should have a transfer function of the form

$$F(s) = \frac{e_f(s)}{e_b(s)} = \frac{1 + \tau s}{s}. \tag{16}$$

This transfer function has been shown[6] to yield optimum loop performance for phase ramp inputs in the presence of white noise, where the performance measure is the mean-square error caused by noise plus the integrated-squared transient error to the ramp input.

The transfer function (16) can be closely approximated by the operational amplifier circuit shown in Fig. 3, which has the transfer function

$$F(s) = -\mu \frac{1 + R_2Cs}{1 + \mu R_1Cs} \tag{17}$$

where $-\mu$ is the amplifier gain under load without feedback. The

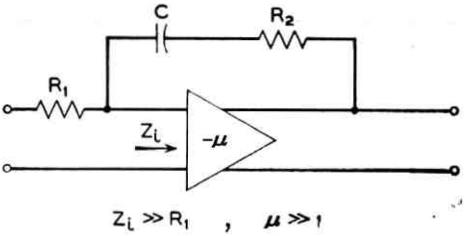


Fig. 3 — Operational amplifier low-pass filter circuit.

assumptions made in deriving (17) are that the input impedance, Z_i, of the amplifier is very large compared to R_1 and that $\mu \gg 1$ (typically 10^6) in the low-frequency range. With $R_1C = 1$ second and μ of the order of 10^6, the transfer function (17) reduces to

$$F(s) \doteq -\frac{1 + R_2Cs}{s} \tag{18}$$

which is the desired form (16), with $R_2C = \tau$. The negative sign in (18) is incidental provided the sign of the total gain around the loop gives negative feedback.

Using (14), (15), and (16), the linear equivalent block diagram for the phase-lock loop is shown in Fig. 4. The objective of the design is to minimize the mean-square value of the random phase error, $\theta_e(t) = \theta_i(t) - \theta_v(t)$, consistent with the requirements on the dynamic tracking capability of the loop.

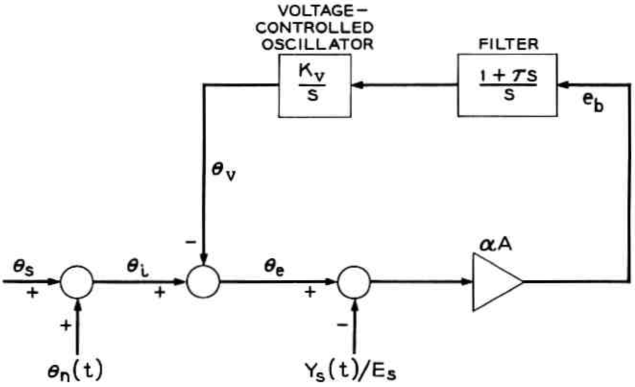


Fig. 4 — Block diagram of phase-lock loop based on linear analysis.

The total transfer function around the loop is

$$G(s) = \frac{\theta_v(s)}{\theta_e(s)} \doteq \alpha K \frac{1 + \tau s}{s^2} \tag{19}$$

where $K = AK_v$. [If R_1C is not unity, as was assumed in (18), then $K = AK_v/R_1C$.]

For the analysis of the phase error, θ_e, due to the noise sources $\theta_n(t)$ and $Y_s(t)$, we let the signal phase θ_s be zero and obtain from Fig. 4 the transfer function for θ_e in terms of $G(s)$ in (19)

$$\theta_e(s) = \left[\frac{1}{1 + G(s)}\right]\theta_n(s) + \left[\frac{G(s)}{1 + G(s)}\right]\frac{Y_s(s)}{E_s}. \tag{20}$$

Since the one-sided power density spectrum of $\theta_n(t)$ is Φ_θ, given in (4), and that of $Y_s(t)/E_s$ is $2\Phi_N/E_s^2$ where Φ_N is given in (3), and since they are uncorrelated random variables, the one-sided power density spectrum of the random phase error, θ_e, is

$$\Phi_{\theta_e} = \left| \frac{1}{1 + G(j\omega)} \right|^2 \frac{2}{\tau_{ce}\,\omega^2} + \left| \frac{G(j\omega)}{1 + G(j\omega)} \right|^2 \frac{2\Phi_N}{E_s^2}. \qquad (21)$$

The mean-square value of this random phase error is then, from (19), (21) and integral tables[7]

$$\sigma_e^2 = \int_0^\infty \Phi_{\theta_e} \frac{d\omega}{2\pi} = \frac{1}{\tau_{ce}} \left(\frac{1}{2\alpha K \tau} \right) + \frac{2\Phi_N}{E_s^2} \left(\frac{\alpha K \tau}{4} + \frac{1}{4\tau} \right), \text{rad}^2. \qquad (22)$$

This expression for the mean-square phase error can be written in terms of the undamped natural frequency, ω_n, and the damping ratio, ζ, of the phase-lock loop, as

$$\sigma_e^2 = \frac{1}{4\zeta\omega_n\,\tau_{ce}} + \frac{2\Phi_N}{E_s^2}\,\omega_n \left(\frac{1 + 4\zeta^2}{8\zeta} \right) \qquad (23)$$

since $\omega_n = \sqrt{\alpha K}$, and $2\zeta = \tau\sqrt{\alpha K}$.

The proper operation of the phase-lock loop depends upon the magnitude of the phase error remaining less than $\pi/2$ radians. For the phase error due to random fluctuations we can require only that the probability of its magnitude exceeding $\pi/2$ radians be very small. A criterion for this which has been chosen[4,8] as a realistic measure of the threshold of the phase-lock loop is that the mean-square value of the total phase error be restricted by

$$\sigma_e^2 \leqq \tfrac{1}{8}\ \text{rad}^2. \qquad (24)$$

For a normally distributed random phase error with zero mean and variance σ_e^2, this criterion implies that the probability of exceeding $\pi/2$ is exceedingly small (about 10^{-5}). This criterion also gives validity to the first assumption made above in obtaining the linear model of the phase-lock loop, namely that $\sin\theta_e \doteq \theta_e$ and $\cos\theta_e \doteq 1$. The error in these approximations is quite small provided

$$|\,\theta_e\,| \leqq 0.57 \text{ radian}$$

which holds with approximately 90 per cent probability when the condition (24) is satisfied.

The second assumption made above for the linear model is not strictly justified in the region of threshold, where the signal-to-noise ratio in the 3-kc bandwidth at the phase detector input will typically be less than unity. However, comparison studies with a digital computer simulation[9] which was implemented for the angle-error detector in the precision tracker system have indicated that the linear model estimate of the mean-square phase error (23) is sufficiently accurate even in the vicinity of threshold to justify its use for the analytical design optimization of the phase-lock loop. The digital computer baseband model of the phase-lock loop includes the in-phase noise term as well as the quadrature noise term in (13), and the sine and cosine operations of the phase detector (see Ref. 9). A comparison of the computer simulation data with the linear analysis data for the design examples considered in Section VI is given at the end of that section.

An important parameter in the phase-lock loop analysis is the effective noise bandwidth, B_L, of the loop, defined by

$$B_L = \int_0^\infty \left| \frac{G(s)}{1 + G(s)} \right|^2 df.$$

The loop noise bandwidth for the system under consideration has already been evaluated in the second terms of (22) and (23), namely

$$B_L = \frac{\alpha K_T}{4} + \frac{1}{4\tau} = \omega_n \left(\frac{1 + 4\zeta^2}{8\zeta} \right) \tag{25}$$

which increases with the limiter suppression factor α and hence increases with the signal-to-noise power ratio at the limiter input. This is the desired adaptive feature which the limiter provides in the phase-lock loop operation, since for small S/N (long-range condition) the loop bandwidth is small, decreasing the mean-square error due to thermal noise, while for large S/N (short-range condition) the loop bandwidth is large, providing improved phase tracking accuracy for the greater Doppler frequency rate of change occurring at short range.

The Doppler frequency variation as a function of time is approximated by a frequency "ramp" input having a constant slope of magnitude $\dot{\omega}$ for a duration T_d, as indicated in Fig. 5. Since this approximate function represents a somewhat more difficult variation for the loop to track than the actual Doppler variation, the evaluation of the loop tracking accuracy and transient behavior based on the approximate input should serve as a conservative basis for design.

Neglecting the thermal noise terms in (13), the phase detector output reduces to

$$e_b(t) = \alpha A \sin [\theta_i(t) - \theta_v(t)] \text{ volts} \tag{26}$$

where $\theta_i(t) = \frac{1}{2}\dot{\omega}t^2$ is the phase input corresponding to the frequency

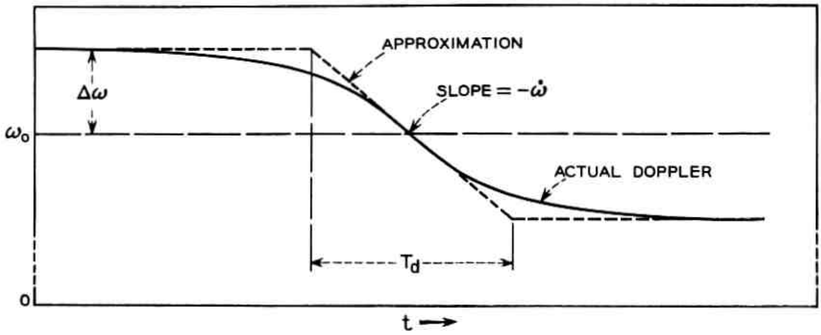


Fig. 5 — Typical Doppler frequency variation and piecewise linear approximation.

"ramp" input discussed above. Since this output is bounded in magnitude by $\pm A$ volts, the low-pass filter which follows the phase detector should have a linear input dynamic range of at least this magnitude. In addition, the voltage-controlled oscillator should have a linear frequency range at least as large as the expected maximum Doppler shift. With these two design requirements satisfied, the only essentially nonlinear element in the phase-lock loop circuit is the phase detector (see Fig. 6a).

An analysis of the response of this nonlinear circuit to a frequency "ramp" input is given in Appendix B. From this analysis it is concluded that for adequate phase-lock tracking of the Doppler shift, the loop gain should satisfy the condition

$$\alpha K > 2\dot{\omega}_{\max} \tag{27}$$

where $\dot{\omega}_{\max}$ is the maximum Doppler rate in rad/sec^2.

When condition (27) is satisfied, the steady-state phase error due to this Doppler rate will not exceed $\pi/6$ radian (see Appendix B), and the loop response will closely approximate that of the linear second-order circuit, shown in Fig. 6(b). For this circuit with the input $\theta_i(t) = \frac{1}{2}\dot{\omega}t^2$, the Laplace transform of the phase error is

$$\theta_e(s) = \frac{\dot{\omega}}{s[s^2 + 2\zeta\omega_n s + \omega_n^2]}$$

where, as before, $\omega_n = \sqrt{\alpha K}$, $2\zeta = \tau\sqrt{\alpha K}$, and $K = AK_v$. The time response of this type of linear second-order system is thoroughly discussed in elementary texts on linear circuits or control system theory.*

* See, for example, Ref. 10.

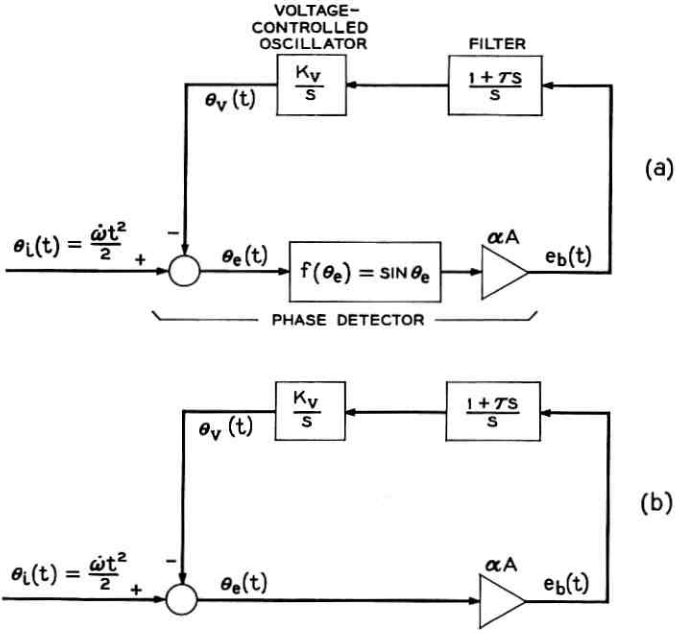


Fig. 6 — Phase-lock loop circuits for Doppler-shift analysis: (a) equivalent circuit including nonlinearity of phase detector; (b) approximate linear circuit, valid for $| \theta_e | \leqq \pi/6$ radian.

The two per cent settling time* for the transient of the phase error is approximately

$$T_s \doteq \frac{8\zeta}{\omega_n} = 4\tau \text{ seconds} \tag{28}$$

for the overdamped case ($\zeta > 1$). This also serves as a good approximation of the duration of the transient for the underdamped case in the range $0.7 < \zeta < 1$.

V. PHASE-LOCK LOOP DESIGN

The design procedure which will be followed is first to ascertain the system requirements necessary for minimum acceptable performance and second to optimize the performance within the range of parameter adjustment available to the designer.

* Defined as the time required for the transient response to settle down to within two per cent of the steady-state value. See Ref. 10.

The primary factors external to the phase-lock loop which affect its performance capability are:

(*i*) *The equivalent coherence time*, τ_{ce}, of the input signal, which characterizes the random frequency fluctuations of the various system oscillators which affect the instantaneous frequency of the signal into the phase-lock loop. The coherence time should be made as large as possible, but is primarily limited by the frequency stability achievable in the small satellite transmitter and hence will be considered a fixed parameter not available for phase-lock loop design adjustment.

(*ii*) *The noise-to-signal power ratio* at the input to the phase-lock loop. Since this ratio varies with the range of the satellite and with the IF bandwidth preceding the loop, it is desirable to characterize the relative "noisiness" of the system independent of range and bandwidth variations. The thermal noise power is given by

$$N = \Phi_N B_{\text{IF}} = kT_{eq}B_{\text{IF}} \text{ watts}$$

while the average signal power is[11,12]

$$S = \frac{E_s^2}{2} = \frac{P_T G_T A_r}{4\pi R^2} \text{ watts}$$

where

P_T = transmitter power, watts,
G_T = transmitter antenna gain,
R = transmitter to receiver range, and
A_r = effective area of receiving antenna (same units as R^2).

The noise-to-signal power ratio can then be expressed as

$$\frac{N}{S} = \frac{2\Phi_N}{E_s^2} B_{\text{IF}} = \left(\frac{4\pi k T_{eq}}{P_T G_T A_r}\right) R^2 B_{\text{IF}}. \tag{29}$$

The factors in parenthesis in (29) are constant* characteristics of the satellite-ground system and will be represented by a single constant called the "receiver noise index," denoted by the symbol k_r, and having the units of seconds/(distance)^2. Then,

$$\frac{N}{S} = \frac{2\Phi_N}{E_s^2} B_{\text{IF}} = k_r R^2 B_{\text{IF}}. \tag{30}$$

The noise index, k_r, is taken as the second fixed parameter† of the

* The transmitter antenna gain, G_T, will not actually be constant unless the radiation pattern is uniform or the satellite is properly attitude controlled. It is assumed essentially constant in this study.

† From (30) it is apparent that the receiver noise index, k_r, corresponds to the unit bandwidth noise-to-signal power ratio in the receiver when the satellite is at unit distance from the receiver.

external system affecting loop design.

(*iii*) *The satellite orbit characteristics.** The orbit characteristics affect the performance capability of the phase-lock loop as follows: First, the maximum and minimum communication range will determine the variation of the noise-to-signal ratio as shown in (30). Second, the maximum range rate, $\dot{R}_{\max}$, will determine the maximum Doppler shift

$$|\Delta\omega|_{\max} = \frac{\omega_b |\dot{R}|_{\max}}{c} \text{ rad/sec} \tag{31}$$

where

$$\begin{aligned} \omega_b &= 2\pi \times \text{satellite beacon frequency, and} \\ c &= \text{velocity of light.} \end{aligned}$$

Finally, the maximum range acceleration, $\ddot{R}_{\max}$, will determine the maximum Doppler rate

$$\dot{\omega}_{\max} = \frac{\omega_b \ddot{R}_{\max}}{c} \text{ rad/sec}^2. \tag{32}$$

Since the maximum $\dot{\omega}$ occurs at the range of closest approach for all possible satellite passes, the limiter suppression factor α will be at its maximum value, giving the largest loop gain, $\alpha_{\max}K$. From (27), the basic lower limit on the fixed loop gain constant, K, is then

$$K > \frac{2\dot{\omega}_{\max}}{\alpha_{\max}} \text{ sec}^{-2}. \tag{33}$$

The total mean-square error due to random fluctuations in the system is given by (22) or (23). Using (25) and (30), the mean-square error can be expressed as

$$\sigma_e^2 = \frac{1 + (1/4\zeta^2)}{8\tau_{ce} B_L} + k_r R^2 B_L, \text{ rad}^2. \tag{34}$$

The second term in (34), due to thermal noise in the system characterized by k_r, increases with the square of the range. This increase is somewhat offset by the reduction of the loop bandwidth, B_L, which decreases approximately as the first power of the range. This reduction of B_L, however, increases the mean-square error due to random phase fluctuations of the signal, given by the first term in (34). Hence, the total mean-square error will be maximum at the longest range condition.

* Derivation of pertinent orbit characteristics is given in Appendix C.

From (24), a requirement which the loop design must satisfy is therefore given by

$$(\sigma_e^2)_{R=R_{\max}} \leqq \tfrac{1}{8}\ \text{rad}^2. \tag{35}$$

As a function of B_L, the minimum possible value of σ_e^2 is, from (34)

$$\min \sigma_e^2 = R\left[\frac{k_r}{2\tau_{ce}}\left(1 + \frac{1}{4\zeta^2}\right)\right]^{\frac{1}{2}} \tag{36}$$

when

$$B_L = \text{opt}\ B_L = \frac{1}{R}\left(\frac{1 + (1/4\zeta^2)}{8\tau_{ce}k_r}\right)^{\frac{1}{2}}. \tag{37}$$

It is apparent from (36) that unless

$$\frac{k_r R_{\max}^2}{\tau_{ce}} < \frac{1}{32} \tag{38}$$

the condition (35) cannot be satisfied for any values of loop bandwidth and damping ratio, or equivalently for any values of loop gain, αK, and time constant, τ. This corresponds to the minimum signal power condition given by Develet.[4]

The two basic design requirements which the external system imposes upon the phase-lock loop design are given by the inequalities (33) and (38). The first requirement does not appear critical, since for satellite communication systems the value of $\dot{\omega}_{\max}$ is unlikely to exceed about 2×10^4 rad/sec^2 (see Appendix C). Since $\alpha_{\max} \doteq 1$, condition (33) requires that

$$K > 4 \times 10^4\ \text{sec}^{-2}$$

while values of K on the order of ten times this lower limit are achievable in present phase-lock loop designs.

The second requirement, (38), is more critical, since it depends entirely upon the fixed parameters of the external system. If the inequality (38) is not satisfied, either the system noise index must be decreased or the coherence time of the satellite transmitter must be increased before satisfactory operation of the phase-lock loop can be achieved at maximum range. When this requirement is satisfied, the optimum loop design with respect of the total mean-square error is achieved by making the loop bandwidth equal to the optimum value, given by (37), and by making the damping ratio, ζ, as large as is consistent with satisfactory transient response of the loop.

Fig. 7 depicts in graphical form the design requirements discussed above. For a given value of

$$\tau_{ce}{}' = \frac{\tau_{ce}}{1 + (1/4\zeta^2)} \tag{39}$$

the contour of $\sigma_e{}^2 = \frac{1}{8}$ rad^2 is given as a function of k_rR^2 and the loop bandwidth B_L. Within this contour the mean-square error is less than the threshold value of $\frac{1}{8}$ rad^2 and has its minimum possible value for a given k_rR^2 on the dashed line associated with each contour. The vertical line defined by $k_rR^2 = k_rR_{\max}{}^2$ must intersect the contour for the appropriate value of $\tau_{ce}{}'$ in order for satisfactory loop design to be achieved. If this critical requirement is satisfied, then the optimum design is given by adjusting the loop bandwidth, B_L, to equal the value obtained from the dashed line within the contour at the particular value of k_rR^2.

The input-adaptive adjustment of B_L by means of the limiter sup-

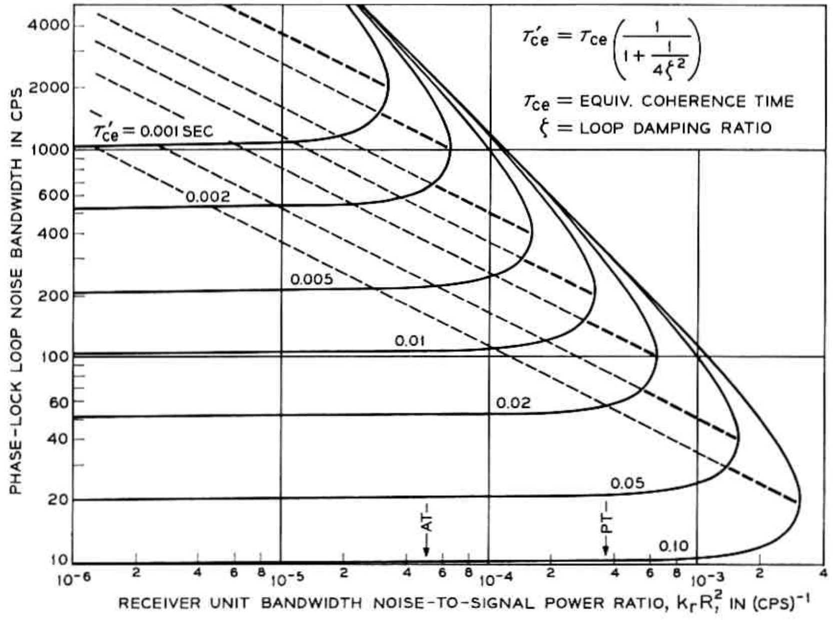


Fig. 7 — Phase-lock loop threshold contours. Inside the contour for a particular $\tau_{ce}{}'$ the mean-square phase error, $\sigma_e{}^2$, is less than the threshold value of 0.125 rad^2, and has its minimum value for a given k_rR^2 on the dashed line associated with the contour. Indicated on the abscissa are the expected maximum values of k_rR^2 for the autotrack and the precision tracker.

pression factor, α, can approximate this optimum adjustment of B_L. The optimum B_L, given by (37), varies inversely with the range, R. The actual value of B_L, given by (25), varies directly with α, but α itself varies inversely with R, as shown in Appendix A.

Before the optimum bandwidth can be determined from (37) or Fig. 7, the value of ζ at maximum range must be chosen. As was pointed out at the end of Section IV, the settling time of the loop is approximately 4τ seconds. If T_M denotes the maximum tolerable settling time, then this requires that $4\tau \leqq T_M$. This also places an upper limit on ζ when the loop bandwidth, B_L, is fixed at the optimum value (37), since, from (25), B_L and ζ are related to τ as follows

$$B_L = \frac{1}{4\tau}(4\zeta^2 + 1).$$

To satisfy the maximum settling time restriction, then

$$4\tau = \frac{4\zeta^2 + 1}{B_L} \leqq T_M \tag{40}$$

which implies that ζ cannot be arbitrarily increased while holding B_L fixed at the optimum B_L. Using (37), (39) and (40), the upper bound on ζ when $B_L = \text{opt } B_L$ can be expressed in terms of T_M and the external system parameters as

$$\zeta^2 \leqq \zeta_M^{\ 2} \equiv \tfrac{1}{8}\left[\left(\frac{T_M^{\ 2}}{2\tau_{ce}k_r R^2} + 1\right)^{\frac{1}{2}} - 1\right]. \tag{41}$$

The lower bound on ζ when $B_L = \text{opt } B_L$ follows from the requirement that min $\sigma_e^{\ 2} < \frac{1}{8}$ rad^2. Using (36) and some algebraic manipulation

$$\zeta^2 > \zeta_m^{\ 2} \equiv \frac{8k_r R^2}{\tau_{ce} - 32k_r R^2}, \tag{42}$$

which is a finite real lower bound on ζ only when the basic requirement (38) is satisfied.

Since ζ varies with range, the upper and lower bounds given above should be evaluated at the same range, R. It is apparent from (41) and (42) that the least upper bound and the greatest lower bound on ζ both occur at $R = R_{\max}$, and that the bounds constrict the range of ζ as the noise index, k_r, increases. As k_r decreases, the limits separate, allowing a wide range of ζ. However, from the point of view of relative stability and fast transient response, ζ should not be less than 0.7 nor greater than about two; also, from the point of view of minimizing $\sigma_\epsilon^{\ 2}$, it can be seen from (36) that there is negligible improvement in increasing ζ beyond about two. These restrictions on the range of ζ in the optimum design of the phase-lock loop can be summarized by

$$\max\,(0.7,\zeta_m) < \zeta \leqq \min\,(2.0,\zeta_M) \tag{43}$$

where ζ_M and ζ_m are the upper and lower bounds defined in (41) and (42), respectively.

Assuming that the set of positive real values of ζ satisfying (43) is not empty, the optimum choice of ζ (with respect to minimizing the mean-square phase error) is the least upper bound value given in (43). Using this value, the optimum value of the loop noise bandwidth, B_L, at maximum range is determined from (37) or Fig. 7.

The selection of the optimum ζ and B_L at maximum range fixes the value of the loop filter time constant, $\tau = R_2C$, as

$$\tau = \frac{4\zeta_{\text{opt}}^{\;2}+1}{4(B_L)_{\text{opt}}} = \begin{cases} T_M/4\;, & \zeta_{\text{opt}} = \zeta_M \\ \dfrac{4.25}{(B_L)_{\text{opt}}}\,, & \zeta_{\text{opt}} = 2 \end{cases} \tag{44}$$

and the value of loop gain at maximum range as:

$$\alpha_{\min} K = \frac{4\zeta_{\text{opt}}^{\;2}}{\tau^2}\,. \tag{45}$$

When the IF bandwidth is specified, $\alpha_{\min}$ is determined from Fig. 11 in Appendix A, with $R = R_{\max}$. Knowing $\alpha_{\min}$, the loop gain constant, K, is then determined from (45). Since K has the lower bound given by (33) and certainly an upper bound dictated by practical equipment considerations, the range of α may have to be controlled through the selection of B_{IF} to give a value of K which is compatible with these bounds.

VI. *TELSTAR* SYSTEM DESIGN EXAMPLES

To illustrate this design approach, sample designs will be considered for both the precision tracker and the autotrack systems for the Telstar experimental program. Based on analysis of the expected orbit (see Appendix C), and on preliminary system data, the system parameters assumed for the design examples are given in Table I. While the IF bandwidth values are not necessarily fixed, the 200-kc bandwidth given in Table I for the 5-mc channels (see Fig. 2) is desirable from practical considerations.

Considering first the dynamic tracking capability in the absence of

TABLE I—ASSUMED SYSTEM PARAMETERS FOR DESIGN EXAMPLES

Equivalent coherence time, $\tau_{ce} = 0.02$ sec
Receiver noise index (see Section V)
PT system ...... $k_r = 1.5 \times 10^{-5}$ sec/knm (kilo-nautical miles)
AT system ...... $k_r = 2.0 \times 10^{-6}$ sec/knm
Orbital data:
Maximum range, $R_{\max} = 5$ knm
Minimum range, $R_{\min} = 0.5$ knm
Maximum Doppler shift, $|\Delta f| \doteq 100$ kc
Maximum rate of shift, $|\dot{\omega}| = 5620$ rad/sec², at 0.5 knm
IF bandwidth preceding limiter, $B_{IF} = 200$ kc (both systems)
Nominal range of loop gain constant, $K = 10^5$ to 10^6 sec⁻²
Maximum tolerable settling time, $T_M = 0.1$ sec

noise, the dynamic range of the voltage-controlled oscillator in the phase-lock loop should be about ±150 kc to avoid saturation effects when the maximum Doppler shift is ±100 kc, and to allow for some drift of the center frequency during operation.

The range of variation of the limiter suppression factor, α, can be determined from Fig. 11 in Appendix A, using the values of k_r, R, and B_{IF} in Table I

	Precision Tracker	Autotrack
Max. range, $\alpha = \alpha_{\min}$:	0.10	0.28
Min. range, $\alpha = \alpha_{\max}$:	0.79	0.97

Since the maximum Doppler rate occurs at minimum range, the condition (33) that the maximum steady-state error be less than $\pi/6$ radian requires that the loop gain constant, K, satisfy

	Precision Tracker	Autotrack
$K > \dfrac{2 \times 5620}{\alpha_{\max}}$	$= 1.42 \times 10^4$	1.15×10^4

Both of these lower limits are well below the lower nominal value of 10^5 sec^{-2} given in Table I. Using this lower nominal value the maximum steady-state error in tracking the Doppler shift will be less than 0.08 radian for both systems [see (50), Appendix B].

The upper and lower bounds on ζ at maximum range condition, using (43) and Table I, are

Precision Tracker	Autotrack
$\left.\begin{matrix}0.61\\0.70\end{matrix}\right\} < \zeta \leqq \left\{\begin{matrix}1.76\\2.0\end{matrix}\right.$	$\left.\begin{matrix}0.15\\0.70\end{matrix}\right\} < \zeta \leqq \left\{\begin{matrix}2.95\\2.0\end{matrix}\right. .$

Taking the least upper bound as the optimum value gives at $R = R_{\text{max}}$

Precision Tracker	Autotrack
$\zeta_{\text{opt}} = 1.76$	$\zeta_{\text{opt}} = 2.0$

The min σ_e^2 and optimum B_L at maximum range, using (36) and (37) and the above values of ζ_{opt}, are

$$\begin{array}{l|l} \min \sigma_e^2 = 0.1 \text{ rad}^2 & \min \sigma_e^2 = 0.036 \text{ rad}^2 \\ \text{opt } B_L = 135 \text{ cps} & \text{opt } B_L = 364 \text{ cps.} \end{array}$$

Finally, using (44) and (45) and the values of $\alpha_{\min}$ given above for the 200-kc IF bandwidth, the optimum values for the phase-lock loop constants τ and K are

$$\begin{array}{l|l} \tau = 0.025 \text{ sec} & \tau = 0.012 \text{ sec} \\ K = 2 \times 10^5 \text{ sec}^{-2} & K = 4 \times 10^5 \text{ sec}^{-2}. \end{array}$$

Since both values of K are greater than the lower bounds given above and are within the nominal range given in Table I, the desired 200-kc bandwidth need not be changed.

The performance of the systems as a function of the range, R, using these design parameters, is summarized by the curves shown in Figs. 8 and 9. These curves show how the input adaptive loop gain, αK, helps provide near-optimum design at ranges other than the maximum range where the optimum design was accomplished.

The mean-square phase error obtained from this linear analysis was compared with values obtained from the more accurate digital computer simulation of the phase-lock loop described in a separate paper.[9] Results for the two design examples considered above, which are shown in Fig. 10, demonstrate that although the error in the linear model estimate increased somewhat as the signal-to-noise ratio decreases, there is no drastic breakdown in the accuracy of the linear model. For these and two other designs tested, the linear analysis estimate of the threshold signal-to-noise ratio was within 1.5 db of the digital computer results.

VII. CONCLUDING REMARKS

The material presented in this paper was part of a design study conducted during the initial construction of the ground station tracking systems for the Telstar program. At that time there was some concern about the capability of the angle-error detector to maintain phase-lock at the longer satellite ranges, due to the small beacon signal power available and the uncertainty about the coherence time of the 4080-mc

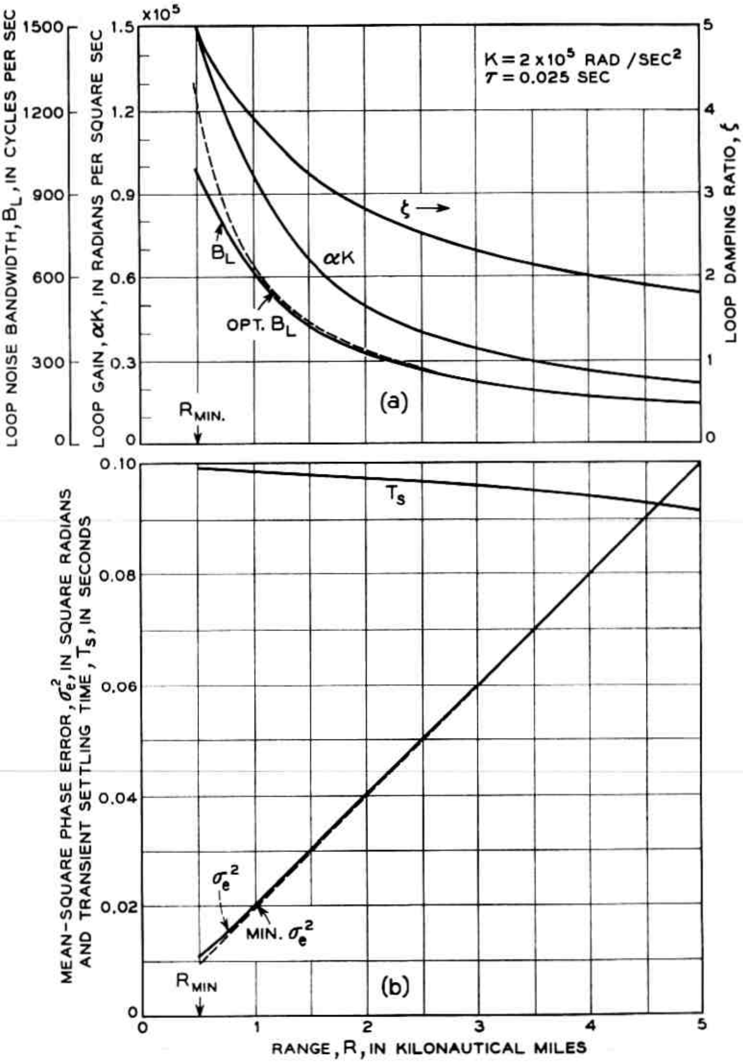


Fig. 8 — Performance curves for precision tracker (PT) optimum design example: (a) phase-lock loop parameter variations with range from ground station to satellite; (b) phase-lock loop performance measures as a function of range.

beacon signal. This was particularly critical in the detection system for the precision tracker because of the significantly lower gain of the precision tracker antenna compared to the horn-reflector antenna. For this reason, the means for achieving optimum design of the phase-lock

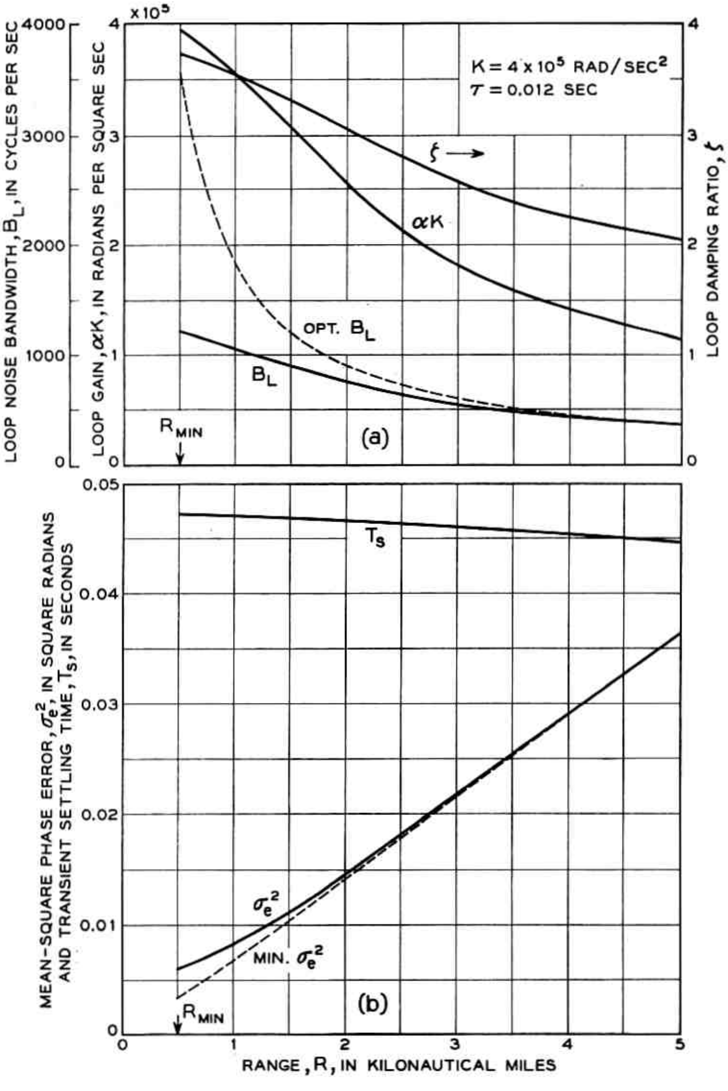


Fig. 9 — Performance curves for autotrack optimum design example: (a) phase-lock loop parameter variations with range from ground station to satellite; (b) phase-lock loop performance measures as a function of range.

loop at maximum range was a crucial consideration in the initial design. One of the fortunate results contributing to the highly successful operation of the first Telstar satellite experiment was the excellent phase stability achieved in the satellite beacon transmitter. Measurements of the mean-square phase error under strong-signal conditions

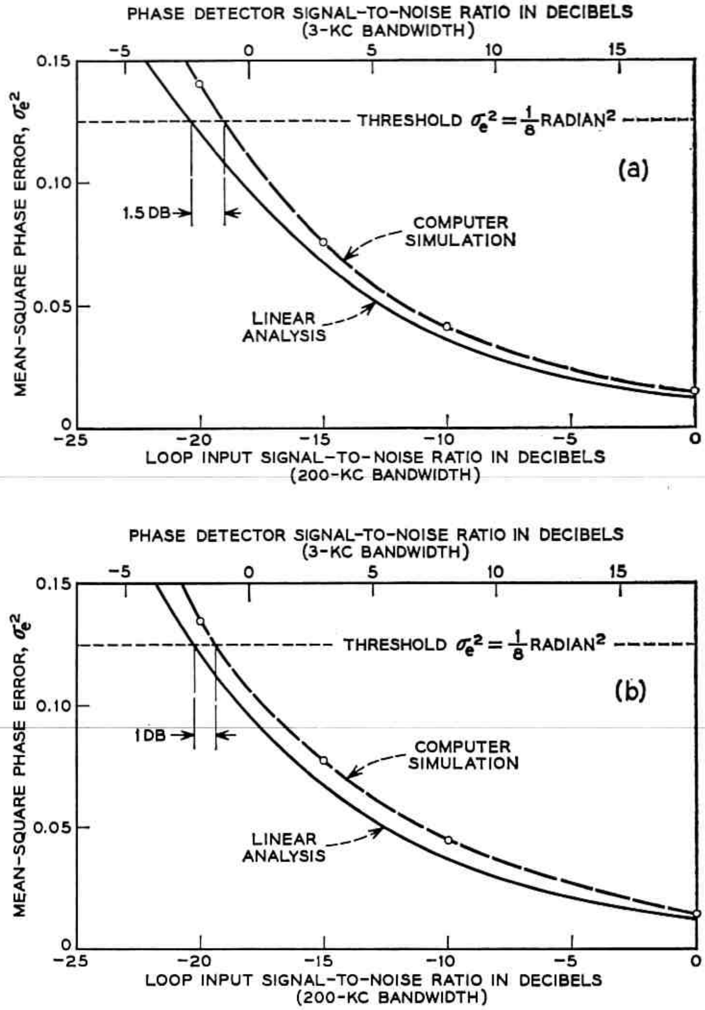


Fig. 10 — Comparison of mean-square error predicted by linear analysis with results from digital computer simulation of phase-lock loop, using design values from the two examples in Section VI.

(where the major contribution to the error is due to phase jitter in the beacon signal) indicated an effective coherence time of about 0.1 second instead of the 0.02-second value assumed for the design examples above. This higher coherence time makes the noise bandwidth and damping ratio adjustment in the phase-lock loop much less critical, as can be seen from the threshold contours in Fig. 7.

More detailed descriptions of the final design and the performance of the tracking systems in the first Telstar experiments are given in a series of papers [1,2,3] appearing in an earlier issue of this journal.

APPENDIX A

Effective Gain of Bandpass Limiter

Two factors which characterize the operation of an ideal bandpass limiter are:

(*i*) The total power output of the limiter remains constant,[11] i.e.

$$S_0 + N_0 = C \tag{46}$$

where S_0 = output signal power, N_0 = output noise power.

(*ii*) When a sinusoidal signal and narrow-band Gaussian noise are applied to the input of a bandpass limiter, the output signal to noise ratio is related to the input signal to noise ratio by[13]

$$\frac{S_0}{N_0} = \lambda \frac{S_i}{N_i} \tag{47}$$

where the factor λ, given in Fig. 5 of Ref. 13, varies from $\pi/4$ to 2 as the input signal to noise ratio varies from zero to infinity.

When no noise is present, we assume that the output signal power equals the input signal power (any fixed gain in the bandpass limiter is absorbed into the loop gain constant, K). Then, from (46), when $N_0 = 0$

$$S_0 = C = S_i$$

so that, when noise is present

$$S_0 + N_0 = S_i\,.$$

A little algebraic manipulation of this expression gives

$$\frac{S_0}{S_i} = \frac{\dfrac{S_0}{N_0}}{1 + \dfrac{S_0}{N_0}}\,.$$

Using (47), we obtain

$$\frac{S_0}{S_i} = \frac{\lambda \dfrac{S_i}{N_i}}{1 + \lambda \dfrac{S_i}{N_i}} \equiv \alpha^2, \tag{48}$$

where α is called the limiter suppression factor.[11]

Thus, the limiter has the effect of reducing the signal power from S to $\alpha^2 S$, and as a consequence the effective loop gain is reduced from K to αK. The factor α defined in (48) varies from 0 to 1 as the input signal-to-noise ratio varies from 0 to ∞. From (30) in Section V, the input signal-to-noise ratio can be expressed as

$$\frac{S_i}{N_i} = (k_r R^2 B_{\text{IF}})^{-1}.$$

Using this expression in (48) gives

$$\alpha = \left[1 + \frac{k_r R^2 B_{\text{IF}}}{\lambda}\right]^{-\frac{1}{2}},$$

which shows the inverse dependence of α on the satellite range, R. With values of λ obtained from Fig. 5 of Ref. 13, the limiter suppression factor, α, is plotted as a function of the input signal-to-noise ratio in Fig. 11.

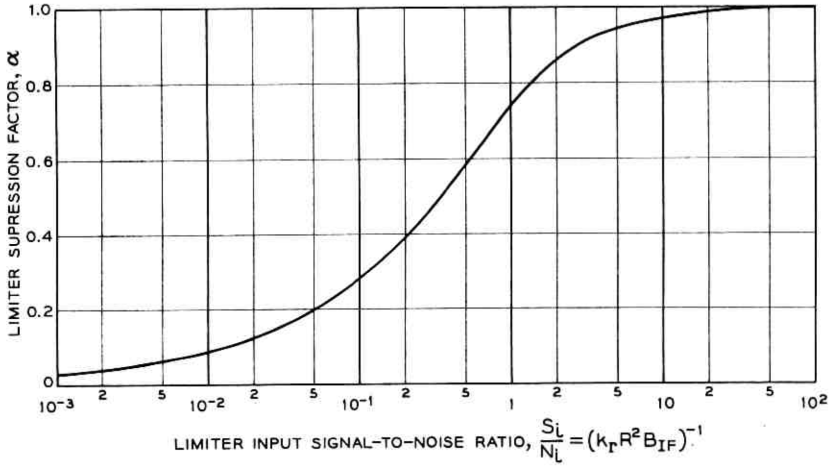


Fig. 11 — Limiter suppression factor, α, as a function limiter input signal-to-noise ratio.

APPENDIX B

Nonlinear Analysis of the Phase-Lock Loop

The phase-lock loop equivalent circuit, which includes the sine-function nonlinearity of the phase detector, is shown in Fig. 6(a). In terms of the phase error, θ_e, and its time derivative, ω_e, the differential equations governing the response of this circuit to the frequency-ramp input, $\omega_i = \dot{\omega} t$, are

$$\begin{aligned} \frac{d\theta_e}{dt} &= \omega_e \\ \frac{d\omega_e}{dt} &= \dot{\omega} - 2\zeta\omega_n\omega_e \cos\theta_e - {\omega_n}^2 \sin\theta_e \end{aligned} \tag{49}$$

where, as in Section IV, we define ${\omega_n}^2 \equiv \alpha K$, $2\zeta \equiv \tau\omega_n$.

The values of θ_e and ω_e which satisfy the equilibrium condition that the right-hand side of (49) vanish are

$$(\theta_e)_{\text{eq}} = \sin^{-1}\left(\frac{\dot{\omega}}{\alpha K}\right), \qquad (\omega_e)_{\text{eq}} = 0. \tag{50}$$

If the phase-lock loop "locks-on" to the frequency ramp input, the frequency error is zero, but there is a steady-state phase error given by equilibrium value in (50). A *necessary* condition for the existence of this phase-locked response is that

$$\alpha K > \dot{\omega}$$

i.e., the total loop gain must exceed the input frequency rate.

To analyze further the response of this circuit, it is convenient to normalize (49) in time and frequency with respect to the parameter ω_n. Defining the symbols

$$\begin{aligned} x &\equiv \omega_n t \\ \nu &= \omega_e/\omega_n \\ r &= \dot{\omega}/{\omega_n}^2 = \dot{\omega}/\alpha K \end{aligned}$$

the differential equation (49) can be written.

$$\begin{aligned} \frac{d\theta}{dx} &= \nu \\ \frac{d\nu}{dx} &= r - 2\zeta\nu\cos\theta - \sin\theta \end{aligned} \tag{51}$$

where $\theta \equiv \theta_e(t)$ is assumed in the remainder of Appendix B.

The solutions of these equations for given initial conditions and given value of normalized input rate, r, describe trajectories in the normalized (θ, ν) state space. The slope of these trajectories in this state space is, from (51)

$$\frac{d\nu}{d\theta} = \frac{r - \sin \theta}{\nu} - 2\zeta \cos \theta. \tag{52}$$

For initial conditions $\theta = 0$, $\nu = 0$ (corresponding to the circuit being in steady-state phase lock with constant frequency input prior to the onset of the frequency-ramp input), two sets of trajectories obtained from numerical integration of (52) are shown in Fig. 12. In Fig. 12(a) a value of $r = 0.966$ is chosen to illustrate the response when the loop gain exceeds the frequency rate only slightly. It can be seen that for $\zeta \geqq 1$, the phase error tends to the steady-state value of 1.31 radians ($= \sin^{-1} 0.966$) with only small overshoot. For $\zeta = 0.707$, however, there is a large overshoot which actually exceeds $\pi/2$ radians, but eventually returns to the steady-state value; for $\zeta < 0.707$ the phase error does not reach the steady-state: i.e., the circuit is unable to "lock-on."

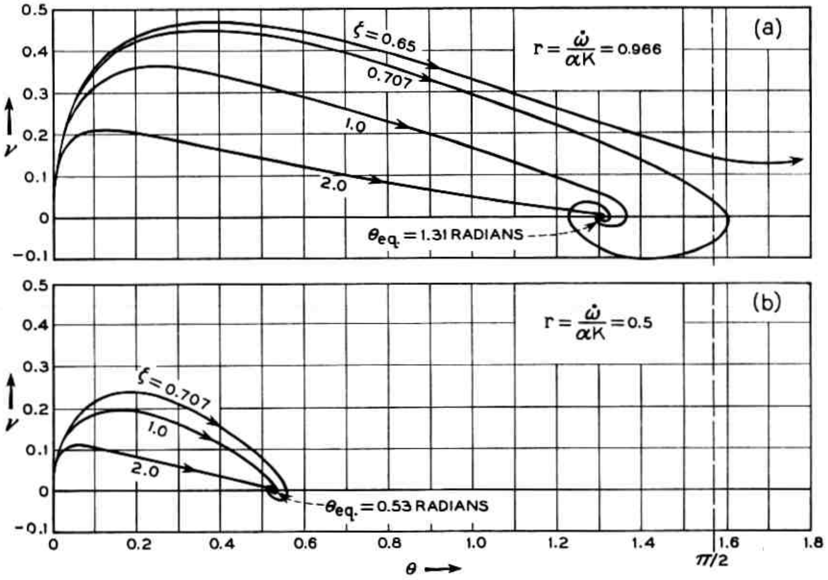


Fig. 12 — Trajectories of the phase-lock loop response to frequency ramp input for various values of the damping ratio: (a) input frequency rate, $\dot{\omega}$, nearly equal to the loop gain, αK; (b) input frequency rate, $\dot{\omega}$, equal to one-half the loop gain, αK.

To avoid large steady-state phase errors and large peak phase errors with the attendant likelihood of random perturbations causing the circuit to fall out of phase lock, the following conditions should be imposed

$$\left.\begin{aligned}\alpha K &\geqq 2\dot{\omega}\\ \zeta &> 0.7\end{aligned}\right\}. \tag{53}$$

Fig. 12(b) shows a set of trajectories with conditions (53) satisfied. The response closely approximates that of the linear second-order system obtained by letting $\sin \theta = \theta$, $\cos \theta = 1$ in (51).

APPENDIX C

Satellite Orbit Characteristics

The parameters of the satellite orbit which affect the design of the phase-lock loop are evaluated in this appendix. The effect of the oblateness of the earth and other perturbations upon the satellite orbit is neglected in this analysis. However, since this effect does cause rotation of the perigee, the maximum range, minimum range and maximum Doppler effects are derived considering all possible locations of the perigee relative to the ground station.

The geometry and terminology of the analysis is shown in Fig. 13(a). It is sufficient to consider only the condition when the ground station is in the plane of the orbit in order to derive all the parameters needed.

(*i*) *Minimum and Maximum Communicating Range.* It is obvious that the minimum possible range occurs when the satellite passes overhead at perigee. Therefore

$$R_{\min} = R_p = \text{perigee altitude}. \tag{54}$$

From the point of view of the satellite, the maximum possible range to any visible point on earth occurs when the satellite is at apogee and the range is taken along the tangent to the earth's surface. The satellite would then appear on the horizon at maximum range to a tracking station anywhere along the locus of these points of tangency.

Since, however, the satellite must be at a small angle, φ_h, above the horizon before communication is feasible at maximum range, the conditions at the maximum possible communication range are as shown in Fig. 13(b). In terms of the angle φ_c in Fig. 13(b)

$$R_{\max}{}^2 = r_a{}^2 + r_0{}^2 - 2r_a r_0 \cos \varphi_c$$

where

$$r_0 = \text{radius of earth}$$

$$r_a = R_a + r_0\,, \qquad R_a = \text{apogee altitude.}$$

Now for φ_h small (less than about 10°), the angle φ_c is very closely given

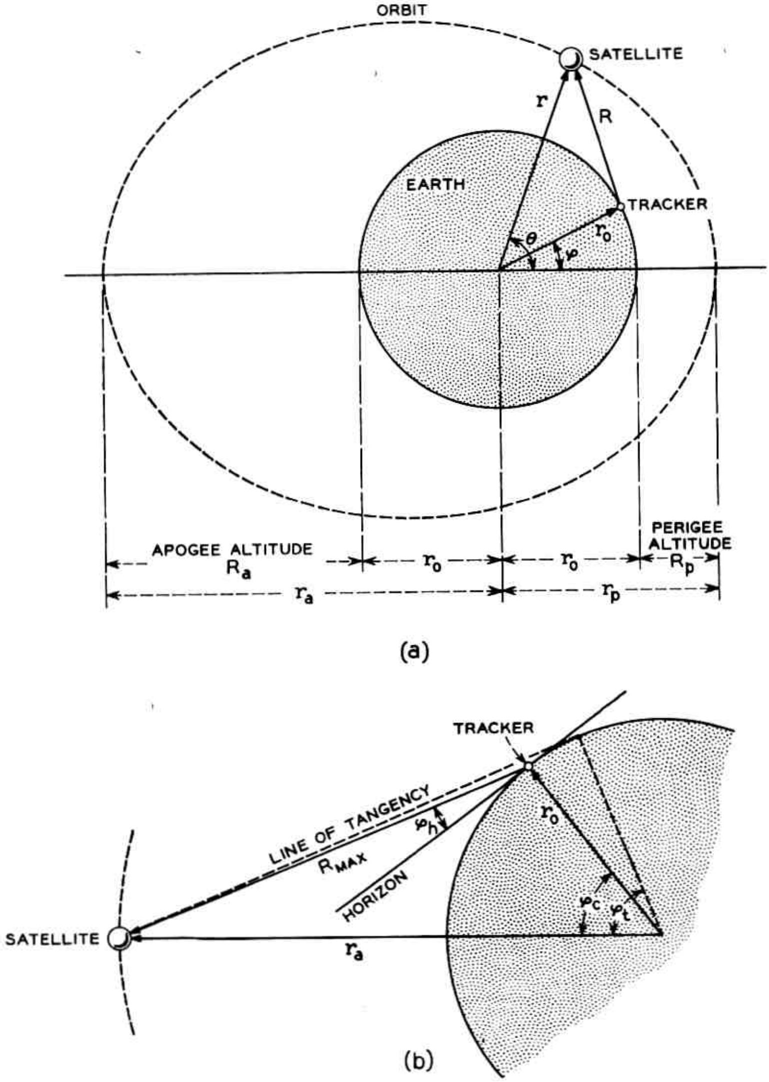


Fig. 13 — Satellite orbit diagrams: (a) geometry and terminology for satellite orbit; (b) conditions for maximum communication range.

by

$$\varphi_c \doteq \varphi_t - \varphi_h , \qquad \varphi_t = \cos^{-1}\left(\frac{r_0}{r_a}\right).$$

Therefore, in terms of known parameters and a given horizon angle φ_h, the maximum possible communication range is

$$R_{\max} \doteq [r_a^2 + r_0^2 - 2r_a r_0 \cos(\varphi_t - \varphi_h)]^{\frac{1}{2}} \tag{55}$$

where

$$\varphi_t = \cos^{-1}(r_0/r_a).$$

(ii) Maximum Doppler Shift and Rate. The orbit parameters needed to determine these Doppler effects are the maximum range rate, $\dot{R}_{\max}$, and the maximum range acceleration, $\ddot{R}_{\max}$. The parametric equation of the satellite orbit in polar form corresponding to the choice of coordinates in Fig. 13(a) is

$$r = \frac{r_m}{1 + \epsilon \cos \theta} \tag{56}$$

where

$$r_m = \frac{2r_a r_p}{r_a + r_p}$$

$$\epsilon = \frac{r_a - r_p}{r_a + r_p} = \text{eccentricity of orbit}$$

$$r_a = R_a + r_0$$

$$r_p = R_p + r_0.$$

Furthermore, from the "law of areas" for motion in a central force field

$$\dot{\theta} = \frac{k}{r^2}\ \text{rad/sec} \tag{57}$$

where

$$k^2 = GMr_m = gr_0^2 r_m$$

G = universal gravitational constant

M = mass of earth

g = acceleration due to gravity at surface of earth.

From Fig. 13(a), the range, R, for a tracker at angle φ is related to the orbit variables, r and θ, by

$$R^2 = r^2 + r_0^2 - 2r_0 r \cos(\theta - \varphi). \tag{58}$$

Differentiating this expression and using (57) gives for the range rate $\dot{R}$

$$\dot{R} = \frac{\dot{r}}{R}[r - r_0 \cos(\theta - \varphi] + \frac{kr_0}{Rr}\sin(\theta - \varphi) \tag{59}$$

and for the range acceleration, $\ddot{R}$

$$\ddot{R} = \frac{\ddot{r}}{R}[r - r_0 \cos(\theta - \varphi)] + \frac{\dot{r}^2 - \dot{R}^2}{R} + \frac{k^2 r_0}{Rr^3}\cos(\theta - \varphi). \tag{60}$$

These expressions depend upon r, $\dot{r}$, and $\ddot{r}$, which are determined as a function of θ by the orbit equations (56) and (57). The evaluation of $\dot{R}_{\max}$ is rather tedious and is most easily obtained for a given orbit by machine or graphical computation. It was evaluated for the expected Telstar satellite orbit using a part graphical and part analytical computation, with the results given at the end of this appendix.

The evaluation of $\ddot{R}_{\max}$ is quite easily obtained, however, since it occurs for the conditions $\theta = \varphi = 0$; i.e., when the satellite passes overhead at perigee. The maximum range acceleration is given by

$$\ddot{R}_{\max} = g\,\frac{r_0^2}{(R_p + r_0)^2}\left(\frac{r_m}{R_p} - 1\right) \tag{61}$$

which occurs when $R = R_{\min} = R_p$.

The maximum Doppler rate varies directly with the maximum range acceleration

$$\dot{\omega}_{\max} = \frac{2\pi f_b}{c}\ddot{R}_{\max} = 2\pi f_b \frac{g}{c}\left[\frac{r_0^2}{(R_p + r_0)^2}\left(\frac{r_m}{R_p} - 1\right)\right] \tag{62}$$

where

f_b = satellite beacon frequency

c = velocity of light.

To estimate the maximum Doppler rate which might be expected for practical communication satellite systems, we take the following conditions as representing practical extremes from the point of view of good communication and satellite lifetime

maximum $f_b = 5 \times 10^9$ cyc/sec

minimum perigee, $R_p = 0.2$ knm

maximum apogee, $R_a = 5.0$ knm.

Using these values in (62) gives as an estimate for the maximum expected Doppler rate

$$\max\,(\dot{\omega}_{\max}) \approx 2 \times 10^{4}\ \text{rad/sec}^{2}.$$

(iii) Numerical Values for the Expected Telstar Satellite. The constants needed for the range and Doppler calculations are:

$$\begin{aligned} R_p &= 0.5\ \text{knm (perigee)} \\ R_a &= 3.0\ \text{knm (apogee)} \\ r_0 &= 3.44\ \text{knm} \\ r_m &= 4.88\ \text{knm} \\ g/c &= 3.27 \times 10^{-8}\ \text{sec}^{-1}; \qquad c = 162\ \text{knm/sec} \\ f_b &= 4.08 \times 10^{9}\ \text{cyc/sec} \\ \varphi_h &= 7.5^\circ\ \text{(acquisition angle above horizon)}. \end{aligned}$$

Using these numerical constants in (54), (55), and (62) gives for $R_{\min}$, $R_{\max}$, and $\dot{\omega}_{\max}$ the values:

$$\begin{aligned} R_{\min} &= 0.5\ \text{knm} \\ R_{\max} &= 5.0\ \text{knm} \\ \dot{\omega}_{\max} &= 5.62 \times 10^{3}\ \text{rad/sec}^{2}. \end{aligned}$$

The maximum value of $\dot{R}$ in (59) for the Telstar satellite orbit was found to occur when $\theta = 340^\circ$, $\varphi = 10^\circ$ and has a magnitude

$$|\dot{R}|_{\max} \doteq 4 \times 10^{-3}\ \text{knm/sec}.$$

The maximum Doppler shift is then given by

$$|\Delta f|_{\max} = \frac{f_b}{c}|\dot{R}|_{\max} \doteq 100\ \text{kc}.$$

REFERENCES

1. Githens, J. A., Kelly, H. P., Lozier, J. C., and Lundstrom, A. A., Antenna Pointing System: Organization and Performance, B.S.T.J., **42,** July, 1963, p. 1213.
2. Anders, J. V., Higgins, E. F., Murray, J. L., and Schaefer, F. J., The Precision Tracker, B.S.T.J., **42,** July, 1963, p. 1309.
3. Cook, J. S., and Lowell, R., The Autotrack System, B.S.T.J., **42,** July, 1963, p. 1283.
4. Develet, J. A., Jr., Fundamental Sensitivity Limitations for Second-Order Phase-Lock Loops, STL Report 8616-0002-NU-000, June 1, 1961.

5. Develet, J. A., Jr., Thermal Noise Errors in Simultaneous Lobing and Conical Scan Angle-Tracking Systems, I.R.E. Trans. on Space Electronics and Telemetry, **SET 7,** June, 1961, pp. 42-51.
6. Jaffe, R., and Rechtin, E., Design and Performance of Phase-Lock Circuits Capable of Near-Optimum Performance Over a Wide Range of Input Signal and Noise Level, I.R.E. Trans. Inf. Theory, **IT 1,** March, 1955, pp. 66–76.
7. James, H. M., Nichols, N. B., and Phillips, R. S., *Theory of Servomechanisms*, Rad. Lab. Series, **25,** McGraw-Hill, New York, 1947, pp. 369–370.
8. Enloe, L. H., unpublished work.
9. Ball, W. H. W., Analysis and Digital Simulation of the *Telstar* Precision Tracker, Paper No. CP-63-368, presented at the IEEE Winter General Meeting, New York, 1963.
10. D'Azzo, J. J., and Houpis, C. H., *Control System Analysis and Synthesis*, McGraw-Hill, New York, 1960, pp. 58–62.
11. Viterbi, A. J., System Design Criteria for Space Television, J. Brit. I.R.E., **19,** Sept., 1959, pp. 561–570.
12. Brockman, M. H., Buchanan, H. R., Choate, R. L., and Malling, L. R., Extra-Terrestrial Radio Tracking and Communication, Proc. I.R.E., **48,** No. 4, April, 1960, pp. 643–654.
13. Davenport, W. B., Jr., Signal-to-Noise Ratios in Band-Pass Limiters, J. Appl. Phys., **24,** June, 1953, pp. 720–727.

Estimates of Error Rates for Codes on Burst-Noise Channels

By E. O. ELLIOTT

(Manuscript received April 8, 1963)

The error structure on communication channels used for data transmission may be so complex as to preclude the feasibility of accurately predicting the performance of given codes when employed on these channels. Use of an approximate error rate as an estimate of performance allows the complex statistics of errors to be reduced to a manageable table of parameters and used in an economical evaluation of large collections of error detecting codes. Exemplary evaluations of error detecting codes on the switched telephone network are included in this paper.

On channels which may be represented by Gilbert's model of a burst-noise channel, the probabilities of error or of retransmission may be calculated without approximations for both error correcting and error detecting codes.

I. INTRODUCTION

The structure in bursts of noise on real communication channels is usually very difficult to describe. As a consequence, no general procedure exists for predicting the performance of error detecting or error correcting codes, and no basic set of parameters exists for describing the channel. Gilbert[1] has shown that a simple Markov model with three parameters provides a close approximation to certain telephone circuits used for the transmission of binary data. When such an approximation is possible, the error rates for codes may be easily calculated from these channel parameters and properties of the code. (See Section V.)

To provide a means for estimating error rates for binary block codes in more general circumstances, a table of probabilities $P(m,n)$ may be employed. $P(m,n)$ is the probability that m bit errors occur in a transmitted block of n bits. It was speculated and later corroborated (as we will show) that equivalent error detecting codes would have rather comparable error rates when employed on the same channel. (Two codes are equivalent if one may be obtained from the other by a permutation of bit positions.) Thus the average error rate for all codes equivalent to a given code may be used as an estimate of the true error rate. This average probability of an undetected error in a single transmission of a word is given by

$$\bar{P}_u = \sum_{m=1}^{n} \frac{w(m)}{\binom{n}{m}} P(m, n)$$

where code word usage is assumed uniform, $w(m)$ is the average number of code words at distance m from a typical code word, and n is the block length of the code.

No definitive statement regarding the accuracy of this estimate can be made at this point. A limited investigation, however, suggests that it will ordinarily be a reasonable estimate.

As an example of the use of this method, a collection of 29 interesting error detecting codes is evaluated, using the recorded error data of the field testing program conducted by the data transmission evaluation task force of the Bell System.[2] The Bose-Chaudhuri (31, 21) code is included in this collection and is analyzed in considerable detail to illustrate the full potentials and limitations inherent in the method.

In the interest of simplicity, the discussion to follow will be limited to binary block codes with particular interest in error detection. The methods employed, however, are not limited to these particular applications, and are open to obvious generalizations.

II. PRELIMINARY DEFINITIONS AND OBSERVATIONS

A binary block code C, hereinafter referred to as a "code," is a collection of binary words of 0's and 1's of length n. N will be used to denote the total number of words in C. The distance $\delta(x,y)$ between two binary words x and y of length n is the number of bit positions in which x and y differ. The weight $|x|$ of x is the distance $\delta(\theta,x)$ between x and the all-zero word θ. The number of ordered pairs of code words x, y such that $\delta(x,y) = m$ is denoted by $W(m)$, and $w(m) = W(m)/N$.

The communication channel is described by a collection of conditional probabilities of the form $P(x \rightarrow y)$, which give the probability that the word y will be received when x is transmitted. A channel is called *metric* whenever $P(x \rightarrow y)$ is a function only of $\delta(x,y)$: i.e., $P(x \rightarrow y) = F(m,n)$, where $m = \delta(x,y)$ and n is the block length. A channel is called *symmetric* whenever $P(x \rightarrow y)$ is a function only of $z = y - x \pmod 2$.

It should be noted that a metric channel is symmetric and that a symmetric memoryless channel is metric. The Gilbert burst-noise channel[1] is an example of a symmetric channel which, because of its memory (i.e., interdependence of error probabilities of neighboring bits), is not metric.

When a code is used for error detection, it will be assumed that error correction is accomplished by retransmissions of any received words which are detected to be in error. The specific manner in which the receiver signals to the transmitter for a retransmission will not be considered. It will be assumed, however, that this backward signaling is error-free, that each retransmission consists of a single word, and that repeated retransmissions of a word are possible. Since very little information is required for the backward signaling for retransmissions, it is not too unrealistic to assume that it is error-free. Most retransmission systems will, however, probably involve delays in retransmissions, and the retransmitted data may consist of a block of several words. Because of the burst nature of noise on many channels, the effect of these retransmission delays is improvement of the channel, and we can then expect codes to perform better than our model indicates.

Thus, for an error detecting code, an (undetected) error occurs if a received word is a code word different from the transmitted word. If x is the transmitted word, then the probabilities of an undetected error, of a word retransmission, and of acceptance of a correct word are, respectively

$$\textstyle\sum_{y(\neq x)\varepsilon C} P(x \rightarrow y) \qquad \sum_{y \notin C} P(x \rightarrow y) \qquad \text{and} \qquad P(x \rightarrow x).$$

Now, if we assume that the words of the code are used with equal frequencies, then the averages of the above probabilities are, respectively

$$P_u = \frac{1}{N} \textstyle\sum_{x \varepsilon C} \sum_{y(\neq x)\varepsilon C} P(x \rightarrow y) \tag{1}$$

$$P_r = \frac{1}{N} \textstyle\sum_{x \varepsilon C} \sum_{y \notin C} P(x \rightarrow y) \tag{2}$$

and

$$P_0 = \frac{1}{N} \textstyle\sum_{x \varepsilon C} P(x \rightarrow x). \tag{3}$$

These probabilities are of some interest in themselves, but for symmetric communication channels the probability P_E that a word is received in error after possible retransmissions is given by

$$P_E = \frac{P_u}{1 - P_r}. \tag{4}$$

This result follows from the definitional equation

$$P_E = \text{Prob (undetected error} \mid \text{received word is accepted)}$$

the definition of conditional probability, and the observations that an undetected error implies acceptance of the received word and that the probability of a received word being accepted is $1 - P_r$.

Suppose the channel is metric, so that $P(x \to y) = F(m,n)$ where $m = \delta(x,y)$ and n is the length of x and y. Then, from (1), (3) and (2)

$$P_u = \sum_{m=1}^{n} w(m)F(m,n), \tag{5}$$

$$P_0 = F(0,n),$$

and

$$P_r = 1 - (P_0 + P_u). \tag{6}$$

It is evident from (5) that on a metric channel equivalent codes have identical values of P_u, since w is invariant under a permutation of the bit positions in a code.

III. $\bar{P}_u$ ON SYMMETRIC CHANNELS

Let $\bar{P}_u$ denote the average value of P_u over all bit-position permutations of the code. If $P(m,n)$ is the total probability of m errors in a block of length n, i.e.

$$P(m,n) = \textstyle\sum_{|y|=m} P(\theta \to y) \tag{7}$$

then

$$\bar{P}_u = \sum_{m=1}^{n} w(m) \frac{P(m,\, n)}{\binom{n}{m}}. \tag{8}$$

This result may be seen as follows: consider a particular code C and channel X. Corresponding to each permutation π of the n bit positions is a permutation of C which we will call πC. Now, using (1)

$$\begin{aligned} \bar{P}_u &= \frac{1}{n!} \textstyle\sum_{\pi} \frac{1}{N} \sum_{x \varepsilon \pi C} \sum_{y(\neq x) \varepsilon \pi C} P(x \to y) \\ &= \frac{1}{N} \textstyle\sum_{x \varepsilon C} \sum_{y(\neq x) \varepsilon C} \frac{1}{n!} \sum_{\pi} P(\pi x \to \pi y). \end{aligned} \tag{9}$$

For a symmetric channel there is a function f such that $P(x \to y) =$

$f(z)$ where $z = y - x \pmod 2$. Then, if z contains m ones

$$\frac{1}{n!} \sum_\pi P(\pi x \to \pi y) = \frac{1}{n!} \sum_\pi f(\pi z).$$

Now any n-place binary sequence having exactly m ones is left invariant by $m!(n\text{-}m)!$ permutations of its digits. The sum just written is therefore equal to

$$\frac{m!(n-m)!}{n!} \sum f(u) = \frac{P(m, n)}{\binom{n}{m}}$$

where the sum is over all distinct n-place binary sequences u having exactly m ones. Equation (8) follows by inserting this result in (9).

Now that $\bar{P}_u$ has been obtained, it is an easy matter to obtain $\bar{P}_r$, the average probability of a retransmission for all permutations of the code. Since $P_0 = P(0,n)$ and $P_0 + P_r + P_u = 1$, it follows that

$$\bar{P}_r = 1 - \bar{P}_u - P(0,n).$$

Our $\bar{P}_u$ estimate is exactly equal to P_u whenever the code in question is invariant under all permutations of bit positions. Thus, accurate results are obtained on symmetric channels for single parity check codes, constant weight codes, etc.

It is of interest to note that in the case of group codes of given block length and redundancy, $w(m)$ has an unevenly weighted average value which may be used to estimate $\bar{P}_u$ in terms of the code's minimum distance D. Consider group codes of block length n and dimension k. For such group codes there are 2^{kc} ways of assigning the k information positions to the check positions of the $c = n - k$ check bits, but for such assignments the resulting codes are not necessarily distinct. Of these, however, it is known (Ref. 3, p. 54) that in $2^{(k-1)c}$ cases a given binary word z will belong to the resulting code, provided the information portion of z does not contain only 0's. Now, there are $\left[\binom{n}{m} - \binom{c}{m}\right]$ binary words of weight m having nonzero information parts whenever $0 < m \leqq c$, and there are $\binom{n}{m}$ such words whenever $c < m \leqq n$. As a consequence, the "average" number $\bar{w}(m)$ of code words of weight m is

$$\bar{w}(m) = \frac{1}{2^{kc}}\left[\binom{n}{m} - \binom{c}{m}\right] 2^{(k-1)c} \qquad \text{when } 0 < m \leqq c$$

and

$$\bar{w}(m) = \frac{1}{2^{kc}}\binom{n}{m}2^{(k-1)c} \qquad \text{when } c < m \leqq n$$

wherein the average is over the multiplicity of group codes of the specified block length and dimension which result from these assignments of information bit positions to check bit positions.

Let

$$\bar{C}(m) = \bar{w}(m) \Big/ \binom{n}{m}$$

then

$$\bar{C}(m) = 2^{-c}\left\{1 - \frac{\binom{c}{m}}{\binom{n}{m}}\right\} \qquad \text{when } 0 < m \leqq c$$

and

$$\bar{C}(m) = 2^{-c} \qquad \text{when } c < m \leqq n.$$

This result may be of use as follows. Suppose we have knowledge only of the minimum distance D of a given group code that we wish to evaluate on some channel. Let us make the bold assumption that the big difference between the given code and the "average" of all codes is the fact that the given code contains no words of weight $1, \cdots, D - 1$. Then (8) yields

$$\bar{P}_u \approx \sum_{m=D}^{n} \bar{C}(m)P(m,n). \tag{10}$$

When the dimension of a code is large, it may be unfeasible to ascertain $w(m)$ because of the immense amount of computation required. It is in such cases that (10) may prove to be a useful approximation.

IV. $\bar{P}_u$ ON ASYMMETRIC CHANNELS

We propose the following reasonably general model of an asymmetric channel. Two channel states are hypothecated: a "good" state in which no errors occur, and a "bad" state in which $0 \rightarrow 1$ errors occur with probability p_0 and $1 \rightarrow 0$ errors occur with probability p_1. The manner in which good and bad states occur will not be specified beyond knowledge of the total probability $S(s,n)$ of being in the bad state for some s bits of the n bits of a block. Particular arrangements of these s bad bits need not be equiprobable.

Let $q_0 = 1 - p_0$ and $q_1 = 1 - p_1$, and make the following definitions when x and y are binary words of length n:

$A(x,y)$ = the number of bit positions where x is 0 and y is 1,
$A'(x,y)$ = the number of bit positions where both x and y are 0,
$B(x,y)$ = the number of bit positions where x is 1 and y is 0,
$B'(x,y)$ = the number of bit positions where both x and y are 1.

Let the state sequence of the channel be described by a binary word v, in which each digit is G or B according as the state of the channel at that digit's position is good or bad. Now define

$A^*(x,y,v)$ = the number of bit positions in which x and y are 0 and v is B, and

$B^*(x,y,v)$ = the number of bit positions in which x and y are 1 and v is B.

The error probabilities for this channel, conditional on the state sequence v, may now be given as follows:

$$P(x \to y \mid v) = \left| \begin{array}{ll} 0 \text{ if at some bit position } v \text{ is } G \text{ and} & \\ \quad x \text{ and } y \text{ are different} & (11) \\ p_0{}^a q_0{}^{a^*} p_1{}^b q_1{}^{b^*} \text{ otherwise} & (12) \end{array} \right.$$

where the values of the previously defined functions are

$$a = A(x,y) \qquad a' = A'(x,y)$$
$$B = B(x,y) \qquad b' = B'(x,y)$$

and

$$a^* = A^*(x,y,v),$$
$$b^* = B^*(x,y,v).$$

Define

$$\bar{P}(x \to y \mid v) = \frac{1}{n!} \sum\nolimits_\pi P(\pi x \to \pi y \mid v) \tag{13}$$

wherein π is the arbitrary permutation of bit positions that we have used before. Notice that by (11), (12) and (13)

$$\bar{P}(x \to y \mid v) = \bar{P}(x \to y \mid \pi v) \tag{14}$$

and therefore that $\bar{P}(x \to y \mid v)$ depends only on how many B's are in v and not on their positions in v. Suppose v contains s B's. We can now say, using (13) and writing $\bar{P}_s(x \to y)$ for $\bar{P}(x \to y \mid v)$, that

$$\bar{P}_s(x \to y) = \frac{1}{n!} \sum\nolimits_\pi P(x \to y \mid \pi v) \tag{15}$$

and from (11) we know $P(x \to y \mid \pi v) = 0$ whenever πv has G's in positions where x and y differ.

We will use (12) in evaluating (15) by first finding the number of permutations π for which $P(x \to y \mid \pi v)$ has a fixed value. Suppose a^* and b^* are such numbers that $a^* + b^* = s - (a + b)$, with $0 \leqq a^* \leqq a'$ and $0 \leqq b^* \leqq b'$. Then there are $\binom{a'}{a' - a^*} = \binom{a'}{a^*}$ arrangements of $a' - a^*$ G's among the a' bit positions where x and y are 0, and there are $\binom{b'}{b' - b^*} = \binom{b'}{b^*}$ arrangements of $b' - b^*$ G's among the b' bit positions where x and y are 1. Hence there are a total of $\binom{a'}{a^*}\binom{b'}{b^*}$ arrangements of the $n - s$ G's among the $a' + b'$ bit positions where x and y are the same. For each such arrangement there are $s!(n - s)!$ permutations π under which the arrangement is invariant. Consequently, the total number of permutations π for which $P(x \to y \mid \pi v) = p_0^a q_0^{a^*} p_1^b q_1^{b^*}$ is given by $s!(n - s)! \binom{a'}{a^*}\binom{b'}{b^*}$. Hence, they contribute

$$\frac{n!\binom{a'}{a^*}\binom{b'}{b^*}}{\binom{n}{s}} p_0^a q_0^{a^*} p_1^b q_1^{b^*}$$

to the sum in (15). We conclude then that

$$\bar{P}_s(x \to y) = \sum_{a^*=\max(0,\, t-b')}^{\min(a',\, t)} \frac{\binom{a'}{a^*}\binom{b'}{t - a^*}}{\binom{n}{s}} p_0^a p_1^b q_0^{a^*} q_1^{t-a^*} \qquad (16)$$

where $t = s - (a + b)$.

If we set $r = \mid x \mid$, then $\bar{P}_s(x \to y)$ may be expressed in terms of b, a, r and s as

$$H(b, a, r, s)$$
$$= \left(\frac{p_0}{q_1}\right)^a \left(\frac{p_1}{q_1}\right)^b q_1^s \sum_{a^*=\max(0,\, s-(a+r))}^{\min(n-(a+r),\, s-(a+b))} \left(\frac{q_0}{q_1}\right)^{a^*} \frac{\binom{n - (a + r)}{a^*}\binom{r - b}{s - (a + b + a^*)}}{\binom{n}{s}}$$

which is just another form of (16). This asymmetric channel may now be compactly described by a function J which gives the probability, averaged over all permutations π of the bit positions, of making b $1 \rightarrow 0$ errors and a $0 \rightarrow 1$ errors in a transmitted word of weight r

$$J(b,a,r) = \sum_{s=a+b}^{n} H(b,a,r,s)S(s,n).$$

For a code C, let us define $I_C(b,a,r)$ to be $1/N$ times the number of ordered code-word pairs (x,y) for which $A(x,y) = a$, $B(x,y) = b$, and $|x| = r$. Then finally the asymmetric analogue of (8) for the average probability $\bar{P}_u$ of an undetected error may be written as

$$\bar{P}_u = \sum_{(b,\,a,\,r):b \leqq r \leqq n, a \leqq n-r} I_C(b, a, r)\, J(b, a, r). \tag{17}$$

V. ERROR PROBABILITIES ON GILBERT BURST-NOISE CHANNELS

Gilbert's model[1] of a burst-noise channel is a binary symmetric channel (with memory) determined by an elementary Markov chain. As in the preceding model for an asymmetric channel, a good (G) and bad (B) state are assumed of the channel. No errors occur in the G state, but in the B state, the probability of a bit error is $(1 - h)$. With the transmission of each bit, the channel has opportunity to change states. The transitions $G \rightarrow B$ and $B \rightarrow G$ have probabilities P and p, respectively, while the transitions $G \rightarrow G$ and $B \rightarrow B$ have probabilities $Q = 1 - P$ and $q = 1 - p$. When Q and q are large, the states G and B tend to persist, simulating features of a burst-noise channel. Gilbert (Ref. 1, p. 1262) has shown how this model approximates the burst noise on two of the calls from the field testing program of the data transmission evaluation task force of the Bell System.

Using conditional probabilities determined by the parameters P,p,h, it is a simple matter to calculate the probability that a transmitted word x be received as y on a Gilbert channel. This probability depends on the modulo 2 difference $z = y - x$ of y and x.

Suppose a is the number of 0's in z which precede the first 1 in z, c is the number of 0's following the last 1, and b_i $(i = 1, \cdots, |z| - 1)$ are the number of 0's between consecutive 1's in z. Then, if $z \neq \theta$

$$P(x \rightarrow y) = P(z) = w(a)\{\textstyle\prod_{i=1}^{|z|-1} v(b_i)\}u(c) \tag{18}$$

where w, v and u are functions such that $w(k) = P(0^k1)$, $v(k) = P(0^k1 \mid 1)$, and $u(k) = P(0^k \mid 1)$ $(k = 0, 1, \cdots)$. Here 0^k denotes k consecutive zeros. Also, if $z = \theta$ then

$$P(x \rightarrow x) = P(\theta) = 1 - \sum_{i=0}^{n-1} w(i). \tag{19}$$

Using generating functions, Gilbert has shown that u, v and w satisfy the following recurrence equations

$$\begin{aligned} u(0) &= 1,\ u(1) = p + hq; \\ u(k) &= (Q + hq)u(k-1) - h(Q-p)u(k-2), \qquad k = 2, 3, \cdots \qquad (20) \\ v(k) &= u(k) - u(k+1), \qquad k = 0, 1, \cdots \\ w(k) &= p_B(1-h)u(k). \end{aligned}$$

Equation (18) results from the obvious composition of the conditional probabilities in the v and u terms. Equation (19) results from the fact that the event not 0^n is the union of the events 1, 01, $0^2 1$, $\cdots$, $0^{n-1}1$. Since these events are disjoint,

$$P(0^n) = 1 - P(\text{not } 0^n) = 1 - \sum_{i=0}^{n-1} P(0^i 1).$$

In the interest of completeness, we shall sketch a proof that u, v and w satisfy the recurrence equation (20).

To see that

$$v(k-1) = u(k-1) - u(k), \qquad k = 1, 2, \cdots$$

note that the event 10^{k-1} is the union of $10^{k-1}1$ and 10^k and that the latter two events are disjoint. Hence

$$\text{Prob } (0^{k-1} \mid 1) = \text{Prob } (0^{k-1}1 \mid 1) + \text{Prob } (0^k \mid 1)$$

and therefore

$$u(k-1) = v(k-1) + u(k).$$

We define $u(0) = 1$. That $u(1) = p + qh$ is obvious. To establish that $u(k+1) = (Q + hq)u(k) - h(Q-p)u(k-1), \quad k = 1, 2, \cdots$ we shall need to introduce

$$u_G(k) = \text{Prob } (0^{k-1}G \mid 1) \qquad \text{and} \qquad u_B(k) = \text{Prob } (0^{k-1}0_B \mid 1)$$

wherein 0_B denotes a zero in the bad state. Clearly

$$u(k) = u_G(k) + u_B(k)$$

and

$$u_B(k) = \frac{h}{1-h} v(k-1).$$

Now, considering transitions, we see that

$$\begin{aligned}u(k+1) &= (Q+Ph)u_G(k) + (p+qh)u_B(k)\\ &= (Q+Ph)\{u(k) - u_B(k)\} + (p+qh)u_B(k)\\ &= (Q+Ph)u(k) - (Q+Ph-p-qh)u_B(k)\\ &= (Q+Ph)u(k) - (Q-p)(1-h)u_B(k)\\ &= (Q+Ph)u(k) - (Q-p)h\{u(k-1) - u(k)\}.\end{aligned}$$

Finally, it is evident that if z' is obtained from z by inverting the order of the bits, then $P(z') = P(z)$. This results from the fact that the forward and backward state transition probabilities are identical. As a consequence,

$$\begin{aligned}w(k) = \text{Prob } (0^k1) = \text{Prob } (10^k)&\\ = p_B(1-h) \text{ Prob } (0^k \mid 1) = p_B(1-h)u(k)&\end{aligned}$$

and the proof is complete.

The performance of error detecting codes on Gilbert channels can now be calculated using (18)–(20) in (1)–(4). For an error correcting group code using coset decoding,[4] the probability of incorrect decoding is given by

$$P_e = 1 - \sum_{i=1}^{n} P(\alpha_i)$$

where the α_i are the coset leaders for the code. These coset leaders would presumably be chosen so as to minimize P_e and therefore may not necessarily be the minimal weight elements of cosets.

It is interesting to note that if a Gilbert channel with parameters (P,p,h) is sampled at every kth bit, then the string of bits obtained has the same structure as the bits on a Gilbert channel with parameters (P',p',h) where

$$P' = \frac{P}{P+p}\{1 - (Q-p)^k\}$$

and

$$p' = \frac{p}{P+p}\{1 - (Q-p)^k\}.$$

The proof of this assertation is given in Ref. 5, p. 385. This result is useful for analysis when time division multiplex encoding is employed.

VI. $P(m,n)$ FOR GENERALIZED GILBERT CHANNELS

The probabilities $P(m,n)$ for a Gilbert burst-noise channel are readily computed by recursive methods. However, it is just as easy to obtain $P(m,n)$ for a slightly more general symmetric channel. In the Gilbert model, an error bit can occur only when the channel is in the bad state. In the model proposed here, an error bit can occur in either the good or the bad state but with different probabilities. Transitions between the good and bad states are the same as in the Gilbert model.

Let k denote the probability of correct reception of a bit when the channel is in the good state, and let $h' = 1 - h$ and $k' = 1 - k$.

Let $G(m,n)$ = Prob (m errors in a block of length n | the channel is in the good state at the first bit) and $B(m,n)$ = Prob (m errors in a block of length n | the channel is in the bad state at the first bit). Then

$$P(m,n) = \frac{p}{P + p} G(m,n) + \frac{P}{P + p} B(m,n)$$

and $G(m,n)$ and $B(m,n)$ may be found recursively from

$$\begin{aligned} G(m,n) &= G(m,n-1)Qk + B(m,n-1)Pk + G(m-1,n-1)Qk' \\ &\quad + B(m-1,n-1)Pk', \\ B(m,n) &= B(m,n-1)qh + G(m,n-1)ph + B(m-1,n-1)qh' \\ &\quad + G(m-1,n-1)ph', \\ &\quad G(0,1) = k \qquad\qquad B(0,1) = h, \\ &\quad G(1,1) = k' \quad \text{and} \quad B(1,1) = h'. \end{aligned}$$

We must also assign the values $G(m,n) = B(m,n) = 0$ when $m < 0$ or $m > n$.

VII. THE BOSE-CHAUDHURI (31, 21) CODE ON THE TELEPHONE NETWORK

As an illustration of the use of the $\bar{P}_u$ estimate for P_u, the performance of a Bose-Chaudhuri (31, 21) code (Ref. 3, p. 166) on the switched telephone network is analyzed. As a source of error statistics for the channels of the telephone network, the records of the field testing program described by Alexander, Gryb and Nast[2] are employed. These give in sequence the numbers of correct bits and error bits for 1010 calls of 10 and 30 minutes' duration over a variety of facilities in the switched telephone network. A detailed summary of the number of

TABLE I — NUMBER OF CALLS

Type of Call	1200 bps		600 bps 10 min.
	30 min.	10 min.	
Long haul	181	34	229
Short haul	102	20	151
Exchange	108	28	157

calls of each type made at 600-bps and 1200-bps transmission rates with the FM digital subset appears in Table I.*

For each call in this program, the probability $P(m,31)$ that m bit-errors occur in a block of length 31 for $m = 0, 1, \cdots, 31$ has been determined. In doing this, each call is divided into consecutive blocks 31 bits long starting at the ith bit in the call ($i = 1, \cdots, 31$) and the number $N_i(m)$ of blocks containing m bit-errors is noted. This corresponds to viewing each call as, in some sense, 31 different calls, depending on the phase with which we enter the call (i.e., which of the first 31 bits we take as first in governing the subdivision). We thus obtain, for each $i = 1, \cdots, 31$, a probability $P_i(m,31) = N_i(m)/N$ that a block in the subdivision contains m bit-errors. (N is the total number of blocks in the subdivision.) We now average over the possible entry phases and take the probability $P(m,31)$ that m errors occur in a block of length 31 to be $(1/31) \sum_{i=1}^{31} P_i(m,31)$.

Examination of the $P(m,31)$ values obtained reveals some interesting facts. For example, on some calls the probability of having numerous errors in a block greatly exceeds the probability of having only a few errors. For many calls, however, $P(m,31)$ is maximum at $m = 1$, decreases with increasing m, and is often zero for m greater than 2 or 3. On still others, $P(m,31)$ is maximum at $m = 1$, decreases for the next few values of m, and then increases to some smaller relative maximum around $m = 15$ to 17 before its final descent to zero. To illustrate this variability among calls, we present in Table II the $P(m,31)$ values for four calls. In Table II, the P_1 entry under the call's number is the over-all bit-error rate for the call.

Properties of the burst nature of errors on calls like No. 1167 are responsible for $P(m,31)$ having its maximum value midway in the range $m = 1, \cdots, 31$. On such calls there are long bursts of errors. When the burst length is shorter, $P(m,31)$ may more closely resemble that for call No. 1641. These effects can be noted also in Table III,

* Consult Refs. 2 and 6 for a description of call types and for further details regarding the field testing program.

TABLE II — SAMPLE $P(m,31)$ VALUES

Call Type	LH/600/10	SH/1200/30	EX/600/10	LH/1200/30
Call No.	1167	1641	2058	2250
P_1	22.50×10^{-4}	2.70×10^{-4}	0.22×10^{-4}	0.03×10^{-4}
$m = 0$	0.99587	0.99276	0.99972	0.99995
1	0.14×10^{-4}	63.20×10^{-4}	1.83×10^{-4}	0.18×10^{-4}
2	0.22	8.32	0.08	0.16
3	0.14	0.69	0.06	0.12
4	0.20	0.06	0.08	0.0
5	0.14	0.05	0.11	
6	0.25	0.02	0.64	
7	0.14	0.0	0.0	
8	0.14			
9	0.50			
10	0.92			
11	1.00			
12	0.75			
13	1.20			
14	3.18			
15	4.29			
16	4.46			
17	4.88			
18	3.84			
19	4.12			
20	3.51			
21	2.54			
22	2.17			
23	1.78			
24	0.78			
$\geqq 25$	0.0			

which gives the average $P(m,n)$ values for all calls of the field test program.

The quantities $w(m) = 0, 1, \cdots, 31$ for the Bose-Chaudhuri (31, 21) code are presented in the Table IV. Since each check bit of this code applies to an odd number of information bits, $w(m)$ is symmetric: i.e., $w(m) = w(31 - m)$, and therefore $w(m)$ is tabulated only for $m = 0, \cdots, 15$.

Using the above $w(m)$ values and the $P(m,31)$ tables in (8) gives a $\bar{P}_u$ estimate for the undetected error rate on each call.

The smoothed cumulative distributions of the percentage of calls over particular facilities having an estimated undetected error rate not exceeding specified values are shown in Figs. 1 and 2 for the two transmission rates used. We have excluded the 10-minute calls at 1200 bps from this summary of the data because of the small size of the sample.

The approximate retransmission probabilities were generally less than 0.1 per cent. On some 7 per cent of the calls, the rate was between 0.1 and 1 per cent. On only three calls did it exceed one per cent.

It is impossible to obtain exact values of P_u for this code on the tele-

TABLE III — AVERAGE $P(m,n)$ VALUES FOR ALL CALLS OF THE FIELD TEST PROGRAM

m	$n = 8$	$n = 15$	$n = 17$	$n = 21$	$n = 23$	$n = 31$*
1	1.24×10^{-4}	2.07×10^{-4}	2.31×10^{-4}	2.77×10^{-4}	3.00×10^{-4}	3.90×10^{-4}
2	2.28×10^{-5}	3.66×10^{-5}	4.06×10^{-5}	4.82×10^{-5}	5.19×10^{-5}	6.72×10^{-5}
3	9.31×10^{-6}	1.27×10^{-5}	1.40×10^{-5}	1.63×10^{-5}	1.75×10^{-5}	2.18×10^{-5}
4	5.15×10^{-6}	6.98×10^{-6}	7.40×10^{-6}	8.62×10^{-6}	9.20×10^{-6}	1.14×10^{-5}
5	3.62×10^{-6}	4.69×10^{-6}	4.75×10^{-6}	5.50×10^{-6}	5.98×10^{-6}	7.30×10^{-6}
6	2.41×10^{-6}	3.59×10^{-6}	3.54×10^{-6}	3.61×10^{-6}	3.82×10^{-6}	4.81×10^{-6}
7	8.25×10^{-7}	2.66×10^{-6}	2.98×10^{-6}	2.58×10^{-6}	2.78×10^{-6}	3.54×10^{-6}
8	1.03×10^{-7}	2.29×10^{-6}	2.37×10^{-6}	2.13×10^{-6}	2.06×10^{-6}	2.63×10^{-6}
9		1.92×10^{-6}	2.04×10^{-6}	2.23×10^{-6}	1.69×10^{-6}	2.11×10^{-6}
10		1.74×10^{-6}	1.83×10^{-6}	2.17×10^{-6}	1.89×10^{-6}	1.65×10^{-6}
11		1.17×10^{-6}	1.68×10^{-6}	1.86×10^{-6}	2.12×10^{-6}	1.31×10^{-6}
12		4.51×10^{-7}	1.30×10^{-6}	1.81×10^{-6}	1.98×10^{-6}	1.16×10^{-6}
13		8.78×10^{-8}	4.73×10^{-7}	1.72×10^{-6}	1.88×10^{-6}	1.03×10^{-6}
14		1.23×10^{-8}	1.19×10^{-7}	1.02×10^{-6}	1.57×10^{-6}	1.15×10^{-6}
15		9.47×10^{-10}	2.83×10^{-8}	6.03×10^{-7}	1.07×10^{-6}	1.33×10^{-6}
16			2.83×10^{-9}	2.28×10^{-7}	5.13×10^{-7}	2.10×10^{-6}
17				7.67×10^{-8}	2.81×10^{-7}	2.31×10^{-6}
18				1.13×10^{-8}	1.04×10^{-7}	1.31×10^{-6}
19				$2.83 \times 10^{-}$	2.17×10^{-8}	8.25×10^{-7}
20					2.83×10^{-9}	5.13×10^{-7}
21						3.26×10^{-7}
22						2.07×10^{-7}
23						1.32×10^{-7}
24						5.10×10^{-8}
25						4.72×10^{-9}

* $P(m,31) = 0$ for $m = 26, \cdots, 31$.

Table IV—$w(m)$ for the Bose-Chaudhuri (31, 21) Code

m	$w(m)$
0	1
1	0
2	0
3	0
4	0
5	186
6	806
7	2635
8	7905
9	18910
10	41602
11	85560
12	142600
13	195300
14	251100
15	301971

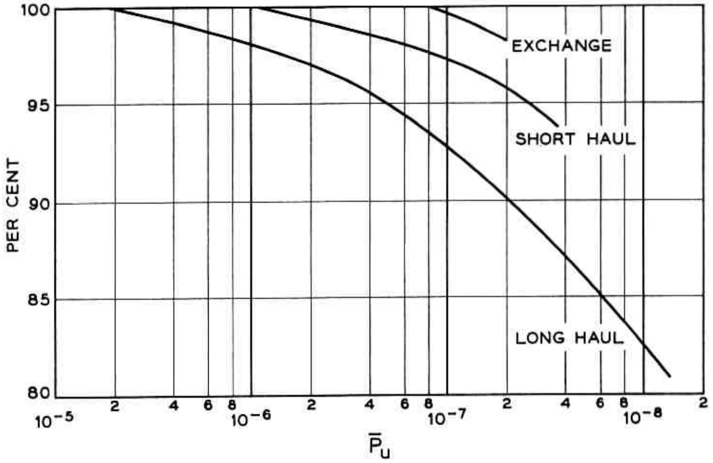


Fig. 1 — Percentage of 10-minute calls at 600 bps with undetected error probabilities not exceeding $\bar{P}_u$.

phone network, since it was not measured during the actual field test program. The records of that program do not allow accurate calculation of it for a variety of reasons.[6] We can, however, think of the recorded bit-error data from the field test program as representing the additive noise of a class of hypothetical channels, and then ask the question, "How well does $\bar{P}_u$ estimate P_u for these hypothetical channels?" To do this, a computer program was written to reconstruct the sequences

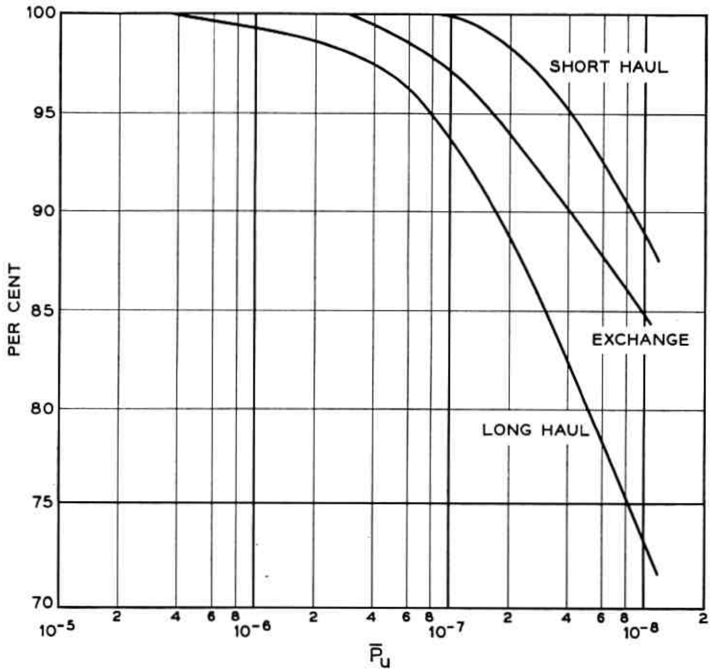


Fig. 2 — Percentage of 30-minute calls at 1200 bps with undetected error probabilities not exceeding $\bar{P}_u$.

of 1's and 0's from the sequential numbers of correct bits and error bits of the task force records. The resulting sequences are then divided into blocks of length 31, and each block is tested to determine if it is the zero word, a code word, or a noncode word. Again each call is treated as 31 calls, according to which of the first 31 bits is chosen first in determining the subdivision into blocks, and the average undetected and detected error rates are calculated.

Of the 1010 test calls in the program, only 10 contained undetected errors. The total number of word-errors was 37 out of a total of 1.06×10^9 words. This corresponds to an over-all undetected word-error rate of 3.5×10^{-8}.

To compare the estimates of $\bar{P}_u$ with the values of P_u obtained from the simulation, we note first that $P_u = 0$ for 1000 calls, whereas $\bar{P}_u$ on these calls varied over a considerable range. On the 10 calls with undetected word-errors the ratios of $P_u/\bar{P}_u$ ranged from 0.83 to 24.8, with an average value of 7.4. On 7 out of the 10 calls, $P_u/\bar{P}_u$ was less than 10. The average value of $\bar{P}_u$ over all calls was 2.8×10^{-8}, which is indeed a good approximation to the over-all error rate noted above for the simulation.

The foregoing example suggests that order-of-magnitude accuracy may be obtained using the $\bar{P}_u$ estimate for P_u in ordinary circumstances. To investigate the question of accuracy further, 35 different codes with block lengths less than 25 bits were analyzed on a variety of Gilbert channels. The exact P_u values and $\bar{P}_u$ estimates were compared and found generally to agree within an order of magnitude except in some extreme cases. In these extreme cases, both $\bar{P}_u$ and P_u are practically zero, yet their ratio is large.

It should be noted that, whereas no definitive statement about the accuracy of $\bar{P}_u$ is presently possible, there are practical advantages associated with its use. First, the analysis of the code and channel are separated so that, once the channel has been analyzed for a given block length, many codes of that block length may be evaluated and compared. Secondly, the amount of computation required is significantly less than that required using various simulation techniques. There is one notable limitation imposed on its use. When the code is very large, the amount of computation required to obtain $w(m)$ may be prohibitive. In such cases the approximation offered by (10) may be useful.

VIII. A SAMPLE SURVEY OF CODES

Table V

To further illustrate the sort of code evaluation programs that the $\bar{P}_u$ estimate may be employed in, a collection of 29 codes of various block lengths and redundancy were evaluated using the $P(m,n)$ data from the field tests as summarized in Table III. The codes are all cyclic codes with exception of the constant-weight 4-out-of-8 code, and they have, for their given block length and redundancy, the largest minimum distance attainable with cyclic codes. They are designated in Table V by the number pair (n,k), where n is block-length and k is the dimension of the code. In most cases, when there are two codes with the same (n,k) but with different $w(m)$ values, both codes are included in the evaluation. The difference between the evaluation of these codes and the previous evaluation of the Bose-Chaudhuri (31, 21) code is that here the average undetected error-rate over all calls is calculated instead of an individual rate for each call. The distribution of call types in the field test program is not ideal for taking such an average as a figure of merit, yet the average does provide a convenient single number for each code, and, moreover, a considerable delineation of requirements

TABLE V — UNDETECTED ERROR-RATE ESTIMATES FOR A SAMPLE COLLECTION OF CODES

$n = 8$		$n = 15$		$n = 17$		$n = 21$		$n = 23$		$n = 31$	
k	$\overline{P}_u$	k	$\overline{P}_u$	k	$\overline{P}_u$	k	$\overline{P}_u$	k	$\overline{P}_u$	k	$\overline{P}_u$
*	1.7×10^{-5}	11	2.6×10^{-6}	9	9.4×10^{-8}	15	6.4×10^{-7}	12	1.0×10^{-8}	21	2.8×10^{-8}
7	3.1×10^{-5}	10	1.0×10^{-6}	8	4.1×10^{-8}	12	4.9×10^{-7}	11	4.6×10^{-9}	21	3.0×10^{-8}
4	1.1×10^{-6}	7	8.4×10^{-8}			11	2.5×10^{-8}			20	1.2×10^{-8}
		6	3.7×10^{-8}			9	4.2×10^{-9}			20	1.1×10^{-8}
		6	3.5×10^{-8}			5	2.0×10^{-10}			16	8.2×10^{-10}
		5	1.2×10^{-8}							15	3.9×10^{-10}
		4	5.3×10^{-9}							15	3.5×10^{-10}
		2	1.7×10^{-9}							11	1.4×10^{-11}
										10	7.4×10^{-12}

* The 4-out-of-8 code.

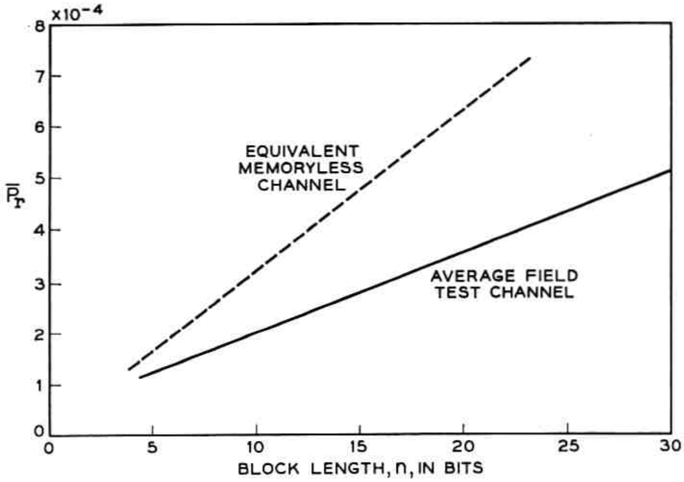


Fig. 3 — Probability of retransmission $\bar{P}_r$ versus block length n.

would be necessary to devise an improved set of weighting factors. There is the further consideration that, when ranked according to error rates, the relative positions of codes would remain almost unchanged by such a refinement.

The probability $\bar{P}_r$ of retransmission is given as a function of code block length in Fig. 3. The slight differences in $\bar{P}_r$ between different codes of the same block length are too small to be noted at three-decimal accuracy. Also plotted in Fig. 3 are the retransmission rates for a memoryless binary symmetric channel having the same average probability $P_1 = 3.2 \times 10^{-5}$ of a bit being in error. This second curve is above the first, since errors are more broadly scattered on the memoryless channel and consequently cause more retransmissions.

IX. CONCLUSIONS

In the search for suitable codes for a given data transmission service, the problem of predicting or evaluating performance is encountered. Several mathematical models of communication channels exist for which the calculation of error rates may be easily performed using parameters associated with the channel. Of such models, we note particularly that Gilbert's burst-noise channel is to be included, and we have outlined the appropriate methods for these calculations. Not all channels, however, admit to a representation by such reasonable models. At this point, models could be abandoned completely and recourse could be taken to actual field testing of a complete system or to the simulation of a complete system using data obtained in field testing. Short of such complete abandonment of models is the method of approximation of code performance factors which has been presented here. Useful mostly for error detecting codes, the method separates the analysis of performance into two parts. The channel is characterized by the probabilities of various numbers of bit errors occurring in a block of given length. A code is characterized by the average number of code words at specified distances from other code words. A simple combination of these two types of quantities gives a useful and economical indication of code performance applicable to general binary block codes and to asymmetric channels with memory. The numbers resulting from such analysis are probably more valuable for a relative indication of performance than they are for an absolute indication. In this connection, it is well to note that when error rates are very low, small differences are operationally of little significance.

As an exemplary application of this method, a collection of 29 codes was evaluated for use on the switched telephone network as error-detecting codes, in conjunction with retransmission as a means of error-correction. The codes in this collection present a wide range in reliability and indicate that it would not be difficult to select appropriate codes for specific data transmission services by suitably enlarging the class of codes examined.

X. ACKNOWLEDGMENTS

The author would like to acknowledge his indebtedness to W. R. Cowell and Robert Morris for their assistance in this undertaking and to Eric Wolman for his support and advice.

REFERENCES

1. Gilbert, E. N., Capacity of a Burst-Noise Channel, B.S.T.J. **39**, September, 1960, p. 1253.
2. Alexander, A. A., Gryb, R. M. and Nast, D. W., Capabilities of the Telephone Network for Data Transmission, B.S.T.J., **39**, May, 1960, p. 431.
3. Peterson, W. W., *Error-Correcting Codes*, John Wiley and Sons, New York, 1961.
4. Slepian, D., A Class of Binary Signaling Alphabets, B.S.T.J., **35**, January, 1956, p. 103.
5. Feller, W., *An Introduction to Probability Theory and its Applications*, Vol. I, second edition, John Wiley and Sons, New York, 1959.
6. Morris, R., Further Analysis of Errors Reported in Capabilities of the Telephone Network for Data Transmission, B.S.T.J., **41**, July, 1962, p. 1399.

Speech Volumes on Bell System Message Circuits—1960 Survey

By KATHRYN L. MC ADOO

(Manuscript received November 7, 1962)

Speech volumes of customers on Bell System message circuits have been measured at class 5 offices. Data are presented for intrabuilding, interbuilding, tandem and toll connections. Average speech volumes are lowest for intrabuilding calls and increase in level for the other types of connections, with volumes on toll calls being the highest. In general, volumes on business calls are higher than those on social calls, and men speak louder than women. Speech volumes remain substantially the same in locations comparable to those in a survey made in 1950.

I. INTRODUCTION

The volume of message signals at various points in the telephone network is of importance to those who design and engineer telephone systems and equipment, and ultimately, of course, to the listener at the far end of the connection. This volume is influenced not only by the speech pressure produced by the talker and by his habits in using the telephone set, but also by the characteristics of the set, the battery supply and loop resistance, and the electrical loss (or gain) between the set and the point at which knowledge of the level is desired.

Speech signals are very complex quantities varying in amplitude from instant to instant. They are measured in a prescribed manner on a standardized meter known as a volume indicator. Data obtained using this technique are called speech volumes and are expressed as volume units (VU) on a db scale. Such measurements are of value to engineers who design equipment, determine crosstalk objectives and permissible noise levels, and otherwise engineer the telephone network.

Since changes in the telephone plant affect transmission performance of the lines, and consequently may affect the customers' habits in the use of the telephone set, up-to-date information on customer speech volumes is necessary. When the last general survey of speech volumes on Bell System message circuits was made in 1950–1951,[1] a large percentage of telephone sets were 200 and 300 types.[2] It is now estimated that 65 per cent of the telephone sets are 500 type;[3] finer-gauge conductors are used in the loop plant; and interoffice trunk and toll circuit losses have been reduced.

This paper presents the results of speech volume measurements made in 1959 and 1961. Observations were made in 1959 in cities larger than 10,000 population, and observations were made in 1961 in smaller communities. The aggregate results of the two groups are referred to as the 1960 survey. The particular cities and offices in which speech volumes were measured were selected as representative of the range of offices in the Bell System, but no rigorous sampling procedure was used. It is believed that the conclusions drawn from the data are sufficiently accurate to serve as a guide for plant design. The measurements were made at class 5 (local or end) offices and were limited to the speech volumes of customers connected directly to that office (near-end talker).

II. SUMMARY

More than 14,000 speech volumes were measured in 30 central offices in 23 cities located throughout the United States. These cities varied in size from single-office cities to large metropolitan centers and their suburban areas. Observations were made on intrabuilding, interbuilding, tandem and toll connections (Fig. 1) in crossbar, step-by-step, panel and Community Dial Offices (CDO's). The locations and office designations are shown in Table I. Some observations were also made in the private branch exchange (PBX) in the Murray Hill, New Jersey, location of Bell Telephone Laboratories. These latter measurements were taken in 1959 when the Murray Hill PBX was of the step-by-step type.

The weighted average speech volumes derived in the 1960 survey are shown in Table II. These averages and all others are obtained by weighting the data according to the population represented by each city unless specifically stated otherwise.

Thirty per cent of the intrabuilding calls and 52 per cent of the interbuilding calls were of a business nature. Fifty-eight and 80 per cent of tandem and toll calls, respectively, were of a business nature.

Averages derived in individual class 5 offices are shown in Figs. 2 and 3.

The large spread or variation in speech volume, as shown by standard deviations of 5.9 to 7.3 db, is caused only in small part by differences in transmission losses of various loop lengths and by different telephone set supply currents. This is supported by consideration of the results ob-

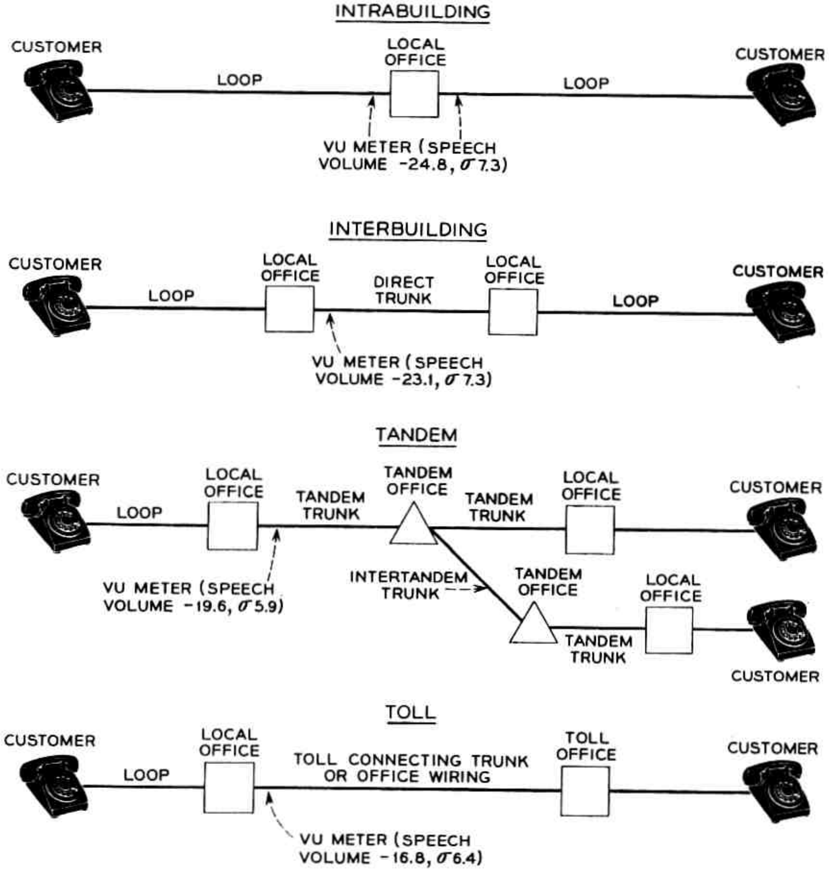


Fig. 1 — Average speech volumes on typical telephone connections (1960 survey).

tained for the PBX at the Murray Hill Bell Telephone Laboratories. In Murray Hill, all extensions had short loops, few of which exceeded 2000 feet. On intra-PBX calls the standard deviation was 5.5 db, indicating that the spread is largely a result of differences in levels and habits of individual speakers.

The variation in the average speech volumes among offices (Fig. 2) is substantial. Examination of the data reveals that speech volumes in New York City average 2 to 3 db higher than in other locations where similar loop plants exist. In general, the higher speech volumes are associ-

TABLE I — SUMMARY OF SPEECH VOLUME DATA

Location	Year Surveyed	1960 Population	Office Designation	Type of Office	Area of Office Sq. Mi.	No. of Stations	Avg. Loop Length* (feet $\times 10^3$)
Atlanta, Ga.	1959	487,455	JA 345	SXS		313,000	8
			CE 7	SXS			8
Auburn, N.Y.	1961	35,249	AL 23	NO5XBR	126.5	20,880	8
Austin, Tex.	1959	186,545	GR 37	SXS	26.3	77,000	6
			HO 5	SXS	49.3		8
Boone, Ia.	1961	12,468	GE 2	NO5XBR	238.0	7,482	10
Cleveland, Miss.	1961	10,249	VI 3	SXS	180.0	4,501	10
Cortland, N.Y.	1959	19,181	36	SXS	159.4	13,000	10
Drew, Miss.	1961	2,143	745	335 CDO	35.0	724	7
Enid, Okla.	1961	38,859	AD 47	NO5XBR	150.0	19,731	9
Ithaca, N.Y.	1959	28,799	234	SXS	134.4	24,000	11
Liberty, Mo.	1961	8,909	STI, THI	NO5XBR	180.0	5,296	7
Medford, N.J.	1961	4,356	OL 4	335 CDO	48.0	4,356	7
Moss Point, Miss.	1961	8,510	GR 5	SXS	195.0	3,972	8
Mount Holly, N.J.	1961	13,271	AM 7	NO5XBR	54.8	6,686	8
New York, N.Y.	1959	7,810,000	WO 4	NO1XBR		4,204,000	3
			SW8, LO5	NO1XBR			4.5
			WA 378	PAN			4.5
Pascagoula, Miss.	1961	17,139	SO 2	NO5XBR	72.0	8,361	8
Plainfield, N.J.	1959	45,330	PL 4567	NO1XBR	49.4	49,660	6
Ridgewood, N.J.	1959	25,391	GI 43	NO5XBR	18.3	34,780	8
San Francisco, Cal.	1959	742,855	Main EX7	NO1XBR	4.8	519,000	4
			Main YU7	PAN	4.8		4
			MO 4, LO 4	NO1XBR	8.2		8
Sioux City, Ia.	1959	89,159	Main 2578	SXS		41,000	8.5
			Morn 6	SXS			7
Skaneateles, N.Y.	1961	2,921	OV 5	SXS	58.0	3,569	9
Trenton, N.J.	1961	114,015	OW 5	SXS		60,000	10
Waukomis, Okla.	1961	516	PL 8	350 CDO	140.0	437	9
Woodland, Cal.	1959	13,524	MO 2	SXS	210.0	8,000	11

* Estimate— except Plainfield and Ridgewood.

ated with the larger cities. Differences in the percentage of business calls are one contributing factor. Others may be talking habits, ambient noise and average length or loss of loops.

As observed in the 1950 survey, speech volumes on long-distance calls increase approximately 1 db for every 1,000 miles.

There is a 4-db variation in the average speech volume of males, depending on the sex of the far-end talker and whether the call is of a

Table I — Summary of Speech Volume Data (Cont.)

Location	Near-End Speech Volumes									
	Intrabuilding		Interbuilding			Tandem			Toll	
	Avg. VU	Std. Dev. db	Trunk Loss-db	Avg. VU	Std. Dev. db	Trunk Loss-db	Avg. VU	Std. Dev. db	Avg. VU	Std. Dev. db
Atlanta, Ga.	−22.6	6.2	4.0	−20.9	6.1				−15.4	5.2
			8.0–9.0	−20.1	5.9					
	−23.2	5.5	5.2	−21.5	6.0	4.6	−21.7	5.0	−16.8	5.1
			8.7	−21.4	5.4					
Auburn, N.Y.	−27.3	8.1	3.9	−23.3	6.5				−23.1	5.7
Austin, Tex.	−24.9	6.2	4.7–5.5	−23.8	6.0				−16.4	5.2
	−26.7	6.3								
Boone, Ia.	−27.3	7.7							−21.0	5.8
Cleveland, Miss.	−26.4	7.4							−20.4	6.5
Cortland, N.Y.	−23.5	6.0								
Drew, Miss.	−27.6	8.1								
Enid, Okla.	−28.4	7.0							−20.4	7.5
Ithaca, N.Y.	−25.4	6.7	5.0–6.0	−21.2	6.5				−15.6	4.9
			5.0–6.0	−23.1	6.4					
Liberty, Mo.	−27.4	8.5	3.3–5.7	−27.3	6.5	2.5	−23.2	6.1	−19.6	6.6
Medford, N.J.	−27.6	5.6	5.5	−25.6	7.3					
Moss Point, Miss.	−27.6	6.9	1.2	−26.8	6.3				−21.3	5.3
Mount Holly, N.J.	−26.6	6.3	5.5	−24.1	7.2					
New York, N.Y.	−18.9	5.6	6.5–8.0	−16.2	5.0	1.8–3.4	−17.4	5.3	−11.0	5.2
						8.6	−17.6	4.9		
	−17.7	6.2	6.9–8.8	−16.4	4.9	4.0	−18.1	5.4	−14.2	5.8
						7.7–9.0	−16.4	5.4		
	−18.8	5.8								
Pascagoula, Miss,	−26.0	7.3	1.2	−25.8	6.9					
Plainfield, N.J.	−22.0	6.2	4.7	−20.1	5.6	3.9	−19.9	4.8	−16.9	5.3
			11.0	−19.2	5.0	10.5	−17.0	5.8		
Ridgewood, N.J.	−22.1	5.6	5.0	−21.0	6.2	3.6–4.0	−18.6	5.8	−15.0	5.1
			9.3–10.0	−19.8	5.4	10.1	−18.0	4.8		
San Francisco,	−21.8	5.4	6.6	−19.4	5.9	3.3–4.9	−19.1	4.9	−14.4	4.8
Cal.	−20.0	5.9	6.6–6.7	−20.2	6.2	7.6–7.8	−18.2	5.9	−14.7	4.9
	−24.9	6.7	6.6	−20.2	5.1	3.0	−20.7	5.4	−17.8	4.7
Sioux City, Ia.	−23.8	6.9	3.9	−22.3	5.9				−14.9	5.9
	−24.5	6.6								
Skaneateles, N.Y.	−24.7	7.2	3.9	−24.6	6.4					
Trenton, N.J.	−25.4	6.3								
Waukomis, Okla.	−25.0	6.7								
Woodland, Cal.	−23.9	6.3							−55.8	5.7

Table II — Summary of Near-End Speech Volumes

Type of Connection	Average VU	Standard Deviation db	Maximum Observed VU	Minimum Observed VU
Intrabuilding	−24.8	7.3	−2.1	< −50.0
Interbuilding	−23.1	7.3	−2.6	−46.0
Tandem	−19.6	5.9	−3.0	−40.4
Toll	−16.8	6.4	+5.3	−39.8

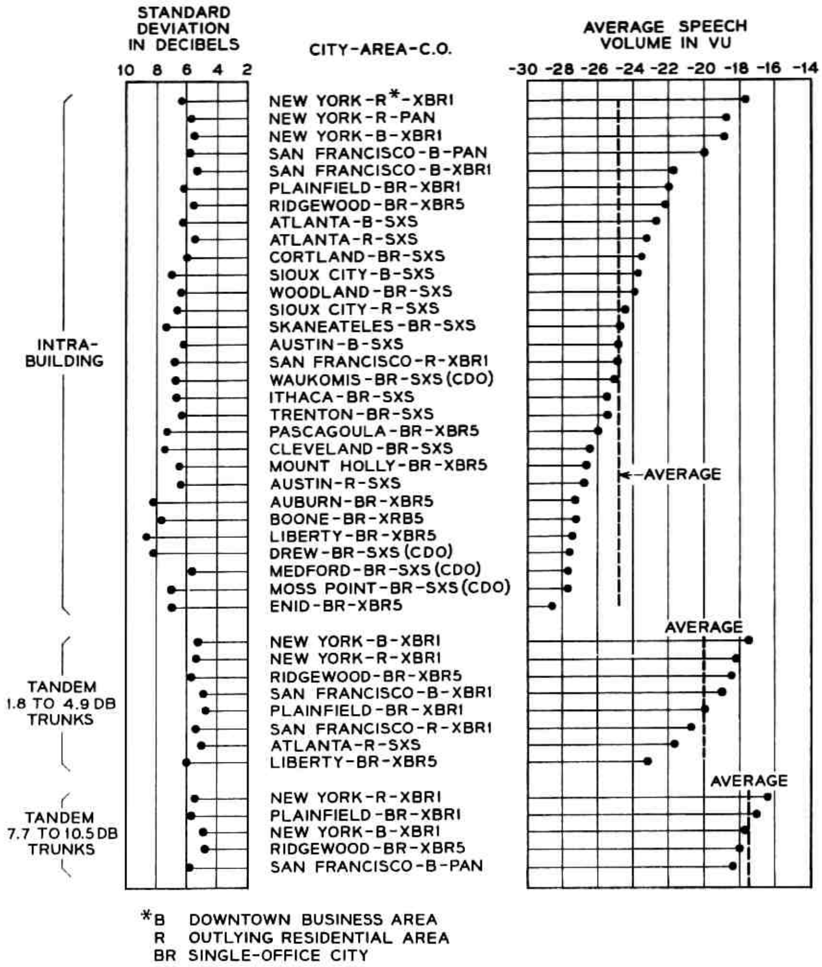


Fig. 2 — Average intrabuilding and tandem speech volumes.

social or business nature. The variation in the average speech volume of females is smaller.

Speech volumes on business calls average slightly higher than those on social calls, partially because business talkers are predominantly men and business calls tend to be over longer distances.

Speech volumes measured in the 1960 survey appear at first glance to be lower than those measured in the 1950 survey, with decreases

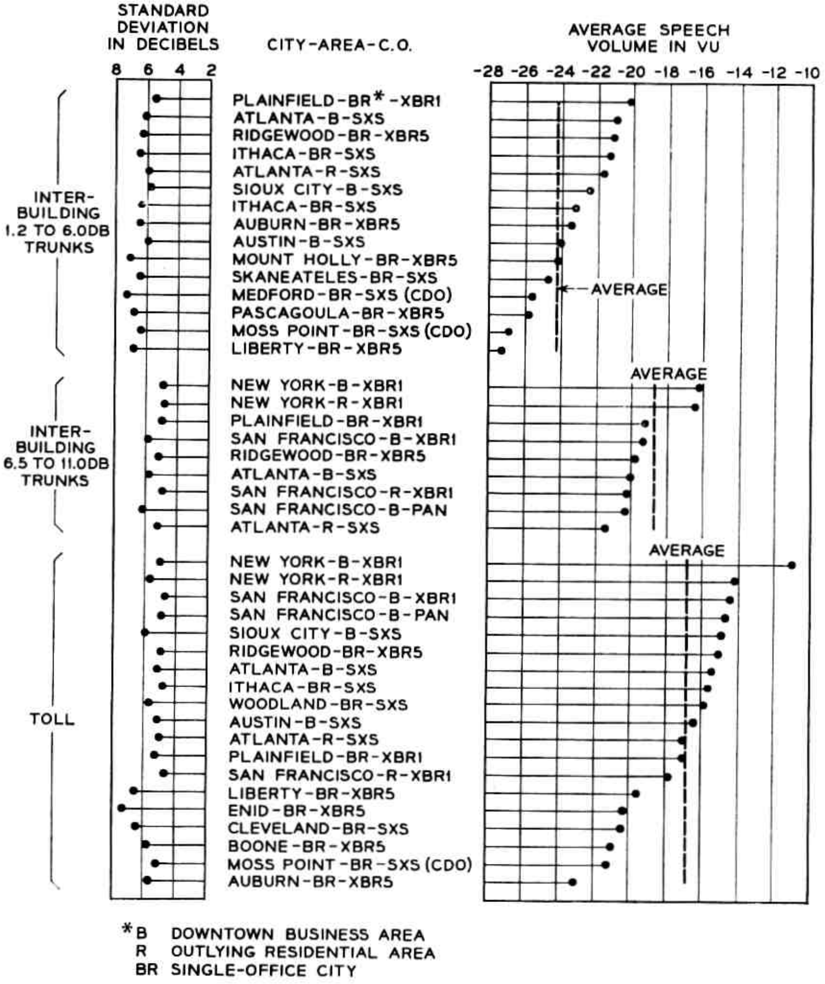


Fig. 3 — Average interbuilding and toll speech volumes.

varying from 2.2 db on tandem calls to 5.8 db on local or intrabuilding calls. However, this is largely due to the fact that New York speech volumes, which are higher than average, comprised more than one-third of the measurements made in 1950. In the wider sample in the present survey, New York City speech volumes account for less than 10 per cent of the total number. For comparable locations, the data of the two surveys are in substantial agreement.

III. DESCRIPTION OF TESTS

Near-end talker volumes were measured in 30 central offices located in 23 cities throughout the United States. Detailed information on the cities and central offices is given in Table I. The communities range in size from 516 people in Waukomis, Oklahoma, to nearly eight million in New York City. The offices included No. 1 crossbar, No. 5 crossbar, step-by-step and panel offices; 350A and 355A Community Dial Offices (CDO's). In larger cities data were obtained in offices in both business and residential areas.

Measurements were made on intrabuilding, interbuilding, tandem and toll connections. The types of connections are illustrated by simple schematics in Fig. 1. In many of the smaller central offices there were neither interbuilding connections nor tandem switching. Toll observations were not made in some locations where the traffic was too slow to warrant spending the amount of time necessary to obtain a complement of measurements.

Records were kept of the sex of the near-end and far-end talker and the nature of the call, whether social or business. Additional information was obtained on loop lengths, trunk losses, and station sets. Most observations were made during the day; however, some observations on toll calls were made in the evening when a greater possibility of obtaining social calls existed.

IV. TEST EQUIPMENT

Measurements were made at convenient circuit locations in each office, using a high-impedance standard volume indicator.[4] Two different models were used, both with a nominal input impedance of 12,500 ohms and a response essentially flat from 50 to 15,000 cps. On one volume indicator the range of volumes which can be read in accordance with the method described below is -32 VU to $+30$ VU. Speech volumes a few db lower than -32 VU can, however, be estimated with reasonable accuracy. The other instrument has a range of -42 to $+20$ VU, thus allowing for greater accuracy in reading the low speech volumes. It also has an optional 60-cps elimination filter, the loss of which is not detectable above 300 cps.

Volume indicators are calibrated to read voltage across 600 ohms. However, the impedances of most exchange telephone circuits generally differ substantially from 600 ohms, thereby causing appreciable errors in volume indicator readings. Corrections were computed from the formula

$$C + 10 \log_{10} \left(\frac{600}{| Z |} \right)$$

where $| Z |$ is the magnitude of the impedance of the circuit into which speech volumes are measured, derived from knowledge of the average gauge, length, loading, and termination of the circuit for the type of call being measured. Trunk and loop data were supplied by operating company personnel. The final correction was a weighted average of the corrections at four or five important frequencies in the speech band.

V. TEST PROCEDURE

At the beginning of an observation, the observers waited for the connection to be established and thereby distinguished the near- and far-end talkers by their salutations. In no case was volume used as the sole criterion in identifying the parties, for difference in volume in many cases exceeded the transmission differences between customers.

The standard procedure for measuring speech volumes on telephone message circuits requires taking the arithmetic average of a series of individual volume measurements on each customer. An individual volume measurement is defined as the visual average of five to six of the highest meter deflections over a 3- to 10-second interval. In so doing, the occasional high peaks and the series of low peaks are ignored. An input attenuator, adjustable in 2-db increments, is set to allow the peaks used in determining the average to fall in the region from 0 to −2 VU on the meter scale. About ten individual measurements were averaged to obtain the speech volume of the customer.

For each type of call at a location, speech volumes of 120 to 160 customers were measured by two observers. Preliminary training of all observers consisted of practice in reading volumes from recordings of the 1939 World's Fair telephone exhibit. Throughout the survey the observers rechecked their methods of reading the volume indicator in order to eliminate the possibility of developing poor habits.

VI. METHOD OF COMBINING DATA

The principal objective of these speech volume measurements was to derive a system-wide average and standard deviation for each of the four types of calls. This involves first assuming that the values obtained for an office are representative of similarly located offices throughout the United States, and then combining the data in accordance with the calling rate in the different kinds of areas. Neither of these factors is

Table III — U. S. Population Statistics 1960: Method of Determining Weighting Factors

Population (Thousands)	Number of Cities	Total Population in Cities	Weighting % of Total Population	City Sampled
>1000	5	17,290,300	9.6	New York, N. Y.
500–1000	15	10,442,300	5.8	San Francisco, Cal.
250–500	31	11,078,300	6.2	Atlanta, Ga.
100–250	76	11,078,500	6.2	Austin, Tex. Trenton, N. J.
50–100	178	12,369,300	6.9	Plainfield, N. J. Ridgewood, N. J. Sioux City, I.
25–50	403	14,815,500	8.3	Auburn, N. Y. Enid, Okla. Ithaca, N. Y.
10–25	1099	17,052,500	9.5	Boone, Ia. Cleveland, Miss. Cortland, N. Y. Mount Holly, N. J. Pascagoula, Miss. Woodland, Cal.
5–10	1381	9,697,300	5.4	Liberty, Mo. Moss Point, Miss.
<5	*	75,498,100	42.1	Drew, Miss. Medford, N. J. Skaneateles, N. Y. Waukomis, Okla.
Total		179,322,100	100.0	

* No estimate available. This category includes unincorporated places less than 1000 population and other rural population not included in other categories.

accurately known, but useful values can be obtained by accepting the measured speech volumes as representative and weighting them in accordance with population.

The population statistics and the weighting factors used to obtain the composite averages are given in Table III. The population statistics are taken from the 1960 *Census of Population — Advance Reports* distributed by the U. S. Department of Commerce.

VII. OBSERVATIONS ON INTRABUILDING, INTERBUILDING, TANDEM AND TOLL CONNECTIONS

Average speech volumes obtained in each office are shown in Figs. 2 and 3 for the four types of calls. These averages, when combined using the weighting factors given in Section VI, yield the Bell System averages. These are shown in Table II.

The data for interbuilding and tandem calls shown in Figs. 2 and 3 are separated into two groups according to trunk losses. This illustrates the effect of trunk loss on speech volume. On the average there is a 1-db increase in speech volume for every 3-db increase in trunk loss.

Locations with high speech volumes on intrabuilding calls have consistently high speech volumes on the other types of calls. Conversely, locations with low speech volumes on intrabuilding calls have low speech volumes on other types of calls. The high speech volumes are, with few exceptions, found in the larger cities. Lower-loss loops and a greater incidence of business calls may be contributing factors. Regional speech characteristics and other factors which cannot be ascertained by measurement (for example, some hypothesis has naturally been made on the effect of the faster pace of urban living than rural living on speech volumes) may contribute to the differences in speech volumes from office to office.

The standard deviation associated with the average tandem speech volume is considerably smaller than the standard deviations for the other kinds of calls. This is probably because tandem switching is largely confined to metropolitan areas and the calling population is more homogeneous than if it were scattered throughout the country. This same factor may account for the high average level of this type of call.

VIII. SPEECH VOLUMES OF MALES AND FEMALES ON SOCIAL AND BUSINESS CALLS

The speech volumes of male and female talkers on social and business calls are interesting to note and may be of use in the design of some special systems. The intrabuilding, interbuilding, tandem and toll speech volumes have been combined without any weighting to give an indication of the relative difference in speech volumes as illustrated in Fig. 4. These are averages for all types of connections and therefore do not indicate the actual levels of measured volumes.

The average speech volume of the female talker remains within a 1-db range, whereas that of the male talker drops as much as 4 db when the far-end changes from male to female. Over-all, men tend to talk slightly louder than women, and business conversations are louder than social ones.

Approximately 73 per cent of the business calls observed were made by male speakers, whereas females made 81 per cent of the social calls. The majority of the tandem and toll calls were made by men, and most of the local telephone calls were made by women.

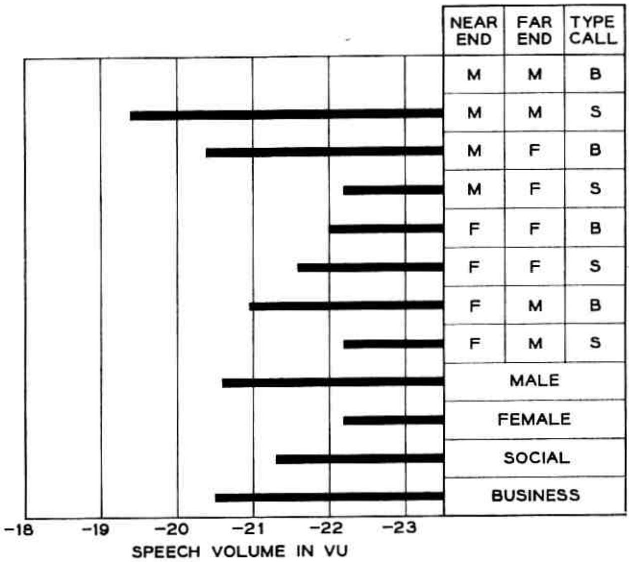


Fig. 4 — Speech volumes of males and females, social and business calls.

IX. DISTANCE EFFECT

In New York special observations were made in several toll centers on circuits to Philadelphia, Chicago and Mexico City. These data are not included in the previously discussed averages. They are summarized in Table IV.

These data illustrate the distance effect observed by V. Subrizi in the 1950 survey. In spite of the lower circuit losses on the long connections, there is an increase in near-end speech volume of approximately 1 db per 1000 miles. This increase may be caused by increased noise and distortion on longer toll connections or may be psychological.

X. PBX OBSERVATIONS

Some preliminary speech volume measurements were made in the 701A PBX at Murray Hill Bell Telephone Laboratories on intra-PBX calls and on tie lines to the West Street, New York, and Whippany, New Jersey, Laboratories. Tie line losses to New York varied from 3.9 to 7.5 db and those to Whippany from 4.0 to 5.0 db. The average speech volumes obtained are shown in Table V.

These volumes are generally higher than the composite averages for local offices. Contributing factors are short loops and the fact that the calls are predominantly made by men talking business. The effect of variation in loop loss on the standard deviation is virtually eliminated in these PBX observations, and variations due to station sets are greatly reduced because their current supply is at a uniformly high level. The standard deviation is still large, indicating that the spread in speech volumes is largely a result of variation in individual habits and speaking levels rather than in loop and station characteristics.

XI. COMPARISON WITH 1950 SPEECH VOLUME SURVEY

One of the interesting questions posed by the surveys in 1950 and 1960 is whether speech volumes are increasing or decreasing. This is a

Table IV — Long Distance Observations at New York Toll Centers at Zero Level Point

Terminal	Speech Average VU	Volume Sigma db	Circuit Loss db	Air Miles
Philadelphia	−15.3	4.8	7.8	80
Chicago	−14.3	4.0	6.0	850
Mexico City	−12.7	4.4	5.0	2094

Table V — Speech Volume Observations at the Murray Hill Laboratories PBX

	Average VU	Standard Deviation db
Intra-PBX	−17.8	5.5
Tie line to Whippany, N. J. Laboratories	−17.7	4.7
Tie line to West Street, N. Y. C. Laboratories	−16.7	4.9

Table VI — Comparison of Speech Volumes in 1950 and 1960

1950 Survey			1960 Survey		
Connection	Speech Volume		Connection	Speech Volume	
	Avg. VU	Std. Dev. db		Avg. VU	Std. Dev. db
Local	−19.0	5.7	intrabuilding	−24.8	7.3
			interbuilding	−23.1	7.3
Tandem	−17.0	5.8	tandem	−19.2	5.9
Toll	12.0*	5.3*	toll	−16.8	6.4

* Measured at toll office, but corrected back to local office by toll connecting trunk loss.

difficult question to answer, since the two surveys varied widely in scope and since some of the office areas measured in both surveys have changed considerably. A summary of both surveys is shown on Table VI.

The averages for 1950 were obtained in Atlanta, Ga.; Cleveland, O.; and New York, N. Y. Local calls as defined in the 1950 survey include intrabuilding calls and short (or with low-loss trunks) interbuilding and tandem calls.

For locations comparable to the three cities observed in 1950, speech volumes now average 0.5 db lower on local calls and vary from a few tenths of a db to 2 db lower on toll calls. These differences are too small to be considered significant.

XII. LIMITATIONS OF SURVEY

Caution is advised against using these data for engineering systems used by private, military or air control personnel. These data apply only to Bell System customers working into the switched Bell System network. Very much higher talker volumes have been observed in limited measurements of military and private-line networks.

XIII. ACKNOWLEDGMENT

Acknowledgment is gratefully made to the many persons in the operating companies and the American Telephone and Telegraph Company whose cooperation made the survey possible.

REFERENCES

1. Subrizi, V., A Speech Volume Survey on Telephone Message Circuits, Bell Laboratories Record, **31,** August, 1953, pp. 292–295.
2. Inglis, A. H., Transmission Features of the New Telephone Sets, B.S.T.J., **17,** July, 1938, pp. 358–380.
3. Inglis, A. H., and Tuffnell, W. L., An Improved Telephone Set, B.S.T.J., **30,** April, 1951, pp. 239–670.
4. American Standard Practice for Volume Measurements of Electrical Speech and Program Waves, American Standards Association, C16.5, Nov., 1954.

A Self-Steering Array Repeater

By C. C. CUTLER, R. KOMPFNER and L. C. TILLOTSON

(Manuscript received February 8, 1963)

A scheme is disclosed whereby an antenna array is automatically directed by a simple intermodulation of signal components. In reception, each array element feeds a pilot signal and the modulated signal to a third-order mixer wherein the phase associated with the signal in that element is automatically cancelled. This allows in-phase addition of the contributions from the many elements irrespective of the array shape or the direction of the incoming signal. For transmission, a pilot signal received from the distant receiver location provides by intermodulation a phase compensation to the signal radiated from each transmitting element so as to automatically direct the radiated signal to the distant receiver. There are no significant restrictions as to the shape of the array or the frequencies used.

The scheme lends itself to multiple-element, low-power circuitry and may be used in either space or terrestrial systems to give a high repeater directivity without requiring stabilized platforms or control of antenna orientation. An experimental verification of the basic principle is described.

I. INTRODUCTION

Antenna directivity has become widely used to provide a high effective radiated power with only modest transmitted power, particularly at microwave frequencies where antennas having apertures of many wavelengths are of reasonable size. The precise aiming of the antennas made necessary by this high directivity requires the use of sturdy towers in terrestrial systems, and the proposed use of stabilized platforms for space applications.

A number of methods of avoiding the requirement of accurate orientation and stabilization based upon the Van Atta array concept[1] have been proposed,[2,3,4,5] but each puts strict requirements upon the array shape, and none provides a common intermediate terminal where one may drop and/or add channels. Another method[6] uses a phase-locked loop or servo control to automatically phase the reception from each element for in-phase addition.

The scheme proposed herein avoids many of the limitations of the earlier ones, puts no restrictions on the shape of the array, and does not involve servo control or feedback — either electrical or mechanical. The new scheme compensates for the relative phase of each array element by an intermodulation (frequency mixing) process like that used in some radio diversity[7,8] receivers.

An experimental circuit has been constructed from available hardware to demonstrate in its simplest form the basic principle — coherent addition of microwave signals regardless of relative phase at the input. In-phase addition was obtained as predicted.

Since state-of-the-art microwave solid-state devices provide low power but at relatively high efficiency (i.e., Esaki diodes, varactor multipliers, and microwave transistors), paralleling the power output from many such units in an array provides four distinct advantages:* (*i*) efficient addition of the power from many repeaters, (*ii*) steerability of the beam, (*iii*) high directivity and antenna gain, and (*iv*) reliability—failure of individual units is of little consequence.

Thus it is expected that by the use of modern solid-state devices and micro transmission line techniques, a very simple lightweight self-steering repeater can be built requiring only a fraction of the power of more conventional repeaters with no necessity for orientation control and with the inherently increased reliability provided by a multiplicity of independent parallel circuits.

II. A SELF-STEERING ARRAY REPEATER — BASIC IDEAS

The pointing angle of a steerable array is determined by the relative phasing of the individual elements. Signals received from a distant transmitter by elements of an array differ in phase by an amount which depends upon the geometry of the array, and the relative phases are distributed in a manner exactly opposite to that required for retransmission back toward the source. Van Atta used this fact to show that by suitable interconnection of the elements of a regular linear array, the phase differences can be canceled out, resulting in a return characteristic from an array much like that of a corner reflector[1,2,3] (see Fig. 1).

An alternative which avoids many of the limitations and difficulties encountered when the basic Van Atta array is used in an actual repeater can be explained with reference to a satellite whose orientation is uncontrolled, as shown schematically in Fig. 2. This figure shows an array

* These advantages were pointed out by R. C. Hanson in Ref. 3 for the case of the conventional active Van Atta array.

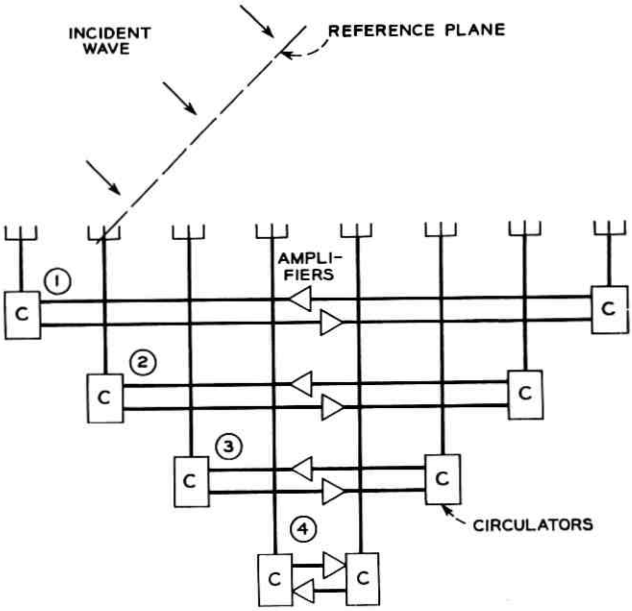


Fig. 1 — An active Van Atta array.

in which the signal received by an element is heterodyned with a locally generated beating oscillator, and the difference frequency is connected back to the same antenna element. All elements are treated alike, and all are excited from a common local oscillator supply. Let the signal received by the ith element be of the form:

$$e_i(t) = \exp j(\omega_R t + \varphi_i) \tag{1}$$

where φ_i is the phase of $e_i(t)$ relative to an arbitrary reference plane normal to the transmission path. The output of the mixer will be:

$$E_i(t) = \exp j[(\omega_B - \omega_R)t - \varphi_i]. \tag{2}$$

If the local oscillator, which is common to all elements of the array, is adjusted to a frequency larger than ω_R, then $(\omega_B - \omega_R)$ is positive, and the phase φ_i will be reversed in sign with respect to that of the received signal, as shown[9] in (2). If, further, ω_B is adjusted to be about twice ω_R

$$(\omega_B - \omega_R) \approx \omega_R \tag{3}$$

and the resultant can be diplexed onto the same array element. The excess phase on transmission just cancels that on reception. This is true

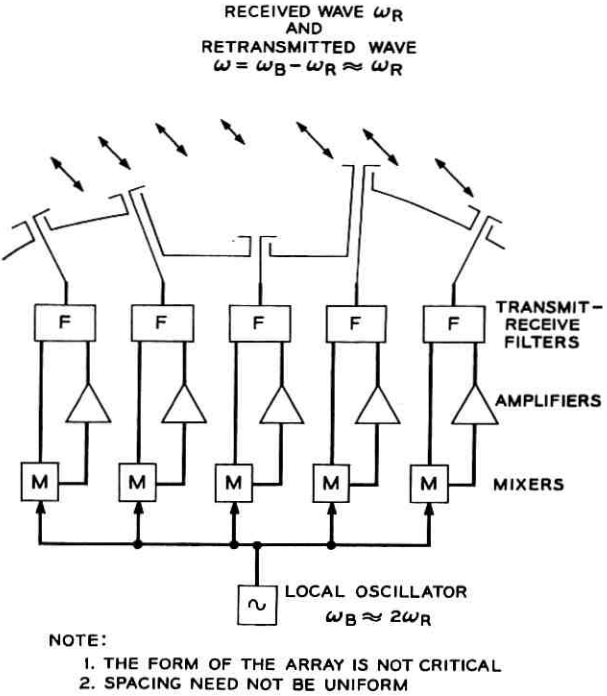

RECEIVED WAVE ω_R
AND
RETRANSMITTED WAVE
$\omega = \omega_B - \omega_R \approx \omega_R$
F
TRANSMIT-RECEIVE FILTERS
AMPLIFIERS
M
MIXERS
LOCAL OSCILLATOR
$\omega_B \approx 2\omega_R$
∿
NOTE:
1. THE FORM OF THE ARRAY IS NOT CRITICAL
2. SPACING NEED NOT BE UNIFORM

Fig. 2 — An elementary form of active converting array.

for all array elements and their associated circuits. Hence the retransmitted signals are phased just right to form a beam directed back toward the distant transmitter. Note that the foregoing is true regardless of the position of the ith element or the shape of the array, and that no interconnection of array elements is necessary except for the common local oscillator.

This cancellation of phase by mixing is basic to all of the systems described herein. In case it is desired to receive independently, the phase mixing can be accomplished as shown in Fig. 3 (see Ref. 8). The received signal is first divided into two parts in a branching filter, and the carrier or a separate pilot frequency is amplified and further separated from the modulation products in a narrow-band amplifier. The three frequencies — i.e., carrier or pilot, modulation, and local oscillator — are then mixed in a third-order mixer (or two separate more conventional mixers) and the third-order product is selected for subsequent demodulation. The phase of the incoming waves (φ_i) is relative to a plane perpendicular to the incoming wave normal, and will be different for each element.

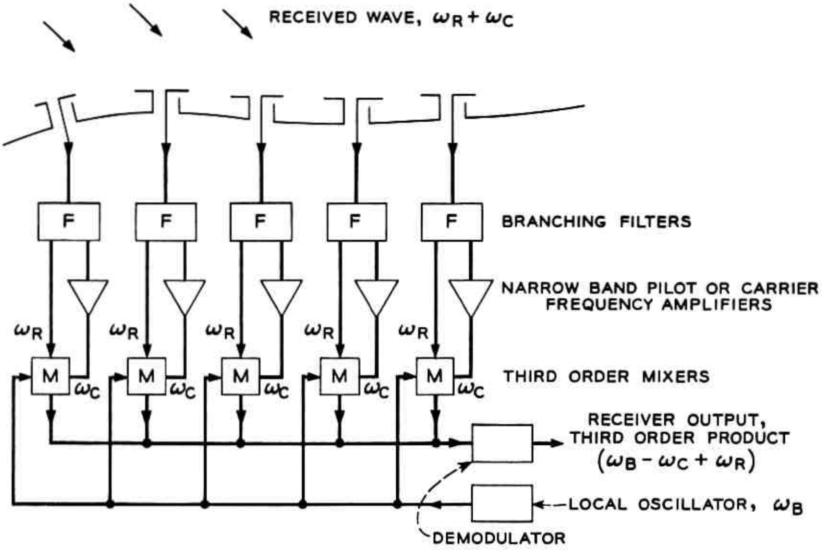

RECEIVED WAVE, $\omega_R + \omega_C$
F
BRANCHING FILTERS
NARROW BAND PILOT OR CARRIER FREQUENCY AMPLIFIERS
ω_R
M
ω_C
THIRD ORDER MIXERS
RECEIVER OUTPUT, THIRD ORDER PRODUCT $(\omega_B - \omega_C + \omega_R)$
LOCAL OSCILLATOR, ω_B
DEMODULATOR

Fig. 3 — An active converting array receiver.

The third-order product is:

$$E_i t = \exp j[(\omega_B - \omega_C + \omega_R)t - \varphi_i + \varphi_i] = \exp j(\omega_B - \omega_C + \omega_R) \quad (4)$$

where ω_B is the local oscillator (radian) frequency, ω_C the received carrier, or a separate pilot frequency, and ω_R the rest of the received signal. Since (4) contains no phase term, evidently the voltages from several such channels can be added.

For transmission of a locally generated modulation, it is evident that modulation applied to the local oscillator of Fig. 2 will be contained in the retransmitted signal. In such a case, the incoming signal acts as a pilot to direct the transmission.

III. ARRAY STEERING BY PILOT FREQUENCY CONTROL

In many applications it is not desired to retransmit in the direction of the received signal. In such a case, a separate pilot signal sent from the distant receiving terminal can serve to define the direction for retransmission, and by the described frequency mixing operation, this can be accomplished automatically. Since the antenna beam is directed or steered toward the distant receiver regardless of its location, the antenna gain can be as large as desired, independent of the changing satellite orientation in space systems or of movement of towers in terrestrial systems.

The advantages of a satellite repeater which does not require orientation control are quite apparent; the advantages to be gained in the application of this scheme to a terrestrial system are also worth noting. Since changes in pointing angles in terrestrial relay systems will be small, the elemental antennas of the array can be relatively high gain, and thus relatively few elements are required to implement a steerable array (STAR) repeater. Hence the advantages of reliability and self-steering can be obtained in terrestrial systems with only small increase in the amount of repeater electronics.

IV. SELF-STEERING SATELLITE REPEATER

Before going into detail, we will describe a prototype repeater embodying the principles described. Since we are interested in partially oriented terrestrial as well as nonoriented satellite repeaters, we will attempt to generalize the discussion to cover both situations. In the terrestrial case, the array elements can profitably use area directivity, and it is desirable to interconnect two separate arrays with elementary repeaters, as shown in Fig. 4. In the satellite case it is more desirable to combine the functions in a single array, or to separate transmitting and receiving functions. The satellite repeater is visualized as spherical and entirely covered with elemental antennas, as shown in Fig. 5.

To insure that all of the antenna elements act in concert as a phased array, and hence as an antenna having an aperture nearly equal to the projected area of the array, it is necessary to combine in-phase the re-

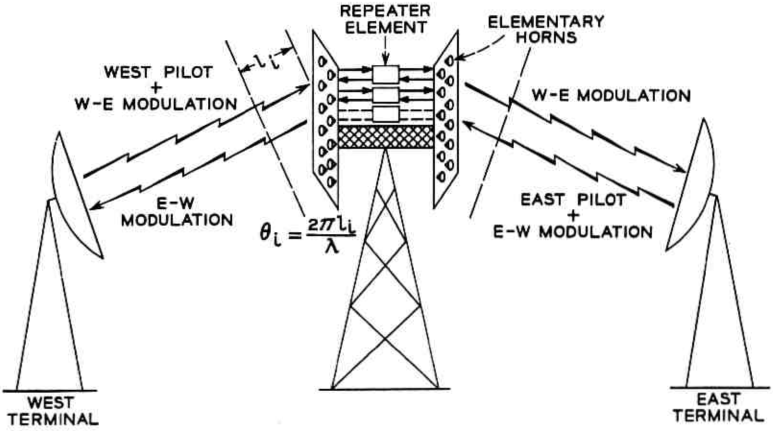

REPEATER ELEMENT
ELEMENTARY HORNS
WEST PILOT + W-E MODULATION
l_i
W-E MODULATION
E-W MODULATION
$\theta_i = \frac{2\pi l_i}{\lambda}$
EAST PILOT + E-W MODULATION
WEST TERMINAL
EAST TERMINAL

Fig. 4 — A possible two-way terrestrial repeater configuration.

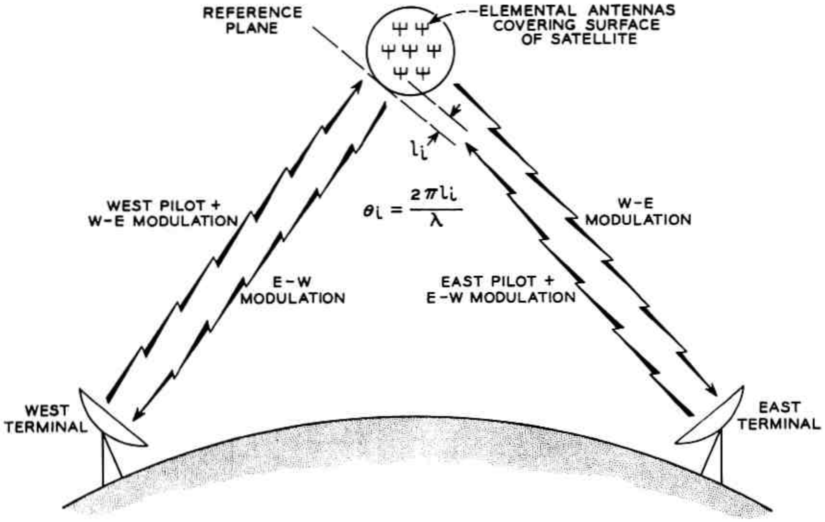

REFERENCE PLANE
ELEMENTAL ANTENNAS COVERING SURFACE OF SATELLITE
l_i
$\theta_i = \frac{2\pi l_i}{\lambda}$
WEST PILOT + W-E MODULATION
W-E MODULATION
E-W MODULATION
EAST PILOT + E-W MODULATION
WEST TERMINAL
EAST TERMINAL

Fig. 5 — A possible two-way satellite repeater configuration.

ceived signals from all the elements. As received from the west terminal, the signals have relative phase shifts

$$\theta_i, \theta_2, \theta_3 \cdots \theta_i, \qquad \text{where } \theta_i = \frac{2\pi l_i}{\lambda},$$

and l_i is the variable distance between the ith element and a reference plane normal to the radius vector to the western terminal, as shown in Figs. 4 and 5. This distance, and hence the phase shift, depends upon the orientation of the array and changes when the array moves relative to the fixed terminals. In-phase addition of the received signals can be accomplished by the use of the pilot beam as follows: Consider first the left-hand part of the two-way repeater shown schematically in Figs. 6 or 7. Signals received from the west terminal by the ith elemental antenna are: (a) the pilot, $\exp j[\omega_P t + \theta_{i,p}]$, and (b) the modulation,

$$\exp j[\omega_{M(\text{W-E})} t + \varphi_{(\text{W-E})} t + \theta_{i,M}].$$

These are passed by the transmit-receive filter and are converted in an intermediate frequency circuit by mixing with a local oscillator in a square-law mixer. The results are

$$A(t) = \exp j[\omega_{LO_1} t - \omega_{P(\text{W})} t - \theta_{i,P}] \qquad (5)$$

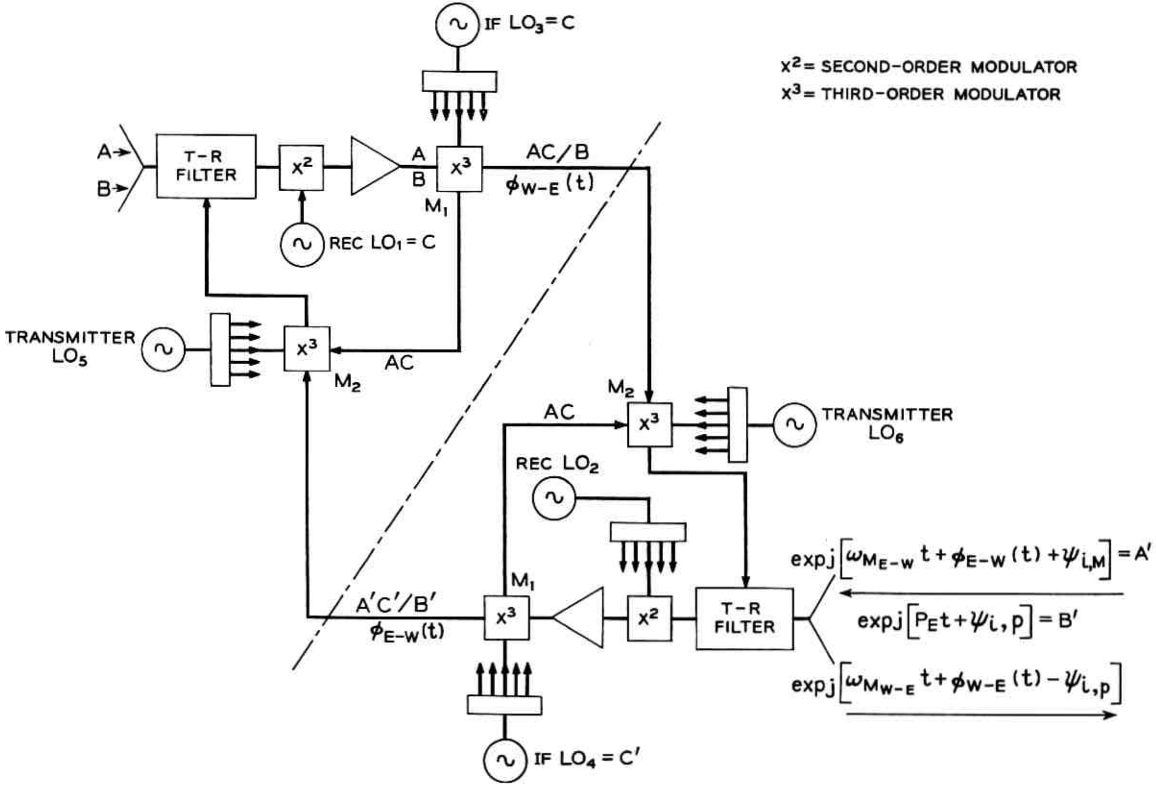

x^2 = SECOND-ORDER MODULATOR
x^3 = THIRD-ORDER MODULATOR
IF $LO_3 = C$
A→
B→
T-R FILTER
x^2
A
B
x^3
M_1
AC/B
$\phi_{W-E}(t)$
REC $LO_1 = C$
TRANSMITTER LO_5
x^3
M_2
AC
M_2
AC
x^3
TRANSMITTER LO_6
REC LO_2
M_1
A′C′/B′
$\phi_{E-W}(t)$
x^3
x^2
T-R FILTER
IF $LO_4 = C'$
$\exp j\left[\omega_{M_{E-W}} t + \phi_{E-W}(t) + \psi_{i,M}\right] = A'$
$\exp j\left[P_E t + \psi_{i,p}\right] = B'$
$\exp j\left[\omega_{M_{W-E}} t + \phi_{W-E}(t) - \psi_{i,p}\right]$

Fig. 6 — Basic elements of a two-way repeater [one of several sections joined with a common local oscillator and attached to separate input and output antenna arrays suitable for terrestrial (fixed) service].

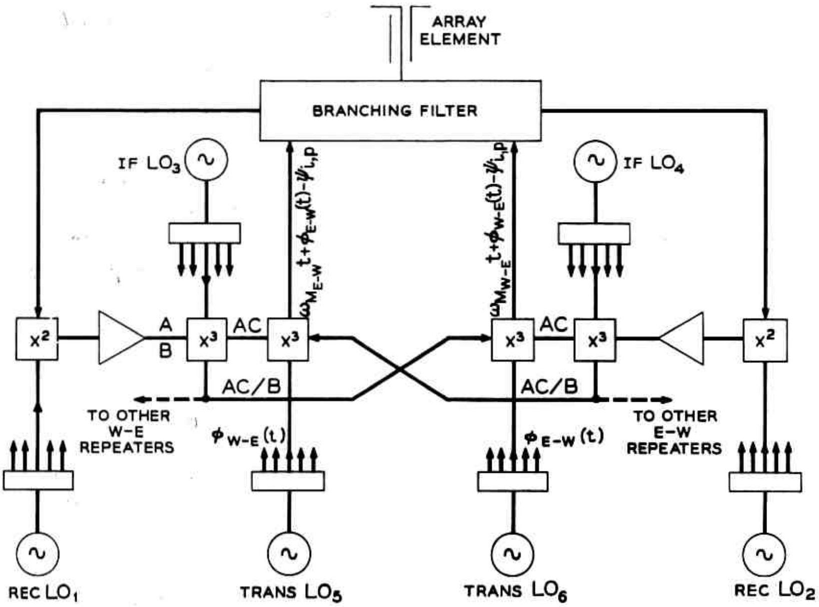


Fig. 7 — Basic elements of a two-way repeater [one of several sections joined by a common local oscillator and diplexed into a common receiving-transmitting array suitable for unstabilized satellite repeater use].

and

$$B(t) = \exp j[\omega_{LO_1}t - \omega_{M(\text{W-E})}t - \varphi_{(\text{W-E})}(t) - \theta_{i,M}] \tag{6}$$

where

ω_{LO_1} = radian frequency of common receiving local oscillator,
$\omega_{P(\text{W})}$ = radian frequency of pilot received from west terminal.

$\theta_{i,M}$ and $\theta_{i,P}$ are the phases relative to the common reference plane, which is equal to $2\pi l_i/\lambda$, where λ is the appropriate wavelength and l_i is the distance between ith antenna element and the reference plane. $\omega_{M(\text{WE})}$ = the radian frequency of the west-east modulation channel carrier and $\varphi_{(\text{W-E})}(t)$ = the angle modulation of the west-east carrier.* These are amplified and put into a third-order mixer along with IF local oscillator signal $\exp j(\omega_{LO_3}t) \equiv C(t)$. From the many modulation products generated in this mixer, two are selected by filtering. The first is

* While the repeater is described in terms of the commonly used frequency modulation, the basic scheme can be used with any modulation technique.

$$\begin{aligned} AC/B &= \exp j[\omega_{LO_3}t + \omega_{LO_1}t - \omega_{P(\mathrm{W})}t - \theta_{i,P} - \omega_{LO_1}t + \omega_{M(\mathrm{W\text{-}E})}t \\ &\qquad + \varphi_{(\mathrm{W\text{-}E})}(t) + \theta_{i,M}] \qquad (7) \\ &= \exp j[\omega_{LO_3} - \omega_{P(\mathrm{W})} + \omega_{M(\mathrm{W\text{-}E})})t + \varphi_{(\mathrm{W\text{-}E})}(t) + (\theta_{i,M} - \theta_{i,P})]. \end{aligned}$$

This is a carrier angle-modulated by $\varphi_{(\mathrm{W\text{-}E})}(t)$ and having the residual relative phase angle $(\theta_{i,M} - \theta_{i,P})$. If the pilot frequency is chosen nearly equal to the modulation frequency, $(\theta_{i,M} - \theta_{i,P})$* will be very small, and the W-E modulation received by the ith antenna will be in phase with that received by all of the other antennas. This completes the receiving functions; the remaining problem is to derive steering information for the outgoing beam. To this end we select the modulation product

$$AC = \exp j[(\omega_{LO_3} + \omega_{LO_1} - \omega_{P(\mathrm{W})})t - \theta_{i,P}] \qquad (8)$$

where the symbols are defined above. Now the relative phase of this wave, $-\theta_{i,P}$, is just right for retransmission toward the west terminal near the frequency $\omega_{P(\mathrm{W})}$, using the same antenna element as for receiving. By mixing this steering signal with the modulation $\varphi_{(\mathrm{E\text{-}W})}(t)$ derived in a similar fashion from the right-half of the repeater, together with a microwave local oscillator LO_5, the retransmission function is complete.

The optional interconnection among amplifying elements, shown by dashed lines on Fig. 7, provides for the situation encountered with low-altitude satellites serving widely separated earth stations. In this case, the body of the satellite "shadows" some of the elements, and the part of the satellite surface seen by both earth stations is a small part of the total. Since it requires a signal from each earth terminal to generate the retransmitted signal, without interconnection only a small number of elements are effective. With interconnection, all satellite elements visible from the transmitting earth station will receive the modulated signal; the contributions from the various elements will then be added in phase at IF and the sum impressed on the outgoing carrier. All satellite elements visible from a receiving earth station will then emit information-bearing waves which will add in-phase in the direction of this earth station receiver. Also, it should be noted that since not all branches receive the same signal levels, some weighting[10] of voltage levels must be accomplished before combining the outputs or there will be a loss in signal-to-noise ratio. This may be accomplished by operation of mixers in a strictly square law region or by auxiliary means beyond the scope of this paper.

* The magnitude of this residual is also affected by the size of the satellite.

The satellite repeater can have an arbitrarily large antenna array gain for transmitting or receiving or both, regardless of satellite altitude or orientation and independent of earth terminal separation. This is not true for any other scheme of which the authors are aware.

There is another factor which should be mentioned at this point. It is important that the pilot frequency be filtered from the surrounding noise or modulation and desirable that it be enhanced to a level well above that of the modulated signal before the second mixer shown in Fig. 6. This is to keep the pilot from bringing noise into the final modulation band and to assure that the desired products predominate over higher-order products.

V. ARRAY SCALING AND FREQUENCY MULTIPLICATION

In the foregoing scheme, it is phase rather than time delay that is compensated by the frequency conversion operation. If the application permits the receive and transmit frequencies to be nearly the same, as was assumed in discussing Fig. 2, a single array of antennas can be used for both transmitting and receiving. However, if a rather large change in frequency is required between repeater input and output, as is frequently the case in both terrestrial and space systems, additional phase compensation must be provided. One possibility is to use a different array for transmitting than for receiving, the arrays to be similar but scaled in proportion to the wavelength. Alternatively, one may compensate by using a step of frequency (and phase) multiplication, as we shall see. Let us consider the four waves associated with reception of information from one distant terminal and retransmitted to the other. The total phase shift ψ_i of a signal wave in passing from a reference wave front (Fig. 8a), through the ith branch of the circuit and to a reference plane perpendicular to the path to the other distant terminal is:

$$\psi_i = \underset{\text{(receiving)}}{2\pi\left[\frac{S_i}{\lambda_R} - \frac{S_i}{\lambda_{P_1}}\right]} + \underset{\text{(transmitting)}}{2\pi\left[-\frac{d_i}{\lambda_{P_2}} + \frac{d_i}{\lambda_T}\right]} \tag{9}$$

where

$2\pi S_i/\lambda_R$ = phase shift of the received modulated signal caused by delay between a reference plane wave front and the ith receiving element,

$2\pi S_i/\lambda_{P_1}$ = phase shift of the pilot associated with the above signal and path,

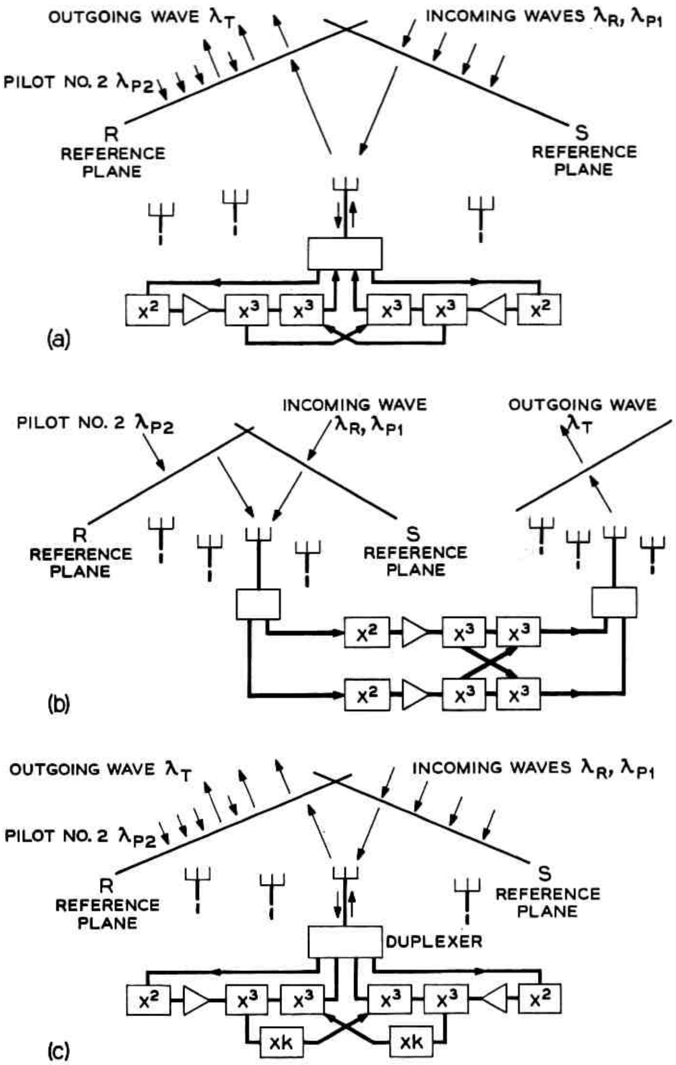


Fig. 8 — Methods of compensating for frequency change in the repeater: (a) reference diagram, (b) array scaling, (c) frequency (and phase) multiplication.

$2\pi d_i/\lambda_{P_2}$ = phase shift of the received pilot signal from the outgoing path, caused by delay between a reference plane wave front of this wave and the ith antenna element, and

$2\pi d_i/\lambda_T$ = phase shift of the retransmitted signal between the associated reference plane and element.

There are a number of possible ways to make $\psi_i = 0$. Let us suppose that all of the waves received are near the same frequency; then

$$\lambda_R \approx \lambda_{P_1} \approx \lambda_{P_2}$$

and the first term in (9) is very small. The transmitter in many applications will be considerably removed from the receiving band, in which case $\lambda_{P_2} \not\approx \lambda_T$. To make $\psi_i \approx 0$, we may scale the transmitting array as in Fig. 8(b), in proportion to the wavelength. Then, letting the primes indicate the scaled dimensions.

$$\frac{d_i'}{\lambda_T} = \frac{d_i}{\lambda_{P_2}}$$

and

$$\psi_i \approx 0.$$

It may not be convenient to scale the array. As an alternative, we can operate on the pilot signal before mixing. Suppose that after the second step of frequency conversion we pass the pilot signal through a frequency multiplier, as shown in Fig. 8(c). This multiplies the pilot frequency term, including phase, by a factor k. The signal out of the multiplier is of the form [from (8)]

$$\exp j[k(\omega_{LO_3} + \omega_{LO_1} - \ )t - \omega_{P(\mathrm{W})}\, k\theta_{i,P} + 2\pi nk]. \tag{10}$$

Now, adding the phase contributions through the repeater we get

$$\psi_i - 2\pi nk = \underset{\text{(receiving)}}{2\pi\left[\frac{S_i}{\lambda_R} - \frac{S_i}{\lambda_{P_i}}\right]} + \underset{\text{(transmitting)}}{2\pi\left[-k\,\frac{d_i}{\lambda_{P_2}} + \frac{d_1}{\lambda_T}\right]} \tag{11}$$

and k can be chosen to make $\psi_i = 0$. In general, however, k will not be an integral and there is a phase ambiguity. A possible way of removing this ambiguity is to lightly couple the frequency multipliers in adjacent channels, so that they prefer to be nearly in phase, and limit the array design so that adjacent elements are not more than $\lambda/2$ apart in the direction of transmission* at both the transmitting and receiving frequencies.

Of course, combinations of array scaling, phase shift multiplication and a judicious choice of pilot, transmitting, mixing and receiving frequencies will be important and interrelated parts of the design of a practical system.

* Elements need not be physically less than $\lambda/2$ apart, but the distance along the direction of propagation should not differ by more than $\lambda/2$ when the line of sight is within the beamwidth of the element.

VI. ARRAY GAIN

Up to this point no limits have been placed on the form of the antenna array. The elements do not have to be arranged with any particular form or symmetry. However, the presence of the satellite itself, in the case of satellite repeaters, and mutual coupling between array elements and bandwidth considerations in either space or terrestrial systems, provide some limits to the form of the array. Since the phase between elements varies, strong coupling between elements would have serious consequences in impedance mismatch. However, element directivity and separation can be used to reduce coupling and also to reduce the shadowing of elements one by another. If array elements are mounted on a conducting surface, an element gain of two is inherent in that the element can only radiate into a hemisphere. A gain of three to five is more practical, can be obtained from small elements, and results in relatively small coupling between elements.

In the case of a satellite without orientation control, elements must point in all directions; but an element having a gain (g) can illuminate only $1/g$th of the total solid angle. Thus, if N elements are distributed more or less uniformly over the surface of a spherical satellite, only $(1/g)N$ will contribute to the received (or transmitted) signal. Also, of the total power radiated, only a fraction $(1/g)$ is delivered to the array elements forming the beam. The remainder is not utilized. Net effective array gain for transmission is the product of element gain and the number of elements effective and the fraction of the total power which is useful, i.e.,

$$G = g\,(N/g)\,(1/g) = (N/g). \tag{12}$$

Thus the net array gain for transmission is equal to the number of elements in the array divided by the element gain. Evidently one should use little element gain on nonoriented satellite repeaters. In the case of terrestrial repeaters and oriented satellites:

$$G = Ng \tag{13}$$

and element gain is limited by more customary factors. In other circumstances which we will not elaborate, $G = N$.

One would like to get equivalent performance in all directions from an unstabilized satellite. This can be accomplished with the present scheme by covering the outside of a sphere or polyhedron with small radiators. It is best that the radiators be close together to reduce side lobes and possible interference.

If each element is assigned one square wavelength, the satellite diameter must be

$$D \geqq \lambda\sqrt{N/\pi}. \tag{14}$$

Alternatively, the elements can be grouped in a ring on a great circle around the satellite, each element having a fan beam with a maximum in a radial direction and radiating with a gain of $(1/g)$ in a direction 90° from the plane of the array. All of the elements of such an array would contribute in the polar direction with an array gain of (N/g). In the equatorial plane, even though only a fraction of the elements contribute, the gain is also (N/g) by the argument used in deriving (14). In intermediate directions, the gain depends upon the detailed characteristics of the elements, but it should be possible to keep it very near (N/g) in all directions.

One should not confuse the round-trip performance of the array with the radiation pattern obtained with a fixed excitation. The former can be truly isotropic but the latter cannot be, and may be a multilobe affair. If the elements are in a ring, as described above, the re-radiation will in general have two large lobes, one above and one below the plane of the array; and if the elements are widely spaced, some minor lobes may be as large as the major. This is of little consequence to the transmission performance of a satellite system, however, because the phasing of the elements automatically assures that a maximum is always directed toward the appropriate earth terminal. The shape of the pattern depends drastically upon the distribution of elements, but to a first order the strength of the major lobe does not. In any case, spacing between array elements and the element gain should be chosen to minimize side lobes in order to lessen the likelihood of interference.

VII. NUMBER OF ARRAY ELEMENTS

How many array elements is it practical to consider for a repeater of the type proposed herein? As has been shown, the directivity gain for a nonoriented repeater for reception and transmission is equal to the number of elements used divided by the element gain, thus the effective radiated power (ERP) or the power which an isotropic source would have to radiate to produce the same received signal is

$$ERP = (N/g)P_R \tag{15}$$

where N = number of elements in array,
g = gain of individual elements,

P_R = total power radiated = NP_2, and
P_2 = power radiated per element.

The power P_2 which must be radiated by each element to provide a given ERP is, from (15)

$$P_2 = P_R/N = ERP(g/N^2). \qquad (16)$$

Thus there appears to be an advantage in using a large number of elements. However, there will be a component of the dc input power used for local oscillators, low-level amplifiers, etc., which is directly proportional to the number of elements and is nearly independent of the RF power output per element. Hopefully, this can be made small through development of suitable solid-state devices. For the minimum dc power consumption consistent with a given repeater performance, there is evidently an optimum number of array elements. If the low-level mixing and amplifying operations can be accomplished with a power consumption p_1 watts per element, and if a high-frequency output power P_2 watts per element can be obtained with a power amplifier efficiency of η, then the dc input power required to radiate a beam having a stated ERP is

$$P_{\mathrm{dc}} = p_1N + (P_2/\eta)N = p_1N + (g/N)ERP/\eta \qquad (17)$$

where the symbols are as defined above. This has a minimum when

$$\frac{\partial}{\partial N}[p_1N + (g/N)ERP/\eta] = 0 \qquad (18)$$

$$p_1 - (g/N^2)ERP/\eta = 0 \qquad (18a)$$

or

$$N = \sqrt{(g/p_1)ERP/\eta}. \qquad (19)$$

We note that from (18a) and (16)

$$p_1 = g/N^2(ERP/\eta) = P_2/\eta \qquad (20)$$

which says that the minimum dc power will be required when the number of elements is chosen to make the power supplied to the output amplifiers equal to that consumed by the low-level devices.

If, in the case of a nonoriented satellite repeater, it is possible to miniaturize the circuitry enough so that the power supply is the principal source of weight, then this is a real optimum. Otherwise, it represents a sort of design objective, and minimizing satellite weight and complexity will require a smaller number of elements. In any case, it is clear that the practicality of the scheme depends heavily upon the degree of miniaturization and power efficiency which can be achieved with solid-state devices and circuits.

VIII. EXPERIMENTAL VERIFICATION

The basic principle upon which the self-steering array depends is the coherent in-phase addition of randomly phased inputs. The principle has been demonstrated by the simple laboratory experiment shown in Fig. 9. The outputs of two oscillators at 6200 and 6034 mc, which may be thought of as representing the received modulation and pilot signals, are combined and fed together to two receiving circuits, and an adjustable phase shifter or line stretcher is inserted in one branch. Each receiving circuit separates the incoming frequencies, heterodynes the 6200 mc with a 6274-mc local oscillator which is common to the two receiving circuits, to produce a 74-mc intermediate frequency, which in turn is amplified and recombined with the 6034-mc signal in a second mixer to produce a 6108-mc output. Now, it is asserted that the phase of the

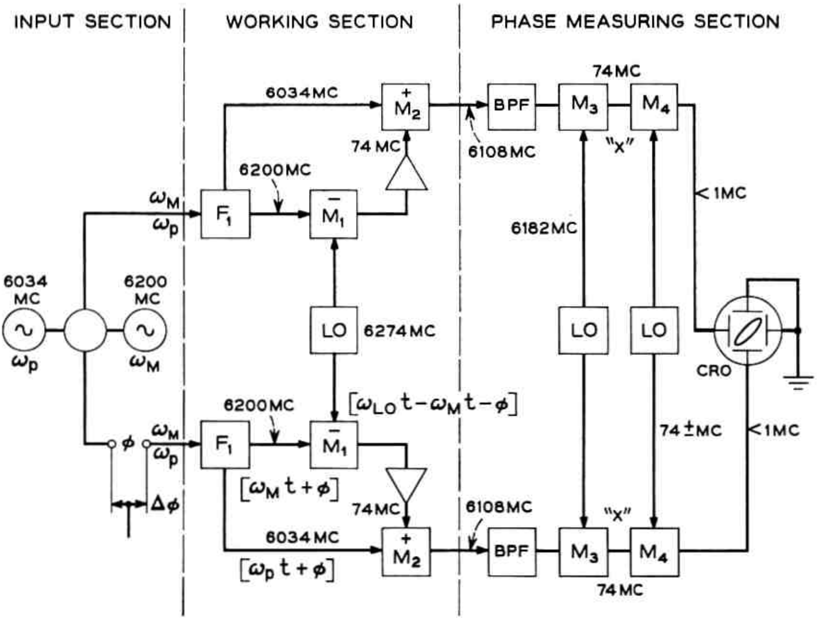


Fig. 9 — Experimental arrangement for testing the principle of phase compensation.

6108-mc wave should be to first order independent of the phase of the combined signals at the input to the branching filter. To check this, the signals from the two branches are compared in the circuit shown to the right of Fig. 9. If the idea is sound, the phase of the output should change very little with movement of the piston in the input section.

The degree of phase change correction or cancellation was observed visually by noting the stability of a Lissajous figure formed by the output sine waves as received over the two branch paths of the test circuit, and was measured using a phase meter. Some change in the phase between the two output branch signals was to be expected because of the difference in frequency between the pilot and the signal. Fig. 10 shows the measured and calculated phase change in the output produced by large changes of phase in the lower branch of the circuit. It will be noticed that the resultant measured output phase change, $\Delta\theta$, is not exactly a linear function of changes in input phase. These variations from linearity were shown to be caused by mismatches in impedance and leakage between various parts of the circuit and were partially corrected with isolators. Calculation of the ratio of wavelengths for WR159 rectangular waveguide for the two frequencies 6200 and 6034 mc gives a value of 1.04 or a difference of 15° for each wavelength, which is in very good agreement with the average measured value. The difference frequency was later reduced to 40 mc, and the output phase variation was reduced proportionally.

Most of the tests described were made under a condition of large signal-to-noise ratio and low gain. The phase correction was found to be

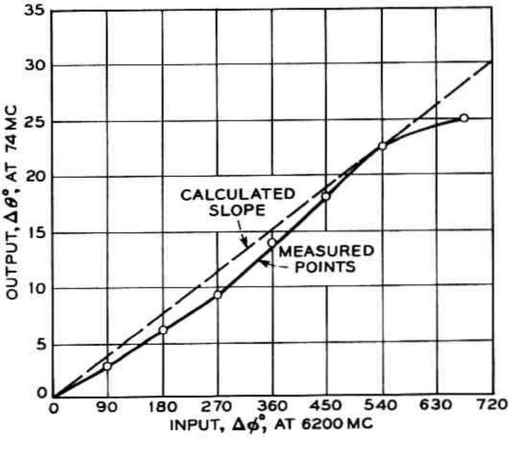


Fig. 10 — Comparison of calculated and measured phase compensation.

very stable with change in time, level and frequency. In order to determine how the scheme would operate for low values of signal-to-noise and with over-all gain typical of an actual repeater, attenuation was added in the lower branch of the circuit to reduce the level of ω_M. This loss was then compensated by adding about 100 db of IF amplification to bring the level back to its former value. When the signal-to-noise ratio was measured at 2 db, the pattern on the scope was much more ragged due to the large random noise present, but it was still stable and indicated the desired cancellation of phase.

IX. CONCLUSIONS

An antenna beam-steering scheme using a pilot tone and phase inversion makes possible large antenna gain even for nonoriented satellite repeaters or movable terrestrial repeaters. A large number of low-power, elemental repeater amplifiers with inputs and outputs connected to like elements in similar arrays, or diplexed onto common elements in a single array, are used. The scheme is particularly suited for use with solid-state devices since the (low) power output of many units is effectively added in-phase. Reliability is provided by the many parallel paths through the repeater; failure of individual units will only slightly degrade performance. Although the radiation is not isotropic, the radiation or sensitivity toward distant terminals can be independent of array orientation, and thus the idea is well suited for use with unoriented satellites. A simple experiment has been performed to demonstrate the basic steering property of the phase inversion scheme.

X. ACKNOWLEDGMENTS

Suggestions by W. C. Jakes, E. A. Ohm and J. C. Schelling were important to the development of these concepts; J. P. Schafer and H. E. Keller carried out the experimental work which verified the soundness of the basic phase-cancellation scheme.

REFERENCES

1. Van Atta, L. C., Electromagnetic Reflector, U. S. Patent No. 2,908,002, October 6, 1959.
2. Sharp, E. D., and Diab, M. A., Van Atta Reflector Array, I.R.E. Trans. on Antennas and Propagation, **AP-8,** July, 1960, pp. 436–438.
3. Hansen, R. C., Communication Satellites Using Arrays, Proc. I.R.E., **49,** June, 1961, pp. 1066–1074.
4. Rutz, E. M., and Kramer, E., Modulated Array for Space Communications, NEREM Record, **4,** November, 1962, pp. 16–17.
5. Davies, D. E. N., Some Properties of Van Atta Arrays and the Use of Two-Way Amplification in the Delay Paths, Proc. IEE (London), **110,** March, 1963, pp. 507–512.

6. Miller, Barry, Self-Focusing Antenna Arrays Developed (News Item), Aviation Week and Space Technology, Aug. 21, 1961, pp. 54–59.
7. Adams, R. T., and Mindes, D. N., Evaluation of Intermediate-Frequency and Baseband Diversity Combining Receivers, I.R.E. Trans. on Communication Systems, **6,** June, 1958, pp. 8–13.
8. Bello, P., and Nelin, B., Predetection Diversity Combining with Selectivity Fading Channels I.R.E. Trans. on Communication Systems, **9,** March, 1962, pp. 32–42.
9. Phase inversion by this means was known and used by Friis in the 1930's; see Friis, H. T., and Feldman, C. B., A Multiple Unit Steerable Antenna for Short-Wave Reception, B.S.T.J., **16,** July, 1937, pp. 337–419, particularly Fig. 28 and accompanying legend.
10. Kahn, L. R., Ratio Squarer (Letter) Proc. I.R.E., **42,** November, 1954, p. 1704.

On the Properties of Some Systems that Distort Signals—I

By I. W. SANDBERG

(Manuscript received March 20, 1963)

This is the first part of a two-part paper concerned with some generalizations and extensions of the Beurling-Landau-Miranker-Zames theory of recovery of distorted bandlimited signals. We present a uniqueness proof that extends Beurling's result and study a class of functional mappings defined on Hilbert space. As an application, we show that the recovery results can be extended to cases in which a known square-integrable corrupting signal is added to the input signal and the result applied to a time-variable device which may be nonlinear. It is proved that an assumption made by the earlier writers is in fact necessary in order that stable recovery be possible. Part II will consider the more complicated situation in which a single time-variable nonlinear element is imbedded in a general linear system.

I. INTRODUCTION

A signal transmission system is a realization of an operator that maps input signals in one domain into output signals in a second domain. When the system contains energy-storage devices as well as time-variable or nonlinear elements, the mapping is usually quite complicated. Very little in the way of a general theory is known concerning the mathematical properties of such mappings.

Of course one of the important properties of a mapping is its invertability or lack of invertability. Some particularly interesting results relating to the existence of the inverse of a special mapping have been obtained by Beurling, Landau, Miranker, and Zames. They consider the situation in which a square-integrable bandlimited signal is passed through a monotonic nonlinear device. Beurling showed, by means of a nonconstructive proof,† that a knowledge of the Fourier transform of the distorted signal on the interval where the transform of the input signal does not vanish is sufficient to uniquely determine the input

† Beurling's proof is given in Refs. 1 and 3.

signal. Landau and Miranker[1] have considered a stable iteration scheme for obtaining the input signal from the bandlimited version of the distorted signal. They assume that the distortion characteristic possesses a derivative bounded above and below by positive constants. A solution of this type was found independently by G. D. Zames.[2] Some material associated with the stability of the iteration scheme and an impressive recovery experiment are discussed by Landau.[3]

This paper is concerned with some generalizations and extensions of the results mentioned above. Our primary objective is to show that the results in Refs. 1 and 2 are special cases of a quite general theory.

Section II considers some mathematical preliminaries. In Section III we discuss the solution of a class of functional equations defined on an arbitrary Hilbert space, and give a uniqueness proof that extends Beurling's result. In the next section two general signal-theoretic applications of the results in Section III are discussed. Theorem IV implies, among other things, that the recovery theory of the earlier writers can be extended to cases in which a known square-integrable corrupting signal is added to the bandlimited input signal and the result applied to a time-variable device which may be nonlinear. Section V concludes Part I with some specialized results that contribute to a deeper understanding of the character of the previous material. In particular it is proved that an assumption made by the earlier writers is in fact necessary in order that stable recovery be possible.

Part II will consider the more complicated situation in which a single time-variable nonlinear element is imbedded in a general linear system. We treat a recovery problem of the type considered by the earlier writers and prove that recovery is possible under quite general conditions. This study may have applications in improving the quality of distorted data obtained, for example, from a malfunctioning transmitter in a space satellite.

II. PRELIMINARIES

Let $\mathcal{R} = [\Theta,\rho]$ be an arbitrary metric space.[4] A mapping $\mathbf{A}$ of the space $\mathcal{R}$ into itself is said to be a contraction if there exists a number $\alpha < 1$ such that

$$\rho(\mathbf{A}x,\mathbf{A}y) \leqq \alpha\rho(x,y)$$

for any two elements $x,y \ \varepsilon\ \Theta$. The contraction-mapping fixed-point theorem[4] is basic to much of the subsequent discussion. It states that every contraction-mapping defined in a complete metric space $\mathcal{R}$ has one and only one fixed point (i.e., there exists a unique element $z \ \varepsilon\ \Theta$

such that $\mathbf{A}z = z$). Furthermore $z = \lim_{n\to\infty} \mathbf{A}^n x_0$, where x_0 is an arbitrary element of Θ.

Throughout the discussion $\mathcal{H}$ denotes a real or complex Hilbert space. If $f,g \; \varepsilon \; \mathcal{H}$, then (f,g), $\| f \| = (f,f)^{\frac{1}{2}}$, and $\| f - g \|$, respectively, denote the inner product of f with g, the norm of f, and the distance between f and g. It is not assumed that $\mathcal{H}$ is separable or that it is of infinite dimension.

The space of complex-valued square-integrable functions with inner product

$$(f,g) = \int_{-\infty}^{\infty} f\underset{\sim}{g} \, dt,$$

where $\underset{\sim}{g}$ is the complex conjugate of g, is denoted by $\mathcal{L}_2$, and $\mathcal{L}_{2\mathrm{R}}$ denotes the intersection of the space $\mathcal{L}_2$ with the set of real-valued functions.

We take as the definition of the Fourier transform of $f(t) \; \varepsilon \; \mathcal{L}_2$:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t} \, dt,$$

and consequently

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)e^{i\omega t} \, d\omega.$$

With this definition, the Plancherel identity reads:

$$2\pi \int_{-\infty}^{\infty} f(t)\underset{\sim}{g}(t)dt = \int_{-\infty}^{\infty} F(\omega)\underset{\sim}{G}(\omega) \, d\omega.$$

Except when indicated otherwise, a function and its Fourier transform are denoted, respectively, by lower and upper case versions of the same symbol.

The symbol $\mathcal{K}$ denotes an arbitrary subspace of $\mathcal{H}$. Hence $\mathcal{H} = \mathcal{K} \dotplus \mathcal{K}'$, the direct sum of $\mathcal{K}$ and $\mathcal{K}'$, where $\mathcal{K}'$ is the orthogonal complement of $\mathcal{K}$ with respect to $\mathcal{H}$. The operator that projects an arbitrary element of $\mathcal{H}$ onto $\mathcal{K}$ is denoted by $\mathbf{P}$. The subspaces of $\mathcal{L}_{2\mathrm{R}}$ of principal interest to us are†

$$\mathcal{B}(\Omega) = \{f(t) \mid f(t) \; \varepsilon \; \mathcal{L}_{2\mathrm{R}} \; ; \qquad F(\omega) = 0, \, \omega \not\varepsilon \, \Omega\}$$

and

$$\mathcal{D}(\Sigma) = \{f(t) \mid f(t) \; \varepsilon \; \mathcal{L}_{2\mathrm{R}} \; ; \qquad f(t) \; = 0, \, t \not\varepsilon \, \Sigma\},$$

† It is a simple matter to verify that the linear manifold $\mathcal{B}(\Omega)$ is in fact a subspace. An obvious modification of the proof in Ref. 1 for the case in which Ω is a single interval suffices.

where Ω and Σ are each the union of disjoint intervals. It is hardly necessary to mention that the class of electrical signals belonging to $\mathcal{B}(\Omega)$ or $\mathcal{D}(\Sigma)$ is of considerable importance in the theory of electrical communication systems.

We shall use the fact that any projection operator defined on a Hilbert space is self adjoint [i.e., that $(f,\mathbf{P}g) = (\mathbf{P}f,g)$ for any $f,g \; \varepsilon \; \mathcal{H}$].

The symbol **I** is used throughout to denote the identity transformation.

III. INVERSION OF A CLASS OF OPERATORS DEFINED ON AN ARBITRARY HILBERT SPACE

As we have said earlier, a signal transmission system is a realization of an operator that maps input signals in one domain into output signals in a second domain. The following theorem relates to the existence of the inverse of a particularly relevant type of nonlinear mapping defined on an arbitrary Hilbert space.

Theorem I: Let **Q** *be a mapping of* $\mathcal{K}$ *into* $\mathcal{H}$ *such that for all* $f,g \; \varepsilon \; \mathcal{K}$*:*

$$Re(\mathbf{Q}f - \mathbf{Q}g, f - g) \geqq k_1 \| f - g \|^2$$
$$\| \mathbf{PQ}f - \mathbf{PQ}g \|^2 \leqq k_2 \| f - g \|^2$$

where k_1 *and* k_2 *are positive constants. Then for each* $h \; \varepsilon \; \mathcal{K}$*, the equation* $h = \mathbf{PQ}f$ *possesses a unique solution* $(\mathbf{PQ})^{-1}h \; \varepsilon \; \mathcal{K}$ *given by* $(\mathbf{PQ})^{-1}h = \lim_{n\to\infty} f_n$ *where*

$$f_{n+1} = \frac{k_1}{k_2}(h - \mathbf{PQ}f_n) + f_n$$

and f_0 *is an arbitrary element of* $\mathcal{K}$*. Furthermore, for all* $h_1, h_2 \; \varepsilon \; \mathcal{K}$

$$\| (\mathbf{PQ})^{-1}h_1 - (\mathbf{PQ})^{-1}h_2 \| \leqq \frac{1}{k_1} \| h_1 - h_2 \|.$$

Proof:

Let $\mathbf{A} = \mathbf{PQ}$ and note first that

$$\begin{aligned}\mathrm{Re}(\mathbf{A}f - \mathbf{A}g, f - g) &= \mathrm{Re}(\mathbf{Q}f - \mathbf{Q}g, \mathbf{P}f - \mathbf{P}g) \\ &= \mathrm{Re}(\mathbf{Q}f - \mathbf{Q}g, f - g) \geqq k_1 \| f - g \|^2\end{aligned}$$

for all $f,g \; \varepsilon \; \mathcal{K}$ since **P** is a self-adjoint transformation.

The equation $h = \mathbf{A}f$ is equivalent to $f = \tilde{\mathbf{A}}f$, where $\tilde{\mathbf{A}}f = ch + f - c\mathbf{A}f$ and c is any nonzero constant. The following calculation shows that $\tilde{\mathbf{A}}$, a mapping of $\mathcal{K}$ into $\mathcal{K}$, is a contraction when $c = k_1(k_2)^{-1}$:

$$\begin{aligned}\|\tilde{\mathbf{A}}f - \tilde{\mathbf{A}}g\|^2 &= \|f - g - c\mathbf{A}f + c\mathbf{A}g\|^2 \\ &= \|f-g\|^2 - 2c\,\mathrm{Re}(\mathbf{A}f - \mathbf{A}g, f - g) + c^2\|\mathbf{A}f - \mathbf{A}g\|^2 \\ &\leqq (1 - 2ck_1 + c^2k_2)\|f-g\|^2, \qquad c > 0.\end{aligned}$$

Since $(1 - 2ck_1 + c^2k_2) \geqq 0$ for all $c > 0$, it follows that† $k_1^2 \leqq k_2$. Hence

$$\|\tilde{\mathbf{A}}f - \tilde{\mathbf{A}}g\|^2 \leqq \left(1 - \frac{k_1^2}{k_2}\right)\|f-g\|^2, \qquad 0 \leqq \left(1 - \frac{k_1^2}{k_2}\right) < 1.$$

The last inequality stated in the theorem follows from an application of the Schwarz inequality. For all $f,g \,\varepsilon\, \mathcal{K}$

$$\|\mathbf{A}f - \mathbf{A}g\| \cdot \|f-g\| \geqq |(\mathbf{A}f - \mathbf{A}g, f - g)| \geqq k_1\|f-g\|^2.$$

Thus

$$\|\mathbf{A}f - \mathbf{A}g\| \geqq k_1\|f-g\|.$$

In particular, with $f = \mathbf{A}^{-1}h_1$ and $g = \mathbf{A}^{-1}h_2$,

$$\|h_1 - h_2\| \geqq k_1\|\mathbf{A}^{-1}h_1 - \mathbf{A}^{-1}h_2\|.$$

3.1 *Uniqueness Theorem*

We show here that the uniqueness property of solutions to equations of the type considered in Theorem I is implied by much weaker hypotheses than those stated in the theorem.

Theorem II: Let $f,g \,\varepsilon\, \mathcal{K}$ and let $\mathbf{Q}$ be a mapping of $\mathcal{K}$ into $\mathcal{H}$ such that $(\mathbf{Q}f - \mathbf{Q}g, f - g)$ vanishes only if $f = g$. Then if the equation $h = \mathbf{PQ}z$ has a solution $z \,\varepsilon\, \mathcal{K}$, it is unique.

Proof:

Assume that $\mathbf{PQ}z_1 = \mathbf{PQ}z_2$ where $z_1, z_2 \,\varepsilon\, \mathcal{K}$. Since $\mathbf{P}$ is self-adjoint,

$$\begin{aligned}(\mathbf{Q}z_1 - \mathbf{Q}z_2, z_1 - z_2) &= (\mathbf{Q}z_1 - \mathbf{Q}z_2, \mathbf{P}z_1 - \mathbf{P}z_2) \\ &= (\mathbf{PQ}z_1 - \mathbf{PQ}z_2, z_1 - z_2) \\ &= 0.\end{aligned}$$

Hence $z_1 = z_2$.

Theorem II is a generalization of the uniqueness theorem due to A. Beurling.[1,3]

† Alternatively, the hypotheses and an application of the Schwarz inequality yields:

$$k_1^2\|f-g\|^4 \leqq |(\mathbf{A}f - \mathbf{A}g, f - g)|^2 \leqq \|\mathbf{A}f - \mathbf{A}g\|^2 \cdot \|f-g\|^2 \leqq k_2\|f-g\|^4.$$

IV. APPLICATIONS

We present two theorems that have specific signal-theoretic interpretations.

Theorem III: Let $\mathcal{H} = \mathfrak{L}_2$ *and let*

$$\mathbf{L}f = \int_{-\infty}^{\infty} l(t - \tau)\, f(\tau)\, d\tau$$

where $l(t) \,\varepsilon\, \mathfrak{L}_2$ *and* $f \,\varepsilon\, \mathcal{K}$. *Suppose that*

$$\sup_{\omega} |\, L(\omega)\,| < \infty, \qquad Re\; L(\omega) \geqq -\alpha$$

where $\alpha < 1$. *Then for any* $h \,\varepsilon\, \mathcal{K}$, $h = f + \mathbf{PL}f$ *has a unique solution* $f \,\varepsilon\, \mathcal{K}$. *Suppose alternatively that*

$$\sup_{\omega} |\, L(\omega)\,| < \infty, \qquad Re\; L(\omega) > 0 \text{ a.e.}$$

Then **PL** *is a mapping of* $\mathcal{K}$ *into itself such that the equation*

$$h = \mathbf{PL}f, \quad h \,\varepsilon\, \mathcal{K}$$

possesses at most one solution $f \,\varepsilon\, \mathcal{K}$.

Proof:

Let $\mathbf{Q} = \mathbf{I} + \mathbf{L}$ and let $z \,\varepsilon\, \mathcal{K}$. Using the Plancherel identity

$$\begin{aligned}\mathrm{Re}(\mathbf{Q}z,z) &= \| z \|^2 + \mathrm{Re}(\mathbf{L}z,z)\\ &= \| z \|^2 + \frac{1}{2\pi}\,\mathrm{Re} \int_{-\infty}^{\infty} L(\omega)\,|\,Z(\omega)\,|^2\, d\omega\\ &\geqq (1 - \alpha)\, \| z \|^2 .\end{aligned}$$

Also,

$$\begin{aligned}\| \mathbf{PQ}z \|^2 &\leqq \| z \|^2 + 2\,\mathrm{Re}(\mathbf{L}z,z) + \| \mathbf{L}z \|^2\\ &\leqq (1 + 2\delta + \delta^2)\, \| z \|^2\end{aligned}$$

where $\delta = \sup_{\omega} |\, L(\omega)\,|$. Hence the hypotheses of Theorem I are satisfied. This establishes the first part of Theorem III. The second part is a direct application of Theorem II since,† in view of the Plancherel identity, it is clear that here $\mathrm{Re}(\mathbf{L}z,z)$ vanishes only if $z = 0$.

If $\mathcal{K} = \mathfrak{D}(\Sigma)$, Theorem III implies that under either of the stated conditions only a knowledge of the output for $t \,\varepsilon\, \Sigma$ of a known linear filter is necessary to completely determine the input to the filter, if it is known that the input vanished for $t \not\varepsilon\, \Sigma$. In addition, if $h(t)$ is any ele-

† The boundedness of $|\, L(\omega)\,|$ is required in order that $\mathbf{L}f \,\varepsilon\, \mathfrak{L}_2$ whenever $f \,\varepsilon\, \mathcal{K}$.

ment of $\mathfrak{D}(\Sigma)$, there exists in the first case a unique input signal in $\mathfrak{D}(\Sigma)$ such that the projection of the output signal is $h(t)$, and this input signal, which can be computed in accordance with Theorem I, depends continuously on $h(t)$. Some related results are discussed in the Appendix.

Definition I: It is assumed throughout that $\varphi(x) = \varphi(x,t)$ *is a real-valued function of the real variables* x *and* t.

Theorem IV: Let $\mathfrak{H} = \mathfrak{K} \dotplus \mathfrak{K}'$ *be a real Hilbert space in which* $| f(t) | \geqq | g(t) |$ *for all* t *implies that* $\| f \| \geqq \| g \|$ *whenever* $f,g \,\varepsilon\, \mathfrak{H}$. *Let* $\varphi(x,t)$ *satisfy*

$$m(x - y) \leqq \varphi(x,t) - \varphi(y,t) \leqq M(x - y) \text{ when } x \geqq y$$

where m *and* M *are positive constants. Let* $\varphi[f] \,\varepsilon\, \mathfrak{H}$, $f \,\varepsilon\, \mathfrak{H}$. *Then for any* $u(t) \,\varepsilon\, \mathfrak{K}$, $v(t) \,\varepsilon\, \mathfrak{K}'$, *there exists a unique* $w(t) \,\varepsilon\, \mathfrak{K}$ *such that*

$$\mathbf{P}\varphi[w(t) + v(t)] = u(t).$$

In fact, $w(t) = \lim_{n\to\infty} w_n$ *where*

$$w_{n+1} = \frac{m}{M^2}\{u - \mathbf{P}\varphi[v + w_n]\} + w_n$$

and w_0 *is an arbitrary element of* $\mathfrak{K}$. *In addition,*

$$\| \mathbf{P}\varphi[v + f] - \mathbf{P}\varphi[v + g] \| \geqq m \| f - g \| ; \qquad f,g \,\varepsilon\, \mathfrak{K},\ v \,\varepsilon\, \mathfrak{K}'$$

and if $\mathbf{P}\varphi[v_a + w_a] = u_a$, $\mathbf{P}\varphi[v_b + w_b] = u_b$ *where* $w_a, w_b, u_a, u_b \,\varepsilon\, \mathfrak{K}$ *and* $v_a, v_b \,\varepsilon\, \mathfrak{K}'$,

$$\| w_a - w_b \| \leqq \frac{1}{m} \| u_a - u_b \| + \frac{M}{m} \| v_a - v_b \|$$

$$\| u_a - u_b \| \leqq M \| v_a - v_b \| + M \| w_a - w_b \|.$$

Proof:

We first show that the hypotheses of Theorem I are satisfied when $\mathbf{Q}$ is defined by $\mathbf{Q}w = \varphi[w + v]$. Let $\hat{\eta} = (\eta - m)$ where

$$\frac{\varphi[v + f] - \varphi[v + g]}{f - g} = \eta; \qquad f,g \,\varepsilon\, \mathfrak{K}.$$

Observe that an application of a well known identity yields (with $z = f - g$):

$$\begin{aligned}(\eta z,z) - m(z,z) &= (\hat{\eta} z,z)\\ &= \tfrac{1}{4} \| (\hat{\eta} + 1)z \|^2 - \tfrac{1}{4} \| (\hat{\eta} - 1)z \|^2\\ &\geqq 0.\end{aligned}$$

Hence $(\varphi[v + f] - \varphi[v + g], f - g) \geqq m \| f - g \|^2$. Since, in addition,

$$\begin{aligned} \| \mathbf{P}\varphi[v + f] - \mathbf{P}\varphi[v + g] \| &\leqq \| \varphi[v + f] - \varphi[v + g] \| \\ &\leqq M \| f - g \| , \end{aligned}$$

the hypotheses are satisfied. The bound on $\| w_a - w_b \|$ is obtained from the inequality:

$$\| w_a - w_b \| \leqq \frac{1}{m} \| \mathbf{P}\varphi[w_a + v_a] - \mathbf{P}\varphi[w_b + v_a] \|.$$

Specifically, the right-hand side is equal to

$$\begin{aligned} &\frac{1}{m} \| \mathbf{P}\varphi[w_a + v_a] - \mathbf{P}\varphi[w_b + v_b] + \mathbf{P}\varphi[w_b + v_b] - \mathbf{P}\varphi[w_b + v_a] \| \\ &\qquad \leqq \frac{1}{m} \| u_a - u_b \| + \frac{1}{m} \| \mathbf{P}\varphi[w_b + v_b] - \mathbf{P}\varphi[w_b + v_a] \| \\ &\qquad \leqq \frac{1}{m} \| u_a - u_b \| + \frac{M}{m} \| v_a - v_b \|. \end{aligned}$$

With $\mathcal{H} = \mathfrak{L}_{2R}$ and $\mathcal{K} = \mathcal{B}(\Omega)$, Theorem IV implies that if a function of time $w(t)$ having frequency components which vanish outside Ω is added to a second function $v(t)$ with frequency components which vanish inside Ω, and if the result is applied to a quite general type of time-variable nonlinear amplifier in cascade with an ideal linear filter having only passbands coincident with the intervals contained in Ω, then the output is sufficient to uniquely determine the signal $w(t)$, assuming of course that $v(t)$, Ω, and the function $\varphi(x,t)$ are known. Furthermore, for each signal $v(t) \varepsilon \mathcal{K}'$, there exists a unique input $w(t) \varepsilon \mathcal{K}$ such that the output is any prescribed element of $\mathcal{K}$. In particular, $w(t)$ depends continuously on the prescribed output and $v(t)$.

If $\mathcal{H}$ is the usual space of real-valued periodic functions of t, and $\varphi(x,t)$ is similarly periodic in t, the theorem possesses a similar interpretation. Of course, all of the results are valid for the interesting special case in which $\varphi(x,t) = x\varphi(1,t)$ (i.e., when the physical operation corresponding to this function is product modulation).

The inequality: $\| \mathbf{P}\varphi[v + f] - \mathbf{P}\varphi[v + g] \| \geqq m \| f - g \|$ in the conclusion of Theorem IV is quite interesting from an engineering viewpoint. For example, let $\mathcal{H} = \mathfrak{L}_{2R}$, $\mathcal{K} = \mathcal{B}(\Omega)$, and suppose that $f \varepsilon \mathcal{B}(\Omega)$ is the input to a time-variable nonlinear amplifier with transfer characteristic $\varphi(x,t)$ which satisfies the assumptions stated and for simplicity $\varphi(0,t) = 0$. Then $\| \mathbf{P}\varphi[f] \| \geqq m \| f \|$, a *lower bound* on that part of the energy of the output signal which is associated with the frequency bands occupied by the input signal.

Remark: It can be shown that Theorem IV remains valid if the words "Hilbert space" are replaced with "Banach space" (and $\mathcal{K}$ denotes an arbitrary subspace of the Banach space with $\mathbf{P}$ the corresponding projection operator). In particular, the existence and uniqueness of the function $w(t)$ follows from an application of the contraction-mapping fixed-point theorem to the equation $w = (\mathbf{P} - c\mathbf{PQ})w + cu$ in which $\mathbf{Q}$ is defined by $\mathbf{Q}w = \varphi[w + v]$ and c is a real constant. Using the fact that $\| \mathbf{P} \| \leqq 1$, it is not difficult to show that there exists a c for which $(\mathbf{P} - c\mathbf{PQ})$ is a contraction.

V. SOME SPECIAL RESULTS

In this section we present some results that contribute to a deeper understanding of the character of the material already described. We shall be concerned throughout with the space $\mathcal{L}_2$.

In the proof of Theorem IV the hypotheses concerning $\varphi(x,t)$ is used to establish the applicability of Theorem I. The following theorem asserts that, for this purpose, the hypotheses can be relaxed somewhat if $\mathcal{H} = \mathcal{L}_{2\mathrm{R}}$ and $\mathcal{K} = \mathcal{B}$, where $\mathcal{B}$ denotes $\mathcal{B}(\Omega)$ when Ω is a single fixed finite interval centered at the origin. The orthogonal complement of $\mathcal{B}$ is denoted by $\mathcal{B}^*$.

Theorem V: Let $\mathcal{H} = \mathcal{L}_{2\mathrm{R}}$ *and* $\mathcal{K} = \mathcal{B}$. *Let* $f \varepsilon \mathcal{B}$, $v \varepsilon \mathcal{B}^*$. *The operator* $\mathbf{Q}$ *defined by* $\mathbf{Q}f = \varphi[f + v]$ *satisfies the hypotheses of Theorem I assuming that*

$$m(x - y) \leqq \varphi(x,t) - \varphi(y,t) \leqq M(x - y) \text{ when } x \geqq y$$

for all $t \varepsilon \Pi$, *where* m *and* M *are positive constants,* Π *is a subset of the real line, and*

$$\delta < \frac{m[1 - \lambda(\Pi)]}{\lambda(\Pi)}$$

in which

$$\delta = \sup_{\substack{t \varepsilon \Pi \\ x,y}} \left| \frac{\varphi(x,t) - \varphi(y,t)}{x - y} \right|$$

and

$$\lambda(\Pi) = \sup_{f \varepsilon \mathcal{B}} \frac{\int_\Pi |f|^2 \, dt}{\|f\|^2}.$$

Proof:

Clearly, $\| \mathbf{P}\varphi[f + v] - \mathbf{P}\varphi[g + v] \| \leqq \| \varphi[f + v] - \varphi[g + v] \| \leqq \max(\delta, M) \| f - g \|$.

Let Π^* be the complement of Π with respect to the real line. Observe that

$$(\varphi[f + v] - \varphi[g + v], f - g) = \int_{\Pi*} (\varphi[f + v] - \varphi[g + v])(f - g)\,dt$$

$$+ m \int_{\Pi} (f - g)^2\,dt + \int_{\Pi} (\varphi[f + v] - \varphi[g + v])(f - g)\,dt$$

$$- m \int_{\Pi} (f - g)^2\,dt \geqq m \int_{-\infty}^{\infty} (f - g)^2 - (m + \delta) \int_{\Pi} (f - g)^2\,dt$$

$$\geqq [m - (m + \delta)\lambda(\Pi)]\, \|f - g\|^2.$$

When Π is any set of finite measure, $\lambda(\Pi)$ is less than unity.†

At this point it is convenient to introduce

Definition II: An operator $\mathbf{A}$ *defined on a Banach space is said to be bounded if there exists a constant* k *such that* $\| \mathbf{A}f - \mathbf{A}g \| \leqq k \| f - g \|$ *for all* f,g *in the domain of* $\mathbf{A}$.

This definition obviously reduces to the usual one in the event that $\mathbf{A}$ is a linear operator. From the viewpoint of implementing a signal recovery scheme (i.e., of constructing a device that reverses the effect of some known operator), it is highly desirable that the inverse operator be known to be bounded, since this situation guarantees that an error in the input signal to the recovery device would produce at most a proportional error in the recovered signal, assuming that the device functions as an ideal realization of the inverse operator. We shall consider the existence of two situations in which a mapping of the type considered earlier does not possess a bounded inverse.

Theorem VI: Let $m(x - y) \leqq \varphi(x,t) - \varphi(y,t) \leqq M(x - y)$ *for* $x \geqq y$ *when* $t \,\varepsilon\, \Pi$ *and* $\varphi(x,t) = 0$ *when* $t \not\varepsilon\, \Pi$, *where* m *and* M *are positive constants and* Π *is a set of finite measure. Let* $\mathbf{A}$ *be the mapping of* $\mathcal{B}$ *into* $\mathcal{B}$ *defined by* $\mathbf{A}f = \mathbf{P}\varphi[f]$, $f \,\varepsilon\, \mathcal{B}$. *Then* $\mathbf{A}$ *does not possess a bounded inverse.*

Proof:

If $\mathbf{A}^{-1}$ existed and satisfied $\| \mathbf{A}^{-1}f - \mathbf{A}^{-1}g \| \leqq k \| f - g \|$ for all $f,g \,\varepsilon\, \mathcal{B}$, it would follow that $\| \mathbf{A}f - \mathbf{A}g \| \geqq (k)^{-1} \| f - g \|$. However, since for any $\epsilon > 0$ there exists a $z \,\varepsilon\, \mathcal{B}$ such that

† This is proved in Ref. 5 for the case in which Π is a single interval. H. J. Landau has pointed out to the writer in a private conversation that the published argument can be extended to apply to an arbitrary set of finite measure.

$$\| z \| = 1 \qquad \text{and} \qquad \int_{\Pi} z^2 \, dt < \epsilon,$$

the following calculation shows that the inequality cannot hold for any finite k:

$$\| \mathbf{P}\varphi[f] - \mathbf{P}\varphi[g] \|^2 \leqq \| \varphi[f] - \varphi[g] \|^2 = \int_{\Pi} (\varphi[f] - \varphi[g])^2 \, dt$$

$$\leqq M^2 \int_{\Pi} (f - g)^2 \, dt.$$

Recall that the mapping described in Theorem IV possesses a bounded inverse and that $\varphi(x,t)$ is assumed to satisfy the Lipschitz condition: $m(x - y) \leqq \varphi(x,t) - \varphi(y,t)$ when $x \geqq y$, where m is a positive constant. The assumption that m does not vanish is essential; the result is obviously not valid if $\varphi(x)$ vanishes throughout a neighborhood of the origin of the x-axis for all t. The following theorem focuses attention on some restrictions imposed on the derivative of $\varphi(x)$ by the requirement that the mapping possess a bounded inverse.

Theorem VII: Let $\varphi(x,t)$ be independent of t and continuously differentiable with respect to x on the interval Ξ. Let $| \varphi(x,t) - \varphi(y,t) | \leqq M \, | x - y |$ and

$$\inf_{x \varepsilon \Xi} \left| \frac{d\varphi(x)}{dx} \right| = 0.$$

Then the mapping $\mathbf{A}$, *of* $\mathcal{B}$ *into* $\mathcal{B}$, *defined by* $\mathbf{A}f = \mathbf{P}\varphi[f]$, $f \varepsilon \mathcal{B}$ *does not possess a bounded inverse.*

Proof:

As in the proof of Theorem VI it suffices to show that for any $\epsilon > 0$ there exist functions $f,g \varepsilon \mathcal{B}$ such that $\| f - g \| = 1$ and $\| \mathbf{P}\varphi[f] - \mathbf{P}\varphi[g] \| < \epsilon$. We need the following result.

Lemma I: Let τ and ϵ be positive constants and let k be a real number. Then there exists a function $g \varepsilon \mathcal{B}$ such that

$$| g(t) - k | < \epsilon, \qquad | t | < \tau.$$

The proof of the lemma is very simple. Let $\hat{g}(t) \varepsilon \mathcal{B}$ such that $\hat{g}(0) \neq 0$. Since $\hat{g}(t)$ is continuous, $| a\hat{g}(t) - k | < \epsilon$, $| t | < b\tau$ for some constants a and b where $b > 0$. If $b < 1$, set $g(t) = a\hat{g}(bt)$. This proves the lemma.

From the hypotheses there exists for any $\epsilon_1 > 0$ an $x_0 \varepsilon \Xi$ such that

$$\left| \frac{d\varphi(x)}{dx} \right| < \epsilon_1, \qquad |x - x_0| < \delta_1$$

where δ_1 is a positive constant that depends on ϵ_1. Choose† h such that‡ $h \varepsilon \mathcal{B}$, $\| h \| = 1$, and $| h(t) | < \frac{1}{2}\delta_1$; and then, for any $\epsilon_2 > 0$, determine T such that

$$\int_{|t|>T} h^2 \, dt \leqq \epsilon_2^2 .$$

Through Lemma I, choose $g \varepsilon \mathcal{B}$ such that $| g - x_0 | < \frac{1}{2}\delta_1$ when $| t | < T$, and set $f = g + h$. Observe that $\| \mathbf{P}\varphi[f] - \mathbf{P}\varphi[g] \|^2 \leqq \| \varphi[f] - \varphi[g] \|^2$ and that the right-hand side is equal to

$$\int_{|t|\leqq T} \{\varphi[g + h] - \varphi[g]\}^2 \, dt + \int_{|t|>T} \{\varphi[g + h] - \varphi[g]\}^2 \, dt$$

$$\leqq \sup_{|t|\leqq T} \left| \frac{\varphi[g + h] - \varphi[g]}{h} \right|^2 \int_{|t|\leqq T} h^2 \, dt$$

$$+ M^2 \int_{|t|>T} h^2 \, dt \leqq \epsilon_1^2 + M^2 \epsilon_2^2 .$$

Since ϵ_1^2 and ϵ_2^2 are arbitrary positive constants, our proof is complete. *Remark:* The proof can easily be extended to cover some situations in which the variation of $\varphi(x,t)$ with t plays an important role. One such situation is that in which

$$\left. \frac{\partial \varphi(x,t)}{\partial x} \right|_{x=\zeta(t)} = 0$$

where $\zeta(t)$ is continuous for all finite t and $\partial\varphi/\partial x$ is uniformly continuous in a neighborhood of the curve $x = \zeta(t)$.

VI. ACKNOWLEDGMENT

The writer is indebted to H. J. Landau and H. O. Pollak for carefully reading the draft.

APPENDIX

Some Results Related to the Previously Mentioned Application of the First Part of Theorem III

Suppose that $\mathbf{L}$ is redefined by

$$\mathbf{L}f = \int_{\Sigma} l(t,\tau) \, f(\tau) \, d\tau, \quad f \varepsilon \mathfrak{D}(\Sigma)$$

† The writer is indebted to H. J. Landau for suggesting this approach.

‡ The function $(\sin kt)/\sqrt{k\pi}\, t$ satisfies the unit norm condition and for sufficiently small k satisfies the other two requirements.

where

$$\int_{\Sigma}\int_{\Sigma} | l(t,\tau) |^2 \, d\tau \, dt < 1.$$

Then **PL** is a mapping of $\mathfrak{D}(\Sigma)$ into itself such that for any $h \varepsilon \mathfrak{D}(\Sigma)$, the equation $h = f + \mathbf{PL}f$ possesses a unique solution $f \varepsilon \mathfrak{D}(\Sigma)$. The proof of this result follows from Theorem I, a two-fold application of the Schwarz inequality which shows that

$$| \operatorname{Re}(\mathbf{L}z,z) | \leqq \| z \|^2 \int_{\Sigma}\int_{\Sigma} | l(t,\tau) |^2 \, d\tau \, dt$$

for all $z \varepsilon \mathfrak{D}(\Sigma)$, and a similar calculation using the Schwarz inequality which establishes that $\mathbf{L}z \varepsilon \mathfrak{L}_2$ whenever $z \varepsilon \mathfrak{D}(\Sigma)$ and that there exists a constant k such that $\| \mathbf{P}(\mathbf{I} + \mathbf{L})z \| \leqq k \| z \|$ for all $z \varepsilon \mathfrak{D}(\Sigma)$.

The result mentioned above can be obtained also from a direct consideration of the pertinent Fredholm integral equation:[6]

$$h(t) = f(t) + \int_{\Sigma} e(t) \, l(t,\tau) \, f(\tau) \, d\tau, \tag{1}$$

where

$$\begin{aligned} e(t) &= 1, \qquad t \varepsilon \Sigma, \\ &= 0, \qquad t \not\varepsilon \Sigma. \end{aligned}$$

In addition when

$$l(t,\tau) = 0, \qquad t < \tau$$

[i.e., when $l(t,\tau)$ is a Volterra kernel], it is known[6] that (1) possesses a solution f if

$$\sup_{t,\tau} | l(t,\tau) | < \infty, \quad \int_{\Sigma} | h(t) | \, dt < \infty,$$

and Σ is a bounded set.

REFERENCES

1. Landau, H. J., and Miranker, W. L., The Recovery of Distorted Band-Limited Signals, Jour. Math. Anal. and Appl., **2**, February, 1961, pp. 97–104.
2. Zames, G. D., Conservation of Bandwidth in Nonlinear Operations, Quarterly Progress Report, MIT Research Laboratory of Engineering, October, 1959, No. 55.
3. Landau, H. J., On the Recovery of a Band-Limiting Signal, After Instantaneous

Companding and Subsequent Band-Limiting, B.S.T.J., **39**, March, 1960, pp. 351–364.
4. Kolmogorov, A. N., and Fomin, S. V., *Elements of the Theory of Functions and Functional Analysis*, Graylock Press, New York, 1957.
5. Landau, H. J., and Pollak, H. O., Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty — II, B.S.T.J., **40**, January, 1961, pp. 65–84.
6. Riesz, F., and Sz.-Nagy, B., *Functional Analysis*, Frederick Ungar Publishing Co., New York, 1955.

A Radiometer for a Space Communications Receiver

By E. A. OHM and W. W. SNELL

(Manuscript received March 28, 1963)

By adding a square wave of noise to the input of an ultra-low-noise receiver via a directional coupler, a radiometer with a sensitivity greater than a Dicke type can be achieved when the basic system temperature is less than 18°K. A noise-adding radiometer is compatible with a communications receiver and has been used (i) to measure and monitor the absolute system temperature and (ii) to check the boresighting of a space communications antenna by detecting and tracking radio stars.

I. INTRODUCTION

A noise-adding radiometer, unlike the Dicke type, does not require the input to be switched to a reference temperature. Since a typical good switch adds 7°K or more to the system temperature, it can cause a relatively large increase in the temperature of an ultra-low-noise receiver, and this in turn will cause a significant decrease in the radiometer and/or communications sensitivity. The use of an input switch can be avoided by using a noise-adding radiometer which, for an ultra-low-noise system temperature, is just as sensitive as a Dicke radiometer. A unique feature is that it can be added to an ultra-low-noise communications receiver without causing a large increase in the system temperature. Thus a sensitive tracking receiver, designed primarily to handle communications,[1] can also be used to monitor, measure, and map the system environment temperature including radio stars. Conversely, the radio stars with known positions can be used to check the boresighting of the antenna.

The major hardware components required for a noise-adding radiometer are readily available; the excess noise temperature, mismatch, and instability problems normally associated with a mechanical or ferrite switch are avoided; and the fluctuations due to imperfect circuit components can be reduced to an acceptable value by using a new high-

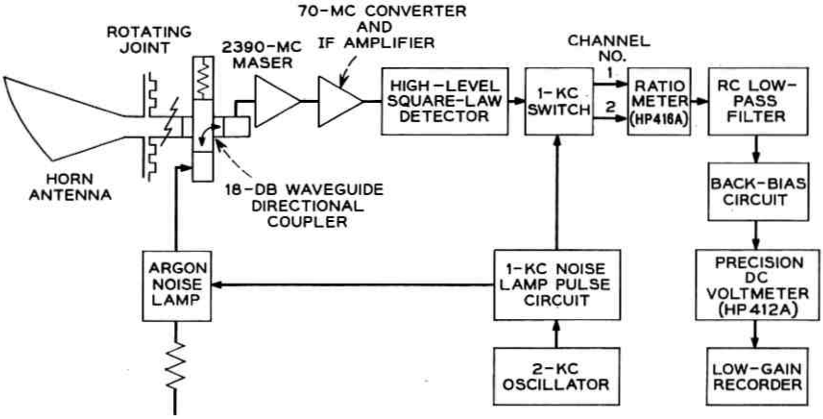


Fig. 1 — Block diagram of a noise-adding radiometer.

output-level square-law detector in combination with an improved noise lamp pulse circuit. In particular, a threshold of $\Delta T = 0.04°\text{K}$ ($2\frac{1}{2}$ times theoretical) has been achieved for periods of 10 seconds when the post-detection time constant is 1 second. For 30-minute periods, a long-term threshold of $\Delta T = 0.12°\text{K}$ (10 times theoretical) has been achieved when the time constant is 15 seconds. Due to the rather small aperture of the Crawford Hill horn-reflector antenna, 26.8 square meters, the corresponding long-term flux or power density threshold is 1.2×10^{-25} watts meter^{-2} cps^{-1}, but this is sufficiently sensitive to detect and track 30 or more radio stars as well as Venus near inferior conjunction. Cassiopeia A and Virgo A have been measured and found to have flux densities of 1.47×10^{-23} and 1.48×10^{-24} watts meter^{-2} cps^{-1}, respectively, at 2390 mc. The uncertainty of measurement is less than 15 per cent. Since absolute system temperatures can be rapidly and precisely recorded for many hours at a time, it was also practical to obtain data for, and prepare, an accurate environment temperature map of the horn antenna site at Crawford Hill, New Jersey.

A block diagram of the noise-adding radiometer is shown in Fig. 1. In the manual mode of operation, a known amount of excess noise from an argon noise lamp is added to the input circuit via a waveguide directional coupler. This gives a ratio, Y, of the system input temperature with the noise added, $T_S + T_A$, to the system input temperature, T_S.

$$Y = \frac{T_S + T_A}{T_S} \tag{1}$$

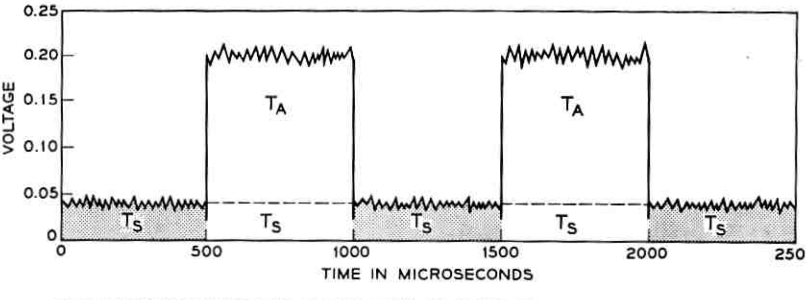
therefore

$$T_S = \frac{T_A}{Y - 1}. \tag{2}$$

To determine T_S, Y can be measured by noting the change in IF attenuation required to keep the IF output power constant when the noise lamp is turned on. Alternatively, since the detector is square-law and the IF amplifier is linear, Y is also given by the ratio of the output voltages of the square-law detector. Since the first method is more accurate, it serves as a calibration check for the second, which is more suitable for a continuous measurement.

The ratio of output voltage, Y, is generated a thousand times per second by pulsing the noise lamp at a 1-kc (50 per cent duty cycle) rate. This produces a rectangular wave at the output of the square-law detector as shown in Fig. 2. The rectangular wave is then passed through a solid-state single-pole, double-throw switch, also operated at 1 kc, where the voltage proportional to $T_S + T_A$ is always switched to channel 1, and the voltage proportional to T_S is always switched to channel 2. The two waveforms are filtered to obtain the fundamental 1-kc components and are then connected in-phase to a ratio-meter. The ratio, Y, is continuously indicated on a calibrated meter, and also by the output voltage of the ratio-meter. Since T_S is a function of Y and the known constant T_A, (2), the continuous outputs can be readily calibrated in terms of the absolute system temperature.

A small change in input temperature can be measured with good accuracy by using observed values of Y and ΔY. Differentiating (2)



THE VOLTAGE UNDER THE CLEAR AREA IS SWITCHED TO CHANNEL 1 AND THAT OF THE SHADED AREA IS SWITCHED TO CHANNEL 2

Fig. 2 — Output voltage of a square-law detector.

$$\frac{dT_S}{dY} = -\frac{T_A}{(Y-1)^2}.$$

Substituting T_S of (2) for $T_A/(Y-1)$

$$\Delta T_S = -\frac{T_S}{Y-1}\Delta Y \tag{3}$$

where T_S, Y, and ΔY can be found from the output voltages of the ratio-meter.

II. MINIMUM DETECTABLE CHANGE OF INPUT TEMPERATURE

It is shown in Appendix A that the theoretical minimum change of input temperature which causes the output voltage, V, to change the same amount as the rms value of the noise fluctuation is

$$\Delta T_S \text{ (theoretical)} = T_S\left(1 + \frac{T_S}{T_A}\right)\frac{\pi}{2}\frac{1}{\sqrt{B\tau}} \tag{4}$$

where: T_S = the total system temperature referred to the waveguide input

T_A = the temperature added when the noise lamp is on

B = the IF (predetection) bandwidth

τ = the RC time constant of the output (post-detection) filter.

Although it may appear anomalous that ΔT_S is reduced as T_A is increased, this can be explained in terms of the relative amplitude of the 1-kc rectangular wave in channel 1 compared to that in channel 2. Referring now to Fig. 2, an increase in system temperature will increase the amplitude of each 1-kc component the same amount. If one is much larger than the other, however, the percentage decrease in ratio, Y, will be larger, and this in turn will cause a larger change in the output voltage of the ratio-meter. At the same time, the theory of a square-law detector shows that the fluctuation in each channel due to noise power is proportional to the 1-kc signal power. Thus the signal-to-noise ratio in each channel is independent of the amplitude of the 1-kc rectangular wave. It follows that the signal-to-noise ratio at the output of the ratio-meter is also independent of the input amplitudes. Since the output signal voltage due to a change in system temperature has been enhanced, and the output signal-to-noise ratio is unchanged, it is thus possible to detect a smaller signal when the 1-kc rectangular waves have a larger difference in amplitude. In this application, the difference is achieved by adding noise T_A to channel 1.

In the limit, T_A and the amplitude of channel 1 are infinite. In this case a change in system temperature cannot affect the amplitude of channel 1, and it can be considered as a reference. Since the radiometer is now insensitive to the system temperature for half the time, it operates in the limit as a Dicke radiometer. For comparison, the sensitivity of a Dicke radiometer is discussed later in connection with (15). A tentative comparison of (15) to (4) with $T_A \rightarrow \infty$ shows that *if* the system temperatures could be made equal, the sensitivities would also be equal.

Since T_A cannot be made infinite, its value must be taken into account when calculating the sensitivity of a noise-adding radiometer. In particular, T_A can be altered over a wide range by changing the coupling, L, of the directional coupler, i.e.,

$$T_A = L\,T_H \tag{5}$$

where T_H is the excess noise available from the noise lamp. When the noise lamp is off, the coupling L also adds room temperature noise from the noise lamp termination to the basic system temperature

$$T_S = T_{\text{basic}} + T_{\text{room}}\,L = T_{\text{basic}} + 290L. \tag{6}$$

Thus T_S will also be altered over a wide range. Upon substitution of (5) and (6) into (4)

$$\Delta T_S(\text{theoretical}) = \left\{(T_{\text{basic}} + 290L) + \frac{(T_{\text{basic}} + 290L)^2}{LT_H}\right\}\frac{\pi}{2}\,\frac{1}{\sqrt{B\tau}}. \tag{7}$$

By differentiating (7) with respect to L and setting the result equal to zero, it can be shown that (7) has an optimum minimum when L has the value

$$L(\text{optimum}) = \frac{T_{\text{basic}}}{(290)^{\frac{1}{2}}(T_H + 290)^{\frac{1}{2}}}. \tag{8}$$

Substituting back into (7)

$$\Delta T_S(\text{optimum}) \approx \left[1 + 2\left(\frac{290}{T_H}\right)^{\frac{1}{2}} + 2\left(\frac{290}{T_H}\right) + \left(\frac{290}{T_H}\right)^{\frac{3}{2}}\right]\frac{\pi}{2}\,\frac{T_{\text{basic}}}{\sqrt{B\tau}}. \tag{9}$$

For a practical value of T_H, 10,200°K,[2] (9) reduces to

$$\Delta T_S(\text{optimum}) = 1.4\,\frac{\pi}{2}\,\frac{T_{\text{basic}}}{\sqrt{B\tau}}. \tag{10}$$

For a communications receiver, it is desirable to minimize L of (6) and have T_S as close as practical to T_{basic}. Using (6) and (8):

$$\frac{T_S}{T_{\text{basic}}} = 1 + \frac{290L}{T_{\text{basic}}} = 1 + \frac{L}{L(\text{optimum})}\left(\frac{290}{T_H + 290}\right)^{\frac{1}{2}}. \qquad (11)$$

For $L = L(\text{optimum})$ and $T_H = 10{,}200°\text{K}$, T_S is 16.6 per cent larger than T_{basic}. For $L = \frac{1}{4} \times L(\text{optimum})$ this can be reduced to a more acceptable 4.2 per cent. By substituting $L = \frac{1}{4} \times L(\text{optimum})$ into (7) and comparing the result with that of (9), it can be shown that ΔT_S(theoretical) is increased 35 per cent. Since ΔT_S(theoretical) is equivalent to the theoretical rms value of noise fluctuation, and since it has been found that other practical sources of fluctuation add $1\frac{1}{2}$ times as much to this value, the percentage increase in ΔT_S(total) due to an increase in ΔT_S(theoretical) is reduced by a factor of $2\frac{1}{2}$. Thus, in practice, the total fluctuation is increased only about 14 per cent. Since this is an acceptable penalty, the recommended value of L is

$$L(\text{for communications receiver}) = \tfrac{1}{4} \times L(\text{optimum}).$$

For the nonoptimum values of T_H and L used here, along with the values of other parameters encountered in the experiment, i.e., for

$$\begin{aligned} T_H &= 6190°\text{K}^* \\ L &= 0.0153(-18.15\ \text{db}) \\ T_A &= L\,T_H = 94.6°\text{K} \\ T_S &= 21.0°\text{K (at the zenith)} \\ 290L &= 4.45°\text{K} \\ T_{\text{basic}} &= T_S - 290L = 16.55°\text{K} \\ B &= 7.75\ \text{mc} \\ \tau &= 1\ \text{sec.} \end{aligned}$$

The nonoptimum theoretical value of ΔT_S can be found by inserting the values of T_S and T_A in (4) which, for convenience, can be written

$$\Delta T_S(\text{theoretical}) = \frac{T_S}{T_{\text{basic}}}\left(1 + \frac{T_S}{T_A}\right)\frac{\pi}{2}\frac{T_{\text{basic}}}{\sqrt{B\tau}} = 1.55\,\frac{\pi}{2}\frac{T_{\text{basic}}}{\sqrt{B\tau}}. \qquad (12)$$

Comparison of (12) and (10) shows that the theoretical threshold

* T_H was relatively small since it came from a coaxial noise lamp and was further reduced by a coaxial line loss. See p. 1087 of Ref. 1.

sensitivity with the above experimental parameters is only 10 per cent less than optimum.* By inserting experimental values for T_{basic} and B in (12), the corresponding numerical value of ΔT_S is

$$\Delta T_S(\text{theoretical}) = 0.015°\text{K} \tag{13}$$

when the post-detection time constant, τ, is one second.

III. COMPARISON WITH A DICKE RADIOMETER

An expression for the threshold sensitivity of a Dicke[3] radiometer in which only one RF sideband is contributing to the receiver output has been worked out in similar terms by Selove.[4] His analysis assumes that the switched reference temperature is small compared to the over-all system temperature. To be valid for an ultra-low-noise receiver, it must be assumed that the switched reference temperature is very low and about equal to the antenna-plus-sky temperature. It is not at room temperature as in the original Dicke radiometer. Thus, from the first two paragraphs of Selove's Appendix,

$$\frac{\Delta T_S}{T_S} = \pi \sqrt{\frac{b}{B}} \tag{14}$$

where: b = the output low-pass-filter (post-detection) bandwidth, and B = the IF (pre-detection) bandwidth. On writing $b = \frac{1}{4}\text{RC} = 1/4\tau$, i.e., in terms of the time constant of an equivalent noise bandwidth, the threshold temperature is

$$\Delta T_S(\text{Dicke}) = T_S \frac{\pi}{2} \frac{1}{\sqrt{B\tau}} \tag{15}$$

where T_S, in this case, is composed of the basic system temperature plus the temperature added by the required input switch.

$$T_S = T_{\text{basic}} + T_{\text{switch}} .$$

The effect of T_{switch} on the sensitivity can be seen by putting (15) in the form

$$\Delta T_S(\text{Dicke}) = \left(1 + \frac{T_{\text{switch}}}{T_{\text{basic}}}\right) \frac{\pi}{2} \frac{T_{\text{basic}}}{\sqrt{B\tau}} .$$

From comparison with the threshold temperature of an optimized noise-adding radiometer, (10),

* For the value of T_H used here and the observed value of T_{basic}, L is very close to the optimum value called for by (8). Thus nearly all the reduction in sensitivity is due to the relatively small value of T_H.

$$\frac{\Delta T_S(\text{Dicke})}{\Delta T_S(\text{noise-adding})} = \left(\frac{1 + (T_{\text{switch}}/T_{\text{basic}})}{1.4}\right). \qquad (16)$$

Thus, when $T_{\text{switch}} \ll T_{\text{basic}}$, the threshold temperature of a noise-adding radiometer is 40 per cent greater than that of the Dicke radiometer. However, the threshold temperature will be less; i.e., the noise-adding radiometer will be more sensitive, if $T_{\text{basic}} < 2.5T_{\text{switch}}$. Since a practical input waveguide switch has an insertion loss of 0.1 db or more, a typical value of T_{switch} is at least 7°K. Thus, if the basic system temperature is 18°K or less, an optimized noise-adding radiometer can be more sensitive than a Dicke radiometer. Since the over-all sensitivity in each case is determined largely by fluctuations added by the required practical circuits, and since the circuits for each radiometer are different, the theoretical comparison at this time merely indicates that the over-all sensitivities are similar.

IV. MINIMUM DETECTABLE POWER DENSITY

Although the sensitivity of a radiometer can be conveniently expressed in terms of ΔT_S, the more important system parameter is the minimum detectable change of power density per cycle of bandwidth. It will now be shown how these are related by the effective area of the antenna. To start,

$$P_{\text{received}} = \tfrac{1}{2} \times P \times A$$

where: P = the incident power flow in watts per square meter, and A = the effective antenna area in square meters. The factor $\frac{1}{2}$ allows for the fact that the receiver is sensitive to only a single polarization. Solving for P,

$$P = \frac{2}{A} \times P_{\text{received}} = \frac{2}{A} KT_SB$$

where: K = Boltzmann's constant, B = the bandwidth in cycles per second, and T_S = the equivalent input temperature. The incident power flow per cycle of bandwidth is therefore

$$\frac{P}{B} = \frac{2K}{A} T_S. \qquad (17)$$

In the radio astronomy literature, the quantity P/B is often called the flux density, S. Small changes in flux density, ΔS, are proportional to changes in the input temperature, ΔT_S, and therefore,

$$\Delta S = \Delta\left(\frac{P}{B}\right) = \frac{2K}{A}\Delta T_S. \qquad (18)$$

Upon substitution of numerical values for K, 1.380×10^{-23} joules/degree, and A, 26.8 square meters at 2390 mc,[5]

$$\Delta S = 1.02 \times 10^{-24} \times \Delta T_S . \tag{19}$$

Upon further substitution of the minimum detectable change of input temperature, 0.015°K from (13), the minimum detectable change in flux density, for $\tau = 1$ second, is

$$\Delta S(\text{theoretical}) = 1.53 \times 10^{-26} \text{ watts meter}^{-2} (\text{cps})^{-1} \tag{20}$$

The experimental value of ΔS is somewhat larger, and the increase is due to other sources of system temperature fluctuation, which will be identified and discussed in the following description of the radiometer parts.

V. COMMUNICATIONS RECEIVER

The Echo receiver had an over-all system temperature of 21.0°K,[1] and its steerable horn-reflector antenna provided an effective area of 26.8 square meters.[5] The area is rather small for observing point-source radio stars, but the disadvantage is compensated, in part, by the small contribution to the system temperature by the far-side and back lobes of a horn-reflector antenna. This minimizes the random change in system temperature as the antenna beam is moved, and this in turn allows (*i*) tracking to achieve a longer observation time and (*ii*) lobing to obtain a more accurate position measurement.

The antenna is connected to the maser package with about 5 feet of assorted waveguide. Included is a rotating joint for mechanically decoupling the antenna from the receiver and a 18.15-db directional coupler for adding noise from a noise lamp. Of the 21.0°K system temperature, about 2.5°K is due to the loss and temperature of the waveguide and 4.5°K is due to the room temperature termination of the directional coupler. Of the 8°K added by the maser package, about 7°K is believed due to the near-room-temperature insertion loss of the input coaxial line.[6] Thus, about 14°K of the system temperature is proportional to room temperature and as such is a source of long-term fluctuation.

$$\Delta T_{SR} = 14 \times \frac{\Delta T(\text{room})}{T(\text{room})} . \tag{21}$$

For $T = 290°\text{K}$ and an observed value of $\Delta T(\text{room}) = \pm 1°\text{K}$, due to air conditioner cycling, the calculated rms value of ΔT_{SR} is 0.034°K. However, the waveguide and coaxial lines have a poor thermal contact with the air and thus a long thermal time constant. Although the period of air conditioner cycling depends on the weather, it is usually relatively short, and therefore the actual value of ΔT_{SR}, in general, will be somewhat less. A typical value is probably about 0.02°K.

In regard to the rotating joint, a gap of 20 mils or less, and an offset error of 15 mils or less,[1] were sufficient to reduce this source of temperature fluctuation to a trivial amount.

A balanced diode converter follows the maser amplifier and contributes a small amount of noise, T_{SC}, to the system temperature

$$T_{SC} = \frac{T_{\text{converter}}}{G_{\text{maser}}} = \frac{2700°K^*}{4000\ (36\,\text{db})} = 0.68°K. \tag{22}$$

T_{SC} will change, however, if either the maser gain or the converter temperature changes. Taking the total differential

$$\Delta T_{SC} = T_{SC}\left(\frac{\Delta T_c}{T_c}\right) - T_{SC}\left(\frac{\Delta G_m}{G_m}\right)$$

where: $\Delta T_c/T_c$ = the estimated change in normalized converter temperature in 30 minutes = ±0.01

$\Delta G_m/G_m$ = the measured change in normalized maser gain in 30 minutes = ±0.045.

Assuming that ΔT_c and ΔG_m are statistically independent

$$\Delta T_{SC} = T_{SC}\left[\left(\frac{\Delta T_c}{T_c}\right)^2 + \left(\frac{\Delta G_m}{G_m}\right)^2\right]^{\frac{1}{2}}. \tag{23}$$

Upon substitution of the numerical values, ΔT_{SC} (rms value) = 0.022°K. This source of fluctuation can be nearly eliminated, i.e., T_{SC} can be reduced toward zero, by using a maser with a larger gain or by using two masers in series.

A net gain of 116 db is provided to drive the high-level square-law detector with an IF noise power of +6 dbm when the noise lamp is on. This is 12 db under the maximum linear IF output power and provides a safe margin for the higher noise peaks. The predetection bandwidth, B, of the radiometer is limited by the converter preamplifier to 7.75 mc. If this is increased to 16 mc, the maser bandwidth, the theoretical sensitivity, from (4), could be increased by a factor of $1\frac{1}{2}$.

VI. HIGH-OUTPUT-LEVEL SQUARE-LAW DETECTOR

By detecting a high-output level of voltage, the separate channel gains which follow the 1-kc switch can be reduced to a minimum. Since

* The interconnecting cable loss, 2.3 db, is included as part of this temperature.

the gains can vary independently and thus increase the system fluctuation, they can and should be reduced to a minimum. The square-law characteristic is needed to convert the IF power, which is proportional to the input temperature, to a linear output voltage. The combination of high-output level and square-law is usually difficult to obtain, but has been achieved with the circuit shown in Fig. 3. As indicated, a relatively high output voltage, 0.5 volt, can be generated by a network of series-parallel diodes when the available input power is +7.5 dbm. For expediency, the detector assembly was matched to the output im-

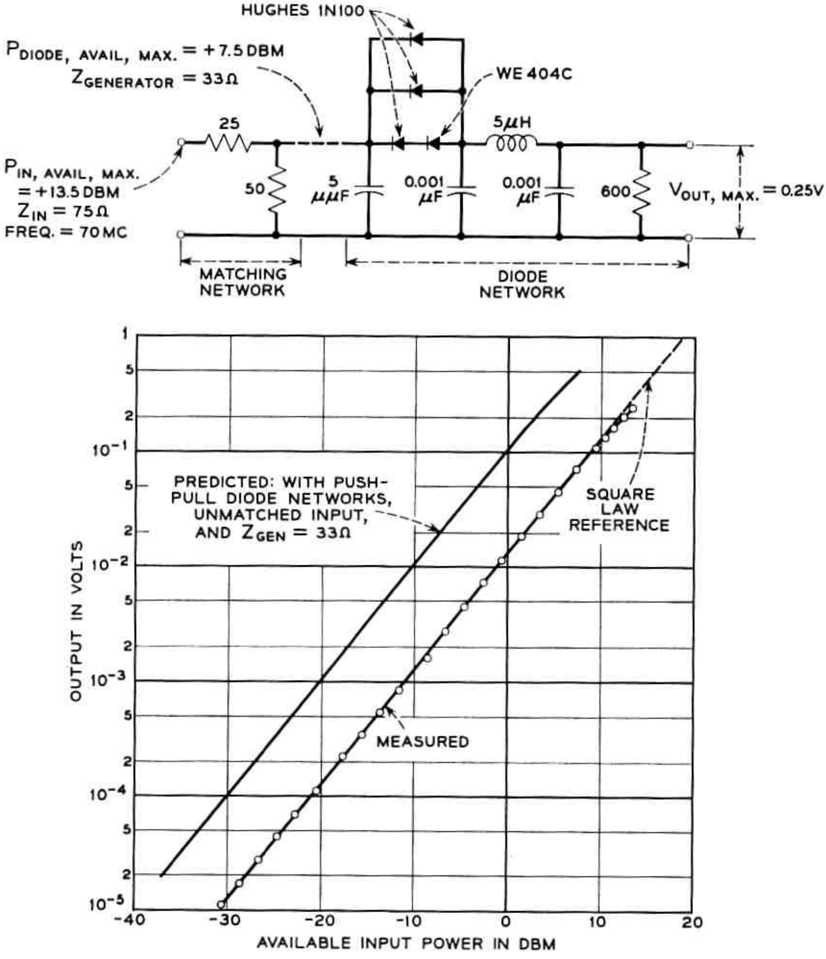


Fig. 3 — Characteristics of a high-output-level square-law detector.

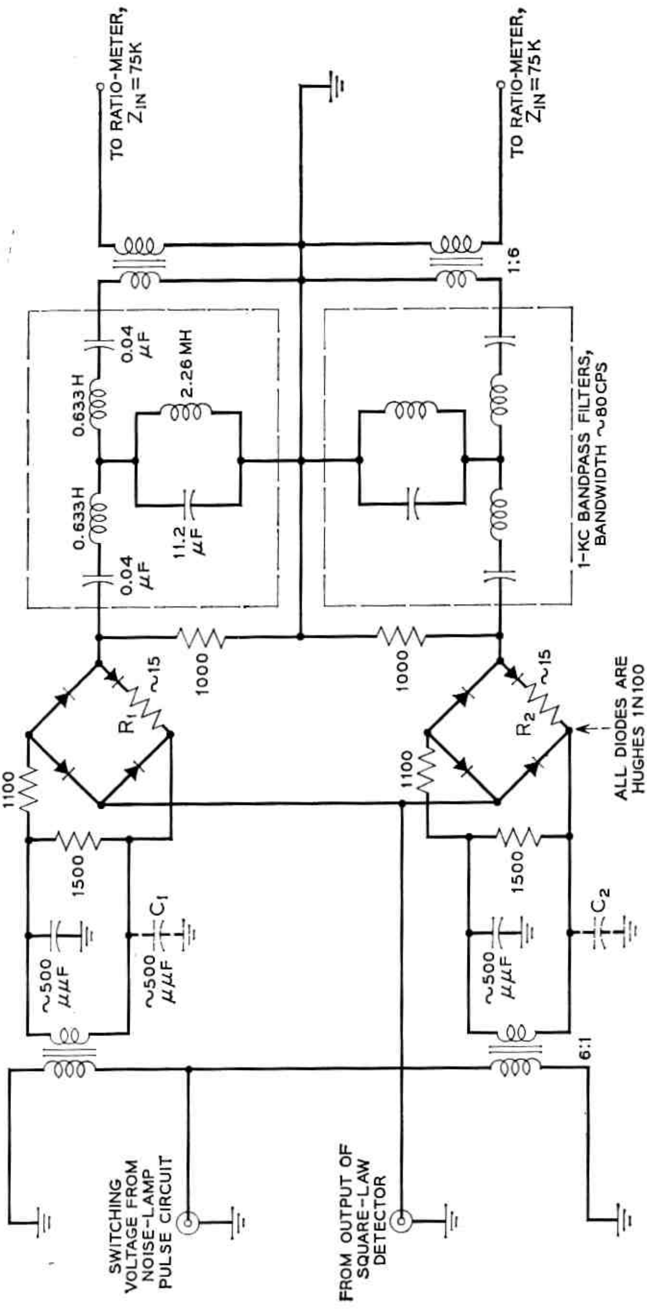

TO RATIO-METER, Z_{IN}=75K
TO RATIO-METER, Z_{IN}=75K
0.04 μF
0.633 H
2.26 MH
0.633 H
11.2 μF
0.04 μF
1-KC BANDPASS FILTERS, BANDWIDTH ~80 CPS
1:6
1000
1000
~15
R_1
~15
R_2
ALL DIODES ARE HUGHES 1N100
1100
1100
1500
1500
C_1
C_2
~500 μμF
~500 μμF
~500 μμF
6:1
SWITCHING VOLTAGE FROM NOISE-LAMP PULSE CIRCUIT
FROM OUTPUT OF SQUARE-LAW DETECTOR

Fig. 4 — One-kc diode switch assembly.

pedance of the IF amplifier, and the required low impedance for driving the diode network was obtained, by using a resistive matching network. The measured characteristic is shown by the lower curve of Fig. 3. It was experimentally verified that the output voltage can be doubled for a given available diode drive power by placing a second diode network of reversed polarity in parallel with the first. Since the insertion loss of the resistive matching network is 6 db, and it can be replaced by a lossless reactive network with the same generator impedance, the doubled output voltage can be achieved with a 6-db reduction in the available input power. The anticipated characteristic is indicated by the upper curve of Fig. 3. For a precision square-law application, the output voltage of a single-ended diode network should not exceed 0.125 volt, and that of a push-pull diode network should not exceed 0.25 volt. The corresponding available input power, from the upper curve of Fig. 3, can then be as low as +4 dbm. Since the maximum linear IF noise power output is +6 dbm, and the circuit of Fig. 3 was used in the experimental radiometer, the corresponding output voltage, from the lower curve of Fig. 3, was on the order of 0.05 volt.

VII. DIODE SWITCH ASSEMBLY

The diode switch assembly consists of two clusters of Hughes 1N100 diodes, each forming a bridge network as shown in Fig. 4. The two bridge networks are energized 180° out of phase and receive a 1-kc switching voltage, a 15-volt, peak-to-peak square wave, from the noise lamp pulse circuit. The switching voltage is isolated from the input and output circuits by the balance resistors R_1, R_2 and capacitors C_1, C_2. Since the isolation tuning is a function of particular diode impedances and driving transformer capacitances (to ground), the values shown are nominal. By potting the diodes in an insulating compound, it has been possible to reduce the system fluctuation due to thermal changes to a negligible amount.*

VIII. RATIO-METER

A ratio-meter (Hewlett Packard 416A) is used to measure the ratio of the sinusoidal voltage amplitudes supplied by the 1-kc switch assembly. One output for indicating the ratio, Y, is a meter on the front

* The diodes are switched continuously, even when the noise lamp is off, in order to provide 1-kc reference signals for the ratio-meter when $Y = 1$. This also improves the long-term stability since possible diode heating, due to switching power, is held constant.

panel. From the theory of operation,[7] the differential voltage across the meter is

$$V_{\text{meter}} \approx C \times \tan^{-1}(1/Y). \tag{24}$$

To take nonlinearities into account, the meter face must be calibrated, and one of the factory calibrated scales, Percent Reflection, is equal to $100\sqrt{1/Y}$. Thus, in principle, Y can be derived from this scale. In practice, however, the ratio indicated by the meter may be somewhat less than the true value due to ignition and deionization times associated with the noise lamp. Therefore, the meter was recalibrated, as discussed in Appendix B, for measured values of Y, and a typical result is shown in Fig. 5.

The other output for indicating Y is a single-ended voltage similar in form to (24)

$$V_{\text{out}} \approx 6 \times \tan^{-1}(1/Y). \tag{25}$$

In order to reduce error due to loading by the measuring and recording circuit, a cathode follower was added as shown in Fig. 6, and its output, V, was used as the ratio-meter output voltage. V was calibrated for measured values of Y, as discussed in Appendix B, and a typical result is shown by the upper left-hand curve of Fig. 7. The corresponding locus of T_S was calculated from (2) for the experimental value of T_A, i.e., 94.6°K.

Thus, from Fig. 7, the absolute system temperature, T_S, can be found by measuring the ratio-meter output voltage. In addition, small changes in system temperature, ΔT_S, can be found by measuring V and ΔV and using these in connection with (3)

$$\Delta T_S = \frac{-T_S}{Y-1}\Delta Y = \frac{-T_S}{Y-1}\frac{\Delta Y}{\Delta V}\Delta V \tag{26}$$

where T_S, Y, and $\Delta Y/\Delta V$ are functions of V and can be found from Fig. 7. For $T_S = 21.0°K$, (26) reduces to

$$\Delta T_S = 12.5 \times \Delta V. \tag{27}$$

The ratio-meter offers good discrimination against a change in RF or IF gain. From the accuracy specification, it can be shown that a change of 1 db will change the output voltage about 0.00001 volt. Upon substitution in (27), the apparent change in system temperature (rms value) is $\Delta T_{SG} = 0.001$°K. Since this is an order of magnitude less than the minimum detectable signal given by (13), this source of fluctuation can be neglected. Other sources affected the ratio-meter, how-

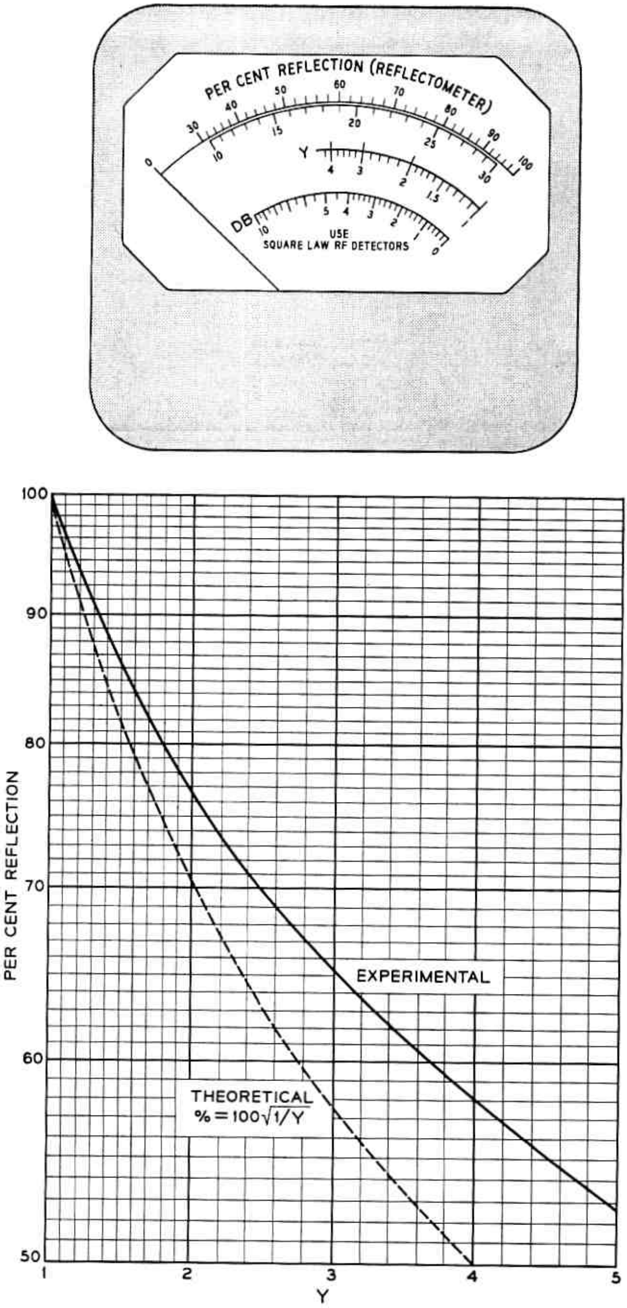
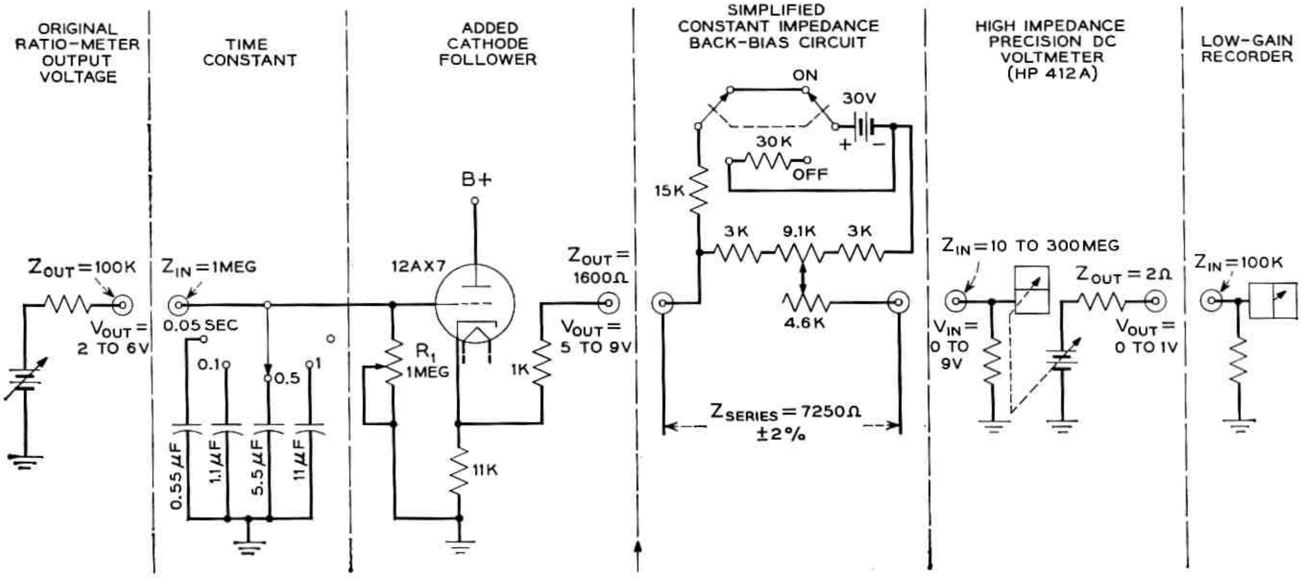


Fig. 5 — A typical ratio-meter "meter" calibration.



THE RATIO-METER OUTPUT VOLTAGE, V, OF FIG. 7 IS REFERRED TO THIS POINT

Fig. 6 — Measuring and recording circuit.

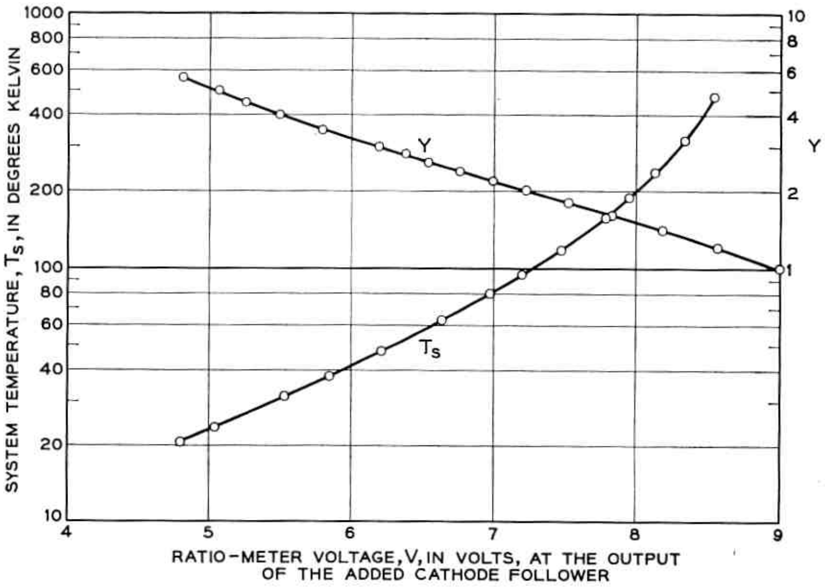


Fig. 7 — Typical ratio-meter voltage calibrations.

ever, and generated relatively large fluctuations, equivalent to 0.5°K. These were traced to changes in line voltage, vibration, and changes in room temperature, and were reduced by (*i*) regulating the line voltage and supplying the filaments with a regulated dc voltage, (*ii*) substituting premium tubes for the standard factory-supplied tubes, and (*iii*) placing the ratio-meter in an oven with a controlled temperature of 105°F. With these modifications, the fluctuations coming from the ratiometer were reduced to

$$\Delta T_{SR} = 0.015^\circ\text{K (short-term)}$$
$$= 0.04^\circ\text{K (long-term)}.$$

The short-term fluctuation is an rms value observed over periods of 10 seconds, and the long-term fluctuation is an additional rms value observed over periods of 30 minutes. Separate channel gains are built into a commerical ratio-meter to allow for low input voltages and to provide for a large dynamic range. Since large voltages are available from a square-law detector, and since the dynamic range required in this application is relatively small, the separate channel gains are not needed and should be reduced and/or eliminated. It is believed that this modification will cause a further significant decrease in the long- and short-term fluctuations contributed by the ratio-meter.

IX. MEASURING AND RECORDING CIRCUIT

The voltage range to be measured and recorded is indicated by the abscissa of Fig. 7. To estimate the required order of stability, the value of ΔV which corresponds to the threshold value of ΔT_S can be found from (26). Solving for ΔV

$$\Delta V = -\frac{Y-1}{T_S}\frac{\Delta V}{\Delta Y}\Delta T_S. \tag{28}$$

Upon substitution of $\Delta T_S = 0.015$°K from (13), $T_S = 21.0$°K, and the values of Y and $\Delta V/\Delta Y$ from Fig. 7 which correspond to $T_S = 21.0$°K, the threshold value of ΔV is 0.0012 volt. In order to measure this change in voltage, a back-bias circuit is needed to buck out the 5 to 9 volts dc; a sensitive dc meter is needed; and the utmost stability is required. These objectives were met by combining a constant impedance back-bias circuit with a precision high-impedance voltmeter (Hewlett-Packard 412A), as shown in Fig. 6.

In regard to the back-bias circuit, note that V, in contrast to ΔV, can be readily measured with the precision voltmeter by turning off the back-bias switch. When this is done, the series impedance (7250 ohms) and the battery drain (1 ma) are both held constant to maintain good stability.

The voltage finally recorded is generated by the precision voltmeter. Since it varies from 0 to 1 volt for any full-scale meter deflection, a low-gain and therefore stable recorder can be used. Since the precision voltmeter drift is less than 0.1 per cent on any scale, and since the drift in back-bias voltage is less than 0.0001 volt, the total fluctuation due to the measuring and recording circuit is negligible.

X. NOISE LAMP PULSE CIRCUIT

The circuit outlined in Figs. 8 and 9 can be used to operate a fluorescent or argon gas discharge tube with a near 50 per cent duty cycle and a repetition frequency from 40 to 2000 cps. The excess noise is very stable; the on current can be varied over a wide range; and the filament is expected to last as long as it would in continuous service. In this application the repetition frequency was adjusted to coincide with the 1-kc center frequency of the ratio-meter.

The key to a stable pulsed noise output is that the high-voltage igni-

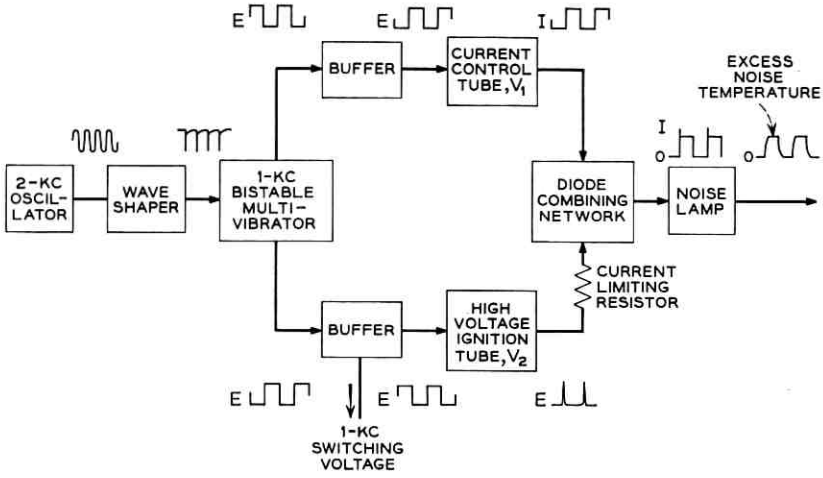


Fig. 8 — Block diagram of an improved noise lamp pulse circuit.

tion spike is generated in parallel with, rather than in series with, the main current path, and it is applied to the noise lamp via a large series resistance. Referring to Fig. 9, the sequence of operation is: (*i*) V_2 is initially conducting and its current stores magnetic energy in the inductance, L_1. V_1 is cut-off. (*ii*) To achieve ignition, V_2 is now cut-off and V_1 is biased into conduction. The slow collapse of the magnetic field maintains a near-constant current, which, since V_2 is cut-off, increases

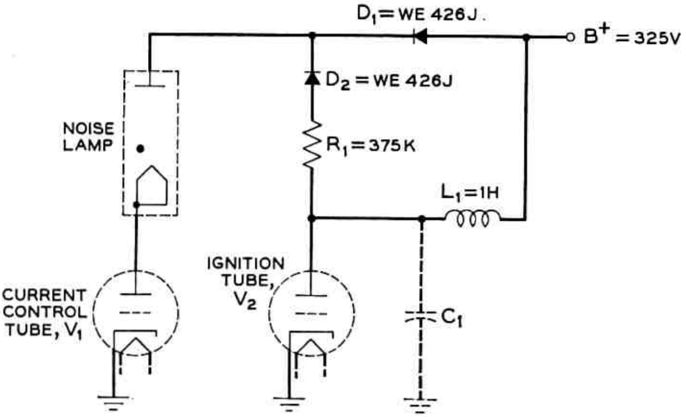


Fig. 9 — Critical parts of an improved noise lamp pulse circuit.

the voltage across C_1 towards a very large value. C_1, incidentally, is due primarily to the stray capacitance associated with L_1. Since the deionized noise lamp has an infinite resistance and since its anode is isolated from $B+$ by diode D_1, the rising voltage of C_1 is also applied across the noise lamp. (*iii*) When sufficient voltage is developed, the noise lamp is ionized and its resistance falls abruptly to 375 ohms. In a conventional circuit in which the anode of the noise lamp is connected to the junction of L_1 and C_1, the charge stored in C_1 is discharged directly through the noise lamp via V_1. This causes an undesirable high-power transient, which in turn causes severe ionic bombardment of the noise lamp filament. In the circuit of Fig. 9, however, the charge stored in C_1 is largely dissipated in the current-limiting resistor, R_1. (*iv*) After ionization, the current control tube, V_1, allows an adjustable amount of current to flow for the rest of the on period via the diode D_1. (*v*) At the beginning of the off period, V_1 is again biased beyond cut-off and V_2 is biased into conduction. This is the initial condition and the sequence is repeated at the start of the next on period.

By accounting for all other output voltage variations, it was estimated that the short-term (10-second) fluctuation at the ratio-meter output due to the noise lamp and its pulse circuit was $\Delta V \approx 0.0008$ volt (rms), and the additional long-term (30-minute) variation was $\Delta V \approx 0.0020$ volt (rms). The equivalent rms changes in system temperature, from (27), are

$$\begin{aligned} \Delta T_{SL} &= 0.010°\text{K (short-term)} \\ &= 0.025°\text{K (long-term)}. \end{aligned}$$

Incidentally, the corresponding change in noise lamp temperature, ΔT_H, can be found by substituting $(T_S + T_A)/T_S$ for Y in (25), and then differentiating with respect to T_A. After rearranging,

$$\Delta T_A = -\frac{(T_A + T_S)^2 + \text{T}_S^2}{6T_S}\Delta V.$$

Since ΔT_A is attenuated from ΔT_H by the waveguide directional coupling, L, and by the transmission coefficient, K, of the coaxial line connecting the noise lamp to the directional coupler,

$$\Delta T_H = -\frac{(T_A + T_S)^2 + T_S^2}{6T_SKL}\Delta V. \tag{29}$$

For $K = 0.73$ and the experimental values of T_A, T_S and L, the change in noise lamp temperature is

$$\Delta T_H = 9900\ \Delta V.$$

For the estimated total value of ΔV, 0.0028 volt (rms), the apparent value of ΔT_H (rms) is only 28°K, i.e., very small compared to T_H = 6190°K.

It is shown by (36) of Appendix A that the ratio-meter, to a first order, responds to only the in-phase component of the 1-kc input voltages. By assuming a model of flat-topped and delayed excess noise, it can further be shown, using Fourier analysis, that the difference between the experimental and theoretical calibration curves, Fig. 5, can be explained by an ionization delay of 160 microseconds and a deionization lag of 64 microseconds. For comparison, the desired excess noise interval is 500 microseconds. The calculated delays are similar to those reported by Kuhn and Negrete.[8] Thus the excess noise is apparently delayed into the off time interval and, in addition, it is on for only 404 microseconds. By delaying the 1-kc switching voltage of Fig. 8 about 110 microseconds, T_A can be centered in the on time slot with a guard time of 50 microseconds on either edge. An analysis of this shows that the experimental calibration curve, Fig. 5, will then be within $\frac{1}{2}$ division of the theoretical. The improvement in timing is highly recommended since it will probably eliminate most of the fluctuations attributed to the noise lamp pulse circuit.

XI. SUMMARY OF FLUCTUATIONS

The system fluctuations, in terms of rms changes in system temperature, are summarized in Table I. The first item is a natural fluctuation which is due to an intrinsic property of the input temperature T_S, as described by (4). It is assumed here that T_S is due to a steady contribution from the atmosphere, antenna, waveguide, maser, and IF converter. In contrast to this, all other fluctuations are due to imperfections in the radiometer parts. Items 1 through 4 are short-term variations which occur in less than 10 seconds, and items 5 through 9 are additional long-term variations which occur in periods of 30 minutes. The post-detection time constant, τ, in each case is one second. Some promising means for decreasing the fluctuations are also indicated. As can be seen, the observed short-term fluctuation, item 4, is about $2\frac{1}{2}$ times greater than theoretical, item 1, and the total long-term fluctuation is about 10 times greater than theoretical. By increasing τ to 15 seconds, item 4 can be reduced to 0.01°K. This, however, does not greatly reduce the slow

TABLE I — SYSTEM FLUCTUATIONS, τ = 1 sec

Source	Amplitude (rms)	Recommended Improvements
1. Thermal noise	0.015°K (calculated)	Increase the IF bandwidth to coincide with the maser bandwidth
2. Ratio-meter	0.015°K (measured)	Eliminate separate channel gains
3. Noise lamp	0.010°K (estimated)	Synchronize 1-kc switch interval with on excess noise
4. Subtotal of short-term fluctuations	0.040°K (measured)	Increase the post-detection time constant
5. Input waveguide and coaxial line	0.022°K (estimated)	Insulate, and reduce changes in ambient temperature
6. IF converter	0.022°K (calculated)	Use a maser with increased gain, or two masers in series
7. IF gain change	0.001°K (calculated)	Improve B^+ regulation
8. Ratio-meter	0.040°K (estimated)	Eliminate separate channel gains, and improve B^+ regulation
9. Noise lamp pulse circuit	0.025°K (estimated)	Synchronize 1-kc switch interval with on excess noise, and improve B^+ regulation
Total long-term fluctuation (sum of items 4 through 9)	0.150°K (measured)	

fluctuations, items 5 through 9. Thus, for a reasonably long time constant, the total long-term fluctuation will not be reduced much below 0.12°K. The corresponding measured flux density threshold, from (19), for τ = 15 seconds, is

$$\Delta S(\text{measured}) = 1.2 \times 10^{-25} \text{ watts meter}^{-2} (\text{cps})^{-1}.$$

With the alterations recommended in Table I, and τ retained at 1 second, it is estimated that the total long-term fluctuation of Table I can be reduced by a factor of 3, to 0.05°K. Using (19)

$$\Delta S(\text{predicted}) = 5 \times 10^{-26} \text{ watts meter}^{-2} (\text{cps})^{-1}.$$

XII. EXPERIMENTAL RESULTS

A Crawford Hill sky-plus-environment temperature map was constructed from data taken during an 8-hour period when the sun and moon were below the horizon on Feb. 15, 16, 1961. By avoiding sun and moon temperature anomalies (the sun can add 20°K via side lobes and the moon has been observed to add 16°K via the main beam), it was possible to identify weaker radio sources, including the center of the galaxy, which adds 4.5°K, and delete these from the temperature map.

The raw data consisted of seventeen constant-elevation scans, taken in elevation increments of 1° between −1° and +10°, plus scans at 12°, 15°, 20°, 25° and 30°. A typical scan is shown in Fig. 10. The data were first replotted in terms of system temperature versus elevation for every two degrees of azimuth. With these curves, it was possible to construct a detailed contour map, of which two sample parts are shown in Fig. 11. The absolute accuracy is ±15 per cent, of which ±5 per cent is due to the data-reducing technique and ±10 per cent is due to the temperature calibration accuracy. The latter is discussed in Appendix B.

An elevation scan was made at an azimuth angle of minimum observed temperature, and the results are plotted in Fig. 12. With this curve, it was possible to calculate the absolute value of the zenith sky temperature with good accuracy (see Ref. 1, pages 1088 and 1089), and the result is 2.3 ± 0.2°K. The theoretical value[9] is also plotted and shows good agreement down to an elevation of 1° where the near-side-lobes of the antenna began to intercept the hot earth.

A drift pass of Virgo A, Fig. 13, was obtained by positioning the antenna so the radio source, due to the earth's rotation, would pass through the antenna beam. This, of course, stabilized the side-lobe temperature contributions. At 2390 mc the value of ΔT_S was found to be 1.44°K, and the corresponding flux density, from (19), is 1.48×10^{-24} watts meter^{-2} (cps)$^{-1}$. The value of ΔT_S for Cassiopeia A, from a similar measurement, was found to be 14.3°K, and the corresponding flux density is 1.47×10^{-23} watts meter^{-2} (cps)$^{-1}$. No effort was made to correct these numbers for other weak, but possibly significant, sources in the antenna beam. The flux density measurement accuracy is limited to ±15 per cent, of which ±10 per cent is due to a possible error in the temperature calibration, Appendix B, and ±5 per cent is due to an uncertainty in the antenna gain measurement.

XIII. CONCLUSIONS

A noise-adding radiometer has been found to be a convenient practical tool for measuring small absolute system temperatures over long periods of time. It is compatible with an ultra-low-noise communications receiver, and can be used to check the boresighting of a satellite communications antenna by tracking radio stars.[12] Although the short-term (10-second) system fluctuation is larger than theory by a factor of only $2\frac{1}{2}$ when $\tau = 1$ second, the long-term (30-minute) fluctuation, which limits the minimum detectable power density, is larger than theory by a factor of 10. The sources of excess fluctuation have been identified, and

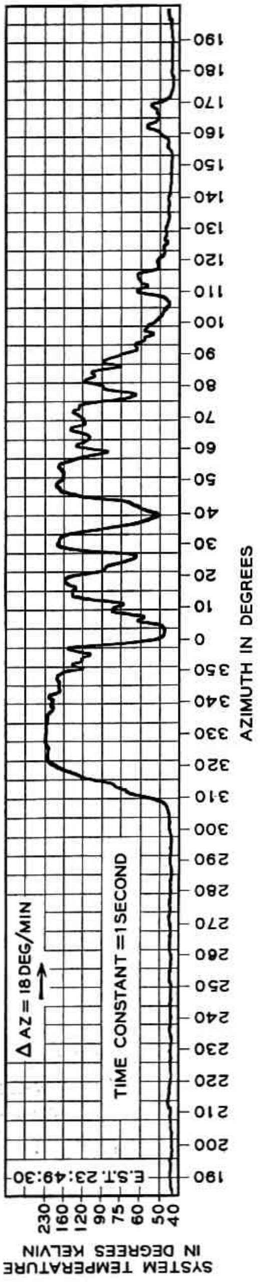


Fig. 10 — System temperature at a 5° elevation.

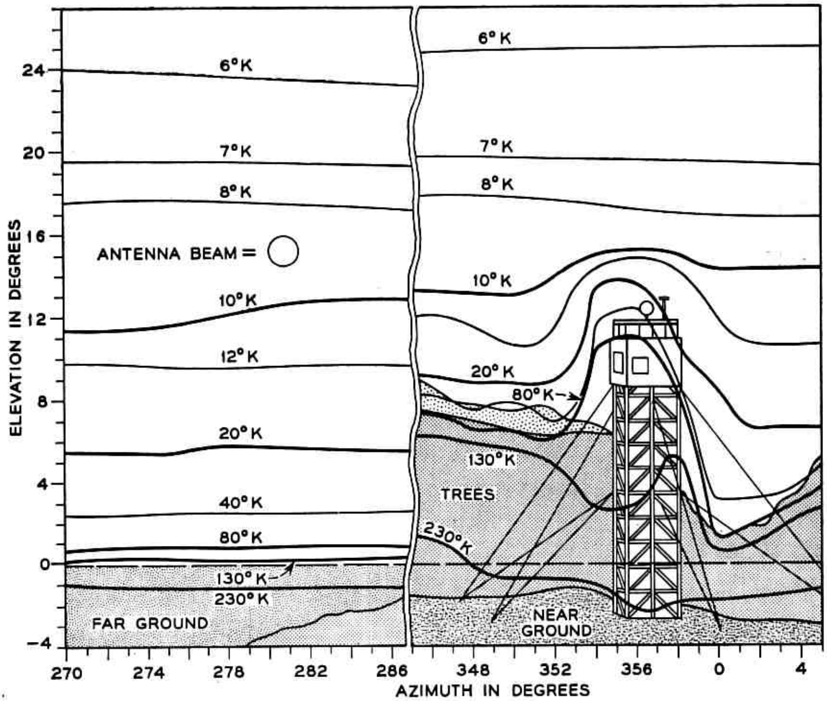


Fig. 11 — Typical environment temperatures of the horn-reflector antenna site at Crawford Hill, N. J. Frequency = 2390 mc; antenna beamwidth = 1.25 degrees.

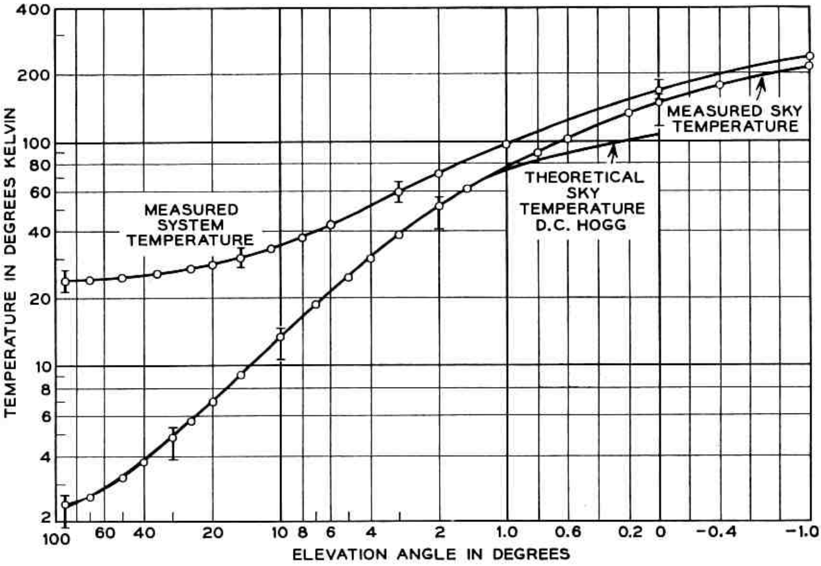


Fig. 12 — Measured sky and system temperatures. Frequency = 2390 mc; azimuth = 190°; time pre-sunrise, February 16, 1961.

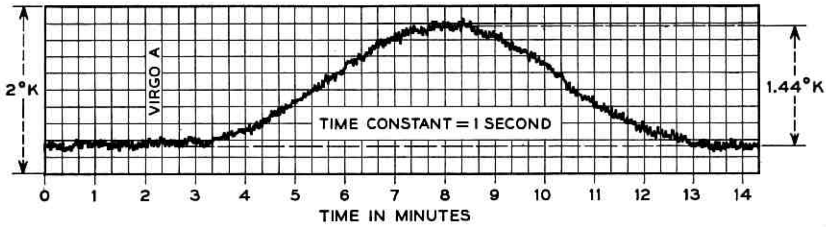


Fig. 13 — Drift pass of Virgo A.

with the suggested alterations, it is believed that the total long-term fluctuation, for $\tau = 1$ second and $B = 16$ mc, can be reduced to 0.05°K, which is larger than theory by a factor of 5.

XIV. ACKNOWLEDGMENTS

The authors are indebted to A. A. Penzias for reducing some sources of receiver fluctuation and to L. H. Enloe for proving the assumptions used in deriving the threshold equation. The authors are also pleased to acknowledge the aid rendered by R. A. Desmond in assembling the noise lamp pulse circuit and the help given by L. Meyers in plotting the data for the environment temperature map.

APPENDIX A

The Minimum Detectable Change in Input Temperature

In the system of Fig. 1, the 1-kc input voltages required by the ratio-meter are derived from the waveforms shown in Fig. 2. From inspection, the voltage switched to each channel has a strong 1-kc component of different amplitude and a small fluctuation due to the input temperature. It is also apparent, but not shown, that a small increase in the system temperature, ΔT_S, will increase each 1-kc component the same amount.

With these inputs, the output of the ratio-meter is a dc voltage on which is superimposed a small fluctuation due to noise. In addition, a slow variation in the dc voltage will occur as in Fig. 13 when noise power received by the antenna increases the system temperature. The theoretical threshold sensitivity will be found by calculating the change of input temperature which causes the dc output voltage to change the same amount as the rms value of the output noise fluctuation.

An examination of the ratio-meter theory of operation[7] shows that the output voltage, V_{out}, is a linear function of the phase angles, φ', which

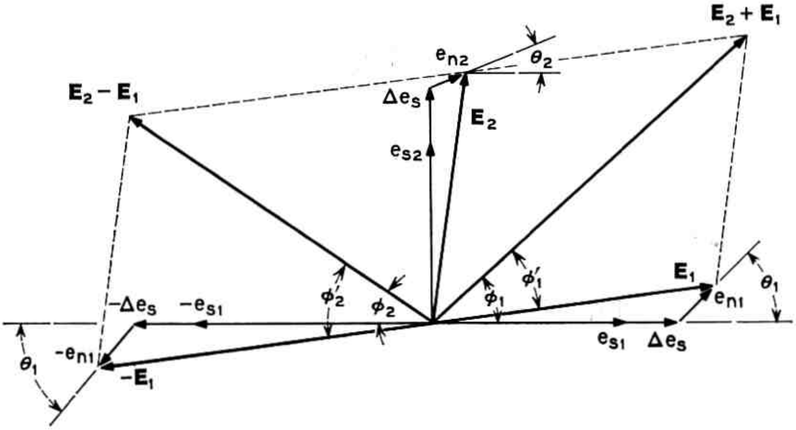


Fig. 14 — Typical ratio-meter input voltages.

result from the vector addition of the input voltages at and around 1 kc. In particular

$$V_{\text{out}} = K\,2(\varphi_1' + \varphi_2')$$

where K is an arbitrary constant and φ_1' and φ_2' are shown in Fig. 14. Note from the geometry that $\varphi_1' + \varphi_2' = \varphi_1 + \varphi_2$, and therefore V_{out} can also be written

$$V_{\text{out}} = K\,2\,(\varphi_1 + \varphi_2). \tag{30}$$

As shown in Fig. 14, φ_1 and φ_2 are functions of the input voltages, $\mathbf{E}_1$ and $\mathbf{E}_2$, each of which consists of three parts.

$$\mathbf{E}_1 = e_{S1} + \Delta e_S + \mathbf{e}_{n1}$$

$$\mathbf{E}_2 = e_{S2} + \Delta e_S + \mathbf{e}_{n2}$$

where: e_{S1} = the 1-kc component in channel 1 due to on-off modulation of the input temperature, $T_S + T_A$.

e_{S2} = the 1-kc component in channel 2 due to on-off modulation of the input temperature, T_S.

Δe_S = an in-phase change in the 1-kc components due to a change in input temperature, ΔT_S.

$\mathbf{e}_{n1}$ = a random fluctuation in channel 1 due to the input temperature, $T_S + T_A$.

$\mathbf{e}_{n2}$ = a random fluctuation in channel 2 due to the input temperature, T_S.

From Fig. 14 it can also be seen that

$$\varphi_1 = \tan^{-1}\left(\frac{e_{S2} + \Delta e_S + e_{n2}\sin\theta_2 + e_{n1}\sin\theta_1}{e_{S1} + \Delta e_S + e_{n2}\cos\theta_2 + e_{n1}\cos\theta_1}\right) \tag{31}$$

$$\varphi_2 = \tan^{-1}\left(\frac{e_{S2} + \Delta e_S + e_{n2}\sin\theta_2 - e_{n1}\sin\theta_1}{e_{S1} + \Delta e_S - e_{n2}\cos\theta_2 + e_{n1}\cos\theta_1}\right). \tag{32}$$

Since Δe_S, e_{n1}, and e_{n2} are small compared to e_{S1} and e_{S2}, (31) and (32) can be written:

$$\begin{aligned}\varphi_1 = \tan^{-1}\frac{e_{S2}}{e_{S1}}\Bigg[1 + \frac{\Delta e_S}{e_{S2}} - \frac{\Delta e_S}{e_{S1}} + \frac{e_{n2}}{e_{S2}}\sin\theta_2 + \frac{e_{n1}}{e_{S2}}\sin\theta_1 \\ - \frac{e_{n2}}{e_{S1}}\cos\theta_2 - \frac{e_{n1}}{e_{S1}}\cos\theta_1\Bigg]\end{aligned} \tag{33}$$

$$\begin{aligned}\varphi_2 = \tan^{-1}\frac{e_{S2}}{e_{S1}}\Bigg\{1 + \frac{\Delta e_S}{e_{S2}} - \frac{\Delta e_S}{e_{S1}} + \frac{e_{n2}}{e_{S2}}\sin\theta_2 - \frac{e_{n1}}{e_{S2}}\sin\theta_1 \\ + \frac{e_{n2}}{e_{S1}}\cos\theta_2 - \frac{e_{n1}}{e_{S1}}\cos\theta_1\Bigg\}.\end{aligned} \tag{34}$$

Using the series expansion for $\tan^{-1}$ and noting with dissimilar brackets the differences in signs of (33) and (34)

$$\varphi_1 = \frac{e_{S2}}{e_{S1}}[1 + \cdots] - \frac{1}{3}\left(\frac{e_{S2}}{e_{S1}}\right)^3[1 + \cdots]^3 + \frac{1}{5}\left(\frac{e_{S2}}{e_{S1}}\right)^5[1 + \cdots]^5 + \cdots$$

$$\begin{aligned}\varphi_2 = \frac{e_{S2}}{e_{S1}}\{1 + \cdots\} - \frac{1}{3}\left(\frac{e_{S2}}{e_{S1}}\right)^3\{1 + \cdots\}^3 \\ + \frac{1}{5}\left(\frac{e_{S2}}{e_{S1}}\right)^5\{1 + \cdots\}^5 + \cdots.\end{aligned}$$

Thus $\varphi_1 + \varphi_2$ for use in (30) is

$$\begin{aligned}\varphi_1 + \varphi_2 = \frac{e_{S2}}{e_{S1}}([1 + \cdots] + \{1 + \cdots\}) \\ - \frac{1}{3}\left(\frac{e_{S2}}{e_{S1}}\right)^3([1 + \cdots]^3 + \{1 + \cdots\}^3) + \frac{1}{5}\left(\frac{e_{S2}}{e_{S1}}\right)^5([1 + \cdots]^5 \\ + \{1 + \cdots\}^5) - \frac{1}{7}\left(\frac{e_{S2}}{e_{S1}}\right)^7([1 + \cdots]^7 + \{1 + \cdots\}^7) + \cdots\end{aligned}$$

The first term of $\varphi_1 + \varphi_2$ reduces to

$$+\frac{2e_{S2}}{e_{S1}}\left(1 + \frac{\Delta e_S}{e_{S2}} - \frac{\Delta e_S}{e_{S1}} + \frac{e_{n2}}{e_{S2}}\sin\theta_2 - \frac{e_{n1}}{e_{S1}}\cos\theta_1\right).$$

The second term reduces to

$$-\,2\left(\frac{e_{S2}}{e_{S1}}\right)^3\left(\frac{1}{3}+\frac{\Delta e_S}{e_{S2}}-\frac{\Delta e_S}{e_{S1}}+\frac{e_{n2}}{e_{S2}}\sin\theta_2-\frac{e_{n1}}{e_{S1}}\cos\theta_1\right).$$

The third term reduces to

$$+\,2\left(\frac{e_{S2}}{e_{S1}}\right)^5\left(\frac{1}{5}+\frac{\Delta e_S}{e_{S2}}-\frac{\Delta e_S}{e_{S1}}+\frac{e_{n2}}{e_{S2}}\sin\theta_2-\frac{e_{n1}}{e_{S1}}\cos\theta_1\right),\ \text{etc.}$$

Therefore $\varphi_1+\varphi_2$ can be written

$$\begin{aligned}\varphi_1+\varphi_2 = 2&\left[\frac{e_{S2}}{e_{S1}}-\frac{1}{3}\left(\frac{e_{S2}}{e_{S1}}\right)^3+\frac{1}{5}\left(\frac{e_{S2}}{e_{S1}}\right)^5-\cdots\right]\\ &+2\left[\frac{e_{S2}}{e_{S1}}-\left(\frac{e_{S2}}{e_{S1}}\right)^3+\left(\frac{e_{S2}}{e_{S1}}\right)^5-\cdots\right]\left[\frac{\Delta e_S}{e_{S2}}-\frac{\Delta e_S}{e_{S1}}\right.\\ &\left.\qquad+\frac{e_{n2}}{e_{S2}}\sin\theta_2-\frac{e_{n1}}{e_{S1}}\cos\theta_1\right].\end{aligned}\tag{35}$$

The first bracket of (35) is the series expansion of $\tan^{-1}(e_{S2}/e_{S1})$. Thus, it is equal to the rest angle, φ_0, which would result from (31) or (32) if the perturbation terms were zero. The first term is accordingly $2\varphi_0$. The coefficient of the second term of (35) can be written

$$\begin{aligned}2\,\frac{e_{S2}}{e_{S1}}\left[1-\left(\frac{e_{S2}}{e_{S1}}\right)^2+\left(\frac{e_{S2}}{e_{S1}}\right)^4-\cdots\right]&=\frac{2(e_{S2}/e_{S1})}{1+(e_{S2}/e_{S1})^2}\\ &=\frac{2(e_{S2}/e_{S1})}{\sqrt{1+(e_{S2}/e_{S1})^2}}\cdot\frac{1}{\sqrt{1+(e_{S2}/e_{S1})^2}}.\end{aligned}$$

Since $\varphi_0 = \tan^{-1}(e_{S2}/e_{S1})$, it follows from trig identities that this is equivalent to

$$2\sin\varphi_0\cos\varphi_0=\sin 2\varphi_0$$

Thus (35) reduces to

$$\varphi_1+\varphi_2=2\varphi_0+\sin 2\varphi_0\left[\frac{\Delta e_S}{e_{S2}}-\frac{\Delta e_S}{e_{S1}}+\frac{e_{n2}}{e_{S2}}\sin\theta_2-\frac{e_{n1}}{e_{S1}}\cos\theta_1\right].$$

Upon substitution in (30), the output voltage of the ratio-meter is given by

$$V_{\text{out}}=K\left\{4\varphi_0+2\sin 2\varphi_0\left[\frac{\Delta e_S}{e_{S2}}-\frac{\Delta e_S}{e_{S1}}+\frac{e_{n2}}{e_{S2}}\sin\theta_2-\frac{e_{n1}}{e_{S1}}\cos\theta_1\right]\right\}\tag{36}$$

where: $\varphi_0=\tan^{-1}(e_{S2}/e_{S1})$.

Since the detector in Fig. 1 has a square-law characteristic, the 1-kc voltage components, e_{S2} and e_{S1}, at the output of the 1-kc switch assembly, are proportional to the input temperatures $T_S + T_A$ and T_S as indicated in Fig. 2. Therefore φ_0 is a function of the temperature, T_A, added by the noise lamp

$$\varphi_0 = \tan^{-1} \frac{e_{S2}}{e_{S1}} = \tan^{-1} \frac{T_S}{T_S + T_A} = \tan^{-1} (1/Y). \tag{37}$$

Although V_{out} of (36) is thus a function of φ_0, note that the signal component, due to Δe_S, and the noise component, due to e_{n1} and e_{n2}, are both proportional to $\sin 2\varphi_0$. Thus the theoretical signal-to-noise ratio is independent of φ_0, and a value of φ_0 less than 45° (typically 12°) merely reduces the signal and noise gain by a factor of (0.4). Incidentally, since only the in-phase components of the noise voltages of Fig. 16 are retained in (36), it can be seen that the output of the ratio-meter responds only to the instantaneous in-phase components of the input voltage amplitudes.

In order to determine the threshold sensitivity from (36), e_{S2}, e_{S1}, Δe_S, e_{n2}, and e_{n1} can be calculated using random noise theory. In particular, since the spectral density of the output of a square law detector with a stationary input of white noise is known,[10] and the correlation time corresponding to the large predetection bandwidth is very small compared to each switched interval, the steady and fluctuating parts of $\mathbf{E}_2$ can be calculated by (*i*) assuming the spectral density of the square-law detector input and output is constant with time and (*ii*) allowing the output of the square-law detector to be switched on and off at a 1-kc rate. Similarly, the steady and fluctuating parts of $\mathbf{E}_1$ can be calculated by assuming a larger constant spectral density, which corresponds to the input temperature when the noise lamp is on. An important consequence of the small correlation time is the noise components $\mathbf{e}_{n1}$ and $\mathbf{e}_{n2}$ are uncorrelated, and thus can be added on a power basis. A more rigorous analysis by L. H. Enloe[11] proves these assumptions and arrives at the same result.

The result of switching (or multiplying) the output of the square-law detector with a 1-kc switch can be readily calculated, since these functions are statistically independent, by convolving the spectra density of the square law detector output with that of the 1-kc switch.

$$S_z(f) = \int_{-\infty}^{+\infty} S_x(\psi)\, S_y(f - \psi)\, d\psi \tag{38}$$

where $S_z(f)$ = output spectral density at frequency f
$S_x(\text{f})$ = spectral density of the 1-kc switch = $\frac{1}{4}$ at dc, $1/\pi^2$ at ±1 kc, $\frac{1}{9}\pi^2$ at ±3 kc, $\frac{1}{25}\pi^2$ at ±5 kc, etc.
$S_y(f)$ = spectral density of the square-law detector output = $4a^2A^2B^2$ at dc + $4a^2A^2(B - |f|)$ where $0 < |f| < B$
a = a scaling constant of the square-law detector } From Ref. 10
A = input spectral density of the square-law detector
B = IF bandwidth.

Since (*i*) the IF bandwidth, $B = 7.75$ mc, is large compared to the switching frequency, $f_0 = 1$ kc, (*ii*) the switch is followed by a bandpass filter of $f_{BP} = f_0 \pm \Delta f/2$, and (*iii*) the switch has a discrete spectral density, (38) reduces to one term for the signal power density at $f_0 = 1$ kc and to a closed series for the bandpass noise power density at $f_0 = 1$ kc.

$$S_z(f_0)\bigg|_{\text{signal}} = 2 \times \frac{1}{\pi^2} \times 4a^2A^2B^2 \tag{39}$$

$$S_z(f_0)\bigg|_{\text{noise}} = 2 \times 4a^2A^2B\left[\frac{1}{4} + \frac{2}{\pi^2} + \frac{2}{9\pi^2} + \frac{2}{25\pi^2} + \cdots\right]$$

$$S_z(f_0)\bigg|_{\text{noise}} = 2a^2A^2B\left[1 + \frac{8}{\pi^2}\left(1 + \frac{1}{9} + \frac{1}{25} + \cdots\right)\right]$$

therefore

$$S_z(f_0)\bigg|_{\text{noise}} = 4a^2A^2B. \tag{40}$$

Since $B \gg f_0$, the noise power density of a frequency near f_0 is equal to that at f_0, and is thus equal to that given by (40). The ratio-meter also acts as a frequency converter in that the bandpass noise power densities are converted to frequencies near dc. For example, the noise power density at $f = f_0 + f_1$ and $f = f_0 - f_1$ (where $0 < f_1 < \Delta f/2$) are both converted to the frequency f_1. Since the noise power densities are equal and uncorrelated, the resultant noise power density at f_1 is doubled.

$$S_z(f_1)\bigg|_{\substack{\text{converted}\\\text{noise}}} = 2S_z(f_0)\bigg|_{\text{noise}} = 8a^2A^2B. \tag{41}$$

The total noise power in the converted band, when limited by the band-pass filter bandwidth, Δf, is the noise power density of (41) times the range of f_1, i.e., times $\Delta f/2$.

$$\text{Noise power (limited by } \Delta f) = 4a^2A^2B\Delta f.$$

In order to reduce the output noise further, the ratio-meter is followed by a narrow low-pass filter of bandwidth b (where $b < \Delta f/2$). In this case

$$\text{Noise power (limited by } b) = 8a^2A^2Bb$$

and the corresponding output noise voltage (rms value) is

$$e_n(\text{rms}) = 2\sqrt{2}\, a\, A\sqrt{Bb}. \tag{42}$$

The signal voltage at 1 kc (peak value) is $\sqrt{2}$ times the square root of the spectral density at 1 kc. Therefore, from (39),

$$e_S(\text{peak}) = \frac{4}{\pi}\, a\, AB. \tag{43}$$

and the corresponding change in signal voltage (peak value) is

$$\Delta e_S(\text{peak}) = \frac{4}{\pi}\, a\, B\Delta A. \tag{44}$$

The quantities required by (36) are given by (42), (43), and (44). However, the spectral density, A, at the input of the square-law detector is different for each channel and is proportional to the assumed input temperature; i.e.,

$$\text{if } A_2 = CT_S\,,$$

$$\text{then } A_1 = C(T_S + T_A);$$

$$\text{therefore } \Delta A_2 = \Delta A_1 = C\Delta T_S\,.$$

Referring now to (36), the Δe_S terms can be written

$$\begin{aligned} \frac{\Delta e_S}{e_{S2}} - \frac{\Delta e_S}{e_{S1}} = \frac{\Delta A}{A_2} - \frac{\Delta A}{A_1} &= \frac{C\Delta T_S}{CT_S} - \frac{C\Delta T_S}{C(T_S + T_A)} \\ &= \frac{\Delta T_S T_A}{(T_S + T_A)T_S}\,. \end{aligned} \tag{45}$$

Since the amplitude and phase of e_{n2} and e_{n1} are random and uncorrelated, the total rms fluctuation due to the noise terms can be found by adding the individual rms values on a power basis:

$$\begin{aligned}\frac{e_{n2}}{e_{S2}}\sin\theta_2 - \frac{e_{n1}}{e_{S1}}\cos\theta_1 &= \left\{\left[\frac{e_{n2}}{\sqrt{2}\,e_{S2}}\right]^2 + \left[\frac{e_{n1}}{\sqrt{2}\,e_{S1}}\right]^2\right\}^{\frac{1}{2}} \\ &= \left\{\left[\frac{e_{n2}(\text{rms})}{e_{S2}(\text{peak})}\right]^2 + \left[\frac{e_{n1}(\text{rms})}{e_{S1}(\text{peak})}\right]^2\right\}^{\frac{1}{2}} = \left\{\left[\frac{\pi}{\sqrt{2}}\sqrt{\frac{b}{B}}\right]^2 \right. \\ &\qquad \left. + \left[\frac{\pi}{\sqrt{2}}\sqrt{\frac{b}{B}}\right]^2\right\}^{\frac{1}{2}} = \pi\sqrt{\frac{b}{B}}.\end{aligned} \tag{46}$$

Inserting (45) and (46) into (36)

$$V_{\text{out}} = K\left\{4\varphi_0 + 2\sin 2\varphi_0\left[\frac{\Delta T_S T_A}{(T_S + T_A)T_S} + \pi\sqrt{\frac{b}{B}}\right]\right\} \tag{47}$$

where: $\varphi_0 = \tan^{-1}(e_{S2}/e_{S1}) = \tan^{-1}(1/Y)$.

The second-from-last term is due to a change, ΔT_S, of input temperature, and the last term is an rms variation due to noise fluctuations. The minimum detectable change of input temperature can be found by equating the last two terms. The result is

$$\Delta T_S(\text{theoretical}) = T_S\left(1 + \frac{T_S}{T_A}\right)\pi\sqrt{\frac{b}{B}}. \tag{48}$$

It can be shown that the noise bandwidth, b, of an RC low-pass filter is equal to $\frac{1}{4}$ RC $= \frac{1}{4}\tau$ where τ is the RC time constant. With this substitution (48) becomes

$$\Delta T_S(\text{theoretical}) = T_S\left(1 + \frac{T_S}{T_A}\right)\frac{\pi}{2}\frac{1}{\sqrt{B\tau}}. \tag{49}$$

APPENDIX B

Calibration

Since Y is the ratio of two noise powers, the output voltage of the ratiometer can be calibrated well in advance without using the antenna or maser preamplifier. To do this, the coaxial noise lamp is connected to the input of the IF converter via an RF level-set attenuator. Since the converter noise temperature is about 1350°K and the excess noise temperature of a coaxial noise lamp is about 8360°K, a value of $Y \approx 5.5$ can be obtained. By adjusting the level set attenuator this can be varied down to $Y \approx 1$. The resulting value of Y is measured precisely by noting the change of IF attenuation required to keep the IF output power constant when the noise lamp is turned on. By pulsing the lamp at a 1-kc rate, the corresponding ratio-meter output voltage can be measured with the precision dc voltmeter. Although the effective value of Y under pulsing conditions may be somewhat less due to ionization and deionization effects, the above method of calibration bypasses this as a source of error. The accuracy of the resulting Y versus V curve, Fig. 7, is limited by (*i*) the precision IF attenuator, to ±3.0 per cent, and (*ii*) the precision dc voltmeter, to ±1.0 per cent, for a subtotal of ±4.0 per cent. Since the T_S versus V curve, Fig. 7, is, in addition, a function of T_A, (2), and the accuracy of T_A, for $T_A = 94.6°K$, is limited by (*iii*) the noise lamp temperature, to ±2.7 per cent, and (*iv*) the directional coupling, to ±3.6 per cent, for a subtotal of ±6.3 per cent, the total accuracy of the absolute system temperature calibration is ±10.3 per cent.

REFERENCES

1. Ohm, E. A., Project Echo Receiving System, B.S.T.J., **40**, July, 1961, p. 1065.
2. DeGrasse, R. W., Hogg, D. C., Scovil, H. E. D., and Ohm, E. A., Ultra-Low-Noise Antenna and Receiver Combination for Satellite or Space Communication, Proc. Nat. Elec. Conf., **15**, Oct., 1959, p. 370.
3. Dicke, R. H., The Measurement of Thermal Radiation at Microwave Frequencies, Rev. Sci. Instr., **17**, July, 1946, p. 268.
4. Selove, W., A DC Comparison Radiometer, Rev. Sci. Instr., **25**, Feb., 1954, p. 120.
5. Crawford, A. B., Hogg, D. C., and Hunt, L. E., A Horn-Reflector Antenna for Space Communications, B.S.T.J., **40**, July, 1961, p. 1107.
6. DeGrasse, R. W., Kostelnick, J. J., and Scovil, H. E. D., The Dual Channel 2390-mc Traveling-Wave Maser, B.S.T.J., **40**, July, 1961, p. 1125.
7. 416A/AR Ratio Meter Operating and Servicing Manual, Hewlett-Packard Co., Nov., 1959, Section III.
8. Kuhn, N. J., and Negrete, M. R., Gas Discharge Noise Sources in Pulsed Operation, I.R.E. International Convention Record, Part 3, Mar., 1961, p. 166.
9. Hogg, D. C., Effective Antenna Temperatures Due to Oxygen and Water Vapor in the Atmosphere, J. Appl. Phys., **30**, Sept., 1959, p. 1417.
10. Davenport, W. B., Jr., and Root, W. L., *Random Signals and Noise*, McGraw-Hill, New York, 1958, p. 256.
11. Enloe, L. H., Sensitivity of a Noise-Adding Radiometer, private communication, Nov. 5, 1962.
12. Jakes, W. C., and Penzias. A. A., unpublished work.

The ALPAK System for Nonnumerical Algebra on a Digital Computer — I: Polynomials in Several Variables and Truncated Power Series with Polynomial Coefficients

By W. S. BROWN

(Manuscript received April 17, 1963)

This is the first of two papers on the ALPAK system for nonnumerical algebra on a digital computer. This paper is concerned with polynomials in several variables and truncated power series with polynomial coefficients. The second paper will discuss rational functions of several variables, truncated power series with rational-function coefficients, and systems of linear equations with rational-function coefficients. The ALPAK system has been programmed within the BE-SYS-4 monitor system on the IBM 7090 computer, but the language and concepts are machine independent.

The available polynomial arithmetic operations are add, subtract, multiply, divide (if divisible), substitute, differentiate, zero test, nonzero test, and equality test. The speed of the system is indicated by the rule of thumb that one man-hour equals one 7090-second. The available space in core is usually sufficient for approximately 8000 polynomial terms.

Section I of this paper consists of a nontechnical description of the system and a brief glimpse into the future. Section II discusses several specific problems to which the ALPAK system has been applied. These two parts do not presuppose any knowledge of computers or computer programming. Section III describes the use and the implementation of the algebraic operations relating to polynomials in several variables and truncated power series with polynomial coefficients. The reader of Section III is assumed to be acquainted with the elements of FAP (FORTRAN Assembly Program) programming, including the use of macros, as described in a series of IBM publications, the latest of which is IBM 7090-7094 Programming Systems MAP (Macro Assembly Program) Language (Form Number C28-6311).

TABLE OF CONTENTS

	Page
I. NONTECHNICAL DESCRIPTION	2082
1.1 *Introduction*	2082
1.2 *Choosing a Canonical Form*	2083
1.3 *Polynomial Arithmetic*	2085
1.4 *Rational Functions and the Future*	2087
II. APPLICATIONS	2087
2.1 *Introduction*	2087
2.2 *On the Zeros of Gaussian Noise*	2088
2.3 *A Queueing System with Priorities*	2089
2.4 *A Single-Server Queue with Feedback*	2090
2.5 *The Triskelion Diagram*	2092
2.6 *Wave Propagation in Crystals*	2093
III. USERS' MANUAL	2095
3.1 *Introduction*	2095
3.2 *Input-Output*	2097
3.3 *Polynomial Arithmetic*	2104
3.4 *Truncated Power Series*	2111
3.5 *The Main Program*	2112
3.6 *Loading Instructions*	2115
3.7 *Diagnostics*	2116
3.8 *Debugging*	2118
IV. ACKNOWLEDGMENTS	2119

I. NONTECHNICAL DESCRIPTION

1.1 *Introduction*

Many theoreticians devote a substantial portion of their time to the routine manipulation of algebraic expressions. It has long been recognized that digital computers are capable in principle of easing this burden. The ALPAK system, which is described herein and in a subsequent paper and has been programmed for the IBM 7090 computer, represents a significant start toward the practical implementation of that capability. It performs a limited set of operations — add, subtract, multiply, divide, substitute, differentiate, zero test, nonzero test, and equality test — on a limited class of expressions: rational functions of several variables and truncated power series with rational-function coefficients. It can also solve (by Gaussian elimination) systems of linear equations with rational-function coefficients. This paper is concerned with polynomials in several variables and truncated power series with polynomial coefficients. The generalizations indicated above will be discussed in a separate paper by B. A. Tague, J. P. Hyde, and the present author.

The ALPAK system is not a "sophomore imitator" or "elementary mathematics system." There are many elementary mathematical operations (e.g., the proving of trigonometric identities) which are beyond its present capabilities. However, when faced with problems within its range of capability, its speed (one man-hour $\approx$ one 7090-second) and power (the available space in core is usually sufficient for approximately 8000 polynomial terms) are impressive.

Neither is the ALPAK system a "symbol manipulation system," because it views a polynomial as an array of coefficients and exponents rather than as a string of numbers, variable names, operation symbols, parentheses, and the like. This is the key to speed and power. Polynomials are stored in a nearly optimal manner, and polynomial operations are reduced to their essentials.

We have been speaking of polynomials and rational functions without being specific about the possible coefficient rings. The coefficients may be integers or they may be elements of any other integral domain for which arithmetic and input-output facilities are available. All operations on coefficients are performed by a small set of macros (user-defined instructions which expand into one or more machine instructions). These are currently defined for integers, but the user may redefine them to suit his own needs. (Of course this requires reassembly of the ALPAK subroutines.) The use of floating-point coefficients is not in keeping with the spirit of symbolic computing and should be avoided if possible. The occurrence of roundoff error causes zero to be nonunique and gives rise to a host of difficult problems which the author has not attempted to solve. It is usually feasible and desirable to replace the nonrational numbers which occur in an expression by literal symbols. These can be treated by the ALPAK system as variables. The result will then involve no roundoff error, and the dependence on these symbols will be explicitly displayed.

To maximize speed and minimize space, the coefficients and exponents of a polynomial are stored in a contiguous block, and the exponents are packed as specified in a user-provided format statement. The names of the variables are kept in the format statement and are referred to as infrequently as possible. Storage allocation is automatic and dynamic, so that the programmer can refer to a polynomial by name without worrying about its size, structure, or location.

In Sections 1.2 and 1.3 we shall discuss the canonical form for polynomials and the implementation of the various polynomial-arithmetic operations. Section 1.4 contains a very brief preview of the rational-function operations and an even briefer mention of some of our hopes for the future.

1.2 *Choosing a Canonical Form*

In the ALPAK system *every* polynomial in storage is *always* kept in a unique canonical form, which we shall describe. Every subroutine, except the input and output subroutines, assumes that its inputs are in canonical form and produces its outputs (if any) in canonical form. On input a polynomial is put into canonical form, and on output it is left in whatever form it is found. Barring trouble, this will always be canonical form.

It is important to recognize that the "best canonical form" for a given class of expressions need not be an approximation to what human beings would call the "simplest form." In fact, the two concepts are in some respects opposite. The simplest form may be defined roughly as "that form which requires the smallest number of symbols." On the other hand, an approximate definition of the best form is "that form into which the general expression of the class can most easily be put." This latter definition clearly favors canonical forms in which expressions are expanded over those in which they are collapsed, because the collapsing of expressions tends to be difficult, while their expansion tends to be easy. For example, in the case of polynomials in several variables we must choose between an "expanded form" in which each polynomial is represented as an ordered sum of terms and a "factored form" in which each polynomial is represented as an ordered product of irreducible factors. In general, the factored form is more compact, but we must reject it because the factoring algorithm† can be extremely time consuming, while the expansion of a factored polynomial into a sum of terms is always simple and fast.

Now a polynomial in n variables can be viewed as a finite n-dimensional array of coefficients. If a majority of them are zero, it is advantageous to represent the polynomial as a list of the nonzero ones together with their coordinate labels (i.e., their exponents). Otherwise, it is preferable to use the entire array. In many practical cases the number of variables and the maximum exponent sizes are all of the order of 10, so an array size as large as 10^{10} would not be unusual. However, it is difficult to imagine a practical case involving more than a few hundred (or conceivably a few thousand) nonzero terms. For generality we are therefore obliged to represent each polynomial as an ordered list of its nonzero terms. It is convenient to order the terms according to the magnitude of the first exponent, and to order those terms having the same first exponent according to the magnitude of the second, etc. The order of the variables is the order in which they appear in the format statement.

† See exercise 15 on page 82 of Ref. 1.

1.3 *Polynomial Arithmetic*

In this section we shall discuss the implementation of the various polynomial-arithmetic operations. Let us begin with a simple illustration of their use. Suppose polynomials A, B, C, and D are in storage, and C is thought to be a divisor of $A*B$. (The asterisk denotes multiplication.) To compute and print

$$F = \frac{A*B}{C} + D \tag{1}$$

we write†

```
POLMPY    F,A,B
POLDIV    F,F,C,NODIV
                              (2)
POLADD    F,F,D
POLPRT    F
```

The first line replaces F by $A*B$. The second replaces F by F/C; that is, by

$$\frac{A*B}{C} \tag{3}$$

This illustrates the fact that an output may overwrite an input. The third line replaces F by $F + D$; that is, by

$$\frac{A*B}{C} + D \tag{4}$$

which is the desired result. Finally, the fourth line causes this result to be printed on the output tape. If the division in the second line is unsuccessful, i.e., if C is not a divisor of $A*B$, control will be transferred to the location called NODIV.

A polynomial is represented on data cards as a sequence of coefficients and exponents, each coefficient being followed by its exponents. It is terminated by the appearance of a zero where a coefficient would other wise be expected. For example the polynomial

$$3x^2 + 2xyz - 5yz^2 \tag{5}$$

might appear as

† Note the similarity to the arithmetic orders of a three-address computer. The prefix "POL" stands for "polynomial."

$$\begin{array}{rl} 3 & 2,0,0 \\ 2 & 1,1,1 \\ -5 & 0,1,2 \\ 0 & \end{array} \tag{6}$$

We have chosen this type of representation primarily because of its appeal to the computer. However, for large polynomials it is also an unexpectedly appealing form for people. On several occasions we have observed geometrical patterns in the computer output which would not be apparent in a conventional human transcription.

The addition of two polynomials in canonical form is analogous to the ordered merging of two ordered subdecks of a deck of playing cards, except that the addition subroutine must also be on the lookout for combinations and cancellations.

The multiplication of a polynomial by a nonzero monomial does not disturb canonical form. When two polynomials are to be multiplied, the longer one is multiplied by each term of the shorter one, and each of these products is added to the sum of all the preceding ones.

The polynomial division subroutine is successful only when the dividend is exactly divisible by the divisor. However, it is programmed so that it can be used as a test for divisibility if that is desired. The divisor and dividend are treated as polynomials in one variable with coefficients in the ring of polynomials in all the remaining variables. Divisions in this ring can be handled by the division subroutine itself,† and the main task is carried out by the familiar process of "long division."

The polynomial substitution subroutine works in the most straightforward possible way — substituting into one term at a time and preserving only the latest partial result. This procedure may involve substantial duplication of effort, but it uses a minimum of working space and a minimum of program, and in most practical cases the running time is reasonable.

The polynomial differentiation subroutine differentiates term by term with respect to a specified variable. It is perhaps worth remarking that this process does not upset the canonical ordering.

A truncated power series with polynomial coefficients can be treated as a polynomial, except that it is necessary to keep track of the order

† A subroutine which calls itself is called "recursive." At the innermost level it must, of course, operate by an independent mechanism. Collisions between the different levels are prevented by saving necessary information in a push-down list. It is perhaps worth remarking that every inductive algorithm can be programmed as a recursive subroutine. In the case of polynomial division the induction is on the number of variables, and the innermost level is simply coefficient division.

and to prevent the appearance of meaningless higher-order terms. The ALPAK system contains only two orders ("truncate" and "multiply and truncate") for dealing with truncated power series. These are sufficient for many applications, but much remains to be done.

1.4 *Rational Functions and the Future*

Every rational function can be represented as the quotient of two polynomials. The extension from polynomial operations to rational-function operations would be trivial except for the problem of removing all common factors from the numerator and denominator of each rational function. This has been accomplished by means of a generalized version of Euclid's greatest-common-divisor algorithm. However, we must caution the reader that Euclid's algorithm is extremely explosive, and the computer will not be able to handle rational functions with numerators and denominators of high degree in many variables until more sophisticated techniques are developed.

Aside from the difficulties mentioned above, the handling of truncated power series with rational-function coefficients and the solution by Gaussian elimination of systems of linear equations with rational-function coefficients are straightforward.

One of the primary problems encountered in the development of the ALPAK system is the problem of automatic dynamic storage allocation. Usually the inputs to a subroutine are polynomials of arbitrary size, and in general the required working space could not be predicted even if the sizes of the inputs were known. Therefore it is imperative to be able to obtain blocks of space as needed and to return idle space to the system. Our storage allocator provides these services in a manner suitable to our current needs, but it is not general or elegant. A general purpose storage allocation system including tracing and other service routines has been developed by Miss D. C. Leagus and the author, and will be described in a forthcoming paper. With this as a foundation, we hope to write a faster and more powerful version of the present ALPAK system, and perhaps to extend it into other areas of mathematics.

II. APPLICATIONS

2.1 *Introduction*

This section is devoted to a few general remarks about the usefulness of symbolic computing. The skeptic will protest that any symbolic calculation too long to be done with pencil and paper is not really worth doing. This sentiment might be expressed in the form of the question, "Who wants to look at a polynomial ten pages long?" The objection is not without merit, but it is worth recalling that similar objections were once raised in connection with numerical calculations. Furthermore it is unmistakably clear that mathematical analyses arising in many different contexts involve substantial amounts of routine algebra which could be done faster and more reliably by a computer. What, then, are the types of problems to which symbolic computing facilities are likely to be applicable?

It often happens that a "straightforward calculation" whose end result is concise and understandable involves many tedious manipulations of lengthy expressions at intermediate stages. Sometimes the end result can also be reached by a shorter route, but the result itself (and the knowledge that it is indeed concise and understandable) may play a decisive role in the discovery of that route.

If the desired output of a calculation is numerical or graphical, it may nevertheless be advantageous (or even essential) to begin the calculation symbolically and allow a numerical program to take over only during the final stages. The problem of error analysis will not arise until these final stages are reached.

A third type of application arises when a simple calculation must be repeated many times with only minor variations, e.g., for all possible values of some set of indices.

Other types of applications may possibly occur to the reader. In the next five sections we shall discuss specific problems to which the ALPAK system has been applied.

2.2 *On the Zeros of Gaussian Noise*

Our first significant test problem arose in a study by D. Slepian[2] of the distribution of zeros of Gaussian noise. It was desired to find the leading term in the power series expansion of the determinant

$$\begin{vmatrix} \rho(ut - vt) & \rho(vt) & \rho(t - vt) & -\rho'(vt) & \rho'(t - vt) \\ \rho(ut) & 1 & \rho(t) & 0 & \rho'(t) \\ \rho(t - ut) & \rho(t) & 1 & -\rho'(t) & 0 \\ -\rho'(ut) & 0 & -\rho'(t) & 1 & -\rho''(t) \\ \rho'(t - ut) & \rho'(t) & 0 & -\rho''(t) & 1 \end{vmatrix} \tag{7}$$

where

$$\rho(t) = 1 - \frac{t^2}{2!} + \frac{at^3}{3!} + \frac{bt^4}{4!} + \frac{ct^5}{5!} + \frac{dt^6}{6!} + \frac{et^7}{7!} + \frac{ft^8}{8!} + \cdots. \tag{8}$$

The algebra is difficult not only because of the order of the determinant, but also because the leading term corresponds to an unexpectedly high power of t. In the general case, $a \neq 0$, the leading term is

$$\frac{a^3t^7}{9} v^2(1-u)^2(3u-v-2uv). \tag{9}$$

When $a = 0$ but $c \neq 0$, it is

$$\begin{aligned}\frac{2(1-b)c^2t^{12}}{(5!)^2} v^2(1-u)^2[2v^3(2u^3-u^2-4u-2) \\ -5uv^2(2u^2-u-4)+5u^2(2u-3)].\end{aligned} \tag{10}$$

Finally, when $a = 0$ and $c = 0$, it is

$$\frac{t^{16}}{144(4!)^2}(b^2+d)(b^3+d^2+f+2bd-bf)u^2v^2(1-u)^2(1-v)^2. \tag{11}$$

These results were obtained by a program written in the ALPAK language by Mrs. W. L. Mammel. Although approximately 2000 polynomial terms were in storage at the floodcrest of the computation, the computing time for all three cases was only 92 seconds.

2.3 *A Queueing System with Priorities*

Another interesting problem arose in a study by J. P. Runyon[3] of a queueing system in which a group of servers handles traffic from two sources, one of which is preferred over the other. It is desired to solve the functional difference equation

$$\begin{aligned}(\alpha - x)(\beta-\alpha)^{n-1}g_n(x) \\ = \alpha(\beta-x)^n g_{n-1}(\alpha) - x(\beta-\alpha)^n g_{n-1}(x) \qquad n \geqq 1\end{aligned} \tag{12}$$

where $g_0(x) = 1$, and $0 < \alpha < \beta$. It follows by induction that for $n \geqq 1$, $g_n(x)$ is a polynomial of degree $(n-1)$ in x, whose coefficients are polynomials in α and β. The value of $g_n(\alpha)$ is of particular interest. By the time this author was ready to attack the problem, Runyon had conjectured and J. A. Morrison[4] had proved that

$$g_n(\alpha) = \sum_{r=0}^{n-1}\binom{n-1}{r}\binom{n}{r}\frac{\beta^{n-r}\alpha^r}{n+1}. \tag{13}$$

Nevertheless, a short program was written to compute as many as possible of the polynomials $g_n(x)$ and the corresponding $g_n(\alpha)$. The program stopped after $87\frac{1}{2}$ seconds because of a coefficient overflow† during the

† The largest allowed coefficient is $2^{35} - 1$.

calculation of $g_{16}(x)$. The polynomial $g_{15}(x)$ has 197 terms and a maximum coefficient of several billion. If the program had been available sooner, it would have spared Runyon the necessity of calculating the first five of the $g_n(x)$ by hand.

2.4 *A Single-Server Queue with Feedback*

Another problem from queueing theory arose in a study by L. Takács[5] of a single-server queueing system with "feedback." The input is a Poisson process of density λ, the service times are determined by a distribution function with moments α_k, and after being served a customer rejoins the queue with probability p or departs with probability $q = 1 - p$.

It is shown by Takács that the rth moment of the total time spent in the system is

$$\beta_r = (-1)^r \Phi_{r0} \tag{14}$$

where

$$\Phi_{ij} = \left[\left(\frac{\partial}{\partial s} \right)^i \left(\frac{\partial}{\partial t} \right)^j \Phi(s,t) \right]_{s=0,\, t=0} . \tag{15}$$

The function $\Phi(s,t)$ is implicitly defined by the equation

$$\Phi(s,t) = (q - \lambda\alpha_1) W(s,t) + p\psi(s + \lambda t)\Phi(s,\omega(s,t)) \tag{16}$$

where†

$$\begin{aligned} \psi(s) &= \sum_{i=0}^{\infty} \frac{(-1)^r \alpha_r s^r}{r!} \\ \omega(s,t) &= 1 - (1 - pt)\psi(s + \lambda t) \\ W(s,t) &= \psi(s + \lambda t) + S(s + \lambda t, \lambda\omega(s,t)) T(\omega(s,t)) \end{aligned} \tag{17}$$

with

$$\begin{aligned} S(x,y) &= \frac{\psi(x) - \psi(y)}{x - y} \\ T(\omega) &= \frac{\lambda\omega(1 - \omega)}{1 - \omega - (1 - p\omega)\psi(\lambda\omega)} . \end{aligned} \tag{18}$$

This last pair of equations can be rewritten in the more useful form

† For convenience we have assumed that all of the service moments α_r are finite. However, for the calculation of β_r it is clearly sufficient to require only the finiteness of α_{r+1} .

$$\begin{aligned} S(x,y) &= \sum_{r=0}^{\infty} \frac{(-1)^{r+1}\alpha_{r+1}}{(r+1)!} C_r(x,y) \\ T(\omega) &= \frac{-\lambda(1-\omega)}{q - \lambda(1-p\omega)\varphi(\lambda\omega)} \end{aligned} \tag{19}$$

where

$$\begin{aligned} C_r(x,y) &= \frac{x^{r+1} - y^{r+1}}{x - y} = \sum_{k=0}^{r} x^k y^{r-k} \\ \varphi(x) &= \frac{1 - \psi(x)}{x} = \sum_{r=0}^{\infty} \frac{(-1)^r \alpha_{r+1} x^r}{(r+1)!} \,. \end{aligned} \tag{20}$$

It is now clear that

$$\begin{aligned} \psi(0) &= \alpha_0 = 1 \\ S(0,0) &= -\alpha_1 \\ T(0) &= \frac{-\lambda}{q - \lambda\alpha_1} \\ W(0,0) &= \frac{q}{q - \lambda\alpha_1} \end{aligned} \tag{21}$$

so from (14)–(16)

$$\beta_0 = \Phi_{00} = \Phi(0,0) = 1 \tag{22}$$

as is required by the definition of the zeroth moment.

Now suppose all of the quantities Φ_{ij} for $i + j < r$, where r is some positive integer, have been calculated and are expressed as rational functions of λ and p (or q) and the service moments α_k. Then by differentiation of (16) we can obtain a system of $r + 1$ linear equations in the $r + 1$ unknowns, Φ_{ij} with $i + j = r$. These equations will also contain the quantities Φ_{ij} with $i + j < r$, which can be replaced by their known values. The solutions of this linear system will again be rational functions of λ and p (or q) and the service moments α_k. Theoretically, this procedure permits the calculation of arbitrarily many of the moments, but in practice the calculations are extremely lengthy.

The first moment can be calculated by hand, with the result

$$\beta_1 = \alpha_1 \left(\frac{1 - \lambda\alpha_1}{q - \lambda\alpha_1}\right) + \frac{\lambda\alpha_2}{2(q - \lambda\alpha_1)} \,. \tag{23}$$

The second moment was calculated with the aid of an IBM 7090 computer and the ALPAK system. The intermediate expressions are extremely lengthy, but the final result is the relatively compact expression

$$\beta_2 = \frac{(2qF - G)(q^2 - 2q)}{6(q - \lambda\alpha_1)^2[q^2 - q(\lambda\alpha_1 + 2) + \lambda\alpha_1]} \tag{24}$$

where

$$\begin{aligned} F &= 6\lambda\alpha_1^3 - 6\alpha_1^2 + 6\lambda\alpha_1\alpha_2 + 3\alpha_2 + \lambda\alpha_3 \\ G &= 12\lambda\alpha_1^3 - 12\alpha_1^2 - 6\lambda\alpha_1\alpha_2 + 2\lambda^2\alpha_1\alpha_3 - 3\lambda^2\alpha_2^2. \end{aligned} \tag{25}$$

For a more detailed discussion of this calculation, see the appendix in Ref. 5.

2.5 *The Triskelion Diagram*

The problem to be considered in this section arose in a study by D. B. Fairlie and the author[7,8] of the analyticity properties of the Feynman amplitudes corresponding to several simple vertex diagrams in quantum field theory. One of these is the triskelion diagram, which is shown in Fig. 1. Here the p's and q's are vectors in space-time, and

$$\begin{aligned} z_i &= p_i^2 \\ a_i &= q_i^2 \\ b_i &= (p_i - q_i)^2 \end{aligned} \tag{26}$$

for $i = 1, 2, 3$. The corresponding Feynman amplitude is the boundary value of an analytic function $H(a,b,z)$ of these nine variables, analytic

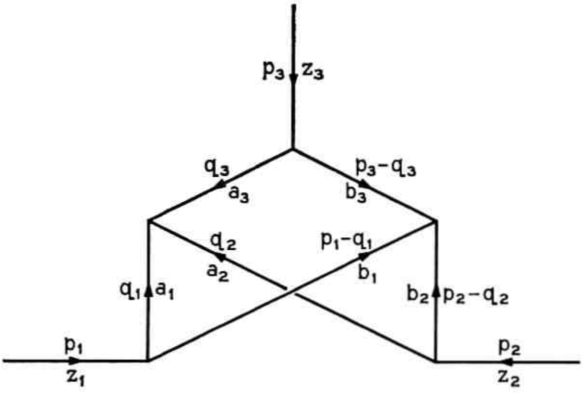


Fig. 1 — The triskelion diagram.

everywhere except on certain manifolds which can be obtained from lower-order "contracted" diagrams, and on the manifold

$$\Psi(a,b,z) \equiv 4D^2(4D + A^2) + 4AB^2(9D + 2A^2) - 27B^4 = 0 \quad (27)$$

where

$$\begin{aligned} A &= \tfrac{1}{8}\lambda(z + a + b) - \tfrac{1}{4}[\lambda(z) + \lambda(a) + \lambda(b)] \\ B &= -\tfrac{1}{4} \det (z,a,b) \\ D &= -\tfrac{1}{8} \sum_{i=1}^{3} \{z_i^2(a_jb_k + a_kb_j) + z_jz_k(a_iB_i + b_iA_i) \\ &\quad + z_i[2a_ib_jb_k + a_jb_kB_k + a_kb_jB_j + 2b_ia_ja_k + b_ja_kA_k + b_ka_jA_j]\}. \end{aligned} \quad (28)$$

Here (i,j,k) is a cyclic permutation of $(1,2,3)$ and

$$\begin{aligned} A_i &= a_i - a_j - a_k \\ B_i &= b_i - b_j - b_k \\ \lambda(x) &= x_1^2 + x_2^2 + x_3^2 - 2x_1x_2 - 2x_1x_3 - 2x_2x_3 . \end{aligned} \quad (29)$$

It is shown in Ref. 8 that Ψ is a homogeneous twelfth-degree polynomial in its nine arguments, and is irreducible over the rationals. Furthermore, it is invariant under permutations of the indices, 1,2,3, permutation of the vectors, a,b,z, and transposition of the matrix of these vectors.

It is natural to ask whether the substitution of (28) and (29) into (27) yields a compact expression or an unwieldy monstrosity. A short program was written to perform the substitutions, but it stopped at an early stage because of insufficient space. However, the polynomial $\Psi(a_1,a_2,a_3;\ b_1,b_2,b_3;\ 0,0,z_3)$ was easily computed (in 50 seconds) and was found to have 2642 terms. Since $\Psi(a,b,z)$ contains all of these terms and many more, we can safely assume that (27) is the most useful way of writing it.

2.6 *Wave Propagation in Crystals*

The problem to be considered in this section arose in a study by R. N. Thurston[9] of wave propagation in crystals under pressure. It is of particular interest to investigate the effect of pressure on propagation velocity. For given temperature T, pressure p (hydrostatic or uniaxial), and propagation direction N (a unit vector), there are in general three modes of propagation, corresponding to three displacement directions which are mutually perpendicular if $p = 0$. In simple cases one of these modes is longitudinal and the other two are transverse. For a given mode, let $V(p,T)$ be the propagation velocity and let ρ_0 be the crystal density at $p = 0$. Then define

$$S' \equiv \rho_0 \frac{\partial}{\partial p} [V^2(p,T)]_{p=0} . \tag{30}$$

It can be shown that

$$S' = U_p U_q D_{pq} \tag{31}$$

(summation convention understood), where U is a unit vector in the direction of particle displacement in the given mode, and where

$$D_{pq} = N_j N_k F_{st} [\delta_{pq} C_{jkst}{}^T + 2\delta_{qs} C_{pjtk}{}^S + 2N_s N_t C_{pjqk}{}^S + C_{pjqkst}]. \tag{32}$$

The C's are elastic constants at zero pressure. The six-index C array has 3^6 entries of which at most 56 are distinct, while each four-index C array has 3^4 entries of which at most 21 are distinct. F_{st} is a symmetric matrix whose entries are rational functions of these elastic constants (and of the direction of pressure in the uniaxial case). Our task is to perform the indicated summations in special cases to get explicit expressions for S'.

The complete analysis for the case of cubic crystals is given in Ref. 9. A program has been written by J. P. Hyde (using the ALPAK system) to evaluate S' and serve as a check for this analysis. In the cubic case, the six-index C array has only six distinct nonzero elements, which are abbreviated as C_{111}, C_{112}, C_{144}, C_{166}, C_{123}, and C_{456}. The four-index C^T array has only three distinct nonzero elements, abbreviated as $C_{11}{}^T$, $C_{12}{}^T$, and C_{44}, and the four-index C^S array has only three distinct nonzero elements, abbreviated as $C_{11}{}^S$, $C_{12}{}^S$, and C_{44}. Note that C_{44} appears in both arrays. The results for the case of hydrostatic pressure and wave propagation along (1,1,0) are as follows: For longitudinal displacement along (1,1,0)

$$S' = 2C_{11}{}^S + 2C_{12}{}^S + 4C_{44} + \tfrac{1}{2}C_{111} + 2C_{112} + C_{144} + 2C_{166} + \tfrac{1}{2}C_{123} . \tag{33}$$

For transverse displacement along (1,−1,0)

$$S' = 2C_{11}{}^S - 2C_{12}{}^S + \tfrac{1}{2}C_{111} - \tfrac{1}{2}C_{123} . \tag{34}$$

And for transverse displacement along (0,0,1)

$$S' = 4C_{44} + C_{144} + 2C_{166} . \tag{35}$$

The computing time to obtain these results was approximately 20 seconds.

A modified version of this program would make possible the corresponding calculations for crystals of lower symmetry, including quartz.

III. USERS' MANUAL

3.1 *Introduction*

This section consists of a brief outline of Section III and a discussion of several basic concepts. The polynomial input-output and arithmetic operations are discussed in Sections 3.2 and 3.3, respectively. Section 3.4 consists of a brief introduction to the theory of truncated power series and a description of the orders for dealing with them. In Section 3.5 the rules for writing main programs (including those governing the use of POLBEG and VARTYP) are described, and two sample programs are presented. Loading instructions for assembly and/or run are given in Section 3.6. Finally, the dumping facilities and diagnostics are described in Section 3.7, and hints for debugging are given in Section 3.8.

3.1.1 *A Polynomial in Core*

A nonconstant polynomial† in core consists of a pointer, a heading, a data block, and a format statement (see Fig. 2). The pointer is a single word containing the heading address. The heading is a three-word block containing the data address, the format address, and the number of terms. The data block contains the terms, stored consecutively in a manner determined by the format statement. The format statement contains the names of the variables and the maximum exponent size in bits associated with each. The name of a polynomial is ordinarily used for the symbolic address of its pointer, and the name of a format statement for its symbolic address.

3.1.2 *Format Compatibility*

A format statement is usually shared by many polynomials. In fact two polynomials cannot be added, subtracted, multiplied, or divided unless they share the same format statement.

3.1.3 *More Than One Pointer to a Heading*

If two or more polynomials are equal, their pointers may point to a common heading. This is especially convenient when arrays of polynomials with many equal elements must be dealt with, but the user must keep in mind that if one of the polynomials is changed the others will be changed in the same way.

† A constant polynomial has only a pointer and a heading. Its value is kept in the heading (see Section 3.2.7), and no format is needed.

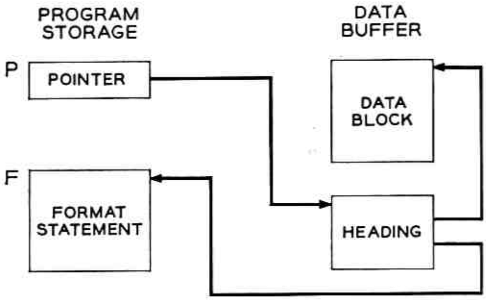


Fig. 2 — A polynomial P with format F.

3.1.4 *Storage Allocation*

Space for headings and data blocks is provided by the storage allocator. Headings are never moved, but the storage allocator is free to move data blocks as necessary.

Space for pointers and format statements must be provided by the user. Each pointer must be a full word, but only its address field is used. This must initially contain zero and will be filled in by the system. The prefix, tag, and decrement fields will be cleared. When a polynomial is read or computed its pointer is tested. If the address field of the pointer contains zero, a heading is created and the pointer is filled in with the heading address. Otherwise it is assumed that the pointer contains the address of a heading which can be overwritten. The data block (if any) previously attached to that heading is left "headless" and thereby becomes "garbage."

3.1.5 *Macros and Subroutines*

The polynomial portion of the ALPAK system consists of a macro deck and two subroutine packages, ALPAK1 and ALPAK2. ALPAK1 consists of input, output, and service subroutines, while ALPAK2 contains the operating subroutines. Together the two packages occupy less than 5000_{10} words of memory. Most of the macros expand into calling sequences for subroutines of the same name. For example the macro

$$\text{POLADD} \qquad \text{R,P,Q} \tag{36}$$

which is represented by the equation

$$R = P + Q \tag{37}$$

("replace *R* by $P + Q$"), expands to

```
TSX         POLADD,4
PZE         R
PZE         P                 (38)
PZE         Q
```

Here P, Q, and R are the symbolic addresses of pointers. When POLADD is executed, the *P* and *Q* pointers must contain the addresses of polynomial headings. The address field of the *R* pointer may contain the address of a heading to be overwritten or it may contain zero. In the latter case, a new heading will be created by the storage allocator and the *R* pointer will be filled in with its address. In either case, a data block for the sum of the polynomials *P* and *Q* will be obtained from the storage allocator and attached to the *R* heading, and the sum will be computed therein.

3.1.6 *Indexing*

This method of communication gives us a natural way of handling indexed arrays of polynomials. For example the set of polynomials

$$R_i = P_i + Q_i \, ; \qquad i = 1, \cdots , n \tag{39}$$

can be computed by writing

```
POLADD      (R,1)(P,1)(Q,1)                    (40)
```

inside a suitable loop (see Section 3.5), where index register 1 corresponds to the index, *i*. The expansion of this macro is simply

```
TSX         POLADD,4
PZE         R,1
PZE         P,1               (41)
PZE         Q,1
```

Clearly, index register 4 cannot be used for this type of indexing, because it has been reserved for the subroutine linkage.

3.2 *Input-Output*

3.2.1 *Summary* (*See Descriptions Section* 3.2.2)

	POLRDF	F	read format	(a)	
F	POLCVF	(X,15,Y,21,Z,36)	convert format	(b)	
	POLRDD	P,F	read data	(c)	
	POLCVD	P,F,H	convert data	(d)	
	POLCLR	P	clear	(e)	
	POLSTZ	P	store zero	(f)	
	POLSTI	P	store identity	(g)	(42)
	POLSTC	P,C	store constant	(h)	
	POLSTV	P,X,F	store variable	(i)	
	POLPRT	P,CC,(NAME)	print	(j)	
	POLPCH	P,(NAME)	punch	(k)	
	POLPRP	P,CC,(NAME)	print and punch	(l)	
	POLRDP	P,F,CC,(NAME)	read and print	(m)	
	POLCVP	P,F,H,CC,(NAME)	convert and print	(n)	

C = constant (symbolic address of constant)
CC = control character for printer
F = format (symbolic address of/for format statement)
H = Hollerith data (symbolic address of data)
NAME = alternative name for polynomial (not exceeding 21 characters)
P = polynomial (symbolic address of pointer)
X = variable (specified in the manner indicated by the last previous VARTYP declaration — see Section 3.5.2).

3.2.2 *Descriptions* (*See Also Sections* 3.2.3–3.2.8)

(a) POLRDF F

Read a polynomial format statement from cards into a block starting at location F. The length of this block must be at least $(2 + 2v + e)$ words where v is the number of variables and e is the number of exponent words per term.

(b) F POLCVF (X,15,Y,21,Z,36)

Assemble the parenthesized polynomial format statement and assign the symbol F to its first location. (F is a location-field argument of the macro.)

(c) POLRDD P,F

Read the polynomial P from cards according to the format F and put P into canonical form. Here, P is the address of a "pointer" for the polynomial, and F is the address of a format statement.

(d) POLCVD P,F,H

Same as POLRDD except that the data is to be found in core in a block of not more than 12 words of binary-coded information (BCI) starting at location H.

(e) POLCLR P

Clear the polynomial *P*.

(f) POLSTZ P

Set *P* equal to zero.

(g) POLSTI P

Set *P* equal to one.

(h) POLSTC P,C

Set *P* equal to the constant C.

(i) POLSTV P,X,F

Set *P* equal to the variable *X* using the format F.

(j) POLPRT P,CC,(NAME)

Print the polynomial *P* using CC for the control character for the first line of print and NAME (not more than 21 characters of BCI) for the name. If NAME is not provided *P* will be used for the name, and if CC is not provided a minus (triple space) will be used for the control character.

(k) POLPCH P,(NAME)

Punch the polynomial *P* on cards using NAME (not more than 21 characters of BCI) for the name. If NAME is not provided, P will be used for the name.

(l) POLPRP P,CC,(NAME)

Same as POLPRT followed by POLPCH.

(m) POLRDP P,F,CC,(NAME)

Same as POLRDD followed by POLPRT.

(n) POLCVP P,F,H,CC,(NAME)

Same as POLCVD followed by POLPRT.

3.2.3 *Polynomial on Cards*

A polynomial is represented on data cards as a sequence of coefficients and exponents separated by blanks and/or commas, each coefficient being followed by its exponents. It is terminated by the appearance of a zero where a coefficient would otherwise be expected. It is customary to use one card for each term and one as an end card. For example, the polynomial

$$3x^2 + 2xyz - 5yz^2 \tag{43}$$

is usually represented as

$$\begin{array}{rl} 3 & 2,0,0 \\ 2 & 1,1,1 \\ -5 & 0,1,2 \\ 0 & \end{array} \tag{44}$$

or

$$\begin{array}{rl} 3 & 2\ 0\ 0 \\ 2 & 1\ 1\ 1 \\ -5 & 0\ 1\ 2 \\ 0 & \end{array} \tag{45}$$

However, it is equally correct to put more than one term on a card

$$\begin{array}{rl} 3,2,0,0 & 2,1,1,1 \\ -5,0,1,2 & 0 \end{array} \tag{46}$$

or to use more than one card for a term

$$\begin{array}{rl} 3 & 2 \\ & 0,0 \\ 2 & 1 \\ & 1,1 \\ -5 & 0 \\ & 1,2 \\ 0 & \end{array} \tag{47}$$

If two commas are adjacent or separated only by blanks, a zero is understood. Similarly if the first (last) character on a card is a comma, a preceding (succeeding) zero is understood. Thus (43) can be represented as

$$\begin{array}{rl} 3 & 2,, \\ 2 & 1,1,1 \\ -5 & 0,1,2, \end{array} \tag{48}$$

or

$$3,2,,,2,1,1,1,-5,,1,2, \tag{49}$$

If identifying comments are desired, they may be printed on the last card, after the blank or comma which terminates the conversion of the final zero, and/or in columns 73–80 of any card.

The data is read from cards, converted, packed into the data buffer, and put into canonical form by the subroutine POLRDD (read data). The manner of packing is determined by a format statement which must be read first. If the polynomial has k variables, the first number in the data sequence is interpreted as a coefficient and the next k numbers are interpreted as exponents. This process is repeated until a zero appears in the position of a coefficient. The reading is then terminated, and the subroutine POLCFM (canonical form) is called to put the polynomial into canonical form.

3.2.4 *Format Statements*

Before discussing the operation of POLCFM it will be necessary to consider in detail the format statements and the representation of polynomials in core. A format statement on card(s) is an alternating sequence of variable names and field widths, starting in column 1 and separated by commas. Each field width must be a positive integer not greater than 36. It is the maximum exponent size in bits of the corresponding variable. Each variable name must be a string of not more than six characters (usually a FAP symbol) containing neither blanks nor commas. It is legal to skip to the next card after any comma, and this makes it possible to use as many continuation cards as necessary. The format statement is terminated by a blank immediately following a field width. Each field width specifies the number of bits to be reserved in each term for the exponent of the corresponding variable, and thereby determines the maximum allowable exponent for that variable. As an example, the format statement

$$\text{X,15,Y,21,Z,36} \tag{50}$$

specifies three variables, X, Y and Z, with field widths of 15, 21 and 36 respectively. This means that the maximum exponent sizes are $2^{15} - 1$,

$2^{21} - 1$ and $2^{36} - 1$ respectively. The sum of the field widths must be an integral multiple of 36, and each smaller multiple (if any) must be included among the partial sums. The card(s) is (are) read by the subroutine POLRDF (read format), which stores the format statement in a block *provided by the user*. POLRDF also counts v, the number of variables, computes e, the number of exponent words per term (the sum of the field widths divided by 36), and constructs a mask for use in exponent addition (see Section 3.3). The mask is a block of e words partitioned into v bit fields as indicated by the format statement with a one at the right end of each bit field. These items are stored as part of the format statement, whose length is $2 + 2v + e$ words. For example the internal format statement (in octal) corresponding to (50) is

000000000002	2 exponent words per term	
000000000003	3 variables	
676060606060	X	
000000000017	15	
706060606060	Y	(51)
000000000025	21	
716060606060	Z	
000000000044	36	
000010000001	MASK	
000000000001		

3.2.5 *Polynomial in Core*

A polynomial term is stored in two or more consecutive locations in a manner determined by the format statement. The coefficient is placed in the first word and the exponents are packed into the remaining words, allowing the specified number of bits for each. For example, the term

$$5x^2y^7z^3 \tag{52}$$

in the format (50) has the octal representation

000000000005	5	
000020000007	2,7	(53)
000000000003	3	

A nonconstant polynomial in core consists of a pointer, a heading, a data block, and a format statement as explained in Section 3.1 (see Fig. 2). The data block contains the terms as in (53) stored consecutively.

3.2.6 *Canonical Form*

We are now prepared to discuss the canonical form subroutine, POLCFM. Its task is to put any given polynomial, stored in the manner described above, into canonical form. More precisely, it must order the terms according to their exponent sets, combining terms with equal exponent sets and discarding any resulting zeros. The terms are to be arranged in increasing order of the first exponent, and terms having the same first exponent are to be arranged in increasing order of the second, etc. If there is only one exponent word per term, this means that the terms can be ordered according to the magnitude of that word treated as an unsigned 36-bit integer. Otherwise they must be ordered according to the magnitude of the first exponent word and subordered according to the magnitude of the second, etc. No working space is required. The ordering is done first, with the aid of the system sort, FAPSTL, and the combinations and cancellations, if any, are then performed. Finally if the result is a constant, it is stored according to the "heading convention" which we shall now describe.

3.2.7 *Heading Convention*

As we mentioned in Section 3.1, each nonconstant polynomial has a fixed heading of three words containing the data address, the format address, and the number of terms, respectively. Since constant polynomials can usually profit from special treatment and in any case the zero polynomial requires it, we have devised a special representation for constants. The first word of the heading contains the code number 5, which cannot possibly be a legal data address, and the second contains the value of the constant. Such a heading has no associated data block, and its third word is never consulted. The code number zero signifies an idle heading, and the numbers one to four are reserved for rational functions.

The macro POLCLR (clear) stores zero in the first word of the heading, thereby marking it as idle and destroying the attached data block (if any). The macros POLSTZ (store zero), POLSTI (store identity), and POLSTC (store constant) store 5 in the first word of the heading and the specified constant in the second word.

3.2.8 *Output*

There is one output subroutine with three entry points — POLPRT (print), POLPCH (punch), and POLPRP (print and punch). Each term of a polynomial is printed (punched) on a single line (card), except that continuation lines (cards) will be used if necessary. All the coefficients are right adjusted to column 22, so that they form a column in the output. The exponents form one or more additional columns headed by the corresponding variable names. In each line the coefficient is separated from the first exponent by two blanks, and the exponents are separated from each other by single blanks. Therefore the exponent columns are not always straight. In printed output the first line contains the name of the polynomial (or any comment not more than 21 characters long) starting in column 2, and the names of the variables (separated by single blanks) starting in column 25. In punched output the first card contains the name or comment, the next card(s) is (are) a complete format statement, the ensuing cards contain the data, and finally an end card including the name is appended.

As an example, suppose the polynomial (43) is in core (in canonical form), and its name (i.e., the symbolic address of its pointer) is P. If P is then printed, the output will be

```
P            X  Y  Z
         -5  0  1  2
          2  1  1  1          (54)
          3  2  0  0
```

If it is punched, the output will be

```
P
X,12,Y,12,Z,12
           -5  0  1  2
            2  1  1  1          (55)
            3  2  0  0
0     END   P
```

where each line represents one card. The second card is a valid format statement, and the last one is a valid END card. A polynomial in many variables may require more than one line (card) for the list of variables (format statement) and/or more than one line (card) per term.

3.3 *Polynomial Arithmetic*

3.3.1 *Summary (See Descriptions Below)*

(*i*) *Basic Operations*

POLADD	R,P,Q	$R = P + Q$	add	(a)
POLSUB	R,P,Q	$R = P - Q$	subtract	(b)

POLMPY	R,P,Q	$R = P*Q$	multiply	(c)
POLDIV	R,P,Q,NODIV	$R = P/Q$	divide (if divisible)	(d)
POLSST	G,F(LISTP) (LISTV)	$G = F$(LISTV = LISTP)	substitute	(e)
POLDIF	Q,P,X	$Q = \partial P/\partial X$	differentiate	(f)
POLZET	P	skip *iff* $P = 0$	zero test	(g)
POLNZT	P	skip *iff* $P \neq 0$	nonzero test	(h)
POLEQT	P,Q	skip *iff* $P = Q$	equality test	(i)
POLDUP	Q,P	$Q = P$	duplicate	(j)
POLCHS	P	$P = -P$	change sign	(k)

(*ii*) *Alternatives for Added Convenience and/or Efficiency*

POLSMP	Q,C,P	$Q = C*P$	scalar multiply	(l)
POLSMO	C,P	$P = C*P$	scalar multiply and overwrite	(m)
POLOMP	Q,M,P	$Q = M?P$	one-term multiply	(n)
POLOMO	M,P	$P = M*P$	one-term multiply and overwrite	(o)
POLSAD	Q,C,P	$Q = C + P$	scalar add	(p)
POLSAO	C,P	$P = C + P$	scalar add and overwrite	(q)
POLADO	P,Q	$P = P + Q$	add and overwrite	(r)
POLDFO	P,X	$P = \partial P/\partial X$	differentiate and overwrite	(s)

(*iii*) *Explanation of Symbols*

F,G,P,Q,R = polynomials (symbolic addresses of pointers)
C = scalar (symbolic address of scalar)
M = monomial (symbolic address of pointer)
X = variable (specified in the manner indicated by the last previous VARTYP declaration — see Section 3.5.2)
LISTP = list of polynomials
LISTV = list of variables.

3.3.2 *Descriptions*

(a) POLADD R,P,Q

P and Q are assumed to be in canonical form. The addition is analogous to the ordered merging of two ordered subdecks of a deck of playing cards, except that POLADD must also perform combinations and cancellations. Suppose P has n terms and Q has m terms. Then a block long enough for $n + m$ terms is reserved for R if space permits. Otherwise all the remaining space is reserved for R, and the subroutine proceeds in the hope that combinations and/or cancellations will compensate for the deficiency. If space runs out, the job will be dumped. The first (next) term of R is found by comparing the exponent sets of the first (next) term of P and the first (next) term of Q. If these differ, the first (next) term of R is the first (next) term of P or of Q, depending on which comes earlier in the canonical ordering. If they are the same, the first (next) term of R is the sum of the first (next) terms of P and Q, unless the sum is zero. In that case the first (next) terms of P and Q cancel, making no contribution to R.

(b) POLSUB R,P,Q

This uses POLCHS (twice) and POLADD. If P and Q have the same heading, it uses POLSTZ instead.

(c) POLMPY R,P,Q

POLMPY multiplies the longer of the polynomials P and Q by each term of the shorter using POLOMP and accumulates these products using POLADD or POLAOE. The latter is a slightly modified version of POLADO, not normally available to the outside world. Its mnemonic is "Add, Overwrite the first argument, and Erase the second."

Suppose P has m terms and Q has n terms with $m \leqq n$. Let P_i be the ith term of P, let $T_i = P_iQ$ be the ith partial product, and let $S_i = \sum_{j=1}^{i} T_j$ be the ith partial sum.

If there is enough space for $(nm + n)$ terms, then the "leapfrog method," a fast method involving no data moving (see Fig. 3), is employed. Imagine the space partitioned into $m + 1$ blocks, each n terms long. The first partial product, T_1, is placed in the mth block and the second, T_2, in the $(m + 1)$st block. POLADD is then directed to add these, starting the sum S_2 at the beginning of the $(m - 1)$st block. This partial sum overwrites a portion (perhaps all) of the mth block as explained in the discussion of POLADO. The next partial product T_3 is then placed in the $(m + 1)$st block, and the next partial sum S_3 is started at the beginning of the $(m - 2)$nd block, overwriting a portion (perhaps all) of S_2. This process is repeated, each partial sum overwriting a portion (perhaps all) of the preceding one, until the final result S_m appears starting at the beginning of the first block.

If there is not enough space for this procedure, then the slower "compact method" (see Fig. 4) is used, requiring only enough space for the final result (or the longest partial sum) and n additional terms. The latest partial sum always starts at the top of the available space. The next partial product is placed immediately below it, and both are then moved down leaving a gap n terms long above the partial sum. The partial product is then added to the partial sum by POLAOE to produce a new partial sum, starting at the top of the available space and overwriting a portion (perhaps all) of the previous partial sum. This process

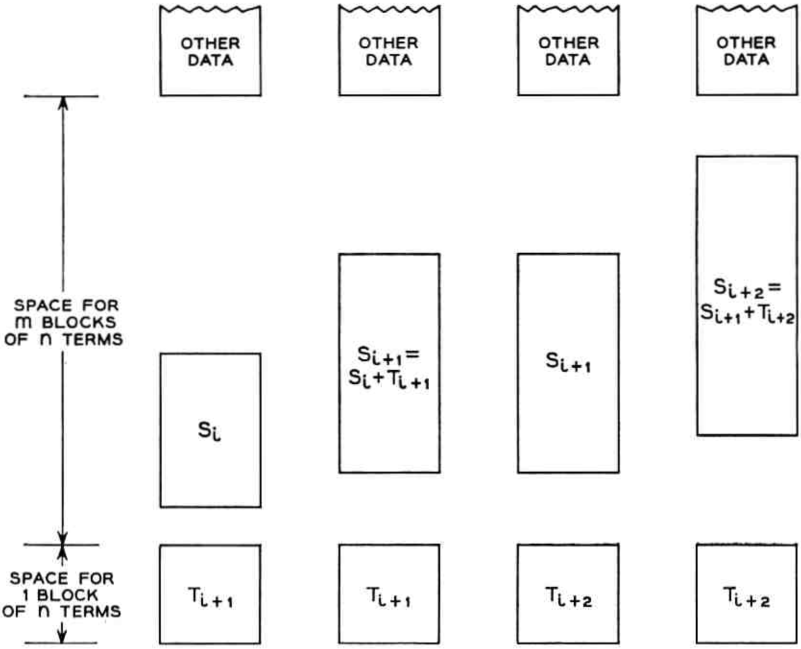

OTHER DATA
OTHER DATA
OTHER DATA
OTHER DATA
SPACE FOR m BLOCKS OF n TERMS
SPACE FOR 1 BLOCK OF n TERMS
S_i
$S_{i+1} = S_i + T_{i+1}$
S_{i+1}
$S_{i+2} = S_{i+1} + T_{i+2}$
T_{i+1}
T_{i+1}
T_{i+2}
T_{i+2}

FIG. 3 — Successive steps in multiplication by the "leapfrog method."

is repeated until the final result is achieved or the available space exhausted.

(d) POLDIV R,P,Q,NODIV

The dividend P and the divisor Q are treated as polynomials in one variable (the first variable that at least one of them depends on) with coefficients in the ring of polynomials in all the remaining variables (if any). Divisions in this ring can be handled by calling POLDIV itself,† and the main task is carried out by the familiar process of "long division." The fourth argument, NODIV, is an address to which control will be transferred if Q does not divide P. If the fourth argument is omitted, the macro will supply ENDJOB in its place.

(e) POLSST G,F(LISTP)(LISTV)

† A subroutine which calls itself is called recursive. At the innermost level it must, of course, operate by an independent mechanism. Collisions between the different levels are prevented by saving necessary information in a push-down list. It is perhaps worth noting that every inductive algorithm can be programmed as a recursive subroutine. In the case of polynomial division the induction is on the number of variables, and the innermost level is simply coefficient division.

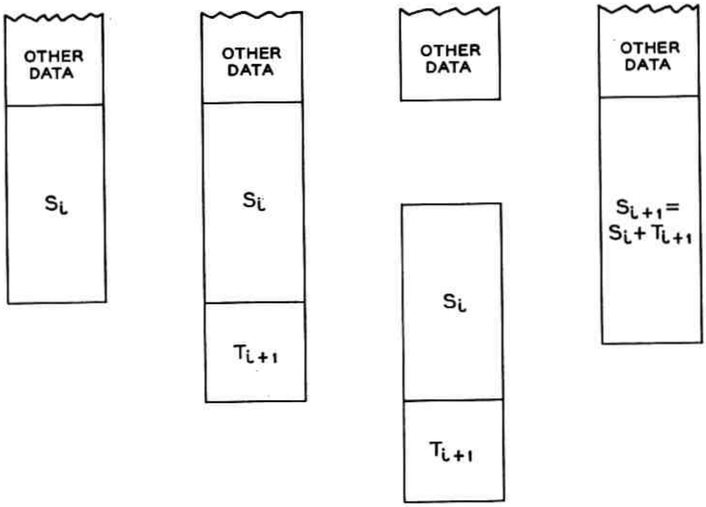


Fig. 4 — Successive steps in multiplication by the "compact method."

Here LISTP is a list of polynomials in a common format† which must include all the variables of F not being replaced, and LISTV is a list of the variables of F which are to be replaced by the polynomials in LISTP. For example if F depends on $X1, \cdots, X10$ and we wish to replace $X3$ and $X4$ by P and Q respectively, we write

POLSST G,F(P,Q)(X3,X4)

The variables in LISTV must be specified in the manner indicated by the last previous VARTYP declaration. If LISTV is not provided, it is understood to be the list of all the variables in the format of F.

POLSST works in the most straightforward possible way — substituting into one term at a time and preserving only the latest partial result. This procedure may involve substantial duplication of effort, but it uses a minimum of working space and a minimum of program, and in most practical cases the running time is reasonable.

(f) POLDIF Q,P,X

P is duplicated using POLDUP, and the copy is then differentiated with respect to X using POLDFO.

(g) POLZET P

† If all the polynomials in LISTP are constants (which have a universal format — see Section 3.2.7), then the format of F is used.

The next instruction is skipped if and only if $P = 0$.

(h) POLNZT P

The next instruction is skipped if and only if $P \neq 0$.

(i) POLEQT P,Q

The polynomials P and Q are considered to be equal if and only if they have the same format address, the same number of terms, and identical data blocks.

(j) POLDUP Q,P

Q is replaced by a copy of P.

(k) POLCHS P

The signs of all the coefficients of P are reversed.

(l) POLSMP Q,C,P

P is duplicated using POLDUP, and the copy is then multiplied by C using POLSMO.

(m) POLSMO C,P

Each coefficient of the polynomial P is multiplied by the scalar C.

(n) POLOMP Q,M,P

P is duplicated using POLDUP, and the copy is then multiplied by M using POLOMO.

(o) POLOMO M,P

Each term of the polynomial P is replaced by its product with the monomial M. To multiply two monomials, it is necessary to multiply their coefficients and add their exponents. In the case of integer coefficients, the coefficient multiplication macro

```
        CMP     Z,X,Y
```

expands to

```
        LDQ     X
        MPY     Y
        TZE     *+2
        REM1
        STQ     Z
```

where REM1 is the REMARK macro (see Section 3.7) for coefficient overflow. The exponent addition macro

```
EAD         Z,X,Y
```

adds the exponents one word at a time even though several exponents may be packed into each word. To check for overflow, EAD uses the appropriate word from the mask in the format statement. Suppose all the exponents are packed into a single word. Then the mask is a word containing a one in the low-bit position of each exponent block and zeros elsewhere. Now EAD expands to

```
CAL         X
ACL         Y
SLW         Z
ERA         X
ERA         Y
ANA         MASK
TZE         *+2
REM2
```

where REM2 is the REMARK macro (see Section 3.7) for exponent overflow. The first three lines compute the sum correctly, provided no overflows occur. After line 5 the low-bit positions in the AC should be zero, since ERA is the same as addition without carry. After line 6 the entire AC should therefore be zero. If it is not, control will pass to REM2 and the AC will contain a one-bit immediately to the left† of each exponent block which has overflowed.

(p)   POLSAD   Q,C,P

P is duplicated using POLDUP, and C is then added to the copy using POLSAO.

(q)   POLSAO   C,P

The scalar C is added (or appended) to the polynomial P.

(r)   POLADO   P,Q

Since P is to be replaced by the sum $P + Q$, it is not necessary to have space for both P and the sum. Instead it is possible to open a gap the size of Q above P, and then to use that gap together with the block occupied by P as a block for the sum. It is easy to see that no term of P can be overwritten by a term of the sum before making its contribution.

† Here we think of the AC as a circular register. An overflow in the leftmost exponent block will leave a one-bit at the right end of the AC.

(s) POLDFO P,X

Each coefficient of P is multiplied by the corresponding exponent of X. If the exponent is zero the term is deleted. Otherwise the exponent is reduced by one.

3.4 *Truncated Power Series*

Let x represent the k-tuple of variables $(x_1, \cdots, x_k)$. A *formal power series in* x is an expression of the form

$$A(x) = \sum_{i_1,\cdots,i_k=0}^{\infty} a_{i_1\cdots i_k} x_1^{i_1} \cdots x_k^{i_k} \tag{56}$$

where the a's are elements of any integral domain. The sum $i = i_1 + \cdots + i_k$ of the exponents in any individual term will be called the *order* of the term. Letting $a_i(x)$ be the (finite) polynomial consisting of all the terms of order i, we have

$$\begin{aligned} A(x) &= \sum_{i=0}^{\infty} A_i(x) \\ A_i(x) &= \sum_{\substack{i_1,\cdots,i_k \geqq 0 \\ i_1+\cdots+i_k=i}} a_{i_1\cdots i_k} x_1^{i_1} \cdots x_k^{i_k}. \end{aligned} \tag{57}$$

A *truncated power series of order* p is a formal power series from which all terms of order higher than p have been dropped. We shall restrict our attention to the case in which the a's are polynomials in a set of variables $y_1, \cdots, y_l$ $(l \geqq 0)$ not including any of the x's. The *sum* of two truncated power series

$$\begin{aligned} A(x) &= \sum_{i=p'}^{p} A_i(x); \qquad A_{p'}(x) \neq 0 \\ B(x) &= \sum_{j=q'}^{q} B_j(x); \qquad B_{q'}(x) \neq 0 \end{aligned} \tag{58}$$

is their polynomial sum truncated to order

$$\min\,(p,q) \tag{59}$$

while their *product* is their polynomial product truncated to order

$$\min\,(p + q', q + p'). \tag{60}$$

The ALPAK system contains two macros for dealing with truncated power series. These are POLTRC (truncate) and POLMPT (multiply and truncate). Addition can be handled with POLTRC and POLADD.

Each truncated power series must be stored as a polynomial in a format whose first k variables are the x's and whose remaining variables, if any, are the y's. The command

```
POLTRC      P,ORD,K                                  (61)
```

causes P to be truncated to order ORD. That is, all terms of order greater than ORD are deleted. The command

```
POLMPT      R,ORDR,P,ORDP,Q,ORDQ,K                   (62)
```

is represented by the equation

$$R = P*Q \tag{63}$$

where P and Q are truncated power series. K is the address of the num ber of power series variables [i.e., the x's of (56) and (57)], ORDP and ORDQ are the addresses of the orders of P and Q respectively, and ORDR is an address for the order of R, which is to be computed by the rule (60).

If it is desired to multiply a truncated power series by a polynomial, the latter should be thought of as a truncated power series of order infinity. It is required that all finite orders be less than 2^{35}, and any number greater than or equal to 2^{35} is treated as infinity. Thus if P is a truncated power series of order 4 in 3 variables and we wish to multiply it by the polynomial Q, we write

```
POLMPT      R,ORDR,P,=4,Q,=-1,=3                     (64)
```

where the order -1 of Q will be interpreted† as $2^{35} + 1$, which is equivalent to infinity.

3.5 *The Main Program*

3.5.1 *POLBEG*

Every main program starts with the macro POLBEG (begin). At assembly time, this reserves a block of storage for the "data buffer" and at execution time it initializes the storage allocator. The command

```
POLBEG      N                                        (65)
```

† In the IBM 7090 computer some operations interpret a word as a signed 35-bit integer and others interpret it as an unsigned 36-bit integer. If a negative integer is examined by one of the latter, the sign bit is assumed to represent a contribution of 2^{35} to the magnitude of the number.

(where N is an integer — not the address of an integer) reserves an N-word block in the "remote program," while the command

```
                    POLBEG      N,COMMON                            (66)
```

reserves an N-word block in "common storage." If COMMON is used, the space occupied by the loader at loading time can be a part of the data buffer at execution time. Therefore the size of the data buffer can be somewhat larger. However, no other program using COMMON can be loaded at the same time without careful use of ORIGIN cards.

3.5.2 *VARTYP*

Every program which uses POLSTV, POLSST, POLDIF, or POLDFO must contain at least one VARTYP declaration. The command

```
                    VARTYP      T                                   (67)
```

indicates that all subsequent references to variables (prior to the next VARTYP declaration if any) are of type T, which may be any of the following

NAM	(name)	(68)
NUM	(number)	
NAM*	(address of name)	
NUM*	(address of number)	

The variables in a format statement are numbered according to the order of their appearance.

For example, if we wish to differentiate the polynomial P with respect to the variable X, we use NAM and write

```
                    POLDIF      Q,P,X                               (69)
```

To differentiate P with respect to the third variable we use NUM and write

```
                    POLDIF      Q,P,3                               (70)
```

To differentiate P with respect to the variable whose name is at location LX, we use NAM* and write

```
                    POLDIF      Q,P,LX                              (71)
```

Finally, to differentiate P with respect to the variable whose number is at location K, we use NUM* and write

```
                    POLDIF      Q,P,K                               (72)
```

Typically, NAM is used in main programs and NUM* in subroutines, since the main programmer usually knows the names of the variables while the subroutine programmer usually knows nothing about the format.

3.5.3 *Sample Programs*

The following program computes $R = P + \partial Q/\partial Y$.

```
        POLBEG    10000
        VARTYP    NAM
FMT     POLCVF    (X,12,Y,12,Z,12)
        POLRDP    P,FMT
        POLRDP    Q,FMT
        POLDIF    DQDY,Q,Y
        POLADD    R,P,DQDY                            (73)
        POLPRT    R,-,(R = P + DQ/DY)
        TRA       ENDJOB
P       PZE
Q       PZE
R       PZE
        END
```

A slightly more complicated example illustrates the use of indexing. To compute

$$R_i = P_i + Q_i \,; \qquad i = 1, \cdots, 10 \tag{74}$$

we write

```
        POLBEG    10000,COMMON
        POLDRF    FMT
        AXT       10,1
RD1     POLRDP    (P,1),FMT
        TIX       RD1,1,1
        AXT       10,1
RD2     POLRDP    (Q,1),FMT
        TIX       RD2,1,1
        AXT       10,1
ADD     POLADD    (R,1),(P,1),(Q,1)
        TIX       ADD,1,1
        AXT       10,1                                (75)
```

```
PRT     POLPRT     (R,1),−,(R(I)=P(I)+Q(I))
        TIX        PRT,1,1
        TRA        ENDJOB
FMT     BSS        20
P       BES        10
Q       BES        10
R       BES        10
        END
```

The storage section of a main program must contain a block for each format statement read by POLRDF and a pointer (whose address field initially contains zero) for each polynomial. For further discussion of these rules see Section 3.2.

3.6 *Loading Instructions*

The polynomial portion of the ALPAK system consists of a macro deck and two subroutine packages, ALPAK1 and ALPAK2. Most of the macros expand into calling sequences for subroutines of the same name, but a few call one or more differently named subroutines and a few others call no subroutines at all. The macro deck is available as a symbolic deck or as a CRUNCH deck with no END card crunched in. ALPAK1 and ALPAK2 are available as binary decks and also as symbolic decks or CRUNCH decks. In their present form these decks can only be used within the BE-SYS-4 monitor system on an IBM 7090 computer.

The following example illustrates the arrangement of decks and control cards for a typical ALPAK assembly:

```
JOB
FAP
UNLIST
MACROS (CRUNCH deck with no end card crunched in)          (76)
LIST
MAIN PROGRAM (Symbolic deck with END card)
```

The UNLIST and LIST cards are normally included in order to suppress the printing of eleven pages of macro definitions. This is a FAP assembly and may be embellished in any way that conforms to the rules of FAP.

The next example shows a typical arrangement of decks and control cards for assembly and run:

JOB
FAP
UNLIST
MACROS (CRUNCH deck with no END card crunched in)
LIST
MAIN PROGRAM (Symbolic deck with END card)
LOAD BATCH (77)
ALPAK1 (Binary deck, preceded by LOAD card and followed by binary transfer card)
ALPAK2 (Binary deck, preceded by LOAD card and followed by binary transfer card)
TRA
DATA

Our final example illustrates a run with a previously assembled main program:

JOB
MAIN PROGRAM (Binary deck preceded by LOAD card and followed by binary transfer card) (78)
ALPAK1 (Binary deck preceded by LOAD card and followed by binary transfer card)
ALPAK2 (Binary deck preceded by LOAD card and followed by binary transfer card)
TRA
DATA

3.7 *Diagnostics*

The ALPAK diagnostic mechanism recognizes the following ten types of failure:

1. COEFFICIENT OVERFLOW. No coefficient or scalar can have magnitude greater than $2^{35} - 1$.
2. EXPONENT OVERFLOW. No exponent can be greater than $2^B - 1$, where B is the corresponding field width (in bits).
3. INSUFFICIENT SPACE. The reporting subroutine was unable to obtain needed space from the storage allocator.
4. ILLEGAL SUBROUTINE ARGUMENT. One of the inputs to the reporting subroutine failed some simple test.
5. INCOMPATIBLE FORMATS. See *Format Compatibility* in Section 3.1.2.
6. INTERNAL INCONSISTENCY. There may be a bug in the reporting subroutine.

7. POLBEG NOT CALLED. Every main program must begin with the macro POLBEG (see Section 3.5.1).
8. ILLEGAL FORMAT CARD. See *Format Statements* in Section 3.2.4.
9. END OF FILE. All the data cards have been read.
10. INPUT READING ERROR. An unrecoverable parity check failure has occurred on input.

Whenever a failure is detected, control is transferred to the REMARK subroutine, which performs the following functions: First it takes a hollerith snapshot of two locations containing the words "REMARK SNAP" in BCD. The purpose of this is to provide a console dump at the time of the failure. It then prints the location of the failure, the type of failure, and the subroutine nesting list. Finally it transfers control to the DUMP section (if any) of the first subroutine on the nesting list, whose function is to print the inputs and perhaps a partial result.

As an example, suppose the multiplication

```
POLMPY     C,A,B                    (79)
```

fails because of insufficient space. This might result in the output

```
LOCATION 1703
POLDUP REPORTS
INSUFFICIENT SPACE

SUBROUTINE NESTING LIST
NAMES AND CALLING LOCATIONS
    POLMPY 00174, POLOMP 03412, POLDUP 03152

FINAL DUMPS FROM POLMPY

R = P*Q. PS = PARTIAL SUM.

P        X Y Z
       1 0 1 0
       1 1 0 0

Q        X Y Z
       1 0 0 2
       1 0 2 0
       1 2 0 0
```

```
PS        X Y Z
        1 0 1 2
        1 0 3 0
        1 2 1 0
```

and the snapshot

```
                    2175      SNAP      H,2410,2411
          AC...MQ...SI...EK...SW...SL...OVF...TM...
          IR1...IR2...IR4...
                    2140      "REMARK"      "SNAP."
```

which will be the last snap prior to post mortems. The output indicates that POLMPY (multiply) was called from location 174 in the main program, POLOMP (one-term multiply) was called from location 3412 in POLMPY, POLDUP (duplicate) was called from location 3152 in POLOMP, and the space shortage was discovered at location 1703 in POLDUP. Furthermore, POLMPY was attempting to compute $R = P*Q$ where

$$\begin{aligned} P &= X + Y \\ Q &= X^2 + Y^2 + Z^2 \end{aligned} \tag{80}$$

and had obtained the partial result

$$PS = X^2Y + Y^3 + YZ^2 = Y(X^2 + Y^2 + Z^2) \tag{81}$$

which is the product of Q and the first term in the canonical ordering of P. Note that P, Q and R in the output are dummy names, which in this case correspond to A, B and C in the user's program [see (79)].

The subroutine nesting list is maintained automatically by the ENTER and EXIT macros, which are used in all but the lowest level subroutines. If a failure is detected in one of these unentered subroutines, its name will appear along with the location of the failure but not on the nesting list.

3.8 *Debugging*

The normal method of debugging an ALPAK program is to run it and see what happens. Most programming errors and all overflows will be located and identified by the diagnostic mechanism, which is described in the preceding section. If difficulties persist, POLPRT (print) orders can be inserted (by reassembly) into the main program, or even into one or more of the ALPAK subroutines. Each POLPRT order is essentially a symbolic snapshot. If an error is detected by POLPRT, a suitable remark (see below) is printed but the flow of control is not affected (unless it depends on the AC, the MQ, or XR4). The following remarks are available:

1. EMPTY POINTER. The pointer contains ±0.
2. NO DATA. The number of terms is +0.
3. GARBAGE. Either the heading is idle (see Section 3.2.7), the data address is outside the data buffer, the number of terms is $\leqq -0$, or the number of exponent words per term is $\leqq 0$.
4. ILLEGAL FORMAT. Since POLRDF and POLCVF do not accept illegal format statements, this remark implies that the format address is wrong or the format statement has been overwritten.
5. DATA OVERFLOW. The data block begins in the data buffer but ends beyond it.

If all else fails, ordinary snaps and/or post mortems can be taken in the usual manner. However, a snap of the data buffer is unusually difficult to comprehend and should be taken only in desperation.

IV. ACKNOWLEDGMENTS

The author is indebted to M. D. McIlroy for many valuable discussions throughout the course of this work, and to B. A. Tague and J. P. Hyde for major contributions to the planning and programming. He also wishes to thank S. P. Morgan and all of the above for their critical reading of the manuscript and their many suggestions for improvements in the presentation.

REFERENCES

1. Birkhoff, G., and MacLane, S., *A Survey of Modern Algebra*, rev. ed., MacMillan Company, New York, 1953.
2. Slepian, D., On the Zeros of Gaussian Noise, Ch. 6 of Proceedings of Symposium on Time Series Analysis, ed. Rosenblatt, M., John Wiley and Sons, Inc., New York, 1963.
3. Runyon, J. P., unpublished work.
4. Morrison, J. A., unpublished work.
5. Takács, L., A Single-Server Queue with Feedback, B.S.T.J., **42,** March, 1963, p. 503.
6. Riordan, J., *Stochastic Service Systems*, John Wiley and Sons, Inc., New York, 1962, p. 50.
7. Brown, W. S., and Fairlie, D. B., Analyticity Properties of the Momentum-Vertex Function, J. Math. Phys., **3,** March–April, 1962, pp. 221–235.
8. Brown, W. S., and Fairlie, D. B., Some Examples from Perturbation Theory, Ch. 14 of *On the Analytical Properties of the Vertex Function with Mass Spectral Conditions*, by Brown, W. S., Princeton University Report, 1961, pp. 251–275.
9. Thurston, R. N., Wave Propagation in Fluids and Normal Solids, Ch. 1 of *Physical Acoustics*, ed. Mason, W. P., Academic Press, New York, 1963.

Discrete Smoothing Filters for Correlated Noise

By J. D. MUSA

(Manuscript received April 18, 1963)

This paper discusses discrete, linear, time-invariant, nonrecursive, finite memory, polynomial smoothing filters for noise that is correlated from sample to sample. The wide-sense Markov process is used as a model for the noise. Analysis and synthesis of the aforementioned filters are discussed in detail and several plots are furnished. A simple method for generating discrete, wide-sense Markov noise for simulation is noted. A noise model composed of a linear combination of wide-sense Markov processes is developed and applied for the case in which the previous model is not sufficiently accurate.

I. INTRODUCTION

A discrete polynomial smoother may be defined in the following terms. Consider the random process $R(nT)$, where n is an integer and T is the period of the samples at which the process will be of interest.* The process will be thought of as comprising a desired component $\bar{R}(nT)$, and a noise component $\tilde{R}(nT)$. It will be assumed that $\bar{R}(nT)$ can be satisfactorily approximated by an rth degree polynomial in nT, $\hat{R}(nT)$. Further, assume that $\tilde{R}(nT)$ is a random process that is wide-sense stationary with respect to the sampling instants nT. The foregoing situation would occur, for example, in the tracking of a moving object whose true position could be represented as an rth degree polynomial in time, and whose measured position included a random error. We will assume that

$$E[\tilde{R}(nT)] = 0 \tag{1}$$

and denote var $[\tilde{R}(nT)]$ = var $[R(nT)]$ by σ_R^2, where E is the expected value operator of probability theory and "var" indicates "variance of".

* Symbols used throughout the paper have been collected and defined in a glossary (Section IX) for ready reference.

Let $\Phi_R(iT)$ represent the autocorrelation* function of $\tilde{R}(nT)$, where i is an integer. A discrete polynomial smoother of the pth order and rth degree is a filter which operates on $R(nT)$ in such a fashion that the output $C(nT)$ and the input $R(nT)$ are related by

$$E\{C(nT)\} = \hat{R}^{(p)}(nT + \Gamma) \tag{2}$$

for all n.† Note that the parenthetical superscript (p) denotes "pth derivative of the estimate with respect to nT." The quantity Γ represents prediction time; if Γ is negative, the operation performed is an interpolation.

We will consider linear, time-invariant smoothers which are nonrecursive and have a finite memory. These conditions may be expressed in terms of the input-output relationship

$$C(nT) = \sum_{i=0}^{N-1} W(iT)R[(n - i)T], \tag{3}$$

where the function $W(iT)$ is the weighting function or impulse response. Note that $W(iT)$ is defined only at a *finite* number of points (N points), that it is independent of the input (hence the smoother is *linear*), and that it is *invariant* with the time nT. No previous values of the output appear in (3); hence the smoother is *nonrecursive*. The latter restriction can often be circumvented, because it is frequently possible to approximate a recursive filter by a nonrecursive one.[1]

The quantity var $[C(nT)] = \sigma_C^2$ is of interest in two respects. First, we may wish to know its value, or better yet, the variance ratio

$$\mu^2 = \frac{\sigma_C^2}{\sigma_R^2}, \tag{4}$$

which is a figure of merit of the smoother. Note that μ^2 is not a function of time, since $R(nT)$ was assumed to be wide-sense stationary, and it follows that $C(nT)$ is also wide-sense stationary by the time-invariance of the smoother. Second, we may wish to find the *optimum* smoother of a class specified by p, r, Γ, N and T; i.e., we may want to determine

* In this paper, the term "autocovariance function" will be used to refer to

$$E[Z_t Z_{t+\tau}],$$

where Z_t represents a zero-mean, wide-sense stationary random process Z evaluated at time t. "Autocorrelation function" will be used to refer to the normalized autocovariance function obtained by dividing the autocovariance function by its value at $\tau = 0$.

† The 0th, 1st, and 2nd order smoothers are often referred to as position, velocity, and acceleration smoothers.

the weighting function $W(iT)$ which yields the minimum value of σ_C^2 (or μ^2) under the preceding conditions.

If $\tilde{R}(nT)$ has an autocorrelation function of arbitrary form, it may be shown, using (1), (3), and (4), that

$$\mu^2 = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} W_i W_j \Phi_R[(i-j)T], \tag{5}$$

where $W_i = W(iT)$ and $W_j = W(jT)$. In general, this is a complicated expression. In previous treatments[2,3,4,5] of discrete polynomial smoothers, simplification of (5) has been achieved by assuming that the power density spectrum of the noise component of the input is white, so that

$$\Phi_R[(i-j)T] = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j). \end{cases} \tag{6}$$

This yields the simple form

$$\mu^2 = \sum_{i=0}^{N-1} W_i^2. \tag{7}$$

However, the assumption that the noise is uncorrelated from sample to sample is not justified for many physical systems because the noise is restricted in its rate of change. This is particularly true for mechanical and electromechanical systems. It will be shown that correlated noise may be represented by the wide-sense Markov process as a first-order approximation, or by a linear combination of such processes as a better approximation, with appreciable simplification of (5) still being obtained. By "represent" we refer to the approximation of one autocorrelation function or power density spectrum by another. In discussing smoothers, our primary interest is in the behavior of the generalized second moment of random processes, and further delineation of the character of these processes is not necessary.

II. WIDE-SENSE MARKOV NOISE MODEL

A rigorous definition for the wide-sense Markov process may be found in Doob.[6] It will be sufficient for our purposes to characterize the wide-sense Markov process in an alternative fashion, which Doob[7] has shown to be equivalent to the original definition. A wide-sense stationary, continuous random process will be called wide-sense Markov if it has the autocorrelation function

$$\Phi(\tau) = \exp(-\Omega\tau), \qquad \tau \geqq 0. \tag{8}$$

The quantity Ω will be called the "noise bandwidth." By using the evenness property for autocorrelation functions of real, wide-sense stationary random processes, (8) may be written as

$$\Phi(\tau) = \exp(-\Omega \mid \tau \mid). \tag{9}$$

If a wide-sense Markov random process is real and Gaussian and has zero mean, then it is also strict-sense Markov. The strict-sense Markov process is defined as a random process for which

$$\Pr[Y(t_n) \leqq \lambda \mid Y(t_1), \cdots, Y(t_{n-1})] = \Pr[Y(t_n) \leqq \lambda \mid Y(t_{n-1})] \tag{10}$$

with probability 1 for each λ, all $t_1 < \cdots < t_n$, and all n. We may say in an intuitive manner that a strict-sense Markov process is a process with a structure such that any value of the process is directly related only to the immediately preceding value.

One might consider higher-order Markov processes ("related" to several preceding values) as a better approximation for correlated noise, but it appears that using a linear combination of the simple wide-sense Markov processes gives a more manageable expression for μ^2.

For a discrete wide-sense Markov process with equally-spaced samples, we may write the autocorrelation function as

$$\Phi(\tau) = \exp(-\Omega \mid \tau \mid) Cb_T(\tau), \tag{11}$$

where Cb_T is the comb function defined by

$$Cb_T(\tau) = \sum_{i=-\infty}^{\infty} \delta(\tau - iT). \tag{12}$$

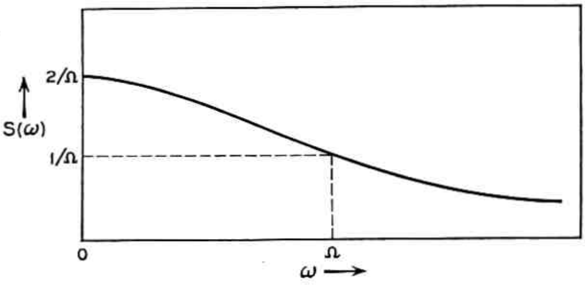


Fig. 1 — Baseband component of normalized power density spectrum for discrete wide-sense Markov process.

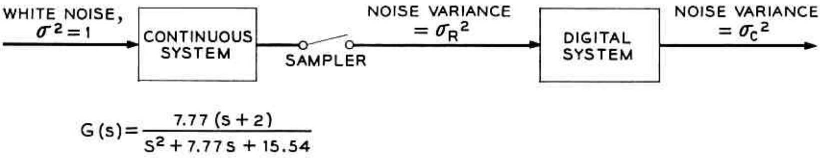

WHITE NOISE, $\sigma^2=1$
CONTINUOUS SYSTEM
SAMPLER
NOISE VARIANCE $=\sigma_R^2$
DIGITAL SYSTEM
NOISE VARIANCE $=\sigma_C^2$
$G(s)=\frac{7.77(s+2)}{s^2+7.77s+15.54}$

Fig. 2 — Control system used in evaluation of wide-sense Markov noise model.

The normalized† power density spectrum, obtained by Fourier transformation of (11), is

$$S(f) = \frac{2\Omega}{(2\pi f)^2 + \Omega^2} * \frac{1}{T} Cb_{1/T}(f), \tag{13}$$

where $*$ indicates convolution. The baseband component of this normalized power density spectrum is illustrated in Fig. 1. Note that the half-power point occurs at $f = \Omega/2\pi$.

Use of the wide-sense Markov noise model reduces (5) to

$$\mu^2 = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} W_i W_j \alpha^{|i-j|}, \tag{14}$$

where

$$\alpha = \exp(-\Omega T) \tag{15}$$

and is called the "intersample correlation." For some weighting functions, (14) can be simplified much further by evaluating the sums, using the finite difference calculus.

As one illustration of the improvement in accuracy obtained by representing correlated noise as wide-sense Markov rather than white, consider the control system of Fig. 2. White noise is filtered by the continuous system such that the normalized power density spectrum at the input to the sampler becomes

$$S(\omega) = \frac{|G(j\omega)|^2}{\sigma^2}, \tag{16}$$

where

$$\sigma^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} |G(j\omega)|^2 \, d\omega = 4.79. \tag{17}$$

† Normalized in the sense that this is the Fourier transform of the autocorrelation function. The power density spectrum is usually defined as the Fourier transform of the autocovariance function. The normalized power density spectrum is equal to the power density spectrum divided by the variance.

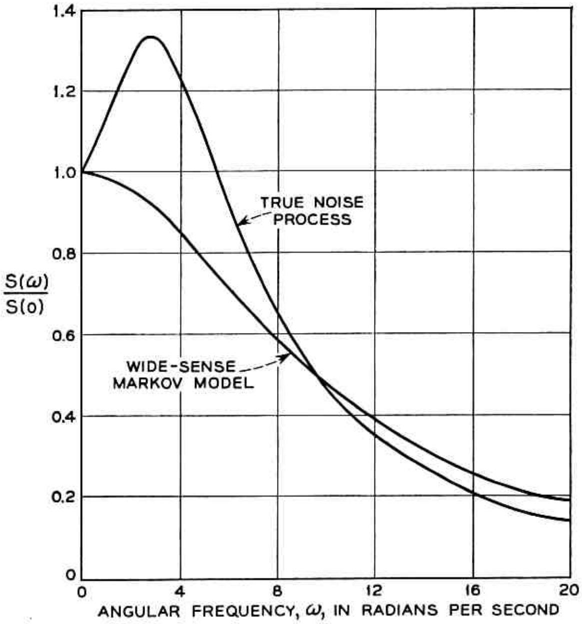


Fig. 3 — Normalized power density spectra.

Hence

$$S(\omega) = \frac{12.6(\omega^2 + 4)}{\omega^4 + 29.3\omega^2 + 241.5}. \tag{18}$$

We can fit models to the true noise process as if all processes were continuous, and following this, introduce the sampling operation. The output-input noise variance ratio μ^2 of the digital system has been computed for the case of a first-order cascaded simple averages smoother† with the following weighting coefficients:

$$W_i = \begin{cases} 0.028257 & (0 \leqq i \leqq 11) \\ 0 & (12 \leqq i \leqq 23) \\ -0.028257 & (24 \leqq i \leqq 35). \end{cases} \tag{19}$$

The true noise process has $\mu^2 = 0.0376$. Use of the wide-sense Markov model yields $\mu^2 = 0.0339$, while use of the white noise model yields $\mu^2 = 0.0192$.

† See Section V for the definition of this smoother.

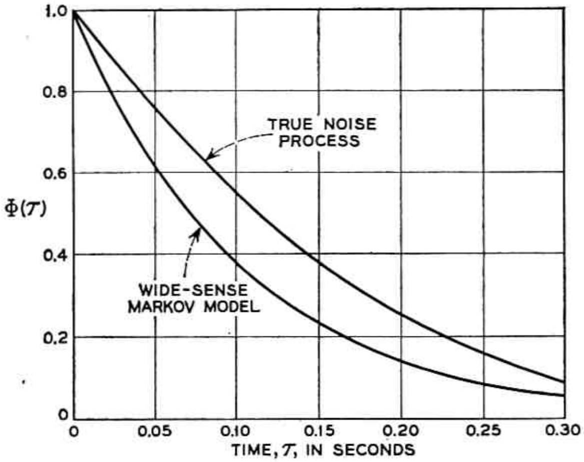


Fig. 4 — Autocorrelation functions.

The normalized power density spectra and autocorrelation functions of the true noise process and the wide-sense Markov model are illustrated in Figs. 3 and 4. The parameter Ω has been picked equal to the half-power point of the power density spectrum of the true noise process, 9.68.

III. MOMENTS OF THE WEIGHTING FUNCTION

The moments of the weighting function of a smoother are important characteristics, since the requirement (2) which specifies the desired output of the smoother is conveniently expressed in terms of them. The moments will be useful in comparing smoothers for equivalence as to meeting (2), and in determining the optimum weighting function for a class of smoothers. The qth moment M_q of the weighting function will be defined as

$$M_q = \sum_{i=0}^{N-1} (iT)^q W_i . \tag{20}$$

To express (2) in terms of moments, we proceed as follows. Substituting (3) and (1) in (2) we obtain

$$\sum_{i=0}^{N-1} W_i \bar{R}[(n - i)T] = \hat{R}^{(p)}(nT + \Gamma). \tag{21}$$

Now $\bar{R}(t)$ will be approximated by $\hat{R}(t)$, which may be expressed in the

Taylor series form

$$\hat{R}(t) = \sum_{q=0}^{r} \frac{\hat{R}^{(q)}(nT)}{q!} (t - nT)^q. \tag{22}$$

Substituting (22) in both sides of (21) and rearranging, we obtain

$$\sum_{q=0}^{r} \frac{(-1)^q \hat{R}^{(q)}(nT)}{q!} \sum_{i=0}^{N-1} (iT)^q W_i = \sum_{q=p}^{r} \frac{\hat{R}^{(q)}(nT)}{(q-p)!} \Gamma^{q-p}. \tag{23}$$

Considering (23) term by term, and using (20), we obtain

$$M_q = \begin{cases} 0 & (0 \leqq q < p) \\ (-1)^p p! & (q = p) \\ \dfrac{(-1)^q q!}{(q-p)!} \Gamma^{q-p} & (p < q \leqq r). \end{cases} \tag{24}$$

It should be noted that the weighting function obviously *has* moments greater than the rth; however, the condition (2) does not fix their values.

IV. OPTIMUM SMOOTHERS

By "optimum smoother" we mean that smoother of the class specified by p, r, Γ, N, and T whose weighting function yields the minimum possible value of μ^2. Optimum smoothers are often not implemented because of the amount of storage and computation required. However, they provide a standard of comparison for the systems that are implemented.

To find the weighting function of the optimum smoother of a class, the quantity μ^2 is minimized under the constraints (24), using Lagrange's method of undetermined multipliers. Blackman[8] has carried out the minimization in matrix form for a general input noise process (any autocorrelation function). The optimum smoother is specified by the matrix equation

$$W = P^{-1}A(\tilde{A}P^{-1}A)^{-1}M, \tag{25}$$

where ~ indicates "matrix transpose." The variance ratio μ^2 for the optimum smoother is given by

$$\mu^2 = \tilde{M}(\tilde{A}P^{-1}A)^{-1}M. \tag{26}$$

The matrix W is a column matrix representing the weighting function at the points $t = iT$, i.e.,

$$W = \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_{N-1} \end{bmatrix}; \tag{27}$$

P is the autocorrelation matrix of the input noise process,

$$P = \begin{bmatrix} 1 & \Phi_R(T) & \Phi_R(2T) & \cdots & \Phi_R[(N-1)T] \\ \Phi_R(T) & 1 & \Phi_R(T) & \cdots & \Phi_R[(N-2)T] \\ \Phi_R(2T) & \Phi_R(T) & 1 & \cdots & \Phi_R[(N-3)T] \\ \vdots & \vdots & \vdots & & \vdots \\ \Phi_R[(N-1)T] & \Phi_R[(N-2)T] & \Phi_R[(N-3)T] & \cdots & 1 \end{bmatrix}; \tag{28}$$

A is the "age" matrix,

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & T & T^2 & \cdots & T^r \\ 1 & 2T & (2T)^2 & \cdots & (2T)^r \\ 1 & 3T & (3T)^2 & \cdots & (3T)^r \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & (N-1)T & [(N-1)T]^2 & \cdots & [(N-1)T]^r \end{bmatrix}; \tag{29}$$

and M is the column matrix of moments,

$$M = \begin{bmatrix} M_0 \\ M_1 \\ \vdots \\ M_r \end{bmatrix}. \tag{30}$$

Unfortunately, (25) and (26) are very difficult to evaluate literally except in the simplest cases. However, they can be evaluated numerically by a digital computer. The inverse of the autocorrelation matrix for wide-sense Markov noise is readily determined to be, in literal form,

$$P^{-1} = \frac{1}{1-\alpha^2}\begin{bmatrix} 1 & -\alpha & 0 & 0 & \cdots & & 0 \\ -\alpha & 1+\alpha^2 & -\alpha & 0 & & & \\ 0 & -\alpha & 1+\alpha^2 & -\alpha & & & \vdots \\ 0 & 0 & -\alpha & 1+\alpha^2 & & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ \vdots & & & & & & 0 \\ & & & & & 1+\alpha^2 & -\alpha \\ 0 & & \cdots & & 0 & -\alpha & 1 \end{bmatrix}. \tag{31}$$

The principal operation, aside from the matrix multiplications, is the inversion of the $(r + 1) \times (r + 1)$ matrix $\tilde{A}P^{-1}A$.

Blackman[8] has evaluated (25) and (26), assuming that the noise is wide-sense Markov, for zero prediction time smoothers with $p = 0$, $r = 0$ and $p = 1, r = 1$. For the former,

$$W_i = \begin{cases} \dfrac{1}{N-(N-2)\alpha} & (i = 0, N-1) \\[2ex] \dfrac{1-\alpha}{N-(N-2)\alpha} & (i = 1, 2, \cdots, N-2) \end{cases} \tag{32}$$

and

$$\mu^2 = \frac{1+\alpha}{N-(N-2)\alpha}. \tag{33}$$

For the latter,

$$W_i = \begin{cases} \dfrac{3}{T}\dfrac{(1+\eta)[1+\eta(N-2)]}{(N-1)\{[1+\eta(N-1)][2+\eta(N-1)]+[1-\eta^2]\}} \\ \hfill (i = 0) \\[2ex] \dfrac{6}{T}\dfrac{\eta^2(N-1-2i)}{(N-1)\{[1+\eta(N-1)][2+\eta(N-1)]+[1-\eta^2]\}} \\ \hfill (i = 1, 2, \cdots, N-2) \\[2ex] -\dfrac{3}{T}\dfrac{(1+\eta)[1+\eta(N-2)]}{(N-1)\{[1+\eta(N-1)][2+\eta(N-1)]+[1-\eta^2]\}} \\ \hfill (i = N-1) \end{cases} \tag{34}$$

and

$$\mu^2 = \frac{1}{T^2}\,\frac{12\eta}{(N-1)\{[1+\eta(N-1)][2+\eta(N-1)]+[1-\eta^2]\}}, \quad (35)$$

where

$$\eta = \frac{1-\alpha}{1+\alpha}. \quad (36)$$

These optimum weighting functions and variance ratios have been plotted in a normalized form in Figs. 5, 6, 7, and 8. The ordinates for the 1st order, 1st degree smoother are given in terms of the smoothing interval $T_S = (N-1)T$. The curves are plotted for the parameter $B = \Omega T_S$, which may be thought of as a noise-smoother "bandwidth ratio." The asymptotes for the above curves, as $N \to \infty$ (with T_S and Ω fixed), are derived in Appendix A.

Let us consider the behavior of these curves from a physical viewpoint. For wide-sense Markov noise, the noise autocorrelation function is positive and monotonically decreasing with time. Hence, if the number of samples smoothed, N, is increased with the smoothing interval

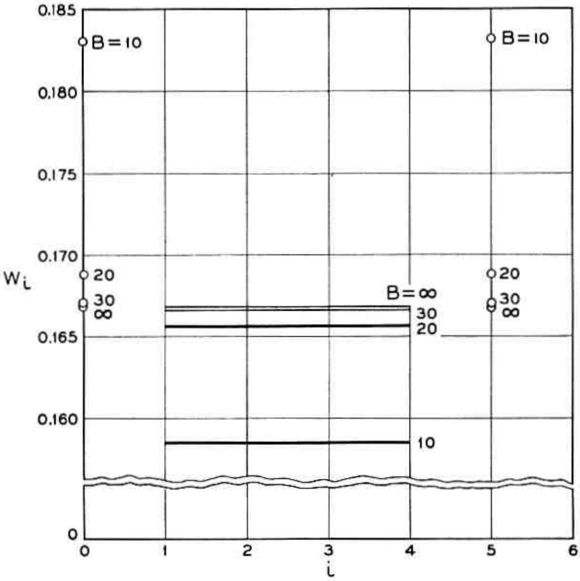


Fig. 5 — Optimum weighting function: 0th order, 0th degree smoother ($\Gamma = 0$, $n = 6$).

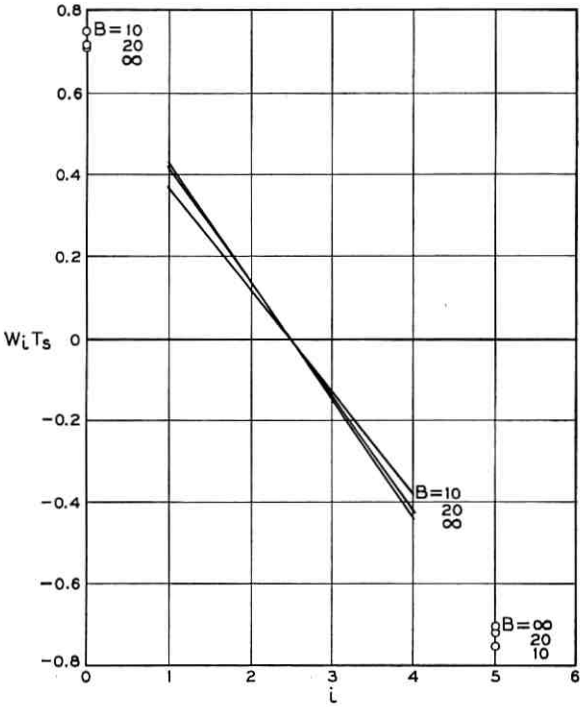


Fig. 6 — Optimum weighting function: 1st order, 1st degree smoother ($\Gamma = 0$, $n = 6$).

T_s and the noise characteristics remaining fixed, the intersample correlation will increase. Although each additional sample provided to the smoother gives additional information, the information added eventually approaches zero due to the increasing correlation. Now a smoother can reduce its variance ratio only by obtaining more information about the noise or by making better use of the information it already has. An optimum smoother makes the best use of the information available to it. Consequently, the variance ratio of an optimum smoother operating on a signal which includes wide-sense Markov noise (or any noise whose autocorrelation function is positive and decreases monotonically with time) must approach a constant as N increases.

V. CASCADED SIMPLE AVERAGES SMOOTHERS

Cascaded simple averages smoothers are a class of smoothers developed by R. B. Blackman.[1] A cascaded simple averages smoother of sth order

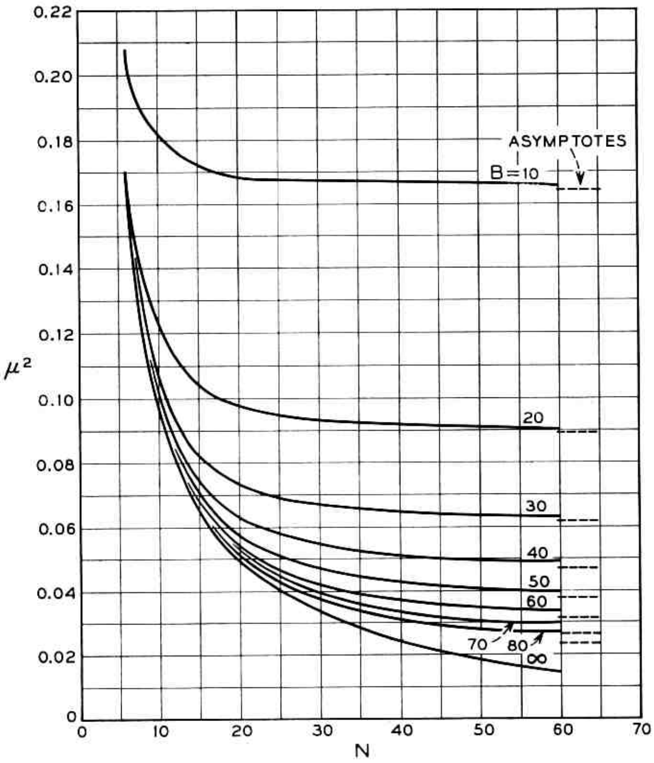


Fig. 7 — Noise variance ratio: optimum 0th order, 0th degree smoother.

approximates an optimum (with respect to white noise) sth order, sth degree, zero prediction time smoother. It may also be used to approximate smoothers that have been optimized with respect to wide-sense Markov noise. The approximation involves using only the values K, $-K$, and 0 for the weighting coefficients, where K is some constant. This smoothing method reduces the amount of storage and the number of arithmetic operations required, at the cost of a slight increase in μ^2 over the optimum method.

The weighting functions of cascaded simple averages smoothers of 0th, 1st, and 2nd orders are as follows (respectively):

$$W_i = \frac{1}{N}, \tag{37}$$

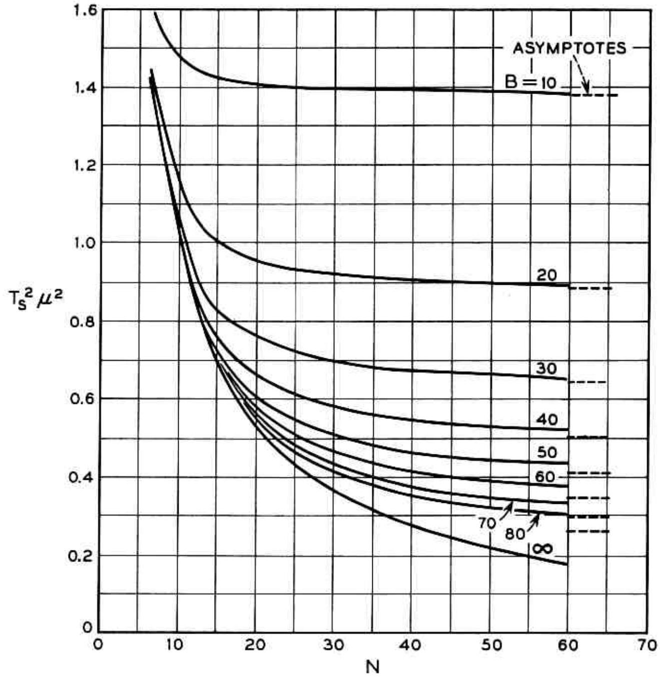


Fig. 8 — Normalized noise variance ratio: optimum 1st order, 1st degree smoother

$$W_i = \begin{cases} \dfrac{4.5(N-1)}{N^2T_S} & \left(0 \leqq i \leqq \dfrac{N}{3} - 1\right) \\ 0 & \left(\dfrac{N}{3} \leqq i \leqq \dfrac{2}{3}N - 1\right) \\ -\dfrac{4.5(N-1)}{N^2T_S} & \left(\dfrac{2}{3}N \leqq i \leqq N - 1\right), \end{cases} \tag{38}$$

and

$$W_i = \begin{cases} \dfrac{36(N-1)^2}{N^3T_S^2} & \left(0 \leqq i \leqq \dfrac{N}{6} - 1\right) \\ 0 & \left(\dfrac{N}{6} \leqq i \leqq \dfrac{N}{3} - 1\right) \\ -\dfrac{36(N-1)^2}{N^3T_S^2} & \left(\dfrac{N}{3} \leqq i \leqq \dfrac{2}{3}N - 1\right) \\ 0 & \left(\dfrac{2}{3}N \leqq i \leqq \dfrac{5}{6}N - 1\right) \\ \dfrac{36(N-1)^2}{N^3T_S^2} & \left(\dfrac{5}{6}N \leqq i \leqq N - 1\right), \end{cases} \tag{39}$$

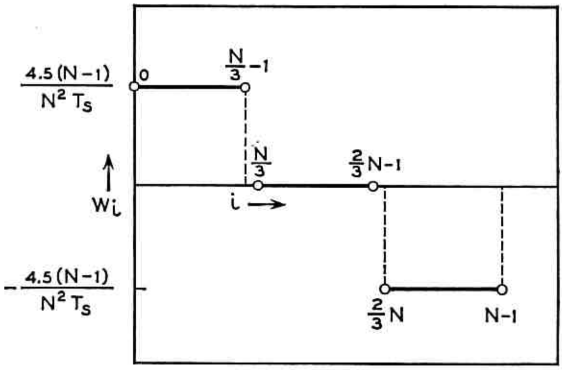


Fig. 9 — Weighting function for 1st order cascaded simple averages smoother.

where N is a multiple of 3 in (38) and a multiple of 6 in (39). The weighting functions for 1st and 2nd order smoothers are plotted in Figs. 9 and 10, respectively.

The variance ratios for 0th, 1st, and 2nd order cascaded simple averages smoothers for a wide-sense Markov noise input are, respectively:

$$\mu^2 = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\alpha^{|i-j|}, \tag{40}$$

$$\mu^2 = \left[\frac{4.5(N-1)}{N^2T_S}\right]^2\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} sgn\, W_i\, sgn\, W_j\, \alpha^{|i-j|}, \tag{41}$$

and

$$\mu^2 = \left[\frac{36(N-1)^2}{N^3T_S^2}\right]^2\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} sgn\, W_i\, sgn\, W_j\, \alpha^{|i-j|}, \tag{42}$$

where

$$sgn\, W_i = \begin{cases} -1 & (W_i < 0) \\ 0 & (W_i = 0) \\ 1 & (W_i > 0). \end{cases} \tag{43}$$

By use of the finite difference calculus, (40), (41), and (42) may be simplified to

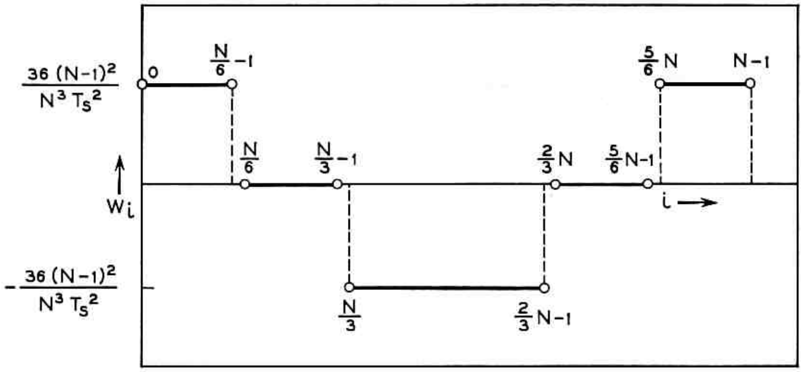


Fig. 10 — Weighting function for 2nd order cascaded simple averages smoother.

$$\mu^2 = \frac{1+\alpha}{N(1-\alpha)} + \frac{2\alpha(\alpha^N - 1)}{[N(1-\alpha)]^2}, \tag{44}$$

$$\mu^2 = 2\left(\frac{4.5}{T_S}\right)^2\left(\frac{N-1}{N}\right)^2 \cdot\left\{\frac{1+\alpha}{3N(1-\alpha)} - \frac{\alpha(\alpha^N - 2\alpha^{2N/3} - \alpha^{N/3} + 2)}{[N(1-\alpha)]^2}\right\}, \tag{45}$$

and

$$\mu^2 = 2\left(\frac{36}{T_S^2}\right)^2\left(\frac{N-1}{N}\right)^4 \cdot\left\{\frac{1+\alpha}{3N(1-\alpha)} + \frac{\alpha(\alpha^N - 2\alpha^{5N/6} - \alpha^{2N/3} + 2\alpha^{N/2} + 3\alpha^{N/3} - 3)}{[N(1-\alpha)]^2}\right\}, \tag{46}$$

respectively.

In Figs. 11, 12, and 13, the variance ratios have been plotted in normalized form for 0th, 1st, and 2nd order cascaded simple averages weighting functions, respectively. The ordinates are μ^2, $T_S^2\mu^2$, and $T_S^4\mu^2$, respectively. The curves are plotted for the noise-smoother "bandwidth ratio" $B = \Omega T_S$. The asymptotes for the above smoothers as $N \to \infty$ (with T_S and Ω fixed) are derived in Appendix A. Note that the expressions simplify appreciably for larger values of B, the exponential terms becoming negligible.

The behavior of these variance ratio curves is somewhat different from those for the optimum smoother. They do not necessarily decrease

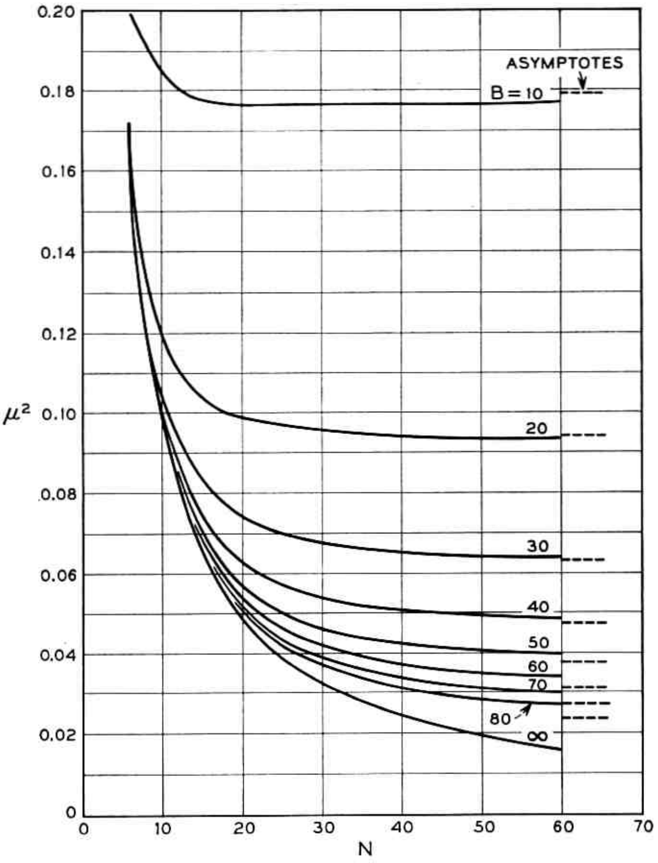


Fig. 11 — Noise variance ratio: 0th order cascaded simple averages smoother.

monotonically with N, even though they have asymptotes similar to the optimum curves. This is due to the fact that the smoothers are not optimum, and therefore the information about the noise is not necessarily utilized in the best manner. Consequently, as N increases, change in variance ratio may be due to changes in the *utilization* of the information available as well as changes in the information available, and the change cannot be readily predicted.

Note that the curves for all three orders of smoothers (Figs. 11, 12, and 13) either have a minimum at some finite value of N or approach a minimum as $N \rightarrow \infty$. These minima are more or less broad. In specifying a smoother, it is advantageous to choose the lowest value of N for

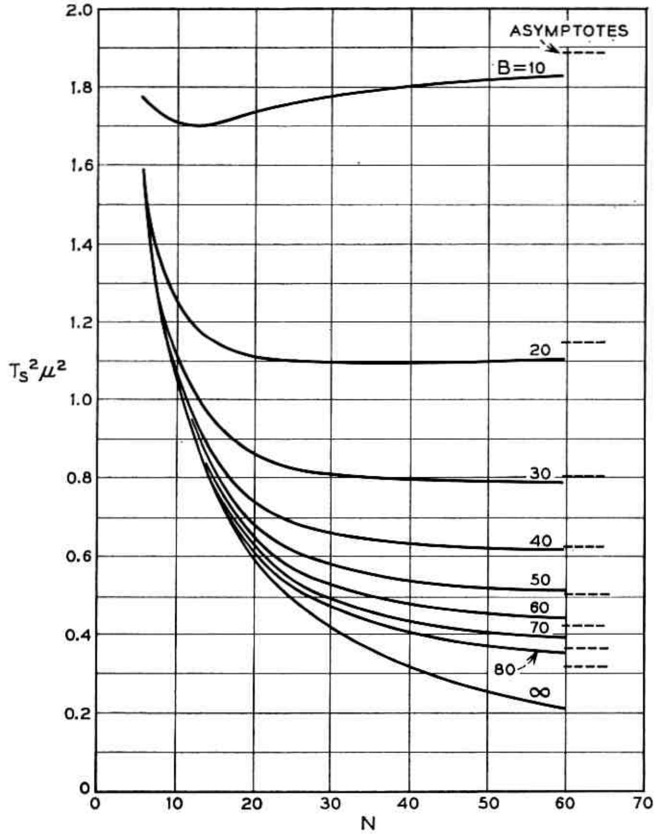


FIG. 12 — Normalized noise variance ratio: 1st order cascaded simple averages smoother.

which μ^2 is reasonably close to the minimum. Note that the neighborhood of the minimum variance ratio as a function of N is reached at lower values of N as B decreases (intersample correlation α increases for fixed T_s). This is reasonable physically, since the value of smoothing a larger number of samples decreases as these samples become more highly correlated.

VI. SYNTHESIS OF POLYNOMIAL SMOOTHERS

In general, the polynomial smoothers we have been discussing are classified by the parameters p, r, Γ, N, and T.† It would be convenient

† The optimum smoother is also classified by the parameter α.

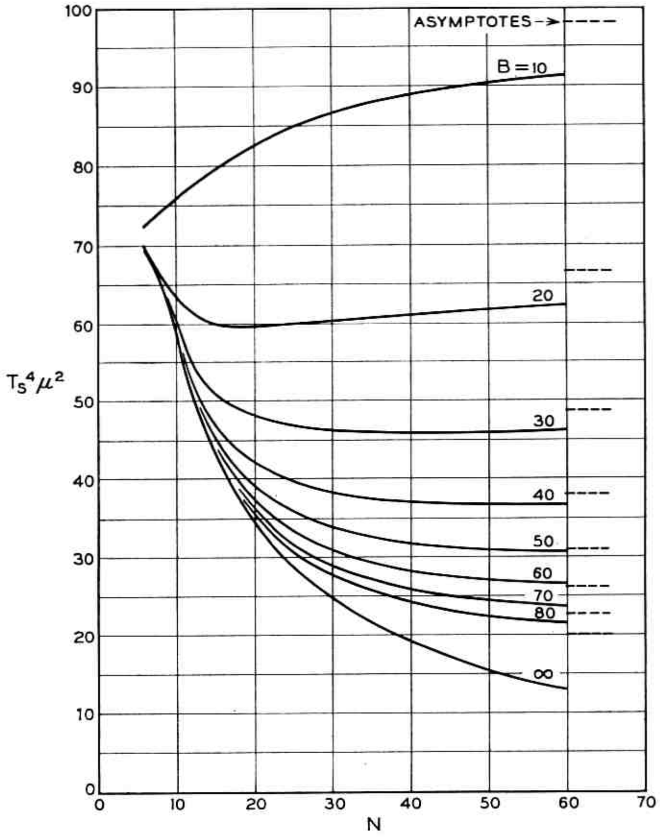


Fig. 13 — Normalized noise variance ratio: 2nd order cascaded simple averages smoother.

to be able to synthesize the smoother *in terms of* sth order, sth degree, zero prediction time components, where $p \leqq s \leqq r$. Note that the components are functions of s, N, and T only; hence their characteristics could be specified fairly simply. Further, several smoothers with different parameters p, r, and Γ but the same N and T could be synthesized with common components by weighting these components differently. Finally, the above breakdown permits any polynomial smoother of the class considered in this paper to be constructed from cascaded simple averages components. The derivation and procedures discussed in this section are valid for discrete polynomial smoothers in general and are not restricted to optimum smoothers or to particular input noise power density spectra.

Consider the linear combination of sth order, sth degree, zero prediction time components shown in Fig. 14. Let W_{si} represent the value of

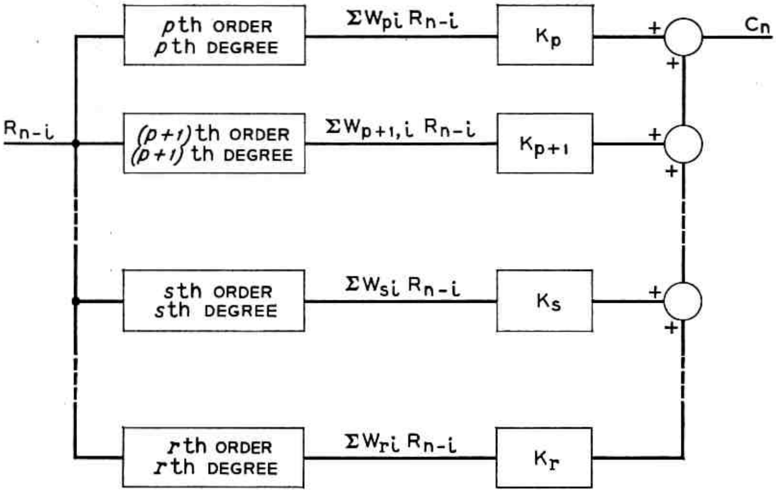


Fig. 14 — Synthesis of pth order, rth degree smoother from sth order, sth degree components.

the weighting function of the sth component at the sample with age iT. Let M_{sq} be the qth moment of the weighting function of the sth component. From Fig. 14 it will be seen that the "over-all" weighting function W_i of the entire smoother is related to the component weighting functions by

$$W_i = \sum_{s=p}^{r} K_s W_{si} . \tag{47}$$

Now, using (47) and (20),

$$M_q = T^q \sum_{i=0}^{N-1} i^q W_i = T^q \sum_{s=p}^{r} K_s \sum_{i=0}^{N-1} i^q W_{si} = \sum_{s=p}^{r} K_s M_{sq} . \tag{48}$$

From (24) we obtain

$$M_{sq} = \begin{cases} (-1)^q q! & (s = q) \\ 0 & (s > q). \end{cases} \tag{49}$$

Substituting (49) in (48) we get

$$M_q = \sum_{s=p}^{q-1} K_s M_{sq} + K_q (-1)^q q!. \tag{50}$$

If the linear combination of components is to be equivalent to the pth order, rth degree smoother, then (24) must be satisfied. It follows that we must have

$$K_s = \begin{cases} 1 & (s = p) \\ \dfrac{\Gamma^{s-p}}{(s-p)!} - \dfrac{(-1)^s}{s!} \displaystyle\sum_{u=p}^{s-1} K_u M_{us} & (p < s \leqq r). \end{cases} \tag{51}$$

It should be noted that a linear combination of optimum components will not necessarily be optimum unless the outputs of the components at a common time are uncorrelated.

The synthesis procedure proper consists of finding a smoother or set of component smoothers which produces the desired output (2) with the least total error ϵ compatible with a simple implementation. In the case of recursive smoothers, stability must be considered; the latter topic is adequately covered in standard texts on control theory.[9] The total error ϵ is given by

$$\epsilon = [\sigma_C^2 + \epsilon_T^2]^{\frac{1}{2}}, \tag{52}$$

where ϵ_T is the truncation error

$$\epsilon_T = \bar{R}^{(p)}(nT + \Gamma) - \hat{R}^{(p)}(nT + \Gamma). \tag{53}$$

Alternatively, using (21), we may write (53) as

$$\epsilon_T = \bar{R}^{(p)}(nT + \Gamma) - \sum_{i=0}^{N-1} W_i \bar{R}[(n - i)T]. \tag{54}$$

Synthesis involves the choice of type of filter (optimum, cascaded simple averages, etc.) and the selection of r, N and T. For convenience, the parameters will be selected in the alternate form r, N, and T_s.

The selection of r is based on the requirement that $r \geqq p$ and the direction of change in ϵ as r increases. Now σ_C^2 increases and ϵ_T decreases (in general) with increasing r. The rate of increase of σ_C^2 with r is such that smoothers with $r > 2$ are seldom used in practice.

The selection of N and T_s will be a trial-and-error process based on achieving a near minimum in ϵ while keeping N as small as possible (for simpler implementation). In the case of a set of components, each component may have a different value of T_s provided the values of T are the same. Figs. 7, 8, 11, 12, and 13 will be useful in calculating σ_C^2. When calculating the over-all output noise of a set of component smoothers, it will be useful to know that the noise outputs of 0th, 1st, and 2nd order cascaded simple averages smoothers are all mutually uncorrelated, though this is not true for all orders.[1]

The problems involved in estimating truncation error have been discussed by Hamming[10] in some detail. We will make the simplifying assumption that the truncation error ϵ_T of an rth degree smoother may be approximated by using (54) with $\bar{R}(t)$ considered as an $(r + 1)$th degree polynomial. Thus $\bar{R}(t)$ may be expressed

$$\bar{R}(t) = \sum_{q=0}^{r+1} \frac{\bar{R}^{(q)}(nT)}{q!} (t - nT)^q. \tag{55}$$

Substituting (55) into (54), and using (22) and (26), we obtain

$$\epsilon_T = \bar{R}^{(r+1)}(nT) \left[\frac{\Gamma^{r-p+1}}{(r-p+1)!} - (-1)^{r+1} \frac{M_{r+1}}{(r+1)!} \right]. \tag{56}$$

Blackman[11] has calculated the $(r + 1)$th moments of 0th, 1st, and 2nd order cascaded simple averages smoothers as $\frac{1}{2}T_s$, $-T_s$, and $3T_s$, respectively. Hence, the truncation errors for these smoothers may be calculated from (56) as $\frac{1}{2}T_s\bar{R}^{(r+1)}(nT)$.

VII. GENERATION OF DISCRETE WIDE-SENSE MARKOV NOISE FOR SIMULATION

It is frequently desired to simulate the performance of discrete smoothing filters and perhaps larger discrete systems of which they may be a part. Standard techniques are available for simulating discrete white noise by generation of a sequence of uncorrelated pseudo-random numbers.[12,13] It is relatively easy to generate discrete wide-sense Markov noise from such a sequence, due to the simple correlation structure of the wide-sense Markov process. The foregoing is another advantage in using the wide-sense Markov model to represent correlated noise.

Let $\{Y_n\}$ be the desired discrete wide-sense Markov noise and $\{X_n\}$ be a sequence of uncorrelated random numbers of zero mean and unit variance. Then Y_n may be generated as

$$Y_1 = \sigma X_1, \tag{57}$$

$$Y_n = \alpha Y_{n-1} + \sigma\sqrt{1 - \alpha^2}\, X_n, \qquad (n > 1), \tag{58}$$

where σ^2 is the variance and α the intersample correlation of the wide-sense Markov noise.

Since Y_n is in effect a linear combination of the X_{n-i}, $i = 0, \cdots, n - 1$, it follows that if the X_{n-i} are jointly Gaussian, then the Y_n are jointly Gaussian.

VIII. DISCRETE SMOOTHING FILTERS BASED ON A MODEL USING A LINEAR COMBINATION OF WIDE-SENSE MARKOV PROCESSES

In some cases the simple wide-sense Markov noise model may not be a sufficiently accurate representation of a physical noise process. A better model may be obtained by approximating the known or assumed noise process by a linear combination of wide-sense Markov noise processes. We may approximate the autocorrelation function by wide-sense Markov autocorrelation functions, or, equivalently, we may approximate the normalized power density spectrum by wide-sense Markov normalized power density spectra. For the purpose of making the preceding approximations, we can work with the process as though it were continuous, later introducing the sampling operation. In a parallel to the use of Fourier series to analyze the behavior of a complicated waveform in a linear system, the wide-sense Markov autocorrelation functions may be used to analyze the behavior of a complicated correlated random noise process in a linear discrete system, by applying the principle of superposition. It is possible to synthesize discrete smoothers using this more complex model.

Further, discrete random noise of arbitrary power density spectrum may be generated in an approximate manner for simulation purposes by a suitable linear combination of wide-sense Markov noise components. In the preceding applications, the use of the wide-sense Markov noise components is simpler and more efficient than use of the actual noise process.

There are two types of approximations that can be made. One is a cut-and-try type of approximation in which one tries various linear combinations of wide-sense Markov noise components with the bandwidths of the components not necessarily being integral multiples of some fundamental bandwidth. The other approach is to use a linear combination of orthonormal functions of wide-sense Markov components. In the latter approach, the bandwidths of the components are integral multiples of a fundamental component. In either case, we may write

$$\Phi(\tau) = \sum_{v=1}^{z} A_v \exp(-\Omega_v \mid \tau \mid) \tag{59}$$

or

$$S(\omega) = \sum_{v=1}^{z} A_v \frac{2\Omega_v}{\Omega_v^2 + \omega^2}. \tag{60}$$

Note that the sum of the coefficients A_v must be equal to 1. In the orthonormal approximation,

$$\Omega_v = v\Omega \tag{61}$$

and the A_v will have a definite form. This is shown in the following section.

8.1 *Orthonormal Approximation*

Laning and Battin[14] and Lee[15] have developed orthonormal approximations for an arbitrary autocorrelation function and an arbitrary normalized power density spectrum. These approximations are in terms of components which will be recognized as wide-sense Markov auto-

TABLE I — VALUES OF COEFFICIENTS c_{kv}

k	c_{k1}	c_{k2}	c_{k3}	c_{k4}	c_{k5}
1	1				
2	2	−3			
3	3	−12	10		
4	4	−30	60	− 35	
5	5	−60	210	−280	126

correlation functions and normalized power spectra, respectively. We shall develop the approximation in somewhat different form.

The set of functions

$$\Phi_k(\tau) = \sum_{v=1}^{k} c_{kv} \sqrt{k\Omega} \exp\left(-v\Omega \mid \tau \mid\right) \tag{62}$$

can be made orthonormal on the interval $-\infty < \tau < \infty$ by proper choice of the coefficients c_{kv}. These coefficients are listed in Table I for values of k up to 5.

These functions may be used to form an orthogonal expansion of any piecewise continuous even function (and hence any piecewise continuous autocorrelation function) on the interval $-\infty < \tau < \infty$. We may write

$$\Phi(\tau) = \sum_{k=1}^{\infty} a_k \Phi_k(\tau), \tag{63}$$

where

$$a_k = \int_{-\infty}^{\infty} \Phi(\tau)\Phi_k(\tau)\, d\tau. \tag{64}$$

If we take z terms of the series expansion and denote the corresponding partial sum $\hat{\Phi}(\tau)$, we may group terms to obtain

$$\Phi(\tau) \approx \hat{\Phi}\,(\tau) = \sum_{v=1}^{z} A_v \exp\left(-v\Omega \mid \tau \mid\right), \tag{65}$$

where

$$A_v = \sum_{k=v}^{z} a_k c_{kv} \sqrt{k\Omega}. \tag{66}$$

The coefficients A_v for $z \leqq 5$ are given by (note that if $z < 5$, then $a_k = 0$ for $k > z$)

$$A_1 = \sqrt{\Omega}\,[a_1 + 2\sqrt{2}\,a_2 + 3\sqrt{3}\,a_3 + 4\sqrt{4}\,a_4 + 5\sqrt{5}\,a_5], \tag{67}$$

$$A_2 = -3\sqrt{\Omega}\,[\sqrt{2}\,a_2 + 4\sqrt{3}\,a_3 + 10\sqrt{4}\,a_4 + 20\sqrt{5}\,a_5], \tag{68}$$

$$A_3 = 10\sqrt{\Omega}\,[\sqrt{3}\,a_3 + 6\sqrt{4}\,a_4 + 21\sqrt{5}\,a_5], \tag{69}$$

$$A_4 = -35\sqrt{\Omega}\,[\sqrt{4}\,a_4 + 8\sqrt{5}\,a_5], \tag{70}$$

$$A_5 = 126\sqrt{\Omega}\,[\sqrt{5}\,a_5]. \tag{71}$$

Now let $S_k(\omega)$ be the Fourier transform of $\Phi_k(\tau)$. Then

$$S_k(\omega) = \sum_{v=1}^{k} c_{kv}\sqrt{k\Omega}\,\frac{2v\Omega}{(v\Omega)^2 + \omega^2}. \tag{72}$$

It can be shown, using Parseval's theorem, that the set of functions $\{(1/\sqrt{2\pi})S_k(\omega)\}$ is orthonormal on the interval $-\infty < \omega < \infty$. Hence we may expand any piecewise continuous even function (and thus any piecewise continuous normalized power density spectrum) on this interval. We may write

$$S(\omega) = \sum_{k=1}^{\infty} a_k S_k(\omega), \tag{73}$$

where

$$a_k = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)S_k(\omega)\,d\omega. \tag{74}$$

From Parseval's theorem it will be seen that the a_k in (74) are the same as those in (64). If we take z terms of the series expansion and denote the corresponding partial sum $\hat{S}(\omega)$, we may group terms as before to obtain

$$S(\omega) \approx \hat{S}(\omega) = \sum_{v=1}^{z} A_v\,\frac{2v\Omega}{(v\Omega)^2 + \omega^2}, \tag{75}$$

where the A_v are given by (67) through (71). Note that (65) and (75) form a Fourier transform pair. Thus, if we have approximated a noise process in terms of normalized power density spectra or autocorrelation functions, the alternative approximation can be immediately obtained. The quantity $v\Omega$ represents the half-power point for each wide-sense Markov component.

Simple rules for the selection of Ω for a particular expansion cannot be established; it is a matter of judgment and perhaps trial and error. The fact that it is the half-power point of the fundamental component of the approximation may be of some help. Also, note that as $\tau \to \infty$, $\hat{\Phi}(\tau) \to A_1 \exp(-\Omega\,|\,\tau\,|)$. We might choose Ω that $\hat{\Phi}(\tau)$ and $\Phi(\tau)$ approach zero at the same rate. However, matching autocorrelation functions by means of their tails is not necessarily a desirable approach.

8.2 *System Analysis and Synthesis*

The noise variance ratio of a linear discrete system for which the input noise autocorrelation function has been approximated by a linear combination of wide-sense Markov autocorrelation functions may be obtained by substituting (59) in (5). We have

$$\mu^2 = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \sum_{v=1}^{z} W_i W_j A_v \exp\left(-\Omega_v T \mid i - j \mid\right). \tag{76}$$

Now let

$$\alpha_v = \exp\left(-\Omega_v T\right). \tag{77}$$

Then

$$\mu^2 = \sum_{v=1}^{z} A_v \left[\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} W_i W_j \alpha_v^{\mid i-j \mid} \right]. \tag{78}$$

The expression in brackets represents the noise variance ratio of the linear discrete system when the vth wide-sense Markov component of the noise is the input. Thus, it is clear that the principle of superposition can be used to find the total noise variance ratio. Figs. 7, 8, 11, 12, and 13 may be applied to the wide-sense Markov components individually.

Use of the linear combination noise model will not be profitable in determining the optimum smoother of a class. There is a matrix inversion required [refer to (25)] which is more easily performed directly with the actual autocorrelation matrix. One should keep in mind that the noise variance ratio of a digital smoother is relatively insensitive to departures of the weighting function from the optimum. Hence a smoother optimized for the simple wide-sense Markov model may be satisfactory.

The synthesis of a polynomial smoother based on the linear combination noise model follows the method of Section VI, except that calculation of ϵ is somewhat more difficult, since σ_C^2 must be calculated using (78) and the relevant plots. It should be noted that the estimate of ϵ_T may not be sufficiently accurate to justify the use of the linear combination noise model. One should consider whether or not the simple wide-sense Markov model might be satisfactory.

8.3 *Noise Generation for Simulation*

Discrete stationary random noise of arbitrary autocorrelation function $\Phi(\tau)$ and variance σ^2 may be approximately generated as a linear combination of independent, wide-sense Markov components. Let

$\hat{Z}_i$ represent the ith sample of the approximating linear combination

$$\hat{Z}_i = \sum_{v=1}^{z} b_v Y_{vi}, \tag{79}$$

where Y_{vi} is the ith sample of the vth wide-sense Markov component. This vth component is generated (refer to Section VII) as

$$Y_{v1} = X_{v1}, \tag{80}$$

$$Y_{vi} = \alpha_v Y_{v,i-1} + \sqrt{1 - \alpha_v^2}\, X_{vi}, \qquad (i > 1). \tag{81}$$

Care must be taken that the normalized uncorrelated random numbers X_{vi} are generated in z similar but mutually uncorrelated sequences $\{X_{1i}\}, \{X_{2i}\}, \ldots, \{X_{zi}\}$ to ensure that the sequences $\{Y_{1i}\}, \{Y_{2i}\}, \ldots, \{Y_{zi}\}$ are mutually uncorrelated. Note that each Y_{vi} will have zero mean and unit variance.

To evaluate the coefficients b_v, approximate the autocorrelation function of the arbitrary random noise process by a linear combination of wide-sense Markov components. Thus, from (59) and (77) we obtain

$$\Phi(\,|\,i - j\,|\,T) \approx \hat{\Phi}(\,|\,i - j\,|\,T) = \sum_{v=1}^{z} A_v \alpha_v^{|i-j|}. \tag{82}$$

Now since the Z process is stationary

$$\Phi(\,|\,i - j\,|\,T) = \frac{\operatorname{cov}(\hat{Z}_i, \hat{Z}_j)}{\operatorname{var}(\hat{Z}_i)} = \frac{\sum_{v=1}^{z} b_v^2 \operatorname{cov}(Y_{vi}, Y_{vj})}{\sigma^2} = \frac{\sum_{v=1}^{z} b_v^2 \alpha_v^{|i-j|}}{\sigma^2}. \tag{83}$$

We have set $\operatorname{var}(\hat{Z}_i) = \sigma^2$ since the arbitrary process and its approximation must be matched in variance. Now, equating terms of (82) and (83), we obtain

$$b_v = \sigma\sqrt{A_v}. \tag{84}$$

Thus,

$$\hat{Z}_i = \sigma \sum_{v=1}^{z} \sqrt{A_v}\, Y_{vi}. \tag{85}$$

IX. GLOSSARY OF SYMBOLS

A_v = Coefficient in approximation of power density spectrum or autocorrelation function by linear combination of wide-sense Markov components

B = ΩT_S = noise-smoother bandwidth ratio for fundamental wide-sense Markov component

C = output signal of smoother

E = expected value operator = $\int_{-\infty}^{\infty} dF$

f = dF = probability density function

M = matrix of moments of weighting coefficients

M_q = $\sum_{i=0}^{N-1} i^q T^q W_i$ = qth moment of weighting function

n = present sample

N = number of samples operated on by nonrecursive discrete smoother

p = order of smoother

P = autocorrelation matrix

r = degree of smoother

R = total input signal to smoother

R_{m-i} = total input signal evaluated at $t = (m - i)T$

$\bar{R}$ = desired component of input signal to smoother

$\bar{R}^{(p)}$ = pth derivative of desired component of input signal

$\hat{R}$ = polynomial approximation to desired component of input signal

$\tilde{R}$ = noise component of input signal

$S(\omega)$ = power density spectrum (normalized sense,

$$\left.\frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)\, d\omega = 1\right) \text{; Fourier transform of } \Phi(\tau).$$

t,τ = time variables (seconds)

T = sampling interval (seconds)

T_S = smoothing interval (seconds)

W_i = weighting function of digital filter evaluated at $t = iT$

$X(t)$ = white noise process

$Y(t)$ = wide-sense Markov noise process

$Z(t)$ = general noise process

$\hat{Z}(t)$ = approximation to general noise process

α = $\exp(-\Omega T)$ = intersample correlation for fundamental wide-sense Markov component

α_v = $\exp(-\Omega_v T)$ = intersample correlation for vth wide-sense Markov component

Γ = prediction time

ϵ = total output error of smoother

ϵ_T = truncation error

μ = σ_C/σ_R = output-input standard deviation ratio

σ^2 = noise variance

σ_C^2 = output noise variance

σ_R^2 = input noise variance

$\Phi(\tau)$ = autocorrelation function

ω = angular frequency variable (radians/sec)

Ω = bandwidth of fundamental wide-sense Markov power density spectrum (radians/sec)

Ω_ν = bandwidth of νth wide-sense Markov power density spectrum component (radians/sec)

X. ACKNOWLEDGMENTS

The author wishes to express his appreciation to all those who, directly or indirectly, have contributed to this paper. Particular thanks are due to N. Sharko, M. Iwama, C. E. Paul, V. O. Mowery, J. A. Cochran, and J. P. O'Donohue of Bell Telephone Laboratories, W. S. Sprackle of Texas Instruments, and D. L. Williams of the University of Michigan for their constructive comments. He is most grateful to Miss C. Ziegler and Miss M. Devine of Bell Telephone Laboratories for the programming and many of the illustrations required for this paper.

APPENDIX

Asymptotic Behavior of Smoothers

We will consider the behavior of μ^2 as $N \to \infty$ with Ω and T_S fixed. We shall first find the limits of two expressions which will be needed in finding the limits of the larger noise variance ratio expressions:

$$\begin{aligned}\lim_{N\to\infty} \alpha^{aN+b} &= \lim_{N\to\infty} \exp\left[-\Omega T(aN+b)\right] = \lim_{N\to\infty} \exp\left[-\frac{\Omega T_S(aN+b)}{N-1}\right] \\ &= \lim_{N\to\infty} \exp\left[-\frac{B(aN+b)}{N-1}\right] = \exp\,(-aB),\end{aligned} \tag{86}$$

$$\begin{aligned}\lim_{N\to\infty} N(1-\alpha) &= \lim_{N\to\infty} \frac{1-\exp\left[-B/(N-1)\right]}{1/N} \\ &= \lim_{N\to\infty} \frac{N^2 B \exp\left[-B/(N-1)\right]}{(N-1)^2} = B.\end{aligned} \tag{87}$$

A.1 *Optimum Smoother—0th Order, 0th Degree*

We have, using (33)

$$\lim_{N\to\infty} \mu^2 = \lim_{N\to\infty} \frac{1+\alpha}{N-(N-2)\alpha} = \lim_{N\to\infty} \frac{1+\alpha}{N(1-\alpha)+2\alpha} = \frac{2}{B+2}. \tag{88}$$

A.2 *Optimum Smoother—1st Order, 1st Degree*

We have, using (35) and (36):

$$\begin{aligned} \lim_{N\to\infty} T_S^2\mu^2 &= \lim_{N\to\infty} \frac{12(1+\alpha)[N(1-\alpha)-1+\alpha]}{[N(1-\alpha)+2\alpha][N(1-\alpha)+1+3\alpha]+4\alpha} \\ &= \frac{24B}{(B+2)(B+4)+4} = \frac{24B}{B^2+6B+12}. \end{aligned} \tag{89}$$

A.3 *Zeroth Order Cascaded Simple Averages Smoother*

Using (44) we have

$$\begin{aligned} \lim_{N\to\infty} \mu^2 &= \lim_{N\to\infty} \left\{ \frac{1+\alpha}{N(1-\alpha)} + \frac{2\alpha(\alpha^N-1)}{[N(1-\alpha)]^2} \right\} \\ &= \frac{2}{B} + \frac{2}{B^2}[\exp(-B)-1] = \frac{2}{B^2}[\exp(-B)+B-1]. \end{aligned} \tag{90}$$

A.4 *First Order Cascaded Simple Averages Smoother*

We have, using (45),

$$\begin{aligned} \lim_{N\to\infty} T_S^2\mu^2 &= \lim_{N\to\infty} 2(4.5)^2 \left(\frac{N-1}{N}\right)^2 \\ &\quad \cdot \left\{ \frac{1+\alpha}{3N(1-\alpha)} - \frac{\alpha(\alpha^N - 2\alpha^{2N/3} - \alpha^{N/3} + 2)}{[N(1-\alpha)]^2} \right\} \\ &= \frac{40.5}{B^2} \left\{ -\exp(-B) + 2\exp(-2B/3) \right. \\ &\qquad \left. + \exp(-B/3) + \frac{2}{3}B - 2 \right\}. \end{aligned} \tag{91}$$

A.5 *Second Order Cascaded Simple Averages Smoother*

Using (46) we have

$$\begin{aligned}\lim_{N\to\infty} T_S{}^4\mu^2 &= \lim_{N\to\infty} 2(36)^2\left(\frac{N-1}{N}\right)^4\left\{\frac{1+\alpha}{3N(1-\alpha)}\right. \\ &\qquad \left. + \frac{\alpha(\alpha^N - 2\alpha^{5N/6} - \alpha^{2N/3} + 2\alpha^{N/2} + 3\alpha^{N/3} - 3)}{[N(1-\alpha)]^2}\right\} \\ &= \frac{2592}{B^2}\left\{\exp(-B) - 2\exp(-5B/6) - \exp(-2B/3)\right. \\ &\qquad \left. + 2\exp(-B/2) + 3\exp -B/3) + \frac{2}{3}B - 3\right\}.\end{aligned} \tag{92}$$

REFERENCES

1. Blackman, R. B., Smoothing and Prediction of Time Series by Cascaded Simple Averages, 1960 IRE Convention Record, **8,** Part 2, March 21–24, pp. 47–54. Also published as Bell System Monograph 3678.
2. Lees, A. B., Interpolation and Extrapolation of Sampled Data, IRE Trans. on Information Theory, **IT-2,** 1, March, 1956, pp. 12–17.
3. Johnson, K. R., Optimum, Linear, Discrete Filtering of Signals Containing a Nonrandom Component, IRE Trans. on Information Theory, **IT-2,** 2, June, 1956, pp. 49–55.
4. Blum, M., An Extension of the Minimum Mean Square Prediction Theory for Sampled Input Signals, IRE Trans. on Information Theory, **IT-2,** 3, September, 1956, pp. 176–184.
5. Blum, M., On the Mean Square Noise Power of an Optimum Linear Discrete Filter Operating on Polynomial Plus White Noise Input, IRE Trans. on Information Theory, **IT-3,** 4, December, 1957, pp. 225–231.
6. Doob, J. L., *Stochastic Processes*, Wiley, New York, 1953, p. 90.
7. Doob, J. L., op. cit., pp. 233–234.
8. Blackman, R. B., *Linear Data Smoothing and Prediction in Theory and Practice,* to be published.
9. Ragazzini, J. R., and Franklin, G. F., *Sampled-data Control Systems*, McGraw-Hill, New York, 1958.
10. Hamming, R. W., *Numerical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1962, pp. 143–152.
11. Blackman, R. B., *Linear Data Smoothing and Prediction in Theory and Practice.*
12. Taussky, Olga, and Todd, J., Generation and Testing of Pseudo-Random Numbers, *Symposium on Monte Carlo Methods*, Wiley, New York, 1956, pp. 15–28.
13. Chartres, B. A., An Exact Method of Generating Random Normal Deviates, T. R. No. 5, Contract No. DA-30-069-SC-78130, Division of Applied Mathematics, Brown University, March, 1959.
14. Laning, J. H., and Battin, R. H., *Random Processes in Automatic Control,* McGraw-Hill, New York, 1956, pp. 381–394.
15. Lee, Y. W., *Statistical Theory of Communication,* Wiley, New York, 1960, pp. 460–469.

Command Guidance of *Telstar* Launch Vehicle

By M. J. EVANS, G. H. MYERS and J. W. TIMKO

(Manuscript received March 22, 1963)

The Telstar I satellite was launched into orbit by a three-stage Delta launch vehicle guided by the Bell Telephone Laboratories command guidance system. The Delta program is a National Aeronautics and Space Administration sponsored series of missile flights designed to place various scientific payloads into orbit around the earth. This paper discusses the theory of the guidance equations employed by the command guidance system in the Delta program.

I. INTRODUCTION

The Telstar I satellite was launched into orbit by a three-stage Delta vehicle on July 10, 1962, at the Atlantic Missile Range. The Bell Telephone Laboratories command guidance system, developed for the Air Force, was employed to guide the Delta vehicle. The guidance system is shown in Fig. 1. The missile-borne equipment, housed in the second stage of the Delta vehicle, serves as a radar beacon to provide return pulses to the tracking radar and as the receiving portion of the command data link between the ground and the missile. The tracking radar functions as the transmitting portion of the data link and also serves as a sensor to determine the slant range, azimuth angle, and elevation angle of the missile during its flight. The precision tracking radar and the missile-borne guidance package were manufactured by the Western Electric Co. The guidance computer was designed and manufactured by the Univac Division of the Sperry Rand Corporation.

The three-stage Delta missile, designed by the Douglas Aircraft Company, consists of two liquid propellant stages and a solid propellant third stage. The powered flight portion of the Telstar satellite trajectory is shown in Fig. 2. The guidance system transmits corrective pitch and yaw steering commands during first- and second-stage powered flight. Second-stage engine cutoff is ordered by the guidance system when the

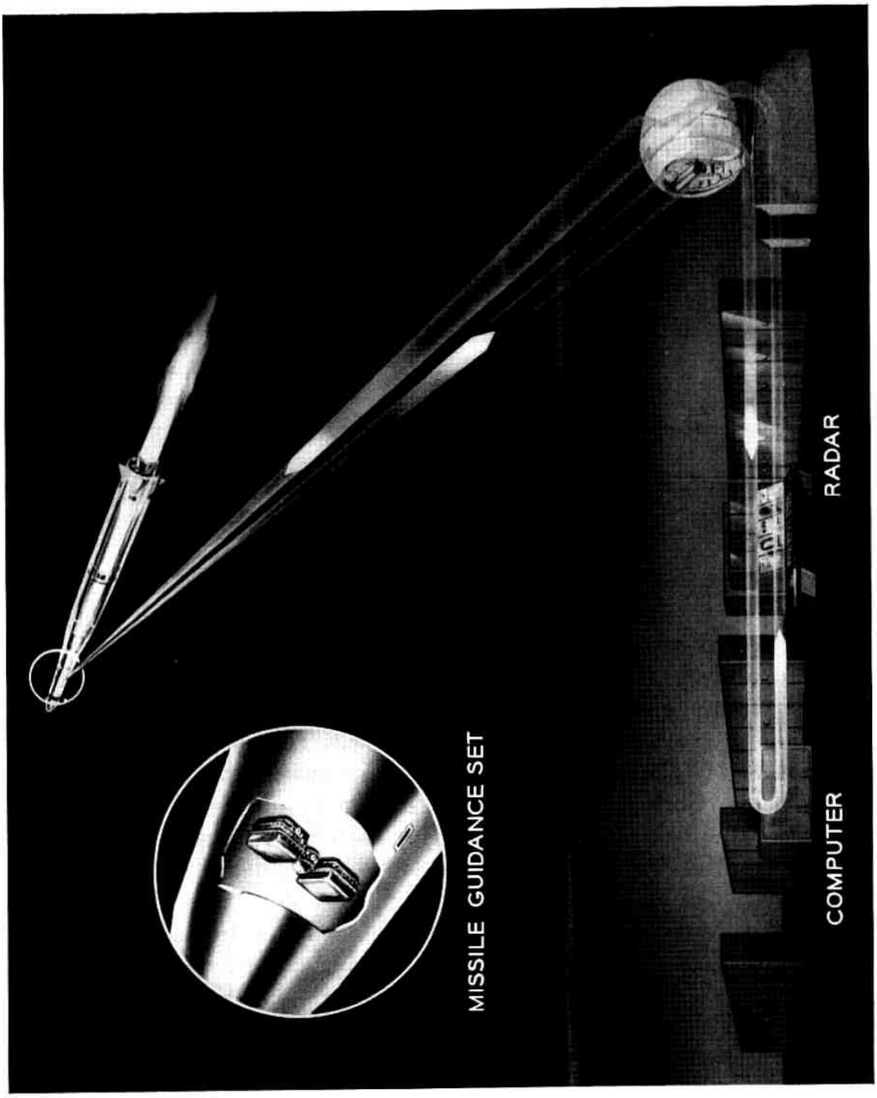


Fig. 1 — Guidance system.

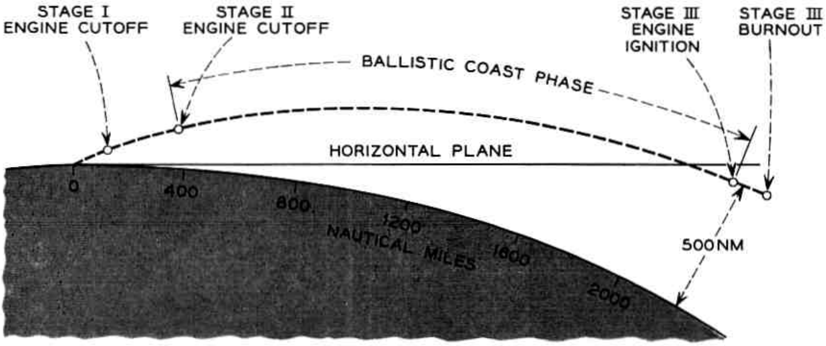


Fig. 2 — Telstar I trajectory.

position and velocity of the missile are such that the addition of the third stage velocity impulse at the end of the ballistic coast phase would yield the desired orbit. The unguided third stage is spin-stabilized to maintain attitude control. For the Telstar satellite trajectory, the third-stage velocity impulse was added at the perigee of the final orbit after a ballistic coast phase of approximately 600 seconds between second-stage cutoff and third-stage ignition. The Telstar satellite was separated from the third stage 120 seconds after third-stage burnout.

The guidance system steering and cutoff commands during the first and second stage ascent are calculated in the guidance computer, using the radar tracking data of the missile's position as the basic input information. The computer is programmed with a set of guidance equations that process the radar data and compute the desired commands to the missile.

This paper contains a description of the theory and design of the guidance equations used in the Telstar satellite flight. Guidance concepts are presented from the point of view of orbital mechanics and control, followed by a description of first- and second-stage guidance. The last section summarizes the results achieved in the Telstar satellite flight.

II. GUIDANCE CONCEPTS

The Keplerian motion of an earth satellite is completely defined by the specification of the vector position and velocity at an epoch. The satellite coordinates at insertion into orbit can be expressed in terms of the spherical coordinate system of Fig. 3 as follows: V_3, γ_3, and β_3 are are the magnitude, the elevation angle above the local horizontal, and the azimuth from North, respectively, of the velocity vector; R_3, λ_3,

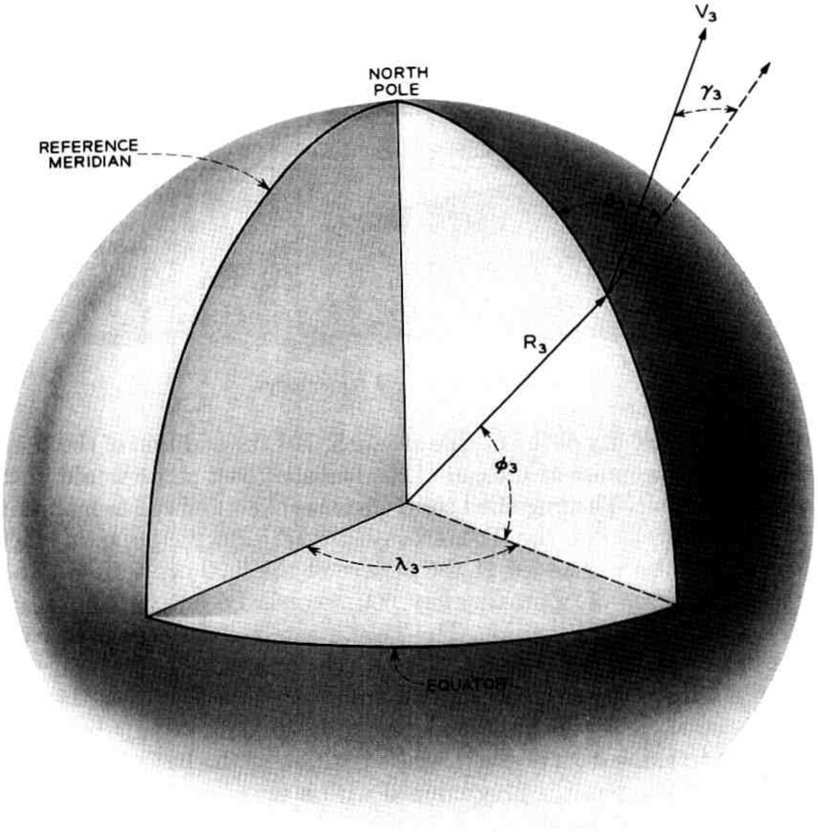


Fig. 3 — Insertion coordinates.

and φ_3 are the radial distance from the earth's center, the longitude, and geocentric latitude, respectively. The elements of the ellipse, i.e., apogee and perigee distances, are determined by V_3, R_3, and γ_3. The orientation of the ellipse relative to the earth in terms of inclination, argument of perigee, and longitude of the ascending node depends, in general, on all the insertion coordinates. It would be necessary to control all six coordinates of the satellite as well as the time of insertion to achieve a specified orbit in inertial space. For earth satellites the requirements on the insertion time are usually not stringent and are largely determined by launch time variations. The problem of guidance thus consists of achieving a specified set of six insertion coordinates.

In the case of the Telstar satellite trajectory, the unguided third stage ignites at a predetermined time on the transfer ellipse following the completion of guidance at second-stage cutoff. Since the characteristics of the third stage are known, fixed relations exist between the insertion coordinates and the coordinates of the missile at second-stage cutoff. The transfer ellipse is defined by the position and velocity vectors, $\mathbf{R}_2$ and $\mathbf{V}_2$, respectively, of the missile at second-stage cutoff. The direction of the velocity impulse added by the third stage is determined by the attitude of the missile's roll axis (axis of thrust application) at second stage cutoff. If we define $\boldsymbol{\varrho}_2$ as a unit vector lying along the roll axis of the missile, the eight independent coordinates of the missile at second-stage cutoff uniquely determine the six insertion coordinates. That is,

$$\{\mathbf{V}_2, \mathbf{R}_2, \boldsymbol{\varrho}_2\} \rightarrow \{V_3, \gamma_3, \beta_3, R_3, \lambda_3, \varphi_3\}. \tag{1}$$

Also the six insertion coordinates specify any six of the cutoff coordinates in terms of the remaining two.

During the first and second stages, guidance of the missile is limited to steering in the pitch and yaw planes and cutting off the second-stage rocket engine. The missile is constrained to fly a predetermined trajectory by pitch and yaw steering during first and second stage. This trajectory, if followed exactly, would yield the coordinates at second-stage cutoff that would produce the desired orbit. However, propulsion system variations, deviations in the attitude control system, and radar noise cause dispersions in the cutoff coordinates from the expected values.

As discussed in Section IV, $|\mathbf{V}_2|$ and $\boldsymbol{\varrho}_2$ can be controlled directly at cutoff to provide direct control over three of the insertion coordinates. Control of $|\mathbf{V}_2|$ by cutoff of the rocket engine provides the most sensitive control of V_3. The two coordinates defining $\boldsymbol{\varrho}_2$ are controlled by pitch and yaw steering. Steering could be based on deviations of position, velocity, or attitude coordinates from the desired reference trajectory. The selection of the coordinates to be controlled by steering is determined by the steering system design, which is based on minimizing the errors in the insertion coordinates. As demonstrated in Section 4.2, control of the pitch and yaw attitude angle, i.e., control of $\boldsymbol{\varrho}_2$, affords the best control over insertion coordinates.

For the Telstar satellite trajectory, $|\mathbf{V}_2|$ and $\boldsymbol{\varrho}_2$ were used to constrain V_3, γ_3, and β_3 at insertion to provide the desired apogee altitude and inclination. The other orbital elements were effectively controlled by steering the missile to the desired reference trajectory during the first-

and second-stage powered flight. The equations relating $|\mathbf{V}_2|$ and $\boldsymbol{\rho}_2$ to the desired orbital conditions, and the equations for pitch and yaw steering derived from the reference trajectory, were programmed into the guidance computer. The targeting task for the Telstar satellite mission was to determine the numerical coefficients for the equations used in first- and second-stage guidance.

III. FIRST-STAGE GUIDANCE

The missile's position as tracked by the radar in slant range, azimuth, and elevation angle is converted to the earth-fixed Cartesian frame shown in Fig. 4. The angle A_0 is selected such that the Y-Z plane is approximately parallel to the pitch plane of the missile and the X axis lies in the yaw plane. The angle E_0 is determined by passing the Y axis through the expected position of the missile at second-stage cutoff.

From approximately 90 seconds after lift-off, pitch and yaw steering orders are transmitted to the missile. Yaw steering is based upon deviations in the $\dot{X}$ velocity from a reference polynomial in $\dot{Y}$. The polynomial is selected to match the desired $\dot{X}$ component of velocity, which is a function of the launch azimuth, the pitch and yaw programmed

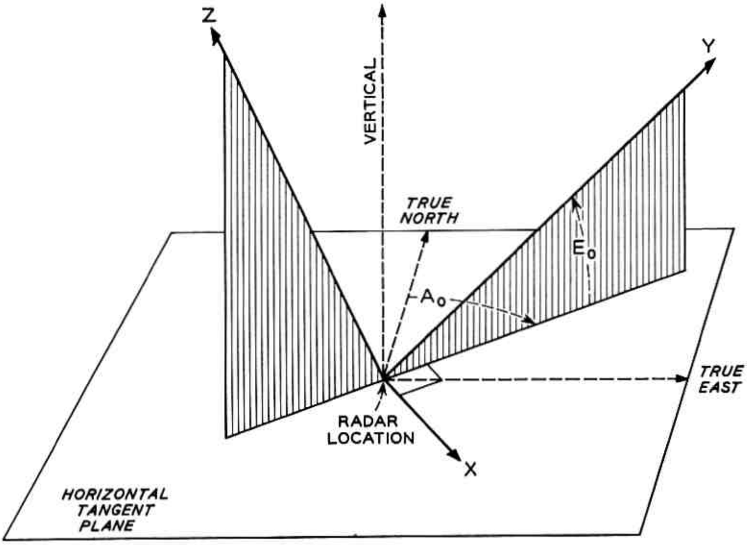


Fig. 4 — Computational coordinate system (X axis is in the horizontal plane; Y axis is at azimuth angle A_0 from true north and at elevation angle E_0 above horizontal plane; Z axis is perpendicular to X and Y axes, completing the right-hand Cartesian coordinate system).

rates in the missile, and the performance of the propulsion system. In a similar manner, pitch steering commands are based on deviations of $\dot{Z}$ from the reference velocity. The steering orders sent to the missile are related to the error signals in a way that balances loop response to missile autopilot errors, propulsion system dispersions, and radar tracking noise. The steering commands are transmitted to the missile via the radar data link at the pulse repetition rate of the radar.

IV. SECOND-STAGE GUIDANCE

4.1 *Second-Stage Cutoff*

In this section the methods used for determining the required coordinates at cutoff will be derived. Since the Telstar satellite was inserted into orbit at perigee, control of apogee distance (r_a) and inclination (i) required that the following relations be satisfied.

$$V_3 = \sqrt{\frac{2K/R_3}{1 + R_3/r_a}} \tag{2a}$$

$$\gamma_3 = 0 \tag{2b}$$

$$\beta_3 = \sin^{-1} (\cos i/\cos \varphi_3). \tag{2c}$$

As discussed in Section II, we can write

$$\alpha_K = f_K (\mathbf{V}_2 , \mathbf{R}_2 , \boldsymbol{\rho}_2), \qquad K = 1 \text{ to } 6 \tag{3}$$

where α_K represents any one of the six insertion coordinates. The six vector functions of (3) depend only on the transfer ellipse and the characteristics of the third stage. If $\mathbf{V}_2$, $\mathbf{R}_2$, and $\boldsymbol{\rho}_2$ are expressed in terms of the coordinate system of Fig. 4, (3) can be expanded in a Taylor series about the expected cutoff coordinates. Linearity studies on the variations in the cutoff coordinates indicate that only the first-order terms in the expansion are significant. Equation (3) simplifies to the form

$$\alpha_K - \alpha_{K_0} = \sum_{i=1}^{8} \frac{\partial \alpha_K}{\partial C_i} (C_i - C_{i_0}) \tag{4}$$

where C_i for $i = 1$ to 8 represents the 8 cutoff coordinates. The partial derivatives in (4) are obtained by perturbing cutoff coordinates and integrating numerically in a digital computer through third-stage burnout to determine the incremental changes in the insertion coordinates. Equation (2) can also be approximated by the first-order terms of a

Taylor expansion and combined with (4), giving three linear equations which can be concisely expressed by the single vector equation below.

$$A_1(\mathbf{V}_2 - \mathbf{V}_{2_0}) + A_2(\mathbf{R} - \mathbf{R}_{2_0}) + A_3(\boldsymbol{\varrho}_2 - \boldsymbol{\varrho}_{2_0}) = 0. \qquad (5)$$

A_1, A_2, and A_3 are 3 × 3 matrices whose (constant) elements are defined by the partial derivatives used in (4) and the expansion of (2). Equation (5) can be solved for any three of the cutoff coordinates as a function of the remaining five. The three coordinates selected are $|\mathbf{V}_2|$, the magnitude of the velocity vector at second-stage cutoff, and the unit vector $\boldsymbol{\varrho}_2$ as defined by the pitch and yaw Euler angles, θ_2 and ψ_2. The missile is cut off when the measured $|\mathbf{V}_2|$ satisfies (5). The attitude constraints on $\boldsymbol{\varrho}_2$ are met by comparing the values of θ_2 and ψ_2 required for the solution of (5) with measured values and commanding the missile to turn by the differences.

4.2 *Steering System Design*

It was shown in the previous section that if orbital elements are to be controlled, the relationships of (5) must be satisfied at second-stage cutoff. Design of the steering system is based not on minimizing the dispersions of the individual variables, $\mathbf{R}_2$, $\mathbf{V}_2$, and $\boldsymbol{\varrho}_2$ at the end of second stage, but rather on minimizing the orbital errors which are caused by errors in these variables.

The pitch and yaw steering system design consists of finding a steering transfer function which minimizes insertion errors due to radar tracking noise and missile dispersions. Missile dispersions include propulsion system variations and errors in the attitude control system of the missile.

4.2.1 *Steering System*

A block diagram of the pitch steering system is shown in Fig. 5. A similar diagram could be drawn for yaw steering.

The computer, operating on the radar data, obtains its measure of the vehicle position in the Z direction, Z_c. After steering has started, the measured trajectory variables are compared with a reference trajectory and corrective pitch turning rates, $\dot{\theta}_g$, are sent to the missile. The pitch rate programmed in the missile, $\dot{\theta}_p$, and the ordered turning rates are the inputs to the autopilot's reference integrating gyro. The function of the autopilot is to align the direction of the acceleration vector, θ, with the desired attitude, as indicated by the gyro output, θ_r.

If the missile's roll axis is aligned with the Y axis of Fig. 4, a small

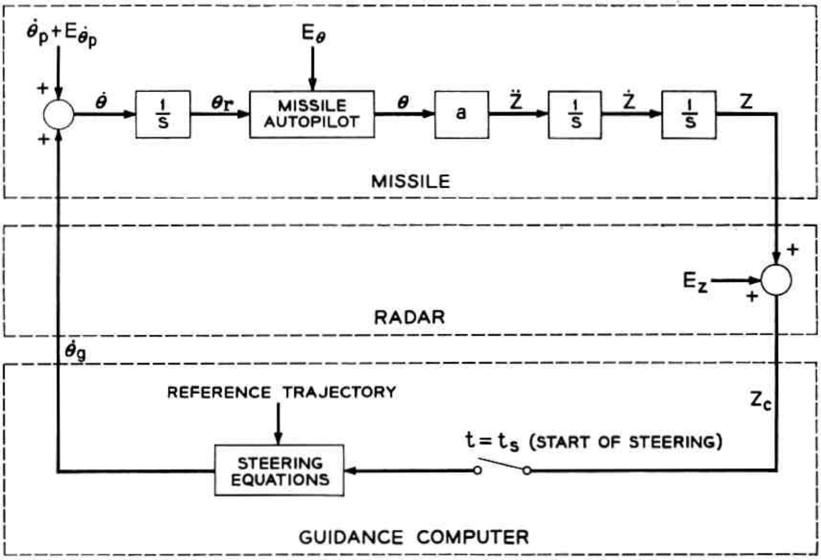


Fig. 5 — Pitch steering system (a is the missile thrust acceleration; $\dot{\theta}_p$ is the programmed turning rate; $E_{\dot{\theta}_p}$ is the error in the programmed rate; E_θ is the error in the thrust vector direction; E_Z is the radar error in measuring the position Z; Z_c is the computer's measurement of Z; and $\dot{\theta}_g$ is the guidance ordered turning rate).

change in pitch attitude, θ, multiplied by the thrust acceleration, a, is the change in Z-direction acceleration, $\ddot{Z}$. Two integrations, represented by their Laplace notation in Fig. 5, then give the vehicle position, Z.

Some of the missile error sources are gyro input errors, propulsion system errors, and misalignments between the direction of the vehicle acceleration, θ, and the reference attitude, θ_r. The gyro input errors are caused by errors in the programmed rate and by electrical or mechanical unbalances. Because the reference trajectory is determined by the programmed turning rates and the expected performance of the propulsion system, propulsion system variations can be replaced in the block diagram by equivalent programmed rate errors. All of the attitude errors, including the integrated rate errors, are caused by slowly varying disturbances.

The radar noise error in Z, E_Z, is the product of the elevation angle error and the distance from the radar to the missile. The radar noise power spectrum has most of its energy at frequencies higher than those encountered in the missile attitude disturbances—a fact important in the steering loop optimization.

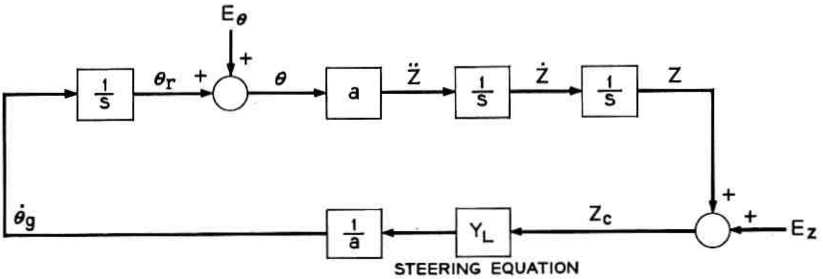


Fig. 6 — Simplified steering loop.

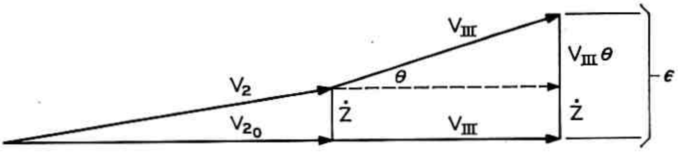


Fig. 7 — Velocity errors.

4.2.2 *Steering Loop Optimization*

In Fig. 6, the programmed pitch rate and the reference trajectory it describes have been removed from the steering loop. All of the missile dispersions are combined in a single attitude error source, E_θ. With the reference trajectory removed, θ, Z, $\dot{Z}$, and $\ddot{Z}$ represent dispersions about the reference values. The quantity to be minimized by steering may be a linear combination of some of the above variables, as will now be shown.

Fig. 7 shows the effect of θ and $\dot{Z}$ on the total insertion velocity vector of the Delta missile with its unguided third stage (assuming θ is small) to be

$$\epsilon = \dot{Z} + V_{III}\theta. \tag{6}$$

The 0 subscript in Fig. 7 indicates reference values, and V_{III} is the magnitude of the velocity increment of the third stage. In complex frequency notation

$$\epsilon(s) = \frac{a}{s}\left(1 + \frac{s}{\omega_2}\right)\theta(s) = \theta(s)\ Y_{\epsilon/\theta} \tag{7}$$

where $\omega_2 = a/V_{III}$, and in general $Y_{a/b}$ is the transfer function from b to a. The total error in terms of the attitude and noise errors can be expressed as

$$\epsilon(s) = Y_{\epsilon/E_Z}\, E_Z + \left(Y_{\epsilon/\theta} + \frac{aY_{\epsilon/E_Z}}{s^2}\right) E_\theta. \tag{8}$$

The total variance of the error, σ_ϵ^2, is the integral over all frequencies of the power spectrum of the error signal. The optimization problem is to determine the transfer function, Y_{ϵ/E_Z}, (which in turn determines the steering equation, Y_L, Fig. 6) that minimizes σ_ϵ^2. The fact that the spectral density of the attitude errors is concentrated at very low frequencies while the radar error power contains higher frequency components suggests that the error may be minimized by a frequency-selective steering equation.

The attitude errors can be approximated by a Markov power spectrum, having a low cutoff frequency,

$$P_{E_\theta}(\omega) = \frac{2\sigma_{E_\theta}^2\,\omega_\theta}{\omega^2+\omega_\theta^2}, \tag{9}$$

where ω_θ is very small. Using techniques described by Bode and Shannon[1] for minimizing the mean squared error, the optimum steering equation is:

$$Y_L = \frac{-s\left[1+\left(2T+\frac{1}{\omega_2}\right)s\right]}{2T^2+\frac{2T}{\omega_2}+\left(T^3+\frac{2T^2}{\omega_2}\right)s+\frac{T^3}{\omega_2}s^2} \tag{10}$$

where

$$T = \left(\frac{\sigma_{E_Z}^2}{a^2\,\sigma_{E_\theta}^2\,\omega_\theta\,\omega_E}\right)^{1/6}. \tag{11}$$

For the Delta second stage, $1/\omega_2$ equals V_{III}/a. If typical numbers for V_{III}, a, and T are inserted, then $1/\omega_2 \gg T$, which leads to the approximation:

$$Y_{L_{\text{II}}} = \frac{-s^2}{2T\left(1+Ts+\frac{T^2}{2}s^2\right)}. \tag{12}$$

The two differentiations of the position data indicated in (12), together with the division by a, Fig. 6, convert the position data to attitude data which is in turn smoothed to give the ordered turning rates.

In the guidance equations the pitch and yaw attitude of the missile is determined by operating on the position data of the missile. The total acceleration of the missile is computed by taking second differences of the position data expressed in the coordinate system of Fig. 4. Gravity, Coriolis and centripetal accelerations are subtracted from the total acceleration to obtain the thrust component of acceleration. Since the thrust acceleration vector is aligned with the missile's roll axis and the missile is roll stabilized, the pitch and yaw attitude angles can be computed. These are compared with the desired values to obtain pitch and yaw attitude errors. The desired attitude angles are computed from polynomials in time designed to match the programmed rates built into the missile's attitude control system.

In this paper, all transfer functions have been given in continuous form (as functions of s), although the transfer functions are actually converted to digital form (functions of $z = e^{sT}$) before being programmed into the digital computer. This conversion does not significantly change any of the relations given here, because the frequency ranges important in the above equations are much less than half of the sampling frequency.

4.2.3 *Accuracy Improvement by Final Value Control*

Equation (12) is the optimum steering equation for minimizing the mean squared error. However, the injection errors are functions of the error at one instant of time, second-stage cutoff. Examination of the system's response suggests means of improving the steering design.

The slowly varying attitude disturbances represented by a Markov distribution may also be represented by initial attitude and velocity errors and a constant gyro drift. The system's response to these errors, derived from (12), is shown in Fig. 8 as a function of time normalized with respect to the steering parameter T.

Since the Delta design is not restricted to a continuous control system, velocity errors at second-stage cutoff in excess of those obtained with the optimum continuous design can be allowed. These velocity dispersions can be measured and used to aim the third stage in accordance with (5). The velocity errors shown in Fig. 8 are either constant or linearly increasing with time and can be measured by filtering the position data. The steady-state attitude error response to gyro drift is a constant. If gyro drift is a major attitude error source, continuous closed loop control of attitude may also be relaxed, and the constant attitude error can be measured with a polynomial filter.[2] The errors can be corrected just prior to cutoff by turning the missile. The reason for relaxing the attitude and velocity dynamic control [making T greater than the "optimum" value in (11)] is that doing so decreases the attitude error caused by the effect of radar noise on the steering system.

The outputs of the attitude and velocity filters used to measure the dynamic residuals also have errors caused by the tracking noise. These decrease as the filter length, T_F, is increased. Since the attitude error signals involve second derivatives of tracking data, the attitude filters require long smoothing times. However, if the filters are started too

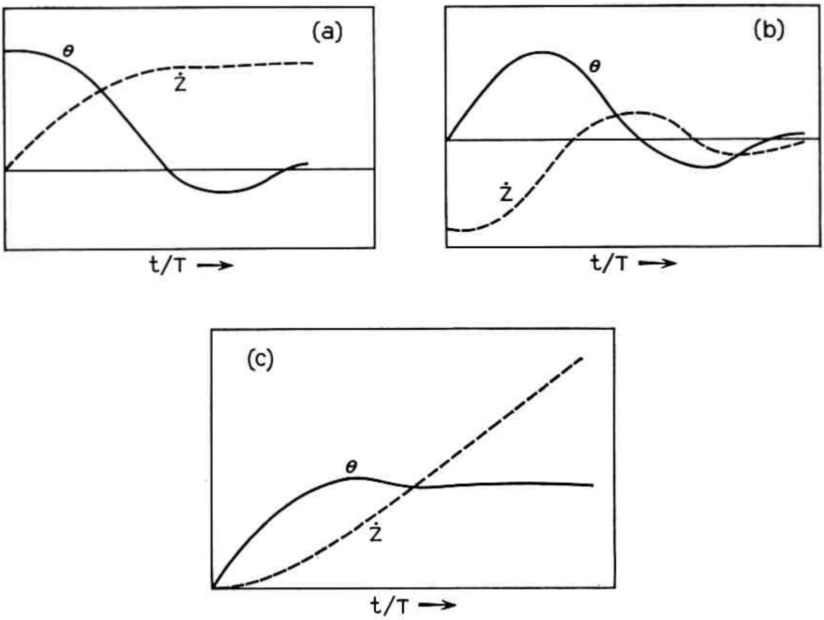


Fig. 8 — Dynamic steering response: (a) response to initial attitude error; (b) response to initial velocity error; and (c) response to gyro drift.

early in the second stage, they are subject to increased dynamic error because the steering system has not yet settled to the steady state (Fig. 8). Thus, optimum filter design is dependent on the steering parameter T.

For the Delta missile, the critical design criterion was that of attitude control. That is, for a system with optimum attitude control, the optimum velocity filter has negligible error, and the measured dispersions can be used as a basis for an attitude correction (aiming the third stage).

For given dynamic error sources and filter length, the total attitude error, after correction of the measured error, is the combination of the filter's dynamic error and the noise error of the filter plus the steering noise error, Fig. 9. The steering system's noise error is a function of T and decreases as T increases. The filter's noise error is independent of T but decreases as the filter length, T_F, increases. The filter's dynamic error is a function of both T_F and the transient response of the steering system as determined by T. Because the steering and filter noise errors are highly correlated, they are added together directly and root sum squared with the filter's dynamic error. T and T_F are chosen from Fig. 9 to minimize the total attitude error.

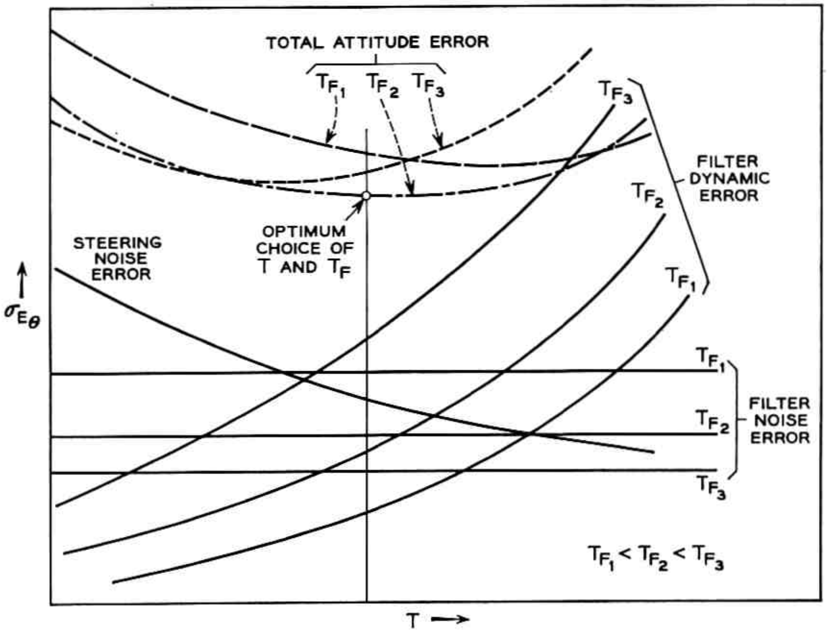


Fig. 9 — Final attitude errors as functions of T and T_f.

In summary, the Delta second-stage guidance equations give closed-loop control of missile attitude. Relatively large dynamic errors in velocity and attitude are permitted in order to decrease noise errors, because the dynamic errors can be measured and corrected. An attitude order to correct the measured attitude error and compensate for velocity and position dispersions in accordance with (5) is sent to the missile shortly before second-stage cutoff.

V. *TELSTAR* I SATELLITE FLIGHT RESULTS

The steering history of the first and second stage of the Telstar I satellite flight is given in Figs. 10 and 11. The maximum steering orders during first-stage guidance were 3.8 degrees in pitch and 1.6 degrees in yaw. Steering in second-stage guidance reached a peak of 0.25 degree in pitch and 1 degree in yaw. Corrective commands of 0.35 and 0.13 degree were issued in pitch and yaw, respectively, just prior to second-stage cutoff. The second-stage engine was cut off at 269.6 seconds from liftoff, as compared to a preflight value based on average propulsion performance of 273.3 seconds. The orbital results as determined by NASA are given in Table I with the preflight reference values.

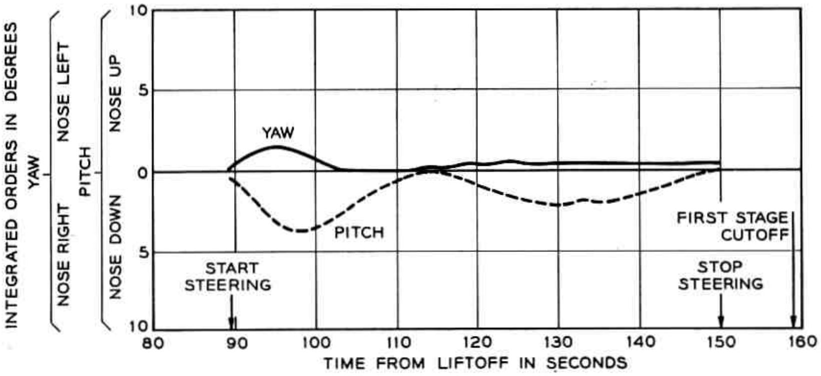


Fig. 10 — First-stage steering.

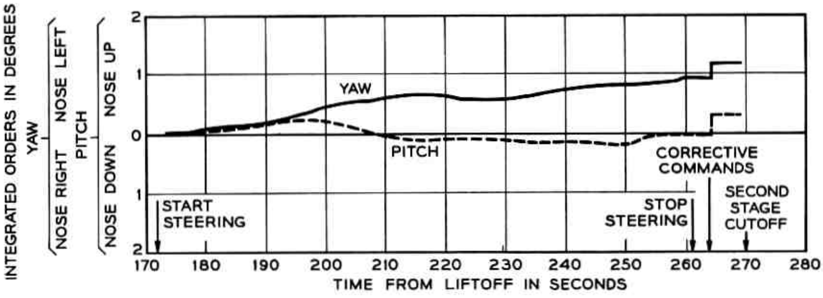


Fig. 11 — Second-stage steering.

Table I — Orbital Results

	Preflight Reference	NASA Minitrack
Altitude		
Apogee (nm)	3000	3044
Perigee (nm)	500	515
Eccentricity	0.2407	0.242
Period (minutes)	156.47	157.82
Inclination (degrees)	45	44.79
Argument of perigee (degrees)	166.6	165
Right ascension of ascending node (degrees)	203.6	204

REFERENCES

1. Bode, H. W., and Shannon, C. E., A Simplified Derivation of Linear Least Square Smoothing and Prediction Theory, Proc. I.R.E., **38,** 1950, pp. 417–425.
2. Blackman, R. B., Smoothing and Prediction of Time Series by Cascaded Simple Averages, Trans. I.R.E., **CT-7,** 1960, pp. 136–143.
3. Myers, G. H., and Thompson, T. H., Guidance of Tiros I, ARS Journal, **31,** 1961, pp. 636–640.

Spin Decay, Spin-Precession Damping, and Spin-Axis Drift of the *Telstar* Satellite

By E. Y. YU

(Manuscript received March 6, 1963)

Dynamical problems of the spin-stabilized Telstar satellite, characterized by spin decay, spin-precession damping, and spin-axis drift, are analyzed in this paper. Both the eddy-current torques and the magnetic torques, which cause the above three phenomena, are evaluated. By extrapolation from the observed data, the characteristic time of the nearly exponential spin decay of the satellite is estimated to be about 330 days. A linear analysis of the precession damper is made, and the results are compared with experiments, showing that the satellite precession angle will diminish by a factor of e in a maximum time of 30 minutes. A qualitative description is given to illustrate the fundamental mechanism of spin-axis drift. Results of these analyses can be applied to any spin-stabilized satellite.

I. INTRODUCTION

For spin stabilization of a communications satellite, it is required that the satellite be statically and dynamically balanced so as to make the principal axis of maximum moment of inertia coincide with the axis of symmetry of the antenna pattern, about which the satellite is given an initial spin. This principal axis, referred to henceforth as the spin axis, is in line with the invariant angular momentum vector and is thus fixed in direction, as desired, in an inertial space, provided there are no external torques acting on the spinning satellite. However, as the satellite is spinning and traveling in the geomagnetic field, eddy-current and magnetic torques continuously act on the satellite so that the angular momentum changes its magnitude and direction, as characterized by spin decay and spin precession. As a consequence of spin decay, the satellite becomes less stable for the same external disturbing torques, and a tumbling motion may eventually result. The precession of the spin axis about the instantaneous angular momentum vector will cause wobbling of the antenna pattern, indicating that the precession should necessarily be dissipated by means of a damping mechanism. Because of the continuous action of the torques, the angular momentum continuously changes its direction in the inertial space; meanwhile, the spin axis precesses about it and, due to the precession damping, tends to align with it. Thus, there results a gradual drift in direction of the spin axis (sometimes called long-term precession), as already observed on the Telstar satellite.

The above dynamics problems of the satellite — namely, spin decay, spin-precession damping, and spin-axis drift — are studied in this paper. In the discussion of spin decay, we will indicate the nature of the retarding torques resulting from eddy currents and magnetic hysteresis losses, analyze the observed spin decay phenomenon, and compute the $1/e$ characteristic time of the exponential decay. For the spin precession, a linear analysis of the precession damping mechanism will be given, and an experimental comparison of the damping time will be outlined. Only a short descriptive analysis is given to the problem of spin-axis drift, since an exact evaluation of the rate and pattern of drift deserves a separate computer study.

It is shown in the following that whenever a spinning rigid body undergoes energy losses (e.g., from internal friction) the axis of maximum moment of inertia or the spin axis, $\hat{z}$, and the angular velocity, $\boldsymbol{\omega}$, will tend to align with the angular momentum, $\mathbf{J} = \boldsymbol{\Phi} \cdot \boldsymbol{\omega}$, where

$$\boldsymbol{\Phi} = I_x \hat{x}\hat{x} + I_y \hat{y}\hat{y} + I_z \hat{z}\hat{z}$$

($\hat{x}$, $\hat{y}$, and $\hat{z}$ are the unit vectors along the principal axes) is the moment of inertia dyadic. If I_z of the spin axis is the minimum, the spinning motion is still stable; however, any energy dissipation will not reduce precession but make the spin axis deviate away from $\mathbf{J}$. A simple proof of the above is given in the following. The kinetic energy, E, of a spinning satellite with precession is written as

$$2E = \boldsymbol{\omega} \cdot \boldsymbol{\Phi} \cdot \boldsymbol{\omega}. \tag{1}$$

By substituting $\mathbf{J} = \boldsymbol{\Phi} \cdot \boldsymbol{\omega}$ and $\boldsymbol{\omega} = \mathbf{J} \cdot \boldsymbol{\Phi}^{-1}$ into the above, E can be expressed in terms of the angular momentum,

$$\mathbf{J} = J(\cos \xi \hat{x} + \cos \eta \hat{y} + \cos \theta \hat{z}),$$

where $\cos \xi$, $\cos \eta$, and $\cos \theta$ are the direction cosines of $\mathbf{J}$ in the body coordinates

$$2E = \mathbf{J} \cdot \boldsymbol{\Phi}^{-1} \cdot \mathbf{J} = J^2 \left(\frac{1}{I_x} \cos^2 \xi + \frac{1}{I_y} \cos^2 \eta + \frac{1}{I_z} \cos^2 \theta \right) \tag{2}$$

or, as $\cos^2 \theta = (1 - \cos^2 \xi - \cos^2 \eta)$

$$2E = J^2 \left[\frac{1}{I_z} + \left(\frac{1}{I_x} - \frac{1}{I_z} \right) \cos^2 \xi + \left(\frac{1}{I_y} - \frac{1}{I_z} \right) \cos^2 \eta \right]. \qquad (3)$$

If there exists energy dissipation at a rate assumed to be so slow as to produce no torque on the satellite, the torque-free rigid body motion of the satellite will tend toward the state of minimum energy. It is observed from (2) and (3) that the minimum kinetic energy state occurs at $\theta = 0°$ (or $\xi \equiv \eta \equiv 90°$) for $I_z > I_x , I_y$, or at $\theta = 90°$ (either $\xi \equiv 0$ or $\eta \equiv 0$) for $I_z < I_x , I_y$. This proves that in order to reduce the precession angle, θ, by means of energy dissipation, I_z of the spin axis should be the maximum.

II. SPIN DECAY

Spin decay results from energy losses in the form of both eddy currents and magnetic hysteresis when the satellite is spinning in the geomagnetic field. It is shown in the following that the hysteresis losses in the magnetic materials are much smaller than the eddy-current losses in the conducting materials at high spin rates simply because the geomagnetic field in the Telstar satellite orbit is relatively weak, ranging from 0.04 to 0.4 oersted.* Magnetic materials are contained in the nickel-cadmium cells of the battery, the magnetic shielding on circuit components, etc. No measurement of the magnetic hysteresis loops has been made on the components actually used in the satellite. However, measurements on similar components contemplated for use on a proposed satellite prior to the Telstar satellite have been made, based on which it was estimated (for different orbit parameters and a different spin-axis attitude from those of the Telstar satellite) that the time-average hysteresis loss is $W = 1.6$ ergs per cycle of rotation. It is believed that the above value can be used for a conservative estimate of spin-decay rate for the Telstar satellite because it contains less magnetic materials than had been anticipated. According to this value of W, the spin decay of the satellite due to hysteresis losses alone is only about 1.5 rpm per year.

Eddy currents are generated essentially in the following parts: (a) the aluminum shell of the electronics chassis, (b) the frames of square magnesium tubing and equatorial antennas, and (c) the magnesium chassis frame assembly. An estimate of the eddy-current torque can be made if the electronics chassis is approximated as a thin spherical shell and the

* These figures are based on a spherical harmonic representation of the geomagnetic field with Hensen and Cain coefficients for the Epoch, 1960.

last two items in the above are approximated as circular loops of wire. The eddy-current torque acting on a thin spherical shell spinning at an angular velocity $\boldsymbol{\omega}$ can be shown to be

$$\mathbf{T}_1 = p_1\mathbf{B} \times (\mathbf{B} \times \boldsymbol{\omega}) \tag{4}$$

where $\mathbf{B}$ is the geomagnetic induction and $p_1 = (2\pi/3)a^4\sigma\, d$, with a = radius, d = thickness, and σ = volume conductivity. The above expression is correct only when the square of the nondimensional quantity, $\frac{1}{3}\mu a\sigma\omega d$ (μ = free-space permeability), is negligibly small compared to unity, which is found to be true in the present case. The above torque can be resolved into two components, i.e., the component parallel to $\boldsymbol{\omega}$

$$T_{\parallel} = -p_1 B_{\perp}^{\ 2}\omega \tag{5}$$ *

which tends to retard $\boldsymbol{\omega}$($B_{\perp}$ = component of $\mathbf{B}$ normal to $\boldsymbol{\omega}$) and the component normal to $\boldsymbol{\omega}$

$$T_{\perp} = p_1 B_{\parallel} B_{\perp}\omega \tag{6}$$

which contributes to the precession of the satellite ($B_{\parallel}$ = component of $\mathbf{B}$ parallel to $\boldsymbol{\omega}$). In the case of a circular loop, it can be easily shown that the time-average eddy-current torque, acting on a circular loop spinning about a diameter, tends only to retard $\boldsymbol{\omega}$; i.e.

$$\mathbf{T}_2 = -p_2 B_{\perp}^{\ 2}\boldsymbol{\omega} \tag{7}$$

where $p_2 = A^2/2R'$ (A = loop area, $R' = l'/\sigma A_w$ = total resistance of the loop of wire, A_w = cross-sectional area of the wire, l' = length of the loop). The above expression is correct only when the square of the nondimensional quantity, $\omega L'/R'$ (L' = inductance), is negligibly small compared to unity, which is found to be the case here. The other component normal to $\boldsymbol{\omega}$ has a zero time-average value.

In general, the eddy-current retarding torque acting on a conducting body spinning in a magnetic field is proportional to $B_{\perp}^{\ 2}$ and ω, or can be written approximately as

$$\mathbf{T}_r = -pB_{\perp}^{\ 2}\boldsymbol{\omega} \tag{8}$$

where p is a constant and is determined by the conducting material and its geometry. Thus, for the Telstar satellite, if the electronics chassis is approximated as a thin spherical shell (of 9.5-inch radius and 0.1-inch thickness) and the frames are approximated as circular loops of wire, then the retarding torque, T_r, acting on the satellite is the sum of $T_{\parallel}$ in

* This expression is the same as that given in Ref. 1, p. 417, problem 12, after the square of the nondimensional quantity, $\frac{1}{3}\mu a\sigma\omega d$, is neglected.

(5) and T_2 in (7) or $p = p_1 + p_2$. It is calculated that $p_1 = 684$, $p_2 = 256$, or $p = 940$ meter^4/ohm in mks units. This value of p is of course too low, because many small conducting parts have not been considered in the calculation. From the expressions of p_1 and p_2, one notices that p_1 is proportional to the fourth power of the radius of a spherical shell and p_2 to the square of the loop area. For this reason, the Telstar satellite was insulated at the equatorial antennas in such a way that the outer shell does not constitute a large continuous surface and that the frames do not form large continuous loops.

The magnitude of p in (8) for the Telstar satellite can be measured by rotating a magnetic field normal to the spin axis while the angular deflection of a torsion wire, which suspends the satellite along the spin axis, is recorded to determine the drag torque. Such an experiment has been devised by M. S. Glass and D. P. Brady. Measurements made on the prototype give $p = 1355$ meter^4/ohm ± 15 per cent. This measured value is believed to be somewhat high, because the magnetic field applied in the measurements was as high as 25 to 100 oersteds (at 23.4 rpm) in order to give significant angular deflection readings of the suspension wire; thus, the measured drag torque unavoidably includes losses due to full hysteresis loops described in the magnetic materials. In the actual case, the magnetic field along the orbit is only 0.04–0.4 oersted, and the losses due to minor hysteresis loops are much smaller. Besides, since the electromagnetic characteristics of the satellite may be different from one model to another, the value of p measured on the prototype may not be applied to the Telstar satellite with good accuracy. Nevertheless, it is believed that the value of p in meter^4/ohm is bounded below by 940 and above by 1560. A later calculation by extrapolation from the observed data on the satellite showed that p is approximately equal to 1110. A further refinement of the evaluation of p might have been obtained from the instantaneous spin-decay rate which can be determined from the telemetry solar aspect data. However, as the obtained instantaneous spin-decay rate was too low to give any significant reading, such an attempt failed to yield any results.

Because of proper functioning of the precession damper, the Telstar satellite is now spinning nearly about its principal z-axis of maximum moment of inertia. In this case, $B_\perp$ in (8) can be approximated as the component of B normal to the z-axis. Obviously, $B_\perp$ is a function of time due to (a) the rotation of the slightly inclined geomagnetic field about the earth's spin axis, (b) the anomalies of the geomagnetic field, (c) the gradual drift in direction of the satellite spin axis, and (d) the variation of orbital parameters due to the oblateness of the earth, notably the apsidal advance and the nodal regression (see Fig. 1). For the same rea-

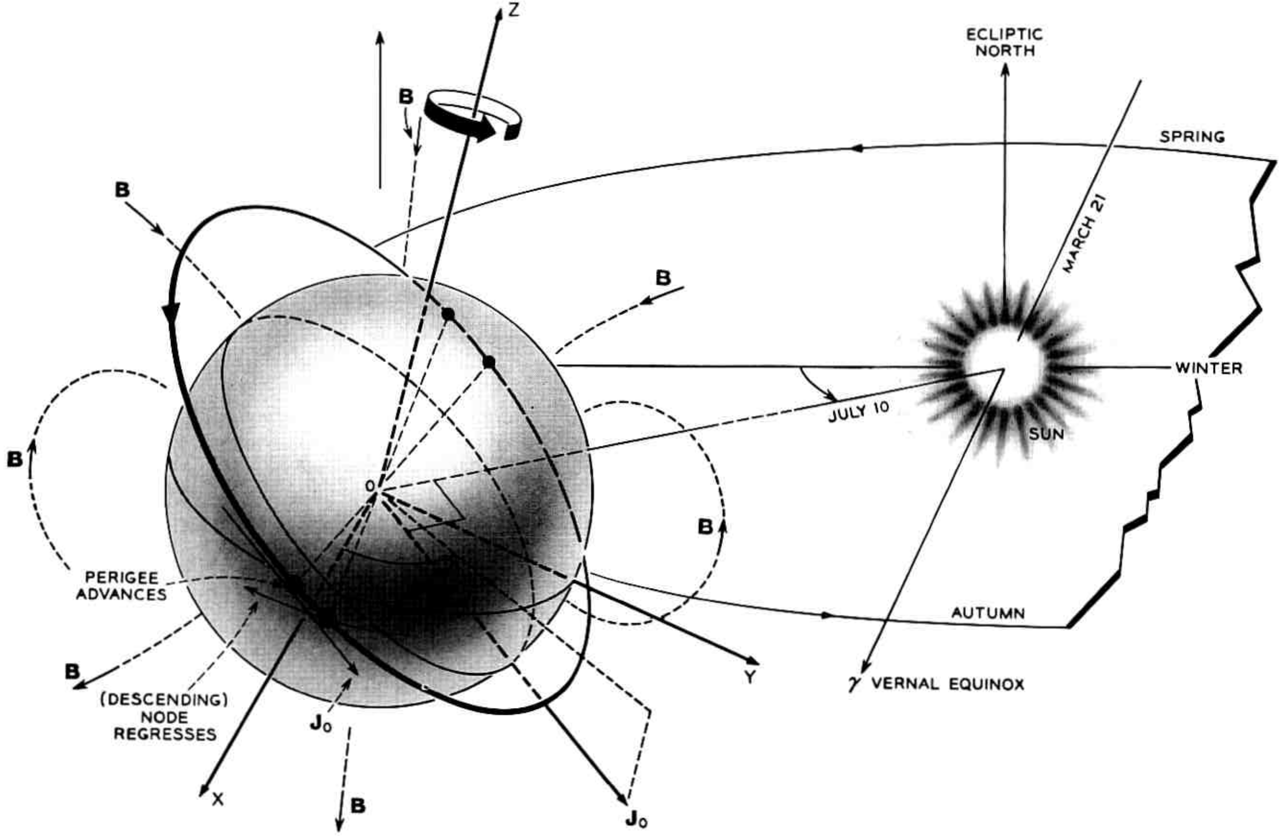


Fig. 1 — Initial Telstar satellite orbit in nonrotating coordinates, O-XYZ.

sons, the magnetic hysteresis loss per cycle of rotation, W, is also a function of time. The decay of spin rate due to both eddy-current and hysteresis losses can be determined from the following equation

$$I_z\dot{\omega} = -pB_\perp^2(t)\omega - \frac{1}{2\pi}W(t) \tag{9}$$

or, upon integration,

$$\omega = \exp\left(-\int_0^t (p/I_z)B_\perp^2(t)\,dt\right) \cdot\left[\omega_0 - \int_0^t \frac{W(t)}{2\pi I_z}\exp\left(\int_0^t (p/I_z)B_\perp^2(t)\,dt\right)dt\right]. \tag{10}$$

where $\omega_0 = \omega(t = 0)$. Let us now define day-average values $\overline{B_\perp^2}$ and $\overline{W}$ as

$$\overline{B_\perp^2} = \frac{1}{t}\int_0^t B_\perp^2(t)\,dt \tag{11}$$

and

$$\overline{W} = \frac{\int_0^t W(t)\exp\left(\int_0^t (p/I_z)B_\perp^2(t)\,dt\right)dt}{\int_0^t \exp\left(\int_0^t (p/I_z)B_\perp^2(t)\,dt\right)dt}. \tag{12}$$

Then (10) can be written as

$$\omega = \left[\omega_0 + \frac{1}{2\pi}\frac{\overline{W}(t)}{p\overline{B_\perp^2}(t)}\right]e^{-t/\tau} - \frac{1}{2\pi}\frac{\overline{W}(t)}{p\overline{B_\perp^2}(t)} \tag{13}$$

where

$$\tau = I_z/p\overline{B_\perp^2}(t). \tag{14}$$

The time t in (13) is in units of days, and $\overline{B_\perp^2}(t)$ and $\overline{W}(t)$ are functions of t. Note that if the term $\overline{W}/2\pi p\overline{B_\perp^2}$, which is much smaller than ω_0 in the case of the Telstar satellite, is neglected in (13), the spin decay is exponential with time with, however, a time-dependent τ.

A plot of $\overline{B_\perp^2}(t)$ is given in Fig. 2 from the day of launch (July 10, 1962) up to December 31, 1962. The day-average $\overline{B_\perp^2}(t)$ is obtained by taking the arithmetic mean of the time-average values of $B_\perp^2$ per pass for approximately nine passes a day. The latter values are computed*

* Computations were provided by J. D. Gabbe.

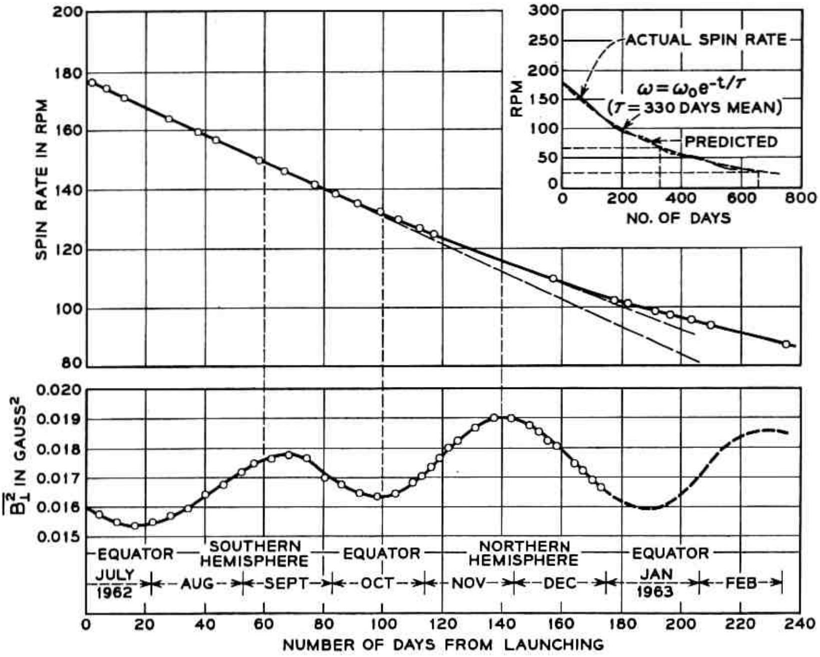


Fig. 2 — Spin rate and $\overline{B_\perp{}^2}$ vs time.

from a spherical harmonic representation of the geomagnetic field, taking into account the continuous variations of the spin-axis direction and of the orbital parameters. Because of these combined effects, the variation of $\overline{B_\perp{}^2}$ is sinusoidal with time with, however, variable amplitude and period. The major contribution to this variation is believed to be due to the apsidal advance in the orbital plane. In order to see this, let us plot in Fig. 3 the time-average values of $B_\perp{}^2$ per pass versus the geographical longitude of the perigee in the beginning of the pass on the days of July 15, 18 and 21, 1962. The latitudes of the perigee on those three days were ±5° within the geographical equator. It is shown on these curves that $B_\perp{}^2$ is relatively low when the perigee falls in the region over South America where the geomagnetic field strength is depressed. The center of this region falls at approximately 25°S latitude and 45°W longitude. (See Ref. 2 for details.) Fig. 3 is a typical example, which shows how the magnitude of $B_\perp{}^2$ depends critically on the position of perigee, although the variation of $B_\perp{}^2$, which also depends on other factors as previously stated, does not necessarily follow the same pattern as in Fig. 3 when the perigee is in other positions. At any rate,

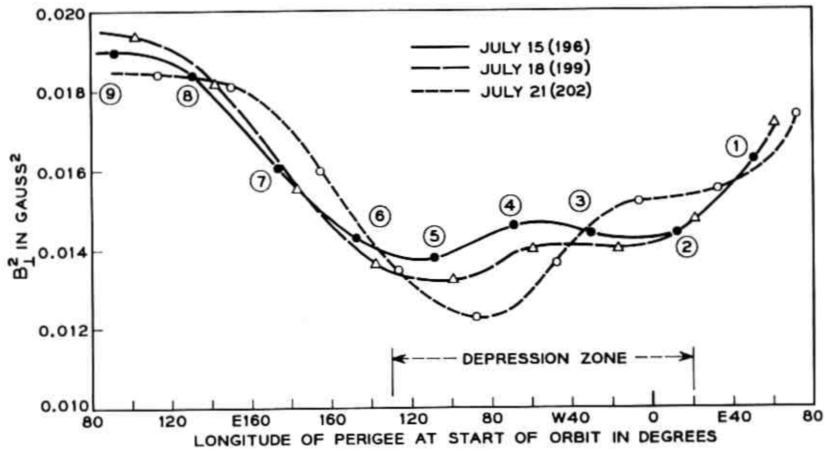


Fig. 3 — $B_\perp^2$ vs longitude of perigee at start of orbit.

the change of position of the perigee is a determining factor for the variation of the day-average $\overline{B_\perp^2}$ and of the observed* spin-rate curve, as plotted in Fig. 2. The initial perigee of the Telstar satellite orbit (593 miles in altitude) was north of the geomagnetic equator on the day of launch, as indicated in Fig. 1. The perigee advances in the direction of the orbital motion at a rate of approximately 2° per day. When the perigee was crossing the geomagnetic equator southward in the orbital plane, $\overline{B_\perp^2}$ first decreased and then increased, as indicated in the initial part of the $\overline{B_\perp^2}$ curve in Fig. 2. In the ascending part of the curve up to September 7, 1962, $t = 60$ days, more spin decay occurred than would result from an exponential decay produced by a constant $\overline{B_\perp^2}$ when the perigee was over the equator. This is why the corresponding part of the spin-rate plot is nearly a straight line instead of an exponential decay curve. From $t = 60$ days to $t = 100$ days, while the perigee was advancing northward toward the geomagnetic equator, $\overline{B_\perp^2}$ was leveling off and then declining, resulting in an exponential decay as shown in the part of the spin-rate plot deviating from the extension of the straight line. From $t = 100$ days to $t = 140$ days, the perigee was entering the northern hemisphere, again getting into a stronger geomagnetic field indicated by the increasing $\overline{B_\perp^2}$. As a result, this part of the spin-rate curve becomes nearly a straight line again, though of a dif-

* The spin rate of the Telstar satellite was measured by J. S. Courtney-Pratt and his coworkers by means of the glint method (see Ref. 3 for details). It was also determined by C. C. Cutler and W. C. Jakes by way of measuring the frequency of the ripple in the amplitude of the radio signal received from the satellite.

ferent slope than the first one. Then the spin decay became exponential again when the perigee was moving southward toward the equator from $t = 140$ to $t = 190$ days. All these indicate that the actual spin-rate curve would wiggle about a mean exponential curve as shown in Fig. 2. Note from Fig. 2 that $\overline{B_\perp^2}$ is lower when the perigee is in the southern hemisphere. This is due to the zone of depressed geomagnetic field strength previously mentioned. It is believed that, as the perigee keeps advancing in the orbital direction, $\overline{B_\perp^2}(t)$ will continue to vary sinusoidally with time. Nevertheless, complete values of $\overline{B_\perp^2}$ cannot be predicted for the entire useful life of the satellite because the spin axis changes its direction continuously because of perturbation of the electromagnetic torques as well as occasional operation of the torque coil,* and because variation of the orbital parameters cannot be predicted accurately. Therefore, exact evaluation of the $1/e$ characteristic time of the nearly exponential spin decay cannot be obtained. Nevertheless, it may be approximately evaluated as follows.

First, let us determine the value of p of the Telstar satellite from the observed spin-decay rate in the first 35 days, which is practically linear with time, with $\omega_0 = 18.67$ rad/sec (178.33 rpm) at $t = 0$ and $\omega = 16.9$ rad/sec (161.2 rpm) at $t = 35$ days. Let us take $\overline{B_\perp^2}$ and $\overline{W}$ as constants in (13) and, as t/τ is small, we may expand $\exp(-t/\tau)$ up to the first-order term in t/τ

$$\omega \approx \omega_0 \left(1 - \frac{p\overline{B_\perp^2}}{I_z} t\right) - \frac{1}{2\pi} \frac{\overline{W}}{I_z} t.$$

Substituting into the above with $\omega = 16.9$, $\omega_0 = 18.67$ rad/sec, $I_z = 5.61$ kg-m^2, $\overline{W} = 1.6 \times 10^{-7}$ joules, and $t = 35$ days $= 3.024 \times 10^6$ sec, we find that

$$p\overline{B_\perp^2} = 1.748 \times 10^{-7} \text{ weber}^2/\text{ohm}.$$

If we take $\overline{B_\perp^2}$ to be 1.572×10^{-10} weber^2/meter^4, then $p = 1110$ meter^4/ohm. Now, we note from Fig. 2 that $\overline{B_\perp^2}$ varies between 0.0154 and 0.0190 gauss^2. If we take the average value of $\overline{B_\perp^2}$ for the entire useful life of the Telstar satellite, denoted as $\overline{\overline{B_\perp^2}}$, to be 0.0177 gauss^2 ± 2 per cent, then the average $1/e$ characteristic time of the exponential spin decay without considering hysteresis losses is found to be

$$\tau = \frac{I_z}{p\overline{\overline{B_\perp^2}}} = 330 \text{ days} \pm 2 \text{ per cent} \tag{15}$$

* The torque coil consists of 200 turns of 32-gauge copper wire wound around the equator of the satellite. When current is turned on at a desired time, the magnetic moment of the coil will interact with the geomagnetic induction to produce torque for correction of the spin-axis direction.

for $p = 1110$ meter^4/ohm. An exponential curve based on this mean $1/e$ time is plotted in Fig. 2; in addition, the actual spin-rate curve is also drawn (not to scale) in order to show that the latter curve is fluctuating about the mean exponential curve at a period of about 180 days. If the hysteresis losses are taken into account, the $1/e$ characteristic time is expressed as

$$\tau_h = \tau \ln \left[(\omega_0 + \overline{\overline{\overline{W}}}/2\pi p\overline{\overline{B_\perp{}^2}})/(\omega_0/e + \overline{\overline{\overline{W}}}/2\pi p\overline{\overline{B_\perp{}^2}})\right] \tag{16}$$

where $\overline{\overline{\overline{W}}}$ is the average value of $\overline{\overline{W}}(t)$ for the entire useful life of the satellite. Let $\overline{\overline{\overline{W}}}$ be 1.6 ergs per cycle of rotation for a conservative estimate, then $\tau_h = 327$ days ± 2 per cent.

Based on the above range of the exponential decay rate, the satellite will spin at about 20 rpm at the end of two years from the day of launch. If the equatorial antennas had not been insulated, a separate calculation indicates that the spin rate after two years would be only about 3 rpm, which seems too low to insure attitude stabilization.

III. SPIN PRECESSION AND PRECESSION DAMPING

Before analyzing precession damping, let us first consider precessional motion of a spinning satellite produced by torques acting transversely to the spin axis. Suppose that the satellite, assumed here to be a rigid body, is initially spinning about its z-axis so that its initial angular momentum is $\mathbf{J}_0 = I_z\boldsymbol{\omega} = I_z\omega\hat{z}$. When a transverse torque $\mathbf{T}$ is acting on the satellite for a time interval Δt, $\mathbf{J}_0$ changes to $\mathbf{J}$ by an amount $\Delta\mathbf{J}$, which is equal to the impulse $\mathbf{T}\Delta t$, as shown in Fig. 4(a). The satellite will then perform precession with the spin axis, $\hat{z}$, and the angular velocity, $\boldsymbol{\omega}$, no longer aligned with $\mathbf{J}$.

The torques causing precession consist of (a) gravitational torque, (b) eddy-current torque, and (c) torque of interaction between the residual magnetic dipole moment, $\mathbf{M}$, and the geomagnetic field, $\mathbf{H}$. The components of the gravitational torque, $3(gR_0^2/\rho^3)\hat{\rho} \times \boldsymbol{\Phi}\cdot\hat{\rho}$ (ρ = geocentric distance, g = gravitational acceleration at earth's surface, and R_0 = earth's radius) normal to the spin axis are proportional to $(I_z - I_x)$ or $(I_z - I_y)$ and are found to have a maximum value of 0.65×10^{-6} ft-lb at perigee of the Telstar satellite orbit (based on the measured values of $I_x = 3.7140$, $I_y = 3.8252$, and $I_z = 4.1412$ slug-ft^2). The eddy-current torque given in (6) can reach a maximum value of 0.56×10^{-6} ft-lb (for $p_1 = 684$ meter^4/ohm and $| B_\parallel B_\perp | = 0.6 \times 10^{-10}$ weber^2/meter^4). The maximum magnitude of the magnetic torque, $\mathbf{M} \times \mathbf{H}$, is as high as 13.2×10^{-6} ft-lb (for $H = 31.84$ amp-turns/m or

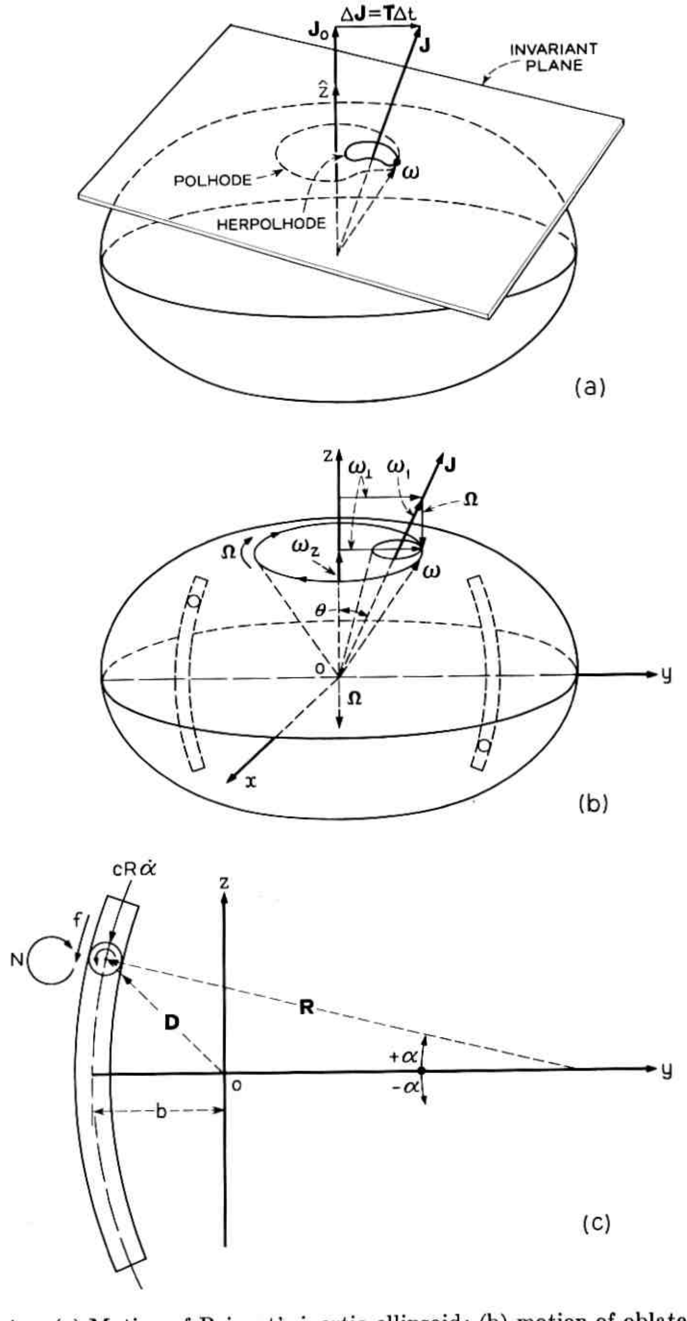


Fig. 4 — (a) Motion of Poinsot's inertia ellipsoid; (b) motion of oblate spheroidal rigid body; and (c) motion of ball in the precession damper.

0.4 oersted and $M = 0.562 \times 10^{-6}$ weber-meter*). Thus, both the gravitational and eddy-current torques are at least one order of magnitude smaller than the magnetic dipole moment torque.

For the analysis of precession damping, let us assume that **J** is temporarily an invariant, since **T** is so small (about 10^{-5} ft-lb maximum, as given above) that it takes a minimum time of about 1.5 days for **J** to change its direction by one degree, whereas the $1/e$ characteristic precession damping time, τ_p, is only of the order of 30 minutes, as will be shown later. Thus, within the time interval comparable to τ_p the precessional motion can be treated as torque-free. Such a motion can be pictured by Poinsot's geometrical construction (see Ref. 4, p. 161) in which the satellite's inertia ellipsoid rolls without slipping on the invariant plane, which is a plane normal to **J** and tangent to the ellipsoid at a fixed distance from the origin of the ellipsoid [see Fig. 4(a)]. The curve traced out by the point of contact on the ellipsoid, known as the polhode, is the locus of the tip of $\boldsymbol{\omega}$ in the body, while the curve on the invariant plane, known as the herpolhode, is the locus of the tip of $\boldsymbol{\omega}$ in the inertial space. To simplify the analysis of the precession damping, we further assume that the satellite is symmetric about its spin axis, or the inertia ellipsoid is an oblate spheroid with a transverse moment of inertia $I(=I_x = I_y) < I_z$. In this case, the precession motion can be visualized, as shown in Fig. 4(b), as a body cone, $zO\omega$, rotating at an angular velocity, $\boldsymbol{\Omega}$, on an immovable space cone, $JO\omega$, which is rotating at an angular velocity, $\boldsymbol{\omega}_1$, along the fixed direction of **J**. The line of tangency between these two cones is the instantaneous axis of rotation of the body or the angular velocity, $\boldsymbol{\omega}$, which is the sum of $\boldsymbol{\Omega}$ and $\boldsymbol{\omega}_1$. The analytic solution of such a torque-free precession motion of an oblate spheroidal body is obtained (see Ref. 4, p. 162) for the angular velocity $\boldsymbol{\omega} = \omega_x\hat{x} + \omega_y\hat{y} + \omega_z\hat{z}$ expressed in the body coordinates with components

$$\begin{aligned} \omega_x &= \omega_\perp \sin \Omega t \\ \omega_y &= \omega_\perp \cos \Omega t \\ \omega_z &= \bar{\omega}_z(=\text{constant}) \end{aligned} \tag{17}$$

where

$$\Omega = \left(\frac{I - I_z}{I}\right)\omega_z \tag{18}$$

* Measured by M. S. Glass and D. P. Brady.

and, as is apparent from Fig. 4(b)

$$\omega_\perp = (\omega_z - \Omega) \tan \theta = \frac{I_z}{I} \omega_z \tan \theta = \text{constant}. \qquad (19)$$

Here, it is obvious that the precession angle, θ, between the spin axis, $\hat{z}$, and **J** is a constant, providing there exists no precession energy dissipation.

The precession energy is the difference between two energy states with and without precession. We use the same assumption as in Section I: that the precession energy dissipation produces negligibly small torque to the rigid body motion. Thus, during the time interval when the precession is substantially damped out, the angular momentum can be treated as an invariant. The kinetic energy in the presence of precession is

$$E = \tfrac{1}{2} I \omega_\perp^2 + \tfrac{1}{2} I_z \omega_z^2$$

where $\omega_\perp^2 = \omega_x^2 + \omega_y^2$, and as shown in Section I the minimum energy state occurs when the precession is completely damped out, i.e.

$$E_{\min} = \tfrac{1}{2} I_z \omega_z'^2$$

where ω_z' can be found from the invariant angular momentum

$$J^2 = I^2 \omega_\perp^2 + I_z^2 \omega_z^2 = I_z^2 \omega_z'^2$$

or

$$\omega_z'^2 = \left(\frac{I}{I_z}\right)^2 \omega_\perp^2 + \omega_z^2 .$$

Therefore, the precession energy is

$$E_p = E - E_{\min} = \frac{1}{2}\left(1 - \frac{I}{I_z}\right) I \omega_\perp^2$$

or, by virtue of (19)

$$E_p = \frac{1}{2}\left(\frac{I_z}{I} - 1\right) I_z \omega_z^2 \tan^2 \theta. \qquad (20)$$

Note that when $I = I_z$ for a spherical satellite there is no precession energy.

To dissipate the precession energy, the satellite is equipped with a pair of curved aluminum tubes filled with neon gas at one atmosphere, each containing a tungsten ball of radius r and mass m [see Fig. 4(c)]. The ball is slightly smaller than the inside diameter of the tube, and the curved tubes are installed concavely towards the spin axis with their bisecting radii of curvature, R, perpendicular to the spin axis at the center of mass of the satellite. When the satellite is rotating precisely about its spin axis, the balls are stationary at the middle of the tubes. However, when precession occurs — i.e., when $\boldsymbol{\omega}$ is precessing along the body cone or following the polhode in the inertia ellipsoid — the balls are forced to move back and forth against the viscous friction of the gas as well as the inviscid friction between the balls and the tubes. Hence, the precession energy is dissipated into heat through the resistances to the motion of the balls, resulting in the attenuation of the precession angle, θ, or in the realignment of the spin axis, $\hat{z}$, and $\boldsymbol{\omega}$ with $\mathbf{J}$.

To derive the equations of motion of the ball, we assume that the ball rolls on the tube without sliding. The equation of the ball's rotational motion about its center of mass can be immediately written in terms of the angle, α, from the y-axis (the tubes are assumed to lie in the yz-plane)

$$\frac{2}{5}\,mr^2\left(\frac{R}{r}\,\ddot{\alpha}\right) = fr - N \tag{21}$$

where f is the force acting at the point of contact, and N the resistance moment due to rolling friction. The position vector of the center of mass of the ball, as shown in Fig. 4(c), is given in the body coordinates as

$$\mathbf{D} = [R(1 - \cos\alpha) - b]\hat{y} + R\sin\alpha\hat{z}. \tag{22}$$

The equation of the translational motion of the ball is then

$$m\,\frac{d^2\mathbf{D}}{dt^2}\cdot\hat{q} = -f - cR\dot{\alpha} \tag{23}$$

where $\hat{q} = \sin\alpha\hat{y} + \cos\alpha\hat{z}$ is the tangential unit vector in the direction of $\dot{\alpha}$ and c the coefficient of viscous friction. In performing the differentiation of $\mathbf{D}$ with respect to time, one should be aware of the fact that the angular velocity, $\boldsymbol{\omega}$, as given in (17) in the body coordinates of a precessing body, is changing in direction in an inertial space and its time derivative is $\dot{\boldsymbol{\omega}} = \boldsymbol{\omega} \times \boldsymbol{\Omega}$, where $\boldsymbol{\Omega} = \Omega\hat{z}$. Therefore, it should be noted, for example, that $(d/dt)\hat{y} = \boldsymbol{\omega} \times \hat{y}$ and $(d^2/dt^2)\hat{y} = (\boldsymbol{\omega} \times \boldsymbol{\Omega}) \times \hat{y} + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \hat{y})$. Upon differentiating $\mathbf{D}$ in (22) twice with respect to time, substituting into (23), and using (21) to eliminate f, we obtain a nonlinear second-order equation for α

$$\begin{aligned}\ddot{\alpha} - \frac{5}{7R}\,\omega_z^{\,2}[R(1 - \cos\alpha) - b]\sin\alpha + \frac{5}{7}\,\omega_y(\omega_z - \Omega)\sin^2\alpha \\ - \frac{5}{7}\,\omega_\perp^{\,2}\sin\alpha\cos\alpha + \frac{5}{7R}\,\omega_y(\omega_z + \Omega)[R(1 - \cos\alpha) - b]\cos\alpha \\ = -\frac{5c}{7m}\,\dot{\alpha} - \frac{\dot{\alpha}}{|\,\dot{\alpha}\,|}\,\frac{5N}{7mrR}\end{aligned} \tag{24}$$

where the sign of N has been chosen in such a way as to make the resistance moment always oppose the rotational motion of the ball, and $\omega_y = \omega_\perp \cos\Omega t = (\omega_z - \Omega)\tan\theta\cos\Omega t$. As the radius of curvature, $R(= 15\text{ ft})$, of the tube used on the Telstar satellite is much larger than its length, $L(\approx 1.4\text{ ft})$, the maximum angle of α is very small, viz., $\alpha_m = L/2R = 0.0465$ radian $\ll 1$, where α_m is the subtended half angle of the curved tube. In order for the balls to move back and forth without bottoming with the ends of the tubes, the precession angle, θ, should be of the same order as α. Thus, for small α and θ, (24) can be linearized to the following

$$\ddot{\alpha} + 2n\dot{\alpha} + P^2\alpha + \frac{\dot{\alpha}}{|\,\dot{\alpha}\,|}\,K = q\cos\Omega t \tag{25}$$

where $2n = 5c/7m$, $P^2 = 5b\omega_z^{\,2}/7R$, $K = 5N/7mrR$, and $q = (5b/7R)\cdot\theta(\omega_z^{\,2} - \Omega^2)$. Because of the presence of the rolling friction term, the above equation can be solved only for each half cycle.

An experiment has been conducted by the author to determine the resistance moment, N, that the tungsten ball encounters when rolling on the aluminum tube. For convenience, N is expressed in terms of a resistance force, F, acting at the center of the ball: i.e., $N = Fr$, and F is determined experimentally to be $F = 0.0002$ lb. In another experiment devised to measure the coefficient of viscous friction, c, for the 0.484-inch ball moving in the tube (nominal inside diameter $= 0.495$ in.) filled with neon gas at one atmosphere, it is found that $c = 0.00193$ lb-sec/ft. The ratio of the energy dissipation per cycle due to the viscous friction (in the steady-state forced oscillation case) and that due to the rolling friction can be shown to be $E_v/E_r = |\,\pi cR\Omega\alpha_m/4F\,|$. With $\alpha_m = 0.0465$ radian, $R = 15$ ft, $|\,\Omega\,| = 1.50$ rad/sec, corresponding to the case of the Telstar satellite at 178.33 rpm, the above ratio is about 8. This indicates that energy dissipation per cycle due to rolling friction is approximately one order of magnitude smaller than that due to viscous friction at the indicated spin rate. If the rolling friction term is neglected, (25) becomes

$$\ddot{\alpha} + 2n\dot{\alpha} + P^2\alpha = q\cos\Omega t \tag{26}$$

for which the steady-state forced oscillation solution is

$$\alpha = \alpha_0 \cos(\Omega t - \beta) \tag{27}$$

where

$$\begin{aligned} \alpha_0 &= \theta\left(1 - \frac{\Omega^2}{\omega_z^2}\right)\left[\left(1 - \frac{\Omega^2}{P^2}\right)^2 + \frac{4n^2\Omega^2}{P^4}\right]^{-\frac{1}{2}} \\ \cos\beta &= (P^2 - \Omega^2)[(P^2 - \Omega^2)^2 + 4n^2\Omega^2]^{-\frac{1}{2}} \\ \sin\beta &= 2\Omega n[(P^2 - \Omega^2)^2 + 4n^2\Omega^2]^{-\frac{1}{2}}. \end{aligned} \tag{28}$$

The energy dissipation due to viscous friction per period of oscillation of the ball (or per period of precession) is

$$E_v = \int_0^T cR^2\dot{\alpha}^2\,dt = cR^2\alpha_0^2\Omega \int_0^{2\pi} \sin^2(\Omega t - \beta)\,d(\Omega t) = \pi c R^2\alpha_0^2\Omega. \tag{29}$$

The time-average rate of energy dissipation per period of precession appears to be

$$\overline{\frac{dE_v}{dt}} = \frac{E_v}{\left(\frac{2\pi}{\Omega}\right)} = \frac{1}{2}\,cR^2\Omega^2\alpha_0^{\,2}. \tag{30}$$

Equating the negative of the above to the rate of change of the precessional energy, E_p, in (20) for the case of a small angle, $\tan\theta \approx \theta$, an equation for the change of the precession angle θ is obtained

$$\left(\frac{I_z}{I} - 1\right) I_z\omega_z^{\,2}\theta\dot{\theta} = -\frac{1}{2}\,cR^2\Omega^2\alpha_0^{\,2}. \tag{31}$$

Substituting into the above with α_0 given by (28) and integrating, we obtain an exponential decay of θ with time

$$\theta = \theta_0 e^{-t/\tau_p} \tag{32}$$

where $\theta_0 = \theta(t = 0)$ and τ_p is the characteristic time given as

$$\tau_p = \frac{5I_z}{7nmR^2(\lambda - 1)\lambda^2(2 - \lambda)^2}\left[\left(1 - \frac{\Omega^2}{P^2}\right)^2 + \frac{4n^2\Omega^2}{P^4}\right] \tag{33}$$

with $\lambda = I_z/I(>1)$. If the time-average rate of energy dissipation per cycle due to rolling friction, $4FR\alpha_0\,|\,\Omega\,|/2\pi$, is included in (31), the solution for θ becomes

$$\theta = (\theta_0 + h)e^{-t/\tau_p} - h \tag{34}$$

where

$$h = \frac{10F}{7\pi mnR\omega_z(\lambda - 1)\lambda(2 - \lambda)}\left[\left(1 - \frac{\Omega^2}{P^2}\right)^2 + \frac{4n^2\Omega^2}{P^4}\right]^{\frac{1}{2}}. \tag{35}$$

The $1/e$ characteristic time in this case is

$$\tau_p' = \tau_p \ln\left(\frac{1 + \dfrac{\theta_0}{h}}{1 + \dfrac{\theta_0}{he}}\right) \tag{36}$$

which involves the initial angle, θ_0. Because of the rolling friction, the precession will be damped out within a finite time, i.e., $\theta = 0$ at $t = \tau_p \ln\,(1 + \theta_0/h)$. A numerical computation, corresponding to the Telstar satellite physical constants, shows that h is considerably smaller than one degree. This indicates that if the precession angle, θ, is substantially greater than one degree, the $1/e$ characteristic time given in (33) for viscous friction alone should be used for convenience, since it is independent of the initial condition.

From (26) it is seen that if the natural frequency, P, is made equal to the frequency of the forcing term, Ω, i.e., if

$$R = \frac{5}{7(\lambda - 1)^2}\,b \tag{37}$$

then the oscillating motion of the ball is in resonance with the precession motion of the satellite. As a consequence, the $1/e$ characteristic time becomes much shorter

$$\tau_{pr} = \frac{28nI_z(\lambda - 1)}{5mb^2\lambda^2\omega_z^2(2 - \lambda)^2}. \tag{38}$$

Unfortunately, such a tuned damper cannot be obtained for the Telstar satellite, since the ratio of moments of inertia ($I/I_z = 0.897$, 0.925 or $\lambda = 1.114$, 1.08) is close to unity, and as the tubes are placed outside of the electronics package, b cannot be made too small. Therefore, in view of (37), a tuned damper for the Telstar satellite would have to be of an exceedingly large radius of curvature ($R = 57$–117 ft for $b = 1.046$ ft). In this case, the tubes become practically straight, and the motion of the balls may not necessarily be at resonance with the precession of the satellite. The equation of motion of a ball in a straight tube can be easily obtained by multiplying (26) through with R and then letting $R \to \infty$ or $P \to 0$; in a similar way, the $1/e$ characteristic time can be obtained. Nevertheless, such a straight or nearly straight tube damper will not be used for the essential reason that, in case of its misalignment with the spin axis, no damping whatsoever will be obtained when the precession angle is smaller than the misalignment angle. Another type of damper can be obtained if the curved tube is placed concavely away from the spin axis. The equation of motion of the ball in such a tube can be easily shown to be the same as (26) except that the sign of the α term is negative, hence forming an unstable system. The ball, which rests at one end of the tube, will move toward the other end at a high speed when a component of $\boldsymbol{\omega}$ is tangent to the former end of the tube and is large enough to make the centrifugal force overcome the static friction. The ball will move back and forth twice in each period of precession. Let the kinetic energy of the ball when it reaches the other end of the tube equal the centrifugal force times the distance traveled by the ball perpendicular to $\boldsymbol{\omega}$. If we assume that the kinetic energy of the ball is completely absorbed by bottoming at each end of the tube, it can be shown that the decay of the precession angle is parabolic with time. Such a damper can effectively reduce the precession even when the ratio I/I_z is close to unity, but it will not damp out a precession angle less than about three degrees; thus, it was not adopted for use with the Telstar satellite.

After comparing the advantages and disadvantages of the several dampers discussed, the untuned concave damper shown in Figs. 4(b) and (c) was finally chosen for the Telstar satellite, although its damping time is somewhat larger than that of the others. This choice was made because the theoretical $1/e$ characteristic damping time, given in (33), is calculated to be a maximum of about three minutes for a ratio of I/I_z up to 0.95, a spin-rate range of 20–180 rpm, and for the parameters given below. Such a damping time is acceptable even if it is one order of magnitude larger, in view of the slow rate of change of the angular momentum due to the small transverse torques previously calculated. The chosen parameters are: $R = 15$ ft (a large value, although the tube still has noticeable curvature), $m = 0.0021$ slug (for two tungsten balls of 0.484 in. diameter; tungsten is chosen for its high density), and $n = 0.65 \text{ sec}^{-1}$ or $c = 0.00193$ lb-sec/ft (corresponding to neon gas, which is chosen for its high viscosity). [Note: For a tuned damper with $1/e$ damping time as given in (38), a gas with low viscosity should preferably be used.]

It is necessary that the theoretical $1/e$ characteristic time τ_p of the untuned concave damper should be compared with experimental results for the following reasons. Formula (33) is obtained from the linearized analysis of the motion of the ball under the assumptions of small amplitude of the motion and an axisymmetric spinning body. In the actual case, the Telstar satellite is dynamically not axisymmetric. Also, as the tube is limited in length, bottoming will occur when the precession angle is larger than about 3.5°; the motion of the ball will then be disturbed and will not follow (26). An experiment has been devised by G. T. Kossyk which consists of an air-bearing supported flywheel ($I_z = 5.14$, $I_{\max} = 4.675$, $I_{\min} = 4.404$ slug-ft^2) mounted with two precession dampers as shown in Fig. 5. The flywheel is driven to reach a certain initial speed about a skew axis making a desired angle with the axis of symmetry, which is the principal axis of maximum moment of inertia. As soon as the drive is released, the spinning flywheel performs a preces-

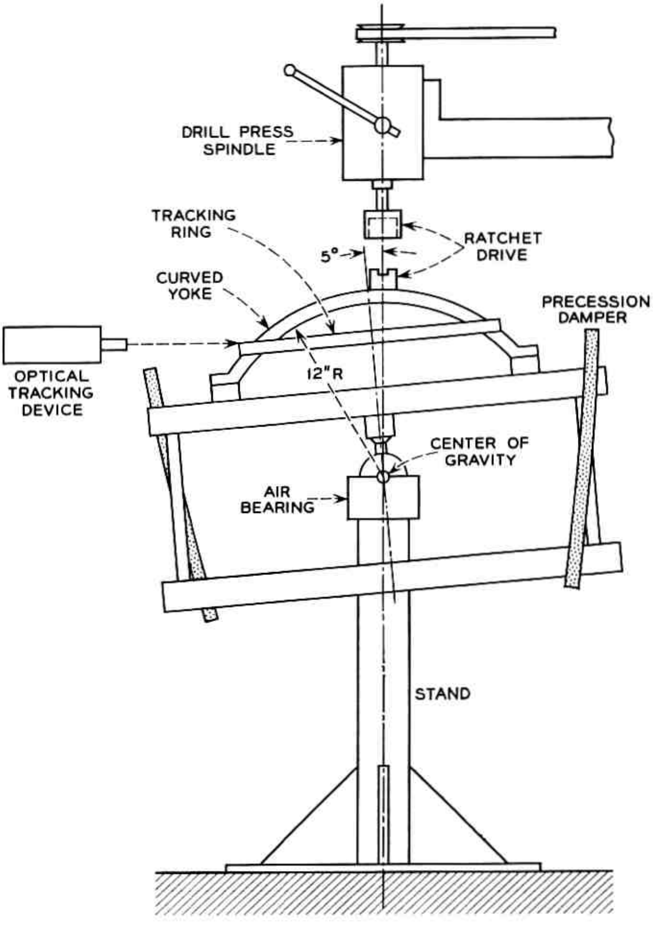


Fig. 5 — Schematic layout of precession damping experiment.

sional motion with known initial angular speed and initial precession angle. The decay of the precession angle is recorded through an optical tracking device for two different cases, with and without the balls in the damper tubes. The difference between these two recordings is the net effect due to the precession damper, excluding all other effects due to air resistance, gravity, etc. The balls were observed to be moving, and bottoming was clearly heard. The experimentally determined τ_p is found to be about four times larger than the theoretical τ_p calculated on the flywheel based on the mean value of the transverse moments of inertia and about nine times larger based on the minimum transverse moment of inertia. Although the Telstar satellite has different moments of inertia from those of the flywheel and a higher $I_{\max}/I_z$ ratio, it is believed that the actual τ_p should not be greater than the theoretical τ_p in (33) by more than one order of magnitude.

For a conservative estimate of the precession damping time of the Telstar satellite, let us multiply (33) by a factor of nine and use the following physical constants: $I_z = 4.1412$, $I_{\max} = 3.8252$ slug-ft², $\lambda_{\min} = I_z/I_{\max} = 1.08$, $m = 0.0021$ slug, $n = 0.65$ sec^{-1}, $R = 15$ ft, $b = 1.046$ ft. Equation (33) is then reduced to

$$\tau_p = 19 \left(0.76 + \frac{4.35}{\omega_z^2} \right) \text{minutes}$$

which is relatively independent of the spin rate in the range of interest. At 178.33, 65, and 24 rpm, the maximum $1/e$ characteristic precession damping times are 14.8, 16.2, and 27.6 minutes, respectively.

IV. DRIFT OF SPIN AXIS

We have shown in the Section III that the major transverse torque causing spin precession is contributed by the residual magnetic dipole moment along the spin axis, **M**. (**M** is found to be pointing toward the rocket-mount end of the Telstar satellite.) The torque produced by the residual magnetic dipole moment normal to the spin axis is mostly averaged out because of the spinning motion; other transverse torques, produced by eddy currents and gravity, are all one order of magnitude smaller than that produced by **M**, as shown previously. Therefore, in the qualitative analysis of the spin-axis drift in this section it is sufficient to take only **M** into account.

The initial direction of the Telstar satellite spin axis on the day of launch was 82.3° right ascension and −65.6° declination, as represented by the initial angular momentum, $\mathbf{J}_0$, (see Fig. 1) in the nonrotating coordinate system O-XYZ, where OX points toward the vernal equinox and OZ is along the earth's spin axis. The geomagnetic field, as shown schematically in Fig. 1, is rotating about OZ at the earth's spin rate. When the satellite is traveling along its orbit, it finds that the geomagnetic induction, **B**, generates a nearly conical surface with respect to $\mathbf{J}_0$ or to O-XYZ, with the axis of the cone pointing in the direction of the orbital angular momentum $\mathbf{J}_{\text{orb}}$ (see Fig. 6). Let us pass a plane through O normal to $\mathbf{J}_0$, project the conical surface (or **B**) onto the plane, and construct the time-average resultant of the projection, $\mathbf{B}_\perp$. Then $\mathbf{B}_\perp$ will interact with the magnetic dipole moment **M** (pointing in the negative direction of **J** or $\boldsymbol{\omega}$) to produce a transverse torque, $\mathbf{T}_\perp$, which causes spin precession. Because of the rotation of the geomagnetic field, which has anomalies, and because of the apsidal advance in the orbit, the conical surface generated by **B** has a different area every orbit.

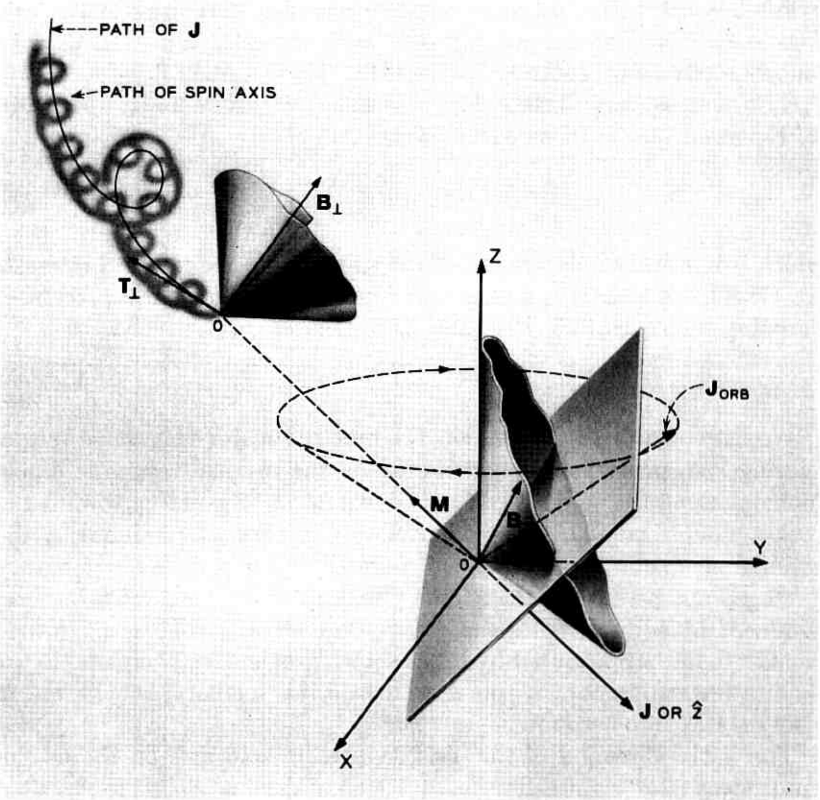


Fig. 6 — Schematic illustration of spin-axis drift.

Furthermore, due to the precession or the nodal regression of the orbital plane, the cone gradually changes its direction in the O-XYZ coordinates, indicated in Fig. 6 as the rotation of $\mathbf{J}_{\text{orb}}$ about OZ. As a result, $\mathbf{B}_\perp$ and hence $\mathbf{T}_\perp$ gradually change in both direction and magnitude, while the angular momentum, **J**, of the satellite follows the pattern of variation of $\mathbf{T}_\perp$, as shown in Fig. 6. Because of spin precession, the spin axis is turning around **J**, yet does not deviate away from **J** as a consequence of precession damping. Therefore, the tip of the spin axis describes a spiral path, shown in an exaggerated way in Fig. 6. Such a phenomenon is termed the drift of the spin axis. The pattern* of the drift is determined by the orbit and by the orientation of **M**, whereas the rate of drift depends on the orbit and the magnitude of **M**.

A rough estimate of the rate of drift of the Telstar satellite spin axis can be given. Let us assume that the transverse torque, $\mathbf{T}_\perp$, keeps acting on the satellite perpendicular to the angular momentum, **J**, despite the fact that **J** continuously changes its direction as time goes on. This assumption is justified by the fact that the precession dampers work properly, so that the spin axis is virtually in line with **J**. Let us disregard the retarding torque at this point. Then, from the principle of angular momentum about the center of mass of the satellite

$$\frac{d\mathbf{J}}{dt} = \mathbf{T}_\perp \tag{39}$$

we find that after a time interval Δt the angular momentum changes to a new position by an angle

$$\Delta\theta = \frac{T_\perp \cdot \Delta t}{J}. \tag{40}$$

In the above we have kept $\mathbf{J} = I_z\omega$, or ω at a constant magnitude, because we have not considered the retarding torque. To evaluate $\Delta\theta$ in a time interval of one week, let us substitute into the above with $\Delta t = 6.048 \times 10^5$ sec, $I_z = 5.61$ kg-m^2, $T_\perp = MH_\perp$, where $M = 0.562 \times 10^{-6}$ weber-meter and $H_\perp = H \cos \gamma$ (H = time-average value of the geomagnetic field on the Telstar satellite orbit and $H_\perp$ is the component of H along the spin axis; γ is the angle between H and the spin axis). Then (40) is reduced numerically to

$$\Delta\theta = 274 \frac{H \cos \gamma}{\omega} \text{ degree}$$

where H is in oersteds and ω in rad/sec. If H is taken to be 0.2 oersted

* For details of the pattern and rate of drift, see Ref. 5.

and γ to be 45°, then at the initial spin rate of 178.33 rpm or $\omega = 18.67$ rad/sec, we find $\Delta\theta = 2.1°$ per week, which is very close to the observed value (approximately 2° per week). This drift rate should increase exponentially with time because of the exponential decay of ω resulting from the action of the retarding torque. It appears from the above estimate that the satellite's residual magnetic dipole moment along the spin axis did not change drastically due to launching.

V. SUMMARY AND CONCLUSIONS

The dynamics problems for the spin-stabilized Telstar satellite, characterized by spin decay, spin-precession damping, and spin-axis drift, have been studied in this paper. In the section on spin decay, the nature of the retarding torque due to eddy-current losses has been analyzed. The observed decay phenomena are largely explained from the computed $\overline{B_\perp^2}$, taking into account the anomalies of the geomagnetic field, the variations of orbital parameters, and the change of the spin-axis direction. The $1/e$ characteristic time of the nearly exponential spin decay is estimated to be about 330 days ±2 per cent by extrapolation from the observed data. This indicates that at the end of two years from the day of launch the Telstar satellite will spin at approximately 20 rpm. It is believed that motion at such a spin rate is still relatively stable with respect to precession.

For the spin precession, it is found that the transverse torque is produced mainly by the residual magnetic dipole moment along the spin axis. The precessional motion of a spinning satellite is illustrated by means of Poinsot's geometrical constructions. A few different types of precession dampers have been considered. Linear analysis of the motion of the ball in the concave type damper has been made, from which explicit expression of the theoretical $1/e$ characteristic precession damping time is obtained. Based on the analysis, it was possible to make a proper design of the damper. An experimental comparison of the theoretical $1/e$ time enables us to estimate the actual $1/e$ time to be about 30 minutes maximum. It is concluded that this damping time is acceptable for the computed magnitude of the transverse torques, and in fact, no precession angle larger than 0.5° has yet been observed on the Telstar satellite.

In discussing the problem of spin-axis drift, only a brief qualitative description is given to illustrate the fundamental mechanism; also, an approximate quantitative analysis is shown for an order-of-magnitude estimate of the drift rate. The observed continuous drift of the spin axis of the Telstar satellite is evidence of proper functioning of the precession dampers.

The above three problems, which are caused essentially by electromagnetic torques, can be summarized into one of the important dynamics design criteria of a spin-stabilized satellite: i.e., evaluation of the maximum allowable eddy-current losses and residual magnetic dipole moment for specified useful life and orbit of the satellite. Other basic dynamical requirements are worth remarking here. The spin axis should necessarily have a maximum moment of inertia because of provision of precessional energy dissipation and because of elastic energy dissipation, since the satellite is not a perfectly rigid body. This moment of inertia should also be made as large as possible for a given weight and size of the satellite, in order to make the satellite more stable and to increase the lifetime for the same initial spin rate. Furthermore, the ratio of the moment of inertia about the spin axis to those about the transverse axes should be made large enough to yield an adequate precession damping time.

VI. ACKNOWLEDGMENTS

The author wishes to thank J. D. Gabbe for providing various computations of the geomagnetic field, M. S. Glass and D. P. Brady for furnishing data on the measurement of the drag torque and magnetic moment on the Telstar satellites, and G. T. Kossyk for making the precession damping experiments. The author is also indebted to L. Rongved for many useful suggestions in the analysis of precession damping and to J. W. West for various constructive comments and technical advice.

REFERENCES

1. Smythe, W. R., *Static and Dynamic Electricity*, McGraw-Hill, New York, 1950, 616 pages.
2. *Satellite Environment Handbook*, ed. Johnson, F. S., Lockheed Missiles and Space Division, Sunnyvale, California, December 1960, part VIII, Geomagnetism, by Dessler, A. J.
3. Courtney-Pratt, J. S., Hett, J. H., and McLaughlin, J. W., Optical Measurements on the *Telstar* satellite to Determine the Orientation of the Spin Axis and the Spin Rate, Jour. Soc. Motion Picture and Television Engineers, **72,** June, 1963, pp. 462–484.
4. Goldstein, H., *Classical Mechanics*, Addison-Wesley Publishing Co., Cambridge, Mass., 1950, 399 pages.
5. Thomas, L. C., The Long-Term Precession Motion of the *Telstar* Satellite, to be published.

A Passive Gravitational Attitude Control System for Satellites

By B. PAUL, J. W. WEST and E. Y. YU

(Manuscript received May 28, 1963)

It is shown how the gravity-gradient effect may be utilized to design a long-lived, earth-pointing satellite attitude control system which requires no fuel supplies, attitude sensors or active control equipment. This two-body system is provided with a magnetic hysteresis damper which effectively damps out oscillations (librations) about the local vertical. The long rods, which must be extended in space from coiled up metal tapes, provide the required large moments of inertia and possess adequate rigidity and sufficient strength to endure the rigors of the extension process. The system is compatible with the requirements of multiple satellite launchings from a single last-stage vehicle. Analysis indicates that the gravitational torques are sufficient to keep the disturbing effects of solar radiation pressure, residual magnetic dipole moments, orbit eccentricity, rod curvature, eddy currents, and meteorite impacts within tolerable limits. It is believed that the high-performance, earth-pointing system described and analyzed in this paper represents an essential step in the development of high-capacity communications satellites requiring long life.

CONTENTS

Page

I. INTRODUCTION 2196
II. DYNAMIC PRINCIPLES AND GENERAL DESCRIPTION OF PGAC 2199
2.1 Principles 2199
2.2 Description 2201
2.2.1 Extensible Rods 2202
2.2.2 Damper Unit 2203
2.2.3 Packaging and Multiple Launching 2205
2.2.4 Weight Breakdown 2205
III. ELASTIC VIBRATIONS VERSUS RIGID BODY MOTION 2206
3.1 Rigidity of Mast Rod and Influence of Distributed Mass on Natural Frequencies 2207
3.2 Natural Frequencies of Deck Assembly 2209
3.3 Torsional Oscillations of Mast 2212
IV. STRESS AND DEFLECTION ANALYSIS OF RODS DURING EXTENSION PHASE .. 2213
4.1 Tumbling 2213
4.2 Spinning 2216
4.3 Umbrella Effect 2217

V. THERMAL LOADING 2219
5.1 Temperature Distribution 2219
5.2 Thermal Bending 2222
VI. SPRING DESIGN FOR MULTIPLE LAUNCH 2223
VII. DAMPER UNIT 2226
7.1 Spring Constants 2227
7.2 Damping Torques 2228
7.3 Hardware Development Program 2229
7.3.1 Spring Design 2229
7.3.2 Damping Torque Test Program 2229
VIII. ANALYSIS OF DISTURBANCES AND ERRORS 2231
8.1 Solar Radiation Pressure 2231
8.2 Residual Magnetic Dipole Moment 2232
8.3 Orbital Eccentricity 2233
8.4 Meteorite Impact 2233
8.5 Cocked Angle due to Rod Deflections 2234
8.6 Miscellaneous Torques 2235
8.7 Summary 2236
IX. CONCLUSIONS 2237
X. ACKNOWLEDGMENTS 2237
REFERENCES 2237

I. INTRODUCTION

An earth-pointing attitude control system offers many advantages for a commercial satellite repeater. By directing the satellite's radiated power to just cover the earth, the satellite's size and weight can be minimized. For example, at a 6000-nm altitude the theoretical gain of an earth-covering conical beam is 14.5 db; however, allowance must be made for inaccuracies of the earth-pointing system and the gain that can be achieved from a practical antenna. Conservative estimates have shown that the achievable antenna gain is at least 10 db higher than with a Telstar-type isotropic antenna.* Hence, with an earth-pointing antenna, the power required from the satellite transmitter is only one-tenth that required with an isotropic antenna. This reduction in power makes the size and weight of communications satellites of high capacity (e.g., two TV channels or 600 two-way voice channels continuously operating) compatible with existing launch vehicles for orbits of interest.

In this paper we will describe a passive gravitational attitude control system (hereafter called PGAC) which provides a particularly attractive way to maintain a satellite axis pointing towards the earth. This system should have an extremely long life since it is entirely passive and requires no power and no active controls or attitude sensors. The system has been designed to be compatible with the launching of several satellites from a single launch vehicle. The importance of this feature becomes

* While the first Telstar satellite satisfied the objectives for a communications experiment, the performance was about 6 db below Bell System objectives.[1] For a higher-altitude commercial satellite, the additional 10 db would be considered essential to assist in meeting systems margins.

apparent when one considers the cost and time required to place perhaps 20 or more satellites into orbit with existing launch vehicles for a medium-altitude satellite system. Unpublished studies at Bell Telephone Laboratories have indicated that three satellites of the capacity mentioned and employing PGAC can be launched by a single Atlas-Agena vehicle in orbits of communications interest.

The most critical part of any passive earth-pointing system is the technique of damping employed to stop tumbling and limit librational motions. A unique feature of the PGAC system described herein is the employment of magnetic hysteresis damping in conjunction with a two-body system that provides large relative motion for damping purposes. Magnetic hysteresis damping is quite effective even at the slow librational rates (approximately a six-hour period at 6000-nm altitude).

PGAC employs long extensible rods to obtain appropriate moments of inertia about the three principal axes. In this respect it is similar to other passive systems[2,3] which also require long rods to obtain a sufficiently large moment of inertia in order for the gravity torque to be effective. Another feature of this PGAC system is that a single trigger or signal separates each satellite and simultaneously causes the rods to extend. This simplicity should enhance reliability of satellite separation and rod extension.

In any passive gravity-gradient orientation system, the satellite is stable with either end pointing towards the earth. In the PGAC system described here, dual antennas are proposed for each end of the satellite, and the appropriate antennas are to be activated by a simple microwave switch. Fig. 1 shows the two possible stable positions of the satellite. Antenna tests have shown that the extended rods do not substantially affect the antenna pattern; the maximum loss due to the rods is about 1 db. However, it may be possible to avoid this loss by properly orienting each satellite initially. This would require precise control of the launching vehicle orientation, satellite tumbling rate during ejection, and speed of extension of the rods.

In Section II of this paper, the dynamic principles of PGAC are described, and a general description of the system is given. Actually, two alternative configurations are described, each of which has its own advantages. Vibration analysis of the system is then given in Section III to demonstrate the validity of certain rigid body assumptions made in the dynamics analysis of the accompanying paper.[4] The stress and deformation of the rods due to dynamic loading during the extension phase and due to thermal effects are analyzed in Sections IV and V.

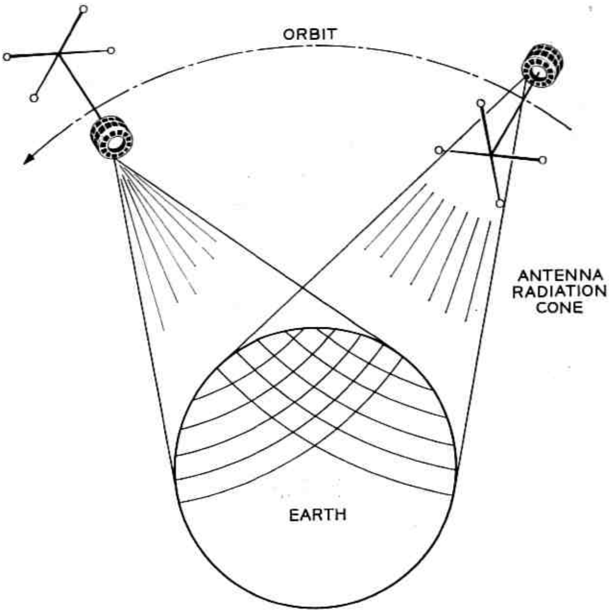


Fig. 1 — Possible satellite orientations.

Various spring designs for satellite separation associated with multiple launch are described in Section VI. The status of the hardware development and tests on the hysteresis damper unit are reviewed in Section VII. Finally, the various disturbing torques which the satellite will encounter in space are reviewed in Section VIII. It is shown that the PGAC system should remain earth-pointing within a few degrees.

Typical computer results have disclosed that for a reasonable initial tumbling rate of the satellite (1 rpm before rod extension, due to ejection from the rocket), the satellite will be earth-pointing within a few degrees of the local vertical in about 10 to 15 orbital periods. The description and discussion of PGAC in this paper is primarily for a satellite in a circular 6000-nm orbit with any inclination. However, with modifications of rod lengths and damping and spring constants, PGAC could be adapted to either higher or lower orbits.

The companion paper[4] in this issue covers the basic dynamics analysis of PGAC. The analysis includes large angle motion (as would be experienced by a satellite due to tumbling after ejection from the launch vehicle), as well as small librational motion. A complete three-dimensional analysis of the satellite motion has been formulated, and stability criteria for the system have been determined. Other dynamics analyses of passive attitude control systems have been reported in the literature.[5,6] These analyses either have not included large angle motions or have been restricted to the pitch motion only.

II. DYNAMIC PRINCIPLES AND GENERAL DESCRIPTION OF PGAC

2.1 *Principles*

The fact that an elongated body in orbit around the earth tends to line up with the local vertical is well known.[7] Just why this should be so is most easily explained by considering a rigid dumbbell with equal tip masses. Fig. 2(a) shows a dumbbell, in orbit around the earth, whose axis makes an angle $\theta (\theta < 90°)$ with the local vertical. Since the gravitational attraction varies inversely as the square of the distance from the geocenter, the lower mass A will experience a gravity force F_A which is slightly larger than the force F_B experienced by the upper mass B.

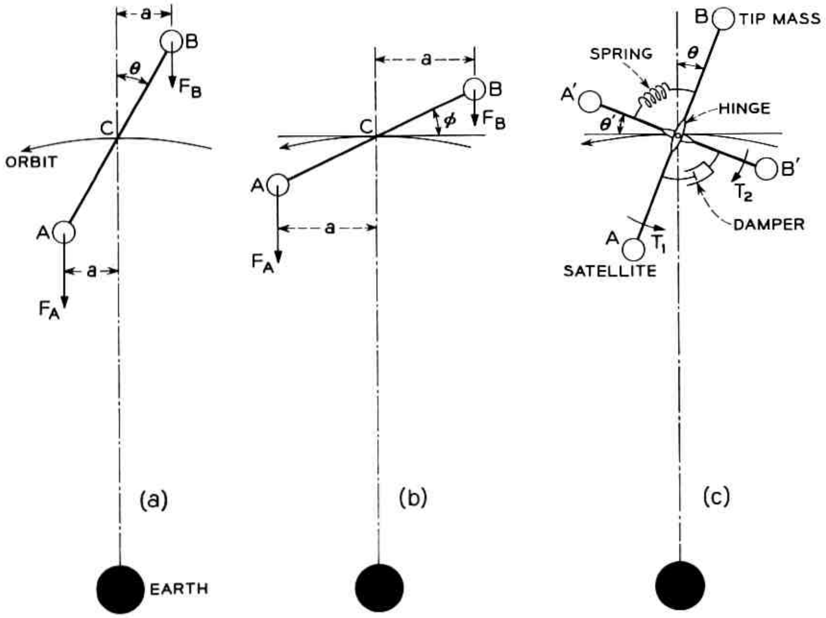


Fig. 2 — (a) Gravity forces acting on a dumbbell in orbit; (b) departure from the unstable equilibrium position; (c) system of primary and secondary dumbbells to produce damping.

The net torque about the mass center C produced by gravity forces is* $(F_A - F_B)a$ where the moment arm a is shown in Fig. 2(a). The gravity torque acts in such a direction as to diminish the angle θ. That is, it is a restoring torque which will rotate the dumbbell axis back to the local vertical. When the dumbbell becomes aligned with the local vertical, the moment arm a vanishes. Hence, the gravity torque becomes zero in the equilibrium position, $\theta = 0$. However, due to the inertia of the masses, the dumbbell does not stop in its equilibrium position but continues to rotate past it, whereupon the gravity torque reverses its direction and acts to restore the dumbbell to the local vertical. This process produces an oscillation or "libration" about the local vertical which would continue indefinitely if not damped out by some energy dissipating mechanism.

It is primarily the method of damping of the libration which distinguishes the various gravity-gradient schemes from each other. One method of damping the librations requires the use of a second dumbbell. In order to understand the function of this second body we should point out that a dumbbell is also in equilibrium (that is, no gravity torque acts upon it) when its axis is perpendicular to the local vertical. However, this equilibrium position is unstable in the sense that when the dumbbell deviates from the local horizontal by an arbitrarily small angle φ, the gravity torque $(F_A - F_B)a$ acts in such a manner (see Fig. 2b) as to increase the angle φ, i.e., to drive the system away from its (unstable) horizontal equilibrium position.

The inherent instability of a horizontal dumbbell may be used to design an efficient oscillation damper shown schematically in Fig. 2(c). In Fig. 2(c) the primary dumbbell AB is connected by means of a frictionless hinge to a secondary dumbbell A′B′. A spring is placed between the two dumbbells which keeps them crossed at right angles when the spring is not stressed. An energy dissipating device (represented in Fig. 2(c) by a piston in a close-fitting cylinder) is placed between the two dumbbells so that any relative motion of the two bodies results in a loss of mechanical energy (mechanical energy converted into heat energy). When the main dumbbell is deflected through an angle θ from the local vertical, it experiences a gravitational torque T_1 which tends to restore it to the local vertical. At the same time, because of the spring, there is a tendency for the secondary dumbbell to be carried along through an angle θ' in the same direction as the angle θ, thereby producing a gravitational torque T_2 on it which tends to increase

* Actually this is a slight oversimplification since F_A and F_B are not exactly parallel, but it adequately describes the main principle involved.

θ' still further. The net effect is that the gravity torques T_1 and T_2 tend to drive the two dumbbells in opposite direction, as shown in Fig. 2(c), thereby dissipating a relatively large amount of energy per cycle in the damping unit. The configuration shown in Fig. 2(c) will damp out oscillations in the plane of the orbit; in order to damp out oscillations perpendicular to the orbital plane, it is only necessary to add a second horizontal dumbbell which is perpendicular to the first one when its spring is unstrained. The two secondary dumbbells may be rigidly connected to each other and still provide damping in both planes of motion.

2.2 *Description*

The previous section describes the basic principles of the two-body system. In this section we discuss one method of reducing these principles to practice and describe the main features of all the major components of the system. These have reached a sufficiently high level of development for us to believe that they may be designed in detail for a specific experimental satellite.

Fig. 3 shows schematically what an actual configuration might look like. The long vertical "mast" connects the satellite to an upper "deck assembly" which serves as the tip mass for the primary "dumbbell" and as the unstable body. The deck assembly consists of two crossed dumbbells which meet at a "hinge unit" that is connected to the mast. This "hinge unit" is actually a universal joint (or Hooke's joint) which also provides elastic restoring forces (springs) and energy dissipation devices (dampers).

It should be mentioned that the deck assembly may be placed much lower on the mast rather than at its extremity as shown in Fig. 3, which illustrates a "high-deck" configuration. If the deck assembly is lowered to the vicinity of the satellite proper, the configuration will be referred to as a "low-deck" configuration (see Fig. 8 below for a schematic drawing of a low-deck configuration). From a dynamics point of view the two systems are identical. The high-deck configuration can be erected in a simple manner by the release of a single trigger which separates the satellite from the launching vehicle and simultaneously initiates extension of all rods. The low-deck arrangement has the advantage of being much stiffer structurally (see Section III) than the high-deck configuration and is much less sensitive in its response to accidental misalignment of the various rods due to initial curvatures, thermal bending, or micrometeorite impacts which might cause plastic deformation. It has the disadvantage of a more complex erection sequence (to provide

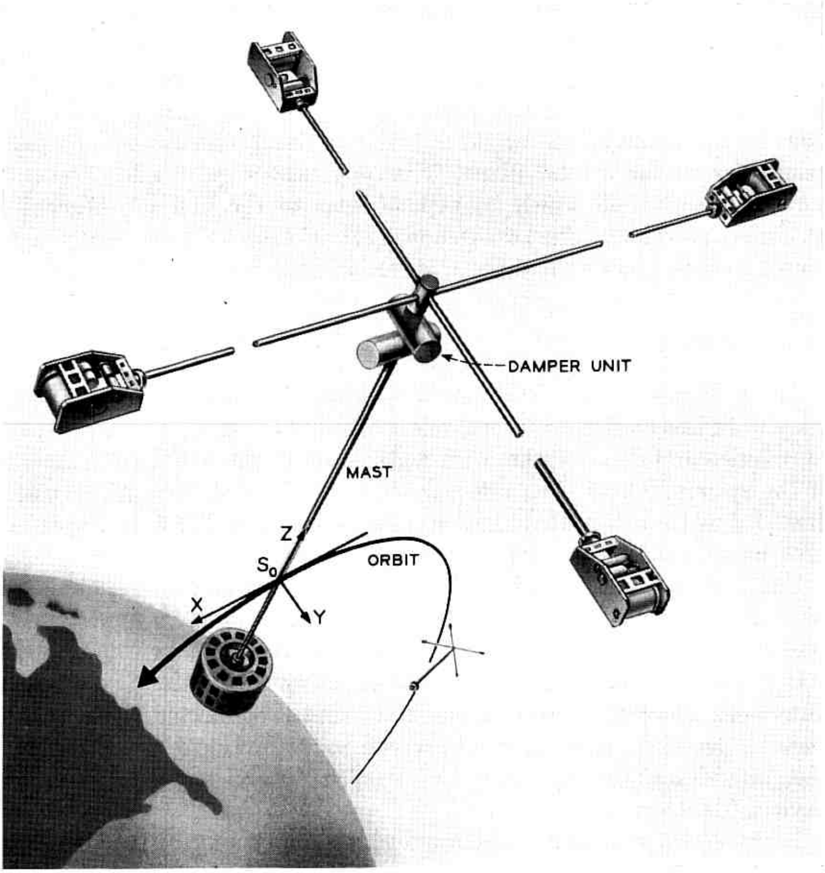


Fig. 3 — A passive gravitational attitude control (PGAC) system configuration.

clearance of the deck rods as they oscillate about the satellite body) and introduces the need to elevate a set of antennas above the height of the deck assembly in order to avoid electromagnetic difficulties.

2.2.1 *Extensible Rods*

A convenient way of erecting the rods in space is to use the STEM (Self-storing Tubular Extensible Member) units designed and developed by DeHavilland Aircraft of Canada, Ltd. These units consist of a beryllium copper tape (0.002 inch to 0.005 inch thick, and 2 inches to 5 inches wide) which is stored on a drum prior to extension, in the same manner as a carpenter's steel tape. However, unlike the carpenter's tape, the STEM tape has been preformed so that it tends to coil into a long straight tube when unwound from the storage drum, as shown in Fig. 4. The tape has a tendency to unwind spontaneously if not restrained from doing so. In fact, it is necessary to supply a governor which limits the extension speed to a safe level or else to provide a motor which drives the tape out at a controlled rate. Whichever is used, the motor or the governor mechanism, it could be located at the extremity of the deck rod to act as part of the necessary tip mass, as shown in Fig. 3.

2.2.2 *Damper Unit*

The damper unit will permit the deck a motion of two degrees of freedom with respect to the mast in the manner of a universal joint. A proposed damper unit is shown in Fig. 5, where the two rotationally symmetric housings are rigidly fixed to one another with their axes crossed at 90°. The deck assembly is free to rotate about the axis of the upper housing while the mast is free to rotate about the axis of the lower housing, thus providing the desired two degrees of freedom. To provide the required restoring torque, the deck assembly is fixed to a rotor whose axis is aligned with that of the housing by means of two fine torsion wires (or ribbons) as shown. These wires are maintained under suitable tension by means of the leaf springs at each end of the housing. This taut wire provides the rotational restoring torque required and also serves to keep the rotor axis aligned with the axis of the housing. A slot is provided so that the connecting rod to the deck assembly may rotate through a total angle of 120° before bottoming on the end of the slot. Although it is not anticipated that the rod will ever hit its stops except for rare periods of tumbling (following injection, or collision with a micrometeorite) one may

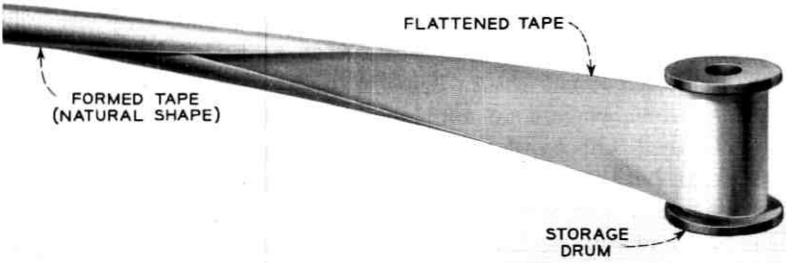


Fig. 4 — Extensible rod element.

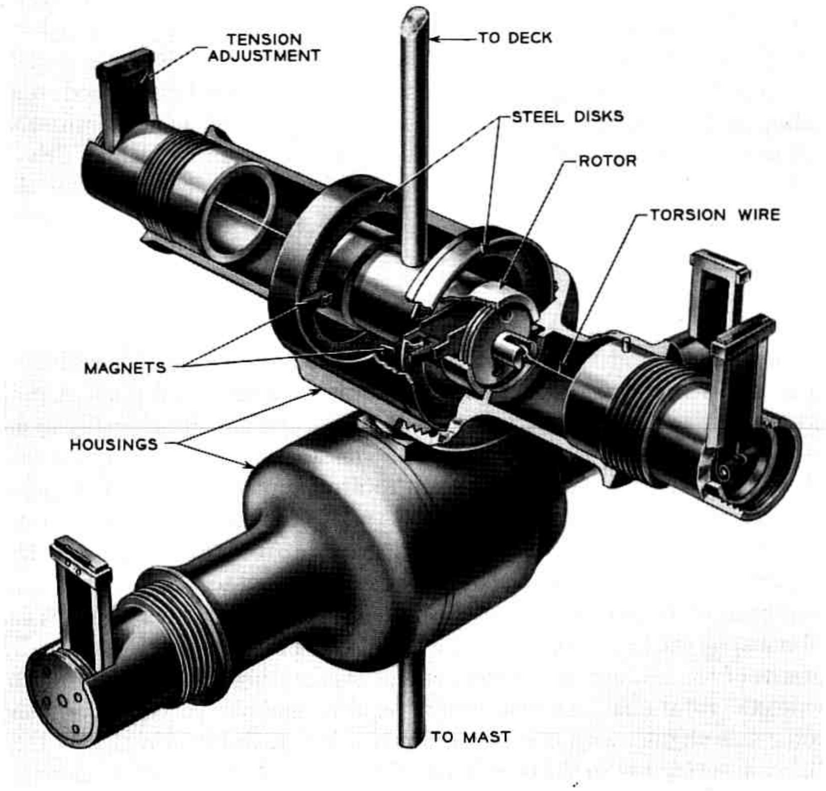


Fig. 5 — Damper unit.

design all stops so that they have no tendency to cold-weld in space. Similar stops are provided to prevent excessive lateral or axial motion of the rotor during periods of tumbling or during the launch phase.

Damping is provided by means of one or more bar magnets fixed along a diameter of the rotor. These magnets have horseshoe-shaped pole pieces which enclose an annular disk of permeable material (e.g. cold-rolled steel) whose outer rim is fixed to the housing. A small gap always exists between the faces of the pole pieces and the permeable disk by virtue of the accurate elastic suspension, and even when the rotor is bottomed during launch or tumbling, the stops maintain a predetermined clearance. As the rotor turns with respect to the housing because of satellite oscillations, the pole pieces rotate magnetic domains in the permeable disk, thereby creating magnetic hysteresis losses. Magnetic hysteresis losses are particularly desirable because they depend essentially upon the amplitude of oscillation rather than the frequency of oscillation and have been found to be effective at the very low libration rates, which are of the same order of magnitude as the orbital frequency (a six-hour period in the case of a 6000-nm altitude).

Damper units of the type described above have been developed which provide the estimated damping torques required for 6000-nm orbits. They appear to be sufficiently rugged to withstand the launching environment and weigh approximately two pounds for the complete damper unit. The feasibility of constructing dampers for higher and lower orbits has been demonstrated. Further details of the damper development program are described in Section VII.

2.2.3 *Packaging and Multiple Launching*

Design layouts indicate that there is no difficulty in packaging a stack of several satellites within the confines of a suitable rocket vehicle in such a manner that they will withstand rocket thrust and vibrations with a minimum amount of additional structure weight. The packaging arrangement is such that each individual satellite may be ejected with the required separation speed (see Section VI). The rod extension process may be triggered by the same explosive bolt mechanism which causes satellite ejection. One possible method of achieving this is indicated in Fig. 6 which shows how the rod constraints, which are needed during the launch phase, are automatically removed when the satellite is injected into orbit.

2.2.4 *Weight Breakdown*

For a 235-lb satellite body, operating at a height of 6000 nm, computer solutions based on the work of Ref. 4 indicate that a good design is achieved if the principal moments of inertia of body 1 (principal dumbbell) are $I_1 = 3333$, $I_2 = 3333$, $I_3 = 10$ lb-ft-sec^2, and the principal moments of inertia* of body 2 (secondary rod configuration) are $I_4 = 450$, $I_5 = 1000$, $I_6 = 1450$ lb-ft-sec^2. These moments of inertia may be achieved by using a 60-ft mast rod, four 40-ft deck rods, and the mass distribution shown in Table I.

* The given values of I_4, I_5 and I_6 differ slightly from those given in the example in Ref. 4; this difference is the result of computer studies made after Ref. 4 was submitted for publication.

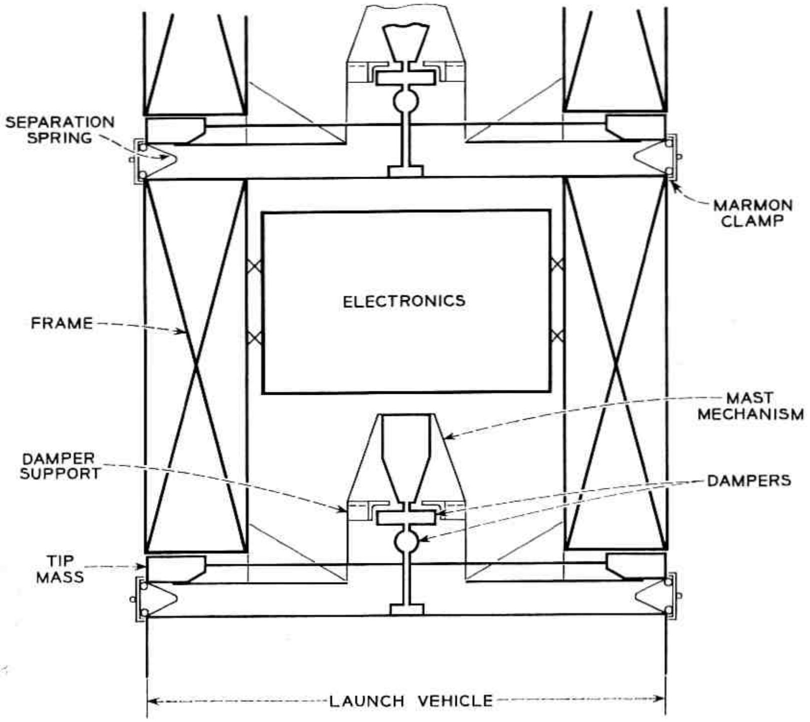


Fig. 6 — Packaging of satellites for multiple launching.

III. ELASTIC VIBRATIONS VERSUS RIGID BODY MOTION

The dynamics analyses[4] have been based on the assumption that the rods behave in an essentially rigid manner under the "gravity-free" conditions and extremely low librational angular speeds which prevail in the anticipated orbits; yet by ordinary earth-bound standards, the long thin rods would seem extremely flexible. To justify the assumption

Table I — Mass Distribution

Deck tip masses	29 lbs
Deck rods	$2\frac{1}{4}$ "
Damper assembly	2 "
Mast rod	$4\frac{1}{4}$ "
Mast motor	$5\frac{1}{2}$ "
Support structure	3 "
Total	46 "

of rigid rods in the basic dynamics studies, it is necessary to estimate the bending and twisting deflections which occur in service.

It is known that under normal circumstances the entire satellite will be oscillating, or librating, at certain well defined frequencies which are of the order of orbital frequency $\Omega/2\pi$. If any part of the system, which is supposed to behave in a rigid manner, happens to have a natural vibration frequency of the order of $\Omega/2\pi$, large deformations might occur, and the results of the rigid body dynamics analyses would be open to serious question. It is the purpose of this section to show that PGAC may be designed so that its smallest natural frequency is well above the orbital frequency $\Omega/2\pi$.

It will be assumed in this section that the satellite has been aligned along the local vertical and that the elastic members of the system are undergoing small vibration with respect to a nonrotating* set of coordinate axes. In order to rigorously calculate the lowest natural frequency of such a system, it would be necessary to consider a set of coupled partial differential equations of considerable complexity. However, it will be shown in Section 3.1 below that the mast rod may be considered to be perfectly rigid in the frequency range of interest. This result enables one to find the bending and twisting frequencies of the deck assembly in a relatively simple manner (involving ordinary rather than partial differential equations) as shown in Section 3.2. Finally, it is shown in Section 3.3 that the torsional mode of the mast rod has the lowest natural frequency of interest but that this frequency is still several times higher than the orbital frequency.

3.1 *Rigidity of Mast Rod and Influence of Distributed Mass on Natural Frequencies*

Let us consider the bending vibrations of a long beam, of length L, flexural rigidity EI, and mass ρ per unit length, whose ends carry two relatively large tip masses, M_1 and M_2, but which are otherwise unconstrained, as shown in Fig. 7. Since the tip masses are so great compared to the beam mass M_3, the tips can never move too far from their equilibrium position (in comparison with the midpoint of the beam). Therefore, the principal mode shape must look somewhat as shown in Fig. 7 with nodal points very near the ends. We can thus consider the problem equivalent to that of a beam, of length $L' \approx L$, whose ends are fixed against displacement and more or less fixed against rotation, depending upon the constraints provided by the tip masses. In any case,

* This is equivalent to neglecting the curvature of the orbital path.

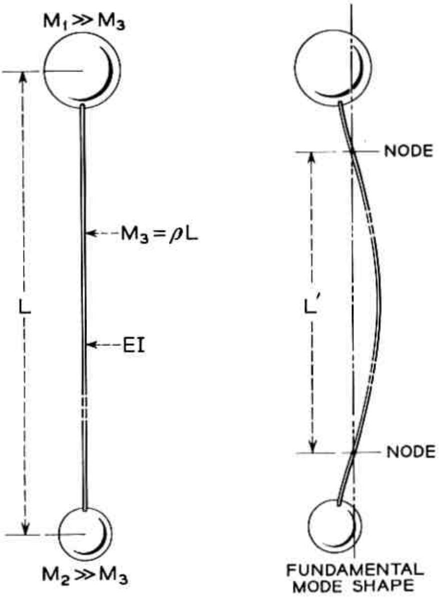


Fig. 7 — Mode shape for light beam carrying heavy tip masses.

the circular frequency p of the fundamental mode is given (Ref. 8, pp. 324–339) by

$$p = (C/L')^2\sqrt{EI/\rho} \qquad (1)$$

where C is a numerical coefficient which depends upon the degree of constraint at the nodes.

Now consider two limiting cases of constraint between the mast and deck assembly. If the springs which connect the deck and mast are extremely soft, no appreciable bending moment can be transmitted to the mast from the deck, so the connection point may be treated as a hinged or simply supported end. If the tip mass at the other end has negligible moment of inertia, that end may also be considered as simply supported and the coefficient[8] $C = 3.14$ (hinged-hinged beam); but if the tip mass has an appreciable moment of inertia, the end may be considered clamped, in which case $C = 3.93$ (hinged-clamped beam). Another limiting case arises if the connecting spring is extremely stiff, in which case the deck assembly with its very large moment of inertia is almost rigidly fixed to the mast rod and essentially prevents rotation of the connected end of the mast. In this case $C = 3.93$ (clamped-hinged beam) or $C = 4.73$ (clamped-clamped beam) accordingly as the tip mass has negligible moment of inertia or infinite moment of inertia.

We thus see that C can only undergo moderate variations despite extreme changes in the manner of end support. If one adopts the most conservative point of view and considers $C = 3.14$, (1) predicts that for a typical mast rod ($L \approx L' = 600$ in, $EI = 38{,}300$ lb in^2, $\rho = 1.51 \times 10^{-5}$ lb sec^2 in^{-2}) the fundamental mode of vibration has a frequency of $p = 1.37$ rad/sec, which is certainly well above the orbital frequency $\Omega = 0.273 \times 10^{-3}$ rad/sec for a 6000-nm orbit. Thus, we see that the influence of the distributed rod mass cannot play an important role in vibrational motions at the low frequencies of interest, and the mast rod may be considered rigid.

3.2 *Natural Frequencies of Deck Assembly*

Having shown that the mast rod is practically rigid, one may consider the satellite and mast rod a single rigid body upon which is mounted the deck assembly via the flexible joints of the damper unit. Because the deck tip masses are so great compared to the mass of the deck rods, one may neglect entirely the deck rod mass and treat the problem according to the standard "lumped mass" point of view. In particular, if one restricts attention for the time being to the low-deck configuration illustrated in Fig. 8, and notes that the center of mass* of the entire system lies fairly close to the plane of the deck assembly, one may introduce a further simplification by considering the deck assembly to be oscillating about a fixed point where the two hinge axes are assumed to cross.

In setting up the equations of motion, we shall use D'Alembert's principle, wherein each moving mass M_i is thought of as loading the structure by a system of "inertia forces": $X_i = -M_i\ddot{u}_i$, $Y_i = -M_i\ddot{v}_i$, $Z_i = -M_i\ddot{\omega}_i$, parallel respectively to the axes of x, y and z; and the rods are loaded by "inertia torques", $T_i = -I_{iM}\ddot{\theta}_i$. Superscript dots denote differentiations with respect to time t, in the usual Newtonian notation. To provide a more flexible and symmetrical notation, we shall frequently speak of the generalized displacements, q_i, and generalized forces, Q_i, which are related to previously defined quantities as in Table II. In this tabulation, generalized masses m_i have been defined for future reference.

Each independent deflection q_i can be found as a function of the

* The center of mass has been assumed to be unaccelerated in inertial space and for our purposes may be considered fixed.

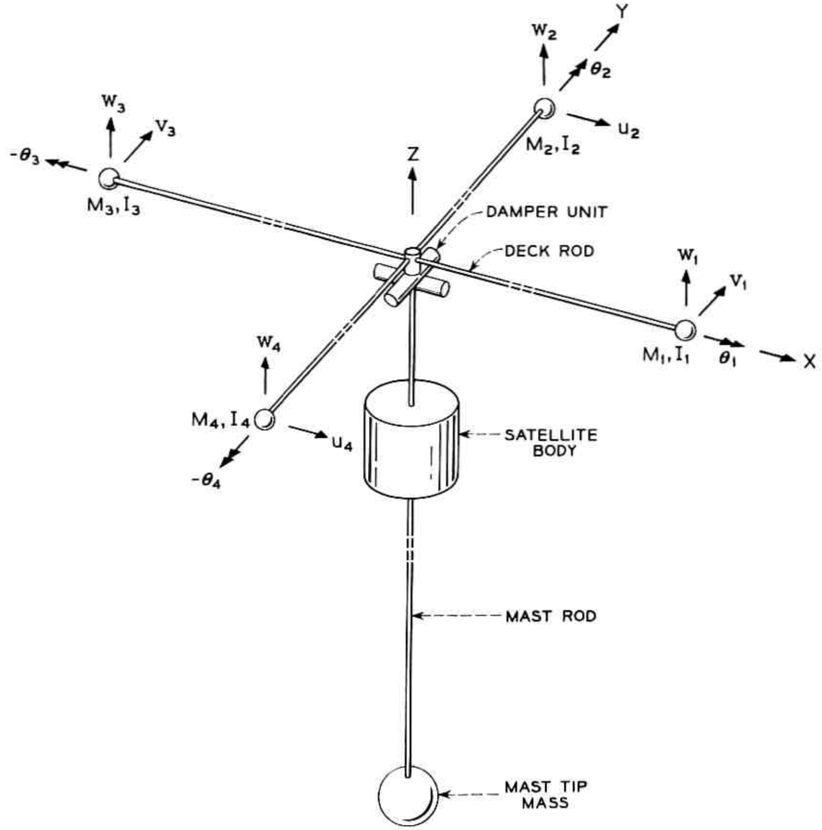


Fig. 8 — Low-deck configuration.

inertia forces applied to the system

$$q_i = \sum \frac{\partial q_i}{\partial Q_j} Q_j. \tag{2}$$

The terms $\partial q_i/\partial Q_j$ are called influence coefficients and are functions of the elastic constants and dimensions of the system. Each influence

Table II — Generalized Displacements, Forces and Masses

i =	1	2	3	4	5	6	7	8	9	10	11	12
q_i =	v_1	w_1	θ_i	u_2	w_2	θ_2	v_3	w_3	θ_3	u_4	w_4	θ_4
Q_i =	Y_1	Z_1	T_1	X_2	Z_2	T_2	Y_3	Z_3	T_3	X_4	Z_4	T_4
m_i =	M_1	M_1	I_{1M}	M_2	M_2	I_{2M}	M_3	M_3	I_{3M}	M_4	M_4	I_{4M}

TABLE III — LOW-DECK CONFIGURATION

Frequency p_n (rad/sec)	Most Significant Vibration
$p_1 = 0.597 \times 10^{-3}$	Oscillation of deck assembly about x (roll) axis
$p_2 = 0.643 \times 10^{-3}$	Oscillation of deck assembly about y (pitch) axis
$p_3 = p_4 = p_5 = 0.401 \times 10^{-1}$	Bending of deck rod parallel to x-axis
$p_6 = p_7 = p_8 = 0.590 \times 10^{-1}$	Bending of deck rod parallel to y-axis
$p_9 = p_{10} = 0.1373$	Twisting of deck rod parallel to x-axis
$p_{11} = p_{12} = 0.2625$	Twisting of deck rod parallel to y-axis

coefficient may be found by an elastic analysis of a statically determinate structure; for the sake of brevity, these coefficients will not be explicitly written out here.

Equation (1) may be written in the form

$$\sum_{j=1}^{12} \left[\frac{\partial q_i}{\partial Q_j} Q_j - \delta_{ij} q_i \right] = 0 \tag{3}$$

where $\delta_{ij} = 1$ for $i = j$, and $\delta_{ij} = 0$ for $i \neq j$. The symbol Q_j represents the generalized inertia force, $-m_j q_j$, and m_j represents the generalized mass (or moment of inertia) defined in Table II. The equations of motion are thus given by

$$\sum_{j=1}^{12} \left[\left(\frac{\partial q_i}{\partial Q_j} \right) (-m_j \ddot{q}_j) - \delta_{ij} q_i \right] = 0. \tag{4}$$

In order to solve the system of differential equations, we may assume that

$$q_i = A_i \cos (pt - \varphi) \tag{5}$$

where A_i, p and φ are as yet unknown quantities. If one substitutes (5) into (4) and follows the standard procedure,[8] one finds a twelfth-degree equation in $1/p^2$ with twelve solutions, $1/p_n^2$ for $n = 1, 2, \cdots, 12$. For these twelve (not necessarily distinct) values of p, (4) is satisfied and the assumption of (5) is justified. Equation (5) shows that the terms p_n represent the circular frequencies of the so-called natural modes of vibration. Upon the introduction of suitable numerical values, one finds the twelve natural frequencies $p_1 \cdots p_{12}$, listed in Table III, for a typical configuration designed to orbit at 6000 nm.

It may be seen from Table III that the lowest natural frequencies are those corresponding to the oscillations about the hinge-spring axes.*

* The frequency of the "rigid body" oscillations of the deck assembly about the pitch and roll hinge axes have been made intentionally close to the libration frequency in order to provide good damping. All other natural frequencies must be kept well above these values to avoid undesired resonances.

These frequencies differ from those which would be obtained with perfectly stiff deck rods by one part in 5000. The bending of deck rods has the next highest natural frequencies, which are about 70 times the pitch spring frequency. This indicates that excitation at a libration frequency ($\approx 0.5 \times 10^{-3}$ rad/sec) would not cause very large unwanted deflections any place in the structure, and that the assumption of rigid rods in the dynamics analysis is well justified for the low-deck configuration.

Although a complete vibration analysis of the high-deck configuration (shown schematically in Fig. 3) has not been made, there is no reason to believe that the natural frequencies of vibrational modes dominated by bending action will differ by orders of magnitude from similar modes in the low-deck configuration.

3.3 *Torsional Oscillations of Mast*

On the other hand, it is to be expected that the frequency of torsional vibration about the mast axis will be considerably less for the high-deck configuration. As a first approximation, one may neglect the bending deformations of the deck rods and consider the system shown in Fig. 3 as a long rod of torsional constant K_m (K_m = torque per unit twist angle) separating two rigid bodies whose moments of inertia are I_b and I_d, respectively. The angular frequency of natural oscillation is given (Ref. 8, p. 12) for such a system by

$$p = \left[K_m \frac{(I_b + I_d)}{I_b I_d} \right]^{\frac{1}{2}} \approx \left[\frac{K_m}{I_b} \right]^{\frac{1}{2}} \tag{6}$$

where the approximation follows from the fact that $I_d \gg I_b$. For a typical case of interest, one would find a torsional oscillation frequency, for the high-deck configuration, of the order of

$$p = 0.0048 \text{ rad/sec.} \tag{7}$$

This value should be compared with a libration frequency (in a so-called higher roll-yaw mode) of $p_{\text{ry}} \approx 0.00049$ rad/sec. If a somewhat more refined analysis is made, which takes into account the elasticity of the deck rods, the improved value of p differs insignificantly from the value given by (7). Although this value is smaller by an order of magnitude than the corresponding frequency of the low-deck configuration, it is still about ten times greater than the largest libration frequency. Some other comparisons between high- and low-deck arrangements have already been discussed in the introduction to Section II.

IV. STRESS AND DEFLECTION ANALYSIS OF RODS DURING EXTENSION PHASE

Since a satellite cannot be injected into orbit with absolutely zero angular velocity, in inertial space the tip masses on the extending rods will tend to cause bending and twisting of the rods during the process of extension. No attempt will be made to examine this problem in full generality, but two important representative cases will be considered. In both cases only the high-deck configuration is considered, since the low-deck configuration would seem to be at least as strong as the high-deck configuration.

Experiments[9,10] have demonstrated that a properly designed spring arrangement is capable of injecting satellites into orbit with tumbling rates below 1 rpm prior to rod extension. It is shown in this section that such rates do not cause excessive stresses or deformation in the rods during the extension process.

4.1 *Tumbling*

The satellite is idealized as shown in Fig. 9 and is assumed to be tumbling at time $t = 0$ with angular speed ω_0 about the body axis x,

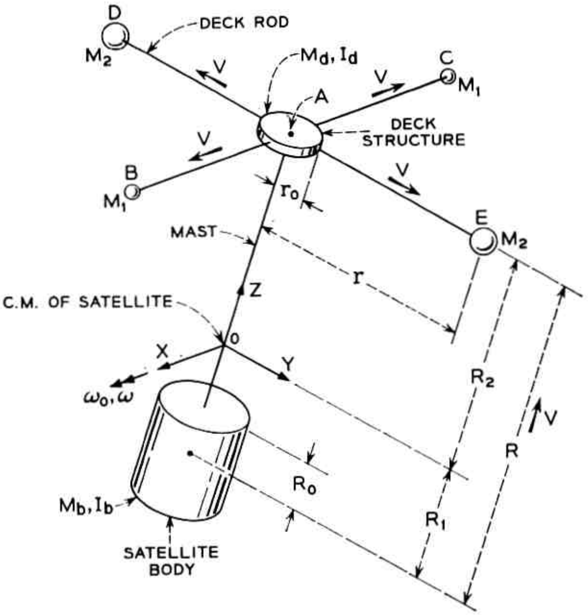


Fig. 9 — Schematic diagram of satellite during extension.

which is parallel to the deck rods carrying tip mass M_1. At time $t = 0$ both mast and deck rods begin to extend with speed v, as indicated in the sketch where r_0 represents the initial distance between the deck tip-masses and the mast axis, and R_0 equals the initial distance, measured along the mast axis, between the satellite body and the deck assembly. All mass in the deck assembly exclusive of the rods and tip masses is considered to be concentrated at the tip of the mast (point A) in the rigid body labelled "deck structure" in Fig. 9. The mass of the "deck structure" is denoted by M_d, and its moment of inertia about a centroidal axis parallel to x is denoted by I_d; similar expressions for the satellite body are denoted by M_b and I_b, respectively.

From Fig. 9 it is seen that the instantaneous distances r, R_1, and R_2 are given by

$$\begin{aligned} r &= r_0 + vt \\ R_1 &= \frac{(2M_1 + 2M_2 + M_d)(R_0 + vt)}{2(M_1 + M_2) + M_d + M_b} \\ R_2 &= \frac{M_b(R_0 + vt)}{2(M_1 + M_2) + M_d + M_b} \end{aligned} \tag{8}$$

and the instantaneous moment of inertia I of the entire system about the x-axis (passing through the instantaneous center of mass) can be shown to be

$$I(t) = I_b + I_d + \bar{M}(R_0 + vt)^2 + 2M_2(r_0 + vt)^2 \tag{9}$$

where $\bar{M}$ is defined by

$$\bar{M} = \frac{M_b(2M_1 + 2M_2 + M_d)}{2M_1 + 2M_2 + M_d + M_b}. \tag{10}$$

The initial value of I is denoted by I_0 and is found from (9) by setting $t = 0$.

We shall now assume that: (*i*) the mass of the rods is negligible; (*ii*) the hinge connection between the mast and deck assembly is rigid; (*iii*) the rods do not bend or twist (until further notice); and (*iv*) the mass center of the system is moving through inertial space with constant velocity. Under these assumptions one may apply the principle of conservation of angular momentum* to find the angular velocity ω and acceleration $\dot{\omega}$ in the form

* Although angular momentum is not strictly conserved in the presence of gravity torque, it can be shown that this effect is not significant.

$$\omega = \frac{\omega_0 I_0}{I}; \quad \dot{\omega} = \frac{-\omega_0 I_0}{I^2}\frac{dI}{dt}. \tag{11}$$

Equations (11) together with (9) fully specify the angular velocity and acceleration during the entire extension phase. The absolute acceleration $\mathbf{a}^P$ of any point P in the system may be found from the vector equation

$$\mathbf{a}^P = \frac{\delta^2 \mathbf{p}}{\delta t^2} + \frac{\delta \boldsymbol{\omega}}{\delta t} \times \mathbf{p} + 2\boldsymbol{\omega} \times \frac{\delta \mathbf{p}}{\delta t} + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \mathbf{p}) \tag{12}$$

where $\mathbf{p}$ is the position vector of the point P measured from the origin shown in Fig. 9, or in terms of the unit vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ along the body axes:

$$\mathbf{p} = p_x\mathbf{i} + p_y\mathbf{j} + p_z\mathbf{k}. \tag{13}$$

By definition:

$$\begin{aligned} \delta\mathbf{p}/\delta t &= \dot{p}_x\mathbf{i} + \dot{p}_y\mathbf{j} + \dot{p}_z\mathbf{k} \\ \delta^2\mathbf{p}/\delta t^2 &= \ddot{p}_x\mathbf{i} + \ddot{p}_y\mathbf{j} + \ddot{p}_z\mathbf{k} \\ \boldsymbol{\omega} = \omega\mathbf{i}; &\qquad \delta\boldsymbol{\omega}/\delta t = \dot{\omega}\mathbf{i}. \end{aligned} \tag{14}$$

Applying (12) successively for the five points A, B, C, D, and E shown in Fig. 9 one may find the components of acceleration $\mathbf{a}^P = a_x{}^P\mathbf{i} + a_y{}^P\mathbf{j} + a_z{}^P\mathbf{k}$ indicated in Table IV.

Table IV, together with (11), gives the absolute acceleration of all the tip masses. The D'Alembert forces acting on masses at points A, B, C, D, and E are respectively $-M_d\mathbf{a}^A$, $-M_1\mathbf{a}^B$, $-M_1\mathbf{a}^C$, $-M_2\mathbf{a}^D$, $-M_2\mathbf{a}^E$. The bending moment M_{mx} at any point z along the mast is given at any time by

$$M_{mx} = -[M_d a_y{}^A + M_1(a_y{}^B + a_y{}^C) + M_2(a_y{}^D + a_y{}^E)](R_2 - z) \\ + M_2 r(a_z{}^E - a_z{}^D) - I_d\ddot{\omega}. \tag{15}$$

Table IV — Components of Acceleration

P (Point)	p_x	p_y	p_z	$a_x{}^P$	$a_y{}^P$	$a_z{}^P$
A	0	0	R_2	0	$-(2\dot{R}_2\omega + R_2\dot{\omega})$	$-R_2\omega^2$
B	r	0	R_2	0	$-(2\dot{R}_2\omega + R_2\dot{\omega})$	$-R_2\omega^2$
C	$-r$	0	R_2	0	$-(2\dot{R}_2\omega + R_2\dot{\omega})$	$-R_2\omega^2$
D	0	$-r$	R_2	0	$r\omega^2 - 2\dot{R}_2\omega - R_2\dot{\omega}$	$-(R_2\omega^2 + 2\dot{r}\omega + r\dot{\omega})$
E	0	r	R_2	0	$-(r\omega^2 + 2\dot{R}_2\omega + R_2\dot{\omega})$	$-R_2\omega^2 + 2\dot{r}\omega + r\dot{\omega}$

Table V — Parameters

M_b = 286/32 lb sec²/ft	I_b = 10 or 20 lb sec² ft
M_d = 8/32 lb sec²/ft	I_d = 0.0845 lb sec² ft
M_1 = 4/32 lb sec²/ft	R_0 = 2 ft
M_2 = 9/32 lb sec² ft	r_0 = 0.75 ft
EI = 13.4 lb ft² (deck rod)	EI = 297 lb ft² (mast rod)

Similar expressions are readily written down for the bending moments at various points along the deck rods but will be omitted here for the sake of brevity. Deflections are found by noting that a tip force F produces a lateral tip deflection $\Delta_F = FL^3/(3EI)$ for a cantilever of length L and bending stiffness EI. Similarly, a tip couple M produces a lateral tip deflection of amount $\Delta_M = ML^2/(2EI)$. The net deflection is found by superposition. For a 6000-nm satellite similar to the one described in Section II, the parameters shown in Table V were used. An investigation of the complete extension history shows that for an initial tumbling rate of 0.1 rad/sec ($\approx$1 rpm) and an extension rate of $v = \frac{1}{2}$ ft/sec, the maximum stresses occur early in the process and decay rapidly thereafter; i.e., maximum moments occur before the rods have extended a distance of 2 ft. The maximum bending moments (which occur at the cantilever root) and the corresponding tip deflections (expressed as a fraction of rod length at the instant of maximum loading) are given in Table VI. Published data[11] indicate that short lengths of the mast rod could sustain a bending moment about a hundred times greater than the maximum value indicated in Table VI, and the deck rods could sustain a value about 30 times larger than the greatest tabulated value. Thus, there appears to be no "stress" problem due to tumbling.

4.2 *Spinning*

Assumptions (*i*) to (*iv*) of the previous section will be retained.

If the entire satellite spins about the mast axis with angular speed

Table VI — Bending Moments and Deflections for $v = \frac{1}{2}$ ft/sec; $\omega_0 = 0.1$ rad/sec

Satellite Moment of Inertia	Maximum Bending Moment, Mast	Maximum Bending Moment, Deck	Mast Tip Deflection	Deck Rod Tip Deflection
(lb sec² ft)	(ft-lb)	(ft-lb)	(Fraction of rod length)	(Fraction of rod length)
10	0.072	0.018	0.0003	0.0012
20	0.113	0.029	0.0006	0.0025

ω_0 at time $t = 0$, the deck assembly will acquire an angular speed $\dot{\theta}_d$ and the satellite body will rotate at speed $\dot{\theta}_b$. Conservation of angular momentum* requires that

$$[I_d' + 2(M_1 + M_2)(r_0 + vt)^2]\dot{\theta}_d + I_b\dot{\theta}_b$$
$$= [I_d' + 2(M_1 + M_2)r_0^2 + I_b]\omega_0 \quad (16)$$

where I_d' denotes the moment of inertia of the "deck structure" about the mast axis. The torque required to produce a unit relative angular displacement between the satellite body and the deck is proportional to $(R - R_0)^{-1}$ and may be represented in the form $k/(vt)$, where k is the torsional rigidity of the mast for unit length. In other terms, the torque on the mast at any time is given by $(\theta_d - \theta_b)k/vt$; this torque is applied directly to the satellite body, so one may write

$$I_b\ddot{\theta}_b = k(\theta_d - \theta_b)/vt. \quad (17)$$

Equations (16) and (17) represent a third-order system of linear differential equations with time-dependent coefficients and with initial conditions of the form $\theta_d(0) = 0$; $\theta_b(0) = 0$; $\dot{\theta}_b(0) = \omega_0$. These equations have been integrated numerically to provide the complete response of the system during the extension phase for various sets of parameters. The solutions indicate that torsional stresses do not become excessive at any time, although the satellite body might rotate, relative to the deck assembly, by as much as 10 revolutions if an extension speed of $v = \frac{1}{2}$ ft/sec is used, and ω_0 is as high as 2 rpm. The bending stresses and deflections produced in the deck rods, under these conditions, are of the same order of magnitude (very safe) as found in the foregoing section on "tumbling." With a nonspinning final-stage vehicle it is unlikely that the initial spin rate ω_0 will reach a value as high as 1 rpm.[9,10]

4.3 *Umbrella Effect*

If all rods are being extended simultaneously, for the high-deck configuration the tip masses on the deck rods will continue to move parallel to the mast axis at the termination of the mast extension phase. This motion will continue until the cantilever bending of the deck rods has converted the tip mass kinetic energy into stored elastic energy. This effect will be referred to as the "umbrella" effect. To compute the maximum tip deflection δ_m and the maximum root bending moment M_m in the deck rods, it should be observed that a lateral tip

* Effect of gravity torque is neglected here, as in Section 4.1.

force P will produce a tip deflection $\delta_m = PL^3/3EI$, where L is the beam length and EI is the flexural rigidity. The stored strain energy U is equal to $(P/2)\delta_m$, which may be expressed as $U = 3EI\delta_m{}^2/2L^3$ by virtue of the linear relationship between P and δ. When the stored energy U is equated to the initial kinetic energy $(Mv^2/2)$ of a tip mass M which moves at speed v, one finds the tip deflection

$$\delta_m = v[ML^3/3EI)]^{\frac{1}{2}}. \tag{18}$$

The root bending moment M_m is found by multiplying the load $P = (3EI\delta_m/L^3)$ by the beam length L to give

$$M_m = v[3EIM/L]^{\frac{1}{2}}. \tag{19}$$

For a typical deck unit with the following parameters, $EI = 13.4$ lb ft^2, $L = 50$ ft, $M = (10/32.2)$ lb sec^2/ft, one finds that for sufficiently small values of v

$$\delta_m \text{ (ft)} = 0.983\, v \text{ (ft/sec)}$$

$$M_m \text{ (ft-lb)} = 0.500\, v \text{ (ft/sec)}.$$

Thus an extension speed of $v = \frac{1}{2}$ ft/sec would produce a bending moment of about $\frac{1}{4}$ ft-lb, which is less than the approximate allowable value of 1 ft-lb. The corresponding tip deflection of about $\frac{1}{2}$ ft is sufficiently small so that the linear bending theory used is adequate. If one wishes to find the maximum extension speed which produces a root bending moment below 1 ft-lb, it is necessary to consider a nonlinear beam theory which allows for large slopes in the deflected beam shape. An approximate treatment of this problem indicates that an extension speed of about 1.5 ft/sec would result in a root bending moment of about 1 ft-lb. Therefore, if one wishes not to exceed the load-carrying capacity of the deck rods, it is essential to keep the extension speed well below 1.5 ft/sec, or else to extend the deck rods after the mast has been fully extended.

Similar considerations show that reasonable differences in the extension speeds of the various deck rods result in tolerable loads on the mast, for practical configurations.

It should be noted that when the oscillating deck masses slam downward, they load the mast axially and tend to produce Euler-type buckling. It may readily be shown that, so long as the mast extension velocities are kept small enough to prevent overloading of the deck rods, the mast will not buckle for the configurations of interest.

V. THERMAL LOADING

The rods used in the proposed design consist of long split overlapping tubes which will experience temperature gradients due to solar heating. These temperature gradients will cause the rods to bend in such a way that the illuminated side becomes convex.

In this section, the lateral deflections of the rods due to solar heating will be calculated. In Section VIII it will be shown that these deflections have a minor influence upon the pointing accuracy of the PGAC System.

5.1 *Temperature Distribution*

It will be assumed that the rod is sufficiently long so that end effects may be ignored; hence the temperature distribution will not vary with length along the rod. Since the heat input depends upon the angle between the collimated solar rays and the axis of the rod, it is implicit in the above statement that the thermally induced curvature of the rod axis is small; this necessary requirement will be verified a posteriori in a numerical example. Confining attention to a unit length of rod as shown in Fig. 10, it may be verified that the cosine of the angle between the solar rays and the normal to a surface element located at an angle θ, measured from the outer edge of the tape, is given by

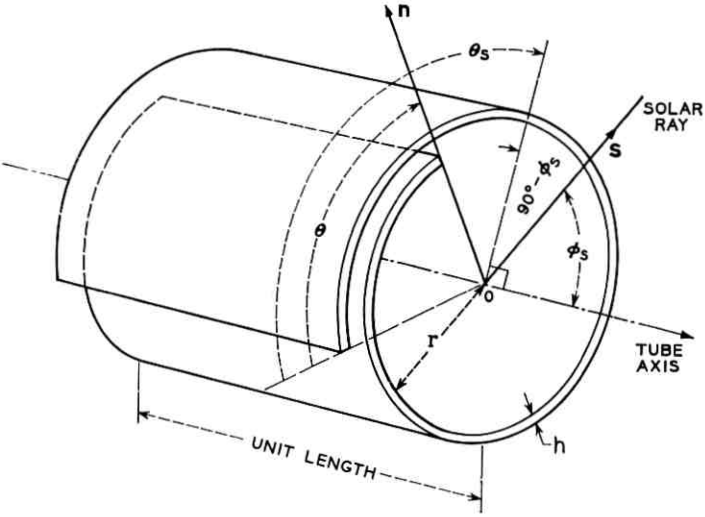


Fig. 10 — Unit length of split overlapping tube illuminated by the sun.

$$\mathbf{s}\cdot\mathbf{n} = \sin\varphi_s \cos(\theta_s - \theta) \tag{20}$$

where **s** is a unit vector pointing to the sun, **n** is a unit surface normal, φ_s is the angle between the solar rays and the tube axis, and θ_s is the angular distance from the outer free edge of the tube to the plane formed by **s** and the tube axis. The heat input per unit time on a unit area of tube surface is given by

$$q_s = a\, S \sin\varphi_s \cos^+(\theta - \theta_s) = a\, \bar{S} \cos^+(\theta - \theta_s) \tag{21}$$

where

a = absorptivity for solar radiation
S = solar constant (442 Btu/hr ft^2)
$\bar{S} = S \sin\varphi_s$ = effective solar constant
$\cos^+ x = \frac{1}{2}(\cos x + |\cos x|)$ (half-rectified cosine wave).

In general, a small element of the tube of arc length $rd\theta$ (where r is the tube radius) gains heat $q_s r d\theta$, in unit time, due to solar heating; the element also gains heat $q_c r d\theta$ by conduction and loses heat $q_e r d\theta$ by emission of radiation. It will be assumed that the overlapping layers do not have an appreciable area in mutual contact. This idealization is useful because of the random nature of the actual contact areas and the uncertainties in the contact pressure and in the associated surface heat transfer coefficients. Any heat conduction which does occur between overlapping layers will tend to reduce temperature gradients and alleviate the thermal bending effect; thus, the neglect of such effects leads to a conservative analysis. Since the walls are very thin, it is permissible to assume that the temperature varies only in the circumferential direction, so that Fourier's law leads to the result

$$q_c = (\kappa h)(d^2T/r^2 d\theta^2) \tag{22}$$

where κ = thermal conductivity, h = wall thickness, and T = absolute temperature. The heat loss by radiation is given by the Stefan-Boltzmann law: $q_e = \epsilon\sigma T^4$, where ϵ = hemispheric emissivity at temperature T, σ = Stefan-Boltzmann constant (1714×10^{-12} Btu/hr ft^2 (°R)^4). For simplicity, the effects of internal radiation will be neglected, so that the above expression for q_e is valid only for $0 \leqq \theta \leqq 2\pi$. The inclusion of internal radiation effects would result in reduced temperature gradients, thereby reducing the thermal bending; thus, this assumption is also conservative.

Upon summing up the three contributions to the thermal balance, one finds that

$$\frac{d^2T}{d\theta^2} - \left(\frac{\epsilon r^2}{\kappa h}\sigma\right) T^4 = -\left(\frac{r^2 a\bar{S}}{\kappa h}\right) \cos^+(\theta - \theta_s); \qquad 0 \leqq \theta \leqq 2\pi \quad (23)$$

$$\frac{d^2T}{d\theta^2} = 0; \quad \theta > 2\pi. \qquad (24)$$

Because no appreciable amount of heat can be radiated over the narrow faces (of area h per unit length) at the edges where $\theta = 0$ and $\theta = \theta_{\max}$, one may write the boundary conditions in the form $dT/d\theta = 0$ at $\theta = 0$ and $\theta = \theta_{\max}$, and observe that both T and $dT/d\theta$ must be continuous at $\theta = 2\pi$. Equation (24) implies a linear temperature distribution in the range $\theta > 2\pi$, but since $dT/d\theta$ vanishes at the edge $\theta = \theta_{\max}$, the temperature must be constant in the range $\theta > 2\pi$. In addition, continuity of $dT/d\theta$ requires that $dT/d\theta = 0$ at $\theta = 2\pi$. Thus, the temperature distribution may be found by solving the nonlinear differential equation (23), subject to the boundary condition $dT/d\theta = 0$ at $\theta = 0$, and at $\theta = 2\pi$. This problem, except for the boundary conditions, is similar to the problem treated by Charnes and Raynor[12] of a continuous (nonsplit) tube. Following their treatment, one may linearize (23) by writing

$$T = T_0 + \tau(\theta) \qquad (\tau \ll T_0) \qquad (25)$$

where T_0 is the mean radiant temperature defined by

$$T_0 = [a\bar{S}/\pi\epsilon\sigma]^{\frac{1}{4}}. \qquad (26)$$

Upon substitution of (25) into (23), one finds

$$\begin{gathered} d^2\tau/d\theta^2 - p^2\tau = -\beta \cos^+(\theta - \theta_s) + \lambda T_0^4 \\ d\tau/d\theta = 0 \text{ at } \theta = 0, \quad \text{and} \quad \text{at } \theta = 2\pi \end{gathered} \qquad (27)$$

where

$$p = 2\sqrt{\lambda T_0^3}; \qquad \lambda = (\sigma\epsilon r^2/\kappa h); \qquad \beta = a\bar{S}r^2/\kappa h \qquad (28)$$

Equation (27) may be solved by a number of standard procedures (e.g., use of Duhamel integral or of Laplace transform) which will be omitted for the sake of brevity. For a typical beryllium-copper rod, the pertinent dimensions are

$$r = 0.225 \text{ in}, \quad h = 0.002 \text{ in}, \quad \kappa = 65 \text{ Btu/hr ft}, \quad \theta_{\max} = 3\pi.$$

Representative values of absorptivity and emissivity, calculated by integration of monochromatic reflectivity measurements, are:

$$a = 0.8, \quad \epsilon = 0.3$$

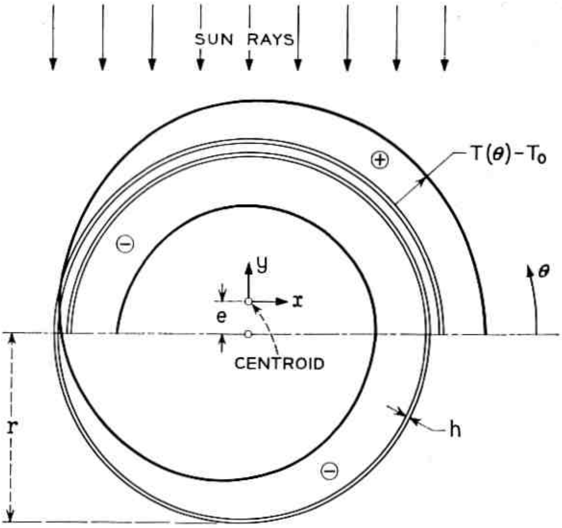


Fig. 11 — Temperature distribution in split overlapping tube (neglecting internal radiation and radial conduction effects).

where ϵ corresponds to a temperature of $T_0 = 684°\text{R}$. If the solar rays are truly normal to the rod axis, one may use $\bar{S} = S = 442 \text{ Btu/ft}^2 \text{ hr}$ and if the rod is oriented with $\theta_s = \pi/2$, as shown in Fig. 11, the temperature distribution will be as shown in the figure, where $\tau = T - T_0$ is plotted radially outward from the tube surface for positive values and inward for negative values of τ. For this example, the temperature drops continuously from $T = 700°\text{R}$ (240°F) at $\theta = 0$ to $T = 667°\text{R}$ (207°F) at $\theta = 3\pi$.

5.2 *Thermal Bending*

Following the method used by Timoshenko and Goodier[13] for bending of a beam of rectangular cross section, one may show that the curvatures developed in the z-x plane and the z-y plane are given, respectively, by

$$\kappa_x = M_y/EI_y\,, \qquad \kappa_y = M_x/EI_x \tag{29}$$

where E is Young's modulus,

$$I_x = \int_A y^2 dA, \qquad I_y = \int_A x^2 dA \tag{30}$$

$$M_x = E\alpha \int_A \tau y dA, \qquad M_y = E\alpha \int_A \tau x dA. \tag{31}$$

In the above expressions, α represents the coefficient of thermal expansion ($\alpha = 9.4 \times 10^{-6}\ {}^{\circ}\mathrm{R}^{-1}$ for numerical example); x and y are measured from the centroid of the cross section which is located a distance e from the axis of the tube, as shown in Fig. 11; x and y have their origin at the centroid; dA represents an element of area; and the integration is made over the entire cross section.

With the temperature distribution shown in Fig. 11 (corresponding to the numerical data given above), one finds that the curvature κ_x is negligible, but κ_y is found to be

$$\kappa_y = 3.18 \times 10^{-3}\ \mathrm{ft}^{-1}\ (R_y = 1/\kappa_y = 314\ \mathrm{ft}).$$

It may readily be shown that one end of a rod of length L bends through an angle $\Delta\Psi = \kappa L$ with respect to the other end and deflects through a lateral distance of $\delta = (\frac{1}{2})L^2/R$. The maximum angular and lateral deviations for a 50-ft length of rod are thus seen to be $\Delta\Psi = 9.1°$, $\delta = 4.1$ ft. A similar calculation shows that $\delta = 3.5$ ft for a rod of $r = 0.45$ in and $h = 0.005$ in.

In view of the conservative nature of the heat transfer analysis used, one may estimate that the actual values of slope and deflection could easily be less than half of the computed values. In any case, the deflections do not cause excessive misalignment from the local vertical (see Section VIII), and the slopes are sufficiently small to justify the initial assumption that the heat input and temperature distribution do not vary appreciably along the axis of the rod.

VI. SPRING DESIGN FOR MULTIPLE LAUNCH

When several satellites are launched from the same rocket vehicle, it is necessary that they be injected with different velocity components along the orbit trajectory; otherwise, all the satellites will have the same period and will appear to be "bunched" together when viewed from the ground. The velocity increments required to "minimize" the undesirable effects of bunching are discussed in Ref. 14, where it is indicated that a relative speed of about 12 ft/sec between the slowest and fastest satellites is desirable for the case of four satellites in a single orbit at 6000 nm. Similar conclusions were reached in unpublished work at Bell Telephone Laboratories for the case of three simultaneously launched satellites.

Velocity increments of 12 ft/sec are readily achieved by mechanical springs. In this section we shall consider the use of ordinary helical springs and of the so-called conical disk-spring (sometimes called a Belleville spring) shown schematically in Fig. 12.

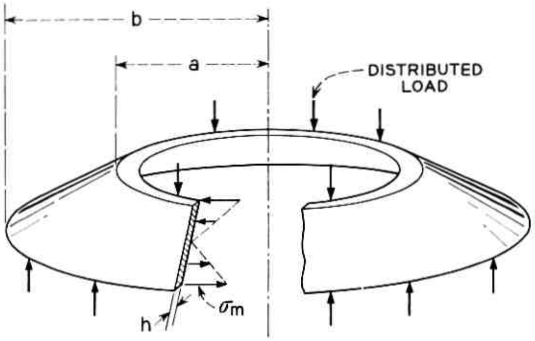


Fig. 12 — Belleville spring showing stress distribution.

It is easily shown[15] that the strain energy per unit volume u_h stored in a close-coiled helical spring with a narrow circular cross section can be expressed in the form

$$u_h = (\tfrac{1}{4})\tau_m^{\,2}/G \tag{32}$$

where τ_m is the maximum shear stress in the spring and G is the shear modulus.

In the case of a Belleville spring, the stresses are distributed[16] in an approximately linear manner over the cross section, as shown in Fig. 12, if the inequality $(b - a)/a \ll 1$ is satisfied and only small deflections are permitted. Under these conditions, it is easy to show that the strain energy per unit volume u_b is given by

$$u_b = (\tfrac{1}{6})\sigma_m^{\,2}/E \tag{33}$$

where σ_m is the maximum tensile stress in the spring and E is the modulus of elasticity. The relative energy-storing efficiencies of helical and Belleville springs may be found from (32) and (33) in the form

$$\frac{u_b}{u_h} = \left(\frac{2}{3}\right)\left(\frac{G}{E}\right)\left(\frac{\sigma_m}{\tau_m}\right)^2 \approx \left(\frac{\sigma_m}{2\tau_m}\right)^2 \tag{34}$$

where use has been made of the well-known relationship (Ref. 15, p. 60) $(E/G) = 2(1 + \nu)$ and of the fact that Poisson's ratio ν is very close to $\frac{1}{3}$ for most structural metals.

If the spring material is to be used most effectively, the stresses τ_m and σ_m should be practically equal to their respective values at the elastic limit. It is seen from (34) that the relative efficiency of Belleville springs versus helical springs depends upon the ratio of (σ_m/τ_m) at the elastic limit. This ratio depends upon the criterion of elastic failure which governs the material. For example, a relatively ductile metal tends to yield (Ref. 16, Section 82) when either the shear stress or the octahedral shear stress reaches a critical value; for such materials it may be shown that $\sigma_m \approx 2\tau_m$. For materials with little ductility, failure generally occurs by fracture and $\sigma_m \approx \tau_m$ (Ref. 17). Therefore, (34) predicts that if the material is stressed up to its useful limit:

$$\frac{u_b}{u_h} \approx 1, \text{ for ductile materials}$$

$$\frac{u_b}{u_h} \approx \frac{1}{4}, \text{ for brittle* materials.}$$

In other terms, both Belleville and helical springs require the same volume of any given ductile metal to store equal amounts of energy; but a Belleville spring can store only $\frac{1}{4}$ the energy stored in an equal-volume helical spring made of the same relatively brittle material. Although there is a tendency towards weight saving in the use of a helical spring made of a relatively brittle material, it may well be that practical geometric considerations, reliability, and the reserve strength of ductile metals would lead one to the choice of a Belleville spring.

To show that reasonable spring weights are required for the present application, let us equate the strain energy in the spring to the kinetic energy $(\frac{1}{2})(W_{\text{sat}}/g)v^2$ required to impart a separation speed v to a satellite of weight $W_{\text{sat}}(g = 386 \text{ in/sec}^2)$. If the volume of spring material is denoted by V_{sp}, (33) leads to the result

$$U_b = u_b V_{\text{sp}} = \frac{\sigma_m^2 V_{\text{sp}}}{6E} = \frac{W_{\text{sat}} v^2}{2g} \tag{35}$$

If one uses the relationship $W_{\text{sp}} = wV_{\text{sp}}$, where W_{sp} is the total weight of the spring and w its specific weight, (35) leads to the conclusion that

$$W_{\text{sp}} = W_{\text{sat}} \frac{3v^2}{g} \left(\frac{wE}{\sigma_m^2}\right). \tag{36}$$

Equation (36) is a compact expression for the weight of a well designed Belleville spring or of a helical spring (for a material with $\sigma_m \approx 2\tau_m$). The material influences the spring weight only through the ratio (σ_m^2/wE), which may be interpreted as twice the elastic energy stored in a unit volume of the material when uniformly stressed at its maximum

* The words ductile and brittle are used in the sense that there either is or is not an appreciable amount of plastic flow between yield and fracture.

TABLE VII — MATERIAL PROPERTIES

Material	E	w	σ_m	wE/σ_m^2
	(psi)	(lb/in^3)	(psi)	(in^{-1})
(*i*) 4130 steel (HT 200,000 psi)	29×10^6	0.282	175,000	2.67×10^{-4}
(*ii*) Aluminum alloy 7075T(6)	10.4×10^6	0.101	73,000	1.97×10^{-4}
(*iii*) Titanium Alloy (Ti-6Al-6V-2Sn)	17×10^6	0.162	190,000	0.763×10^{-4}

allowable value of σ_m. Table VII shows (wE/σ_m^2) for some typical materials of interest. Thus, if one wished to impart a velocity of $v = 12$ ft/sec to a satellite weighing $W_{\text{sat}} = 280$ lb, (36) shows that with the three materials described above, the spring weight W_{sp} would be 12.1 lb, 8.8 lb, and 3.4 lb for materials (*i*), (*ii*) and (*iii*), respectively.

VII. DAMPER UNIT

The damper unit was described in qualitative terms in Section 2.2.2. In this section it will be shown in what respects the PGAC damper differs from other dampers that have been proposed in the literature, and how the damping torques and spring torques must be chosen in order to meet the system requirements outlined in the Introduction, Section I. The hardware development program for the damper units is also described.

Since the means of damping libration motions is perhaps the single most important feature which distinguishes the various attitude control systems that have been proposed by several authors, it would seem worthwhile to indicate the various methods that have been considered. These fall under two main categories: (a) velocity-dependent damping, and (b) amplitude-dependent damping.

In the first category, one finds schemes which depend upon viscous fluids[2,5,18,19] or eddy currents. It has not been demonstrated that practical difficulties concerning seals, viscosity, temperature and adverse rheological effects have been overcome in lightweight systems utilizing fluids. Calculations have shown that effective eddy-current damping requires a considerably greater weight of material than does the magnetic hysteresis unit under discussion.

In the second category of damping methods, the energy loss per cycle is independent of velocity but depends only upon the amplitude of motion. Included in this category are methods based upon Coulomb friction, internal friction,[3,6] and magnetic hysteresis, as described in Section

II of this paper. Coulomb friction (also called dry sliding friction) depends upon physical and chemical surface properties which are notoriously hard to control under conditions of high vacuum and thermal cycling; it is also difficult to control the normal force between the sliding bodies, which greatly influences the level of friction. Solid internal friction, which depends upon energy losses developed in the microstructure of the material, is quite temperature-dependent but does not depend upon unreliable surface properties and should not be unduly influenced by high vacuum.

In addition to the virtues of velocity independence, insensitivity to surface conditions, and lack of rubbing parts, the magnetic hysteresis damper proposed here has been shown to exhibit relative insensitivity to wide temperature fluctuations.

7.1 *Spring Constants*

It is shown[4] that because the deck oscillates about an unstable position of equilibrium, it is necessary to satisfy certain stability criteria. This requires that the torsional spring constants k_1 and k_2 exceed certain critical values k_1^* and k_2^* given by

$$\begin{aligned} k_1^*/I_2\Omega^2 &= 0.625 \text{ (roll)} \\ k_2^*/I_2\Omega^2 &= 1.3 \quad \text{(pitch)} \end{aligned} \tag{37}$$

for the satellite specified in Section 2.2.4, or, for 6000-nm altitude,

$$\begin{aligned} k_1^* &= 0.155 \times 10^{-3} \text{ ft-lb/rad} \\ k_2^* &= 0.324 \times 10^{-3} \text{ ft-lb/rad.} \end{aligned}$$

It is also found that k cannot be too large, since the two-body system becomes so stiff at large k that very small relative displacements between the two bodies are developed and the energy dissipation due to amplitude-dependent damping is reduced. To guarantee stability, it has been decided to keep k_1 and k_2 at least 10 per cent above their critical values. Computer studies indicate that the variation in damping time is relatively small in the range:

$$\begin{aligned} k_1 &= 0.175 \times 10^{-3} \text{ to } 0.375 \times 10^{-3} \text{ ft-lb/rad,} \\ &\qquad\qquad (k_1/I_2\Omega^2 = 0.7 \text{ to } 1.5) \\ k_2 &= 0.36 \times 10^{-3} \text{ to } 0.76 \times 10^{-3} \text{ ft-lb/rad,} \\ &\qquad\qquad (k_2/I_2\Omega^2 = 1.45 \text{ to } 3.06). \end{aligned} \tag{38}$$

From computer solutions it has been noted that if k is smaller than k^*, both the satellite body and the deck body will oscillate about cocked equilibrium positions. This verifies the stability criteria for the spring constants. If the spring constants are much larger than the maximum values given in the above ranges, the relative angular displacements become very small and little energy dissipation occurs.

7.2 *Damping Torque*

Damping torque is produced by rotational hysteresis losses obtained from relative displacement between a magnet, fixed along a diameter of the rotor, and an annular thin disk of cold-rolled steel, fixed to the housing (see Fig. 5). The magnetic fluxes of the magnet pass from the north pole of the magnet through the disk on both halves and back to the south pole, constituting a closed circuit. Except possibly for a small leakage, the unit does not act like a magnetic dipole with respect to the outside field. In the part of the disk near the poles of the magnet there is a relatively high and nonuniform magnetic field. Let the magnetic field in a magnetic domain i be H_i, and let the induced magnetization in the same domain be I_i, which is generally making an angle φ_i with H_i. The magnitude of the retarding torque can be represented by

$$\bar{T}_d = -\sum_i H_i I_i \sin \varphi_i . \tag{39}$$

The minus sign means that the torque tends to oppose the relative displacement between the disk and the magnet.

Provided that the spring constants lie in the ranges specified by (38), computer solutions have shown that there is a relatively small variation in damping time if the damping torques are in the ranges:

$$\begin{aligned} \bar{T}_{d1}/I_2\Omega^2 &= 0.12 \text{ to } 0.29 \text{ (roll)} \\ \bar{T}_{d2}/I_2\Omega^2 &= 0.16 \text{ to } 0.45 \text{ (pitch)} \end{aligned} \tag{40}$$

for the specified satellite. At 6000 nm, the numerical values of $\bar{T}_d$ are

$$\begin{aligned} \bar{T}_{d1} &= 0.30 \times 10^{-4} \text{ to } 0.72 \times 10^{-4} \text{ ft-lb} \\ \bar{T}_{d2} &= 0.40 \times 10^{-4} \text{ to } 1.12 \times 10^{-4} \text{ ft-lb.} \end{aligned} \tag{41}$$

If the damping torques are much lower than the minimum values given above, the satellite will become earth-pointing only after a large number of orbits, as indicated by computer solutions. On the other hand, if the damping torques are much larger than the maximum values, the relative displacements become small and the satellite will keep tumbling for many orbits. In the case of large angle motion the damping time is also found to be greater for values of $\bar{T}_d$ above the ranges given in (41).

7.3 *Hardware Development Program*

Damper units have been developed whose spring constants and damping torques fall in the range given by (38) and (41). These constants are suitable for the 6000-nm satellite previously described. However, the mechanical design is such that the spring constants and damping torques may be adjusted for use with satellites at different altitudes (e.g., between 600 nm and 19,360 nm).

7.3.1 *Spring Design*

Successful torsion spring designs have evolved using both steel wires and flat beryllium-copper ribbons. The torsion springs must be under sufficient tension to prevent the lateral forces (due to gravity differentials, rotational motions, and environmental effects such as solar radiation, etc.) from deflecting the rotor laterally beyond the established clearance, thereby avoiding rubbing or sticking against the housing stops. The lateral forces have been calculated to be smaller than 10^{-3} lb for the 6000-nm satellite previously described. An axial tension force of 6 lb will be more than adequate to resist forces of this level.

Two high-strength steel wires, each of 2-in. length and 0.008-in. diameter, will meet all of the specified requirements and provide a torsional spring constant of 0.36×10^{-3} ft-lb/rad. With a suitably designed end support for the springs, a number of torsional fatigue tests have shown that the springs are capable of withstanding in excess of $\frac{1}{4}$ million cycles at an amplitude of 60°. The static axial tension was 6 to 10 lb. This number of cycles is equivalent to 5 times the expected numbers of libration periods in a 20-year useful life of the satellite.

7.3.2 *Damping Torque Test Program*

The torque-displacement relationship (e.g., T_{d1} versus α for pitch displacement) has been measured for a damper unit at angular speeds between 0.5×10^{-4} and 2×10^{-4} rad/sec (corresponding to a range of angular speeds of $0.18\ \Omega$ to $0.73\ \Omega$ at an altitude of 6000 nm). No dependence of damping torque on the angular speed has been observed in any of the tests, thereby verifying the assumption of velocity-independent magnetic hysteresis damping. Measurements have been made on a number of annular disks of various thicknesses made of cold-rolled

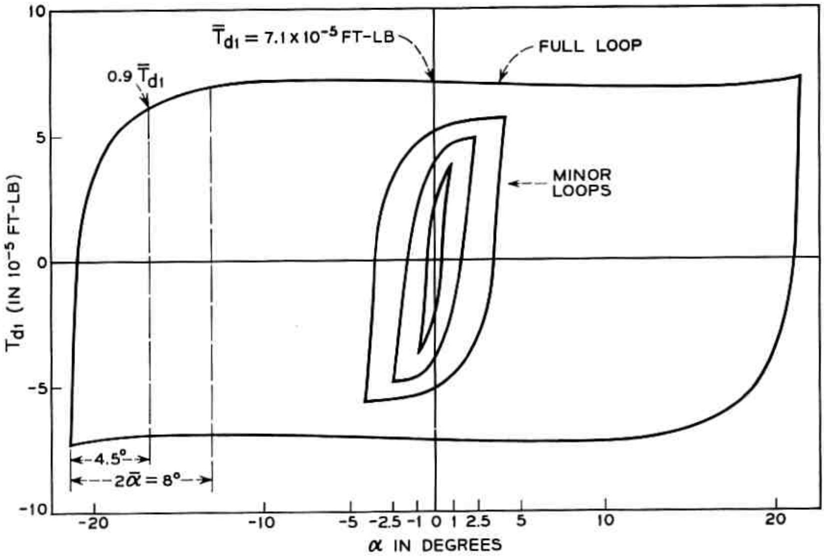


Fig. 13 — Rotational magnetic hysteresis loops.

steel, annealed and unannealed, and of V-Permendur. The maximum torque, $\bar{T}_d$, depends on the volume of the disk, the applied magnetic field and the degree of cold working of the material. A typical T_{d1}-α curve measured on an unannealed cold-rolled steel (1010) is reproduced in Fig. 13. The energy dissipated per cycle is proportional to the area enclosed by the loops in the T_{d1}-α diagram. The slanted part of the curve extends over an angular displacement, $2\bar{\alpha} \approx 8°$, as shown. When the amplitude of the oscillation motion is less than $\bar{\alpha} = 4°$, minor loops as shown in Fig. 13 will be traced out. An appreciable loop area is still obtained even when $\bar{\alpha}$ is as low as 1°.

Measurements have been made on the permeable disks after the disks have been irradiated by an electron flux of $10^{17}/\text{cm}^2$, which is roughly equivalent to the highest electron radiation level anticipated within the Van Allen belt for a period of 30 years. The results indicate that electron radiation has very little effect on the damping torque. The effects of proton bombardment at a flux of 3×10^{12} protons/cm^2 have also been found insignificant on the unannealed cold-rolled steel disks of 0.004- to 0.008-inch thickness without shielding. This flux is equivalent to the highest proton radiation level at 6000 nm for a period of 6 years.

It has also been experimentally observed that the damping torque is relatively insensitive to wide temperature changes. The torque increases only 10 per cent at −40°F and decreases only 15 per cent at +250°F, from its value at room temperature. The temperature of the damper can be controlled within much closer limits in space by appropriate coatings, if desired.

Vibration tests at a 20-g level over a wide frequency band show that the damper will withstand launch conditions.

VIII. ANALYSIS OF DISTURBANCES AND ERRORS

In this section, we shall study the nature and magnitude of disturbing torques which produce forced librational motion of a gravitationally oriented satellite. In Sections 8.1 to 8.6, calculated values are given for each disturbing torque and its corresponding libration angle.

In Section 8.7, the accumulative effects of all the disturbing torques are summarized. It will be seen that the satellite has been so designed that the gravitational torque dominates all disturbing torques at the altitudes of interest.

It is obvious that the amplitude of steady-state librational motion should be kept to a minimum in order that maximum gain can be achieved from the earth-pointing antenna. For example, the theoretical (solid angle) gain at 6000 nm is 14.5 db with no allowance for librational motion. Allowance of a conservative tolerance of 10° on an antenna half angle, to accommodate 10° libration amplitude, results in a 3-db reduction of theoretical antenna gain. However, it will be shown that the steady-state librational amplitude will be less than 10°.

8.1 *Solar Radiation Pressure*

An incident photon beam from the sun to a surface element will be partly absorbed, partly diffusely reflected and partly specularly reflected by the surface, resulting in an exertion of forces in directions normal and tangential to the surface element. These forces produce a net torque about the center of mass of the satellite. A detailed enumeration of the torques contributed by different surface elements on the two-body satellite shown in Fig. 3 indicates that a net maximum solar radiation torque of 0.5×10^{-4} ft-lb will act on the satellite. The magnitude of the gravitational torque at 6000-nm altitude is 0.13×10^{-4} ft-lb per degree of angle, θ, off the local vertical in the orbital plane for small libration angles ($T_{\theta_{\max}} = \frac{3}{2}\Omega^2(I_1 - I_3) = 0.37 \times 10^{-3}$ ft-lb at $\theta = 45°$). Thus, statically the solar torque is balanced by the gravitational torque at $\theta = 4°$, when the sun is in its most unfavorable position. From computer solutions of the dynamics analysis in the pitch case, it is found that the satellite, which is provided with damping, will perform oscillations about the local vertical of a maximum amplitude not greater than 4°. Both the static and dynamic analyses neglected the rod deflections due to solar heating, and the accumulative effects due to solar torque and rod thermal bending will be discussed in the summary, Section 8.7. To be certain that there are no effects which would cause the libration amplitude to appreciably exceed 4°, it would be necessary to perform a three-dimensional analysis.

The foregoing analysis was for a 6000-nm orbit employing the high-deck configuration. For significantly higher orbits the low-deck configuration would be preferred since it would experience less solar torque and a correspondingly smaller deviation from the local vertical.

8.2 *Residual Magnetic Dipole Moment*

The traveling-wave tube employed in a communications satellite such as the Telstar satellite contains two permanent magnets of equal size with the opposite poles placed against each other, thus constituting a quadrupole. Because of possible unequal strength of the two magnets and of inhomogeneous magnetic shielding outside of the traveling-wave tube, there would exist a net residual magnetic dipole moment in the satellite. Both the dipole and the quadrupole moments will interact with the geomagnetic field to produce torques. It can be shown that the torque produced by the quadrupole moment is only about 1 per cent of that produced by the residual dipole moment, when the moments of the two magnets are off by as little as 0.1 per cent. The magnetic moment of the Telstar satellite (produced mainly by the traveling-wave tube) was largely cancelled by the addition of compensating magnets. The residual dipole moment was 10^{-6} weber-meter, the magnitude of which does not seem to have changed much after the satellite was launched into orbit. The use of two traveling-wave tubes, as might be needed in the commercial system, would not appreciably change the satellite's residual magnetic moment.

Other magnetic dipole moments, which exist in the hysteresis damper units and the electric motors of the rod extension units, have been measured to be about 1.4×10^{-6} weber-meter (a value obtained by adding all the moments scalarly). Therefore, a total magnetic dipole moment of 2.4×10^{-6} weber-meter may be expected in the satellite. This value may be reduced by "compensating" the motor dipoles and by further refinement of the cancellation techniques used on the Telstar satellite. Assuming a maximum geomagnetic field of 2.71 amp-turn/meter (0.034 oersted) at 6000-nm altitude, we obtain a maximum torque of about 0.65×10^{-5} newton-meter (0.48×10^{-5} ft-lb). Upon balancing this torque against the gravitational torque in the manner indicated in the preceding Section 8.1, one finds a static libration angle of 0.4°.

8.3 *Orbital Eccentricity*

The eccentricity, ϵ, of an elliptic orbit introduces a forcing torque on the satellite. If the eccentricity is not excessive, the satellite will settle down, as a result of damping, from an initial tumbling motion to a steady-state forced librational motion. In this case, the forcing torque due to eccentricity, $2\epsilon I\Omega^2 \sin(\Omega t + \varphi_0)$ (where $I = I_2$ or $= I_5$), occurs only in the equations of pitch libration. In the case of viscous damping, the steady-state pitch librational angle has been found;[20] however, because of the complexity of the resulting mathematical formulas, they will not be reproduced here. Since an analytical solution has not been obtained in the case of magnetic hysteresis damping, the steady-state libration angle of the earth-pointing body has not been evaluated exactly. However, it may be computed approximately by replacing the hysteresis damping by an equivalent viscous damping for the same energy dissipation per cycle (good only for small oscillations). In so doing, it is found that the libration amplitude $\theta \approx 3.6\epsilon$ radian, corresponding to the numerical data given in Section 2.2.4 and for spring constants and damping torques which fall in the range given in Section VII. A few computer solutions for the case of hysteresis damping indicate that $\theta \approx 5\epsilon$ radian. Guided final-stage vehicles are believed to be capable of achieving orbit eccentricities below 0.005, which would result in libration amplitude of about 1.5°.

8.4 *Meteorite Impact*

Based on Whipple's data,[21] the meteorite flux rate to a spherical surface in the neighborhood of the earth can be shown to be

$$\Phi = C/M \text{ meteorites per meter}^2 \text{ per year} \tag{42}$$

for meteorites of mass $\geqq M$ gram in the range of 10^{-11} to 10^{-1} gram. The constant C (in gram/meter²-year) is found to be $C = 4.16 \times 10^{-5}$ according to Whipple and to be $C = 20.8 \times 10^{-5}$ according to Dubin.[22] For meteorites hitting the deck body, which is at a distance L from the center of mass of the satellite, it can be shown that the expected number of meteorite collisions per year which result in satellite tumbling is

$$\hat{N}_1 = \frac{\pi}{4} LA \frac{nvC}{\sqrt{3pI_y\Omega}} \tag{43}$$

where v is taken to be the average speed of meteorites, A to be the average surface area of the deck body (neglecting the shadowing effect of the earth), I_y is the maximum moment of inertia of the composite satellite, $p = (I_x - I_z)/I_y$ based on the composite satellite, and n is a factor determined by the momentum transfer (<1 for penetration, $= 1$ for completely inelastic impact, $= 2$ for perfectly elastic impact, and >2 for hypervelocity impact when the material is blown backward out of a nearly hemispherical crater). Equation (43) is based on an analysis of the planar pitch motion of a single rigid body satellite, which indicates that if an initial angular speed greater than $\sqrt{3p}\Omega$ is suddenly imparted to a satellite, which is already in line with the local vertical, the satellite will overcome a potential crest and turn over. Based on the result for pitch motion, the expected number of meteorite collision per year to give rise to angles of disturbance from the local vertical in the range from θ_1 to θ_2 ($\leqq 90°$) has been found to be

$$\hat{N}_2 = \frac{\pi}{4} LA \frac{nvC}{\sqrt{3p}I_y\Omega}\left(\frac{1}{\sin\theta_1} - \frac{1}{\sin\theta_2}\right). \tag{44}$$

Both (43) and (44) were derived in a simple manner by approximating the deck as a spherical body. A more exact result can be obtained if the meteorite flux rate is defined with respect to the projected area of a body. In this case the resulting expressions for $\hat{N}_1$ and $\hat{N}_2$ are similar to (43) and (44), respectively, except that the coefficient $(\pi/4)LA$ is replaced by complicated integrals involving the projected area element of various bodies and its distance from the mass center of the satellite.

Numerical values of $\hat{N}$ given in (43) and (44) calculated for the two-body satellite with $C = 4.16 \times 10^{-5}$ are tabulated in Table VIII, from which it is noted that the expected number of turnovers is 0.044 per year (or once in about 23 years). If $C = 20.8 \times 10^{-5}$ is used, based on Dubin's[22] data, all values of $\hat{N}$ in Table VIII should be multiplied by a factor of 5, and the expected turnover rate is 0.22 per year or once in about 4.5 years. These disturbances will be hysteretically damped down to a librational motion with amplitude of 5° in a reasonably short time. For example, computer solutions indicate that the pitch amplitude will be reduced from 45° to 5° in two to four orbital periods. In view of the uncertainty of the meteorite flux rate, all numerical values calculated in this section are to be interpreted as order of magnitude estimates.

8.5 *Cocked Angle Due to Rod Deflections*

The extensible rods will be bent in a natural way and due to the thermal effects, as analyzed in Section V. Consequently, the axes of principal moments of inertia of both bodies will deviate from their positions in the case of perfectly straight rods, and the center of mass of the composite satellite will not lie in the mast rod. As a result, the two bodies may not be perpendicular to each other (or the springs between them may not be in a neutral position) and the desired earth-pointing axis of the satellite may be off from the local vertical by a small cocked angle when the satellite is in a stable equilibrium position. Techniques have been developed for measuring rod straightness so that the natural rod bending in a gravity-free condition will not exceed a predetermined value. It has been found possible to select rods so that the rod bending will not exceed 1 foot for a rod length of 60 feet. This cocked angle could be evaluated if we could find the position vector of the hinge joint in the distorted configuration. A general error analysis has not been made, since it is not known a priori in what way the rods might be deflected. Based on the case of pitch libration, it is found that a deflection of 1 ft of a 60-ft long mast rod will cause a cocked angle of about 0.7° in the case of the high-deck configuration. Assuming that the lateral tip deflections of both the mast and deck rods occur in the same direction, the total cocked angle will be approximately 1.5°. For the case of the low-deck configuration, the cocked angle would be appreciably less.

Rod bending can be caused by solar heating. As has been covered in Section 5.2, the deflection for a 50-ft long rod is expected to be about 2 feet when the sun is perpendicular to the rod. There is a cumulative effect due to the mast and one pair of deck rods when the tips all bend away from the sun during certain periods of the year. The cumulative effects of these rods being bent produce a maximum cocked angle of 3°. This cocked angle can be reduced to less than 1° by silver-plating the rod exterior (the low absorptivity of silver, $\alpha \approx 0.1$, would significantly reduce thermal rod bending due to a reduction of temperature differential across the rod cross section). As will be discussed in the summary, Section 8.7, the effects of rod bending and solar torques are not additive.

8.6 *Miscellaneous Torques*

Torques due to self-gravity and eddy currents have been found to be much smaller than those discussed above. It can be shown that the self-gravity torque acting on one body due to the attraction of the other body is negligibly small compared to the gravitational torque at the altitude of interest. This is due to the fact that the sizes of the two bodies are not significantly different and that their masses are much smaller than that of the earth. Eddy-current losses induced in the conducting materials of the satellite are small because of the low geomagnetic field at the altitudes of interest and because of the satellite's low angular speed. The torques due to self-gravity and eddy currents are less than 10^{-8} ft-lb. The torques produced by plasma effects due to the motion of the satellite in the Van Allen radiation belts are believed to be small. However, it is intended to make a detailed analysis of plasma effects.

8.7 *Summary*

In Table VIII, the maximum libration angles are summarized for each individual disturbance for the high-deck configuration. By simply superposing the individual effects of the various torques, the maximum steady-state libration amplitude is about 10° for the high-deck configuration. However, the various maximum individual effects cannot simply be added to obtain expected maximum libration amplitude. For example, the effect of solar torque and solar rod bending are not additive. The solar torque causes the deck assembly to rotate about the center of mass of the composite satellite in a direction away from the sun, whereas the rod bending due to solar heating causes an effect in the opposite direction. A quantitative analysis is being made of the compensating effects

TABLE VIII — EFFECTS OF DISTURBANCES FOR A 6000-NM SYSTEM

Sources	Maximum Magnitude (ft-lb)	Approximate Librational Angle
Pitch gravity torque: 1.3×10^{-5} per degree off the local vertical		
Solar radiation*	5×10^{-5}	4°
Rod deflection*	See text	
Natural bending		1.5°
Solar heating		3.0°
Orbital eccentricity	See text	1.5° for $\epsilon = 0.005$
Magnetic dipole moment	0.48×10^{-5}	0.4°
Self-gravity and eddy current	Negligible	Negligible

Effects* of Meteorite Impact

Angle from the Local Vertical	$\hat{N}$, Expected Number of Occurrences per Year	Period of Occurrence in Years
5°-15°	0.336	3
15°-30°	0.082	24
30°-50°	0.031	33
50°-70°	0.011	94
70°-90°	0.003	355
5°-90°	0.460	2
>90° (turnover or tumbling)	0.044	23

* Computed for high-deck configuration.

of these two disturbances. Should it develop that the disturbances do not substantially compensate for each other, silver-plating the rods would reduce the effects of both disturbances to less than 5°, rather than 7°, which is the sum of the two disturbance angles. The effects of the other disturbances — natural rod bending, orbit eccentricity, magnetic dipole moment — would not be added to give a total angle of 3.4°. Hence for the high-deck configuration, the final librational angle is expected to be well under 10°. For the low-deck configuration, the librational angle would be expected to be somewhat smaller than that for the high-deck configuration.

IX. CONCLUSIONS

The theoretical feasibility of the proposed PGAC system has been amply demonstrated by more than one hundred computer runs based on the dynamics analysis[4] of the ideal two-rigid-body system. Computer simulations made with a wide variety of initial conditions showed that the system stopped tumbling and then, within about 7 orbital periods, settled down to a state of small oscillations about an earth-pointing direction. It has also been indicated that disturbing influences, such as solar radiation pressure and orbital eccentricity, produce oscillations of less than 10° for a 6000-nm orbit.

The rods have been shown to possess adequate rigidity, to be fully capable of withstanding the loads imposed during the extension phase, and not to undergo excessive bending due to solar heating.

On the basis of comprehensive studies and tests, it is believed that the PGAC system described in this paper is fully capable of meeting all its design objectives, including compatibility with multiple launch procedures, and that it will provide a significant advance in communications satellites practice.

X. ACKNOWLEDGMENTS

The authors wish to acknowledge fully the technical contributions of L. Rongved* and his leadership during the formative stages of the project; H. J. Fletcher* also contributed numerous ideas and mathematical results. Among present Members of Staff at Bell Telephone Laboratories the authors are indebted to J. G. Engstrom, J. W. Stafford, and B. A. Unger who contributed greatly to various analytical and experimental investigations, to L. S. Goldmann who contributed to Section 3 2, and to R. G. Murray for assistance in certain test proce-

* Now with Bellcomm, Inc.

dures. Thanks are also due to Miss E. B. Murphy, Mrs. C. M. Kimme, and Mrs. W. L. Mammel for programming various computations on the IBM 7090 computer.

REFERENCES

1. Hoth, D. F., O'Neill, E. F., and Welber, I., The *Telstar* Satellite System, B.S.T.J., **42,** July, 1963, p. 765.
2. Kamm, L. J., An Improved Satellite Orientation Device, ARS Journal, **32,** No. 6, 1962, pp. 911–913.
3. Fischell, R. E., The TRAAC Satellite, APL Technical Digest, **1,** No. 3, 1962, pp. 2–9.
4. Fletcher, H. J., Rongved, L., and Yu, E. Y., Dynamics Analysis of a Gravitationally Oriented Satellite, this issue, pp. 2239–2266.
5. Nowak, G. H., et al., Unclassified Study of Vertistat Orientation for Communication Satellites, Final Report, Contract NAS5-1898. GD/A Report No. AE 62-0808, 15, September, 1962.
6. Paul, B., Planar Librations of an Extensible Dumbbell Satellite, AIAA Journal, Vol. 1, No. 2, 1963, pp. 411–418.
7. Pierce, J. R., Orbital Radio Relays, Jet Propulsion, **25,** 1955, pp. 153–157.
8. Timoshenko, S., and Young, D. H., *Vibration Problems in Engineering,* D. Van Nostrand, Princeton, N. J., 3rd ed., 1955.
9. Heydon, D. A., Final Report, OGO/Agena B, Separation System Development Tests, STL Document No. 2319-6029-TU-000, 9521.23-128, 26 December, 1962.
10. Aichroth, W. W., Test Report OGO Separation Test, 19V–21, STL Doc. No. 2319-6030-TU-000, 7 January, 1963.
11. Warren, H. R., DeHavilland Antenna Erection Unit, Proc. 5th MIL-E-CON Conference, IRE, 1961, pp. 392–400.
12. Charnes, A., and Raynor, S., Solar Heating of a Rotating Cylindrical Space Vehicle, ARS Journal, **30,** No. 5, May, 1960, pp. 479–484.
13. Timoshenko, S., and Goodier, J. N., *Theory of Elasticity,* McGraw-Hill, New York, 1951, pp. 399–404.
14. Rinehart, J. D., and Robbins, M. F., Characteristics of the Service Provided by Communication Satellites in Uncontrolled Orbits, B.S.T.J., **41,** September, 1962, pp. 1621–1670.
15. Timoshenko, S., *Strength of Materials,* Part I, D. Van Nostrand, Princeton, N. J., 3rd ed., 1958, p. 313.
16. Timoshenko, S., *Strength of Materials,* Part II, D. Van Nostrand, Princeton, N.J., 3rd ed., 1959.
17. Paul, B., A Modification of the Coulomb-Mohr Theory of Fracture, Jour. Appl. Mech., **28,** June, 1961, pp. 259–268.
18. Lewis, J. A., Viscous Damping of Gravitationally Stabilized Satellites, Proc. 4th U.S. Nat. Congr. Appl. Mech., Berkeley, June 18–21, 1962, Am. Soc. Mech. Engrs., 1962, pp. 251–254.
19. Zajac, E. E., Damping of a Gravitationally Oriented Two-Body Satellite, ARS Journal, **32,** December, 1962, pp. 1871–1875.
20. Yu, E. Y., Long Term Coupling Effects between Librational and Orbital Motions of a Satellite, to be published.
21. Whipple, F. L., The Meteoritic Risk to Space Vehicles, in *Vistas in Astronautics,* **1,** M. Halperin and M. Stern, Eds., Pergamon Press, New York, 1958.
22. Dubin, M., IGY Micrometeorite Measurements, in *Space Research,* H. Kallmann, Ed., North-Holland Publ. Co., Amsterdam, 1960, pp. 1042–1058.

Dynamics Analysis of a Two-Body Gravitationally Oriented Satellite

By H. J. FLETCHER,[†] L. RONGVED[†] and E. Y. YU

(Manuscript received July 27, 1962)

The rigid body motion of a two-body satellite under the action of gravitational torques is analyzed. The satellite consists of two rigid bodies connected by a universal joint where damping is provided in the two journals. The motion of the satellite relative to the mass center thus has five degrees of freedom, two of which are provided with energy dissipation. It appears that the rigid body motion of such a composite satellite will automatically converge upon a motion in which a given axis of the satellite is earth-pointing.

The equations of motion are derived directly from those of Euler. Necessary stability criteria are established. Numerical solutions for a practical scheme are presented.

I. INTRODUCTION

This paper deals with the analysis of the rotational motion of a satellite consisting of two rigid bodies connected by a hinge mechanism of universal joint type. The rotational motion of the satellite thus has five degrees of freedom; the two degrees of freedom that involve the relative motion between the two bodies are provided with energy dissipation. It is found that any motion of the satellite with respect to the local vertical always involves relative motion between the two bodies. Therefore, the damping at the hinge joint dissipates not only the relative motion of the two bodies but also the motion of the satellite with respect to the local vertical. The satellite will then converge upon a stable motion in which a specified axis of the satellite will remain close to the local vertical.

The equations of motion are derived directly from those of Newton and Euler. This approach naturally suggests several additional dependent variables and results in numerically workable equations. This is not the case in the Lagrangian formulation.

There are several practical problems involved in this scheme of pas-

[†] Bellcomm, Inc.

sive gravitational orientation. One problem is to make the gravitational torque dominate over all other disturbing torques. A novel solution to this problem, which employs extensible rods, has been given by Kamm.[1] Another problem is the development of the hinge dissipative mechanism. A viscous mechanism is described by Kamm,[1] whereas a hysteresis mechanism is suggested in this paper. These practical matters are not the substance of this paper; they are used as illustrations for the numerical treatment of a practical design.

II. GENERAL EQUATIONS OF MOTION

Consider a satellite which is constructed of two rigid bodies, with masses m_1 and m_2, hinged at a point H. The centers of mass of the two bodies are denoted S_1 and S_2, and the center of mass of the composite satellite is denoted S_0. Let the earth's center be O and let P_1 and P_2 be arbitrary points of body 1 and 2. Also, denote $OP_1 = \mathbf{R}_1$, $OS_1 = \boldsymbol{\varrho}_1$, $OS_0 = \boldsymbol{\varrho}$, $OS_2 = \boldsymbol{\varrho}_2$, $OP_2 = \mathbf{R}_2$, $S_1P_1 = \mathbf{r}_1$, $S_2P_2 = \mathbf{r}_2$, $HS_1 = \mathfrak{L}_1$, $HS_2 = \mathfrak{L}_2$ (see Fig. 1). (Note: $\mathfrak{L}_1$ and $\mathfrak{L}_2$ represent vectors, while ℓ_1 and ℓ_2 which appear later represent their respective magnitudes. See Appendix for list of symbols.)

Let us introduce the following notations:

$\boldsymbol{\omega}_{\text{I}}$, $\boldsymbol{\omega}_{\text{II}}$ = angular velocity of body 1, 2,

$\mathbf{T}_H$ = reactive torque transmitted through the joint on body 1,

$\mathbf{F}_H$ = reactive force transmitted through the joint on body 1,

$\mathbf{T}_1$, $\mathbf{T}_2$ = resultant torque on body 1, 2 exclusive of $\mathbf{T}_H$,

$\mathbf{F}_1$, $\mathbf{F}_2$ = resultant force on body 1, 2 exclusive of $\mathbf{F}_H$,

m_1, m_2 = mass of body 1, 2,

$\bar{m} = m_1m_2/(m_1 + m_2)$ = reduced mass of the system,

$m = m_1 + m_2$ = total mass of the system,

$\boldsymbol{\Phi}_1$, $\boldsymbol{\Phi}_2$ = moment of inertia dyadic of body 1, 2.

Newton's and Euler's equations can now be written as

$$\mathbf{F}_1 + \mathbf{F}_H = m_1\ddot{\boldsymbol{\varrho}}_1 \tag{1a}$$

$$\mathbf{F}_2 - \mathbf{F}_H = m_2\ddot{\boldsymbol{\varrho}}_2 \tag{1b}$$

$$\boldsymbol{\Phi}_1\cdot\dot{\boldsymbol{\omega}}_{\text{I}} + \boldsymbol{\omega}_{\text{I}} \times \boldsymbol{\Phi}_1\cdot\boldsymbol{\omega}_{\text{I}} = \mathbf{T}_1 + \mathbf{T}_H - \mathfrak{L}_1 \times \mathbf{F}_H \tag{1c}$$

$$\boldsymbol{\Phi}_2\cdot\dot{\boldsymbol{\omega}}_{\text{II}} + \boldsymbol{\omega}_{\text{II}} \times \boldsymbol{\Phi}_2\cdot\boldsymbol{\omega}_{\text{II}} = \mathbf{T}_2 - \mathbf{T}_H + \mathfrak{L}_2 \times \mathbf{F}_H \tag{1d}$$

where the dots indicate time derivatives with respect to an inertial frame. Because of the constraint imposed by the hinge, the following relations are satisfied

$$\boldsymbol{\rho}_1 = \boldsymbol{\rho} + (\mathfrak{L}_1 - \mathfrak{L}_2)\,\frac{m_2}{m} \tag{2a}$$

or

$$\boldsymbol{\rho}_2 = \boldsymbol{\rho} + (\mathfrak{L}_2 - \mathfrak{L}_1)\,\frac{m_1}{m}. \tag{2b}$$

Addition of (1a) and (1b) yields the following vector equation which governs the motion of the mass center S_0 :

$$\mathbf{F}_1 + \mathbf{F}_2 = m\ddot{\boldsymbol{\rho}}. \tag{3}$$

Using (1a), (2a), and (3) we may solve for $\mathbf{F}_H$

$$\mathbf{F}_H = \frac{m_1}{m}\,\mathbf{F}_2 - \frac{m_2}{m}\,\mathbf{F}_1 + \bar{m}(\ddot{\mathfrak{L}}_1 - \ddot{\mathfrak{L}}_2). \tag{4}$$

Inserting (4) in (1c) and (1d) and using the fact that

$$\dot{\mathfrak{L}}_1 = \boldsymbol{\omega}_{\mathrm{I}} \times \mathfrak{L}_1 \quad \text{and} \quad \ddot{\mathfrak{L}}_1 = \dot{\boldsymbol{\omega}}_{\mathrm{I}} \times \mathfrak{L}_1 + \boldsymbol{\omega}_{\mathrm{I}} \times (\boldsymbol{\omega}_{\mathrm{I}} \times \mathfrak{L}_1)\,,$$

etc., equations (1c) and (1d) become

$$\begin{aligned}\boldsymbol{\Phi}_1'\cdot\dot{\boldsymbol{\omega}}_{\mathrm{I}} + \boldsymbol{\omega}_{\mathrm{I}} \times \boldsymbol{\Phi}_1'\cdot\boldsymbol{\omega}_{\mathrm{I}} &= \mathbf{T}_1 \;\; + \mathbf{T}_H \\ &+ \mathfrak{L}_1 \times \left\{\frac{m_2}{m}\,\mathbf{F}_1 - \frac{m_1}{m}\,\mathbf{F}_2 + \bar{m}[\boldsymbol{\omega}_{\mathrm{II}} \times (\boldsymbol{\omega}_{\mathrm{II}} \times \mathfrak{L}_2) + \dot{\boldsymbol{\omega}}_{\mathrm{II}} \times \mathfrak{L}_2]\right\}\end{aligned} \tag{5a}$$

$$\begin{aligned}\boldsymbol{\Phi}_2'\cdot\dot{\boldsymbol{\omega}}_{\mathrm{II}} + \boldsymbol{\omega}_{\mathrm{II}} \times \boldsymbol{\Phi}_2'\cdot\boldsymbol{\omega}_{\mathrm{II}} &= \mathbf{T}_2 - \mathbf{T}_H \\ &+ \mathfrak{L}_2 \times \left\{\frac{m_1}{m}\,\mathbf{F}_2 - \frac{m_2}{m}\,\mathbf{F}_1 + \bar{m}[\boldsymbol{\omega}_{\mathrm{I}} \times (\boldsymbol{\omega}_{\mathrm{I}} \times \mathfrak{L}_1) + \dot{\boldsymbol{\omega}}_{\mathrm{I}} \times \mathfrak{L}_1]\right\}\end{aligned} \tag{5b}$$

where $\boldsymbol{\Phi}_i' = \boldsymbol{\Phi}_i + \bar{m}(\ell_i^2\mathbf{I} - \mathfrak{L}_i\mathfrak{L}_i)$, $i = 1, 2$, and $\mathbf{I}$ is the unit dyadic.

III. GRAVITATIONAL FORCE

The earth's gravitational field is taken to be radially symmetric. The gravitational force, $d\mathbf{G}_i$, acting on an infinitesimal mass dm_i at P_i is then

$$d\mathbf{G}_i = -\frac{\mu\, dm_i}{R_i^3}\,\mathbf{R}_i \tag{6}$$

where $\mu = gR_E^2$ with g being the gravitational acceleration at the earth's surface and R_E being the earth's radius. From Fig. 1

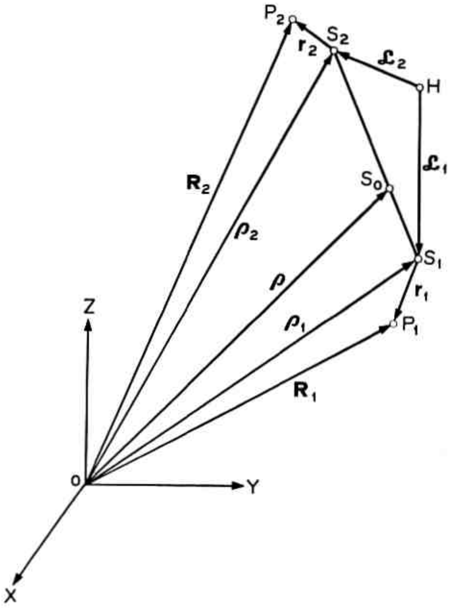


Fig. 1 — Vector displacement diagram of a two-body satellite.

$$d\mathbf{G}_i = -\frac{\mu\, dm_i}{\rho_i^3}(\boldsymbol{\varrho}_i + \mathbf{r}_i)\left(1 + \frac{2\boldsymbol{\varrho}_i\cdot\mathbf{r}_i}{\rho_i^2} + \frac{r_i^2}{\rho_i^2}\right)^{-\frac{3}{2}}$$

$$= \left[-\frac{\mu\, dm_i}{\rho_i^3}\boldsymbol{\varrho}_i - \frac{\mu\, dm_i}{\rho_i^3}\mathbf{r}_i + \frac{3\mu\boldsymbol{\varrho}_i}{\rho_i^5}(\boldsymbol{\varrho}_i\cdot\mathbf{r}_i)\, dm_i\right]\left[1 + O\left(\frac{l^2}{\rho_i^2}\right)\right] \quad (7)$$

where the last quantity represents terms of order l^2/ρ_i^2 and higher and l is the maximum linear dimension of the satellite. These higher-order terms are neglected in the analysis. Since S_i is the center of mass of body i

$$\int_{m_i} \mathbf{r}_i\, dm_i = 0.$$

Hence $\mathbf{G}_1$, the gravitational force on body 1, is

$$\mathbf{G}_1 = -\frac{\mu m_1\boldsymbol{\varrho}_1}{\rho_1^3}\left[1 + O\left(\frac{l^2}{\rho_1^2}\right)\right]$$

or, by (2a)

$$\mathbf{G}_1 = \left[-\frac{\mu m_1\boldsymbol{\varrho}}{\rho^3} + \frac{\mu\bar{m}}{\rho^3}(\mathfrak{L}_2 - \mathfrak{L}_1)\cdot(\mathbf{I} - 3\hat{\rho}\hat{\rho})\right]\left[1 + O\left(\frac{l^2}{\rho^2}\right).\right] \quad (8a)$$

Similarly,

$$\mathbf{G}_2 = \left[-\frac{\mu m_2 \boldsymbol{\rho}}{\rho^3} - \frac{\mu \bar{m}}{\rho^3} (\mathfrak{L}_2 - \mathfrak{L}_1) \cdot (\mathbf{I} - 3\hat{\rho}\hat{\rho}) \right] \left[1 + O\left(\frac{l^2}{\rho^2}\right) \right] \tag{8b}$$

where the symbol "ˆ" denotes a unit vector. Using (7), the gravitational torque acting on body i about the center of mass is given by

$$\mathbf{T}_{Gi} = \int \mathbf{r}_i \times d\mathbf{G}_i = \frac{3\mu}{\rho^3} \hat{\rho} \times \boldsymbol{\Phi}_i \cdot \hat{\rho} \left[1 + O\left(\frac{l}{\rho}\right) \right], \qquad i = 1, 2. \tag{9}$$

Let $\mathbf{F}_i = \mathbf{F}_i' + \mathbf{G}_i$, $\mathbf{T}_i = \mathbf{T}_i' + \mathbf{T}_{Gi}$, $i = 1, 2$. Substituting in (5*a,b*) with the gravitational torques in (9) and the gravitational forces in (8) with terms of $O(l/\rho)$ and $O(l^2/\rho^2)$ neglected, the general equations of rotational motion of two hinged-connected rigid bodies become

$$\begin{aligned} \boldsymbol{\Phi}_1' \cdot \dot{\boldsymbol{\omega}}_{\mathbf{I}} + \boldsymbol{\omega}_{\mathbf{I}} \times \boldsymbol{\Phi}_1' \cdot \boldsymbol{\omega}_{\mathbf{I}} = {} & \frac{3\mu}{\rho^3} \hat{\rho} \times \boldsymbol{\Phi}_1' \cdot \hat{\rho} \\ & + \frac{\mu \bar{m}}{\rho^3} (\mathfrak{L}_1 \times \mathfrak{L}_2 - 3\mathfrak{L}_1 \times \hat{\rho}\hat{\rho} \cdot \mathfrak{L}_2) + \mathbf{T}_1' + \mathbf{T}_H \\ & - \frac{m_1}{m} \mathfrak{L}_1 \times \mathbf{F}_2' + \frac{m_2}{m} \mathfrak{L}_1 \times \mathbf{F}_1' \\ & + \bar{m}(\boldsymbol{\omega}_{\mathbf{II}} \cdot \mathfrak{L}_2 \mathfrak{L}_1 \times \boldsymbol{\omega}_{\mathbf{II}} - \boldsymbol{\omega}_{\mathbf{II}}^2 \mathfrak{L}_1 \times \mathfrak{L}_2 + \mathfrak{L}_1 \cdot \mathfrak{L}_2 \dot{\boldsymbol{\omega}}_{\mathbf{II}} - \mathfrak{L}_2 \mathfrak{L}_1 \cdot \dot{\boldsymbol{\omega}}_{\mathbf{II}}), \end{aligned} \tag{10a}$$

$$\begin{aligned} \boldsymbol{\Phi}_2' \cdot \dot{\boldsymbol{\omega}}_{\mathbf{II}} + \boldsymbol{\omega}_{\mathbf{II}} \times \boldsymbol{\Phi}_2' \cdot \boldsymbol{\omega}_{\mathbf{II}} = {} & \frac{3\mu}{\rho^3} \hat{\rho} \times \boldsymbol{\Phi}_2' \cdot \hat{\rho} \\ & + \frac{\mu \bar{m}}{\rho^3} (\mathfrak{L}_2 \times \mathfrak{L}_1 - 3\mathfrak{L}_2 \times \hat{\rho}\hat{\rho} \cdot \mathfrak{L}_1) + \mathbf{T}_2' - \mathbf{T}_H \\ & - \frac{m_2}{m} \mathfrak{L}_2 \times \mathbf{F}_1' + \frac{m_1}{m} \mathfrak{L}_2 \times \mathbf{F}_2' \\ & + \bar{m}(\boldsymbol{\omega}_{\mathbf{I}} \cdot \mathfrak{L}_1 \mathfrak{L}_2 \times \boldsymbol{\omega}_{\mathbf{I}} - \boldsymbol{\omega}_{\mathbf{I}}^2 \mathfrak{L}_2 \times \mathfrak{L}_1 + \mathfrak{L}_2 \cdot \mathfrak{L}_1 \dot{\boldsymbol{\omega}}_{\mathbf{I}} - \mathfrak{L}_1 \mathfrak{L}_2 \cdot \dot{\boldsymbol{\omega}}_{\mathbf{I}}). \end{aligned} \tag{10b}$$

Note that $\mathbf{T}_1'$ and $\mathbf{T}_2'$ are the resultant torques imposed on body 1 and 2 by some external sources other than gravity. They do not include torques arising from the reaction of one body upon the other. Similarly, $\mathbf{F}_1'$ and $\mathbf{F}_2'$ are the resultant forces on body 1 and 2 due to external sources other than gravity. They do not include the reaction of one body upon the other. Thus both these torques and forces are worked out as though the bodies were not connected. Various environmental disturbances, like solar radiation pressure or interaction of a magnetic moment in the satellite with the geomagnetic field, may be taken into account by assigning appropriate values to $\mathbf{T}_1'$, $\mathbf{T}_2'$, $\mathbf{F}_1'$, and $\mathbf{F}_2'$. This subject is not treated here to conserve space.

From (8a) and (8b) it is seen that

$$\mathbf{G}_1 + \mathbf{G}_2 = -\frac{\mu m \boldsymbol{\varrho}}{\rho^3}. \tag{11a}$$

If the gravitational forces are the only ones in $\mathbf{F}_1$ and $\mathbf{F}_2$, then (3) becomes

$$\ddot{\boldsymbol{\varrho}} = -\frac{\mu \boldsymbol{\varrho}}{\rho^3}. \tag{11b}$$

The solution of this vector equation is an elliptical orbit of S_0, independent of the rotations of the satellite, because of the fact that terms of $O(l^2/\rho^2)$ have now been neglected. As the earth's gravitational field is assumed to be radially symmetric, the orbital plane is fixed in the inertial space.

IV. COORDINATE SYSTEMS

Four reference frames are used to describe the motions of the satellite.

The first frame has its origin at the geocenter O with the Z-axis through the perigee of the orbit and with the Y-axis in the direction of the orbital angular momentum. The X-axis is chosen to form a right-handed set of axes (see Fig. 2). This coordinate system is taken to be inertial.

The second is an earth-pointing frame. It has its origin at the satellite's center of mass, S_0, with the z-axis along OS_0 making an angle ψ with the Z-axis. The y-axis is parallel to the Y-axis, and the x-axis is chosen to form a right-handed system. The relationship between the unit vectors of the coordinate systems O-XYZ and S_0-xyz is

$$\begin{pmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{pmatrix} = \begin{pmatrix} C\psi & 0 & -S\psi \\ 0 & 1 & 0 \\ S\psi & 0 & C\psi \end{pmatrix} \begin{pmatrix} \hat{X} \\ \hat{Y} \\ \hat{Z} \end{pmatrix} \tag{12}$$

where S and C are abbreviations of sine and cosine.

The third frame has its origin at S_1 with axes S_1-$x_1y_1z_1$ along the principal axes of inertia of body 1. Euler parameters[2] are employed to describe the motion of S_1-$x_1y_1z_1$ relative to S_0-xyz. The transformation is given by

$$\begin{pmatrix} \hat{x}_1 \\ \hat{y}_1 \\ \hat{z}_1 \end{pmatrix} = (a_{ij}) \begin{pmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{pmatrix} \tag{13a}$$

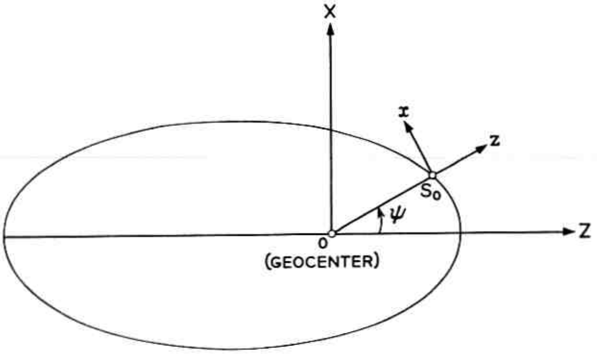


Fig. 2 — Coordinates of the rotating and nonrotating frames.

where

$$(a_{ij}) = \begin{pmatrix} \xi^2 - \eta^2 - \zeta^2 + \chi^2 & 2(\xi\eta + \zeta\chi) & 2(\xi\zeta - \eta\chi) \\ 2(\xi\eta - \zeta\chi) & -\xi^2 + \eta^2 - \zeta^2 + \chi^2 & 2(\xi\chi + \eta\zeta) \\ 2(\xi\zeta + \eta\chi) & 2(-\xi\chi + \eta\zeta) & -\xi^2 - \eta^2 + \zeta^2 + \chi^2 \end{pmatrix} \tag{13b}$$

$i, j = 1, 2, 3$ representing rows and columns respectively, and

$$\xi^2 + \eta^2 + \zeta^2 + \chi^2 = 1. \tag{13c}$$

The fourth frame has an origin at S_2 with axes $S_2\text{-}x_2y_2z_2$ along the principal axes of inertia of body 2. If a universal joint is used, the relative rotation of the second body can be completely specified with only two angles, namely α, the rotation of the journal in body 1, and β, the rotation of the journal of body 2. When these two journals are directed respectively along $\hat{x}_1$ and $\hat{y}_2$, then the transformation from $S_1\text{-}x_1y_1z_1$ to $S_2\text{-}x_2y_2z_2$ is given by

$$\begin{pmatrix} \hat{x}_2 \\ \hat{y}_2 \\ \hat{z}_2 \end{pmatrix} = (b_{ij}) \begin{pmatrix} \hat{x}_1 \\ \hat{y}_1 \\ \hat{z}_1 \end{pmatrix} \tag{14a}$$

where

$$(b_{ij}) = \begin{pmatrix} C\beta & S\alpha\, S\beta & -C\alpha\, S\beta \\ 0 & C\alpha & S\alpha \\ S\beta & -S\alpha\, C\beta & C\alpha\, C\beta \end{pmatrix}. \tag{14b}$$

The constraint equation $\hat{x}_1 \cdot \hat{y}_2 = 0$ is automatically satisfied by the introduction of the two coordinate parameters α and β. The angular velocities of the two bodies are

$$\boldsymbol{\omega}_{\mathrm{I}} = \dot{\psi}\hat{y} + \lambda_1\hat{x}_1 + \lambda_2\hat{y}_1 + \lambda_3\hat{z}_1 \tag{15a}$$

where

$$\lambda_1 = 2(\chi\dot{\xi} + \zeta\dot{\eta} - \eta\dot{\zeta} - \xi\dot{\chi}) \tag{15b}$$

$$\lambda_2 = 2(-\zeta\dot{\xi} + \chi\dot{\eta} + \xi\dot{\zeta} - \eta\dot{\chi}) \tag{15c}$$

$$\lambda_3 = 2(\eta\dot{\xi} - \xi\dot{\eta} + \chi\dot{\zeta} - \zeta\dot{\chi}) \tag{15d}$$

and

$$\boldsymbol{\omega}_{\mathrm{II}} = \boldsymbol{\omega}_{\mathrm{I}} + \dot{\alpha}\hat{x}_1 + \dot{\beta}\hat{y}_2 \,. \tag{15e}$$

V. SPECIALIZED EQUATIONS OF MOTION

Let us specialize our satellite so that $\mathcal{L}_2 = 0$ and $\mathcal{L}_1 = -\ell_1\hat{z}_1$. We assume gravity to be the only external force, i.e., $\mathbf{T}_1'$, $\mathbf{T}_2'$, $\mathbf{F}_1'$ and $\mathbf{F}_2'$ in (10) are taken to be zero. Then equations (10) are equivalent to those derived from two bodies connected at their centers of mass except that the inertia dyadic $\mathbf{\Phi}_1$ is replaced by $\mathbf{\Phi}_1'$ defined in (5) ($\mathbf{\Phi}_2' = \mathbf{\Phi}_2$ as $\ell_2 = 0$). The two bodies are connected by a universal joint, which is characterized by an interposed weightless body, having two perpendicular journals as previously described. The torque $\mathbf{T}_H$, transmitted through the universal joint, consists of the constraint torque $\mathbf{T}_c$, the elastic restoring torque $\mathbf{T}_r$, and the dissipative torque $\mathbf{T}_d$. The components of the latter two along the journals x_1 and y_2 are specified by subscripts 1 and 2 respectively. Hence $\mathbf{T}_H$ can be written as

$$\mathbf{T}_H = T_c\hat{x}_1 \times \hat{y}_2 + (T_{r1} + T_{d1})\hat{x}_1 + (T_{r2} + T_{d2})\hat{y}_2 \,. \tag{16}$$

Let

$$I_1 = \mathbf{\Phi}_1' \cdot \hat{x}_1 \tag{17a}$$

$$I_2 = \mathbf{\Phi}_1' \cdot \hat{y}_1 \tag{17b}$$

$$I_3 = \mathbf{\Phi}_1' \cdot \hat{z}_1 \tag{17c}$$

$$I_4 = \mathbf{\Phi}_2 \cdot \hat{x}_2 \tag{17d}$$

$$I_5 = \mathbf{\Phi}_2 \cdot \hat{y}_2 \tag{17e}$$

$$I_6 = \mathbf{\Phi}_2 \cdot \hat{z}_2 \tag{17f}$$

$$\omega_i(i = 1, 2, 3) = \text{components of } \boldsymbol{\omega}_{\mathrm{I}} \text{ along } \hat{x}_1\,, \hat{y}_1\,, \hat{z}_1 \tag{17g}$$

$$\omega_i(i = 4, 5, 6) = \text{components of } \boldsymbol{\omega}_{\mathrm{II}} \text{ along } \hat{x}_2\,, \hat{y}_2\,, \hat{z}_2 \,. \tag{17h}$$

From the orbit equation (11b), the following relations can be derived

$$\dot{\psi} = \frac{\Omega}{(1 - \epsilon^2)^{\frac{3}{2}}} (1 + \epsilon C\psi)^2 \tag{18}$$

$$G = \frac{3\mu}{\rho^3} = \frac{3\Omega^2}{(1 - \epsilon^2)^3}(1 + \epsilon C\psi)^3 \tag{19}$$

where

ϵ = eccentricity of the orbit

Ω = 2π divided by the orbital period.

Euler's equations of motion (10), simplified for the specialized satellite, are written out as

$$I_1\dot{\omega}_1 = (I_2 - I_3)(\omega_2\omega_3 - Gn_2n_3) + T_{r1} + T_{d1} \tag{20a}$$

$$I_2\dot{\omega}_2 = (I_3 - I_1)(\omega_3\omega_1 - Gn_3n_1) + (T_{r2} + T_{d2})C\alpha - T_cS\alpha \tag{20b}$$

$$I_3\dot{\omega}_3 = (I_1 - I_2)(\omega_1\omega_2 - Gn_1n_2) + (T_{r2} + T_{d2})S\alpha + T_cC\alpha \tag{20c}$$

$$I_4\dot{\omega}_4 = (I_5 - I_6)(\omega_5\omega_6 - Gn_5n_6) - (T_{r1} + T_{d1})C\beta + T_cS\beta \tag{20d}$$

$$I_5\dot{\omega}_5 = (I_6 - I_4)(\omega_6\omega_4 - Gn_6n_4) - T_{r2} - T_{d2} \tag{20e}$$

$$I_6\dot{\omega}_6 = (I_4 - I_5)(\omega_4\omega_5 - Gn_4n_5) - (T_{r1} + T_{d1})S\beta - T_cC\beta \tag{20f}$$

where

$$n_i = a_{i3} \qquad \text{(see 13b)} \qquad i = 1, 2, 3 \tag{20g}$$

$$n_{i+3} = \sum_{k=1}^{3} b_{ik}a_{k3} \qquad \text{(see 14b)} \qquad i = 1, 2, 3. \tag{20h}$$

Because of the constraint $\hat{x}\cdot\hat{y} = 0$, a relation must exist among the six ω_i's. Such a relation, i.e., $(\boldsymbol{\omega}_{\mathrm{I}} - \boldsymbol{\omega}_{\mathrm{II}})\cdot(\hat{x}_1 \times \hat{y}_2) = 0$, can be obtained from (15e). This yields the following relationship:

$$\omega_2S\alpha - \omega_3C\alpha - \omega_4S\beta + \omega_6C\beta = 0. \tag{21}$$

If (21) is differentiated and equations (20) are substituted, the unknown T_c is found to be

$$T_c = \left(\frac{S^2\alpha}{I_2} + \frac{C^2\alpha}{I_3} + \frac{S^2\beta}{I_4} + \frac{C^2\beta}{I_6}\right)^{-1}$$

$$\left\{\frac{S\alpha}{I_2}[(I_3 - I_1)(\omega_1\omega_3 - Gn_1n_3) + (T_{r2} + T_{d2})\,C\alpha]\right.$$

$$- \frac{C\alpha}{I_3}[(I_1 - I_2)(\omega_1\omega_2 - Gn_1n_2) + (T_{r2} + T_{d2})\,S\alpha]$$

$$- \frac{S\beta}{I_4}[(I_5 - I_6)(\omega_5\omega_6 - Gn_5n_6) - (T_{r1} + T_{d1})\,C\beta]$$

$$+ \frac{C\beta}{I_6}\left[(I_4 - I_5)(\omega_4\omega_5 - Gn_4n_5) - (T_{r1} + T_{d1})\, S\beta\right]$$

$$\left. + \dot{\alpha}(\omega_2 C\alpha + \omega_3 S\alpha) - \dot{\beta}(\omega_4 C\beta + \omega_6 S\beta)\right\}. \tag{22}$$

Equations (20) could be considered as a system of six second-order equations in six unknowns ξ, η, ζ, χ, α, β, while ψ is determined from (18) and the ω_i's from (15). For computation purposes it is convenient to also leave the ω_i's as dependent variables. Equations (15) give

$$\dot{\xi} = \tfrac{1}{2}(\chi\lambda_1 - \zeta\lambda_2 + \eta\lambda_3) \tag{23a}$$

$$\dot{\eta} = \tfrac{1}{2}(\zeta\lambda_1 + \chi\lambda_2 - \xi\lambda_3) \tag{23b}$$

$$\dot{\xi} = \tfrac{1}{2}(-\eta\lambda_1 + \xi\lambda_2 + \chi\lambda_3) \tag{23c}$$

$$\dot{\chi} = \tfrac{1}{2}(-\xi\lambda_1 - \eta\lambda_2 - \zeta\lambda_3) \tag{23d}$$

$$\dot{\alpha} = -\omega_1 + \omega_4 C\beta + \omega_6 S\beta \tag{23e}$$

$$\dot{\beta} = -\omega_2 C\alpha - \omega_3 S\alpha + \omega_5 \tag{23f}$$

$$\lambda_i \equiv \omega_i - a_{i2}\dot{\psi}, \qquad i = 1, 2, 3. \tag{23g}$$

There are now 12 first-order equations, (20 a-f) and (23a-f), in the unknowns ξ, η, ζ, χ, α, β, ω_1, ω_2, ω_3, ω_4, ω_5, and ω_6. If Euler angles had been used instead of Euler parameters, there would be certain positions of the body for which the derivatives of the angles have a singularity. However, no singularities occur when Euler parameters are used, as can be seen from (23). It should also be noticed from (13b) that the matrix fixing the body position is not changed if the coordinates (ξ, η, ζ, χ) are replaced by $(-\xi, -\eta, -\zeta, -\chi)$.

VI. DISSIPATIVE AND ELASTIC TORQUES IN THE UNIVERSAL JOINT

To completely define the problem it is necessary to specify the elastic and dissipative torques $\mathbf{T}_r$ and $\mathbf{T}_d$.

6.1 *Damping Torques*

Two types of damping torques are considered here. The first is viscous damping of the linear velocity type; the torque on body 1 has two components

$$\mathbf{T}_{d1} = C_1\dot{\alpha}\hat{x}_1 \tag{24a}$$

$$\mathbf{T}_{d2} = C_2\dot{\beta}\hat{y}_2 \tag{24b}$$

where C_1 and C_2 are viscous damping coefficients.

The second is of magnetic hysteresis type. The damping is furnished by hysteresis losses produced by the relative motion of a permanent magnet and a permeable material. The torque in the x_1-direction might be approximately expressed by the following process. If $\dot{\alpha} > 0$, the torque would be represented in region I in Fig. 3 by

$$T_{d1} = T_{d1}^* + \bar{T}_{d1} \frac{\alpha - \alpha^*}{\bar{\alpha}} \tag{25}$$

as long as $| T_{d1} | < \bar{T}_{d1}$ where $\bar{\alpha}$, $\bar{T}_{d1}$ are constants and α^*, T_{d1}^* are the values of α, T_{d1} when $\dot{\alpha}$ last changed sign. After $| T_{d1} |$ reaches $\bar{T}_{d1}$ then T_{d1} remains at $\bar{T}_{d1}$ as long as $\dot{\alpha}$ does not change sign. This is represented as region II in Fig. 3. If $\dot{\alpha}$ changes sign, then (25) applies and the process is repeated. This is represented by region III of Fig. 3. T_{d2} is defined by replacing α by β and subscript 1 by 2 in (25). According to this idealized hysteresis, no energy is dissipated in region III. In an actual device, energy would also be dissipated in this region because of minor hysteresis loops. The chief advantage of magnetic hysteresis damping is that it is amplitude dependent instead of velocity dependent, since the librational frequency, which is of the order of the orbital frequency, is too low to make the velocity damping effective. Other merits of the magnetic hysteresis damper will be stated in the descriptions of a practical design for a numerical computation.

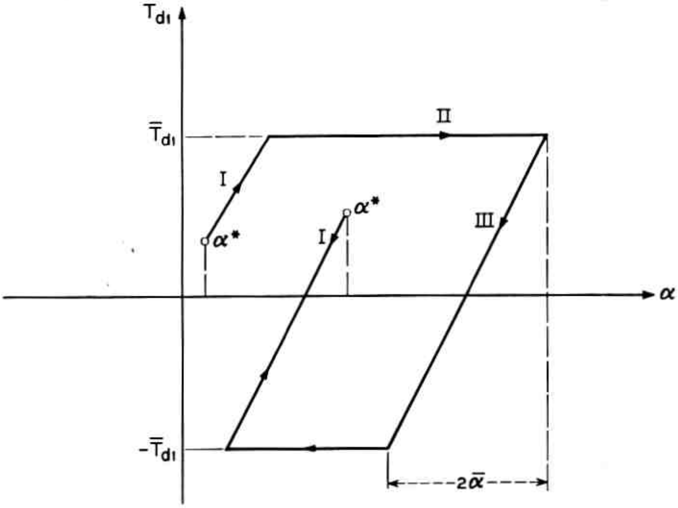


Fig. 3 — Magnetic hysteresis damping torque produced by a magnetic device on the x_1-journal.

6.2 *Elastic Torques*

It is assumed that each journal is furnished with a linearly elastic restoring torque produced, for example, by the torsion of a wire. The torque acting on body 1 is given by

$$\mathbf{T}_{r1} = k_1 \alpha \hat{x}_1 \tag{26a}$$

$$\mathbf{T}_{r2} = k_2 \beta \hat{y}_2 \, . \tag{26b}$$

where k_1 and k_2 are spring constants.

VII. MISCELLANEOUS TORQUES

Many other torques such as those due to interaction of the satellite's magnetic moments with the geomagnetic field, solar radiation, self-gravitation between two bodies, and plasma effects will act as forcing terms in the equations of motion. By proper design, these torques can be made small compared to the gravitational torque. However, since the gravitational torque varies inversely as the cube of the geocentric distance, it may not necessarily dominate in the orientation of satellites in very high orbits. Also, in very low orbits, aerodynamic drag may be big enough to upset the orientation. If long rods are used with weights on the ends, the gravitational torque can be made to dominate for a certain range in altitude.

VIII. EQUILIBRIUM AND STABILITY

Let us consider only the equilibrium position

$$(\xi,\eta,\zeta,\chi,\alpha,\beta) = (0,0,0,1,0,0)$$

in which the x_1, y_1, z_1 axes are lined up with the x,y,z axes. For viscous damping, the stability criteria for the position (0,0,0,1,0,0) can be found by linearizing the equations of motion about this position. The same stability criteria are obtained for equilibrium positions found by rotations of 180° around the x, y, and z axes, i.e., (1,0,0,0,0,0), (0,1,0,0,0,0), (0,0,1,0,0,0). For hysteresis damping, there will be an infinite number of stable equilibrium positions. All of these can, however, be made sufficiently close together to either one of the above four equilibrium positions, thus maintaining an axis in the satellite nearly in line with the local vertical.

From the definition of Euler parameters, the infinitesimal angles of rotation about the x_1, y_1, and z_1 axes are $\xi_1 = 2\xi$, $\eta_1 = 2\eta$, $\zeta_1 = 2\zeta$, defined as the roll, pitch and yaw angles. If ξ_2, η_2, ζ_2 are the infinitesimal angles that the principal axes of body 2 make with respect to the rotating coordinate system S_0-xyz, and α and β are small, then

$$\xi_2 = \alpha + \xi_1 \tag{27a}$$

$$\eta_2 = \beta + \eta_1 \tag{27b}$$

$$\zeta_2 = \zeta_1 \, . \tag{27c}$$

In the linearization process, we take the eccentricity of the orbit, ϵ' to be small in order to insure the realization of the infinitesimal angles˙ This is necessary in view of the well known result of the satellite pitch motion that the angular excursion produced by the eccentricity is of the same order of magnitude as the eccentricity itself. From (18) ψ becomes, with zero phase angle,

$$\psi = \Omega t + 2\epsilon S\Omega t + 0(\epsilon^2). \tag{28}$$

To linearize the general equations of motion given by (10), let us assume viscous damping as expressed in (24) and linear restoring torques as given in (26). The perturbing torques and forces, $\mathbf{T}_i'$ and $\mathbf{F}_i'$ $(i = 1, 2)$, are neglected. Also, let $\mathfrak{L}_1 = -\ell_1\hat{z}_1$ and $\mathfrak{L}_2 = \ell_2\hat{z}_2$. Then, equations (10) are linearized to the following:

$$\ddot{\eta}_1 + L_1\ddot{\eta}_2 + C_1'(\dot{\eta}_1 - \dot{\eta}_2) + d_1\eta_1 - k_1'\eta_2 = 2\epsilon\Omega^2(1 + L_1)S\Omega t \tag{29a}$$

$$\ddot{\eta}_2 + L_2\ddot{\eta}_1 + C_2'(\dot{\eta}_2 - \dot{\eta}_1) + d_2\eta_2 - k_2'\eta_1 = 2\epsilon\Omega^2(1 + L_2)S\Omega t \tag{29b}$$

$$\ddot{\xi}_1 + N_1\ddot{\xi}_2 + C_1''(\dot{\xi}_1 - \dot{\xi}_2) + q_1\Omega\dot{\zeta} + u_1\xi_1 - \bar{k}_1\xi_2 = 0 \tag{29c}$$

$$\ddot{\xi}_2 + N_2\ddot{\xi}_1 + C_2''(\dot{\xi}_2 - \dot{\xi}_1) + q_2\Omega\dot{\zeta} + u_2\xi_2 - \bar{k}_2\xi_1 = 0 \tag{29d}$$

$$\ddot{\zeta} + (1 - f_1 - f_2)\Omega^2\zeta - \Omega f_1\dot{\xi}_1 - \Omega f_2\dot{\xi}_2 = 0 \tag{29e}$$

where

$$L_1 = \bar{m}\ell_1\ell_2/I_2\,, \qquad L_2 = \bar{m}\ell_1\ell_2/I_5$$

$$C_1' = C_2/I_2\,, \qquad C_2' = C_2/I_5$$

$$k_1' = k_2/I_2\,, \qquad k_2' = k_2/I_5$$

$$d_1 = 3\Omega^2(I_1 - I_3)/I_2 + 3\Omega^2L_1 + k_1'$$

$$d_2 = 3\Omega^2(I_4 - I_6)/I_5 + 3\Omega^2L_2 + k_2'$$

$$N_1 = \bar{m}\ell_1\ell_2/I_1\,, \qquad N_2 = \bar{m}\ell_1\ell_2/I_4$$

$$C_1'' = C_1/I_1\,, \qquad C_2'' = C_1/I_4$$

$$q_1 = (I_1 + I_3 - I_2)/I_1$$

$$
\begin{aligned}
q_2 &= (I_4 + I_6 - I_5)/I_4 \\
u_1 &= 4\Omega^2(1 - q_1 + 3N_1/4) + k_1/I_1 \\
u_2 &= 4\Omega^2(1 - q_2 + 3N_2/4) + k_1/I_4 \\
\bar{k}_1 &= k_1/I_1 - \Omega^2 N_1\,, \qquad \bar{k}_2 = k_1/I_4 - \Omega^2 N_2 \\
f_1 &= (I_1 + I_3 - I_2)/(I_3 + I_6), \qquad f_2 = (I_4 + I_6 - I_5)/(I_3 + I_6).
\end{aligned}
$$

It should be noticed that the pitch equations (29a,b) do not depend on ξ_1, ξ_2, ζ and are decoupled from the roll and yaw equations (29c,d,e). The eccentricity enters as the amplitude of a forcing term in pitch but not in roll and yaw. The transient part of the pitch libration can be solved from (29a,b), excluding the forcing terms, by substituting with

$$\eta_i = B_i e^{st}, \qquad i = 1,2.$$

The resulting characteristic equation in s is then

$$
\begin{aligned}
(1 - L_1L_2)s^4 &+ (C_1' + C_2' + C_1'L_2 + C_2'L_1)s^3 \\
&+ (d_1 + d_2 + k_1'L_2 + k_2'L_1)s^2 \\
+ (d_1C_2' + d_2C_1' - C_1'k_2' &- C_2'k_1')s + (d_1d_2 - k_1'k_2') = 0. \qquad (30)
\end{aligned}
$$

The pitch motion is damped about (0,0,0,1,0,0) if and only if the Routh-Hurwitz conditions[3] are satisfied. This insures that the real parts of the roots of (30), representing the damping constants for the two principal modes, are negative. These give

$$I_x > I_z \qquad (31a)$$

$$k_2 > -3\Omega^2 \frac{(I_1 - I_3 + \bar{m}\ell_1\ell_2)}{I_x - I_z}(I_4 - I_6 + \bar{m}\ell_1\ell_2) \qquad (31b)$$

$$\frac{I_1 - I_3 + \bar{m}\ell_1\ell_2}{I_2 + \bar{m}\ell_1\ell_2} \neq \frac{I_4 - I_6 + \bar{m}\ell_1\ell_2}{I_5 + \bar{m}\ell_1\ell_2} \qquad (31c)$$

where

$$
\begin{aligned}
I_x &= I_1 + I_4 + 2\bar{m}\ell_1\ell_2 \\
I_y &= I_2 + I_5 + 2\bar{m}\ell_1\ell_2 \\
I_z &= I_3 + I_6\,.
\end{aligned}
$$

I_x, I_y, I_z represent the moments of inertia of the composite body about S_0. Condition (31a) is the same as that of a single rigid body. Condition (31b) states that k_2 must be larger than a certain critical value if one body is unstable (e.g., $I_4 - I_6 + \bar{m}\ell_1\ell_2 < 0$). This value is zero if both bodies are stable by themselves. It can be shown that there cannot exist a cocked equilibrium position in pitch if the parameters are such as to make the position (0,0,0,1,0,0) stable. Condition (31c) implies that there exists an undamped motion if the equality sign holds. This rigid body motion has a frequency ω_r, given by

$$\omega_r^{\,2} = 3\Omega^2 \frac{(I_1 - I_3 + \bar{m}\ell_1\ell_2)}{I_2 + \bar{m}\ell_1\ell_2}\,. \tag{32}$$

The roll and yaw equations (29c,d,e) are all coupled. This justifies the use of a damper only for roll. Up to first-order terms, there are no forcing terms due to eccentricity. The characteristic equation is

$$b_6 s^6 + b_5 s^5 + b_4 s^4 + b_3 s^3 + b_2 s^2 + b_1 s + b_0 = 0 \tag{33}$$

where

$$\begin{aligned}
b_0 &= \Omega^2(1 - f_1 - f_2)\,(u_1 u_2 - \bar{k}_1\bar{k}_2)\\
b_1 &= \Omega^2(1 - f_1 - f_2)\,(C_2'' u_1 + C_1'' u_2 - \bar{k}_1 C_2'' - \bar{k}_2 C_1'')\\
b_2 &= \Omega^2(1 - f_1 - f_2)\,(u_1 + u_2 + N_1\bar{k}_2 + N_2\bar{k}_1) + u_1 u_2 - \bar{k}_1\bar{k}_2\\
&\qquad + \Omega^2(f_1 q_2 \bar{k}_1 + f_2 q_1 \bar{k}_2 + f_1 q_1 u_2 + f_2 q_2 u_1)\\
b_3 &= \Omega^2(1 - f_1 - f_2)\,(C_1'' + C_2'' + C_2'' N_1 + C_1'' N_2)\\
&\qquad + C_1''(u_2 + \Omega^2 f_1 q_2 + \Omega^2 f_2 q_2 - \bar{k}_2)\\
&\qquad + C_2''(u_1 + \Omega^2 f_2 q_1 + \Omega^2 f_1 q_1 - \bar{k}_1)\\
b_4 &= \Omega^2(1 - f_1 - f_2)\,(1 - N_1 N_2) + u_1 + u_2 + N_1\bar{k}_2 + N_2\bar{k}_1\\
&\qquad + \Omega^2(f_1 q_1 + f_2 q_2 - f_1 q_2 N_1 - f_2 q_1 N_2)\\
b_5 &= C_1''(1 + N_2) + C_2''(1 + N_1)\\
b_6 &= 1 - N_1 N_2\,.
\end{aligned}$$

The Routh-Hurwitz stability criteria are

$$b_0 > 0, \qquad b_1 > 0, \qquad \begin{vmatrix} b_1 & b_0 \\ b_3 & b_2 \end{vmatrix} > 0,$$

$$\begin{vmatrix} b_1 & b_0 & 0 \\ b_3 & b_2 & b_1 \\ b_5 & b_4 & b_3 \end{vmatrix} > 0, \qquad \begin{vmatrix} b_1 & b_0 & 0 & 0 \\ b_3 & b_2 & b_1 & b_0 \\ b_5 & b_4 & b_3 & b_2 \\ 0 & b_6 & b_5 & b_4 \end{vmatrix} > 0,$$

$$\begin{vmatrix} b_1 & b_0 & 0 & 0 & 0 \\ b_3 & b_2 & b_1 & b_0 & 0 \\ b_5 & b_4 & b_3 & b_2 & b_1 \\ 0 & b_6 & b_5 & b_4 & b_3 \\ 0 & 0 & 0 & b_6 & b_5 \end{vmatrix} > 0, \qquad b_6 > 0.$$

If these are satisfied, there will be three more modes of damped librations. Due to the coupling between the roll and yaw librations, the yaw libration can be damped out by the roll damping, as can be observed from (29c,d,e), although no yaw damping mechanism is provided in the present scheme. Hence, all modes can be damped out and the satellite will oscillate with some steady-state amplitude about an equilibrium

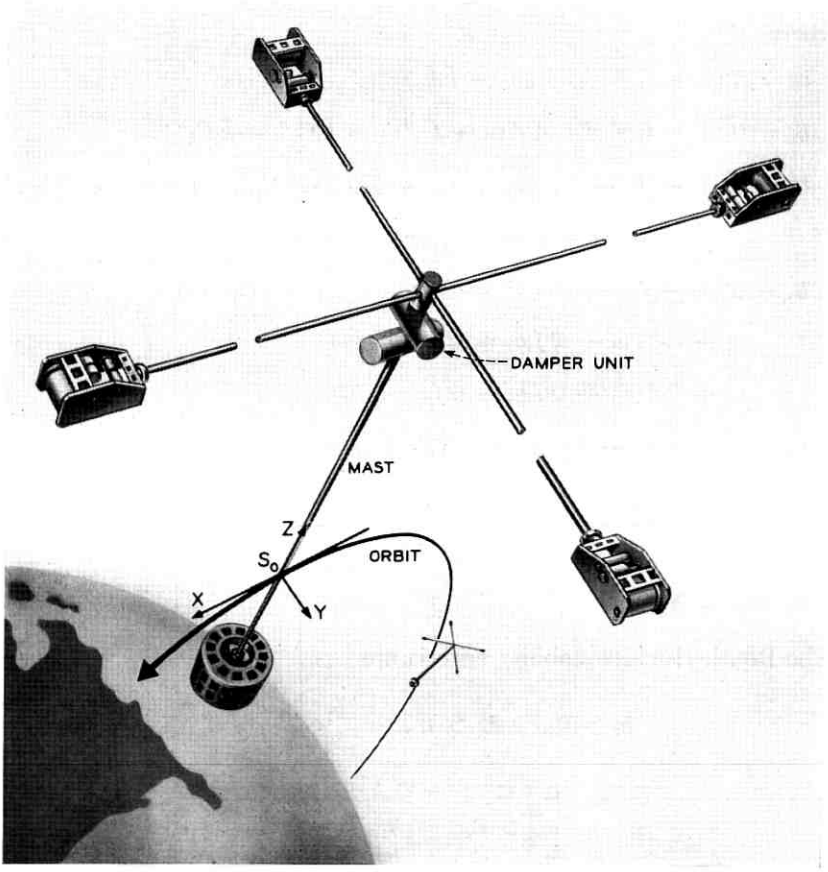


Fig. 4 — Gravitationally oriented two-body satellite with extensible rods.

position. Some of these conditions are too complicated to give any physical insight. However, some are quite simple and are given below.

Since the parts of b_0, b_2 and b_4 which involve k_1 are $(k_1/C_1)b_1$, $(k_1/C_1)b_3$, $(k_1/C_1)b_5$ respectively, multiplying the odd columns of the Hurwitz determinants by k_1/C_1 and adding to adjacent columns will eliminate the k_1 terms. Hence the only condition on k_1 is $b_0 > 0$, i.e.,

$$k_1 > \Omega^2 \frac{[4(I_2 - I_3 + \frac{3}{4}\,\bar{m}\ell_1\ell_2)(I_5 - I_6 + \frac{3}{4}\,\bar{m}\ell_1\ell_2) - \frac{1}{4}\,\bar{m}\ell_1^{\,2}\ell_2^{\,2}]}{I_z - I_y} \tag{34}$$

As k_1 and k_2 approach infinity, the satellite becomes one rigid body. Since the stability conditions are not changed by an increase of k_1 (and k_2),

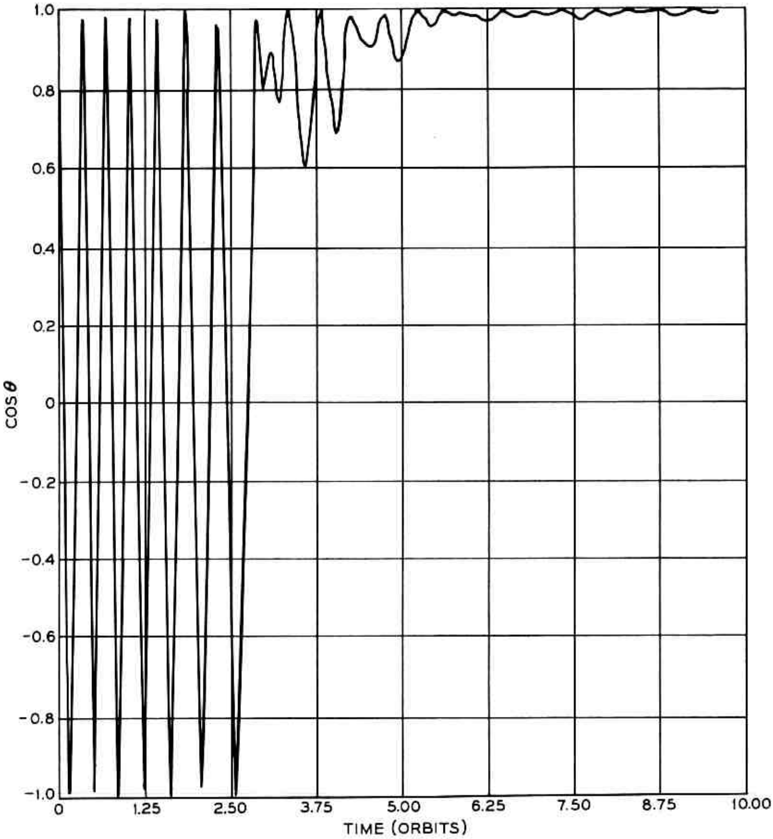


Fig. 5 — Angular variation between the z_1-axis of the satellite and the local vertical for a hysteresis damper, cos θ.

it appears that the single rigid body criteria for roll and yaw stability are necessary. These are

$$P_x P_z > 0 \tag{35a}$$

$$1 + 3P_x + P_x P_z > 4\sqrt{P_x P_z} \tag{35b}$$

where

$$P_x = \frac{I_y - I_z}{I_x}, \qquad P_z = \frac{I_y - I_x}{I_z}$$

and I_x, I_y, and I_z are given in (31). Condition (35a) can be verified from the inequality $b_1 > 0$. Other necessary conditions in the case of $\ell_2 = 0$ are found from the third-order Hurwitz determinant to be

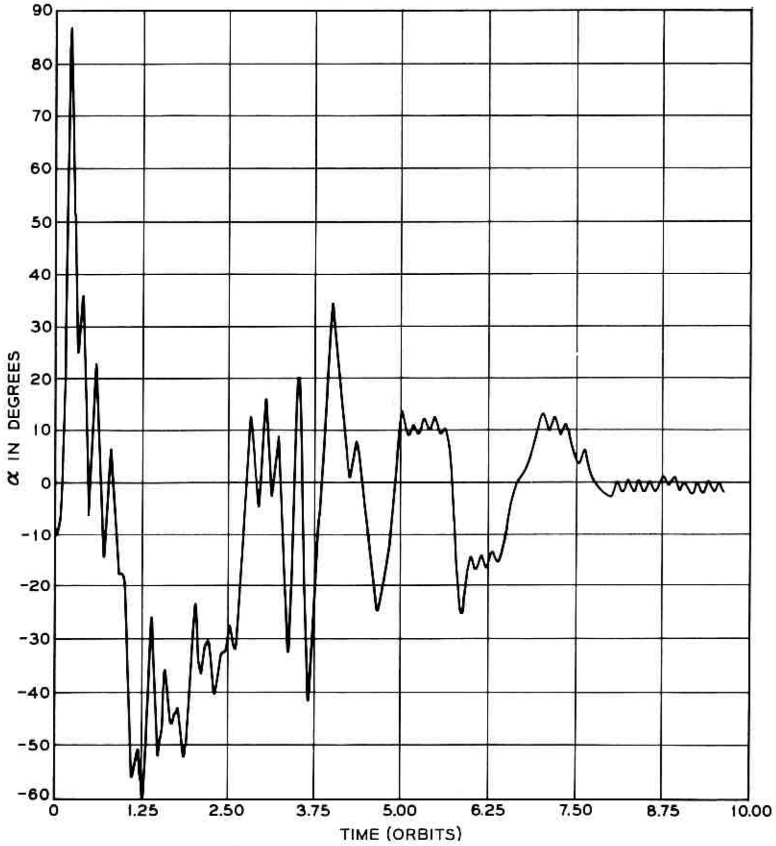


Fig. 6 — Relative angle about the x_1-journal for a hysteresis damper, α.

$$(I_y - I_x)\,(I_y - \tfrac{3}{4}I_z) > 0 \tag{35c}$$

and

$$\frac{I_5 - I_6}{I_4} \neq \frac{I_2 - I_3}{I_1}. \tag{35d}$$

IX. BISTABILITY

The satellite is in a stable equilibrium position if the z_1-axis is in line with the local vertical (i.e., the z-axis) pointing in either direction. If a directional device such as an antenna or a camera is used along the negative z_1-axis, it may point at or away from the earth. The equi-

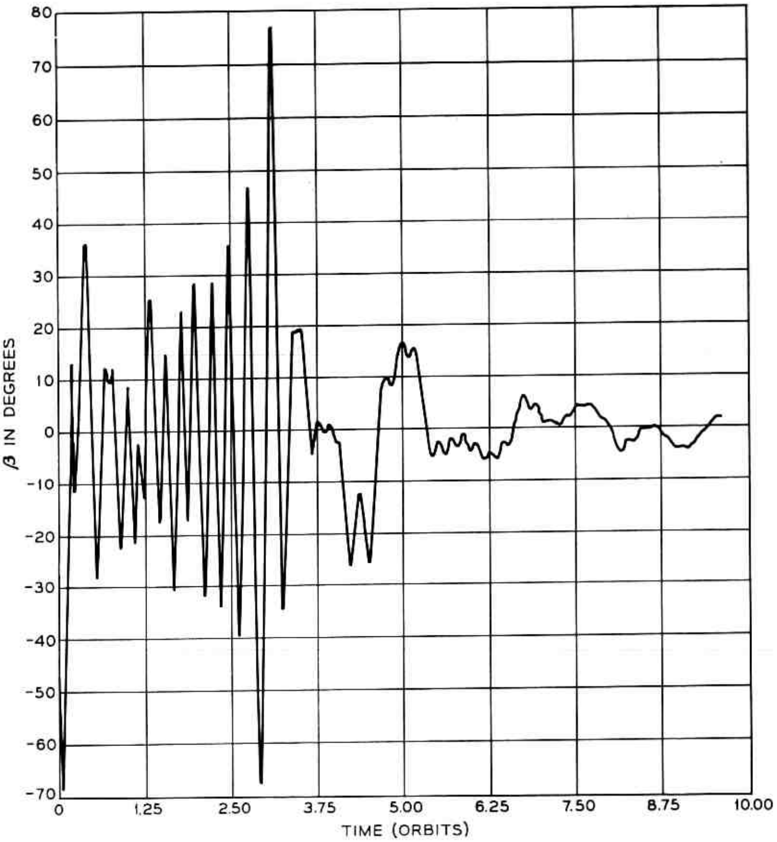


Fig. 7— Relative angle about the y_2-journal for a hysteresis damper, β.

librium positions (0,0,0,1,0,0) and (0,0,1,0,0,0) correspond to the device pointing toward the earth, whereas (1,0,0,0,0,0) and (0,1,0,0,0,0) correspond to the device pointing away from the earth. In the latter case an inertia wheel in the satellite can be activated with a predetermined number of turns, and the satellite can be rotated 180 degrees so that the device will be earth-pointing. The equations governing this turning are given by (10), where the applied torque on body 1 is approximately

$$\mathbf{T}_1' = -\frac{d}{dt}[J_m(C\delta\hat{x}_1 + S\delta\hat{y}_1)] \tag{36}$$

where J_m is the angular momentum of the inertia wheel and δ is the angle between the x_1-axis and the axis of the inertia wheel. Another scheme

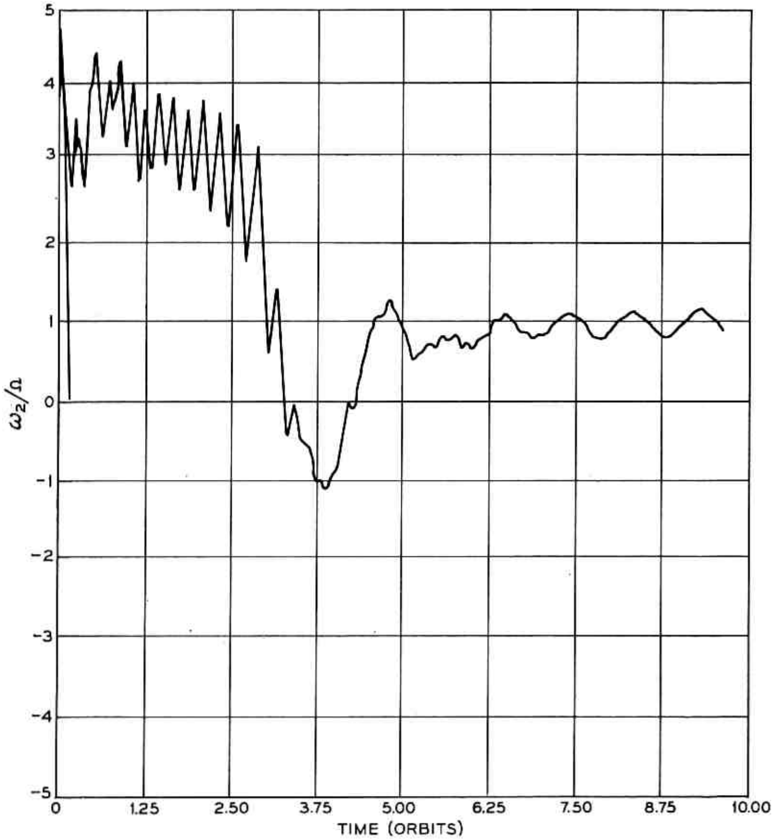


Fig. 8 — Component of angular velocity of the satellite along the y_1-axis for a hysteresis damper, ω_2/Ω.

would be to use two devices, one on each side of the satellite, directed along the positive and the negative direction of the z_1-axis respectively. Only the one that is earth-pointing would be activated.

X. NUMERICAL RESULTS OF A PRACTICAL SCHEME

A practical scheme, as shown in Fig. 4, is suggested here for a communications satellite. The particular construction, employing extensible rods and tip masses, is to effect large moments of inertia so that the gravitational torque will dominate over all disturbing torques. Body 1 of the satellite, which consists of the satellite's main structure (with directional antennas) and a mast rod, is to be earth-pointing. Body 2, being an auxiliary body for attitude-control purpose only, is constructed of two rods and is in an unstable position with respect to the local vertical. These rods are extended, upon ejection from the launching vehicle's

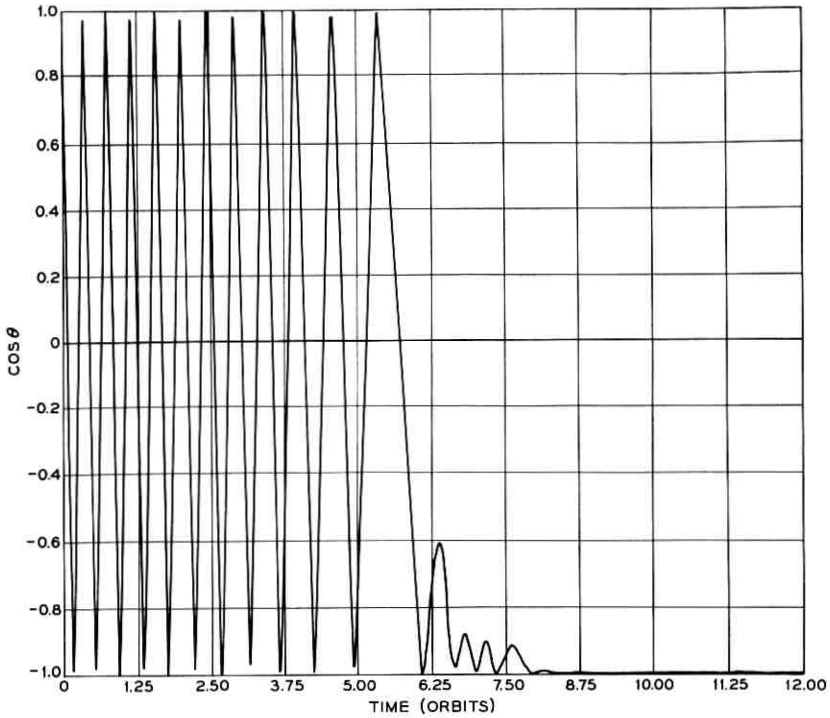


Fig. 9 — Angular variation between the z_1-axis of the satellite and the local vertical for a viscous damper, cos θ.

final stage, by unrolling from sheet metal drums. The universal joint employs torsion wires to produce elastic restoring torques and provides hysteresis damping by relative displacement between magnets and a permeable material. (See Fig. 5 of companion paper.[4])

The advantages of magnetic hysteresis damping are that it is amplitude dependent, insensitive to temperature variation, involves no sliding parts and requires little weight. Coulomb friction damping, while also amplitude dependent, is less desirable because of possible cold welding of sliding parts in the high vacuum of space. Velocity-dependent damping by employing viscous fluids is believed to provide lower damping for a given weight, and the viscous fluids involve questions of temperature sensitivity.

All the parameters are chosen based on the adjusted moment of inertia, I_1, of body 1 subject to stability criteria and other necessary considerations. The stability criteria (31b) and (34) specifying the critical values

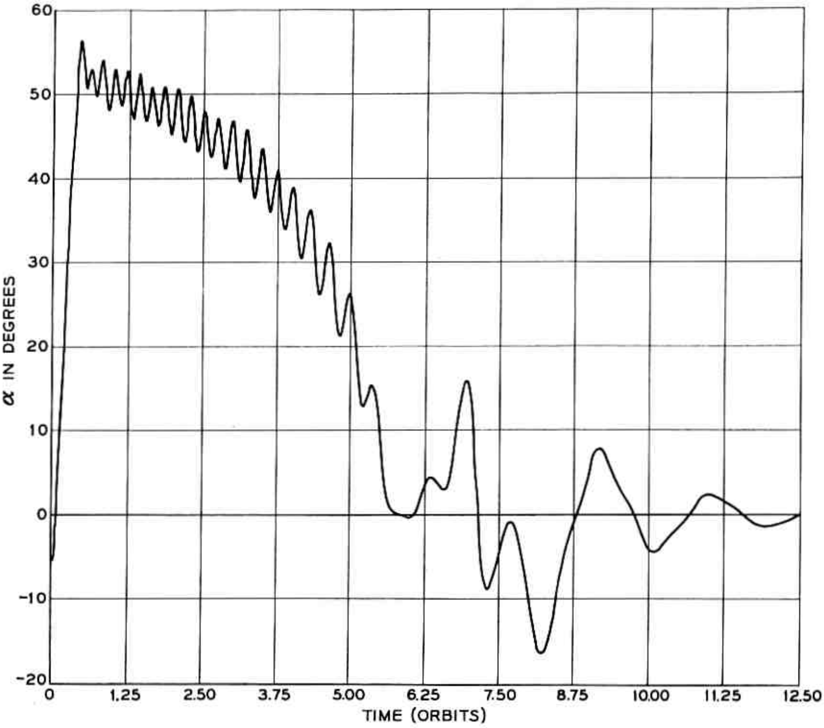


Fig. 10 — Relative angle about the x_1-journal for a viscous damper, α.

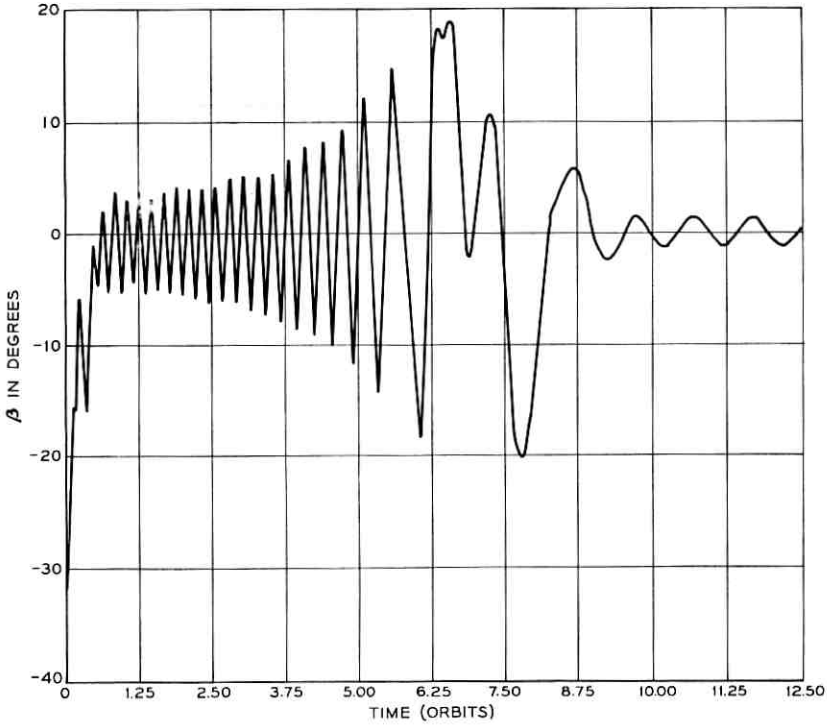


Fig. 11 — Relative angle about the y_2-journal for a viscous damper, β.

of k_2 and k_1, respectively, which are derived from viscous damping, are found to apply approximately also in the case of hysteresis damping. These parameters are: $I_i/I_1 = 1.00, 0.003, 0.159, 0.381, 0.540$ $(i = 2, \cdots, 6)$; $k_i/I_1\Omega^2 = 1.131, 2.238$ $(i = 1,2)$; for a hysteresis damper: $\bar{T}_{di}/I_1\Omega^2 = 0.159, 0.216$ $(i = 1,2)$, $\bar{\alpha} = \bar{\beta} = 2°$; for a viscous damper: $C_i/I_1\Omega = 0.870, 1.281$ $(i = 1,2)$. With the above value of the viscous constant C_2, the amplitude of the lower mode of pitch libration can be reduced according to (30) by a factor of e in 0.22 orbit, which is close to the optimum. The optimum in the case of pitch motion was found by Zajac[5] to be 0.137 orbit. Equations of motion (20)† depend only on the above dimensionless parameters and are independent of I_1 and Ω as long as t is measured in fractions of an orbital period. Some initial conditions which might simulate a micrometeoroid impact or the

† Equations (20) were programmed on an IBM 7090 by Mrs. W. L. Mammel.

motion after the erection of the rods are at $t = 0$: $\xi = \eta = \zeta = \alpha = \beta = 0$, $\chi = 1$, $\omega_1 = \Omega$, $\omega_2 = 5\Omega$, $\omega_5 = \Omega$, $\omega_3 = \omega_4 = \omega_6 = 0$. Figs. 5–8 represent the computer solution of equations (20) using a magnetic hysteresis damper. In Fig. 5, θ is the angle between the z_1-axis and the local vertical. The satellite stops tumbling after four orbits and settles to within 10° of the local vertical after six orbits. The satellite librates about a cocked equilibrium position indefinitely due to the forcing torque of orbital eccentricity ($\epsilon = 0.01$). The pitch angular speed of body 1, ω_2, approaches one revolution per orbit, which is the proper speed for an earth-pointing satellite. Figs. 9–12 show similar results of a viscous damper. In this case the satellite ended up in an inverted position.

Effects of the environmental disturbing torques, such as those due to solar radiation and the interaction of the magnetic moment in the satellite with the geomagnetic field, have been investigated, although the results are not included here. Cases with various other initial conditions

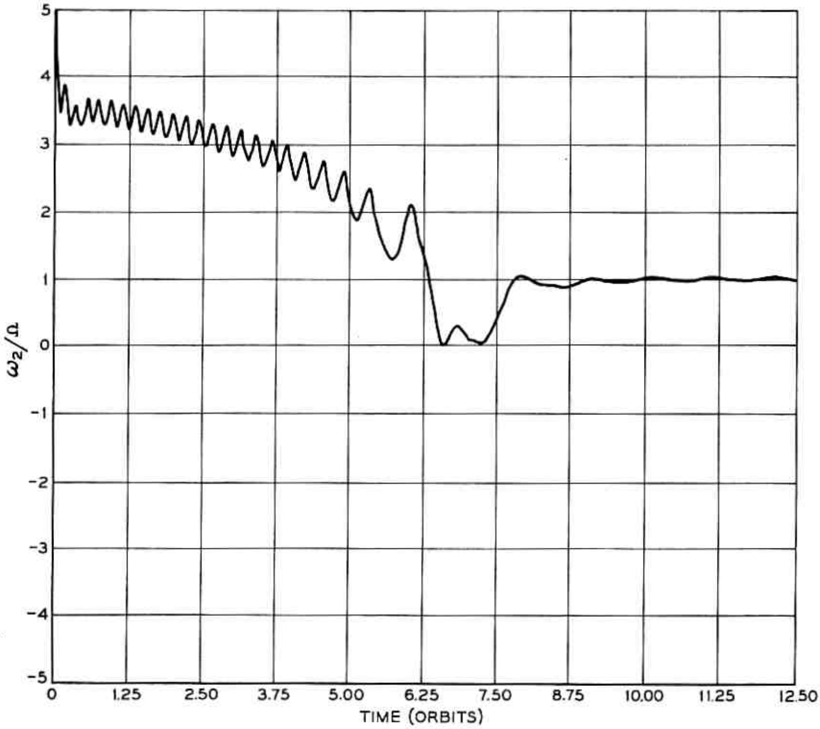


Fig. 12 — Component of angular velocity of the satellite along the y_1-axis for a viscous damper, ω_2/Ω.

have also been computed. All these results indicate that gravitational orientation of a two-body satellite is feasible.

APPENDIX

Nomenclature

A.1 *Latin Symbols*

a_{ij} = direction cosines of S_1-$x_1y_1z_1$ frame with respect to S_0-xyz frame $(i,j = 1,2,3)$
b_{ij} = direction cosines of S_2-$x_2y_2z_2$ frame with respect to S_1-$x_1y_1z_1$ frame $(i,j = 1,2,3)$
b_i = coefficients of characteristic equation of ξ_1, ξ_2, ζ $(i = 0, 1, \cdots, 6)$
B_i = complex constant of η_i $(i = 1,2)$
C = cosine operator
C_i = viscous damping constants of α,β $(i = 1,2)$
C_i', C_i'' = adjusted damping constants of α,β defined in equations (29) $(i = 1,2)$
d_i = coefficients defined in equations (29) $(i = 1,2)$
f_i = moment of inertia coefficients defined in equations (29) $(i = 1,2)$
$\mathbf{F}_H$ = force on body 1 due to reaction of hinge
$\mathbf{F}_i$ = resultant force on body i exclusive of $\mathbf{F}_H$ $(i = 1,2)$
$\mathbf{F}_i'$ = resultant force on body i exclusive of gravity and $\mathbf{F}_H$ $(i = 1,2)$
g = acceleration of gravity on the earth's surface
G = quantity defined in equation (19)
$\mathbf{G}_i$ = gravitational force on body i $(i = 1,2)$
H = hinge point
$\mathbf{I}$ = unit dyadic
I_i = adjusted moments of inertia $(i = 1, \cdots, 6)$
I_x, I_y, I_z = moments of inertia of composite body about the common center of mass
J_m = angular momentum of inertia wheel
k_i = spring constants producing torques in x_1, y_2 directions $(i = 1,2)$
$k_i',\bar{k}_i$ = adjusted spring constants defined in equations (29) $(i = 1,2)$
l = maximum linear dimension of the satellite

$\mathfrak{L}_i$ = position vector of center of mass of body i from hinge ($i = 1,2$)

ℓ_i = magnitude of $\mathfrak{L}_i$ ($i = 1,2$)

L_i = coefficients defined in equations (29) ($i = 1,2$)

m_i = mass of body i ($i = 1,2$)

m = total mass of satellite

$\bar{m}$ = reduced mass

n_i = direction cosines of z-axis on S_1-$x_1y_1z_1$ and S_2-$x_2y_2z_2$ frames ($i = 1, \cdots, 6$)

N_i = coefficients defined in equations (29) ($i = 1,2$)

O = center of the earth

P_i = arbitrary point in body i ($i = 1,2$)

P_x, P_y, P_z = ratio of moments of inertia in equations (35)

q_i = coefficients defined in equations (29) ($i = 1,2$)

$\mathbf{r}_i$ = position vector of P_i from center of mass of body i ($i = 1,2$)

$\mathbf{R}_i$ = position vector of P_i from O ($i = 1,2$)

R_E = mean radius of the earth

S = sine operator

s = variable in characteristic equations

S_0 = center of mass of satellite

S_i = center of mass of body i ($i = 1,2$)

t = time variable

$\mathbf{T}_H$ = reaction torque transmitted through the joint on body 1

$\mathbf{T}_i$ = resultant torque on body i exclusive of $\mathbf{T}_H$ ($i = 1,2$)

$\mathbf{T}_i'$ = resultant torque on body i exclusive of $\mathbf{T}_H$ and gravitational torque ($i = 1,2$)

$\mathbf{T}_{Gi}$ = gravitational torque on body i ($i = 1,2$)

$\mathbf{T}_c$ = constraint torque of joint on body 1

$\mathbf{T}_d$ = dissipative torque of joint on body 1

$\bar{\mathbf{T}}_{di}$ = magnitude of saturated hysteresis torque of magnet i ($i = 1,2$)

T_{di}^* = value of T_{di} when $\dot{\alpha}$ ($i = 1$) and $\dot{\beta}$ ($i = 2$) last changed sign

$\mathbf{T}_r$ = elastic restoring torque of joint on body 1

u_i = coefficients defined in equations (29) ($i = 1,2$)

X,Y,Z = fixed frame coordinates

x,y,z = rotating frame coordinates

x_1, y_1, z_1 = body 1 coordinates

x_2, y_2, z_2 = body 2 coordinates.

A.2 *Greek Symbols*

α = relative angle of rotation of body 2 about x_1-axis

$\bar{\alpha}$ = constant of magnet 1

α^* = values of α when $\dot{\alpha}$ last changed sign

β = relative angle of rotation of body 2 about y_2-axis

$\bar{\beta}$ = constant of magnet 2

β^* = values of β when $\dot{\beta}$ last changed sign

δ = angle between x_1-axis and the inertia wheel axis

ϵ = eccentricity of the orbit

ζ = Euler parameter

ζ_i = infinitesimal angle about z_i-axis ($i = 1,2$)

η = Euler parameter

η_i = infinitesimal angle about y_i-axis ($i = 1,2$)

θ = angle between z_1-axis and the local vertical or z-axis

λ_i = components of the relative angular velocity of body 1 with respect to rotating frame ($i = 1,2,3$)

μ = a gravitational constant of the earth

ξ = Euler parameter

ξ_i = infinitesimal angle about x_i-axis ($i = 1,2$)

$\boldsymbol{\varrho}$ = position vector of S_0 from O

$\boldsymbol{\varrho}_i$ = position vector of S_i from O ($i = 1,2$)

$\boldsymbol{\Phi}_i$ = moment of inertia dyadic of body i ($i = 1,2$)

$\boldsymbol{\Phi}_i'$ = quasi moment of inertia dyadic of body i ($i = 1,2$)

χ = Euler parameter

ψ = true anomaly of ellipse

Ω = mean orbital angular speed of satellite

$\boldsymbol{\omega}_{\mathrm{I}}$, $\boldsymbol{\omega}_{\mathrm{II}}$ = angular velocity of body 1,2

ω_1, ω_2, ω_3 = components of $\boldsymbol{\omega}_{\mathrm{I}}$ along x_1, y_1, z_1 axes

ω_4, ω_5, ω_6 = components of $\boldsymbol{\omega}_{\mathrm{II}}$ along x_2, y_2, z_2 axes

ω_r = natural frequency of an undamped roll libration.

A.3 *Notes*

$\hat{}$ = unit vector

$\cdot$ = time derivative in an inertial frame $\left(= \dfrac{d}{dt}\right)$

boldface characters indicate tensors and vectors (it is assumed that dropping the boldface means the magnitude of the vector; i.e., $\rho = |\boldsymbol{\varrho}|$).

REFERENCES

1. Kamm, L. J., "Vertistat", An Improved Satellite Orientation Device, A.R.S. Journal, **32,** No. 6, June, 1962, pp. 911–913.
2. Whittaker, E. T., *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, Dover Publications, New York, 1944, p. 8 and p. 16.
3. Cesari, L., *Asymptotic Behavior and Stability Problems in Ordinary Differential Equations*, Springer-Verlag, Berlin, 1959, p. 21 and p. 34.
4. Paul, B., West, J. W., and Yu, E. Y., A Passive Gravitational Attitude Control System for Satellites, B.S.T.J., this issue, pp. 2195–2238.
5. Zajac, E. E., Damping of a Gravitationally Oriented Two-Body Satellite, A.R.S. Journal, **32,** No. 12, December, 1962, pp. 1871–1875.

Innage and Outage Intervals in Transmission Systems Composed of Links

By S. O. RICE

(Manuscript received April 3, 1963)

This note is of the nature of an addendum to a recent paper on satellite communication systems. It is concerned with the distribution and average durations of innages and outages occurring in transmission systems composed of a number of links. The links of such a composite system may be either in series, as in a radio relay system, or in parallel, as in a many-satellite system. Several results regarding composite transmission systems, including some due to D. S. Palmer, are reviewed, restated, and extended.

I. INTRODUCTION

This note is in the nature of an addendum to a recent paper of mine on satellite communication systems.[1] It is concerned with the same general problem, namely the reliability of transmission systems composed of links which fail independently. Various published results are reviewed and extended. A large part of these results is due to D. S. Palmer,[2] whose excellent work was overlooked in my satellite paper. The approach given here differs somewhat from that used by Palmer.

Incidentally, questions similar to those discussed here have also appeared in connection with coincidences in counting devices.

The notation to be followed is illustrated in Fig. 1. Suppose that a link in a transmission system is always either in one or the other of two possible states, state (a) or state (b). For example, if the link is a satellite, (a) may be taken as the state of being out of sight and (b) the state of being visible. Again, if the link is one of a series of links in tandem making up a transmission line, we may choose (a) to be the state of working order and (b) the state of breakdown. In a satellite system the links are in parallel and in the transmission line they are in series. Fig. 1 applies to both cases.

The light portions of the top line in Fig. 1 represent the intervals dur-

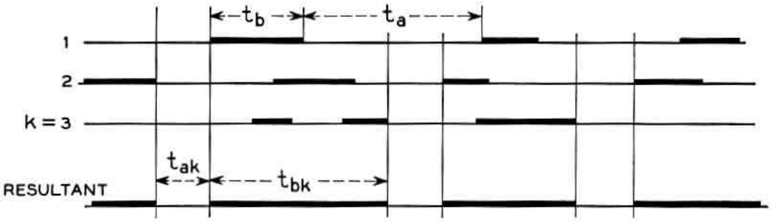


Fig. 1 — Combination of k independent alternating sequences to form the resultant alternating sequence.

ing which Link No. 1 is in state (a), and the heavy portions the intervals of state (b). Similarly, the second and third lines represent the state intervals of Links No. 2 and No. 3. This is a k-link system with $k = 3$. The last line represents the system state intervals. In system state (ak) all k links are in state (a). In state (bk) at least one link is in state (b). State (ak) corresponds to the "intersection" of type (a) intervals and state (bk) to the "union" of type (b) intervals.

For the satellite system, states (ak) and (bk) correspond to "outage" and "innage," respectively. For the transmission line they correspond to "working order" and "breakdown." This reversal of interpretation for links in parallel and for links in series has been mentioned by Palmer.[2]

The problem is to find the distributions of the durations t_{ak} and t_{bk} of states (ak) and (bk). The lengths t_a , t_b of the intervals shown in Fig. 1 are supposed to be independent random variables with given probability densities $p_a(t)$, $p_b(t)$. Usually $p_a(t)$, $p_b(t)$ will be the same for all k links, but in the more general case the densities associated with the ith link will be denoted by $p_a^{(i)}(t)$, $p_b^{(i)}(t)$. The links are assumed to operate independently of each other. It is also assumed that the system has been operating long enough to reach statistical equilibrium.

Although t_a and t_b are independent, t_{ak} and t_{bk} need not be. An example is given just below equation (25) in Section IV.

The results given here do not apply to the case where the pattern of intervals in two or more links shows periodicities. For example, if all type (a) intervals of Links No. 1 and No. 2 are of length 1 and all type (b) intervals are of length 3, then (depending on the relative phase) there may be no type (ak) intervals and just one infinitely long type (bk) interval.

The distribution of t_{ak} depends only on $p_a(t)$. It is obtained in Section II for general $p_a(t)$. The expected value $\bar{t}_{bk}$ of t_{bk} depends only on the expected values $\bar{t}_a$, $\bar{t}_b$ and is given in Section III. At present there seems to be no practicable method, other than simulation on a high-speed computer, of obtaining the distribution of t_{bk} for general $p_a(t)$, $p_b(t)$. For exponential $p_a(t)$ a method due to Palmer[2] and Takács (outlined in Ref. 1) may be used, but even this is difficult unless $p_b(t)$ is also exponential. This method is developed in Section IV and illustrated in Section V. Sections VI and VII are concerned with the special case $k = 2$ but general $p_a(t)$, $p_b(t)$. Now the determination of the distribution of t_{b2} depends upon the solution of an integral equation. A vexing problem which I have been unable to solve is to show that when $p_a(t)$ is exponential the integral equation leads to the same distribution as does setting $k = 2$ in the method of Section IV.

I am indebted to John Riordan, David Slepian, and Lajos Takács for helpful comments.

II. THE DISTRIBUTION OF t_{ak}

It is convenient to set

$$F_a(t) = \int_t^\infty p_a(\tau)\, d\tau \tag{1}$$

$$A_a(t) = \int_t^\infty F_a(\tau)\, d\tau/\bar{t}_a \tag{2}$$

$$\bar{t}_a = \int_0^\infty \tau p_a(\tau)\, d\tau = \int_0^\infty F_a(\tau)\, dt. \tag{3}$$

Here $F_a(t)$ is the probability that $t_a > t$, $\bar{t}_a$ is the expected value of t_a, and $A_a(t)$ is closely related to C. Palm's[3] "next-arrival" distribution. If $\alpha(s)$ is the Laplace transform of $p_a(t)$, i.e.

$$\alpha(s) = \int_0^\infty e^{-st} p_a(t)\, dt \tag{4}$$

then

$$\begin{aligned} \int_0^\infty e^{-st} F_a(t)\, dt &= \frac{1 - \alpha(s)}{s} \\ \int_0^\infty e^{-st} A_a(t)\, dt &= \frac{1}{s}\left[1 - \frac{1 - \alpha(s)}{s\bar{t}_a}\right]. \end{aligned} \tag{5}$$

For the special case $p_a(t) = ae^{-at}$

$$\begin{aligned} F_a(t) &= A_a(t) = e^{-at}, \qquad \bar{t}_a = 1/a \\ \alpha(s) &= a/(a + s). \end{aligned} \tag{6}$$

To interpret $A_a(t)$, consider Fig. 2, which shows a line in Fig. 1 corresponding to a typical link. Choose a point $t = x$ at random [this means that when the choice is from the very long interval $(0,T)$ the chance that x falls between t, $t + dt$ is dt/T]. Let l be the distance to the end of the interval (which may be of either type) in which x falls. Then[2,4,5] $A_a(\tau)$ is the probability that $l > \tau$, given that x fell in an (a) interval.

It should be noted that expression (2) for $A_a(t)$ holds even when successive t_a's and t_b's are correlated. This point is important in the proof of (7), since the intervals t_{ak}, t_{bk} may be correlated. The only requirement is that as $T \to \infty$ the distribution of the lengths of the (a) intervals in $(0,T)$ approaches a definite distribution $F_a(t)$ possessing an average t_a which is neither zero nor infinite. For emphasis we sketch a proof of (2) which is tailored to Fig. 2, in which, for the moment, t_a and t_b may be correlated. The chance that $\tau < l < \tau + d\tau$ is the limit as $T \to \infty$ of the ratio

$$\frac{[\text{number of } (a) \text{ intervals longer than } \tau \text{ in } (0,T)](d\tau)}{\text{total length of } (a) \text{ intervals in } (0,T)}.$$

In the limit this ratio approaches $NF_a(\tau)d\tau/N\bar{t}_a$, where N is the number of (a) intervals in $(0,T)$. Cancelling the N's and integrating τ from t to ∞ then gives (2).

To find the probability $F_{ak}(t)$ that $t_{ak} > t$, suppose that all links are in state (a) at the randomly chosen time x. Since the links are independent, the chance that none has changed to state (b) by time $x + t$ is $[A_a(t)]^k$. Hence the function $A_{ak}(t)$ corresponding to the complete system is $[A_a(t)]^k$. This $A_{ak}(t)$ is related to $F_{ak}(t)$ by an equation obtained from (2) by replacing the subscripts "a" by "ak." Differentiation gives

$$\begin{aligned} F_{ak}(t) &= -\bar{t}_{ak} \frac{d}{dt} [A_a(t)]^k \\ &= (\bar{t}_{ak}/\bar{t}_a) k [A_a(t)]^{k-1} F_a(t) \end{aligned} \qquad (7)$$

where $\bar{t}_{ak}$ denotes the expected length of intervals of type (ak).

Setting $t = 0$ in (7) and using $F_{ak}(0) = A_a(0) = F_a(0) = 1$ leads to

$$\bar{t}_{ak} = \bar{t}_a/k. \qquad (8)$$

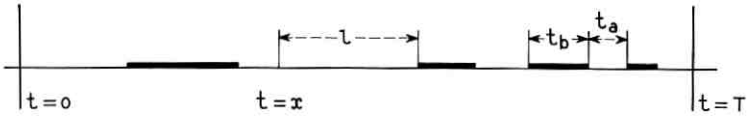


Fig. 2 — $A_a(t)$ is the chance that $l > t$.

When the individual links have probability densities $p_a^{(i)}(t)$, $p_b^{(i)}(t)$, $i = 1, 2, \cdots, k$, the chance that the length of a type (ak) interval exceeds t is

$$F_{ak}(t) = -\bar{t}_{ak} \frac{d}{dt} \prod_{i=1}^{k} A_a^{(i)}(t) \tag{9}$$

which implies

$$(\bar{t}_{ak})^{-1} = \sum_{i=1}^{k} (\overline{t_a^{(i)}})^{-1} \tag{10}$$

just as (7) implies (8). The A_a's and $\bar{t}_a$'s are related by equations corresponding to (1), (2) and (3). These results are due to Palmer,[2] who obtains (7) by a different argument.

III. THE EXPECTED LENGTH OF INTERVALS OF VARIOUS TYPES, INCLUDING TYPE (bk)

The expected value $\bar{t}_{bk}$ of the length of the type (bk) intervals is related to $\bar{t}_{ak}$ by the equation

$$\bar{t}_{bk} = \left(\frac{1}{p_{ak}} - 1\right) \bar{t}_{ak} \tag{11}$$

where p_{ak} [not to be confused with the probability density $p_{ak}(t)$] is the chance that the random point x will fall in an interval of type (ak). This follows almost immediately from

$$\bar{t}_{bk}/\bar{t}_{ak} = p_{bk}/p_{ak} , \tag{12}$$

$$p_{ak} + p_{bk} = 1. \tag{13}$$

A careful discussion of the probability p_{ak} has been given by Weiss.[6]

A relation similar to (11) also holds for $\bar{t}_b$, $\bar{t}_a$, and the probability p_a that the random point x will fall in a type (a) interval. Solving for p_a gives

$$p_a = \bar{t}_a(\bar{t}_a + \bar{t}_b)^{-1} \tag{14}$$

and when this is combined with $p_{ak} = p_a^k$, which follows from the independence of the links, (11) becomes

$$\bar{t}_{bk} = (p_a^{-k} - 1)\bar{t}_{ak} = \left[\left(1 + \frac{\bar{t}_b}{\bar{t}_a}\right)^k - 1\right] \frac{\bar{t}_a}{k}. \tag{15}$$

Palmer's generalization of (15) can be written as

$$
\begin{aligned}
\bar{t}_{bk} &= [(\prod p_a^{(i)})^{-1} - 1]\bar{t}_{ak} \\
p_a^{(i)} &= \bar{t}_a^{(i)}(\bar{t}_a^{(i)} + \bar{t}_a^{(i)})^{-1}
\end{aligned}
\tag{16}
$$

where $\bar{t}_{ak}$ is given by (10).

Einhorn[7] has given the instances $k = 2$ of (10) and (16). He also gives a generalization in which all k links are alike and attention is fixed on the average length $\bar{t}_{ar,k}'$ of the periods during which r or more of the links are in state (a).* He takes both $p_a(t)$ and $p_b(t)$ to be exponential, but his expression for $\bar{t}_{ar,k}'$ appears to hold for general distributions. Thus, let state j be the state of the system in which exactly j of the k links are in state (a), and let

$$
p_{aj,k} = \binom{k}{j} p_a^{\,j} p_b^{\,k-j} \tag{17}
$$

be the fraction of time the system spends in state j. Here $\binom{k}{j}$ is a binomial coefficient, $p_b = 1 - p_a$, and p_a is given by (14). Then Einhorn's results may be stated as

$$
\begin{aligned}
\bar{t}_{ar,k}' &= \bar{t}_{ar} \sum_{j=r}^{k} p_{aj,k}/p_{ar,k} \qquad (18)\\
&= \sum_{j=r}^{k} \binom{k}{j} (\bar{t}_a)^j (\bar{t}_b)^{k-j} \Big/ r \binom{k}{r} (\bar{t}_a)^{r-1} (\bar{t}_b)^{k-r} \\
\bar{t}_{ar,k}'' &= \bar{t}_{ar} \sum_{j=0}^{r-1} p_{aj,k}/p_{ar,k} \qquad (19)
\end{aligned}
$$

where $\bar{t}_{ar} = \bar{t}_a/r$ and $\bar{t}_{ar,k}''$ is the average length of the intervals during which $j < r$. Setting $r = k$ in (18) and (19) gives (8) and (15), respectively.

To establish (18) note that, in the very long interval $(0,T)$, the amount of time the system spends in states for which $j \geqq r$ is $T \sum_{r}^{k} p_{aj,k}$.

The number of periods in $(0,T)$ during which $j \geqq r$ is equal (to within one) to the number of periods during which $j < r$, and both are equal to the number of times the system jumps from state r to state $r - 1$. As shown in the next paragraph, the number of these jumps is $Tp_{ar,k}/\bar{t}_{ar}$, and (18) follows by division.

Divide the links into two groups, Group I consisting of the first r

* A similar problem has been considered in unpublished work by my colleague H. Coo, in which account is also taken of repairs at periodic intervals.

links and Group II of the last $k - r$ links. The fraction of time all $k - r$ links in II are in state (b) is p_b^{k-r}. The number of times all links in I are in state (a) is $p_a^r T/\bar{t}_{ar}$. This is also the number of times Group I jumps from state r to state $r - 1$. Since the two groups operate independently and no periodicities exist, we assume that p_b^{k-r} gives the fraction of these jumps occurring while all links in II are in state (b). Thus $[p_a^r T/\bar{t}_{ar}]p_b^{k-r}$ is the number of jumps the complete system makes from state r to state $r - 1$ when a specified set of $k - r$ links (namely the last $k - r$) remain in state (b). Since the set may be chosen in $\binom{k}{k - r}$ ways, the complete number of jumps is $Tp_{ar,k}/\bar{t}_{ar}$, as stated.

Incidentally, it may be shown that

$$F_{aj,k}(t) = -\bar{t}_{aj,k} \frac{d}{dt} A_a^j(t) A_b^{k-j}(t) \tag{20}$$

$$(\bar{t}_{aj,k})^{-1} = j(\bar{t}_a)^{-1} + (k - j)(\bar{t}_b)^{-1}$$

give the distribution and average length of the state j intervals.

IV. THE DISTRIBUTION OF t_{bk}

In the first part of this section several auxiliary distributions are discussed. They correspond to arbitrary $p_a(t)$, $p_b(t)$ and are more general than needed here, where ultimately $p_a(t)$ is required to be exponential. However, they are used in Section VI.

Consider the probability $Q_{aa}(t)$ that $x + t$ falls in a type (a) interval, given that the random point x falls in a type (a) interval. We have

$$Q_{aa}(t) = A_a(t) - \int_0^t P_{ba}(t - \tau) \frac{d}{d\tau} A_a(\tau)\, d\tau \tag{21}$$

in which $A_a(t)$ is the chance that the original (a) interval lasts beyond $x + t$ and $[-dA_a(\tau)/d\tau]\, d\tau$ the chance that it ends in $x + \tau$, $x + \tau + d\tau$. The end of the original (a) interval marks the beginning of a type (b) interval, and $P_{ba}(t')$ is the probability that a type (a) interval exists at time t', given that a type (b) interval began at time 0.

Let the Laplace transforms of $p_a(t)$, $p_b(t)$ be $\alpha(s)$, $\beta(s)$. Then the transforms of $F_a(t)$, $A_a(t)$ are given by (5). Weiss,[6] and Brooks and Diamantides,[8] have shown that the transform of $P_{ba}(t)$ is

$$s^{-1}[1 - \alpha(s)]\beta(s)/[1 - \alpha(s)\beta(s)].$$

This result is also developed in my paper[1] in ignorance of the earlier work of Weiss. Since the integral in (21) represents a convolution, its transform is the product of the transforms of $dA_a(t)/dt$ and $P_{ba}(t)$. When the transform of $Q_{aa}(t)$ is computed from (21), it is found to be

$$\begin{aligned} \phi_a(s) &= \int_0^\infty e^{-st} Q_{aa}(t)\, dt \\ &= \frac{1}{s} - \frac{[1 - \alpha(s)][1 - \beta(s)]}{s^2 \bar{t}_a [1 - \alpha(s)\beta(s)]} . \end{aligned} \tag{22}$$

This result is given by Palmer[2] and, independently, by Brooks and Diamantides.[8] It is also given, together with a number of related results, by Cox (Ref. 5, Ch. 7).

The argument in the two preceding paragraphs is concerned with type (a) intervals. It applies equally well to intervals of type (ak) *when the* (ak) *and* (bk) *intervals are independent.* When the links are alike, in place of $Q_{aa}(t)$ we have $[Q_{aa}(t)]^k$ for the chance that a type (ak) interval exists at time $x + t$, given that one exists at the randomly chosen time x. In place of the probability densities $p_a(t)$, $p_b(t)$ and their transforms $\alpha(s)$, $\beta(s)$ we have $p_{ak}(t)$, $p_{bk}(t)$ and their transforms $\alpha_k(s)$, $\beta_k(s)$. Equation (22) goes into an expression for the Laplace transform of $[Q_{aa}(t)]^k$

$$\frac{1}{s} - \frac{[1 - \alpha_k(s)][1 - \beta_k(s)]}{s^2 \bar{t}_{ak} [1 - \alpha_k(s)\beta_k(s)]} = \int_0^\infty e^{-st} [Q_{aa}(t)]^k\, dt. \tag{23}$$

Since $\alpha_k(s)$ may be computed from (7) and $Q_{aa}(t)$ from (22), $\beta_k(s)$ is the only unknown in (23). In principle, if not in practice, (23) may be solved for $\beta_k(s)$ and then $p_{bk}(t)$ obtained by inversion.

It should be remembered that (23) is based on the assumption that the (ak) and (bk) intervals are independent. They are independent when $p_a(t)$ is exponential, since then $p_{ak}(t)$ is also exponential, and a knowledge of the lengths of the (ak) intervals tells us nothing about the (bk) intervals and vice versa.

Thus when $p_a(t) = ae^{-at}$, so that $\alpha_k(s) = ka[ka + s]^{-1}$, (23) reduces to

$$\frac{1}{s + ka - ka\beta_k(s)} = \int_0^\infty e^{-st} [Q_{aa}(t)]^k\, dt \tag{24}$$

where $Q_{aa}(t)$ has the transform $[s + a - a\beta(s)]^{-1}$. More generally, when $p_a^{(i)}(t) = a_i \exp(-a_i t)$ and $p_b^{(i)}(t)$ is arbitrary the equation for $\beta_k(a)$ becomes

$$\frac{\bar{t}_{ak}}{s\bar{t}_{ak} + 1 - \beta_k(s)} = \int_0^\infty e^{-st} \prod_{i=1}^{k} Q_{aa}^{(i)}(t)\, dt \tag{25}$$

where $\bar{t}_{ak} = [\sum a_i]^{-1}$ and $Q_{aa}{}^{(i)}(t)$ has the transform $[s + a_i - a_i\beta^{(i)}(s)]^{-1}$. This is essentially the result obtained by Palmer[2] and Takács. (Takács' version is given in my paper.[1])

The only case in which the (ak) and (bk) intervals are obviously independent seems to be that for exponential $p_a(t)$. On the other hand, the following example shows that successive (ak) and (bk) intervals may be correlated even though the (a) and (b) intervals for the individual links are not. Let t_a and t_b be uniformly distributed between 1.00, 1.01 and 2.00, 2.01 respectively. Let $k = 2$. Then, given an (ak) interval of length $t_{a2} = 0.5$, we can infer that the length of the following (bk) interval lies between 2.5 and 2.52.

Hence, so far as the discussion given in this section goes, (23) is no more general than (24). The question now arises as to the form taken by $p_{bk}(t)$ when $p_a(t)$ and $p_b(t)$ are arbitrary. Some information on this is given in Section VI for the case $k = 2$.

V. DISCUSSION AND EXAMPLES

When k tends to infinity, with $p_a(t)$, $p_b(t)$ fixed but arbitrary, $\bar{t}_{ak}$ tends to zero and $\bar{t}_{ba}$ to infinity. When t becomes small, (2) shows that $A_a(t)$ tends to $1 - t/\bar{t}_a$ and its kth power to $\exp(-tk/\bar{t}_a)$. It follows that when t is small and k is large, the chance that the length of an (ak) interval exceeds t is

$$F_{ak}(t) \approx \exp(-t/\bar{t}_{ak}). \tag{26}$$

It may be conjectured that when $k \to \infty$ the chance $F_{bk}(t)$ that $t_{bk} > t$ also tends to an exponential

$$F_{bk}(t) \to \exp(-t/\bar{t}_{bk}). \tag{27}$$

Here t is supposed to be many times larger than $\bar{t}_a$ and $\bar{t}_b$. Indeed, consider a (bk) interval to be in progress and k to be large. Over all of the interval, except for a negligibly small fraction near the beginning, the process will be uncorrelated with the initial conditions, and the chance that the interval will end in $(t, t + dt)$ is independent of t. This leads to the exponential form (27). When $\bar{t}_b/\bar{t}_a \ll 1$, k may have to be extremely large before (27) begins to hold. This is because the argument pictures a great deal of overlap of (b) type intervals.

Next we turn to the case where the k links are different and

$$\begin{gathered} p_a{}^{(i)}(t) = a_i e^{-a_i t}, \qquad p_b{}^{(i)} = b_i e^{-b_i t}, \qquad i = 1, 2, \cdots, k \\ \bar{t}_a{}^{(i)} = a_i^{-1}, \qquad \bar{t}_b{}^{(i)} = b_i^{-1}. \end{gathered} \tag{28}$$

This is one of the few cases in which the work may be carried forward to even a moderate degree. From Section II

$$\begin{aligned} F_a^{(i)}(t) &= A_a^{(i)}(t) = e^{-a_i t} \\ \bar{t}_{ak} &= [\textstyle\sum a_i]^{-1} \\ F_{ak}(t) &= \exp\,[-t/\bar{t}_{ak}] = \exp\,[-t \textstyle\sum a_i] \\ p_{ak}(t) &= [\textstyle\sum a_i] \exp\,[-t \textstyle\sum a_i]\,. \end{aligned} \tag{29}$$

These expressions for $F_{ak}(t)$ and $p_{ak}(t)$ also hold when $p_b^{(i)}(t)$ is arbitrary. From Section III

$$\begin{aligned} \bar{t}_{bk} &= \left(\frac{1}{p_{ak}} - 1\right) \bar{t}_{ak}\,, \\ p_{ak} &= \prod_{i=1}^{k} p_a^{(i)}, \qquad p_a^{(i)} = b_i(a_i + b_i)^{-1}. \end{aligned} \tag{30}$$

The first step in using (25) is to compute $Q_{aa}^{(i)}(t)$ by inverting its Laplace transform. The result is given in the last line of Table I in Section VI and leads to

$$\prod_{i=1}^{k} Q_{aa}^{(i)}(t) = \prod_{i=1}^{k} \left[\frac{b_i + a_i e^{-(a_i+b_i)t}}{b_i + a_i}\right] = \sum_j c_j e^{-d_j t} \tag{31}$$

where in the general case the sum on j contains 2^k terms. The right-hand member of (25) becomes $\sum c_j(s + d_j)^{-1}$, and it follows that the Laplace transform of $F_{bk}(t)$ is equal to

$$\frac{1 - \beta_k(s)}{s} = \frac{\bar{t}_{ak}}{s \sum_j c_j(s + d_j)^{-1}} - \bar{t}_{ak}\,. \tag{32}$$

The transform of $F_{bk}(t)$ is thus a rational function of s. Its poles are at the zeros of $\sum c_j(s + d_j)^{-1}$, and these zeros lie on the negative real axis between the points $s = -d_j$, the rightmost of which is $s = 0$. The nth moment, $\overline{t_{bk}^n}$, of t_{bk} is $n!(-1)^{n-1}$ times the coefficient of s^{n-1} in the power series expansion of (32). Hence for $n > 0$

$$\overline{t_{bk}^n} = n\bar{t}_{ak} \left[\left(-\frac{d}{ds}\right)^{n-1} \left\{\frac{1}{s \sum c_j(s + d_j)^{-1}} - 1\right\}\right]_{s=0}. \tag{33}$$

When the links are alike, some of the d_j's are equal, and

$$\sum_j c_j e^{-d_j t} = \left[\frac{b + a e^{-(a+b)t}}{b + a}\right]^k = \sum_{n=0}^{k} \binom{k}{n} \frac{b^{k-n} a^n}{(b + a)^k}\, e^{-n(a+b)t}.$$

In this case, the results are those used in Ref. 1, namely

$$\begin{aligned} F_{ak}(t) &= e^{-tka}, \qquad \bar{t}_{ak} = 1/ka \\ F_{bk}(t) &= \frac{(1+\rho)^{k+1}}{k\rho} \sum_{m=0}^{k-1} \frac{\exp\left[(a+b)z_m t\right]}{z_m f'(z_m)} \\ \bar{t}_{bk} &= [(1+\rho)^k - 1]/ka, \qquad \rho = a/b \end{aligned} \tag{34}$$

where $f_k'(z) = df_k(z)/dz$, and $z_0, z_1, \cdots, z_{k-1}$ are the zeros of

$$f_k(z) = \sum_{n=0}^{k} \binom{k}{n} \frac{\rho^n}{z+n}. \tag{35}$$

These zeros lie between the poles at 0, -1, -2, $\cdots$, $-k$. The first few terms in the power series for $F_{bk}(t)$ are given by (14) of Ref. 1. In present notation

$$\begin{aligned} F_{bk}(t) = 1 - \frac{bt}{1!} + \frac{[(k-1)a+b]bt^2}{2!} \\ - [(k-1)^2a^2 + 4(k-1)ab + b^2]\frac{bt^3}{3!} + \cdots. \end{aligned} \tag{36}$$

The nth moment, $n > 0$, is

$$\overline{t_{bk}^n} = \frac{n(1+\rho)^{k-n+1}}{k\rho b^n}\left[\left(-\frac{d}{dz}\right)^{n-1}\left\{\frac{1}{zf_k(z)} - (1+\rho)^{-k}\right\}\right]_{z=0}. \tag{37}$$

In particular

$$\overline{t_{bk}^2} = \frac{2(1+\rho)^{k-1}}{akb}\sum_{n=1}^{k}\binom{k}{n}\frac{\rho^n}{n} \tag{38}$$

a result given by Palmer[2] for $a = b = 1$.

When $k = 2$ and the links are alike, (32) gives

$$\frac{1 - \beta_2(s)}{(s)} = \frac{s + a + 2b}{s^2 + (3b + a)s + 2b^2}. \tag{39}$$

Results of the sort given here have occurred in the reliability studies of R. S. Dick[9] and others.

VI. THE DISTRIBUTION OF t_{bk} FOR $k = 2$

Here we consider the determination of $p_{bk}(t)$ for the case of $k = 2$ links when $p_a(t)$ and $p_b(t)$ are arbitrary. The following probabilities, which are related to those mentioned in the first part of Section IV, will be needed:

$P(t)$ = chance that an (a) interval exists at time t, given that an (a) interval starts at time 0;

$Q(t)$ = chance that an (a) interval exists at time $x + t$, given that the random x of Fig. 2 falls in an (a) interval. $Q(t)$ is equal to the $Q_{aa}(t)$ of Section IV;

$R(t)\,dt$ = chance that an (a) interval ends in $t, t + dt$, given that an (a) interval starts at time 0; and

$S(t)\,dt$ = chance that an (a) interval ends in $x + t, x + t + dt$, given that the random x falls in an (a) interval.

Some information regarding these probabilities is summarized in Table I. Here $\alpha(s)$ and $\beta(s)$, the Laplace transforms of $p_a(t)$ and $p_b(t)$, are written for brevity as α and β. The entries may be obtained by the methods indicated in Section IV. Closely related results have been given in unpublished work by my colleague H. E. Rowe.

An expression for the distribution of the length of a $(b2)$ interval which consists of, say, three (b) intervals can be obtained by examining Fig. 3. The probability that the $(b2)$ interval shown in Fig. 3 has a length between t and $t + dt$ can be obtained by integrating

$$\begin{aligned} &p_b(t - x_2)\,dx_2 \cdot R(x_2 - x_3)\,dx_3 \cdot p_b(x_3)\,dt \\ &S(t - x_1)\,dx_1 \cdot p_b(x_1 - x_4)\,dx_4 \cdot P(x_4) \end{aligned} \tag{40}$$

over the permissible values of the x's, namely $0 \leqq x_4 \leqq x_3 \leqq x_2 \leqq x_1 \leqq t$.

Starting with the probability $p_b(t)Q(t)\,dt$ that a $(b2)$ interval consists of one (b) interval and is of length $t, t + dt$, one can write a series for $p_{b2}(t)\,dt$. The first term is $p_b(t)Q(t)\,dt$, and the later terms are multiple integrals of the type obtained by integrating (40). An examination of the series shows that

$$p_{b2}(t) = p_b(t)Q(t) + \int_0^t dx_1 S(t - x_1)q(t,x_1) \tag{41}$$

where $q(t,x)$ satisfies the relation

$$\begin{aligned} q(t,x_1) = \int_0^{x_1} &dx_2 p_b(t - x_2) \\ &\cdot \left[p_b(x_1)P(x_2) + \int_0^{x_2} dx_3\, R(x_2 - x_3)q(x_1, x_3) \right]. \end{aligned} \tag{42}$$

The series for $p_{b2}(t)$ obtained by repeated substitution of (42) in (41) is sometimes useful for small values of t.

TABLE I — INFORMATION ON AUXILIARY DISTRIBUTIONS

	$P(t)$	$Q(t)$	$R(t)$	$S(t)$
Laplace transforms	$\dfrac{1-\alpha}{s(1-\alpha\beta)}$	$\dfrac{1}{s}-\dfrac{(1-\alpha)(1-\beta)}{s^2\bar{t}_a(1-\alpha\beta)}$	$\dfrac{\alpha}{1-\alpha\beta}$	$\dfrac{1-\alpha}{s\bar{t}_a(1-\alpha\beta)}$
Exponential $p_a(t)$, $\alpha = a(s+a)^{-1}$	$\varphi_P = \dfrac{1}{s+a-a\beta}$	φ_P	$a\varphi_P$	$a\varphi_P$
Exponential $p_b(t)$, $\beta = b(s+b)^{-1}$	$\dfrac{(b+s)(1-\alpha)}{s(s+b-b\alpha)}$	$\dfrac{1}{s}-\dfrac{1-\alpha}{s\bar{t}_a(s+b-b\alpha)}$	$\dfrac{(s+b)\alpha}{(s+b-b\alpha)}$	$\dfrac{(s+b)(1-\alpha)}{s\bar{t}_a(s+b-b\alpha)}$
Exponential $p_a(t)$ and $p_b(t)$	$\varphi_P = \dfrac{s+b}{s(s+a+b)}$	φ_P	$a\varphi_P$	$a\varphi_P$
$P(t), \cdots$ for exponential $p_a(t)$ and $p_b(t)$	$P = \dfrac{b+ae^{-(a+b)t}}{b+a}$	P	aP	aP

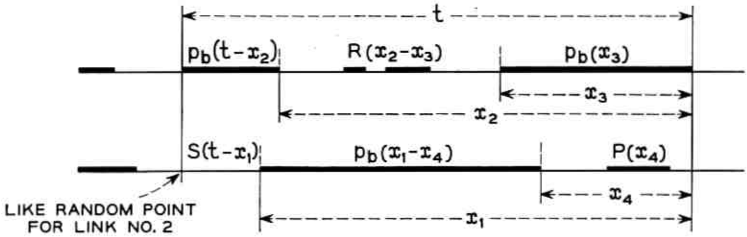


Fig. 3 — Sketch illustrating quantities in expression (40).

Equation (42) is of the form

$$q(t,x) = g(t,x) + \int_0^x K(t - x, x - y) q(x,y)\, dy, \tag{43}$$

where g and K are known functions, and is an integral equation to determine $q(t,x)$ in the region $0 \leqq x \leqq t$. In this region the values of $q(t,x)$ along the line $x = x_1$ are expressed in terms of its values on the line $t = x_1$.

It appears difficult to solve (42) for general $p_a(t)$ and $p_b(t)$. Two special cases are discussed in Section VII.

It should be possible to verify that when $p_a(t) = a \exp(-at)$, (41) and (42) lead to the same result as does (24) with $k = 2$. However, I have been unable to do this for general $p_b(t)$. The problem may be stated as follows: show that the $\beta_2(s)$ defined by setting $k = 2$ and $Q_{aa}(t) = P(t)$ in (24) is the Laplace transform of

$$p_{b2}(t) = p_b(t)P(t) + a \int_0^t P(t - x_1) q(t,x_1)\, dx_1 \tag{44}$$

where $P(t)$ has the transform $[s + a - a\beta(s)]^{-1}$ and $q(t,x_1)$ satisfies

$$q(t,x_1) = \int_0^{x_1} dx_2 p_b(t - x_2) \cdot \left[p_b(x_1)P(x_2) + a \int_0^{x_2} dx_3 P(x_2 - x_3) q(x_1, x_3) \right].$$

When the two links have different statistics, the distribution of the lengths of ($b2$) intervals which begin with a (b) interval of Link No. 1, say, may be expressed as an integral similar to (41). However, there are now two integral equations similar to (42) which must be solved simultaneously.

VII. SPECIAL CASES FOR $p_{b2}(t)$

Here two special cases are given in which the integral equation (42) may be solved. The second case has been used to study the problem mentioned in connection with (44).

(*i*) *Exponential $p_b(t)$ and general $p_a(t)$*. When $p_b(t) = b \exp(-bt)$, (42) shows that $q(t,x_1)$ is of the form $f(x_1) \exp(-bt)$ where

$$f(x_1) = b \int_0^{x_1} dx_2 e^{-b(x_1 - x_2)} \left[bP(x_2) + \int_0^{x_2} R(x_2 - x_3) f(x_3)\, dx_3 \right]. \tag{45}$$

This goes into

$$F(s) = \frac{b}{s + b} [b\varphi_P(s) + \varphi_R(s) F(s)] \tag{46}$$

where $F(s)$, $\varphi_P(s)$, and $\varphi_R(s)$ are the Laplace transforms of $f(t)$, $P(t)$, and $R(t)$. Similarly, the transform of (41) is

$$\beta_2(s) = b\varphi_Q(s + b) + \varphi_S(s + b) F(s + b). \tag{47}$$

From (46), (47), and Table I

$$F(s) = \frac{b^2[1 - \alpha(s)]}{s[s + b - 2b\alpha(s)]}, \tag{48}$$

$$\beta_2(s) = \frac{b}{s + b} - \frac{sb[1 - \alpha(s + b)]}{(s + b)^2 \bar{t}_a [s + 2b - 2b\alpha(s + b)]}. \tag{49}$$

Inversion of $\beta_2(s)$ and $[1 - \beta_2(s)]s^{-1}$ now gives $p_{b2}(t)$ and $F_{b2}(t)$.

When the two links have different exponential $p_b(t)$'s and general $p_a(t)$'s, the two simultaneous integral equations mentioned at the end of Section VI may be solved by a somewhat similar procedure.

(*ii*) *$p_a(t)$ exponential and $p_b(t)$ the sum of two exponentials*. For

$$p_a(t) = ae^{-at}, \qquad p_b(t) = c_1 e^{-b_1 t} + c_2 e^{-b_2 t}, \qquad c_1 b_1^{-1} + c_2 b_2^{-1} = 1 \tag{50}$$

(42) shows that $q(t,x)$ is of the form

$$c_1 e^{-b_1 t} f_1(x) + c_2 e^{-b_2 t} f_2(x).$$

Instead of $f_j(x)$, $j = 1, 2$, it is more convenient to deal with

$$g_j(x) = P(x) + a \int_0^x P(x - y) f_j(y)\, dy$$

having the Laplace transform $G_j(s)$. When Laplace transforms are introduced, (42) goes into two simultaneous equations for $G_1(s + b_1)$, $G_2(s + b_2)$. Solving these leads to

$$
\begin{aligned}
\beta_2(s) &= c_1G_1(s + b_1) + c_2G_2(s + b_2) \\
&= \frac{c_2B_1 + c_1B_2 + 2ac_1c_2(s + b_1 + b_2)^{-1}}{B_1B_2 - a^2c_1c_2(s + b_1 + b_2)^{-2}}, \\
B_j &= \frac{1}{\varphi_P(s + b_j)} - \frac{ac_j}{s + 2b_j}, \qquad \varphi_P(s) = \frac{1}{s + a - a\beta(s)}, \\
\beta(s) &= \frac{c_1}{s + b_1} + \frac{c_2}{s + b_2}.
\end{aligned} \tag{51}
$$

The equation for $\beta_2(s)$ obtained from (24) in this case is

$$
\begin{aligned}
[s + 2a - 2a\beta_2(s)]^{-1} &= \int_0^\infty e^{-st}P^2(t)\,dt \\
&= \frac{1}{2\pi i}\int_L \varphi_P(s - s')\varphi_P(s')\,ds' \\
&= \frac{\varphi_P(s)b_1b_2}{s_1s_2} + \frac{\varphi_P(s - s_1)(s_1 + b_1)(s_1 + b_2)}{s_1(s_1 - s_2)} \\
&\quad + \frac{\varphi_P(s - s_2)(s_2 + b_1)(s_2 + b_2)}{s_2(s_2 - s_1)}
\end{aligned} \tag{52}
$$

where the path of integration L runs from $-i\infty$ to $+i\infty$ so that the singularities of $\varphi_P(s - s')$ lie on its right and those of $\varphi(s')$ on its left. The integral has been evaluated by writing $\varphi_P(s)$ as

$$\varphi_P(s) = \frac{(s + b_1)(s + b_2)}{s(s - s_1)(s - s_2)},$$

$$(s - s_1)\,(s - s_2) = s^2 + (a + b_1 + b_2)s + b_1b_2 + a(b_1 + b_2 - c_1 - c_2)$$

closing L by an infinite semicircle on the left and evaluating the residues at the poles $s' = 0, s_1, s_2$.

The task of verifying that (51) and (52) give the same value of $\beta_2(s)$ appears to be a lengthy one and has not been carried out. Several numerical checks have been made and show no discrepancy.

REFERENCES

1. Rice, S. O., Intervals between Periods of No Service in Certain Satellite Communication Systems, B.S.T.J., **41,** Sept., 1962, pp. 1671–1690.
2. Palmer, D. S., A Theoretical Study of the Statistics of Working Spells and Periods of Breakdown for a Number of Radio Links in Series, in *Statistical Methods in Radio Wave Propagation*, ed. Hoffman, W. C., Pergamon Press, New York, 1960.
3. Palm, C., Intensitätsschwankugen im Fernsprechverkehr, Ericsson Technics, **44,** Ericsson, L. M., Stockholm, 1943.

4. Riordan, J., *Stochastic Service Systems*, John Wiley and Sons, New York, 1962, p. 12.
5. Cox, D. R., *Renewal Theory*, Methuen, London, 1962, Ch. 5 and 7.
6. Weiss, G. H., A Note on the Coincidence of Some Random Functions, Quart. Appl. Math., **16,** 1956, pp. 103–107.
7. Einhorn, J. J., Reliability Prediction for Repairable Redundant Systems, Proc. IEEE, **51,** Feb., 1963, pp. 312–317.
8. Brooks, F., and Diamantides, N., A Probability Theorem for Random Two-Valued Functions, with Applications to Auto-Correlations, SIAM Review, **5,** Jan., 1963, pp. 33–40.
9. Dick, R. S., The Reliability of Repairable Complex Systems: Part A, Proc. Fifth Natl. Conv. Prof. Group on Military Electronics, Washington, D. C., June, 1961, pp. 111–150.

Wide-Angle Radiation Due to Rough Phase Fronts

By C. DRAGONE and D. C. HOGG

(Manuscript received March 8, 1963)

Nonuniformities in the phase fronts of electromagnetic and acoustical waves give rise to radiation in directions other than that desired. The magnitude of this effect is discussed here with special reference to quasi-random roughness. It is found that the level of wide-angle radiation is a strong function of the phase deviations and that reflecting surfaces, for example, should be held to tolerances of about $\pm 0.01\lambda$ to prevent the level of the wide-angle radiation from exceeding twice that due to a perfectly smooth reflector.

I. INTRODUCTION

A rough or nonuniform phase front, be it acoustical, radio, or optical, usually degrades the desired performance of components which transmit, reflect, or receive the wave. The effect is well known in the field of microwave antennas, where lenses are required to have sufficient homogeneity of dielectric constant and reflectors sufficient smoothness of surface to produce uniform phase fronts. Likewise, the quality of optical components is specified, among other things, by ability to reproduce or modify wavefronts in a prescribed manner without undue distortion.

If roughness is introduced into a wavefront by a component, some of the power is no longer radiated in the desired specular direction; it propagates at angles well removed from that direction. This effect can be described by a system of modes in the radiating aperture, each of which radiates in a specified direction.*

In practice, it is difficult to describe the roughness properly. Consider, for example, the reflector of a microwave antenna. If large, seldom is it constructed from a single sheet of metal. More often, sheets are cut and shaped to form modules of given dimension, these then being assembled with the desired precision to construct the antenna. One might expect, therefore, that the power spectrum of the wavefront would have a com-

* Often called the angular power spectrum.

ponent at a spatial wavelength related to the module dimension. Inevitably, there is also a somewhat random component. Only in special cases can one estimate the predominant spatial wavelengths in the random case. For example, if an optical component is ground with particles of given average diameter, then the roughness may be expected to have a corresponding spatial period. In the discussion that follows, we will be concerned mainly with the problem in its relationship to microwave antennas.

Degradation of the radiation patterns of microwave antennas in the vicinity of the main beam due to various phase errors has been studied quite thoroughly.[1,2] From these studies it is often concluded that $\lambda/16$ is a suitable tolerance for reflector surfaces as far as the main beam and immediate side lobes are concerned. The purpose here is to investigate the effect of phase error on that portion of the radiation pattern well removed from the main beam, i.e., the far or wide-angle lobes. For example, we question whether $\pm\lambda/16$ is a suitable tolerance for receiving antennas at earth stations of space communications systems, since in this application the far side lobes control significantly the amount of noise that enters the antenna due to radiation from the earth. At the same time, these lobes influence the amount of man-made interference that such a receiving antenna will withstand. Likewise, one requires that the radiation pattern of transmitting antennas be as clean as possible, thus permitting only the least possible radiation to be propagated in directions other than that of the main beam.

A description of the radiating modes is given first (in one dimension). The circular aperture is then discussed as a specific example. Single sinusoidal phase errors and quasi-random errors constructed from multiple sinusoids are treated. From these calculations, it is concluded that $\pm 0.01\lambda$ is a desirable tolerance for reflecting surfaces of good quality. Finally, the calculation is compared with some experimental data.

II. THE ONE-DIMENSIONAL CASE

Consider a rectangular aperture with sides of length a, b. The center of the aperture is taken as the origin of a Cartesian system of coordinates, and the x, y axes lie in the plane of the aperture. For descriptive purposes, the electric field existing within the aperture is assumed of constant amplitude and directed in the y direction, $E_{ay} = E_0 e^{j\psi(x)}$, where $\psi(x)$ is the phase error. The exponential $e^{j\psi(x)}$ can be expanded in its Fourier series

$$e^{j\psi(x)} = \sum_{n=-\infty}^{\infty} A_n \exp \frac{jn2\pi x}{a} \qquad (1)$$

n being a positive integer. One has

$$e^{j\psi(x)}e^{-j\psi(x)} = 1. \tag{2}$$

By means of (1), one can express the field over the aperture as a sum of the partial fields E_{an}

$$E_{an} = E_0 \left(A_n \exp \frac{jn2\pi x}{a} + A_{-n} \exp \frac{-jn2\pi x}{a} \right) \tag{3}$$

n being the number of sinusoidal phase deviations across the aperture. These elementary modes satisfy the orthogonality relation

$$\int_{-a/2}^{a/2} E_{an}E_{am}\, dx = 0 \qquad \text{if } n \neq m.$$

Then the power P radiated by the aperture is given by the sum of the powers P_n radiated by each mode.

$$P = \sum_0^\infty P_n .$$

The radiation field associated with the nth mode can be derived from the magnetic Hertzian potential directed along the z axis[3]

$$(\Pi_z)_n = \frac{-a}{n2\pi\omega\mu} e^{-jh_nz} \left[A_n \exp \frac{jn2\pi x}{a} - A_{-n} \exp \frac{-jn2\pi x}{a} \right] \tag{4}$$

where $h_n = k_0\sqrt{1 - (n/n_0)^2}$, $k_0 = 2\pi/\lambda$ and $n_0 = a/\lambda$. The component of the magnetic field associated with E_{an} in the plane of the aperture is given by $(H_x)_n(Z_z)_n = (E_y)_n$, $(Z_z)_n$ being the wave impedance of the nth mode measured in the z direction and $(E_y)_n = j\omega\mu\, \partial(\Pi_z)_n/\partial x$. That is

$$Z_n = (E_y/H_x)_n = (h_n/k_0)^{-1}Z_0 = Z_0(\sqrt{1 - (n/n_0)^2})^{-1}. \tag{5}$$

Z_0 is the intrinsic impedance for a plane wave. After integrating the z component of the Poynting vector over the aperture, one has

$$P_n = b \int_{-a/2}^{a/2} | E_{an} |^2\, 1/Z_n\, dx. \tag{6}$$

When $n > n_0$, the wave impedance is imaginary and no real power is radiated in the nth mode, in which case P_n corresponds to production of a storage field. When $n < n_0$, P_n is real, and the nth mode radiates principally in the directions

$$\sin\theta = \pm n/n_0 \tag{7}$$

where θ is measured with respect to the z axis in the x-z plane. This fact can be readily seen from (3) by noting that the nth mode can be thought to be generated in the aperture plane by the superposition of two plane waves traveling in the two directions $\pm\theta$. The amplitudes of these two plane waves are given by E_0A_n, E_0A_{-n}. If one substitutes $n_0 = a/\lambda$ in (7), the grating formula $\sin\theta = \pm n\lambda/a$ results.

In the geometrical optics approximation, these two components would radiate all the power in the two mentioned directions. Indeed, according to the diffraction phenomena, the radiation pattern of E_{an} has two main lobes of amplitudes A_nE_0, $A_{-n}E_0$, directed in the above-mentioned directions. Further, one finds that the patterns of all other components have zeroes there.

The radiation pattern generated by E_{a0} is then the only one to contribute to the power radiated per unit solid angle along the axis of the aperture. This power is proportional to $|A_0|^2$. Remembering that $A_0 = 1$ when $\psi(x) = 0$, one finds that $1 - |A_0|^2$ gives the decrease in gain of the aperture caused by the phase error $\psi(x)$. From (6), the power radiated by the higher-order modes ($n > 0$) is given by

$$\begin{aligned} P_n &= \frac{abE_0^2}{Z_0}(|A_n|^2 + |A_{-n}|^2)\sqrt{1-(n/n_0)^2} \\ &= P_a(|A_n|^2 + |A_{-n}|^2)\sqrt{1-(n/n_0)^2} \end{aligned} \tag{8}$$

where $P_a = abE_0^2/Z_0$ is the power radiated by the aperture when $\psi(x) = 0$.*

III. THE CIRCULAR APERTURE

Consider the circular aperture of Fig. 1. Let the field distribution be given by

$$[1 - \alpha(2\rho/D)^2]e^{j\psi} \tag{9}$$

where ψ is the phase error and a square-law taper of amplitude α has been considered. The radiation pattern of the aperture of diameter D is then given by[1]

$$g(u,\varphi) = D^2/4\int_0^{2\pi}\int_0^1 (1 - r^2\alpha)re^{j\psi}\,e^{jur\cos(\varphi-\varphi')}\,dr\,d\varphi' \tag{10}$$

* If $n \ll n_0$, then $Z_n \approx Z_0$ and $P_n \approx P_a(|A_n|^2 + |A_{-n}|^2)$. The total power radiated by the higher modes represents an increase in the power radiated in the side lobe region of the radiation pattern. Let P_s be this power; then

$$P_s \approx P_a \sum_{n\neq 0} |A_n|^2.$$

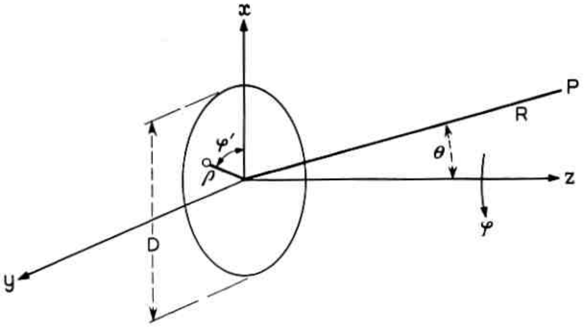


Fig. 1 — Aperture and coordinates.

where $u = (\pi D/\lambda) \sin \theta$ and $r = 2\rho/D$ is the normalized radial coordinate, θ being the angle between the antenna axis and the field point.*

Let us now consider a radial phase error $\psi = \psi(\rho)$. In this case the pattern is symmetrical about the aperture axis, and after performing the integration over φ', (10) becomes

$$g(u) = \pi D^2/2 \int_0^1 (1 - \alpha r^2) \cdot r \cdot e^{j\psi} \cdot J_0(ur)\, dr. \tag{11}$$

When $\psi = 0$, integration of (11) is readily performed and the result after normalization is

$$g_1(u) = 2(1 - \alpha) J_1(u)/u + 4\alpha J_2(u)/u^2 \tag{12}$$

where $J_p(u)$ is the Bessel function.

Different types of phase error ψ will now be considered, and the resulting radiation patterns will be compared in the far side lobe region with the ideal case given by $\psi = 0$. In all cases, since we are considering specifically paraboloidal antennas, the amplitude of the aperture field will be assumed to taper in square-law fashion to -10 db at the edge. After substituting the appropriate value of 0.684 for α in (12), one obtains the ideal pattern with no phase error shown in Fig. 2.

3.1 *Effect of Single Sinusoidal Errors*

Rather than expand the function $e^{j\psi}$ in a series as described in Section II, we choose here to expand the phase itself in a series for purpose of

* There are criticisms of the Huygens-Kirchhoff diffraction theory, especially when calculations are made at angles well removed from the axis; although some of the criticisms are known to be valid, we ignore them for these calculations. The $(1 + \cos \theta)$ factor is neglected; it can readily be multiplied in for any given antenna.

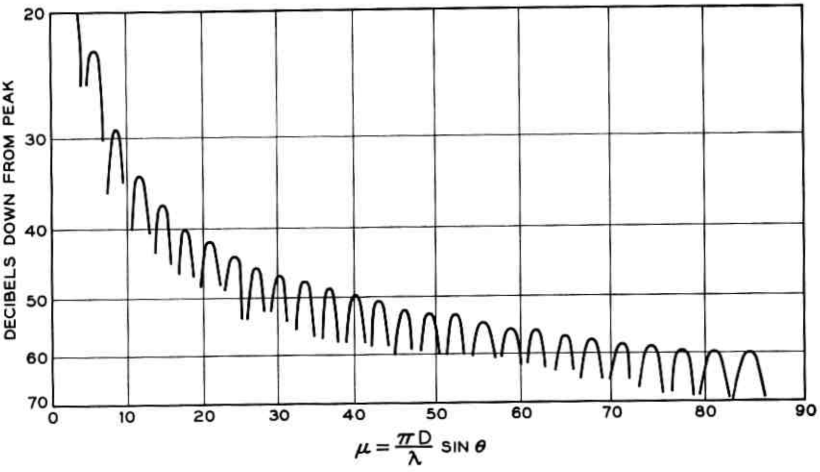


Fig. 2 — Radiation pattern for zero phase error; illumination 10-db taper.

computation. Thus by means of (11), $g(u)$ is calculated for single sinusoidal phase errors of various periods; that is, for

$$\psi = \Phi_m \cos(2\pi m r) \tag{13}$$

m being the number of periods along the aperture radius. The cases $m = 6$ and 12 for $\Phi = 2\pi/16$ have been computed and the results are plotted* in Fig. 3.

The curves of Figs. 3(a) and 3(b) show that a sinusoidal phase error causes a large disturbance only in a relatively small angular region of the radiation pattern and that these regions are located at angles which increase with the number of fluctuations in the phase distribution. The level of the disturbance depends upon the amplitude of Φ and is sensibly independent of m for the cases considered. Clearly, these disturbances are simply related to the various radiation modes discussed in Section II. For example, in Fig. 3(b) where $m = 12$, a disturbance occurs at $u = 75$; this same value is calculated by substituting $n = 2m = 24$ in (7) for $\sin\theta$, thereby evaluating the appropriate u. Thus Fig. 3(b) shows the effect of mode Π_{z24} described by (4).

Data other than those given in Fig. 3 show that there is little overlapping of adjacent disturbances, for example, between patterns for $m = 6$ and $m = 7$. In Fig. 3(a), note that a small "second harmonic" disturb-

* The patterns are plotted in decibels below the peak value. Since the peak value (the gain) is reduced only slightly due to the small phase errors considered here, the correction is neglected.

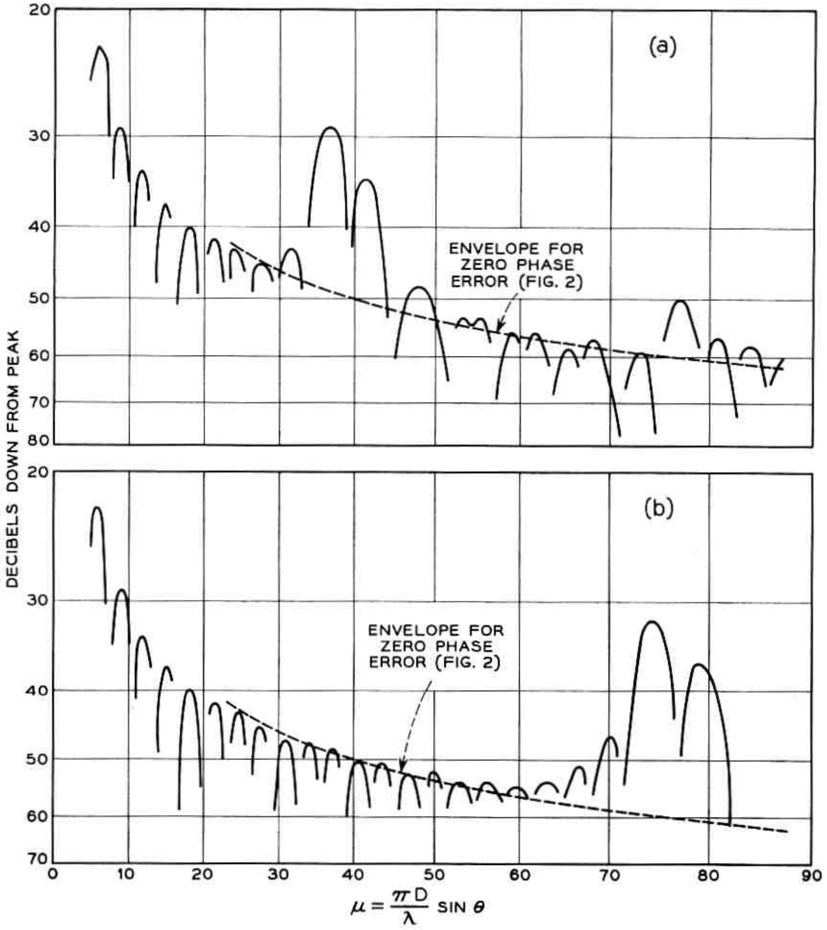


Fig. 3 — Radiation patterns — single sinusoidal phase error. (a) $m = 6$; maximum phase error, $\lambda/8$ peak-to-peak; illumination 10-db taper. (b) $m = 12$; maximum phase error, $\lambda/8$ peak-to-peak; illumination 10-db taper.

ance is evident at $u \approx 80$; this is the position of a second-order fringe if the aperture is interpreted as a weak grating.

3.2 *The Effect of a Typical Phase Error*

Let us now construct from many sinusoidal components a phase error given by the Fourier series

$$\psi = \Phi(\sum_{m=1}^{M} \sin\,(2\pi m[r - (m - 1)/2M - \tfrac{1}{4}])) \tag{14}$$

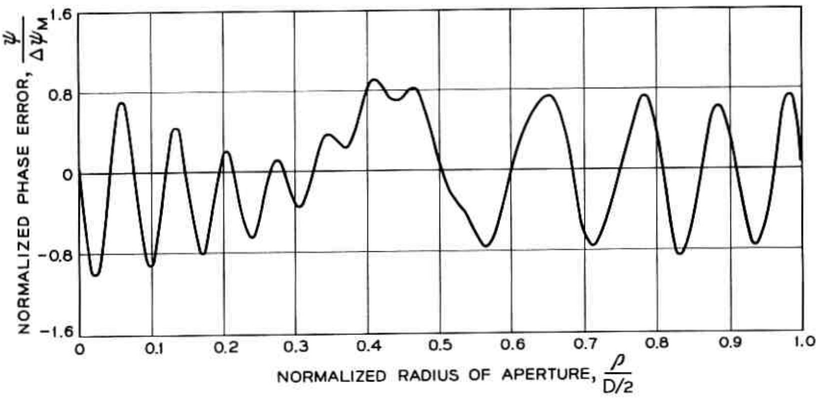


Fig. 4 — Typical phase error derived from 15 sinusoidal components of equal amplitude.

in which the M coefficients are of constant amplitude $\Phi_m = \Phi$. If the spectrum of roughness were known, the Φ_m would be written as a function of r at this point. By means of the term $(m - 1)/2M$, the phase of each component is (somewhat arbitrarily) shifted so that ψ does not have excessive irregularities and has a reasonable peak-to-peak value, $\Delta\psi_M$, between the maximum and minimum value of ψ. The phase distribution resulting from (14) is plotted in Fig. 4 for the case $M = 15$; the resulting peak-to-peak phase for this example turns out to be $\Delta\psi_M = 9.95\varphi$. Of course, one can see certain regularities in the function; but as mentioned previously, in practice these would be related to the design of any given reflector.

Radiation patterns have been computed using the phase function of Fig. 4; these are plotted in Figs. 5, 6, and 7 for peak-to-peak phase ($\Delta\psi_M$) of 0.31λ, 0.155λ, and 0.1λ, respectively.

In Fig. 8, the envelopes of the patterns shown in Figs. 5, 6, and 7 are compared with that of the ideal pattern, $\Delta\psi_M = 0$. Beyond $u \approx 30$, Fig. 8 shows, for example, that wide-angle radiation is 10 to 15 db above the ideal case if the peak-to-peak phase error is 0.155λ (roughly equivalent to a reflector with tolerance ±λ/25). Also shown in Fig. 8 is the envelope for $\Delta\psi_M = 0.048\lambda$, in which case the far side lobes are about 3 db above the ideal case.

Since data have been computed for five values of phase error (including zero), we may plot degradation in side lobe level relative to the ideal case versus peak-to-peak phase error for various angles from the desired direction of propagation. That plot is shown in Fig. 9 for angles corresponding to $u = 50$.

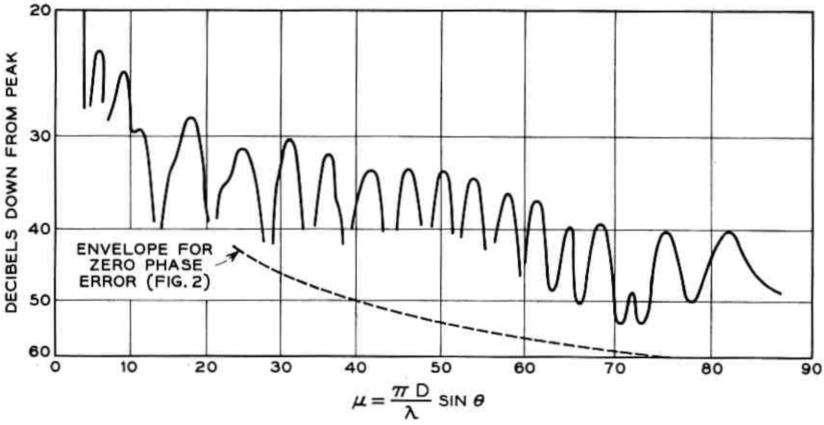


Fig. 5 — Radiation pattern: typical phase error — maximum phase error 0.31λ peak-to-peak; illumination 10-db taper.

Let us now define a good radiation pattern for a real reflector by stipulating that the wide-angle radiation be less than 3 db above the ideal case. Referring to Fig. 9, one sees that for this condition to obtain, $\Delta\psi_M \approx 0.05\lambda$; therefore, since phase disturbances are roughly doubled in reflection from, say, a relatively shallow paraboloid, surface tolerance must be held to about $\pm 0.0125\lambda$.

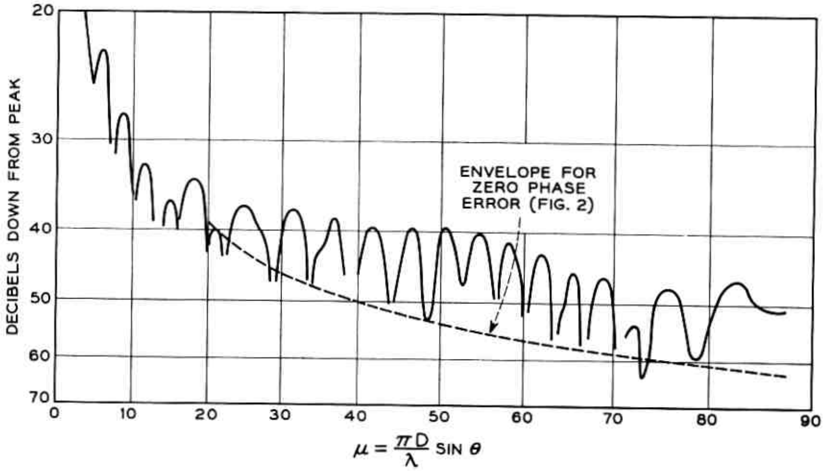


Fig. 6 — Radiation pattern: typical phase error — maximum phase error 0.155λ peak-to-peak; illumination 10-db taper.

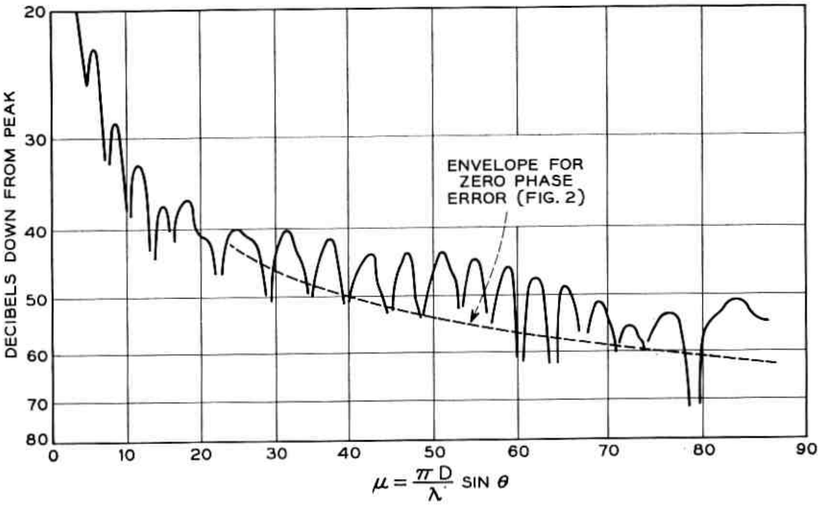


Fig. 7 — Radiation pattern: typical phase error — maximum phase error 0.1λ peak-to-peak; illumination 10-db taper.

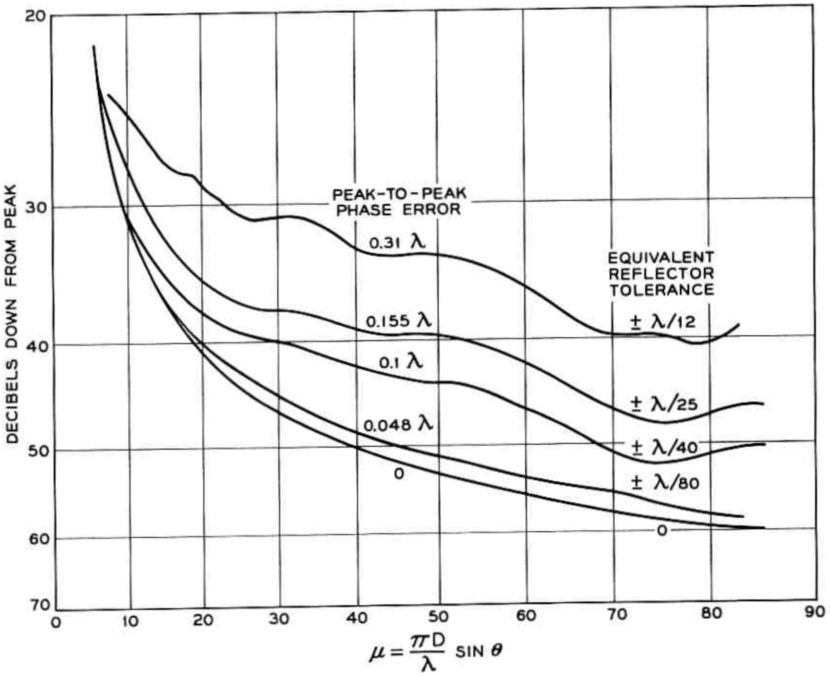


Fig. 8 — Comparison of envelopes of radiation patterns. Peak-to-peak phase errors: 0, 0.048, 0.1, 0.155, and 0.31λ; illumination 10-db taper.

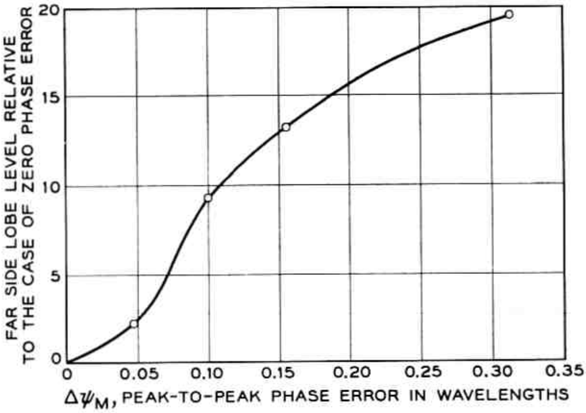


Fig. 9 — Degradation of level versus phase error for lobes at angles corresponding to

$$u = \frac{\pi D}{\lambda} \sin \theta = 50.$$

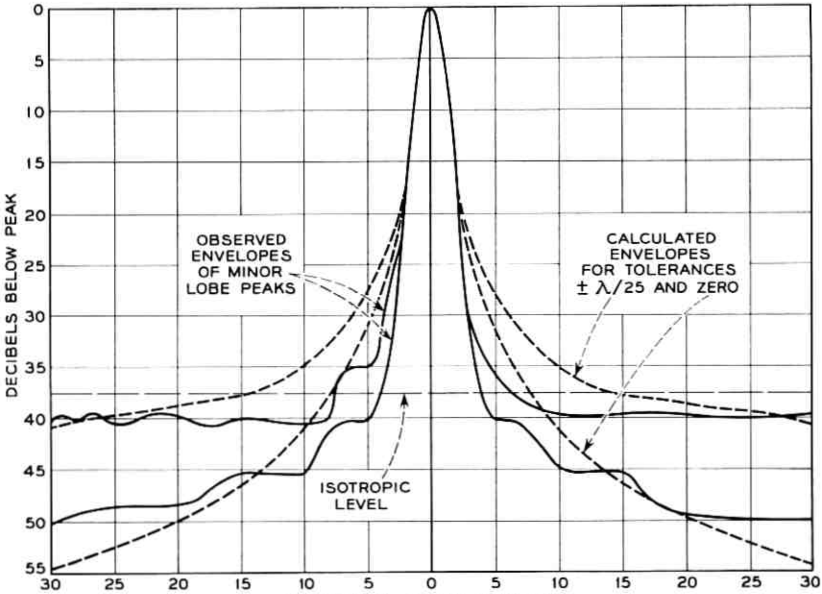


Fig. 10 — Comparison of calculated with measured patterns for precision reflector and wooden reflector with tolerance ≈λ/25.

3.3 *Comparison with Experiment*

In Fig. 10 the results of an experiment reported some years ago by H. T. Friis and W. D. Lewis[4] are shown as full curves. In that experiment the radiation patterns of two paraboloidal antennas were measured at centimeter wavelengths; one employed a searchlight mirror as a precision reflector, the other a carefully constructed metallized wooden paraboloid with the same diameter and nominal contour. The diameter of the aperture was 36 wavelengths in both cases. The data in Fig. 10 clearly show that the precision antenna has a lobe level about 10 db below that of the wooden antenna for angles $\theta = 20°$.

In Fig. 10 the measured patterns are compared with those calculated for the typical phase error of peak-to-peak value $\Delta\psi_M = 0.155\lambda$ (Fig. 6). One sees that it is possible to account for the measured increase in the side lobe level of the wooden antenna if one assumes a tolerance of about $\pm\lambda/25$ in its surface.

Of course, we realize that actual antennas, especially paraboloids, are beset with other deficiencies, such as spillover, edge currents, reradiation by feed supports, and aperture blocking, all of which give rise to wide-angle radiation. However, in the comparison experiment under discussion, these other factors are believed to be the same in the two cases.

IV. CONCLUSION

If one demands that wide-angle radiation from a reflector with quasi-random roughness be of the same order as that for a perfectly smooth surface, calculation shows that the reflector tolerance should be held to $\pm 0.01\lambda$. The calculation is verified, in part, using data obtained on paraboloidal reflectors at centimeter wavelengths. Roughness of the type described results in interference-prone microwave antennas; it also degrades their noise performance.

REFERENCES

1. Silver, S., *Microwave Antenna Theory and Design*, M.I.T. Rad. Lab. Series, 12, McGraw-Hill, New York, 1959.
2. Born, M., and Wolf, E., *Principles of Optics*, Pergamon Press, New York, 1959, p. 458.
3. Stratton, J. A., *Electromagnetic Theory*, McGraw-Hill, New York, 1941, pp. 351–364.
4. Friis, H. T., and Lewis, W. D., Radar Antennas, B.S.T.J., **26**, April, 1947, p. 265.

The TL Radio Relay System

By S. D. HATHAWAY, D. D. SAGASER and J. A. WORD

(Manuscript received May 15, 1963)

The TL radio relay system is a telephone message broadband microwave facility that operates in the 10.7- to 11.7-gc common carrier frequency band. It has been specifically designed for high reliability, low maintenance and small power consumption by the exclusive use of solid-state circuitry except for the transmitter and local oscillator klystron tubes. Transmission performance and over-all system description are presented, as well as a description of some early field applications.

CONTENTS

Page

I. INTRODUCTION 2298
II. OBJECTIVES 2298
2.1 Applications 2298
2.2 Development Objectives 2299
2.3 Transmission Objectives 2300
III. TRANSMISSION PLAN 2301
IV. SYSTEM DESCRIPTION 2303
4.1 General 2303
4.2 Radio Transmitter-Receiver 2307
4.3 Diversity Switching Arrangements 2320
4.4 Order Wire and Alarm 2327
4.5 Power System 2331
4.6 System Characteristics 2337
V. EQUIPMENT FEATURES 2340
5.1 General 2340
5.2 Transportable Equipment Shelter 2344
5.3 Towers 2346
5.4 Waveguide Moisture Problems 2348
VI. MAINTENANCE AND TEST EQUIPMENT 2348
6.1 General 2348
6.2 Built-in Test Features 2348
6.3 Test Sets 2349
6.4 Spare Equipment 2351
VII. APPLICATIONS OF TL RADIO SYSTEMS 2351
7.1 Billings-Hardin System 2351
7.2 Denver-Limon-Burlington System 2353
VIII. ACKNOWLEDGMENT 2353

I. INTRODUCTION

In the past, short-haul microwave radio relay development in the Bell System has progressed along the general lines of heavy-route, cross-country systems. The use of high towers on substantial plots of land located on prominent sites often requiring expensive access roads and power lines was typical. Buildings were sufficiently large to accommodate not only a large number of radio equipment bays but also 5- to 10-kilowatt rotating machinery to provide backup in the event of power failure. Radio equipment costs seldom approached one-half of the total costs of such repeaters.

Today, ideas originally proposed by the Radio Research Department of Bell Telephone Laboratories to reduce over-all microwave station costs have become a reality.[1] The TL radio relay is a new microwave system designed specifically for short-haul service. Low equipment costs, minimum engineering and installation effort, outdoor housing, low-cost antenna support structures, small power consumption, and moderate telephone circuit capacity all contribute to substantially lower over-all TL radio station costs. A high degree of reliability is provided by extensive use of solid-state circuitry, float-charged reserve battery power supply system, and simplified maintenance procedures.

II. OBJECTIVES

2.1 *Applications*

There are a great number of uses for a flexible, economical and reliable short-haul microwave system such as TL. It is intended to have applications in three broad areas: namely, short-haul toll or tributary trunk routes, supplementary circuits along heavier routes, and special service telephone routes. Nondiversity TL can provide economical relief of overburdened open-wire and cable routes. Additional possible applications of TL radio are: (1) open wire replacement, (2) difficult geographical situations, (3) suburban industrial areas, (4) metropolitan and intersuburban trunk groups, (5) routes for the military services, (6) alternate routing, (7) broadband or high-speed data services, (8) short-term or seasonal services, (9) emergency service restoration, (10) spurs on other routes, and (11) order wire and alarm circuits for existing heavy-route microwave systems.

2.2 *Development Objectives*

Concurrent with the beginning of the TL development program, a set of objectives was specified reflecting the best judgment at that time as to the features and capabilities necessary for the new system. These objectives have been reviewed and modified from time to time as the development of the new system progressed and, in most instances, the original requirements for the TL system have been met. The latest change in the requirement, for TL to transmit TV, is not reflected in TL at this time but will appear in later versions of this equipment.

2.2.1 *Frequency Band and Allocation Plan*

The TL system operates in the 11-gc common carrier frequency band and provides telephone message service for end-link or short-haul application. The TJ frequency allocation plan has been adopted, which permits the operation of six two-way TL channels on the same route with common antennas for transmitting and receiving. One-for-one frequency diversity has been provided on an optional basis.

2.2.2 *Telephone Message Capacity*

The TL system has been engineered to transmit 48 channels of N carrier, 96 channels of ON carrier or 300 channels of L carrier multiplex over 10 hops for a total distance of about 200 miles.

2.2.3 *DC Power Drain*

The TL system employs all solid-state circuitry except for the transmitter and receiver local oscillator klystrons, and requires 170 watts of power. The continuously charged battery power supply has sufficient reserve to carry the system for 20 hours under average ambient conditions in the event the ac power fails.

2.2.4 *Order Wire and Alarm*

A simple but effective order wire and alarm system has been provided for the TL system which will operate over the radio. This saving of space and equipment cost has been a substantial factor in the economy of the system.

2.2.5 *Economics*

One of the primary objectives of this development has been to provide system arrangements and operating features at the lowest first cost and annual charges consistent with meeting Bell System transmission requirements. Particular emphasis has been placed on equipment arrangements to minimize job engineering and installation expense. Suitable outdoor housing has been provided with its attendant savings in building and land costs. Ease of maintenance is of prime importance in its relation to annual charges, and equipment arrangements have been devised with this in view.

2.3 *Transmission Objectives*

2.3.1 *General*

The TL hops should be engineered to have a large carrier-to-noise ratio during periods of free-space transmission in order to provide adequate margin over first circuit noise during periods of signal attenuation caused by propagation, especially rainfall. Reliable protection against selective fades and equipment failure outage can be obtained with frequency diversity. Protection against rain attenuation can be assured by engineering sufficient fading margin into the system by using path lengths appropriate to the particular area of the country. Path lengths will range from 10 miles or less in the heavy rain areas to 25 to 30 miles in the dry areas.[2]

2.3.2 *Telephone*

The TL system should be engineered to meet short-haul toll circuit noise objectives of 31 dba at the 0-db transmission level point for 10 hops. This 31 dba includes both radio and multiplex terminal noise. Of this noise, 30 dba is assigned to the radio circuit and 24 dba allocated to the multiplex. Shorter systems should be engineered to correspondingly tighter over-all requirements so that if extended later to 10 hops, they would meet the 31-dba objective. Single-hop performance can be derived from the knowledge of how intermodulation and fluctuation noise add as the number of repeaters is increased. For the TL system, the total noise power increases proportionally with the number of repeaters.

2.3.3 *Stability*

The objective for the short-term net loss variations in a telephone channel is less than ±0.25 db, and the long-term net loss variations should not normally exceed ±1.5 db. These limits are necessary for the system to meet direct distance dialing and other similar Bell System requirements.

III. TRANSMISSION PLAN

The TL radio system offers a maximum of six two-way broadband channels. For radio systems paralleling other communication facilities or transmitting only a modest number of telephone circuits, TL may be used on a nondiversity basis. However, to provide the high degree of reliability needed for systems carrying large numbers of telephone circuits, one-for-one frequency diversity protection may be used, with automatic switching at each repeater offering a maximum of three two-way broadband channels.

The radio signals are transmitted to a dual polarized antenna by RF channelizing and duplexing arrangements. For repeater locations requiring high towers, it is expected that most systems will use a "periscope" type of antenna arrangement to minimize the loss associated with long waveguide runs. To meet these needs, five- and ten-foot paraboloidal antennas and 6 × 8, 8 × 12 and 10 × 15-foot reflectors are available. The paraboloidal antennas may also be used alone as direct radiators in those systems employing short towers on natural elevations.

A block schematic of a two-section, nondiversity TL system is shown in Fig. 1. The multiplex and control signals are combined through high-pass – low-pass filters and transmitted on the radio channel. At the receiver, similar filter arrangements separate the multiplex from the control signals. A 2600-cps pilot, continuously transmitted over the radio system, is used to determine the alarm status of the various repeater and terminal stations.

The TL frequency plan is shown in Fig. 2. Because of the expected use of the "periscope" antenna system, the plan is based on the use of four frequencies for each two-way radio channel. The 10.7- to 11.7-gc common carrier band is divided into 24 channels, each about 40 mc wide. In a given repeater section, only 12 of these are used, resulting in 80-mc spacing between midchannel frequencies. These channels are further divided into two groups of six for transmission in each direction. Polarization of the channels alternates between vertical and horizontal to provide 160-mc separation between signals having the same polarization, thereby substantially easing requirements on the channel-separation networks. The remaining 12 channel assignments are used in adjacent repeater sections. These frequencies are repeated in alternate hops. Potential "overreach" interference is reduced by reversing the polarization of the third section with respect to the first section. Co-channel interference from adjacent repeater stations is eliminated by the use of the four-frequency plan. At a given repeater, adequate frequency separation between transmitters and receivers is achieved by using the upper half of the band for transmitting and the lower half for receiving. This arrangement is inverted in adjacent sections.

Actual TL route cross sections may vary from a single two-way,

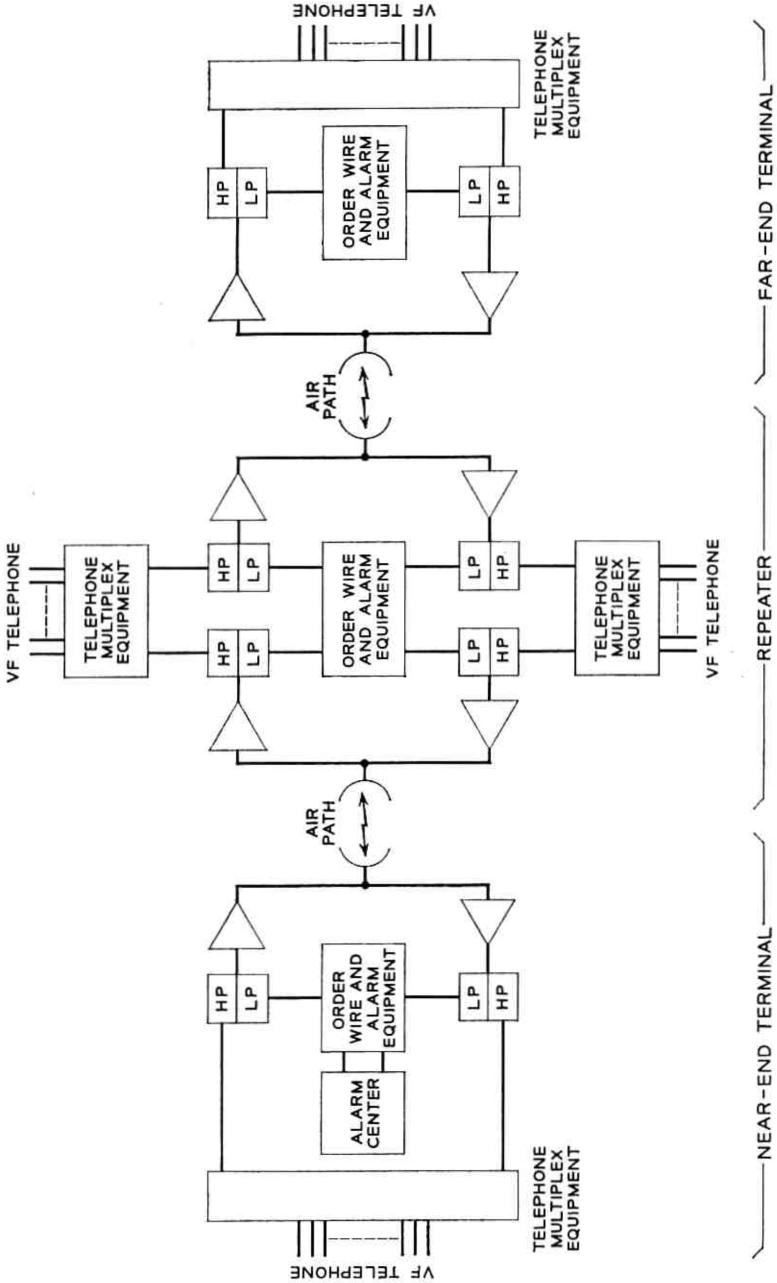
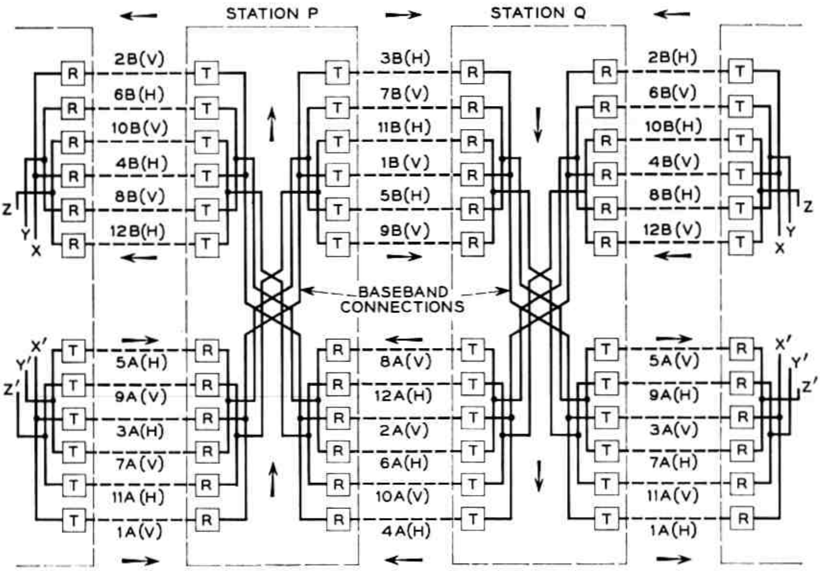


Fig. 1 — Block schematic of two-section TL system.



Channel Number	Transmitter Frequency, kmc	Beat Oscillator Frequency, kmc	Channel Number	Transmitter Frequency, kmc	Beat Oscillator Frequency, kmc
4A	10.715	10.785	9B	11.245	11.315
1A	10.755	10.825	12B	11.285	11.355
10A	10.795	10.865	5B	11.325	11.395
11A	10.835	10.905	8B	11.365	11.435
6A	10.875	10.945	1B	11.405	11.475
7A	10.915	10.985	4B	11.445	11.515
2A	10.955	10.885	11B	11.485	11.415
3A	10.995	10.925	10B	11.525	11.455
12A	11.035	10.965	7B	11.565	11.495
9A	11.075	11.005	6B	11.605	11.535
8A	11.115	10.045	3B	11.645	11.575
5A	11.155	11.085	2B	11.685	11.615

Fig. 2 — TL frequency allocation plan.

nondiversity channel up to a full system of three protected two-way channels. Additional radio channels may be added in the future to a diversity system whose initial requirements are less than its maximum capabilities without disrupting service on the working channels.

IV. SYSTEM DESCRIPTION

4.1 *General*

The basic unit of the TL system is the transmitter-receiver bay. It consists of a radio transmitter-receiver panel, an order wire and alarm

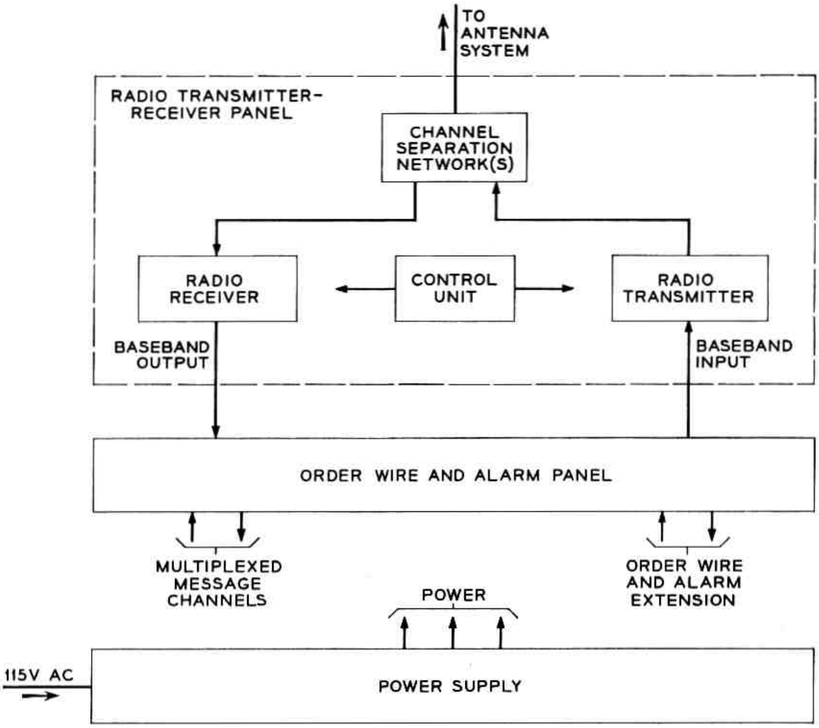

TO ANTENNA SYSTEM
RADIO TRANSMITTER-RECEIVER PANEL
CHANNEL SEPARATION NETWORK(S)
RADIO RECEIVER
CONTROL UNIT
RADIO TRANSMITTER
BASEBAND OUTPUT
BASEBAND INPUT
ORDER WIRE AND ALARM PANEL
MULTIPLEXED MESSAGE CHANNELS
POWER
ORDER WIRE AND ALARM EXTENSION
115V AC
POWER SUPPLY

Fig. 3 — TL bay block diagram.

panel or a diversity switch panel, a power supply panel, and batteries. This equipment is mounted on relay-rack type framework for indoor installations or in a weather-proof cabinet for outdoor installations. With appropriate antenna systems, this basic unit forms a nondiversity terminal; two units may form either a nondiversity repeater, or a diversity terminal, etc., with only minor variations in the make-up of the bay. A block diagram of the basic bay is shown in Fig. 3. Fig. 4 illustrates how the basic bays may be interconnected at microwave frequencies by means of channel separation networks and a polarizer to form the minimum and maximum capacity terminals. A repeater would consist, in effect, of two terminals back-to-back with the receiver output from one bay feeding the transmitter of the other bay at baseband frequencies. These configurations are used for either frequency diversity or nondiversity systems. In diversity systems, one bay of a diversity pair would be on one polarization, the other on the opposite polarization.

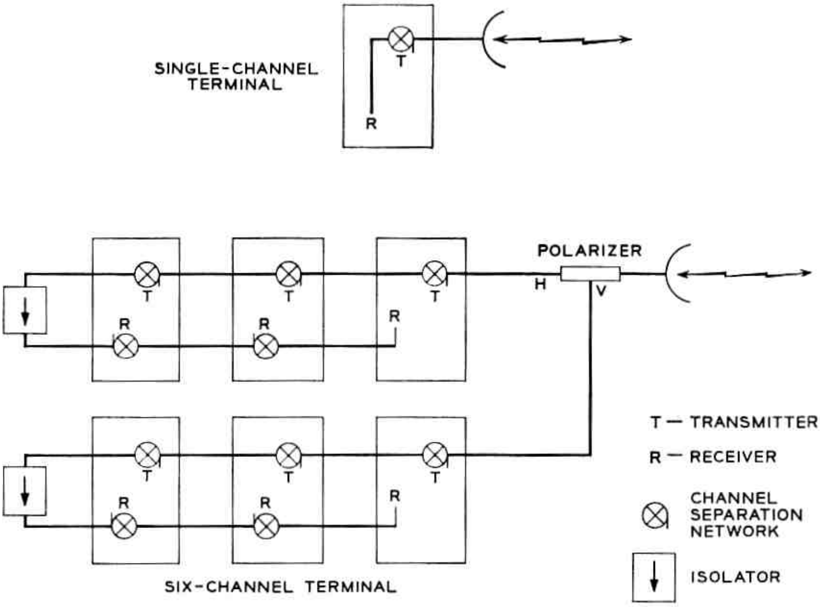


Fig. 4 — TL bay waveguide interconnection diagram

This permits waveguide maintenance on one group of bays without service interruption on the other. The bays would be identical except that one would contain a diversity switch; the other, the order wire and alarm panel for that channel.

To explain the operation of the basic bay in general terms, consideration will be given to the middle bay on vertical polarization of the six-channel terminal of Fig. 4. The vertically polarized received signal would be transmitted from the antenna essentially unattenuated through the following: (1) the port of the polarizer aligned with vertical polarization; (2) the three channel separation networks connecting the lower transmitters to the antenna system; (3) the isolator (purpose to be described later); (4) one receiving channel separation network; and (5) the path to the receiver in the selected bay through the receiving channel-separation network, it being tuned to this particular frequency. (The principle of operation of the channel separation networks has been described elsewhere.[3] Representative transmission characteristics of this waveguide network are given in Fig. 5).

In the radio receiver selected, the received signal is filtered and heterodyned with the output from a local oscillator to produce an intermediate

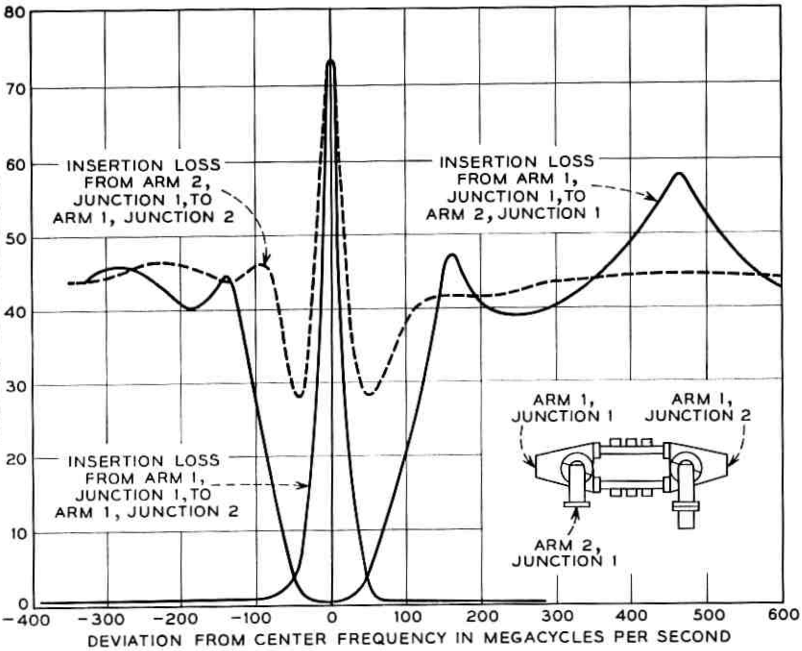


Fig. 5 — Transmission characteristic of channel separating networks.

frequency (IF) of 70 mc. The IF signal is amplified and detected to yield the original baseband intelligence, which is then further amplified. In a diversity system, the signal is then fed to the diversity switch, which selects either this signal or the one from the other receiver of a diversity pair to be applied to the order wire and alarm panel. In this panel, the baseband is split into two parts by high-pass–low-pass filters, the high frequencies being used for the multiplexed message channels and the low frequencies for order wire and alarm purposes. At a terminal, these portions of the baseband are applied to appropriate terminal equipment. At a repeater, any dropping or adding of message circuits that is desired is done at this point, and the order wire and alarm operations are performed. The two portions of the baseband are then recombined with another set of high-pass–low-pass filters and supplied to the transmitter, or transmitters for a diversity system, via a splitting pad.

A transmitter baseband amplifier increases the signal voltage to be applied to the repeller of the transmitting klystron. The resulting frequency-modulated RF signal is combined with outputs of the other transmitters by means of channel separation networks and connected to the antenna via the polarizer.

The baseband-type of repeater just described is especially useful and economical in short-haul microwave systems. Message circuits are frequently dropped at repeater points. Having the message multiplex frequencies available without requiring special terminal equipment, as would be required in an IF-type repeater, is economically advantageous. It also facilitates the order wire and alarm appearances.

4.2 *Radio Transmitter-Receiver*

The radio transmitter-receiver panel consists of the frequency-modulated (FM) transmitter, a heterodyne-type FM receiver, a control unit to provide certain metering and adjustment features, and one or more channel separation networks appropriate to the application of the bay. The radio transmitter section of the panel is shown in the more detailed block diagram of Fig. 6. The transmitter baseband amplifier is a three-stage feedback amplifier using Western Electric 15C germanium diffused-base transistors. It provides a nominal voltage gain of 31 db from the 75-ohm unbalanced input to the high impedance of the klystron repeller, supplying a maximum voltage of eight volts peak-to-peak required to modulate the klystron ±6 mc from rest frequency. Adjustable over a gain range of ±4 db to accommodate the modulation sensitivity of all klystrons, the amplifier has a frequency characteristic flat to ±0.4 db from about 100 cycles to 6 mc. A photo of the amplifier and a typical characteristic are shown in Fig. 7.

The transmitter output is obtained from a Western Electric 457A

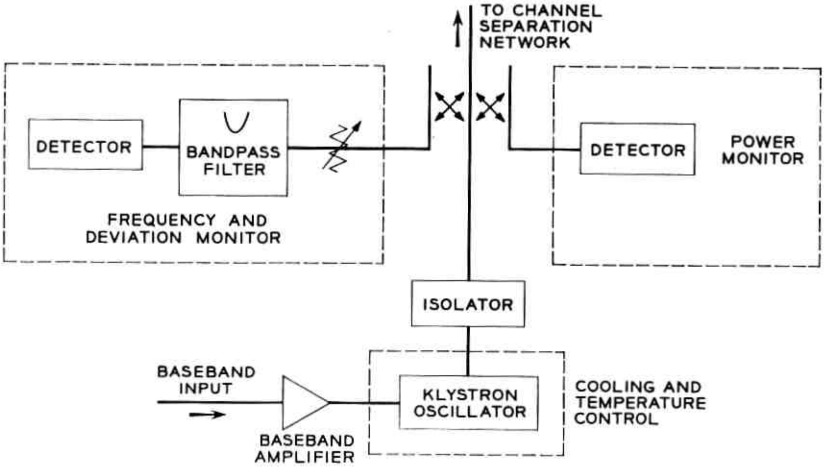


Fig. 6 — Transmitter block schematic.

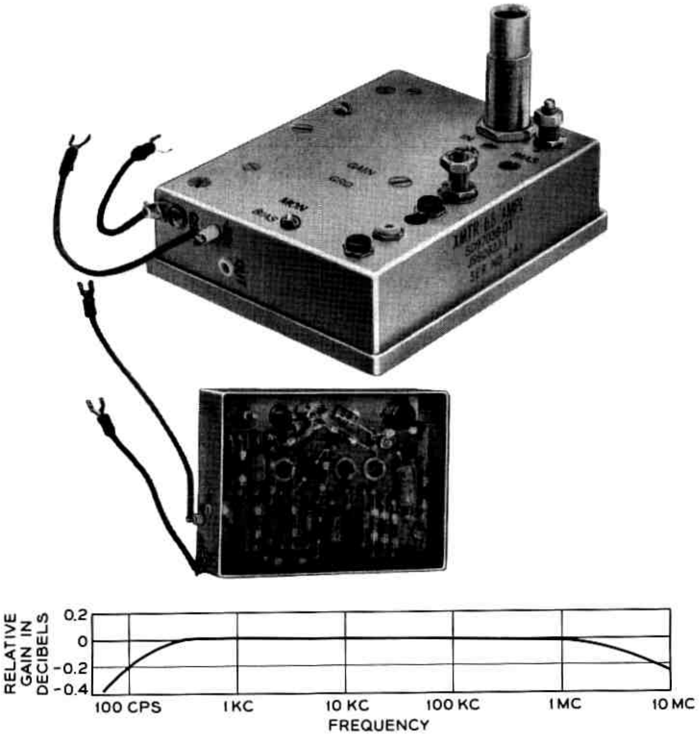


Fig. 7 — Transmitter baseband amplifier.

klystron oscillator, which is illustrated in Fig. 8. This tube was developed specifically for the TL system with special emphasis in the design on obtaining long life and a low frequency-vs-temperature coefficient.[4] The same tube is used for the receiver local oscillator. Since both tubes are operated in the $3\frac{3}{4}$ mode, only one set of voltages is required from the power supply. Typical operating characteristics are summarized in Table I.

The average dependence of the power output of the tube upon its operating frequency is shown in Fig. 8, from data on a typical tube.

The desired frequency stability for the transmitter and receiver is achieved by controlling three important parameters of klystron operation: (1) the frequency-temperature coefficient of the tube; (2) the electrode voltages; and (3) the klystron temperature environment. The first is determined by the design of the tube itself. A low coefficient of 0.15 mc/°F or less has been achieved. The second is accomplished by the design of an extremely stable power supply, aided by the fact that

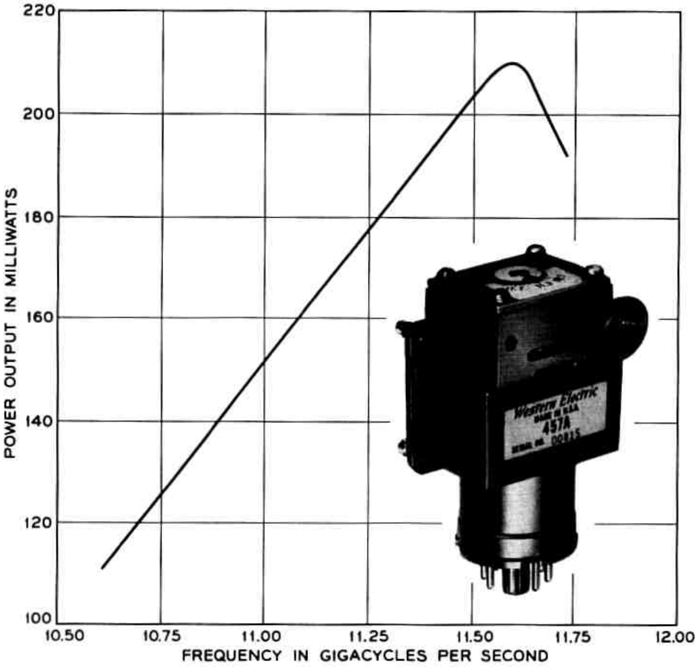


Fig. 8 — 457A klystron.

the voltage-frequency coefficients of the repeller and resonator of the tube are to a large extent canceling and that the small variation of the −600 and −400-volt outputs of the power supply, as a function of temperature, are in the same direction.

The third factor is controlled by an extremely well performing, yet

TABLE I — WE 457A KLYSTRON: TYPICAL OPERATING CONDITIONS

RF power output	100 milliwatts (minimum)
Resonator voltage	400 volts
Resonator current	40 ma
Repeller voltage	115 volts
Repeller modulating sensitivity	1.5 mc/volt (minimum)
Electronic tuning range	116 mc at 10.7 gc 80 mc at 11.7 gc
Mechanical tuning range	10.7 - 11.7 gc
Repeller capacity	2.5 μμf, typical
Mechanical tuning sensitivity	1/3 mc/angular degree
Output	matched to WR90 waveguide
Heater current	0.9 amp
Heater voltage	6.3 volts
Oscillating mode	$3\frac{3}{4}$
Frequency-temperature coefficient	<±0.15 mc/°F, mid-band
Anticipated life	>40,000 hours

simple and economical, cooling system called the "vapor phase cooler" (VPC). The system operates on the physical principle that a liquid boils at a constant temperature at a given pressure. If the heat input to the liquid increases or decreases, it simply boils more or less vigorously, but at the same temperature. In the TL VPC, the liquid is a fluorochemical which boils at approximately 214°F at sea level. It is contained in a small copper boiler designed to permit vigorous boiling of the liquid without restricting the flow of vapor out of the boiler into the condensing system or of the condensed liquid in the opposite direction. The two klystrons are clamped to the sides of the boiler for good heat transfer, as shown in Fig. 9. The heat input to the boiler is obtained from the dissipation in the klystrons, which is sufficient to keep the liquid boiling even at outside temperatures below −40°F. This is insured by enclosing boiler and klystrons in a well-insulated box. Heat transfer by conduction away from the boiler is minimized by using stainless steel tubing having low heat conductivity for connection between the boiler and the condensing system.

The condenser consists of a copper tube clamped to the aluminum panel on which the radio transmitter and receiver equipment are mounted. It thus has a large heat sink for dissipation of the heat released during condensation of the fluorochemical. The copper tube is terminated in a flexible neoprene compound bag, especially formulated to be resistant to passage of the fluorochemical gas through its walls and to remain flexible for many years without drying out with consequent cracking. The flexible bag permits substantial changes in volume of the enclosed system without appreciable changes in pressure. This satisfies the fundamental condition for a constant boiling temperature, even though there may be considerable changes in the heat input to the boiler. These changes are caused primarily by large variations in the ambient temperature.

The performance of the tube, power supply, and VPC system in maintaining good over-all frequency stability as a function of ambient temperature is illustrated in Fig. 10. This figure also shows the performance of the VPC system alone in stabilizing the klystron temperature. These data were taken on a run when the entire transmitter-receiver bay was exposed to the indicated temperature. Changes in barometric pressure of 2 inches of mercury would cause corresponding changes of operating temperatures of approximately 3.5°F.

The body of the klystron, being clamped to the boiler, is operated at ground potential. This minimizes exposure to high voltage on the part of operating personnel and eliminates the necessity for protective interlock switches. Precautions against klystron damage by positive repeller-

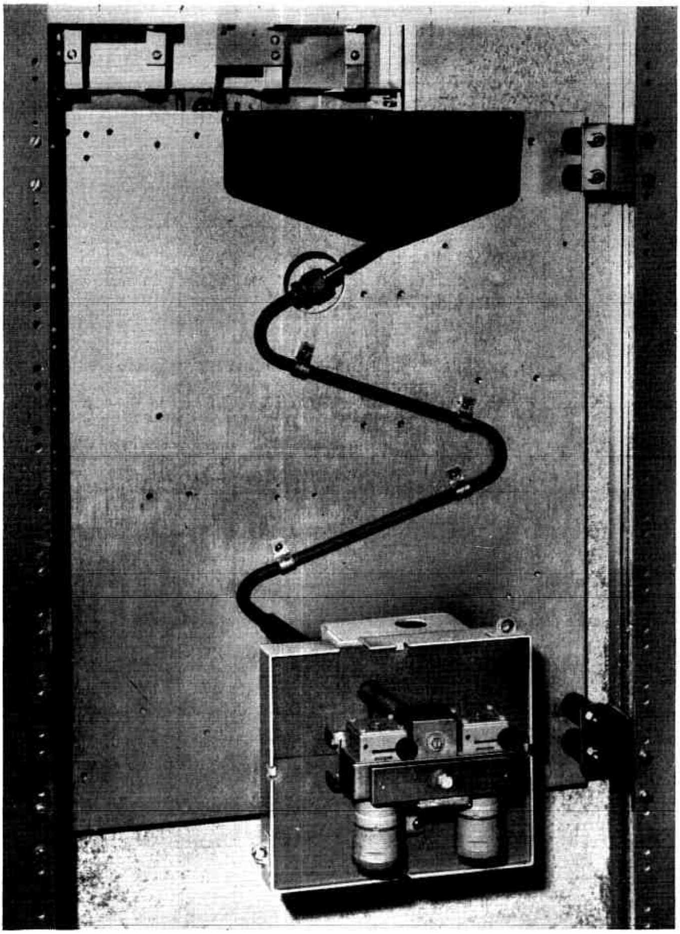


Fig. 9 — Klystrons and vapor phase cooling system.

to-cathode voltage have been included in both the transmitter and receiver klystrons in the form of clamping diodes between these electrodes.

The output of the transmitter klystron is first fed through a Western Electric 1B isolator, shown in Fig. 11. This high-performance field displacement type ferrite device practically eliminates any effect on linearity of frequency-changes vs repeller-voltage-changes caused by reflections in the antenna feed. Performance characteristics of the isolator are shown in Fig. 12.

Connected to the output of the isolator is a 20-db double directional

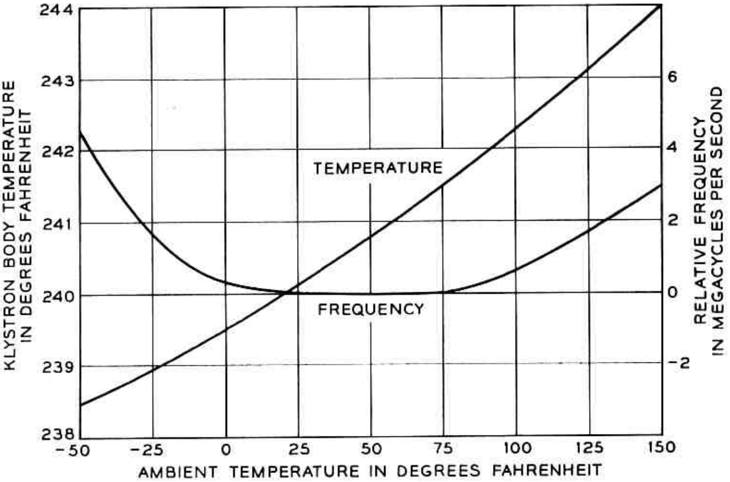


Fig. 10 — Frequency stability of 457A klystron, power supply and VPC vs ambient temperature.

coupler which gives two samples of the transmitted energy for monitoring purposes. One sample is immediately detected for power monitoring; the other is transmitted through a calibrated attenuator, high-Q cavity filter, and then to a detector for frequency monitoring and deviation adjustment purposes. This latter function will be described in the section on maintenance and test equipment.

From the output of the directional couplers, the transmitted signal is applied to the antenna system via the channel separation networks previously described.

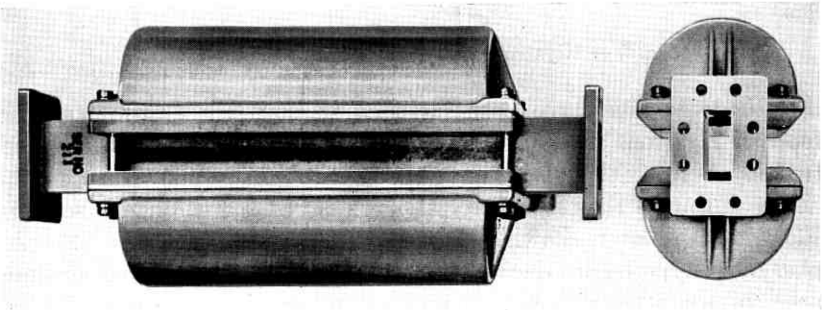


Fig. 11 — Isolator.

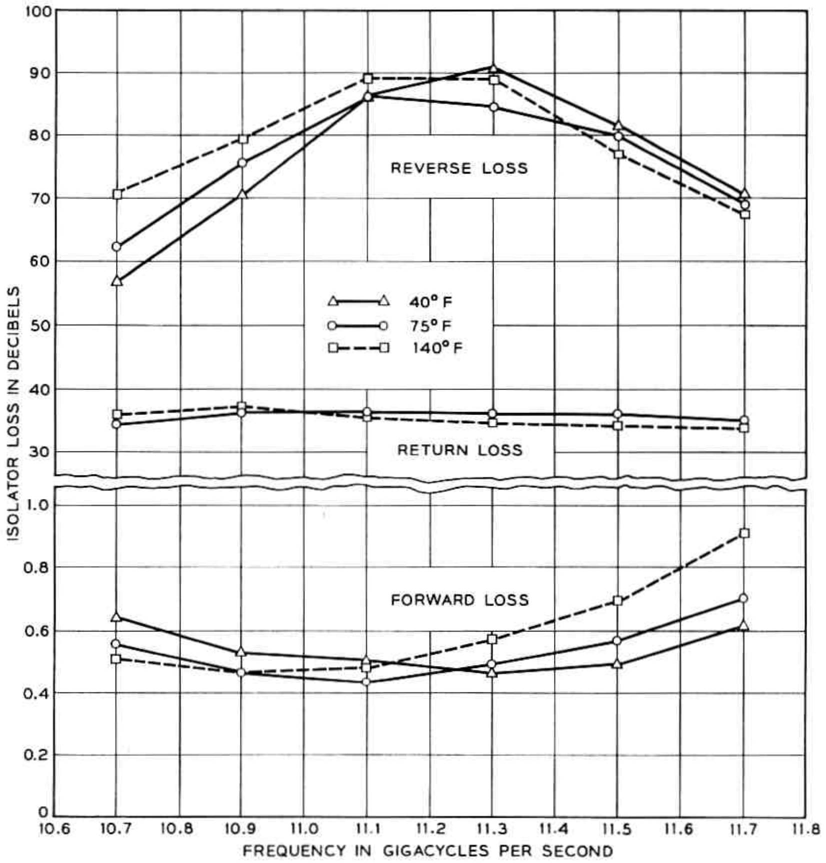


Fig. 12 — Forward and reverse loss characteristic of the 1B isolator.

The radio receiver section of the transmitter-receiver panel is shown in Fig. 13. The incoming RF signal from the antenna is selected and routed by the channel separation networks to the proper receiving modulator through a bandpass filter, a waveguide tuner, and a waveguide spacer. The filter provides attenuation to interfering out-of-band signals, and improves the noise figure of the modulator by reflecting out-of-band modulation products back into the converter in the proper phase. The proper phase relationship is maintained at the different channel frequencies by choosing a suitably dimensioned waveguide spacer, which determines the electrical path length traversed by the modulation products between the converter and filter. The modulator input im-

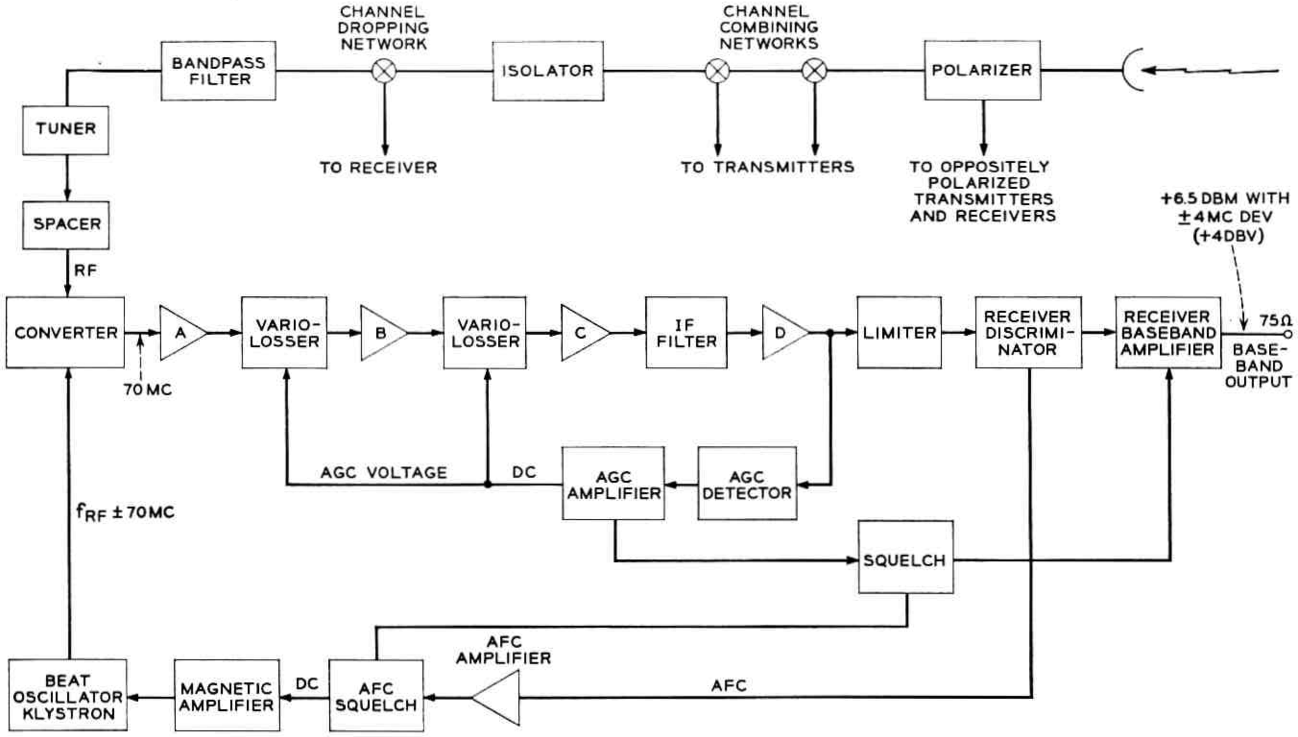


Fig. 13 — Receiver block diagram.

pedance must be closely matched to the waveguide impedance to minimize reflections between the bandpass filter and modulator input. This is achieved by the adjustable two-stub tuner located between the bandpass filter and the converter.

Two types of bandpass filters are used, one having three, and the other four, resonant cavities. As shown in Fig. 4, the last receiver in a bay line-up does not require a channel dropping network, since at this point the number of RF channels has been reduced to one. In this last receiver, the out-of-band attenuation not provided by the channel separation network is obtained by using a four — rather than a three — cavity bandpass filter. Typical transmission characteristics of the two types of filters are shown in Fig. 14.

The modulator is a balanced hybrid junction assembly having diodes in the junction reversed with respect to each other. The balanced structure greatly reduces noise from the local oscillator; the reversed diodes permit paralleled unbalanced output connections, giving the desired 75-ohm unbalanced coaxial output at 70 mc. Waveguide inputs are provided for the incoming signal and the local oscillator input, as shown in the schematic of Fig. 15.

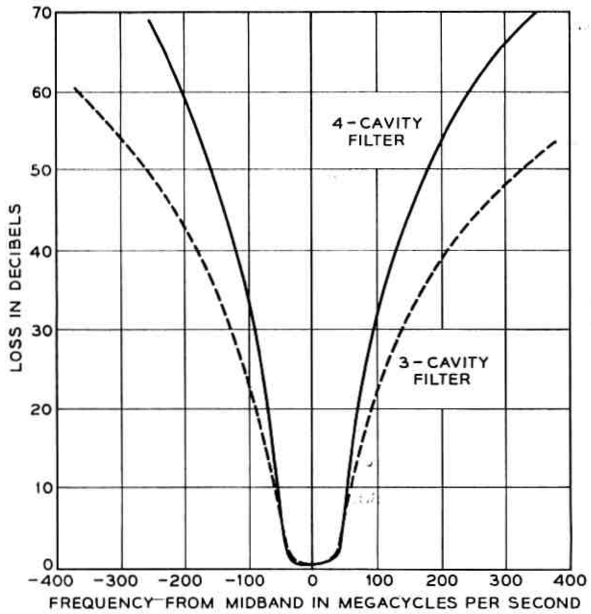


Fig. 14 — Transmission characteristic of receiving bandpass filter.

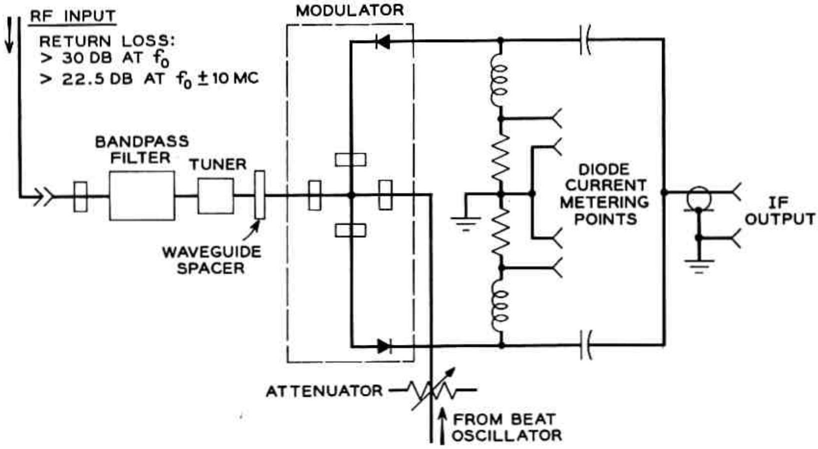


Fig. 15 — Receiving modulator schematic.

As mentioned above, the local oscillator is a Western Electric 457A klystron operated in the same fashion as the transmitting klystron. It feeds the modulator through an adjustable attenuator and two waveguide-to-coaxial cable transducers for mechanical convenience. The attenuator permits adjusting the power input to the modulator to approximately 0 dbm. The hybrid balance of the modulator limits the local oscillator leakage back towards the bandpass filter to −25 dbm. The frequency plan is such that this leakage causes no interference in the receivers on the same polarization (see Fig. 4). To avoid beating oscillator interference with receivers on the opposite polarization, a 1B isolator is placed in the waveguide path of a four- or six-bay line-up to sufficiently reduce the level of the leakage from the beating oscillator. The frequency of the beating oscillator is maintained at 70 mc from the incoming signal by means of the AFC circuit described below.

The 70-mc, frequency-modulated output from the modulator is first amplified and regulated in level in the preamplifier section of the IF and baseband unit.[5] The preamplifier consists of three sections of broadband transistor amplifier stages with two diode variolossers dividing the three sections. The variolossers are controlled to maintain a nearly constant input to a delay-equalized bandpass filter separating the preamplifier from the main IF amplifier. The filter limits the bandwidth of the receiver to 20 mc at the 3-db down points. The main amplifier increases the signal level sufficiently to be limited by a Ruthroff-type limiter[6] for amplitude modulation suppression. The signal is then detected by the discriminator and is amplified by the receiver baseband amplifier. The gain of the receiver baseband amplifier, adjustable by approximately ±2.5 db, is such as to deliver +10 dbm to the 75-ohm unbalanced load with a frequency deviation of ±6 mc.

Two dc amplifiers and a squelch circuit are also included in the IF and baseband unit. One is the AGC amplifier, which applies the detected and differentially compared output of the main IF amplifier to the variolossers; the second is the AFC amplifier which supplies the dc output of the discriminator (a measure of the position of the incoming carrier in the IF band) to other circuitry for control of the receiver beating oscillator. The output of the AGC amplifier is a measure of the received signal strength. This voltage is used to actuate the comparator circuit of the diversity switch in a diversity system. The squelch circuit operates from the AGC output and biases off the first stage of the receiver baseband amplifier when the receiver input level falls below a predetermined low level. This avoids the possibility of the high baseband output noise, characteristic of an FM receiver with no input, from interfering with other aspects of system operation, such as interfering with other message circuits that might be introduced into the system at some subsequent repeater. The level at which the squelch operates is adjustable, but is normally set to operate at a receiver RF input of −83 dbm.

Automatic frequency control of the receiver beating oscillator is achieved by first applying the AFC amplifier output to the control winding of a magnetic amplifier, as shown in Fig. 16. The power input to the magnetic amplifier is an 1800-cps square wave obtained from the power supply. The output is rectified and applied in series with the nominal −200 volts, supplied by the power supply between the resonator and

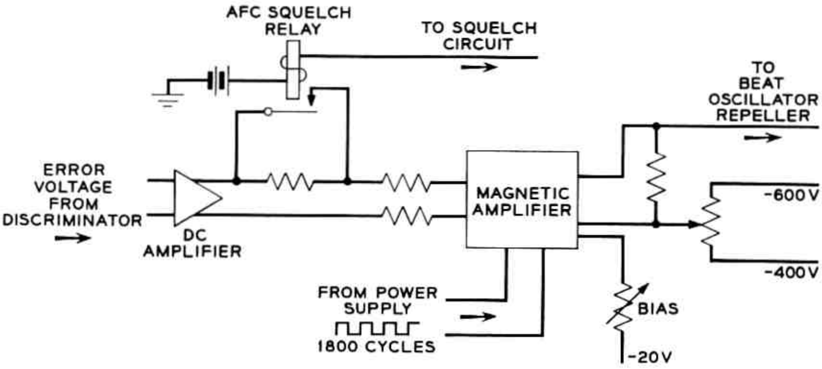


Fig. 16 — Receiver AFC functional schematic.

the repeller of the klystron, in proper phase to form a stable feedback loop and maintain the intermediate frequency essentially at the discriminator crossover frequency. If the crossover frequency is not at exactly 70 mc, an adjustable bias in the AFC amplifier input corrects for the slight error. The magnetic amplifier is particularly applicable here because it readily provides the isolation required between the approximately −600 volts of the repeller circuit and the low voltage of the AFC amplifier, and at the same time provides dc amplification. The magnetic amplifier has a nominal transfer impedance of one megohm and has an overload characteristic designed to act as a clamp on the AFC. This limits the output voltage to a value such that the receiver beating oscillator frequency cannot be changed to such an extent that the receiver can lock onto signals of the adjacent channel. The AFC loop has a gain of approximately 32 db, thus reducing a potential frequency change of, say, 5 mc caused by transmitting and receiver beating oscillator klystron changes, to an actual change in the receiver IF of 125 kc. The characteristic of the AFC loop from discriminator output to beating oscillator repeller voltage is shown in Fig. 17. Since the frequency of the beating oscillator may be below or above the incoming signal frequency, depending upon the particular channel, a phase reversal must be available in the AFC loop to adjust for the condition

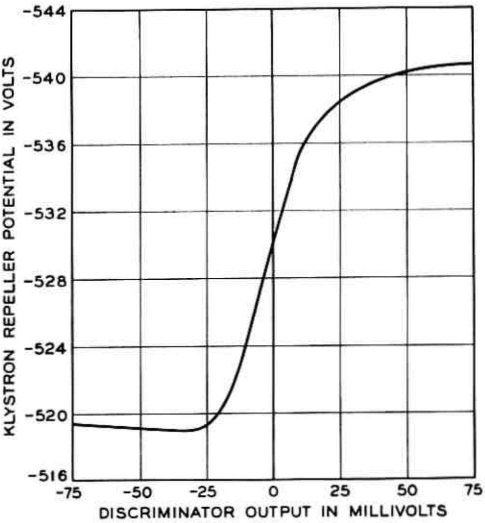


Fig. 17 — Characteristic of AFC loop.

that obtains. This is made by an optional turnover in the wiring of the balanced input to the magnetic amplifier.

When the receiver loses its input signal for any reason, the preamplifier gain goes to its maximum value as the vario-lossers are driven to their minimum loss condition by the AGC amplifier. This causes a high noise output from the discriminator. The dc output from the discriminator in this condition may not be exactly zero, since the noise or the discriminator may be slightly unbalanced. Furthermore, a correction for the crossover frequency not being at exactly 70 mc may have been introduced by the adjustable bias in the AFC amplifier input. To avoid having the resulting residual dc output from the AFC amplifier swing the beating oscillator frequency away from its normal rest frequency, an AFC squelch relay is introduced. This relay is controlled by the previously described squelch circuit. Normally operated (when a received signal is present), the relay releases upon loss of signal and inserts a large series resistor at the input to the magnetic amplifier, greatly reducing the AFC loop gain. This permits the beating oscillator to remain fairly close to its normal rest frequency, so that the signal to the IF amplifier will be within the IF passband when it reappears and will be captured by the AFC circuit. At the same time, some AFC control is retained so that system operation will not be seriously affected should the squelch circuit or relay malfunction.

The IF and baseband unit is shown in Fig. 18. It is made up of four subassemblies: (1) the preamplifier section; (2) the filter section; (3) the IF main amplifier section; and (4) the limiter, discriminator, and baseband amplifier section. Each subassembly, H shaped in cross section, has the transmission elements in one side of the H and the dc supply section on the opposite side. Power is fed to the transmission section through feed-through capacitors. The signal progresses back and forth through the four sections; maximum shielding is obtained by placing transmission components on opposite sides of adjacent sections. The complete unit is easily removable for maintenance. All connections are through plugs and jacks.

The third major component of the transmitter-receiver panel is the control unit. This plug-in unit (see Fig. 19) contains the components required for controlling the transmitter and receiver, testing and monitoring important operating parameters, and powering the klystrons. It also contains the magnetic amplifier and squelch relay of the receiver AFC circuit. A detailed description of the monitoring and control features will be found in the section on maintenance and test equipment.

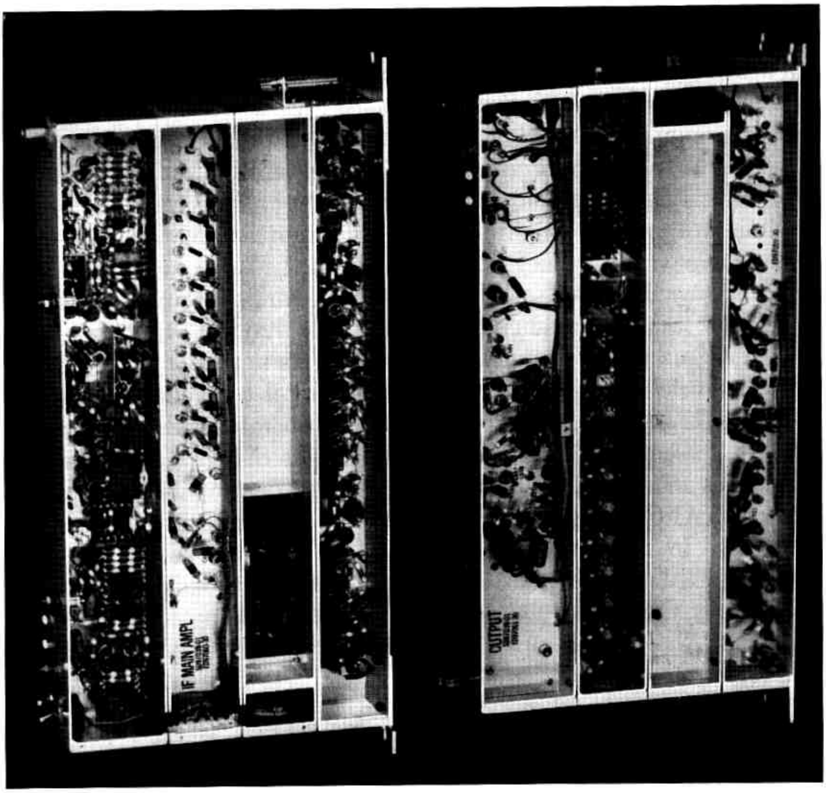


Fig. 18 — IF and baseband unit.

4.3 *Diversity Switching Arrangements*

4.3.1 *General*

It is expected that there will be many applications of TL radio which take advantage of the most economical arrangement possible, that of the nondiversity system. However, where the additional reliability is required, a one-for-one frequency diversity system provides protection against service interruptions caused by multipath fading and equipment failures. In addition, alternate facilities are then available during maintenance periods. With diversity, these periods can be scheduled at convenient times.

The unit which selects the better of the two radio channels for diversity operation is the diversity switch panel (Fig. 20). At each repeater

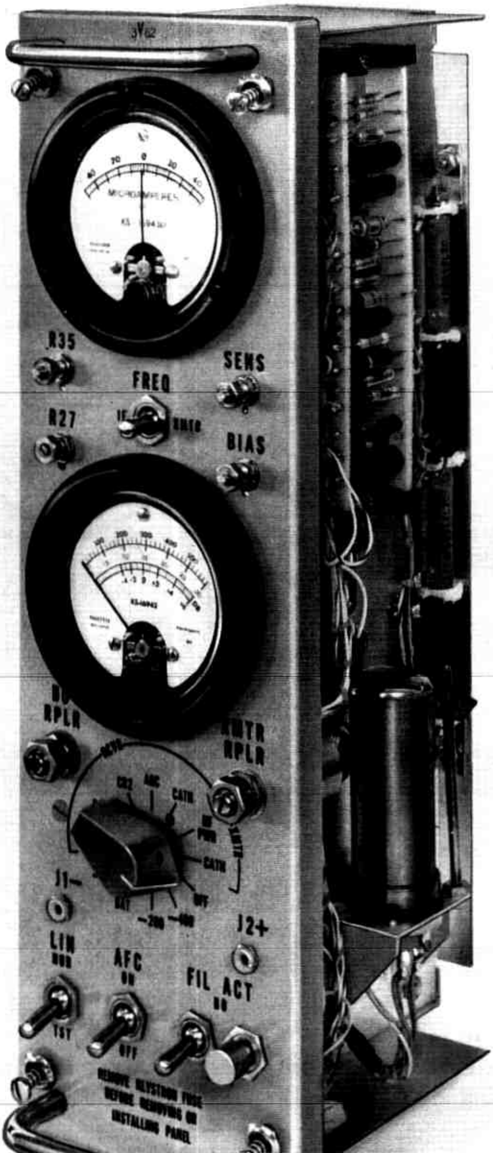


Fig. 19 — Control unit.

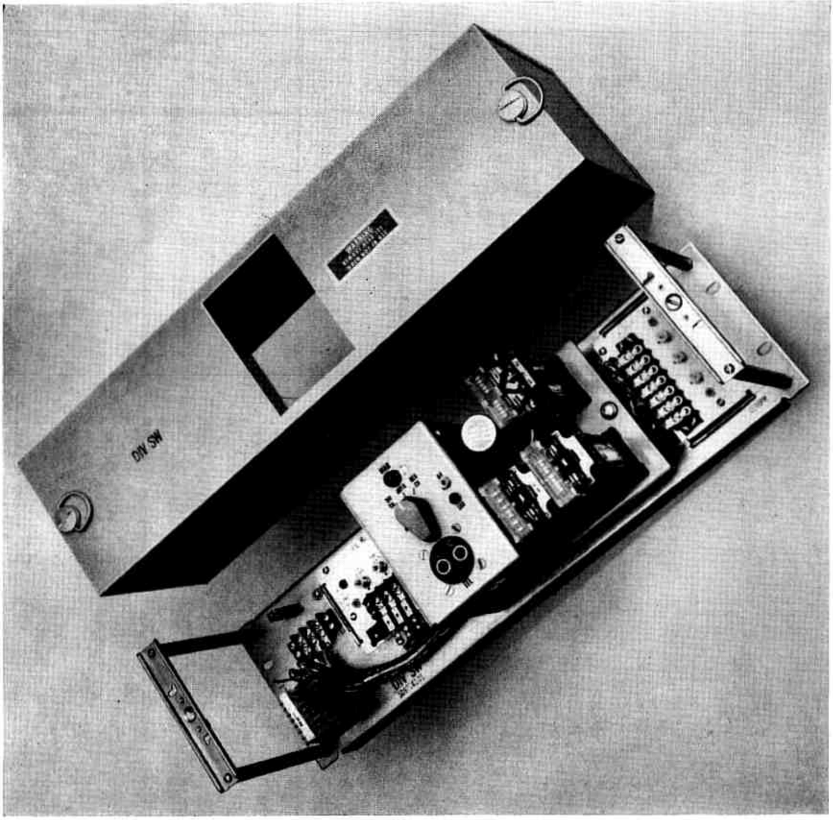


Fig. 20 — Diversity switch panel.

station, this panel selects one baseband output from the two receivers and in turn supplies the selected output to the order wire and alarm panel. The signals are split by means of a three-way pad and applied simultaneously to two transmitters operating on two frequencies separated by 240 mc. The selection of the better signal from the receiver output is controlled by a logic circuit utilizing information from two pilot monitoring circuits and a signal comparator circuit.

4.3.2 *Detailed Description of Diversity Switch Panel*

A block diagram of the diversity switch panel is shown in Fig. 21. A pilot monitor having a high impedance input is bridged across the baseband output of each radio receiver to sense the presence of the 2600-

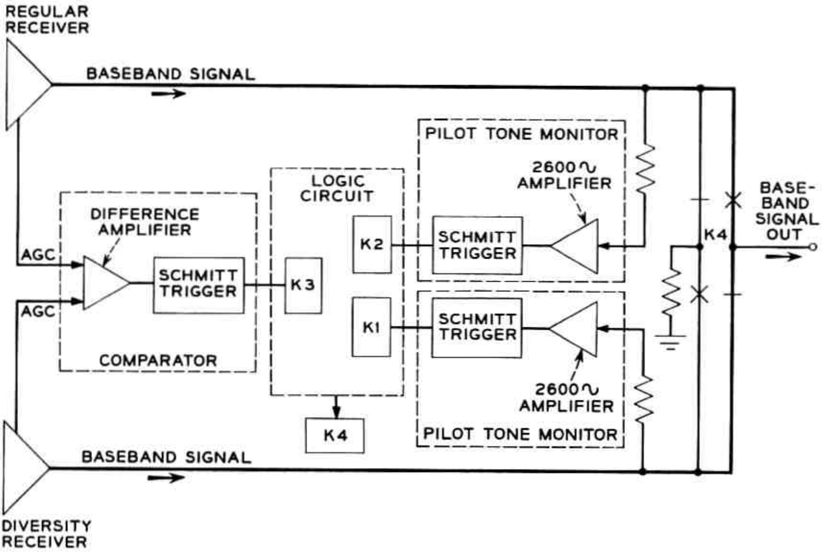


Fig. 21 — Block diagram of diversity switch panel.

cps pilot tone transmitted over the system for alarm and switching purposes (described fully in Section 4.3). Each monitor consists of a feedback amplifier, a full-wave rectifier, and a bistable trigger circuit. The amplifier uses three diffused-base silicon transistors in the common emitter configuration with a sharply tuned 2600-cps network in the beta circuit, removing most of the feedback at this frequency. The output of the resulting highly selective 2600-cps amplifier is rectified in a bridge rectifier. An adjustable dc voltage of the opposite polarity is added to the rectifier output, and the result is applied to the trigger circuit. The trigger circuit then controls a relay in the logic circuit. Improper operation of the pilot monitor from 2600-cps components of talkers on the order wire circuit is prevented by band elimination filters which remove those components at each head set connection. A functional schematic of the pilot monitor and its characteristics are shown in Fig. 22.

The comparator circuit compares the received signal level of the two radio receivers by means of the receiver AGC voltages. A difference amplifier, sensitive to the difference between the AGC voltages but insensitive to the absolute values, provides a dc output voltage, the magnitude of which is a function of the relative signal strength being received by each receiver. The output voltage controls a trigger circuit; it in turn controls a relay in the logic circuit. Controlled hysteresis in the trigger circuit avoids excessive switching on minor differences in signal level.

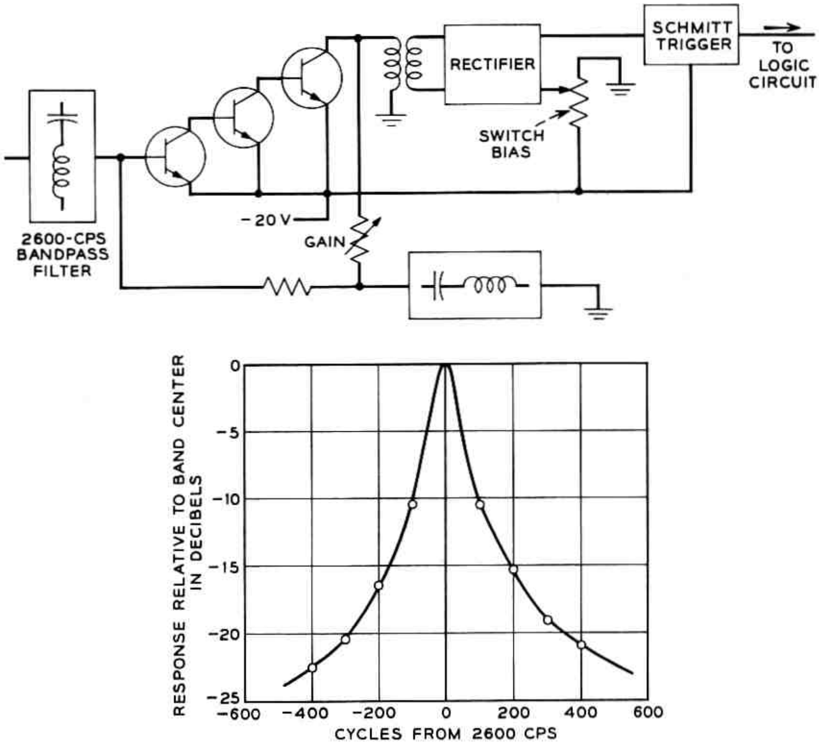


Fig. 22 — Functional schematic and selectivity characteristic of pilot monitor.

A simplified schematic of the comparator and its characteristics are given in Fig. 23.

The information supplied to the logic circuit from the two pilot monitors and the comparator is sufficient to determine which of the two receivers should be connected to the succeeding transmitters. Table II gives the system conditions and the resulting signal to the diversity switch.

If the pilot tones are present or absent simultaneously on both channels, fading controls the switch; if the pilot tone is absent on one channel, fading is disregarded. Removal of the pilot tone from both channels simultaneously for signaling or alarm purposes does not cause a switch. A manual control is also provided to permit a maintenance man to select either receiver output without regard to pilot or fading conditions.

The switch itself is a wire-spring relay with make-before-break contacts. During switchover, the contacts are bunched for approximately

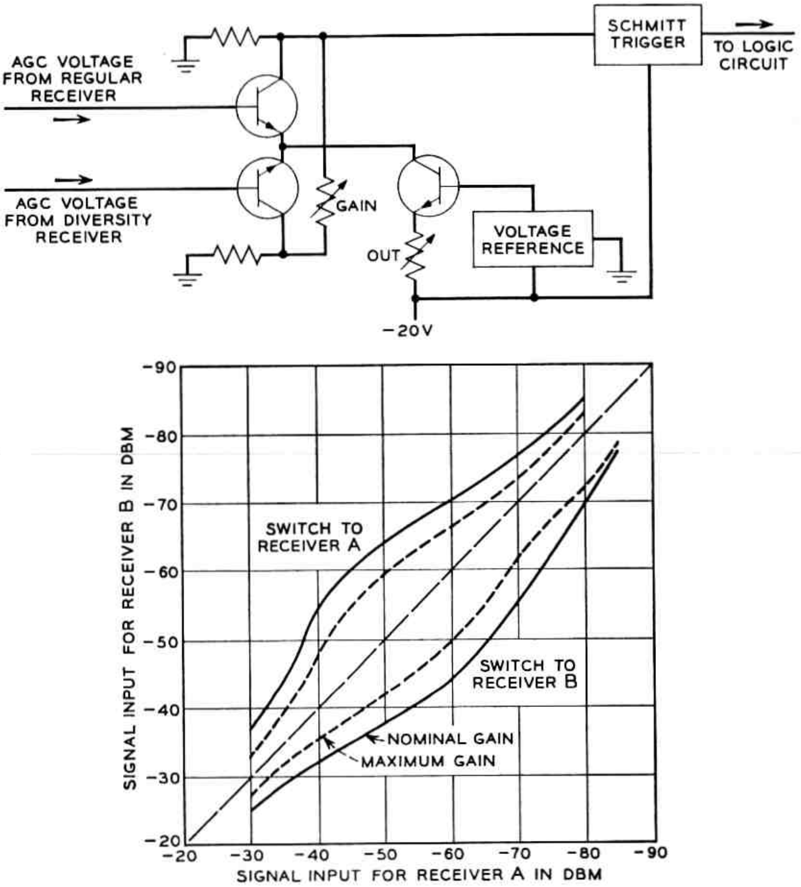


Fig. 23 — Functional schematic and operating characteristic of comparator.

TABLE II — LOGIC TABLE FOR DIVERSITY SWITCH

Regular Pilot Tone	Diversity Pilot Tone	Signal Level in Regular Channel	Signal Level in Diversity Channel	Switch Instruction from Logic Circuit
Present	Present	Faded	Normal	Switch to diversity channel
Present	Present	Normal	Faded	Switch to regular channel
Present	Present	Same as diversity	Same as regular	No switch
Absent	Absent	Faded	Normal	Switch to diversity channel
Absent	Absent	Normal	Faded	Switch to regular channel
Absent	Absent	Same as diversity	Same as regular	No switch
Present	Absent	Any condition		Switch to regular channel
Absent	Present	Any condition		Switch to diversity channel

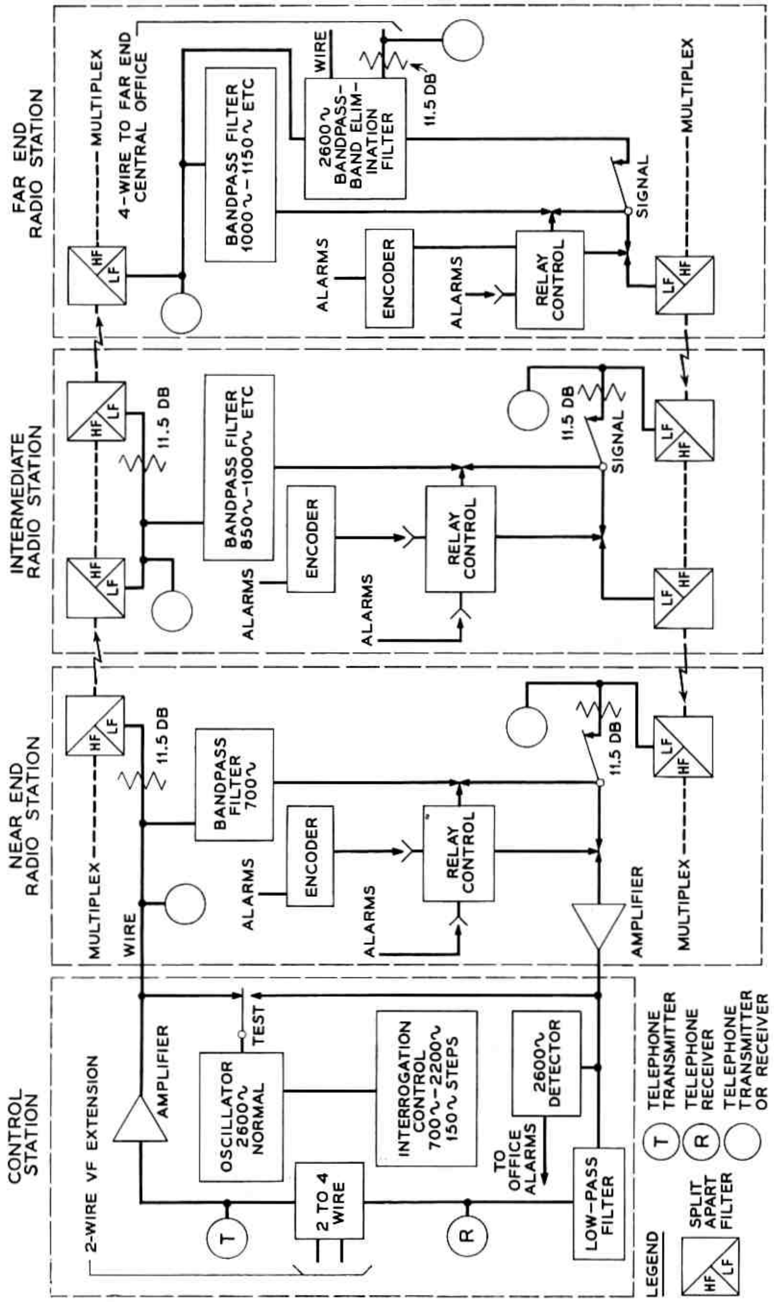


Fig. 24 — Order wire and alarm block diagram.

one millisecond, paralleling the outputs of the two receivers. If the receiver outputs are in phase and of equal magnitude, paralleling the outputs will cause no change in level. With fading, the condition of equal magnitudes will generally be met, since a fade must be very severe to change the receiver baseband output level. Thus, with proper adjustment, switches caused by fading will not cause transmission interruption. The same is true for manual switches made during system maintenance.

For equipment failure causing instantaneous loss of the pilot tone (and signal) in one channel, recognition of the loss by the pilot monitor, operation of the trigger circuit, the logic relay, and the switch to the other channel occurs in approximately 35 milliseconds. This causes negligible interruption to message circuits. Errors would occur in data circuits but because of the infrequent occurrence of equipment failures, this is not considered to be a serious condition.

4.4 *Order Wire and Alarm*

TL radio stations are normally unattended. In order to insure reliability of service and meet the requirements for unattended operation, an alarm system is necessary to report equipment failures and abnormal conditions to an attended control point from which maintenance personnel may be dispatched. A voice-frequency telephone facility to enable communication between radio stations is also needed to expedite maintenance. The TL radio alarm and order wire which makes use of the baseband frequency spectrum below 4000 cycles provides this service. However, in keeping with the objective of a low-cost equipment, the design only incorporates the reporting of those alarms which are essential to maintenance of service. These are:

1. *commercial ac power failure*. This is important since the batteries will discharge within 10 to 24 hours, depending on temperature, and result in system failure.
2. *transmission failure*. Short fades are timed out and not reported to the control center.
3. *low battery voltage and lightning arrestor failure.*
4. *failure of the air-navigation tower warning lights.*
5. *recovery of tower warning lights* when due to resumption of commercial power.
6. *"Signal In" by maintenance personnel* from one of the radio stations.

In this system, (see Fig. 24) a 2600-cycle tone is originated at the control station and is continuously transmitted over the outward radio path to the far radio terminal where, by means of a bandpass filter, it is looped back over the return radio path to the control station. If this tone fails to return, or if it is interrupted, office alarms are activated, indicating a trouble or abnormality. Except in the case of a transmission failure in a nondiversity system, the 2600-cycle tone is interrupted only sufficiently long enough to lock up and hold in the office alarms. Approximately 30 seconds is required for this operation because of built-in delay to avoid unnecessary alerting of personnel due to short-term fades. After the interruption, the alarm loop restores to normal so that indications from other stations may be reported. The system is limited to a single trouble report from each station except where an encoder is used, in which case higher-priority troubles at a station may be reported.

Once an alarm is registered, the next operation is to locate and determine the trouble. The 2600-cycle oscillator can be tuned to other frequencies by means of control keys. These other frequencies, called interrogation tones, are 700 to 2200 cycles in 150-cycle steps. At the first radio station a 700-cycle bandpass filter is bridged between the two directions of transmission, at the second station an 850-cycle filter, at the third station, a 1000-cycle filter, and so on to the far end of the system up to a maximum of eleven stations.

If an interrogation tone key is operated at the control station, an interrogation tone will go out over the system to be routed back at the station having the filter tuned to that frequency, and will be heard by the control station attendant in his telephone head set. Fig. 25 shows the loss characteristic of the adjacent filters 850, 1000 and 1150 cycles. When the 1000-cycle path is opened, the 1000-cycle leakage through the adjacent filters is approximately 53 db down from the normal level. Recognition of the particular frequency is not necessary, since each key marks a definite station. In a nondiversity system having a transmission failure, the tones will return from all stations on the near side of the failure and, of course, from no stations beyond the failure. Where transmission failures are made good by a diversity switch, or for other types of trouble, the interrogation filter bridge path is either opened continuously or on a pulsed basis under relay control, marking that station as a trouble point.

On interrogation, either a tone, no tone, or a pulsed tone will be heard at the control location. A continuous tone indicates no trouble. No tone indicates transmission failure. A pulsed tone indicates commercial power failure or low battery voltage. Where air-navigation lights are installed, an encoder is provided to code-pulse the tone to obtain additional information required by the air-navigation authorities, namely:

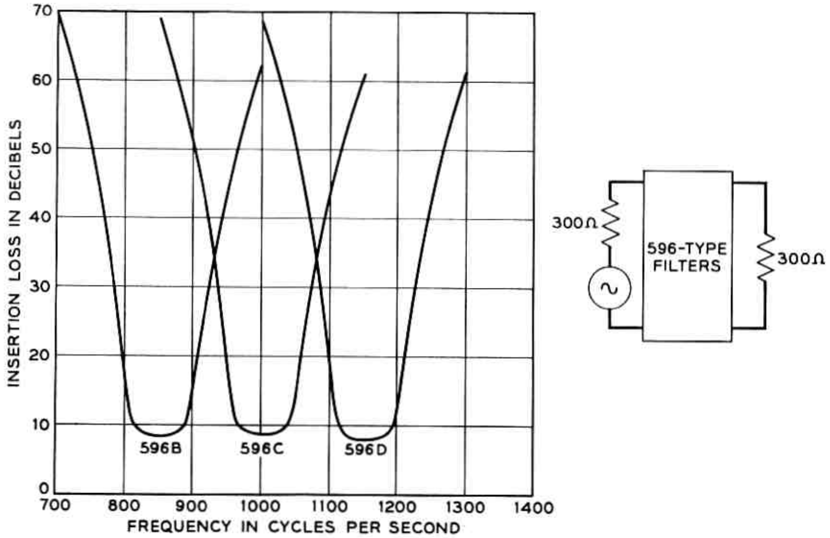


Fig. 25 — Order wire and alarm filter characteristic.

1. Both tower lights. This also indicates commercial ac power failure . . . . (3 shorts)
2. Top light flasher failure . . . . (2 shorts)
3. Low battery voltage or lightning arrestor failure . . . . (pulses)
4. Tower side light or single top light . . . . (1 short and 1 long).

The encoder provides a pulse generator and an eight-step counting and coding circuit operating under control of a directing or start relay for each of the above alarm codes. This circuit is arranged to give priority to the alarms in the above numerical order. If a low-priority alarm is present, a higher-priority alarm will originate a new alarm indication at the control point. Likewise, if the high-priority alarm clears, the lower-priority alarm will be maintained.

A transmission failure cannot be pin-pointed to a particular station because the system is not able to detect whether a failure exists in a transmitter in one station or in the associated receiver in the next station. Therefore, in a nondiversity system, the failure may be in the last station returning an interrogation tone or in the next station which does not return the tone. In a diversity system, the failure may be in the receivers of the station reporting the failure or in the transmitters in the adjacent stations.

The order wire is on a four-wire basis and shares the low-frequency part of the baseband with the alarm signals. The telephone set transmitter and receiver at a radio station under key control may connect to either direction of transmission to enable a conversation either way from the station. An amplifier in the receiver circuit is required to provide adequate listening level. The net loss between transmitter output and receiver input is approximately 15 db. Band elimination filters in the telephone receiver circuit provide about 27 db discrimination against the 2600-cycle alarm tone, which is essentially always present. Filtering is also provided in the transmitter output to attenuate any 2600-cycle components of voice energy entering the alarm loop.

Provision is made to serve spur routes as long as the stations involved do not exceed eleven and as long as no more than one spur is involved at a junction point. Through use of a six-hybrid bridge and proper placement of band elimination filters, the 2600-cycle tone traverses the total loop through all stations, while communication and interrogation are handled on a bridged basis.

In the design of this equipment, every effort has been made to reduce maintenance by simplifying the radio station equipment and confining it insofar as possible to passive circuit elements. The more complex active elements such as the oscillator and detector are located at the attended control point.

If more than one regular two-way radio channel is provided on a route, each is provided with an order wire and alarm facility. Thus a fully loaded three-channel diversity system would require three separate order wire and alarm facilities.

At near and far radio terminals, the order wire and alarm equipment consists of a single $5\frac{1}{4}$-inch × 19-inch panel arranged for single side maintenance (see Fig. 26). At repeater stations, two similar sized panels are required; one is located in the near repeater cabinet or bay and the other in the far repeater cabinet or bay. At a repeater point, the transmitter-receiver operates on a back-to-back basis. The microwave transmitter-receiver equipment facing the near terminal is called the near repeater bay or cabinet, and the one facing the far terminal is called the far repeater bay or cabinet. The first panel incorporates the "near" split-apart filter, the interrogation filter, the relay controls, and the telephone set circuit. The second panel has the "far" split-apart filter, transmission pads and a multiple of the telephone set. It also incorporates the spur hybrid arrangements when required.

The control station equipment, which includes the oscillator, detector and attendant's telephone set circuit, is mounted on a single panel $12\frac{1}{4}$

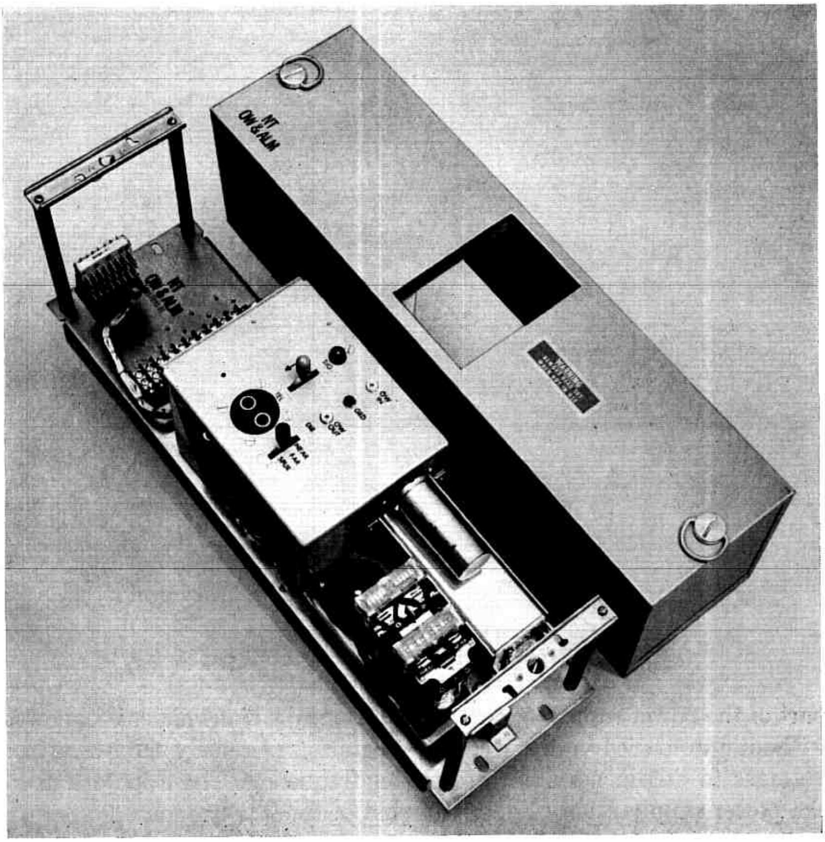


Fig. 26 — Order wire and alarm panel.

inches high by 19 inches wide. This panel is arranged for double side maintenance and has been designed for flush mounting on a bay or console (see Fig. 27).

4.5 *Power System*

4.5.1 *General*

One of the major features of the TL system is that it is not necessary to provide for emergency ac power generation during commercial power outages of normal duration. Continuous service is maintained by operating the repeater or terminal bay from storage batteries which form a

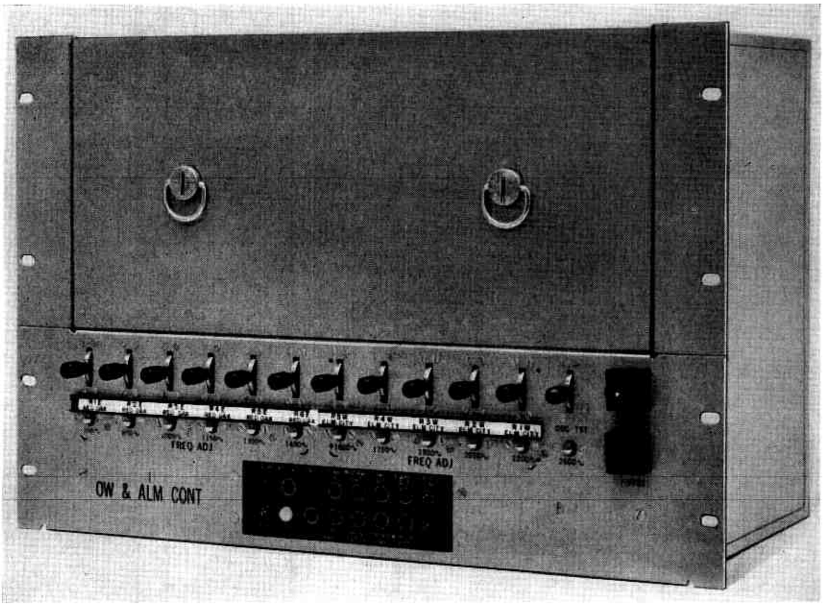


Fig. 27 — Order wire and alarm control panel.

part of the power supply system. In addition, the transmitter is operated without automatic frequency control, requiring extremely stable klystron voltages to obtain a stable transmitter frequency. These factors make the power supply a very important part of the TL system.

As shown in the simplified block diagram of Fig. 28, a ferroresonant transformer and rectifiers, fed from 117-volt commercial ac power, supplies a stabilized dc voltage which float-charges four 6-volt batteries in series. From the batteries, power for the transistor circuits and a dc-dc converter is supplied through two regulators. The converter provides the klystron voltages and the power for the AFC magnetic amplifier. Also included is a low battery voltage alarm. If the bay is operated in an office, the supply may be operated from the −24-volt office battery, omitting the charger and the four batteries normally provided in the bay.

4.5.2 *Detailed Description*

The ferroresonant transformer has two separate toroidal cores having characteristics such as to maintain a secondary voltage relatively insensitive to line voltage. Two silicon diodes in a full-wave center tap

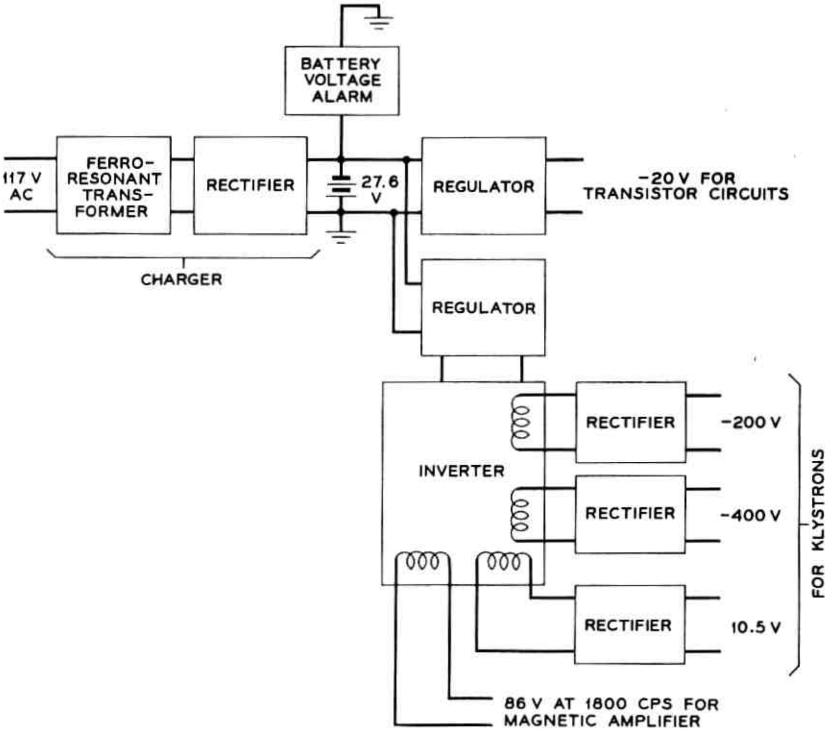


Fig. 28 — Block diagram of power supply.

arrangement give a regulated dc output, which is applied to the batteries for floating or charging.

The four batteries are high specific gravity lead-acid types connected in series, and float at −27.6 volts. They were chosen to provide maximum reserve at an economical cost, in an environment which involves a wide temperature range. Most of the time the batteries are being float-charged. This desirable condition contributes to very low battery maintenance. The high specific gravity aids in protecting the batteries at low temperature: at −40°F they do not freeze, even if fully discharged.

The −20-volt supply for the transistor circuits is obtained through a series regulator transistor which is controlled by an error-voltage amplifier. The reference voltage is obtained from a temperature-compensated voltage reference diode.

The dc-dc converter consists of two transistors and a saturable transformer switching at a rate of 1800 cps. The square-wave output is stepped up in a power transformer and rectified to provide voltages of −400

TABLE III — POWER SUPPLY CHARACTERISTICS

Voltage	Stability*	Current
−600 dc	±0.45%	10 ma
−400 dc	±0.45%	100 ma
10.5 dc	±0.45%	1.75 amps
43.0 ac (rms)	±0.45%	10 ma
−20.0 dc	±1%	750 ma

* Over the temperature range of −40 to +140°F and ac line voltage input of 95 to 135 volts.

volts for the klystron resonators and −200 volts (which is added to the −400 volts to obtain −600 volts for the repellers). Other windings on the power transformer provide an 86-volt center-tapped square wave for the receiver AFC magnetic amplifier, and a 10.5-volt dc supply for the klystron heater circuits.

Regulation of these supplies is accomplished in the same manner as for the −20-volt supply. The feedback is obtained from the −400-volt output, advantage being taken of the fact that the output from the other windings of the transformer will closely follow its output.

The battery voltage alarm circuit is a single transistor circuit which operates a relay when the magnitude of voltage becomes less than 25 volts. This initiates an alarm whenever the ac power fails, the charger fails, a fuse blows, or any event occurs that causes the batteries to cease being charged.

The characteristics of the power system are shown in Table III and in Fig. 29.

The power supply panel (approx. 15$\frac{3}{4}$ inches × 19 inches), Fig. 30, is designed for maintenance from the front side only. To expedite replacement in the field, the regulator and invertor for the klystrons, the transistor circuit inverter, and the battery voltage alarm sections are mounted on removable panels using twist-type fasteners. The electrical connections are made by screw terminals. The battery voltage alarm section is not used in central offices where power is supplied from an existing office battery.

The major heat-producing components (battery charger) are mounted on the rear of the panel to obtain maximum cooling through unrestricted air flow.

4.5.3 *AC Distribution and Grounding*

Arrangements are provided at each station for the distribution of ac commercial power to the radio equipment and to associated facilities

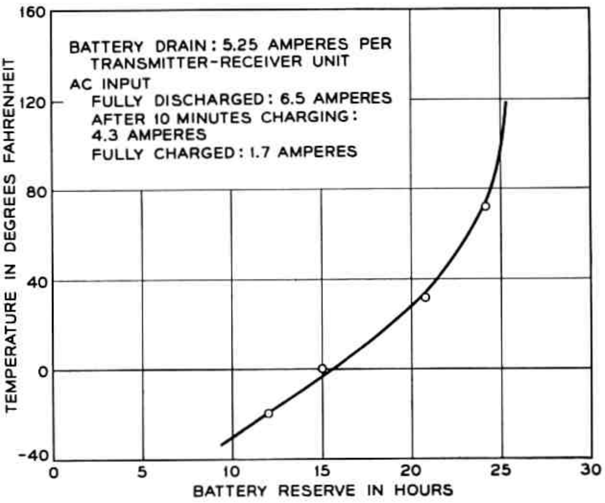


Fig. 29 — Battery reserve vs temperature.

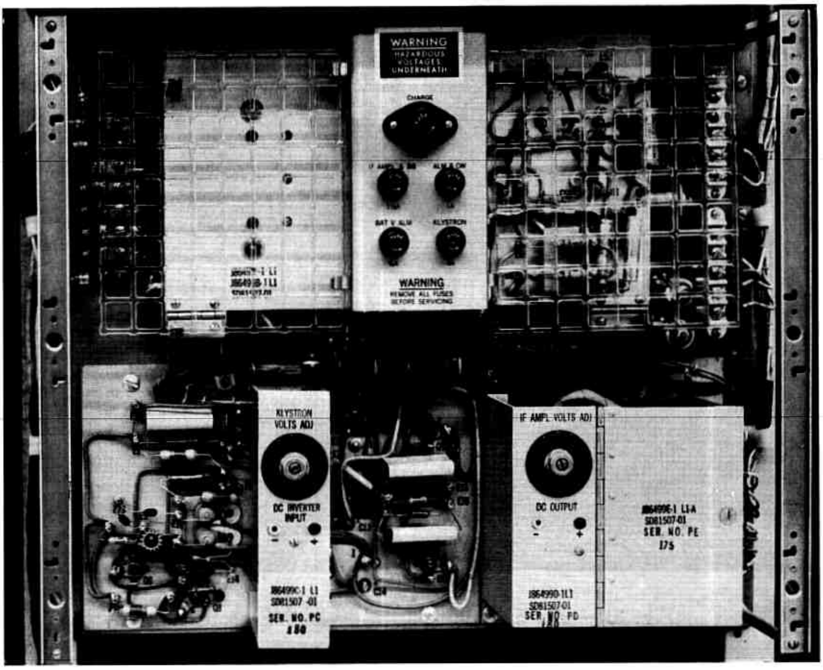


Fig. 30 — Power supply panel.

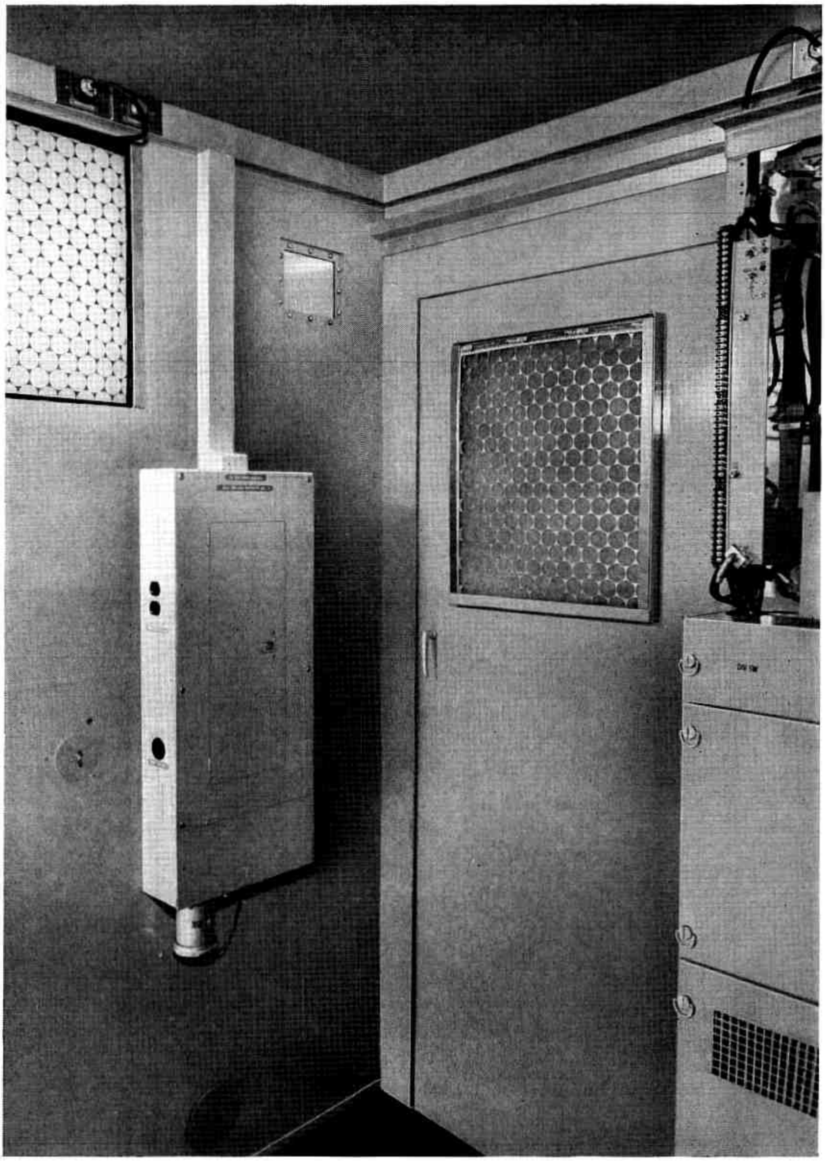


Fig. 31 — AC distribution and lightning protection.

and for adequate grounding to protect both equipment and personnel. Features provided are:

1. *a double-pole primary power disconnect switch* which enables connection to an emergency engine alternator (mobile or transportable unit) in the event of prolonged commercial power failures.

2. *distribution of ac power with overload protection* to the several transmitter-receiver equipments, to the antenna heaters, and to the aircraft warning light circuits on the tower.

3. *a secondary voltage protection circuit* which uses protectors with 6-mil gaps to protect the equipment against voltage surges on the ac power lines in excess of 1100 volts. This circuit provides alarm indications in the event of failure of its components. Fig. 31 illustrates an integrated design for housing the hardware for items 1, 2 and 3 above in a small shelter. A comparable size waterproof housing is available for outdoor cabinet installations.

4. *an aircraft warning light control circuit* which provides for flashing and steady lights on the tower and for alarm indications in the event of light or flasher failure.

5. *a grounding system* which provides for grounding all exposed metal parts in and around the station to a well established grounding grid. In addition to the electrical grounds, this includes grounding of such items as lightning rods, frameworks, building steel, towers, waveguide runs and guy lines.

4.6 *System Characteristics*

The gain-frequency characteristic of a typical single TL link is shown in Fig. 32. The transmitter input for nominal maximum carrier frequency deviation of ±6 mc is -10.5 dbm. The receiver output for this deviation is $+10$ dbm, a difference of 20.5 db. A 6-db splitting pad for feeding a diversity transmitter is incorporated in the system, so that the nominal usable system gain is 14.5 db.

Since many TL repeaters will be exposed to rather large temperature variations, the stability of the system gain as a function of temperature is an important performance parameter. Fig. 33 shows this from data taken on one hop during laboratory environmental tests.

The message-carrying capacity of the system was measured by applying a band of white noise simulating a number of single-sideband message circuits. A band rejection filter at the transmitting end clears a "slot" of the noise near the top channel of the number being simulated. At the receiving end, the noise introduced into this slot by the thermal

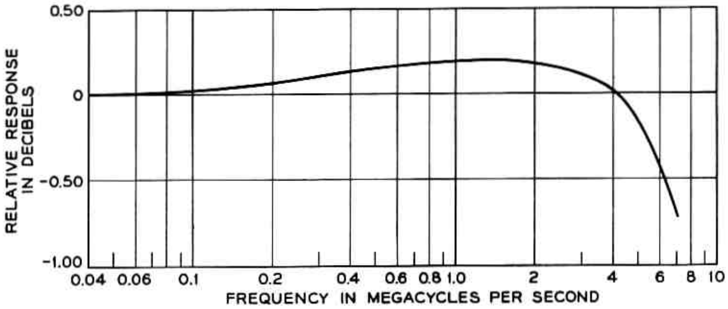


Fig. 32 — Single-hop gain-frequency characteristic.

and cross-modulation characteristics of the system is measured as a function of the level of the "message circuits" applied to the transmitter. Fig. 34 gives typical results of such tests on the Conway-Harrison, Arkansas system.

The contribution of thermal noise to the total noise in a message channel output will depend upon the carrier-to-noise ratio in the IF amplifier at the limiter. The noise figure of the receiver, including the three- or four-cavity bandpass filter preceding the modulator, is a maximum of 15 db at low RF signal levels, and averages somewhat less than this. The relationship between the RF carrier level and the thermal noise in the top channel of 240 channels in a typical link is shown in Fig. 35 for a peak frequency deviation of ±6 mc and a deviation per channel 17.5 db below 6 mc for a 0-dbm signal at the zero transmission level point.

An important characteristic of the system from the standpoint of the maintenance man is the performance of the order wire. The frequency

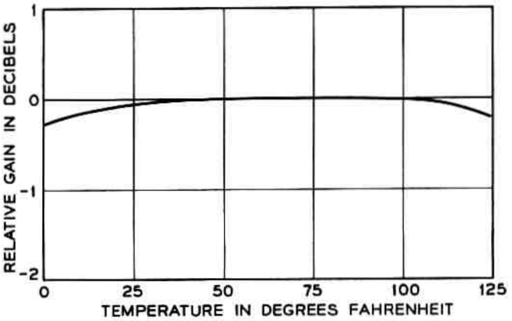


Fig. 33 — Net loss stability vs temperature for one hop.

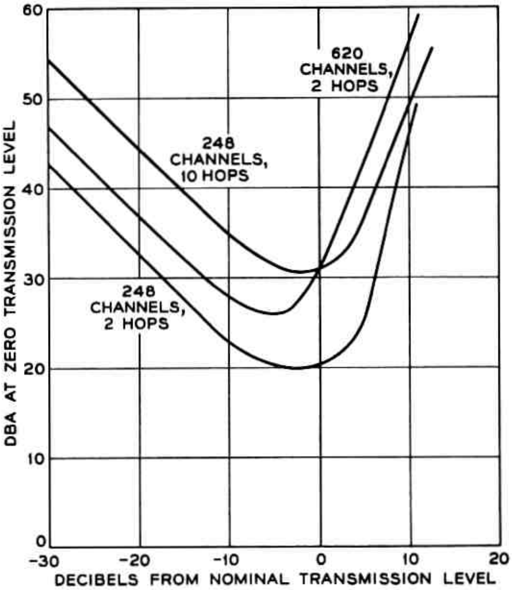


Fig. 34 — Typical noise loading characteristic of TL system.

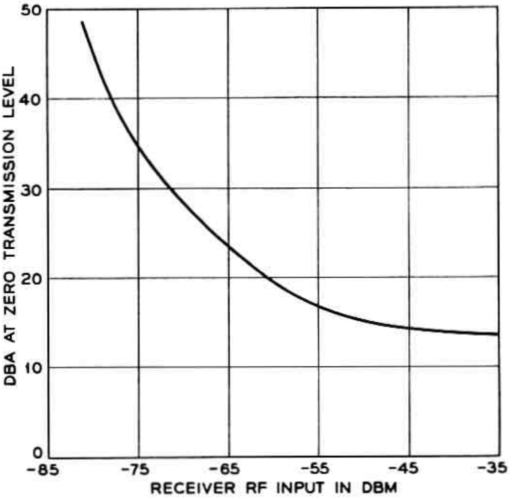


Fig. 35 — Thermal noise vs RF input signal level.

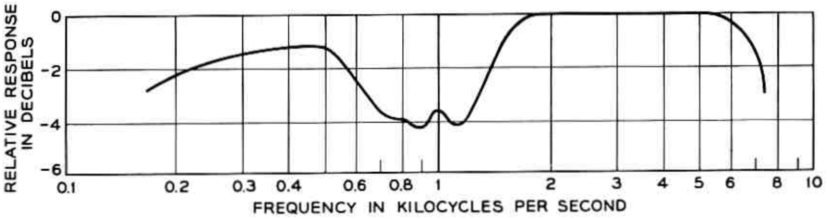


Fig. 36 — Order wire gain-frequency characteristic, four hops.

characteristic is shown in Fig. 36 for four hops. The notches caused by the bridging on of the interrogation filter at each repeater are evident.

V. EQUIPMENT FEATURES

5.1 *General*

The TL radio transmitter-receiver equipment is provided in either a weatherproof pole-mounted type of cabinet for outside, or in bay frameworks for indoor applications.

The cabinet arrangement illustrated in Fig. 37 is 46¼ inches wide by 62¼ inches high by 18½ inches deep. It incorporates two parallel frameworks, each accommodating standard 19-inch wide panels. In the right-hand framework from the bottom up are the power supply unit, the order wire and alarm panel, and the transmitter-receiver panel equipment. Above this frame supported by the cabinet structure are two ventilating fans. These fans are thermostatically controlled to turn on at 105°F and off at 80°F. They furnish approximately 80 cubic feet of air per minute through a 1-inch spun glass filter. This holds the air temperature within the cabinet to approximately 6°F above the outside air temperature. In the lower half of the left-hand frame, two shelves provide for storage of the four batteries, which have protective covers over the terminals. Space in the upper half of this frame is reserved for carrier line equipment such as the line amplifiers and combining equipment for ON carrier. Typical of the method of support for these cabinets is the "H" pole structure for a repeater station of the Billings-Hardin, Montana, system illustrated in Fig. 38.

Two bay arrangements are available, a seven-foot high floor-supported channel iron frame (see Fig. 39) for small unattended stations or shelters, and a nine-foot high standard channel frame for central offices. The equipment layout is similar to the cabinet except for the battery location in the lower part of the frame. Ventilation becomes a building problem, and any required carrier equipment will be mounted on its

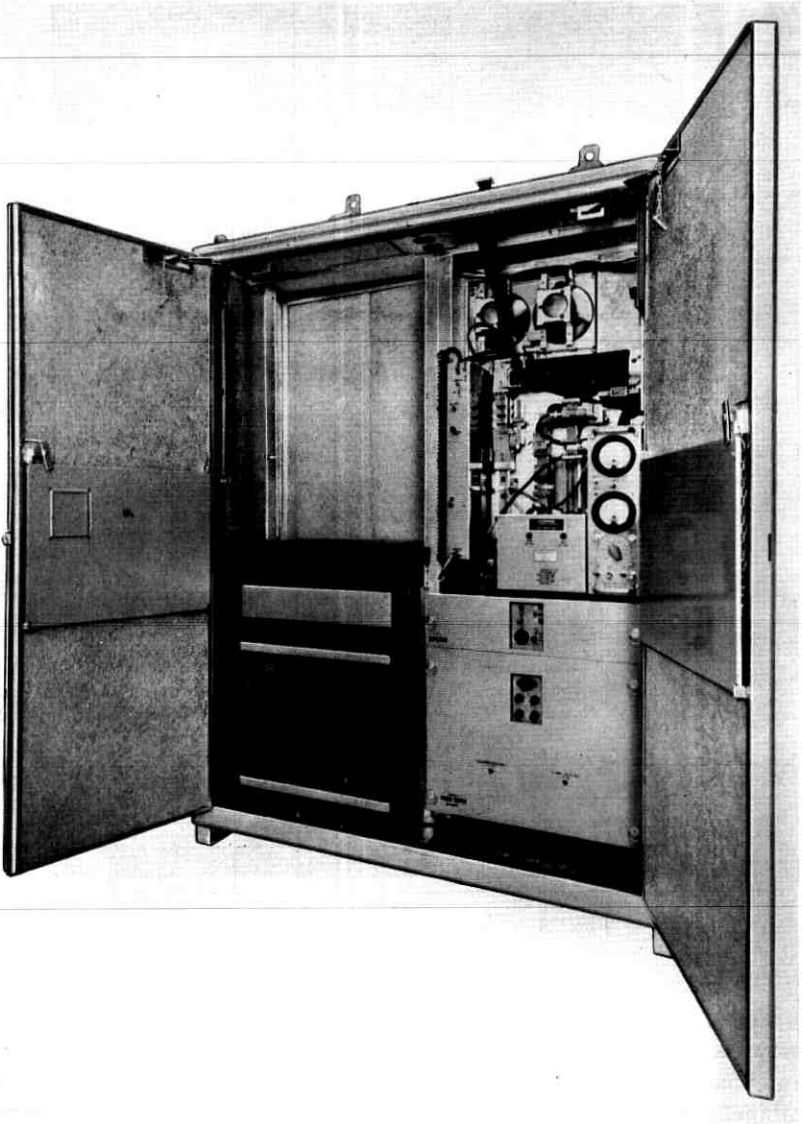


Fig. 37 — Pole-mounted cabinet for outdoor applications.

own separate framework. In a diversity system, the cabinets or bays serving the diversity channel have the diversity switch panel located just above the power supply in the space allocated to the order wire and alarm panel in regular channel equipments.

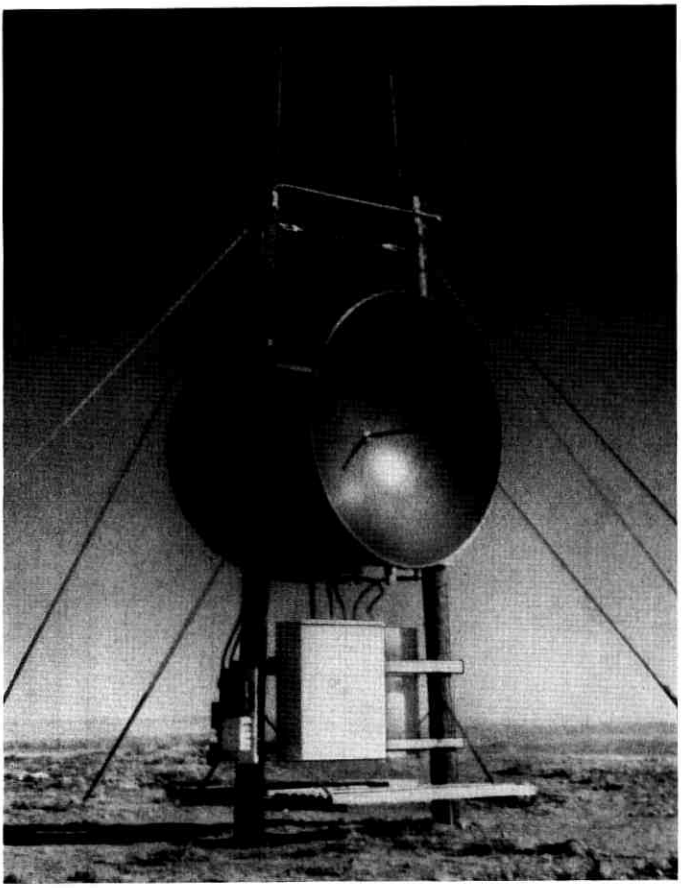


Fig. 38 — Typical "H" frame installation.

The primary difference between cabinet and bay equipments is the arrangement of the channel dropping networks. The cabinet has a single network mounted vertically with a single waveguide entrance to the antenna. A single-channel diversity terminal arrangement involving two cabinets is obtainable, since the individual waveguides run separately to a polarizer where the signals are combined in opposite polarities into the single antenna. Two such arrangements on a back-to-back basis, involving four cabinets, provide a diversity repeater. The bay arrangement mounts one or two networks horizontally at the top of the frame, making it possible to connect as many as three adjacent bays together to obtain the maximum three-channel system. The diversity bays have

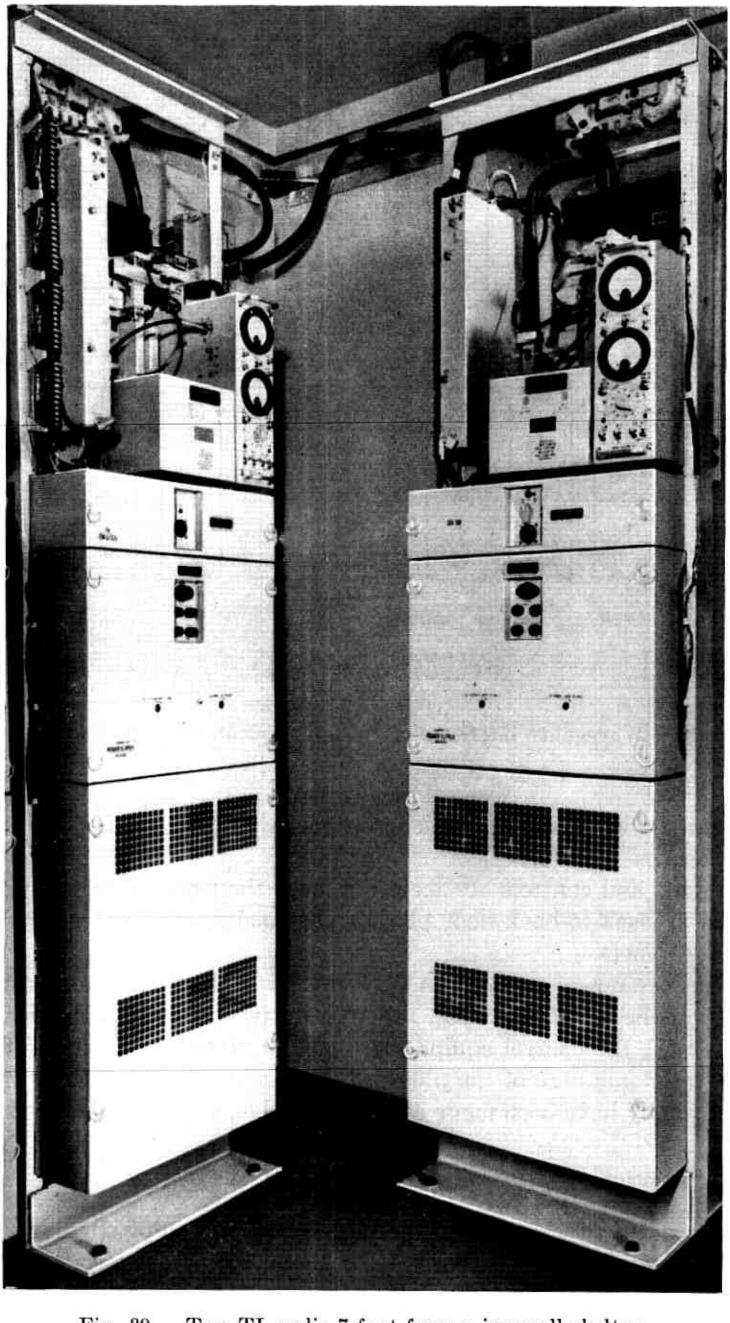


Fig. 39 — Two TL radio 7-foot frames in small shelter.

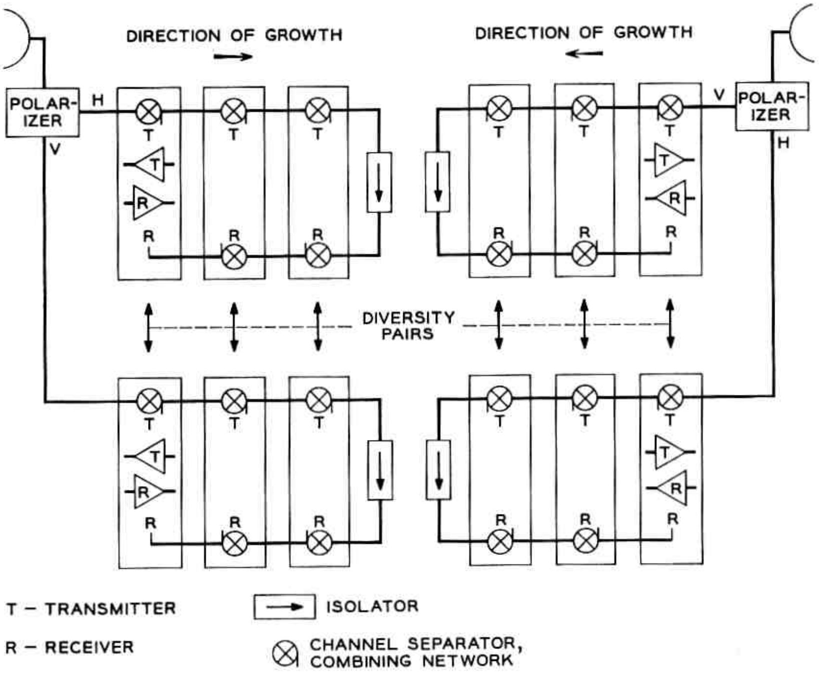


Fig. 40 — Waveguide interconnection arrangement for a three-channel diversity system.

a separate waveguide run which connects with the regular run at the polarizer (see Fig. 40).

All units and controls are accessible from the front, permitting back-to-wall or back-to-back floor plan arrangements and front door access for the cabinets.

To reduce out-of-service time, to simplify maintenance and to minimize "on the spot" servicing, the transmitter baseband amplifier, the receiver IF, the control equipment, and the diversity switch pilot and comparator units are of the plug-in type. Likewise, critical units in the power supply have been made easily removable, as previously described.

5.2 *Transportable Equipment Shelters*

Objectives in the provision of communication equipment have been to minimize effort on the part of the customer in job engineering and during the installation interval. This program of providing an essentially entire service has been termed "Turn Key"; thus the customer is provided a commercial system that is completely installed and tested, ready for him to turn on or up for service. In support of this program as applied to TL radio, a transportable equipment shelter is available which is equipped and wired at the factory, in accordance with customer requirements, prior to shipment to the radio sites. The objectives of this development are:

1. *assemble, wire, and test the major portion of the equipment at the manufacturing plant with experienced labor; minimize job engineering and field installation; and perform the work under controlled factory environment unaffected by outdoor climatic conditions.*

2. *minimize packaging of numerous items to reduce shipping costs; and reduce confusion resulting from many packages arriving at various intervals and the possibility of delayed and lost items.* This will allow more positive installation scheduling and reduced installation intervals.

3. *design the shelters so that they may be transported by standard commercial trucks on the roads of the various states.*

4. *minimize weight, use fireproof materials; incorporate adequate structural strength; provide ventilation and insulation; simplify installation and handling procedures; and incorporate such other features as are essential to the radio station.*

Two sizes of shelters are available:

Small shelter — intended to house a single-channel diversity repeater or for applications requiring a maximum of four 7-foot high transmitter-receiver bays and one miscellaneous bay. It is expected that this smaller shelter will be used rather than the pole-mounted cabinets in areas where general climatic conditions make outside maintenance impractical. This structure is approximately 7 feet wide, 7 feet, 6 inches long, and 8 feet high. Approximate weight of the unequipped shelter is 1400 lbs; maximum fully equipped for shipment, 2600 lbs; maximum in-place weight with batteries, 3800 lbs. Fig. 41 shows this small shelter equipped with two 5-foot paraboloidal antennas vertically beamed for use with a periscopic reflector. Fig. 42 illustrates how this shelter may be transferred from a truck to a prepared foundation (concrete piers) using a relatively small sign-erecting, truck-mounted crane.

Large shelter — intended for the ultimate three-channel diversity repeater or applications requiring a maximum of 12, 7-foot high transmitter-receiver bays and two miscellaneous bays. This shelter is approximately 7 feet wide, 16 feet, 4 inches long, and 8 feet high. Approximate weight of the unequipped shelter is 2500 lbs; maximum fully equipped for shipment. 6000 lbs; and maximum in-place weight with batteries, 10,000 lbs.

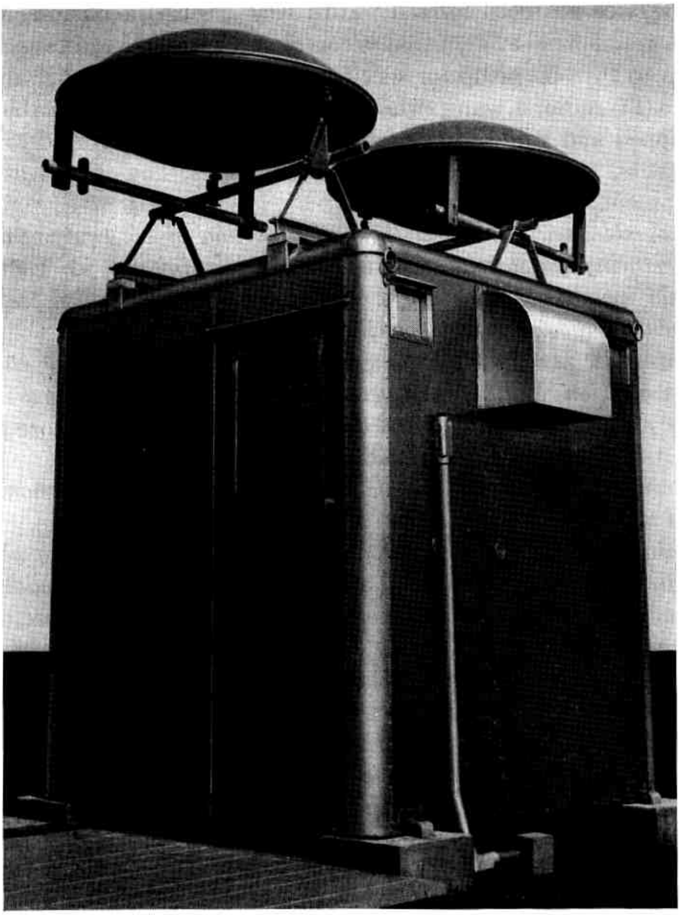


Fig. 41 — 7-foot × 7-foot shelter equipped with two 5-foot paraboloidal antennas.

5.3 *Towers*

An "H" frame wood-pole antenna tower specifically designed for TL radio, as shown in Fig. 38, is available. This guyed structure, having a maximum height of 60 feet, is particularly adaptable to installation by telephone company plant personnel, since handling this type of available pole hardware is a commonplace everyday task with them. This frame supports the antenna and reflector, and as many as four cabinets (single-channel diversity repeater) can be accommodated. Where high

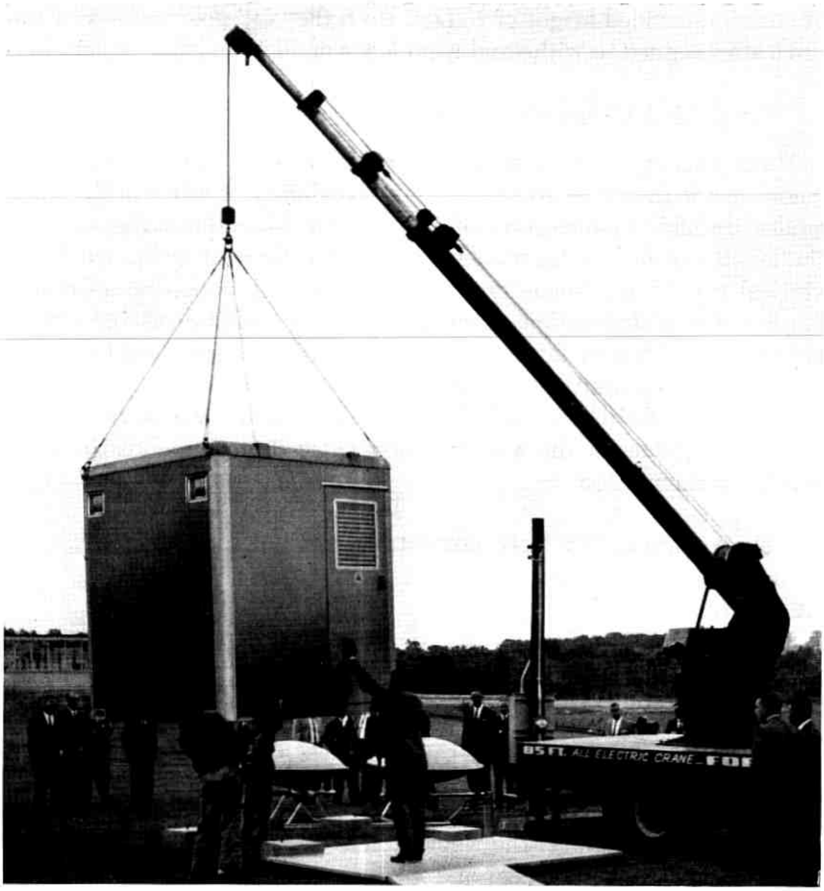


Fig. 42 — Installation of small shelter.

steel towers are required, a short "H" wood pole stub structure is used to support the outdoor cabinets and vertically beamed antennas.

A new series of general-use lightweight steel towers, either guyed or self-supporting ("C" type), is also available for use with TL Radio. These towers vary in height up to 105 feet and are triangular in cross section. The self-supporting tower is tapered from a 4-foot face at the top to approximately 17 feet at the bottom of a 105-foot tower. The guyed tower has parallel sides 2 feet in face width. Both of the above towers are designed to support two 6 × 8-foot reflectors up to maximum height or two 5-foot or 10-foot paraboloidal dish antennas up to a maximum recommended height of 75 feet. Both the "H" pole and steel structures are designed to withstand wind loads of 30 pounds per square foot.

5.4 *Waveguide Moisture Problems*

Many microwave stations provide equipment for charging the waveguide runs with dry air to minimize the possibility of water in the waveguides through condensation or leakage. In TL radio stations, where the lengths of horizontal waveguide runs are short as compared to the vertical runs, use is made of a waveguide "T" junction as a drain to prevent the accumulation of condensation or moisture at the bottom of the vertical waveguide run. In central offices where long horizontal runs may be encountered, available dry air systems may be used. In the TL radio shelter, a barrier (waveguide pressure window) is placed in the waveguides at the wall entrance plates to prevent breathing and possible condensation.

VI. MAINTENANCE AND TEST EQUIPMENT

6.1 *General*

The basic objectives of the TL system design, from the maintenance viewpoint, were to keep the testing to a minimum, to use simple procedures, and to require no elaborate test equipment. For the more complex circuitry, this was achieved by designing stability into the circuits, making the circuits easily replaceable, and requiring only simple over-all measurements for the determination of proper operation. If a unit is determined to be faulty, the whole unit is replaced. Many of the measurements are simple voltage and current measurements, so the means for accomplishing them have been built-in to achieve a high degree of convenience.

6.2 *Built-In Test Features*

A quick and convenient determination of the operating condition of a radio transmitter-receiver panel can be made by observing the two meters on the control unit. One is a zero-center meter which is used to monitor the IF frequency or the transmitter frequency. The second is a multi-scaled meter which can be used with a multiposition switch to monitor the klystron voltages and currents, battery voltage, modulator, diode currents, RF power output, received signal level, and the regulated transistor circuit supply voltage.

As mentioned in the section describing the transmitter, a diode detector, a high-Q invar cavity bandpass filter, and a two-position attenuator are used for monitoring transmitter frequency and adjusting transmitter deviation. A sample of the transmitter output is passed through the cavity filter for frequency monitoring. The filter is tuned to the particular transmitting frequency of the bay. When the transmitter is "on frequency," the reading on the control unit meter will be a maximum. When the transmitted RF carrier is frequency modulated, the detected output through the cavity filter will decrease because the sidebands that are generated (and take some of the power previously in the unmodulated carrier) are attenuated by the cavity filter. The reduction obtained can be calibrated for a given deviation, modulating frequency, and cavity filter. This calibration is made in the two-position attenuator in its "loss" position. Thus, by applying a particular frequency (100 kc is used for TL) at a given level, the transmitter baseband amplifier gain control is adjusted to give the same reduction in output from the cavity filter as is obtained when the attenuator is in its calibrated loss position and the carrier is unmodulated. Without this adjustment, klystrons having different modulation sensitivities being periodically put into the system would cause excessive noise in the message channels because of improper deviations.

Another transmitter adjustment built into the control panel is a simple means for improving klystron deviation linearity. It was determined that near-optimum linearity could be obtained by offsetting the repeller voltage by a small fixed amount from that which gives maximum output power. A final frequency adjustment is then made with the klystron cavity tuning.

Other built-in maintenance aids are pin jacks on the order wire and alarm panel, IF and baseband unit, transmitter baseband amplifier, and power supply which bring out important operating points primarily for trouble location.

6.3 *Test Sets*

The main item of test equipment for the TL system is an especially designed portable unit comprising three main sections: (1) the signal-generating section, which supplies IF and baseband frequencies for various tests at calibrated levels; (2) the voltmeter section, which permits measurement of baseband levels; and (3) the attenuator section, which permits adjustment of IF and baseband signal levels. Fig. 43 is a photograph of the unit, which measures approximately 10 × 10 × 16 inches and weighs less than 30 pounds.

This unit uses solid-state circuitry and is ac powered. With it, the

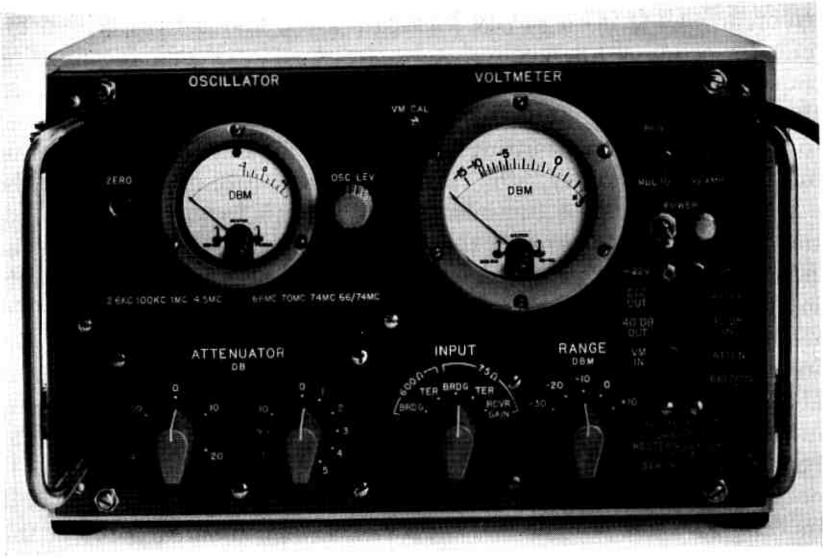


Fig. 43 — Test set.

important tests not built into the control panel can be performed. Some of these tests are: generation of an FM signal at IF for adjustment of the receiver baseband amplifier gain control; measurement of the 2600-cps pilot level; generation and measurement of baseband signals for gain-frequency characteristics; and generation of IF signals for measurement of received signal level, and IF amplifier and discriminator measurements.

The signal generator section has a pushbutton-controlled RC oscillator for generation of 2600 cps, 100 kc, 1 mc, and 4.5 mc; and crystal-controlled oscillators for 66, 70, and 74 mc. A circuit which switches between the 66- and 74-mc outputs at a 100-kc rate gives an FM IF signal which permits setting the gain control of the receiver baseband amplifier. A detector circuit monitors the oscillator outputs and permits accurate adjustment of the level.

The voltmeter measures baseband signals up to 4.5 mc at levels from −40 to +13 dbm. It provides for 75-ohm and 600-ohm terminated measurements and has a high impedance input for bridging measurements.

Other items of test equipment required are an accurate dc voltmeter for power supply adjustments and a general purpose volt-ohm-milliammeter.

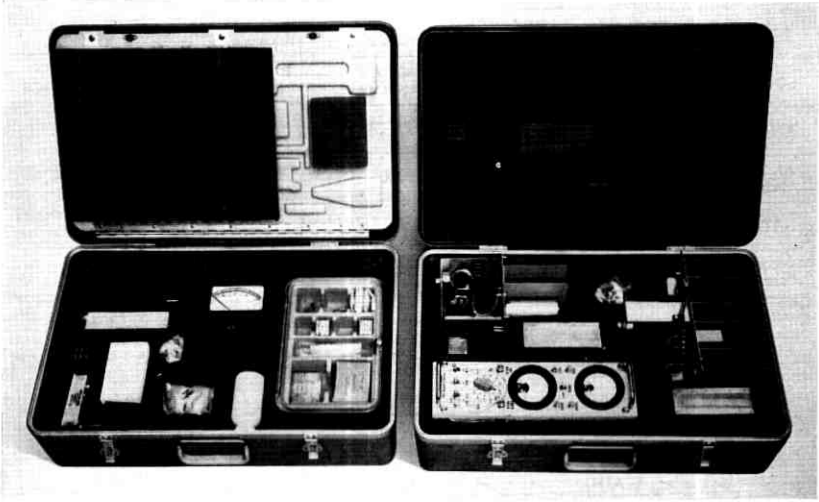


Fig. 44 — Spare parts case.

6.4 *Spare Equipment*

Special carrying cases for the test equipment, spare equipment units and components, and tools have been designed for convenient and safe transport of these items with the service man. These are shown in Fig. 44.

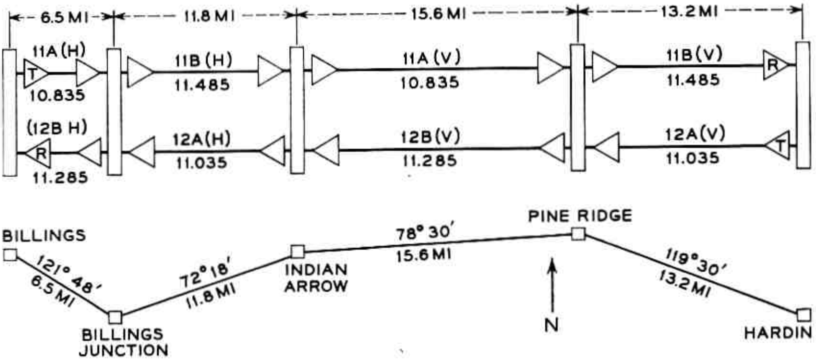
VII. APPLICATIONS OF THE RADIO SYSTEMS

7.1 *Billings-Hardin System*

The Billings-Hardin, Montana, TL radio route, shown in Fig. 45, extends eastward from Billings 48 miles to Hardin. This nondiversity radio system is multiplexed with 32 channels of ON carrier, providing additional facilities and back-up for an open wire line between Billings and Hardin. The radio repeaters lie along a plateau whose elevation is from 1000 to 2000 feet above the two terminals. Short wood-pole "H" towers, outside cabinets, and 5- and 10-foot paraboloidal antennas are used at the repeaters.

Performance on this system has been excellent. There have been no system failures or lost circuit time on this nondiversity system in its 15 months of operation.

STATION	BILLINGS MAIN	BILLINGS R1SE	BILLINGS R2SE	BILLINGS R3SE	HARDIN
LOCATION	BILLINGS	BILLINGS JUNCTION	INDIAN ARROW	PINE RIDGE	HARDIN CENTRAL OFFICE
ANTENNA	5′ DISH	(2) 5′ DISH	(2) 5′ DISH	(2) 10′ DISH	10′ DISH
ANTENNA HEIGHT	ROOF	40′	20′	20′	71′ APPROX.
TOWER	A FRAME	POLE H FRAME	H FRAME	H FRAME	B SELF SUPPORT (ROOF 16′)
TOWER HEIGHT	—	40′	20′	20′	60′



TERMINALS	LATITUDE	LONGITUDE	GROUND ELEVATION
BILLINGS	45° 46′ 33″	108° 30′ 34″	3126′
BILLINGS JUNCTION	45° 43′ 44″	108° 23′ 42″	4003′
INDIAN ARROW	45° 46′ 48″	108° 09′ 24″	4038′
PINE RIDGE	45° 49′ 42″	107° 50′ 42″	4081′
HARDIN	45° 43′ 52″	107° 35′ 57″	2902′

Fig. 45 — Billings-Hardin TL system.

7.2 *Denver-Limon-Burlington System*

Originally, this 8-hop TL system, shown in Fig. 46, extended eastward from Denver 85 miles to Limon and from Limon eastward 78 miles to Burlington. This system has now been extended two more hops to Cheyenne Wells. Initially equipped with 24 channels of ON, this rapidly growing system now carries 36 channels between Denver and Limon and 52 channels between Limon and Burlington. This system employs one-for-one frequency diversity protection.

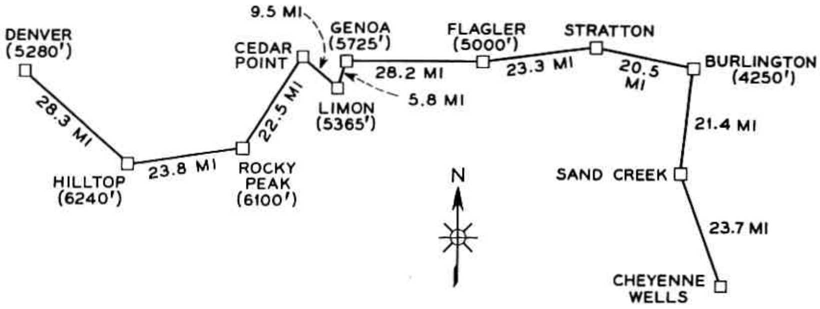


Fig. 46 — Denver-Limon-Burlington TL system.

VIII. ACKNOWLEDGMENT

The system described here is the result of the efforts of many departments of Bell Telephone Laboratories, including research, systems development, device development, outside plant, and system engineering.

REFERENCES

1. Ruthroff, C. L., and Tillotson, L. C., An Experimental "Short Hop" Microwave System, Bell Laboratories Record, **38,** June, 1960, p. 202.
2. Hathaway, S. D., and Evans, H. W., Radio Attenuation at 11 kmc and Some Simplifications Affecting Relay System Engineering, B.S.T.J., **38,** January, 1959, pp. 73–97.
3. Lewis, W. D., and Tillotson, L. C., A Nonreflecting Branching Filter for Microwaves, B.S.T.J., **27,** January, 1948, pp. 83–95.
4. Gucker, G. B., Long-Term Frequency Stability for a Reflex Klyston without the Use of External Cavities, B.S.T.J., **41,** May, 1962, pp. 945–958.
5. Ballentine, W. E., Saari, V. R., and Witt, F. J., The Solid-State Receiver in the TL Radio System, B.S.T.J., **41,** November, 1962, pp. 1831–1864.
6. Ruthroff, C. L., Amplitude Modulation Suppression in FM Systems, B.S.T.J., **37,** July, 1958, pp. 1023–1046.

Spectral Density and Autocorrelation Functions Associated with Binary Frequency-Shift Keying

By W. R. BENNETT and S. O. RICE

(Manuscript received May 15, 1963)

General equations are derived for the spectral density and autocorrelation functions of a wave train consisting of sine-wave segments with constant amplitude. The frequency of a segment may be either f_1 or f_2. At regularly spaced intervals the frequency is switched or not switched according to a random choice. This type of wave occurs when a random series of marks and spaces is sent by frequency-shift keying. The results fall into two main classes — namely, that of discontinuous phase at the transitions, which is the typical situation in switching between two independent oscillators; and that of continuous phase at the transitions, which is more usually applicable when the frequency of a single oscillator is changed. Individual treatment is given of the various special cases which arise when integral relationships between the marking, spacing, signaling, and shift frequencies exist. No restriction is made on the relative magnitudes of the different frequencies involved.

I. INTRODUCTION

The spectral density function, or power spectrum, of a random sequence of signals defines the distribution of average signal power versus frequency. This information is useful in system design because it indicates the frequency band of most importance, the amount of average total power in any frequency interval, and the interference which may result in other systems. It does not tell us how much distortion the signals suffer when the channel does not pass all the frequencies represented, nor does it tell us about important spectral components which may be associated with unlikely but possible specific signal sequences. Keeping these limitations in mind, we still find the spectral density to have merit as a descriptive parameter of the system.

Another important function is the autocorrelation, which is the time domain analog of the spectral density. Because of the Fourier transform relationship between the two functions, either can be used as an auxiliary step in computing the other. The autocorrelation function is useful in its own right in signal analysis and can be made the basis of control operations.

The present paper presents results on the spectral density and autocorrelation functions associated with binary frequency modulation systems. There are two general types of operation, which in terms of apparatus may be classified as (a) switching between two oscillators, and (b) changing the frequency of a single oscillator. The mathematical distinction between these two cases will be considered here as shifting the frequency with discontinuous or continuous phase respectively.

The case of discontinuous phase, which is appropriate when we switch between independent marking and spacing oscillators, is the simpler one to analyze. We assume that the oscillators deliver equal amplitude and that each oscillator preserves its own coherence in time, i.e., that the two frequencies are constant. The waveforms in intervals containing the marking frequency, say, are segments of a sine wave having the marking frequency and extending throughout all time. One would intuitively expect, therefore, to find discrete spectral lines at the marking and spacing frequencies. Analysis verifies that if mark and space signals have independent equal probabilities, each of the two discrete components has half the amplitude of the complete FSK wave.

In addition there is a continuous spectrum consisting of switching function spectra centered at the marking and spacing frequencies. Since the signal wave is discontinuous at the switching instants, the associated voltage spectra fall off only as $1/f$ at high frequencies. This means that the spectral density function ultimately falls off as the inverse square of the frequency. Degenerate cases arise when there are commensurable relationships among the marking, spacing, and signaling frequencies. If these special relations are such as to produce continuous phase at the transitions, the resulting continuity in the signal wave leads to an ultimate inverse fourth-power variation of spectral density with frequency. If in addition the derivative of the phase is continuous, an inverse sixth power is obtained at remote frequencies.

In the case in which the frequency shifting is done with continuous phase, the signal wave is continuous at all times. The spectral density function must, therefore, fall off at least as fast as the inverse fourth power at frequencies remote from the center of the signal band. This is in accordance with the well-known fact that frequency-shift keying with continuous phase does not produce as much interference outside the signal band as the discontinuous case. The analysis of the continuous-phase case is considerably more difficult, even though the final results are of fairly simple form.

It was found necessary to distinguish between four possible cases. In the most general of these, in which there are no degenerate relationships among the three frequencies involved, the line spectra completely disappear and the spectral density function is continuous at all frequencies. When the difference between the marking and spacing frequencies is a multiple of the signaling frequency, defined as the sum of the number of marks and number of spaces per second, line spectral terms appear at the marking and spacing frequencies, and the continuous part of the spectral density function changes its form. The continuous part of the spectrum in this case is found to depend also on whether or not the sum of marking and spacing frequencies is a multiple of the signaling rate. A curious behavior occurs when the frequency shift is an odd multiple of half the signaling rate. Here no discrete components appear in the spectrum, but the continuous spectral distribution undergoes a sudden change relative to that at infinitesimally close values of frequency shift not possessing the critical property.

Table II given in Section VI lists the various cases together with the corresponding equation numbers for the associated spectral densities and autocorrelation functions. Section V gives illustrative curves of these functions for various relations among the marking, spacing, and signaling frequencies.

II. DISCONTINUOUS PHASE

Before considering the frequency-shift keying problem of this section, it is helpful to develop some general results applicable to random switching between two waves. Let $y(t)$ be a given function of time which is bounded for all $t \geqq 0$ and such that the limits later encountered exist. Let $x(t)$ be a random telegraph wave defined by

$$x(t) = x_n , \qquad nT \leqq t < (n+1)T \tag{1}$$

where $n = 0, 1, 2, \cdots$, and the x_n's are independent random variables which assume the values ± 1 with equal probability. We calculate the spectral density $w_v(f)$ and the autocorrelation $R_v(\tau)$ of

$$v(t) = x(t)y(t). \tag{2}$$

We note that $v(t)$ is generated from the source of $y(t)$ by inserting a reversing switch for which a random choice between positions is made at instants T apart.

The spectral density of $v(t)$ is given by the limit of $2\langle| S_N(f) |^2\rangle/NT$ as $N \rightarrow \infty$. Here $\langle \quad \rangle$ denotes "ensemble average" and

$$\begin{aligned} S_N(f) &= \int_0^{NT} e^{-j\omega t}\, v(t)\, dt \qquad \omega = 2\pi f \\ &= \sum_{n=0}^{N-1} x_n e^{-j\omega nT} \int_0^T dt\, e^{-j\omega t}\, y(nT + t). \end{aligned} \tag{3}$$

Since $\langle x_n x_m \rangle$ is 1 if $n = m$ and 0 if $n \neq m$

$$w_v(f) = \lim_{N\to\infty} \frac{2}{NT} \sum_{n=0}^{N-1} \left| \int_0^T dt\, e^{-j\omega t}\, y(nT + t) \right|^2. \tag{4}$$

The autocorrelation function $R_v(\tau)$ may be calculated either by averaging $\langle v(t)v(t + \tau) \rangle$ over all $t \geqq 0$ with τ held fixed, or by taking the Fourier transform of (4); thus

$$R_v(\tau) = \int_0^\infty w_v(f) \cos \omega\tau\, df. \tag{5}$$

Both methods show that $R_v(\tau) = 0$ for $| \tau | \geqq T$ and

$$R_v(\tau) = \frac{1}{T} \int_0^{T-\tau} dt \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} y(nT + t)\, y(nT + t + \tau) \tag{6}$$

for $0 \leqq \tau < T$.

Return now to the frequency-shift keying problem and consider the signal wave

$$\begin{aligned} u(t) = \begin{matrix} u_1(t) \\ \text{or} \\ u_2(t) \end{matrix} \qquad & \begin{matrix} nT \leqq t < (n + 1)T \\ \\ n = 0, 1, 2, \cdots \end{matrix} \\ u_k(t) = A \cos (\omega_k t + \theta_k) \qquad & k = 1, 2 \end{aligned} \tag{7}$$

where the choice is made independently and with equal probability for each interval of length T. The signaling frequency is $\omega_s = 2\pi/T$ rad/sec. The wave $u(t)$ may be written as

$$u(t) = u_+(t) + x(t)u_-(t) \tag{8}$$

where $x(t)$ is defined by (1) and

$$\begin{aligned} u_+(t) &= \tfrac{1}{2}u_1(t) + \tfrac{1}{2}u_2(t) \\ u_-(t) &= \tfrac{1}{2}u_1(t) - \tfrac{1}{2}u_2(t). \end{aligned} \tag{9}$$

Thus, $u(t)$ is the sum of two steady-state cosine waves, given by $u_+(t)$, and a random component given by

$$v(t) = u(t) - u_+(t) = x(t)u_-(t). \tag{10}$$

The random component assumes the values $\pm u_-(t)$.

When $u_-(t)$ is identified with $y(t)$, the spectral density $w_v(f)$ of $v(t)$ can be obtained from (4). Since the algebra is rather tedious the procedure will be merely sketched. The integral in (4) is now

$$\int_0^T dt\, e^{-j\omega t}\, u_-\,(nT + t) = \frac{A}{4}\, e^{-j\omega T/2} \tag{11}$$
$$\{g(\omega - \omega_1) \exp\,[j\theta_1 + j\omega_1 T(n + \tfrac{1}{2})] + \cdots\}$$

where the braces contain four terms similar to the first and

$$g(a) = \frac{\sin\,(aT/2)}{(a/2)}. \tag{12}$$

The cosines in $u_-(nT + t)$ are expressed in exponential form before integrating.

Multiplying (11) by its conjugate complex gives 16 terms. Substitution in (4) gives rise to expressions of the form

$$\gamma_N(\lambda) = \frac{1}{N}\sum_{n=0}^{N-1} e^{j\lambda n} = \frac{1 - e^{j\lambda N}}{N(1 - e^{j\lambda})}$$
$$\gamma(\lambda) = \lim_{N\to\infty} \gamma_N(\lambda) = \begin{cases} 1, \text{ if } \lambda = 2\pi m \\ 0, \lambda \text{ real and } \neq 2\pi m \end{cases} \tag{13}$$

where m is an integer. A typical value of λ is $2\omega_1 T$. At this stage $w_v(f)$ is given by

$$\begin{aligned}\frac{8Tw_v(f)}{A^2} &= g^2(\omega - \omega_1) + g^2(\omega + \omega_1) + g^2(\omega - \omega_2) + g^2(\omega + \omega_2) \\ &+ 2\cos\,(2\theta_1 + \omega_1 T)g(\omega - \omega_1)g(\omega + \omega_1)\gamma(2\omega_1 T) \\ &+ 2\cos\,(2\theta_2 + \omega_2 T)g(\omega - \omega_2)g(\omega + \omega_2)\gamma(2\omega_2 T) \\ &- 2\cos\left[\theta_2 + \theta_1 + \left(\frac{\omega_2 + \omega_1}{2}\right)T\right][g(\omega - \omega_1)g(\omega + \omega_2) \\ &+ g(\omega + \omega_1)g(\omega - \omega_2)]\gamma(\omega_2 T + \omega_1 T) - 2\cos\left[\theta_2 - \theta_1\right. \\ &\left. + \left(\frac{\omega_2 - \omega_1}{2}\right)T\right][g(\omega + \omega_1)g(\omega + \omega_2) \\ &+ g(\omega - \omega_1)g(\omega - \omega_2)]\gamma(\omega_2 T - \omega_1 T).\end{aligned} \tag{14}$$

Some further reduction leads to the final result

$$\begin{aligned}
2TA^{-2}w_v(f) &= \frac{G(\omega-\omega_1)}{(\omega-\omega_1)^2}+\frac{G(\omega+\omega_1)}{(\omega+\omega_1)^2}+\frac{G(\omega-\omega_2)}{(\omega-\omega_2)^2}+\frac{G(\omega+\omega_2)}{(\omega+\omega_2)^2}\\
&\quad+\frac{2\cos 2\theta_1 G(\omega-\omega_1)}{\omega^2-\omega_1^2}\gamma(2\omega_1T)+\frac{2\cos 2\theta_2 G(\omega-\omega_2)}{\omega^2-\omega_2^2}\gamma(2\omega_2T)\\
&\quad-2\cos(\theta_2+\theta_1)\left[\frac{G(\omega-\omega_1)}{(\omega-\omega_1)(\omega+\omega_2)}\right. \qquad (15)\\
&\quad\left.+\frac{G(\omega-\omega_2)}{(\omega+\omega_1)(\omega-\omega_2)}\right]\gamma(\omega_2T+\omega_1T)-2\cos(\theta_2-\theta_1)\\
&\quad\cdot\left[\frac{G(\omega-\omega_1)}{(\omega-\omega_1)(\omega-\omega_2)}+\frac{G(\omega+\omega_2)}{(\omega+\omega_1)(\omega+\omega_2)}\right]\gamma(\omega_2T-\omega_1T)
\end{aligned}$$

where

$$G(a)=\sin^2(aT/2) \qquad (16)$$

and $\gamma(2\omega_1T)$, $\gamma(2\omega_2T)$, $\gamma(\omega_2T+\omega_1T)$, $\gamma(\omega_2T-\omega_1T)$ are zero except when $2\omega_1/\omega_s$, $2\omega_2/\omega_s$, $(\omega_2+\omega_1)/\omega_s$, $(\omega_2-\omega_1)/\omega_s$, respectively, are integers (in which case the corresponding γ's are unity).

In the special case in which $2\omega_1/\omega_s = m$ and $2\omega_2/\omega_s = l$, with l and m integers and $l+m$ even, all of the γ's in (15) are equal to unity and the G's are equal to $\sin^2(\omega T/2)$ if m is even and to $\cos^2(\omega T/2)$ if m is odd. In particular, when $\theta_1=\theta_2=0$ the following simple result is obtained

$$w_v(f)=\frac{2A^2(\omega_2^2-\omega_1^2)^2\omega^2}{T(\omega^2-\omega_1^2)^2(\omega^2-\omega_2^2)^2}\times\begin{Bmatrix}\sin^2\dfrac{\omega T}{2}, m\text{ even}\\[2ex]\cos^2\dfrac{\omega T}{2}, m\text{ odd}\end{Bmatrix}. \qquad (17)$$

This spectral density function varies as $1/\omega^6$ for large ω, showing that the waveform of the signal is continuous and has a continuous derivative.

The spectral density function of $u(t)$ as defined by (7) differs from that of $v(t)$ by the presence of sinusoidal components of amplitude $A/2$ and frequencies ω_1 and ω_2. That is

$$w_u(f)=w_v(f)+\frac{A^2}{8}\delta(f-f_1)+\frac{A^2}{8}\delta(f-f_2). \qquad (18)$$

It can be shown from (6) and (18) that the autocorrelation functions $R_u(\tau)$ and $R_v(\tau)$ of $u(t)$ and $v(t)$ respectively are given by the following set of equations

$$
\begin{aligned}
R_u(\tau) &= R_v(\tau) + \frac{A^2}{8}\cos\omega_1\tau + \frac{A^2}{8}\cos\omega_2\tau \\
R_v(\tau) &= 0, \qquad |\tau| > T \\
8TA^{-2}R_v(\tau) &= (T-\tau)(\cos\omega_1\tau + \cos\omega_2\tau) \\
&\quad - \frac{\sin\omega_1\tau}{\omega_1}\cos 2\theta_1\,\gamma(2\omega_1T) - \frac{\sin\omega_2\tau}{\omega_2}\cos 2\theta_2\,\gamma(2\omega_2T) \\
&\quad + \frac{2(\sin\omega_2\tau + \sin\omega_1\tau)}{\omega_2+\omega_1}\cos(\theta_2+\theta_1)\,\gamma(\omega_2T+\omega_1T) \\
&\quad + \frac{2(\sin\omega_2\tau - \sin\omega_1\tau)}{\omega_2-\omega_1}\cos(\theta_2-\theta_1)\,\gamma(\omega_2T-\omega_1T) \qquad 0 \leqq \tau \leqq T \\
R_v(-\tau) &= R_v(\tau).
\end{aligned}
\tag{19}
$$

III. CONTINUOUS PHASE

Consider the signal wave

$$
u(t) = \begin{matrix} A\cos(\omega_1 t + \theta_n) & nT \leqq t < (n+1)T \\ \text{or} & \\ A\cos(\omega_2 t + \phi_n) & n = 0, 1, 2, \cdots \end{matrix}
\tag{20}
$$

where the choice is made independently and with equal probability for each interval of length T. The initial values at $t = 0$ of the phase are $\theta_0 = \phi_0 = \phi$, and the succeeding values θ_n, ϕ_n are to be chosen so as to make the phase of $u(t)$ continuous at the transition points.

Let

$$\alpha = \tfrac{1}{2}(\omega_2 + \omega_1) \qquad \beta = \tfrac{1}{2}(\omega_2 - \omega_1). \tag{21}$$

Then

$$\omega_1 = \alpha - \beta \qquad \omega_2 = \alpha + \beta. \tag{22}$$

Set

$$u(t) = A\cos B_n(t), \quad nT \leqq t < (n+1)T, \qquad n = 0, 1, 2, \cdots \tag{23}$$

$$B_0(t) = (\alpha + x_1\beta)t + \phi \tag{24}$$

$$B_n(t) = \alpha t + x_{n+1}\beta(t - nT) + \phi + \beta T\sum_{r=1}^{n} x_r, \qquad n > 0. \tag{25}$$

We assume $x_1, x_2, \cdots$ to be independent random variables, each of which is equally likely to have the value $+1$ or -1. We verify that within the interval beginning at $t = nT$, the frequency is $\alpha + x_{n+1}\beta$, which is equal to ω_2 if $x_{n+1} = +1$ and equal to ω_1 if $x_{n+1} = -1$. Therefore, the function $B_n(t)$ satisfies the condition of an equiprobable choice between the two frequencies in each interval. We also note that the phase at the beginning of the typical interval is

$$B_n(nT) = \alpha nT + \phi + \beta T \sum_{r=1}^{n} x_r \tag{26}$$

while the phase at the end of the previous interval is

$$\begin{aligned} B_{n-1}(nT) &= \alpha nT + \phi + x_n\beta[nT - (n - 1)T] + \beta T \sum_{r=1}^{n-1} x_r \\ &= B_n(nT). \end{aligned} \tag{27}$$

Thus the function $B_n(t)$ also satisfies the required condition of continuous phase.

We have evaluated the spectral density function of $u(t)$ by two different methods, namely

(a) by taking the Fourier transform of the autocorrelation function, and

(b) by direct evaluation from the Fourier transform of the signal wave over a long time interval. The two methods are of comparable difficulty. The same results are finally obtained by both procedures, although the agreement is not immediately evident from the expressions which emerge naturally from the two sets of calculations.

To evaluate the autocorrelation function, we first calculate the average value of $u(t)u(t + \tau)$ over the ensemble at fixed t. Set $\tau \geqq 0$ and define k as the member of the set $0, 1, 2, \cdots$ satisfying the inequality

$$(n + k)T \leqq t + \tau < (n + k + 1)T \tag{28}$$

with n defined by $nT \leqq t < (n + 1)T$. Let $E_k(t, \tau)$ represent the mathematical expectation of $u(t)u(t + \tau)$ with t and τ fixed. If $k = 0$, the values of t and $t + \tau$ lie in the same signaling interval and the same function $B_n(t)$ applies to both. When $k > 0$ we use the function $B_n(t)$ for $u(t)$ and the function $B_{n+k}(t)$ for $u(t + \tau)$.

We calculate

$$
\begin{aligned}
\frac{2}{A^2} E_0(t, \tau) & \\
&= 2\langle \cos B_n(t) \cos B_n(t + \tau) \rangle \\
&= \operatorname{Re} \langle [\exp (jB_n(t + \tau) - jB_n(t)) + \exp (jB_n(t + \tau) + jB_n(t))] \rangle \\
&= \operatorname{Re} \Big\langle \Big[\exp (j\alpha\tau) \exp (jx_{n+1}\beta\tau) + \exp (j\alpha(2t + \tau) + j2\phi) \\
&\qquad \cdot \exp (jx_{n+1}\beta(2t + \tau - 2nT)) \prod_{r=1}^{n} \exp (j2x_r\beta T) \Big] \Big\rangle .
\end{aligned} \tag{29}
$$

Since the x's are independent, the average of a product of functions in which the variables appear separately is equal to the product of the averages of the individual functions. Since each x has only two possible values with a probability of one-half for each, we evaluate the expectations of individual terms by inserting the sum of the two possible functions with weighting factor one-half. Performing the necessary operations, we find

$$
\begin{aligned}
\frac{2}{A^2} E_0(t, \tau) &= \cos \alpha\tau \cos \beta\tau \\
&\qquad + \cos (2\alpha\tau + \alpha\tau + 2\phi) \cos \beta(2t + \tau - 2nT) \cos^n 2\beta T.
\end{aligned} \tag{30}
$$

The corresponding calculation for $k > 0$ leads to the result

$$
\begin{aligned}
\frac{2}{A^2} E_k(t, \tau) &= \cos \beta(t + \tau - nT - kT) \\
\cos^{k-1} \beta T [\cos \alpha\tau & \cos \beta(t - nT - T) + \cos (2\alpha t + \alpha\tau + 2\phi) \\
&\qquad \cos \beta(t - nT + T) \cos^n 2\beta T].
\end{aligned} \tag{31}
$$

The lag interval τ is bounded between adjacent multiples of T by defining m as the member from the set $0, 1, 2, \cdots$ which satisfies $mT \leqq \tau < (m + 1)T$. We observe that the number k defined by (28) is related to m as follows

$$
k = \begin{cases} m, & nT \leqq t < (m + n + 1)T - \tau \\ m + 1, & (m + n + 1)T - \tau \leqq t < (n + 1)T. \end{cases} \tag{32}
$$

The autocorrelation function $R_u(\tau)$ is the average of $\langle u(t)u(t + \tau) \rangle$ taken from $t = 0$ to $t = \infty$, i.e.

$$\begin{aligned}R_u(\tau) &= \lim_{N\to\infty} \frac{1}{NT}\int_0^{NT} \langle u(t)u(t+\tau)\rangle\, dt \\ &\qquad = \lim_{N\to\infty} \frac{1}{NT}\sum_{n=0}^{N-1}\left[\int_{nT}^{(m+n+1)T-\tau} E_m(t,\tau)\, dt\right. \\ &\qquad\qquad \left.+ \int_{(m+n+1)T-\tau}^{(n+1)T} E_{m+1}(t,\tau)\, dt\right] \qquad (33) \\ &= \lim_{N\to\infty} \frac{1}{NT}\sum_{n=0}^{N-1}\left[\int_0^{(m+1)T-\tau} E_m(t+nT,\tau)\, dt\right. \\ &\qquad\qquad \left.+ \int_{(m+1)T-\tau}^{T} E_{m+1}(t+nT,\tau)\, dt\right].\end{aligned}$$

As indicated in (30) and (31), the case of $m = 0$, i.e., $0 < \tau < T$, requires a separate treatment from that of $m > 0$. In order to perform the integrations indicated in (33) it is convenient to expand (30) and (31) into the sum of terms in which t appears only once. The expanded equations are, with $0 \leqq t < T$

$$\begin{aligned}\frac{2}{A^2} E_0\,(t+nT,\,\tau) &= \cos\alpha\tau\cos\beta\tau \\ &\quad + \tfrac{1}{2}\cos^n 2\beta T \cos\left[(\alpha+\beta)(2t+\tau) + 2\phi + 2n\alpha T\right] \\ &\quad + \tfrac{1}{2}\cos^n 2\beta T \cos\left[(\alpha-\beta)(2t+\tau) + 2\phi + 2n\alpha T\right] \qquad (34)\end{aligned}$$

$$\begin{aligned}\frac{4}{A^2} E_k(t+nT,\,\tau) &= \cos\alpha\tau\cos^{k-1}\beta T\,[\cos\beta(\tau + T - kT) \\ &\qquad + \cos\beta(2t+\tau-T-kT)] \\ &\quad + \cos^{k-1}\beta T\cos^n 2\beta T\,\{\cos\beta[\tau-(k+1)T]\cos[\alpha(2t+\tau) \qquad (35) \\ &+ 2\phi + 2n\alpha T] + \tfrac{1}{2}\cos[(\alpha+\beta)(2t+\tau) + 2\phi + 2n\alpha T - (k-1)\beta T] \\ &\quad + \tfrac{1}{2}\cos[(\alpha-\beta)(2t+\tau) + 2\phi + 2n\alpha T + (k-1)\beta T]\}.\end{aligned}$$

We note the possibility that some of the terms will not contribute anything to the result after the limiting process in (33) is performed. The expressions in (34) and (35) can be divided into two classes: those which do not contain n and those which contain n in the form of a factor of type $\cos^n 2\beta T \cos(\psi + 2n\alpha T)$. In the former group, we sum N equal terms and divide by NT; hence the summing and limiting operations are equivalent to a division by T. In the other group, the limit is zero if the sum remains finite as N becomes indefinitely large. The only case in which the sum does not remain finite is that in which the terms are equal for all n. Equality occurs if αT and βT are both multiples of π and also if αT and βT are both odd multiples of $\pi/2$. If neither of these conditions exists, the contributions of the terms which depend on n are zero. We shall call this the incommensurable case and shall treat it first. Afterward, we shall consider the effect of commensurable relations.

Except for the cases we have specifically excluded, we can now write for $0 < \tau < T$

$$\begin{aligned} R_u(\tau) &= \frac{A^2}{2T}\int_0^{T-\tau} \cos\alpha\tau\cos\beta\tau\,dt \\ &+ \frac{A^2}{4T}\int_{T-\tau}^{T} \cos\alpha\tau[\cos\beta\tau + \cos\beta(2t+\tau-2T)]\,dt \\ &= \frac{A^2}{4\beta T}[\beta(2T-\tau)\cos\beta\tau + \sin\beta\tau]\cos\alpha\tau. \end{aligned} \tag{36}$$

For $\tau > T$

$$\begin{aligned} R_u(\tau) = \frac{A^2}{4T}\int_0^{(m+1)T-\tau} &\cos\alpha\tau\cos^{m-1}\beta T[\cos\beta(\tau+T-mT) \\ &+ \cos\beta(2t+\tau-T-mT)]\,dt \\ + \frac{A^2}{4T}\int_{(m+1)T-\tau}^{T} &\cos\alpha\tau\cos^{m}\beta T\,[\cos\beta(\tau-mT) \\ &+ \cos\beta(2t+\tau-2T-mT)]\,dt \\ = \frac{A^2}{4\beta T}\cos\alpha\tau\cos^{m-1}&\beta T[\beta T\cos\beta(\tau+T-mT) \\ &+ \beta(\tau-mT)\sin\beta T\sin\beta(\tau-mT) \\ &+ \sin\beta T\cos\beta(\tau-mT)]. \end{aligned} \tag{37}$$

The spectral density function $w_u(f)$ is given by

$$w_u(f) = 4\int_0^{\infty} R_u(\tau)\cos\omega\tau\,d\tau. \tag{38}$$

It is convenient to evaluate the integral in two parts

$$\begin{aligned} 4\int_0^T R_u(\tau)\cos\omega\tau\,d\tau = \frac{A^2}{2\beta T}\int_0^T &[\beta(2T-\tau)\cos\beta\tau + \sin\beta\tau] \\ &\cdot[\cos(\omega+\alpha)\tau + \cos(\omega-\alpha)\tau]\,d\tau \end{aligned} \tag{39}$$

and

$$
\begin{aligned}
4 \int_T^\infty & R_u(\tau) \cos \omega\tau \, d\tau \\
&= \frac{A^2}{\beta T} \sum_{m=1}^{\infty} \int_{mT}^{(m+1)T} \cos \alpha\tau \cos^{m-1} \beta T [\beta T \cos \beta(\tau + T - mT) \\
&\qquad + \beta(\tau - mT) \sin \beta T \sin \beta(\tau - mT) \\
&\qquad + \sin \beta T \cos \beta(\tau - mT)] \cos \omega\tau \, d\tau \\
&= \frac{A^2}{\beta T} \int_0^T [\beta T \cos \beta(\tau + T) + \beta\tau \sin \beta T \sin \beta\tau \\
&+ \sin \beta T \cos \beta\tau] \sum_{m=1}^{\infty} \cos \alpha(\tau + mT) \cos \omega(\tau + mT) \cos^{m-1} \beta T \\
&= \frac{A^2}{2\beta T} \int_0^T [\beta T \cos \beta(\tau + T) + \beta\tau \sin \beta T \sin \beta\tau \\
&\qquad + \sin \beta T \cos \beta\tau][G(\omega + \alpha, \tau) + G(\omega - \alpha, \tau)] \, d\tau
\end{aligned}
\tag{40}
$$

where

$$
G(y,\tau) = \frac{\cos y(\tau + T) - \cos \beta T \cos y\tau}{1 + \cos^2 \beta T - 2 \cos \beta T \cos yT}. \tag{41}
$$

The summation is performed by writing

$$
2 \cos \alpha(\tau + mT) \cos \omega(\tau + mT) \cos^{m-1} \beta T
$$

as the real part of

$$
[\exp (j(\omega + \alpha)(\tau + mT)) + \exp (j(\omega - \alpha)(\tau + mT))] \cos^{m-1} \beta T
$$

and thereby obtaining a geometric series. The series fails to converge when the absolute values of the individual terms are unity. For this reason, we must now exclude the case of βT equal to a multiple of π no matter what αT is. We have thus accumulated three special cases of commensurability to be given individual treatment later.

The remainder of the calculation is straightforward, but appears to lead into a morass of complication. The key to an end result of pleasing simplicity, which the autocorrelation method tends to conceal, is to arrange the work as follows

$$
\begin{aligned}
w_u(f) = \frac{A^2}{T} \Bigg[& \frac{H(\omega + \alpha)}{1 + \cos^2 \beta T - 2 \cos \beta T \cos (\omega + \alpha)T} \\
& + \frac{H(\omega - \alpha)}{1 + \cos^2 \beta T - 2 \cos \beta T \cos (\omega - \alpha)T} \Bigg].
\end{aligned}
\tag{42}
$$

The integral (40) splits naturally into this form. In the integral (39) we associate the term $\cos(\omega + \alpha)\tau$ with the first part of (42) and the term $\cos(\omega - \alpha)\tau$ with the second part. Then

$$\begin{aligned}
H(y) = \frac{1}{2\beta}\int_0^T & \{[1 + \cos^2 \beta T - 2\cos \beta T \cos yT] \\
& \cdot [\beta(2T - \tau)\cos \beta\tau + \sin \beta\tau] \cos y\tau \\
& + [\beta T \cos \beta(\tau + T) + \beta\tau \sin \beta T \sin \beta\tau \\
& + \sin \beta T \cos \beta\tau][\cos y\,(\tau + T) - \cos \beta T \cos y\tau]\}\, d\tau \\
= \frac{1}{2\beta}\int_0^T & \{(1 + \cos^2 \beta T)[\beta(2T - \tau)\cos \beta\tau + \sin \beta\tau] \\
& - \cos \beta T[\beta T \cos \beta(\tau + T) + \beta\tau \sin \beta T \sin \beta\tau \\
& + \sin \beta T \cos \beta\tau]\} \cos y\tau\, d\tau \\
& + \frac{1}{2\beta}\int_0^T \{\beta T \cos \beta(\tau + T) + \beta\tau \sin \beta T \sin \beta\tau \\
& + \sin \beta T \cos \beta\tau - \cos \beta T[\beta(2T - \tau)\cos \beta\tau \\
& + \sin \beta\tau]\} \cos y\,(\tau + T)\, d\tau \\
& - \frac{\cos \beta T}{2\beta}\int_0^T [\beta(2T - \tau)\cos \beta\tau \\
& \qquad\qquad + \sin \beta\tau] \cos y\,(\tau - T)\, d\tau.
\end{aligned} \tag{43}$$

In the second integral, substitute $\tau + T = \tau'$ and in the third integral substitute $T - \tau = \tau'$. Dropping the primes after the substitution and combining terms where possible, we then find that the result can be written in the form

$$H(y) = \frac{1}{2\beta}\int_0^T h_1(\tau) \cos y\tau\, d\tau + \frac{1}{2\beta}\int_T^{2T} h_2(\tau)\cos y\tau\, d\tau \tag{44}$$

where

$$\begin{aligned}
h_1(\tau) = 2\beta(T - \tau)\cos \beta\tau - \beta\tau \cos \beta(\tau - 2T) + \sin \beta(\tau - 2T) \\
+ 2 \sin \beta\tau
\end{aligned} \tag{45}$$

$$h_2(\tau) = \beta(\tau - 2T)\cos \beta(\tau - 2T) - \sin \beta(\tau - 2T). \tag{46}$$

The integration in (44) leads to the result

$$H(y) = 2 \sin^2 \frac{y+\beta}{2} T \sin^2 \frac{y-\beta}{2} T \left(\frac{1}{y-\beta} - \frac{1}{y+\beta}\right)^2. \quad (47)$$

The complete equation for the spectral density can now be written in the form

$$\begin{aligned} w_u(f) = {} & \frac{2A^2 \sin^2\left(\frac{\omega - \omega_1}{2}\right) T \sin^2\left(\frac{\omega-\omega_2}{2}\right) T}{T[1 - 2\cos(\omega-\alpha)T\cos\beta T + \cos^2\beta T]} \\ & \qquad \cdot \left[\frac{1}{\omega-\omega_1} - \frac{1}{\omega-\omega_2}\right]^2 \\ & + \frac{2A^2 \sin^2\left(\frac{\omega + \omega_1}{2}\right) T \sin^2\left(\frac{\omega+\omega_2}{2}\right) T}{T[1 - 2\cos(\omega+\alpha)T\cos\beta T + \cos^2\beta T]} \\ & \qquad \cdot \left[\frac{1}{\omega+\omega_1} - \frac{1}{\omega+\omega_2}\right]^2. \end{aligned} \quad (48)$$

The intermediate step converting the original integral (43) to the form (44) is very helpful in reducing the labor required to obtain the final form (47). Incidentally, (44) shows that the function $H(y)$ is the Fourier transform of a function of τ which is time-limited to the range 0 to $2T$. We also point out that discrete components do not appear in $u(t)$. The spectral density function is continuous at all frequencies and varies as the inverse fourth power of the frequency at frequencies remote from ω_1 and ω_2. The latter property must exist because the waveform of the signal is continuous at all times.

We now return to the three cases of commensurability which we found necessary to avoid in deriving the general result of (48). When βT is a multiple of π, an examination of (25) shows that $B_n(t)$ differs from $\alpha t + x_{n+1}\beta t + \phi$ by a multiple of 2π, and hence

$$u(t) = A \cos(\alpha t + x_{n+1}\beta t + \phi), \qquad nT \leqq t < (n+1)T. \quad (49)$$

Comparison with (7) shows that we now have one of the degenerate cases in which the generally discontinuous phase becomes continuous: i.e., that case in which $\theta_1 = \theta_2 = \phi$ and $\omega_2 - \omega_1 = r\omega_s$, r being an integer.

When βT is a multiple of π and αT is not a multiple of π, we have $\omega_2 - \omega_1 = r\omega_s$, $\omega_2 + \omega_1 \neq l\omega_s$, and it follows that $2\omega_1/\omega_s$, $2\omega_2/\omega_s$ are not integers. Setting $\gamma(2\omega_1 T)$, $\gamma(2\omega_2 T)$, $\gamma(\omega_2 T + \omega_1 T)$ to zero in (15) and using

$$G(\omega - \omega_1) = \sin^2\left[\left(\frac{\omega - \omega_1}{2}\right)T\right] = \sin^2\left[\left(\frac{\omega - \alpha}{2}\right)T + \frac{r\pi}{2}\right]$$

$$= \begin{matrix} \sin^2\left[\dfrac{\omega - \alpha}{2}\right]T, & r \text{ even} \\ \cos^2\left[\dfrac{\omega - \alpha}{2}\right]T, & r \text{ odd} \end{matrix} \tag{50}$$

$$G(\omega - \omega_2) = G(\omega - \omega_1) \tag{51}$$

together with the corresponding expressions for $G(\omega + \omega_1)$, $G(\omega + \omega_2)$, gives $w_v(f)$ in

$$w_u(f) = w_r(f) + \frac{A^2}{8}\left[\delta(f - f_1) + \delta(f - f_2)\right] \tag{52}$$

where the notation is the same as in (18). When r is an even integer we calculate

$$w_v(f) = \frac{A^2}{2T}\left\{\sin^2\left(\frac{\omega - \alpha}{2}\right)T\left[\frac{1}{\omega - \omega_1} - \frac{1}{\omega - \omega_2}\right]^2 + \sin^2\left(\frac{\omega + \alpha}{2}\right)T\left[\frac{1}{\omega + \omega_1} - \frac{1}{\omega + \omega_2}\right]^2\right\} \tag{53}$$

and when r is odd

$$w_v(f) = \frac{A^2}{2T}\left\{\cos^2\left(\frac{\omega - \alpha}{2}\right)T\left[\frac{1}{\omega - \omega_1} - \frac{1}{\omega - \omega_2}\right]^2 + \cos^2\left(\frac{\omega + \alpha}{2}\right)T\left[\frac{1}{\omega + \omega_1} - \frac{1}{\omega + \omega_2}\right]^2\right\}. \tag{54}$$

In the next case both βT and αT are multiples of π; i.e., $\omega_2 - \omega_1 = r\omega_s$, $\omega_2 + \omega_1 = l\omega_s$, and $2\omega_1/\omega_s$, $2\omega_2/\omega_s$ are integers. Now all of the terms in (15) must be considered and

$$\begin{aligned} G(\omega - \omega_1) &= G(\omega - \omega_2) \\ &= G(\omega + \omega_1) = G(\omega + \omega_2) = \begin{cases} \sin^2 \dfrac{\omega T}{2} \text{ for } l - r \text{ even} \\ \cos^2 \dfrac{\omega T}{2} \text{ for } l - r \text{ odd.} \end{cases} \end{aligned} \tag{55}$$

Combining terms in (15) gives

$$w_v(f) = \frac{A^2}{2T}\begin{Bmatrix}\sin^2\dfrac{\omega T}{2} \text{ for } l - r \text{ even}\\ \\ \cos^2\dfrac{\omega T}{2} \text{ for } l - r \text{ odd}\end{Bmatrix}\left[\left(\frac{1}{\omega-\omega_1}-\frac{1}{\omega-\omega_2}\right)^2 + \left(\frac{1}{\omega+\omega_1}-\frac{1}{\omega+\omega_2}\right)^2 + 2\left(\frac{1}{\omega-\omega_1}-\frac{1}{\omega-\omega_2}\right)\left(\frac{1}{\omega+\omega_1}-\frac{1}{\omega+\omega_2}\right)\cos 2\phi\right]. \tag{56}$$

In the last of the exceptional cases, both αT and βT are odd multiples of $\pi/2$. That is,

$$\begin{aligned} 2\alpha T &= (2l+1)\pi, \qquad l = 0, 1, 2, \cdots \\ 2\beta T &= (2r+1)\pi, \qquad r = 0, 1, 2, \cdots. \end{aligned} \tag{57}$$

Equivalently

$$\begin{aligned} \omega_2/\omega_s &= (l + r + 1)/2 \\ \omega_1/\omega_s &= (l - r)/2. \end{aligned} \tag{58}$$

The value of E_0 obtained by substituting (57) in (34) is

$$\begin{aligned} \frac{2}{A^2}E_0(t + nT, \tau) = \cos\alpha\tau\cos\beta\tau &+ \tfrac{1}{2}\cos[(\alpha+\beta)(2t+\tau)+2\phi] \\ &+ \tfrac{1}{2}\cos[(\alpha-\beta)(2t+\tau)+2\phi]. \end{aligned} \tag{59}$$

Substituting $k = 1$ in (35) and then inserting the special conditions of (57) we obtain

$$\begin{aligned} \frac{4}{A^2}E_1(t + nT, \tau) &= \cos\alpha\tau[\cos\beta\tau - \cos\beta(2t+\tau)] \\ &- \cos\beta\tau\cos[\alpha(2t+\tau)+2\phi] + \tfrac{1}{2}\cos[(\alpha+\beta)(2t+\tau)+2\phi] \\ &+ \tfrac{1}{2}\cos[(\alpha-\beta)(2t+\tau)+2\phi]. \end{aligned} \tag{60}$$

Since βT is an odd multiple of $\pi/2$, the value of $\cos\beta T$ is zero. For $k > 1$ the right-hand member of (35) contains $\cos\beta T$ as a factor. Hence $E_k(t + nT, \tau)$ vanishes for $k > 1$. It follows that the autocorrelation function vanishes for $\tau > 2T$ and there can be no discrete sinusoidal components in the spectral density function.

A better understanding of this remarkable behavior can be obtained by examination of a particular signaling interval in which we are equally likely to find one of the two possible waves $A \cos(\omega_1 t + \psi_1)$ and

$A \cos(\omega_2 t + \psi_2)$. At the next switching instant, say $t = nT$, we either continue with the same wave or shift to the other frequency with continuous phase. There are thus four possible waves in the succeeding interval, viz.,

$$\begin{aligned} &(i)\ A \cos(\omega_1 t + \psi_1) \\ &(ii)\ A \cos[\omega_2 t + \psi_1 - (\omega_2 - \omega_1)nT] \\ &(iii)\ A \cos[\omega_1 t + \psi_2 + (\omega_2 - \omega_1)nT] \\ &(iv)\ A \cos(\omega_2 t + \psi_2). \end{aligned}$$

Since $(\omega_2 - \omega_1)T$ is an odd multiple of π, the second and third terms can be written as $(-)^n A \cos(\omega_2 t + \psi_1)$ and $(-)^n A \cos(\omega_1 t + \psi_2)$ respectively. Waves (i) and (ii) are possible when the initial frequency is ω_1, and waves (iii) and (iv) are possible when the initial frequency is ω_2. Now examining the possibilities after the next subsequent switching instant, $t = (n + 1)T$, we find that for the original frequency equal to ω_1, the waves (i) and (ii) can change to the four possible waves:

$$\begin{aligned} &(i)\ A \cos(\omega_1 t + \psi_1) \\ &(ii)\ (-)^{n+1} A \cos(\omega_2 t + \psi_1) \\ &(iii)\ (-)^n A \cos(\omega_2 t + \psi_1) \\ &(iv)\ -A \cos(\omega_1 t + \psi_1). \end{aligned}$$

It will be noted that the first and fourth waves are the same except for opposite signs and likewise for the second and third. In other words, for any observed value of frequency in a specified interval, the possible waveforms in the second succeeding interval can be divided into equally likely positive and negative matching pairs. This behavior, once established, must continue into all succeeding intervals. Hence the average lag product over the ensemble at fixed t and τ must vanish for τ greater than $2T$.

This may also be seen by noting that when $2\beta T = (2r + 1)\pi$, the quantity $\beta T(x_1 + \cdots + x_n - nx_{n+1})$ appearing in the definition (25) of $B_n(t)$ is an even or odd multiple of π according to whether $(x_1 + \cdots + x_n - nx_{n+1})/2 = r_n$ is even or odd. Hence

$$u(t) = \begin{matrix} A(-)^{r_n} \cos(\omega_2 t + \phi), & x_{n+1} = 1 \\ A(-)^{r_n} \cos(\omega_1 t + \phi), & x_{n+1} = -1. \end{matrix} \tag{61}$$

For any observed set of $x_1, \cdots x_{n+1}$ leading to a definite $u(t)$ in the nth interval, the waveforms in the second succeeding interval can be written as

$$\begin{aligned} &A(-)^{r_n+(r_{n+2}-r_n)} \cos(\omega_2 t + \phi), \qquad x_{n+3} = 1 \\ &A(-)^{r_n+(r_{n+2}-r_n)} \cos(\omega_1 t + \phi), \qquad x_{n+3} = -1. \end{aligned} \tag{62}$$

Since $r_{n+2} - r_n$ contains x_{n+2} only through the term $x_{n+2}/2$, the forms in (62) are of random sign independent of the original form (61). Hence the waveform in the second succeeding interval is entirely independent (in sign and frequency) of the waveform in the original interval.

Since $E_0(t + nT, \tau)$ and $E_1(t + nT, \tau)$ in (59) and (60) are found not to depend on n, we can evaluate the autocorrelation function by averaging over t in a single signaling interval as follows

$$\begin{aligned} R_u(\tau) &= \frac{1}{T}\int_0^{T-\tau} E_0(t + nT, \tau)\, dt, \\ &\qquad + \frac{1}{T}\int_{T-\tau}^{T} E_1(t + nT, \tau)\, dt \qquad 0 < \tau < T \end{aligned} \tag{63}$$

$$R_u(\tau) = \frac{1}{T}\int_0^{2T-\tau} E_1(t + nT, \tau)\, dt, \qquad T < \tau < 2T. \tag{64}$$

For $0 < \tau < T$, we calculate

$$\begin{aligned} \frac{4T}{A^2} R_u(\tau) &= \left[(2T - \tau)\cos\beta\tau + \frac{\sin\beta\tau}{\beta}\right]\cos\alpha\tau \\ &\quad + \cos 2\phi\left[\frac{\cos\beta\tau\sin\alpha\tau}{\alpha} - \frac{\sin(\alpha+\beta)\tau}{2(\alpha+\beta)} - \frac{\sin(\alpha-\beta)\tau}{2(\alpha-\beta)}\right]. \end{aligned} \tag{65}$$

For $T < \tau < 2T$,

$$\begin{aligned} \frac{4T}{A^2} R_u(\tau) &= \left[(2T - \tau)\cos\beta\tau + \frac{\sin\beta\tau}{\beta}\right]\cos\alpha\tau \\ &\quad - \cos 2\phi\left[\frac{\cos\beta\tau\sin\alpha\tau}{\alpha} - \frac{\sin(\alpha+\beta)\tau}{2(\alpha+\beta)} - \frac{\sin(\alpha-\beta)\tau}{2(\alpha-\beta)}\right]. \end{aligned} \tag{66}$$

The spectral density function is then found to be

$$\begin{aligned} w_u(f) &= 4\int_0^{2T} R_u(\tau)\cos\omega\tau\, d\tau \\ &= A^2\frac{\sin^2\omega T}{2T}\left[\left(\frac{1}{\omega-\omega_1} - \frac{1}{\omega-\omega_2}\right)^2 + \left(\frac{1}{\omega+\omega_1} - \frac{1}{\omega+\omega_2}\right)^2\right. \\ &\qquad \left. + 2\left(\frac{1}{\omega-\omega_1} - \frac{1}{\omega-\omega_2}\right)\left(\frac{1}{\omega+\omega_1} - \frac{1}{\omega+\omega_2}\right)\cos 2\phi\right]. \end{aligned} \tag{67}$$

The difference between this result and the limit of the general expression (48) when the special case (57) is substituted consists of the term containing 2ϕ. This exceptional case thus has the property of remembering the initial phase angle even though the spectral density is continuous. The reason is that the phase of the wave with respect to the signaling interval must remain fixed throughout all time and is, therefore, not subject to averaging as in the more general case.

IV. FOURIER TRANSFORM METHOD

Expressions (48) and (67) for $w_u(f)$ have been obtained by first computing $R_u(\tau)$ and then taking its Fourier transform. As mentioned earlier, $w_u(f)$ can also be obtained by working with a Fourier-type integral of $u(t)$ taken over a long time interval. This method will now be sketched for the case in which $2\beta T$ is not a multiple of π. Most of the intermediate steps are omitted. If they were included, the length of the derivation would be comparable to the derivation based on $R_u(\tau)$.

The spectral density $w_u(f)$ of $u(t) = A \cos B_n(t)$ is the limit of $2\langle | S_N(f, NT) |^2\rangle/NT$ as $N \to \infty$. In this expression

$$S_N(f, NT) = \int_0^{NT} e^{-j\omega t} u(t)\, dt = \sum_{n=0}^{N-1} s_n \tag{68}$$

$$s_n = Ae^{-j\omega nT} \int_0^T e^{-j\omega t} \cos B_n(t + nT)\, dt \tag{69}$$

where s_n is a function of f. The ensemble average of

$$\begin{aligned} | S_N(f, NT) |^2 &= \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} s_n s_m^* \\ &= \sum_{n=0}^{N-1} s_n s_n^* + \sum_{k=1}^{N-1} \sum_{n=0}^{N-k-1} (s_n s_{n+k}^* + s_{n+k} s_n^*) \end{aligned} \tag{70}$$

is the sum of terms of the form

$$\begin{aligned} \langle s_{n+k} s_n^* \rangle = A^2 e^{-j\omega kT} \int_0^T e^{-j\omega t}\, dt \int_0^T e^{j\omega\tau} \\ \langle \cos B_{n+k}(t + nT + kT) \cos B_n(\tau + nT)\rangle\, d\tau. \end{aligned} \tag{71}$$

In these equations s_n^* denotes the conjugate complex of s_n .

The procedure used to obtain expressions (30) and (31) for $E_0(t, \tau)$ and $E_k(t, \tau)$ leads to

$$
\begin{aligned}
2\langle \cos B_n(t+nT)\cos B_n(\tau+nT)\rangle &= \cos\alpha(t-\tau)\cos\beta(t-\tau) \\
&\quad + \cos[\alpha(t+\tau+2nT)+2\phi]\cos\beta(t+\tau)\cos^n 2\beta T
\end{aligned}
\tag{72}
$$

$$
\begin{aligned}
2\langle \cos B_{n+k}(t+nT+kT)\cos B_n(\tau+nT)\rangle &= \cos\alpha(t-\tau+kT) \\
\cos\beta t\cos\beta(\tau-T)\cos^{k-1}\beta T &+ \cos[\alpha(t+\tau+2nT+kT) \\
&+ 2\phi]\cos\beta t\cos\beta(\tau+T)\cos^{k-1}\beta T\cos^n 2\beta T
\end{aligned}
\tag{73}
$$

where $k > 0$ in the last equation.

We shall consider only the case in which $2\beta T$ is not a multiple of π. Then the terms in (72) and (73) containing $\cos^n 2\beta T$ contribute nothing to the left-hand sides of

$$
\begin{aligned}
&\lim_{N\to\infty} N^{-1}\sum_{n=0}^{N-1}\langle s_n s_n{}^*\rangle \\
&\qquad = 2^{-1}A^2\int_0^T e^{-j\omega t}\,dt\int_0^T e^{j\omega\tau}\,d\tau\cos\alpha(t-\tau)\cos\beta(t-\tau)
\end{aligned}
\tag{74}
$$

$$
\begin{aligned}
&\lim_{N\to\infty} N^{-1}\sum_{n=0}^{N-k-1}\langle s_{n+k}s_n{}^*\rangle \\
&= 2^{-1}A^2e^{-j\omega kT}\cos^{k-1}\beta T\int_0^T e^{-j\omega t}\,dt\int_0^T e^{j\omega\tau}\cos\alpha(t-\tau+kT) \\
&\qquad\qquad \cos\beta t\cos\beta(\tau-T)\,d\tau.
\end{aligned}
\tag{75}
$$

Expression (48) for the spectral density $w_u(f)$ is now obtained by performing the integrations and then summing with respect to k as indicated in (70).

V. ILLUSTRATIVE CURVES

Fig. 1 shows a typical curve for the spectral density function when the phase is discontinuous at the instants of transition. The curve is calculated from

$$
\begin{aligned}
w_u(f)/A^2 &= \frac{\delta(f-f_1)}{8}+\frac{\delta(f-f_2)}{8} \\
&\qquad + \frac{G(\omega-\omega_1)}{2T(\omega-\omega_1)^2}+\frac{G(\omega-\omega_2)}{2T(\omega-\omega_2)^2}
\end{aligned}
\tag{76}
$$

which is an approximation to (15) and (18). It holds when $\omega_2 - \omega_1$ is not a multiple of ω_s, and in addition ω_1 and ω_2 are so large that the portion of the spectrum folded back from $\omega = 0$, i.e., the portion de-

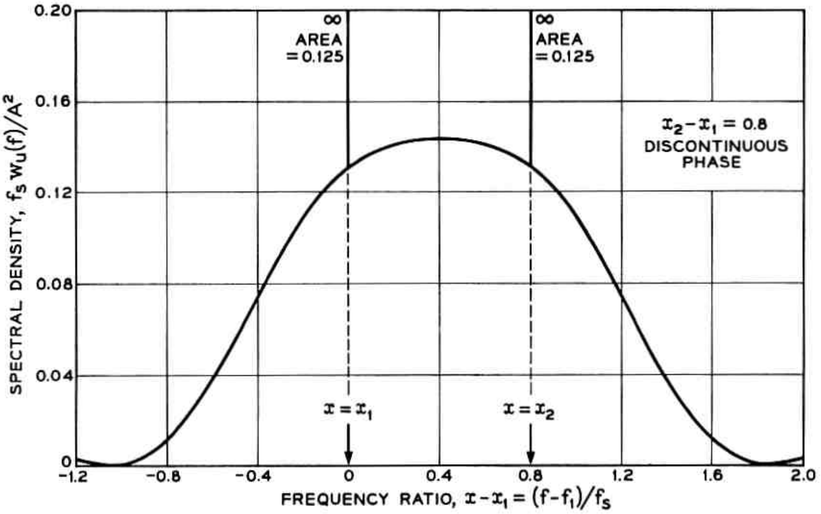


Fig. 1 — Spectral density of random binary FSK wave with discontinuous phase at transitions. Frequency shift = 0.8 times signaling frequency.

pending on inverse powers of $\omega + \omega_1$ and $\omega + \omega_2$, is negligible. The contribution of the neglected terms becomes appreciable only when the marking or spacing frequency is less than the signaling frequency. It is convenient to let $x = \omega/\omega_s = f/f_s$, $x_1 = f_1/f_s$, and $x_2 = f_2/f_s$.

Fig. 1 is calculated for the case $\omega_2 - \omega_1 = 0.8\omega_s$, i.e., $x_2 - x_1 = 0.8$. The abscissa is $x - x_1$. The ordinate is $f_s w_u(f)/A^2$. The curve is symmetrical about $x - x_1 = 0.4$. The steady-state terms are represented by spikes of infinite height and infinitesimal width at $x = x_1$ and $x = x_2$. Each of these spikes has an area of $\frac{1}{8}$. The area under the continuous curve is $\frac{1}{4}$. The total area is $\frac{1}{2}$, which is the mean-square value of the signal wave per unit of squared amplitude.

Fig. 2 shows a case of continuous phase corresponding to a frequency shift equal to 0.8 times the signaling frequency. This curve was computed from (48) where, again, the terms containing inverse powers of $\omega + \omega_1$ and $\omega + \omega_2$ are assumed to be negligibly small. Since the infinite spikes representing steady-state components are absent, the total area under the curve must be $\frac{1}{2}$. We note that peaks of finite height and width appear just outside the interval bounded by the marking and spacing frequencies. These peaks become more pronounced and move toward the marking and spacing frequencies as we approach the commensurable case in which the frequency shift $f_2 - f_1$ is exactly equal to the signaling rate f_s. Fig. 3 shows the curve for frequency shift equal to 0.95 times

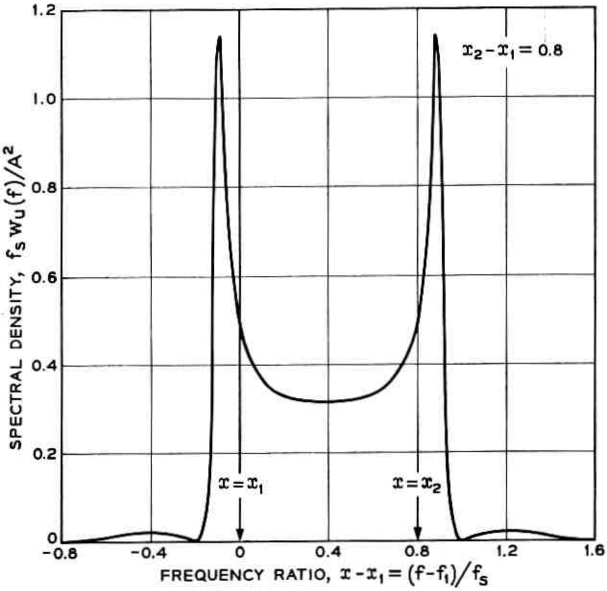


Fig. 2 — Spectral density of random binary FSK wave with continuous phase at transitions. Frequency shift = 0.8 times signaling frequency.

the signaling rate. Here the peaks are almost twenty times as high as in Fig. 2. The limiting case of $x_2 - x_1 = 1$ is exhibited in Fig. 4. The finite spikes of Fig. 2 and 3 have now become full-fledged impulses of infinite height, infinitesimal width, and area $\frac{1}{8}$. They represent the mean-square value of steady-state components at the marking and spacing frequencies. The continuous part of the curve was calculated from (54), noting that $(\omega - \alpha)T = 2\pi(x - x_1 - \frac{1}{2})$ in this case. Fig. 5 shows a representative curve on the other side of the limiting case, with the frequency shift taken equal to 1.2 times the signaling rate. The finite peaks now appear inside the interval between marking and spacing frequencies.

It is instructive to study the transition from Figs. 2 and 3 to Fig. 4. Since the curves are symmetrical about $x - x_1 = (x_2 - x_1)/2$, it is sufficient to consider the region of rapid change near $x = x_1$. Setting $x_2 - x_1 = x_d$, we approximate (48) for $w_u(f)$ in this region by

$$w_u(f) \approx \frac{A^2 \sin^2 (x - x_1)\pi \sin^2 (x - x_2)\pi}{2\pi^2 f_s (x - x_1)^2 [1 - 2 \cos (2x - x_2 - x_1)\pi \cos x_d\pi + \cos^2 x_d\pi]} . \qquad (77)$$

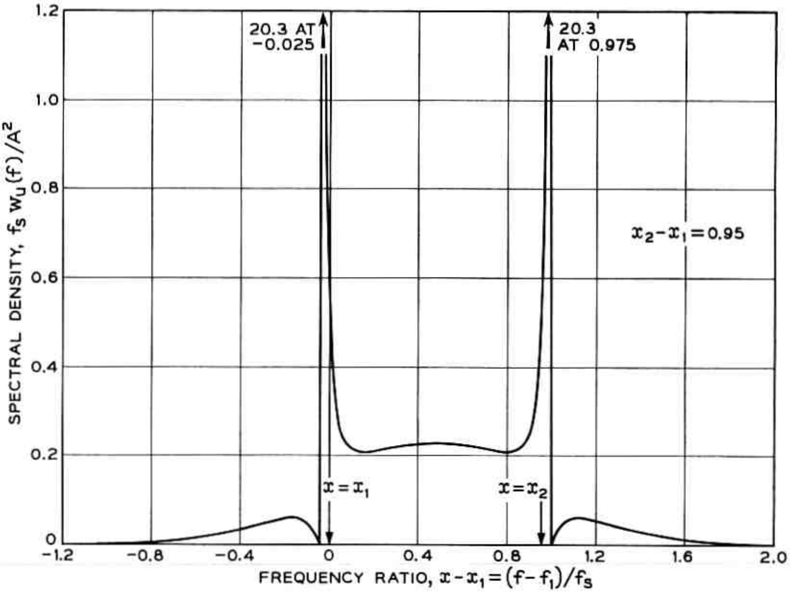


Fig. 3 — Spectral density of random binary FSK wave with continuous phase at transitions. Frequency shift = 0.95 times signaling frequency.

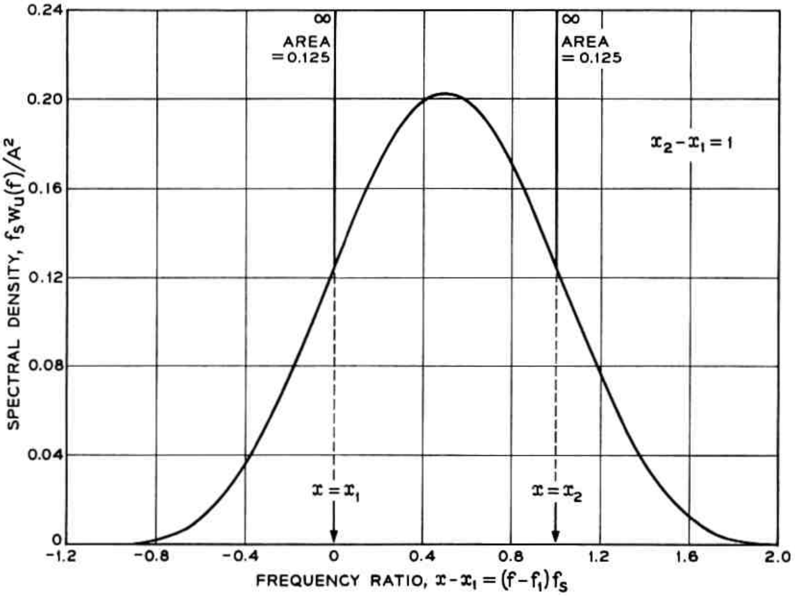


Fig. 4 — Spectral density of random binary FSK wave with continuous phase at transitions. Frequency shift = signaling frequency.

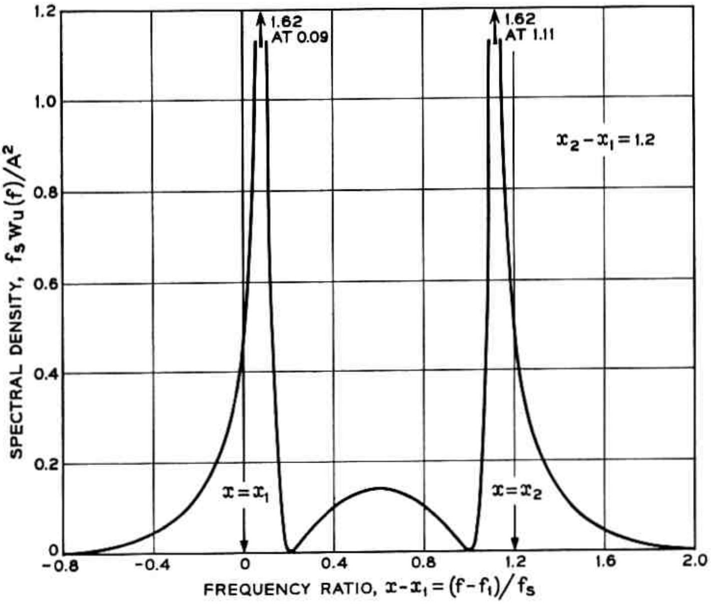


Fig. 5 — Spectral density of random binary FSK wave with continuous phase at transitions. Frequency shift = 1.2 times signaling frequency.

We are interested in the behavior of (77) as $x_2 - x_1$ approaches unity. Setting $x_2 - x_1 = 1 - \epsilon$ and $x - x_1 = y$, we find that when both ϵ and y are small compared with unity we can approximate (77) by

$$Y = f_s w_u(f)/A^2 \approx \frac{1}{8} \frac{(y + \epsilon)^2}{(\pi\epsilon^2/4)^2 + (y + \epsilon/2)^2}. \tag{78}$$

It is seen that Y depends on y approximately as shown in Table I. The value $y = \pm\infty$ corresponds to several positive or negative multiples of ϵ, and therefore actually becomes small in absolute value as ϵ approaches zero. It follows that Y hitches onto the value 0.125 shown at $x = x_1$ in Fig. 4. The curves shown in Figs. 2 and 3 correspond to $\epsilon = 0.2$ and $\epsilon = 0.05$. They show the behavior indicated by (78). In particular,

Table I — Approximate Ordinates of Spectral Density in the Neighborhood of Peak

y	$-\infty$	$-\epsilon$	$-\epsilon/2$	0	$+\infty$
Y	$\frac{1}{8}$	0	$1/(2\pi^2\epsilon^2)$	$\frac{1}{2}$	$\frac{1}{8}$

Y obtains its peak value of approximately $1/(2\pi^2\epsilon^2)$ near $y = -\epsilon/2$ and drops down to about half the peak value at $y = -\epsilon/2 \pm \pi\epsilon^2/4$. When $\epsilon = 0.05$, the peak value of Y is 20.3. The area under the peak, as measured by the integral of Y taken from $y = -\epsilon$ to $y = +\epsilon$, approaches $\frac{1}{8}$ as ϵ approaches zero. This agrees in the limit with the area of the impulses shown in Fig. 4.

The work of Sunde[1] has indicated that the special case in which the frequency shift is equal to the bit rate has a theoretical advantage in that intersymbol interference can be suppressed at the sampling instants in the output of an ideal frequency detector. The results presented here show one method by which such a frequency lock can be attained. Since the two principal peaks of the spectral density function reach maximum height when the condition $x_2 - x_1 = 1$ is attained, the output of a spectral analyzer can be used to determine the proper bias on the tuning control of the keyed oscillator. Another possible instrumentation can be devised by use of the autocorrelation function. When the signal wave with continuous phase transitions is multiplied by itself delayed by a large multiple of the bit interval, the average value of the product tends toward zero except when the frequency shift is locked to the bit rate. In practice, the advantage of a rigid lock-in to the theoretical optimum has not proved to be very significant. The actual reduction of intersymbol interference which could be achieved by choosing the best value of frequency shift would typically be masked by other departures from the ideal conditions.

Fig. 6 illustrates the case in which αT and βT are odd multiples of $\pi/2$. The significantly different properties exhibited are not very pronounced except when marking and spacing frequencies are sufficiently low to be comparable with the signaling rate. The case shown in Fig. 6 applies when the marking frequency is half the signaling frequency and the spacing frequency is equal to the signaling frequency. The frequency shift is half the signaling frequency. The curves are calculated from (67) for three different values of the initial phase angle. Since there are no steady-state components, the area under each curve must be 0.5. The peak of the spectral density function changes from 0.763 to 0.857 as the cosine of the initial phase angle is varied from -1 to $+1$. The ordinates become zero at $x = \frac{3}{2}, 2, \frac{5}{2}, \cdots$. As x approaches infinity the maximum values of the intervening loops decrease as x^{-6} when $\cos \phi = 1$ and as x^{-4} for other values of $\cos \phi$.

When the values $f_1 = f_s/2$, $f_2 = f_s$ corresponding to Fig. 6 are substituted in the general equation (48) for $w_u(f)$, the result is the case $\cos 2\phi = 0$ shown in Fig. 6. This is to be expected since when the con-

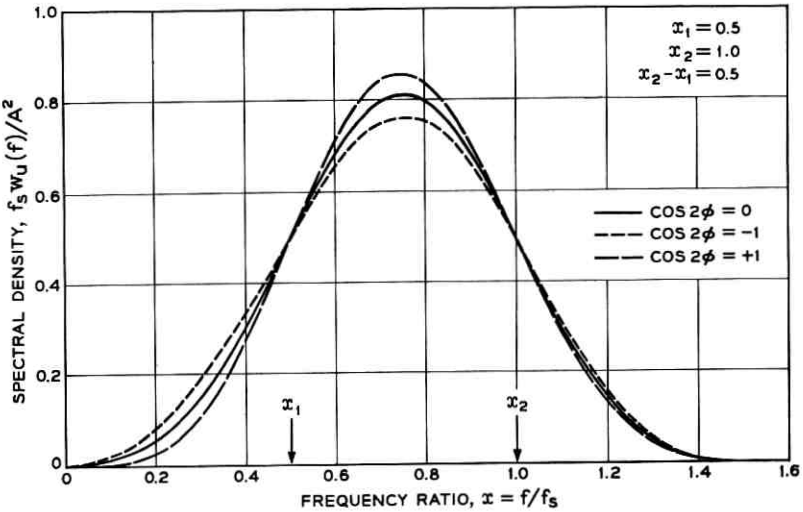


Fig. 6 — Spectral density of random binary FSK wave with continuous phase at transitions. Frequency shift = 0.5 times signaling frequency. Marking frequency = signaling frequency. Spacing frequency = 0.5 times signaling frequency.

ditions $f_1 = f_s/2$, $f_2 = f_s$ are almost (but not quite) satisfied, the phase of $u = A \cos B_n(t)$ changes by a small amount, or by π plus a small amount, from one transition point to another. Over a long period of time these small changes accumulate and have the same effect as replacing $\cos 2\phi$ by its average value 0.

Figs. 7–12 inclusive show the normalized autocorrelation functions corresponding to the cases of Figs. 1–6 respectively. To avoid making the curves depend on the values of marking and spacing frequencies, we have indicated the envelope of the high-frequency oscillations in Figs. 7–11. The autocorrelations are obtained by multiplying the solid line curves by $\frac{1}{2}A^2/\cos[(\omega_2 + \omega_1)\tau/2]$, which in terms of the lag time τ is a cosine wave at the midband frequency. The resulting oscillation has the value unity at $\tau = 0$ and is contained within the solid and dashed curves. Fig. 12, which is drawn for specified marking and spacing frequencies, shows an actual autocorrelation function.

The typical case of discontinuous phase, which is illustrated in Fig. 7, has a linearly damped envelope until τ reaches the value T. At time T the envelope changes continuously to that of the sum of two cosine waves at the marking and spacing frequencies. The latter envelope is a cosine wave at half the difference frequency and it persists with undiminished amplitude throughout all values of τ greater than T. Fig. 8

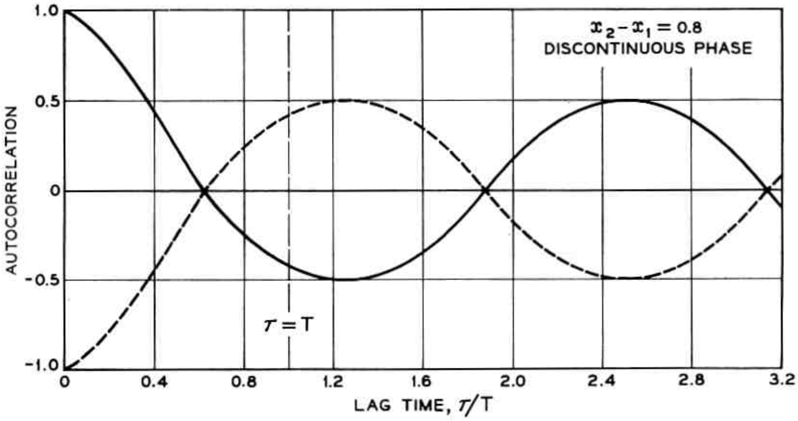


Fig. 7 — Envelope of autocorrelation function when phase is discontinuous at transitions and $f_2 - f_1 = 0.8f_s$. Actual autocorrelation is full line curve multiplied by $(A^2/2) \cos [(\omega_2 + \omega_1)\tau/2]$.

represents the same case as Fig. 7 except that the phase is continuous. The effect is that the envelope of the autocorrelation in Fig. 8 decays to zero at infinite lag time instead of oscillating with constant amplitude. The decay in each multiple of T after the second one is produced by a multiplication of the corresponding values in the preceding interval by $\cos (x_2 - x_1)\pi$, which has the value -0.809 in Fig. 8. As $x_2 - x_1$ approaches unity, the multiplying factor produces only a slight reduction

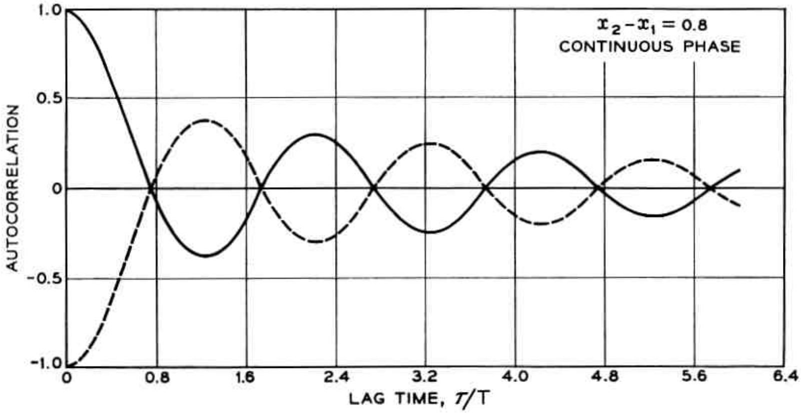


Fig. 8 — Envelope of autocorrelation function when phase is continuous at transitions and $f_2 - f_1 = 0.8f_s$. Actual autocorrelation is full line curve multiplied by $(A^2/2) \cos [(\omega_2 + \omega_1)\tau/2]$.

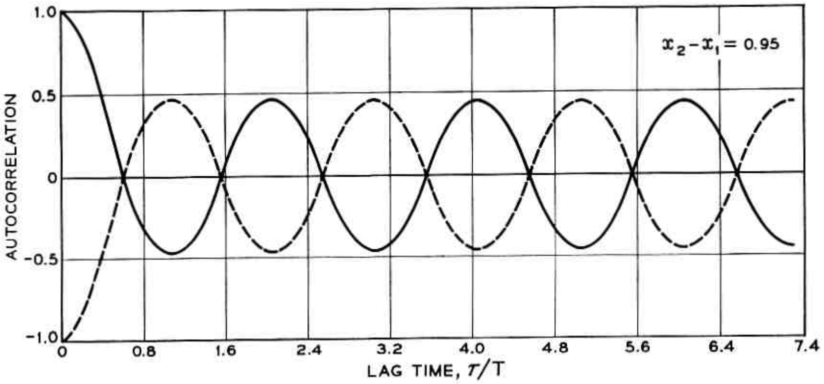


Fig. 9 — Envelope of autocorrelation function when phase is continuous at transitions and $f_2 - f_1 = 0.95f_s$. Actual autocorrelation is full line curve multiplied by $(A^2/2) \cos [(\omega_2 + \omega_1)\tau/2]$.

in each interval and the oscillations retain appreciable amplitude for very large lag times. Such behavior is emphasized in Fig. 9 for the case of $x_2 - x_1 = 0.95$, $\cos (x_2 - x_1)\pi = -0.9877$. The very slow departure from constant amplitude oscillations indicates that the signal wave contains components which are very nearly sinusoidal. The time domain analysis thus agrees with the sharp high peaks found in the frequency domain analysis, as shown in Fig. 3 for $x_2 - x_1 = 0.95$.

Fig. 10 shows the limiting case in which the phase is continuous and

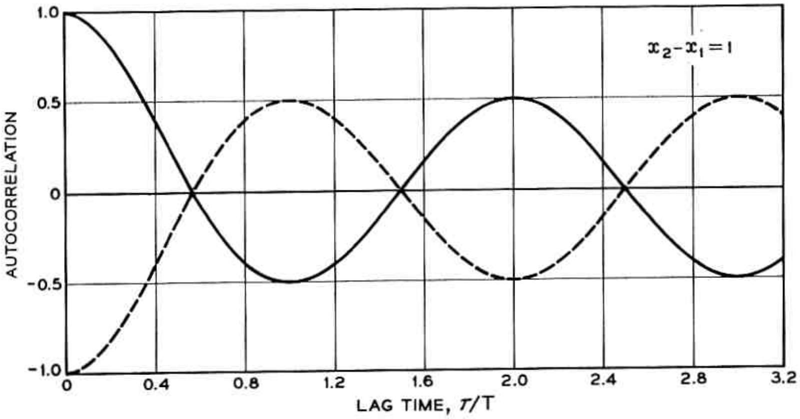


Fig. 10 — Envelope of autocorrelation function when phase is continuous at transitions and $f_2 - f_1 = f_s$. Actual autocorrelation is full line curve multiplied by $(A^2/2) \cos [(\omega_2 + \omega_1)\tau/2]$.

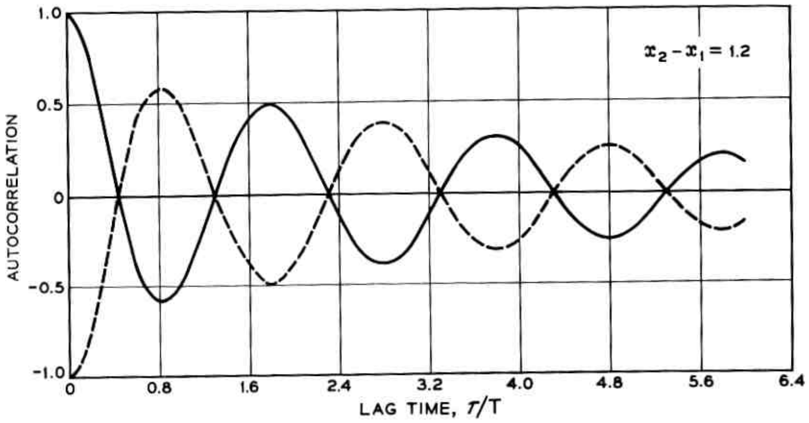


Fig. 11 — Envelope of autocorrelation function when phase is continuous at transitions and $f_2 - f_1 = 1.2f_s$. Actual autocorrelation is full line curve multiplied by $(A^2/2) \cos [(\omega_2 + \omega_1)\tau/2]$.

$x_2 - x_1 = 1$. The appearance of the line spectral terms is indicated by the constancy of the amplitude of oscillations for $\tau > T$. Fig. 11 for $x_2 - x_1 = 1.2$ corresponds to Fig. 5. The decay rate is the same as in Fig. 8, since $\cos 1.2\pi = \cos 0.8\pi$. The period of the oscillations is decreased.

Fig. 12 shows the singular case in which the sum and difference frequencies are both odd multiples of half the signaling frequency. The values chosen are the same as those of Fig. 6. The autocorrelation func-

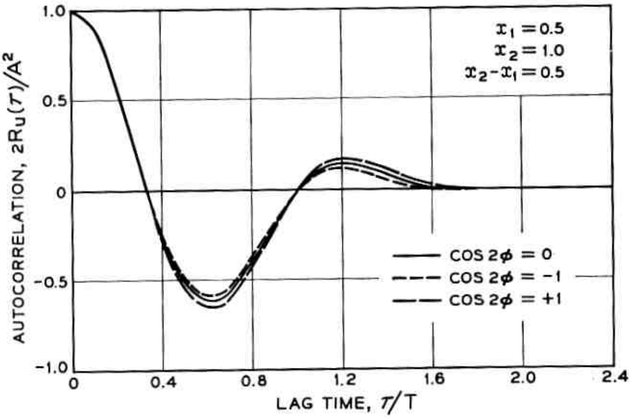


Fig. 12 — Normalized autocorrelation function when phase is continuous, $f_1 = f_s/2$, $f_2 = f_s$, and $f_2 - f_1 = f_s/2$.

tion is time limited, vanishing for all values of τ greater than $2T$. The dependence on initial phase is indicated by the three curves, which are drawn for the cases of $\cos 2\phi = 0$, 1, and -1 as in Fig. 6. The total variation in the height of the negative peak with ϕ is 0.07.

Fig. 7 differs from Figs. 8–12 in that the discontinuity in phase of the signal wave produces a discontinuity in the slope of the autocorrelation curve at $\tau = T$. This is consistent with the decay of the spectral density of Fig. 1 with the inverse square of frequency at high frequencies. The spectral densities of Figs. 2 to 6 vary ultimately as the inverse fourth power of frequency, which requires not only the slope but the second derivative of the corresponding autocorrelation functions, Figs. 8 to 12, to be continuous at all values of τ.

VI. SUMMARY OF RESULTS FOR SPECTRAL DENSITY AND AUTOCORRELATION

Table II lists the equation numbers of the expressions giving $w_u(f)$ and $R_u(\tau)$ for the various cases which can arise. Let $f_s = 1/T$ be the signaling frequency and f_1, f_2 be the marking and spacing frequencies. Also, let l, r denote integers.

TABLE II — LIST OF EQUATIONS FOR SPECTRAL DENSITY AND AUTOCORRELATION OF FSK WAVE

Case	Equation Numbers	
	$w_u(f)$	$R_u(\tau)$
Discontinuous phase:		
(a) general case	(15), (18)	(19)
(b) degenerate cases	(17), (18)	(19)
Continuous phase:		
(c) $f_2 - f_1 = rf_s$, $f_2 + f_1 \neq lf_s$	(52), (53), (54)	(19)
(d) $f_2 - f_1 = rf_s$, $f_2 + f_1 = lf_s$	(52), (56)	(19)
(e) $\left.\begin{matrix} f_2 - f_1 = (r + \frac{1}{2})f_s \\ f_2 + f_1 = (l + \frac{1}{2})f_s \end{matrix}\right\}$	(67)	(65), (66)
(f) all other continuous phase cases	(48)	(36), (37)

VII. OTHER RELATED PUBLICATIONS

Jenks and Hannon[2] have given spectral density curves for the case of frequency shift equal to bit rate which they state have been taken from a forthcoming paper by Pushman in the Journal of the British Institute of Radio Engineers. The curves shown are in agreement with ours for the same case. We have not seen the complete work. Our interest in the problem was initially stimulated by discussions with I. Dorros, who has made use of some of our results in a study[3] of the transmission of binary data by FM over a band-limited channel. Since completing the work, we have become aware of a publication by Postl,[4] who has calculated the spectral density for binary continuous phase narrow-band FSK in which the midband frequency is large compared with both the frequency shift and the signaling rate.

REFERENCES

1. Sunde, E. D., Ideal Binary Pulse Transmission by AM and FM, B.S.T.J., **38,** November, 1959, pp. 1357–1426.
2. Jenks, F. G. and Hannon, D. C., Comparison of the Merits of Phase and Frequency Modulation for Medium Speed Serial Binary Digital Data Transmission Over Telephone Lines, J. Brit. I.R.E., **24,** July, 1962, pp. 21–36.
3. Dorros, I., Performance of a Binary FM System as a Function of the Channel, Columbia University Doctoral Dissertation, November, 1962.
4. Postl, W., Die Spektrale Leistungsdichte bei Frequenzmodulation eines Trägers mit einem Stochastischen Telegraphiesignal, Frequenz, **17,** March, 1963, pp. 107–110.

Binary Data Transmission by FM over a Real Channel

By W. R. BENNETT and J. SALZ

(Manuscript received April 26, 1963)

Formulas are derived for probability of error in the detection of binary FM signals received from a channel characterized by arbitrary amplitude- and phase-vs-frequency distortion as well as additive Gaussian noise. The results depend on the signal sequence and can be presented in terms of averages over all signal sequences or as bounds for the most and least vulnerable ones. Illustrative examples evaluated include Sunde's method of suppressing intersymbol interference in band-limited FM. The effects of various representative channel filters are also analyzed. A solution is given for the problem of optimizing the receiving bandpass filter to minimize error probability at constant transmitted signal power. It is found that a performance from 3 to 4 db poorer than that theoretically attainable from binary PM is realizable over a variety of filtering situations.

I. INTRODUCTION

This paper undertakes to refine and extend the state of knowledge concerning performance of FM systems for binary data transmission over real-life channels. The particular aim is application to facilities such as exist in the telephone plant. Efficient use of the available channels constrains the bandwidth allowed for a given signaling speed. The luxury of a bandwidth sufficient to permit frequency transitions without amplitude variations and without dependence of present waveform on past signal history would in general imply an unjustifiably low information rate for the frequency range occupied. We therefore concentrate our attention on the band-limited channel with its inherent distortion of the FM data wave.

We assume a linear time-invariant transmission medium specified by its amplitude- and phase-vs-frequency functions and the statistics of its additive noise sources. The limiting noise environment in the telephone plant is typically nongaussian and not well defined even in a statistical sense. Nevertheless, with the usual apology, we shall perform our analysis in terms of additive Gaussian noise. Justification of the relevancy is based on the following considerations:

(a) Laboratory tests on data transmission systems are made at present by adding Gaussian noise and counting errors. Good performance in terms of low error rate as a function of signal-to-noise ratio under such test conditions is found to be indicative of good performance on actual channels.

(b) Identification and removal of nongaussian disturbances is a feasible and continuing process which should eventually lead to a more nearly Gaussian description of the residue.

Our measure of performance is expressed in terms of error probability vs the ratio of average transmitted signal power to average Gaussian noise power. In most of the work we assume white Gaussian noise is added at the receiver input. A convenient reference is then the average noise power in a band of frequencies having width equal to the transmitted information rate in bits per second.

II. STATEMENT OF PROBLEM

A block diagram of the transmission system under study is shown in Fig. 1. The data source emits a sequence of binary symbols which for full information rate are independent of each other and have equal probability. The analysis can be generalized without analytical inconvenience to assign a probability m_1 to one of the two binary symbols and $1 - m_1$ to the other. In conventional binary notation the symbols are 1 and 0. It is convenient to express binary frequency modulation of

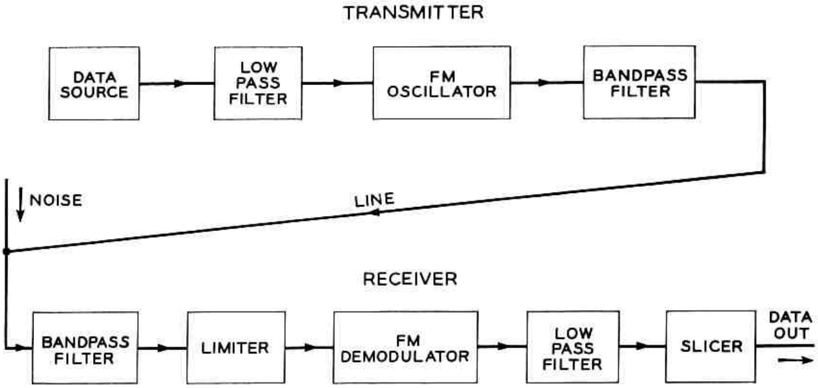


Fig. 1 — Binary FM transmission system.

an oscillator in terms of positive and negative frequency deviations. The combination of data source and low-pass filter is accordingly defined by the shaped baseband data wave train

$$s(t) = \sum_{n=-\infty}^{\infty} b_n g(t - nT) \tag{1}$$

where

$$b_n = 2a_n - 1. \tag{2}$$

The values of a_n represent the data sequence in binary notation. The probability is m_1 that the typical a_n is unity, and $1 - m_1$ that it is zero. The value of b_n is $+1$ if a_n is unity, and -1 if a_n is zero. The function $g(t)$ represents a standard pulse emitted by the low-pass filter for a signal element centered at $t = 0$.

Ideally, the oscillator frequency follows the baseband signal wave $s(t)$. This would imply an output voltage from the FM oscillator specified by

$$V(t) = A \cos \left[\omega_c t + \theta_0 + \mu \int_{t_0}^{t} s(\lambda)\, d\lambda \right]. \tag{3}$$

Here, A is the carrier amplitude, ω_c is the frequency of the oscillator with no modulating signal applied, t_0 is an arbitrary reference time, θ_0 is the phase at $t = t_0$, and μ is a conversion factor relating frequency displacement to baseband signal voltage. The instantaneous frequency of the wave (3) is defined as the derivative of the argument of the cosine function. It can be written in the form $\omega_c + \omega_i$, where ω_i, the deviation from midband, is ideally expressed by

$$\omega_i = \mu s(t). \tag{4}$$

In the practical case, the transmitting bandpass filter restricts the frequency-modulated wave to the range of frequencies passed by the channel. The purpose of this filter is to prevent both waste of transmitted power in components which will not reach the receiver and contamination of the line at frequencies assigned to other channels. The result is a transformation of the voltage wave (3) to a band-limited form, which must depart in more or less degree from the ideal conditions of constant amplitude and linear relationship between frequency and baseband signal. The line also inserts variations in amplitude- and phase-vs-frequency which cause further departures from the ideal. For our purposes it is sufficient to combine the line characteristics with those of the transmitting filter into a single composite network function determining the wave presented to the receiving bandpass filter.

The receiving bandpass filter is necessary to exclude out-of-band noise and interference from the detector input. It also shapes the signal waveform and can include compensation for linear in-band distortion suffered in transmission. Two contradictory attributes are sought in the filter — a narrow band to reject noise and a wide band to supply a good signal wave to the detector. An opportunity for an optimum design thus exists and will be explored in this paper.

The frequency detector is assumed to differentiate the phase with respect to time. The post-detection filter can do further noise rejection and shaping in the baseband range, but its only function in our present analysis is to separate the wave representing the frequency variation from the higher-frequency detection products. The slicer delivers positive voltage when the detected frequency is above midband and negative voltage when the detected frequency is below midband. The slicer output is sampled at appropriate instants to recover the binary data sequence.

The noise-free input to the detector will be written in the form

$$V_r(t) = P(t) \cos (\omega_c t + \theta) - Q(t) \sin (\omega_c t + \theta). \tag{5}$$

$P(t)$ and $Q(t)$ represent in-phase and quadrature signal modulation components respectively, which are associated with a carrier wave at the midband frequency ω_c with specified phase θ. Such a resolution can always be made, even though the details in actual examples may be burdensome. The added noise wave at the detector input is assumed to be Gaussian with zero mean and can likewise be written as

$$v(t) = x(t) \cos (\omega_c t + \theta) - y(t) \sin (\omega_c t + \theta). \tag{6}$$

If $v(t)$ represents Gaussian noise band-limited to $\pm 2\omega_c$, $x(t)$ and $y(t)$ are also Gaussian and are band-limited to $\pm\omega_c$. If the spectral density of $v(t)$ is $w_v(\omega)$, the spectral densities of $x(t)$ and $y(t)$ are given by[1]

$$w_x(\omega) = w_y(\omega) = w_v(\omega_c + \omega) + w_v(\omega_c - \omega), \qquad |\omega| < \omega_c \tag{7}$$

In general, $x(t)$ and $y(t)$ are dependent, with cross-spectral density

$$w_{xy}(\omega) = j[w_v(\omega_c - \omega) - w_v(\omega_c + \omega)] \tag{8}$$

and cross-correlation function expressed in terms of $R_v(\tau)$, the autocorrelation function of $v(t)$, by

$$R_{xy}(\tau) = -2R_v(\tau) \sin \omega_c \tau. \tag{9}$$

The cross correlation vanishes at $\tau = 0$, and hence the joint distribution of $x(t)$, $y(t)$ at any specified t is that of two independent Gaussian variables.

We shall also require the joint distribution of x and y with their time derivatives $\dot{x}$ and $\dot{y}$. The latter are Gaussian with spectral densities

$$w_{\dot{x}}(\omega) = w_{\dot{y}}(\omega) = \omega^2 w_x(\omega). \tag{10}$$

The cross-spectral densities are

$$w_{x\dot{x}}(\omega) = w_{y\dot{y}}(\omega) = j\omega w_x(\omega) \tag{11}$$

$$w_{x\dot{y}}(\omega) = j\omega_{xy}(\omega) = \omega[w_v(\omega_c + \omega) - w_v(\omega_c - \omega)] = -w_{\dot{x}y}\,. \tag{12}$$

The cross correlations are

$$\begin{aligned} R_{x\dot{x}}(\tau) &= \int_{-\infty}^{\infty} w_{x\dot{x}}(\omega)e^{j\tau\omega}\,d\omega = -\int_{-\infty}^{\infty} \omega w_x(\omega)\sin\tau\omega\,d\omega \\ &= R_{y\dot{y}}(\tau) \end{aligned} \tag{13}$$

$$\begin{aligned} R_{x\dot{y}}(\tau) &= -R_{\dot{x}y}(\tau) = \int_{-\infty}^{\infty} w_{x\dot{y}}(\omega)e^{j\tau\omega}\,d\omega \\ &= \int^{\infty} \omega[w_v(\omega_c + \omega) - w_v(\omega_c - \omega)]\cos\tau\omega\,d\omega. \end{aligned} \tag{14}$$

The cross correlation of x and $\dot{x}$ as well as of y and $\dot{y}$ vanish at $\tau = 0$, and hence at any instant $\dot{x}$ is independent of x, and $\dot{y}$ is independent of y. The cross correlations of x and $\dot{y}$, and of $\dot{x}$ and y, do not vanish in general, but do vanish in the special case in which

$$w_v(\omega_c + \omega) = w_v(\omega_c - \omega). \tag{15}$$

This is the case of a noise spectrum which is symmetrical with respect to the midband and represents a reasonable objective in system design. Since the simplification in computational details is quite considerable when the condition of symmetry is imposed, and since the departures caused by lack of symmetry are not of primary interest, we shall assume henceforth that (15) is satisfied. The four variables x, $\dot{x}$, y, and $\dot{y}$ are then independent and have the joint Gaussian probability density function

$$p(x, y, \dot{x}, \dot{y}) = \frac{1}{4\pi^2\sigma_0^2\sigma_1^2}\exp\left[-\frac{x^2 + y^2}{2\sigma_0^2} - \frac{\dot{x}^2 + \dot{y}^2}{2\sigma_1^2}\right] \tag{16}$$

$$\sigma_0^2 = \int_{-\infty}^{\infty} w_x(\omega)\,d\omega = 2\int_{-\infty}^{\infty} w_v(\omega_c + \omega)\,d\omega \tag{17}$$

$$\sigma_1^2 = \int_{-\infty}^{\infty} w_{\dot{x}}(\omega)\, d\omega = 2 \int_{-\infty}^{\infty} \omega^2 w_v(\omega_c + \omega)\, d\omega. \tag{18}$$

The noise-free detector input wave (5) can be written in the equivalent form

$$V_r(t) = R(t) \cos [\omega_c t + \phi(t)] \tag{19}$$

where

$$R^2(t) = P^2(t) + Q^2(t) \tag{20}$$

$$\tan \phi(t) = Q(t)/P(t). \tag{21}$$

The frequency detector and post-detection filter combine to deliver a wave proportional to the instantaneous frequency deviation from mid-band. Taking the constant of proportionality as unity, we write for the output wave

$$\phi'(t) = \frac{d}{dt} \arctan \frac{Q(t)}{P(t)} = \frac{P(t)Q'(t) - Q(t)P'(t)}{P^2(t) + Q^2(t)}. \tag{22}$$

With the functional dependence on t understood, we write this equation in the form

$$\phi'(t) = \dot{\phi} = (P\dot{Q} - Q\dot{P})/R^2. \tag{23}$$

When the noise is added, the detected frequency is changed to

$$\psi'(t) = \dot{\psi} = \frac{(P + x)(\dot{Q} + \dot{y}) - (Q + y)(\dot{P} + \dot{x})}{(P + x)^2 + (Q + y)^2}. \tag{24}$$

Assuming that the system does not make errors in the absence of noise, we can express the probability of error in a given sample of instantaneous frequency taken at the time $t = nT$ as the probability that $\psi'(nT)$ is negative if $\phi'(nT)$ is positive or the probability that $\psi'(nT)$ is positive if $\phi'(nT)$ is negative. Since the system has memory, the values of P, Q, $\dot{P}$, and $\dot{Q}$ at any sampling instant depend on the entire signal sequence. Our procedure is first to show how the error probability can be evaluated at any sampling instant for any sequence. We then calculate error rates for specific sequences and establish bounds for most and least vulnerable sequences.

Since the denominators of (23) and (24) are inherently positive, the decisions are made entirely on the basis of the signs of the numerators. Therefore, we do not require the distribution function of the instantaneous frequency itself. In fact if we let

$$\begin{aligned} x + P &= x_1, & \dot{x} + \dot{P} &= \dot{x}_1 \\ y + Q &= y_1, & \dot{y} + \dot{Q} &= \dot{y}_1 \end{aligned} \tag{25}$$

we require only one value of the distribution function of the variable z defined by

$$z = x_1\dot{y}_1 - y_1\dot{x}_1 . \tag{26}$$

The error probability is fully determined in any specific case either by the probability that z is negative or by the probability that z is positive. That is, if $F(z)$ is the distribution function of z, we only require the value of $F(0)$.

We shall derive a general expression for $F(0)$ in terms of a single definite integral. From this integral we shall then obtain definite integrals representing bounds for the error probability when arbitrary binary data sequences are transmitted. No restrictions on range of signal-to-noise ratios are made. The results will be applied to special cases of practical interest. One is Sunde's binary FM system which avoids intersymbol interference in a finite band in the absence of noise. When noise is added in this system, the detected samples become dependent on past signal history. It has been found possible to give a complete treatment of the Sunde method, including optimization of the receiving filter for minimum probability of error with fixed average transmitted signal power. The other cases analyzed in detail are based on design parameters actually in use on FM data transmission terminals.

III. GENERAL SOLUTION

Our first observation is that when x_1 and y_1 are fixed, the variable z of (26) is defined by a linear operation on the two independent Gaussian variables $\dot{x}_1$ and $\dot{y}_1$. Hence the conditional probability density function $p(z \mid x_1, y_1)$ of z when x_1 and y_1 are given is Gaussian with readily determined parameters. We accordingly write

$$p(z \mid x_1 , y_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(z - z_0)^2}{2\sigma^2}\right]. \tag{27}$$

The mean z_0 is the sum of the means of $x_1\dot{y}_1$ and $-y_1\dot{x}_1$, that is,

$$z_0 = x_1 \text{ av } \dot{y}_1 - y_1 \text{ av } \dot{x}_1 = x_1\dot{Q} - y_1\dot{P}. \tag{28}$$

The variance σ^2 is the sum of the variances of $x_1\dot{y}_1$ and $y_1\dot{x}_1$; hence

$$\sigma^2 = (x_1^2 + y_1^2)\sigma_1^2. \tag{29}$$

The complete probability density function $p(z)$ for z is obtained by averaging the conditional probability density function over x_1 and y_1. This is done by multiplying (27) by the joint probability density function of x_1 and y_1 and then integrating over all x_1 and y_1. Calling the latter function $q(x_1, y_1)$, we can express its value by substituting the values of x and y from (25) in (16) and integrating out the $\dot{x}$ and $\dot{y}$ terms. The result is

$$q(x_1, y_1) = \frac{1}{2\pi\sigma_0^2} \exp\left[-\frac{(x_1 - P)^2 + (y_1 - Q)^2}{2\sigma_0^2}\right]. \tag{30}$$

Then

$$p(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(z \mid x_1, y_1) q(x_1, y_1)\, dx_1\, dy_1. \tag{31}$$

The probability of error when the noise-free sample of frequency deviation is positive is

$$P_+ = \int_{-\infty}^{0} p(z)\, dz = \int_0^{\infty} p(-z)\, dz. \tag{32}$$

Likewise, when the noise-free sample is negative, we obtain a probability of error

$$P_- = \int_0^{\infty} p(z)\, dz. \tag{33}$$

The problem is thus reduced to the evaluation of the triple integral obtained by combining (27), (30), and (31) with either (32) or (33). It is shown in Appendix A that the result of these operations can be expressed in the following form

$$\begin{aligned} P_+ &= \frac{1}{2} \operatorname{erfc} \frac{R}{\sqrt{2}\sigma_0} \\ &\quad + \frac{R}{2\sigma_0\sqrt{2\pi}} \int_{-1}^{1} \exp\left(-\frac{R^2x^2}{2\sigma_0^2}\right) \operatorname{erfc} \frac{R\phi(1 - x^2)^{\frac{1}{2}} - \dot{R}x}{\sqrt{2}\sigma_1}\, dx. \end{aligned} \tag{34}$$

The value of P_- is obtained by subtracting the right-hand member of (34) from unity. We note that ϕ is positive for P_+ and negative for P_-. The symbol $\dot{R}$ is used for dR/dt where R is given by (20). In a pure FM wave, $\dot{R} = 0$, but this condition cannot be maintained in a finite bandwidth.

Differentiating partially with respect to $\dot{R}$ and rearranging, we obtain

$$\begin{aligned} \frac{\partial P_+}{\partial \dot{R}} &= \frac{R}{\pi\sigma_0\sigma_1} \int_0^1 x \exp\left[-\frac{R^2x^2}{2\sigma_0^2} - \frac{R^2\phi^2(1 - x^2) + \dot{R}^2x^2}{2\sigma_1^2}\right] \\ &\qquad \sinh \frac{R\dot{R}\phi x(1 - x^2)^{\frac{1}{2}}}{\sigma_1^2}\, dx. \end{aligned} \tag{35}$$

We note that $\partial P_+/\partial \dot{R}$ vanishes when $\dot{R} = 0$ and at no other value of $\dot{R}$. The latter follows from the fact that the integrand of (35) cannot change sign in the interval of integration. We also find that $\partial^2 P_+/\partial \dot{R}^2$ is positive when $\dot{R} = 0$. We conclude that P_+ is minimum with respect to $\dot{R}$ when and only when $\dot{R} = 0$. A lower bound on the probability of error for any fixed R and ϕ is therefore obtained by setting $\dot{R} = 0$, giving

$$P_l = \frac{1}{2} \operatorname{erfc} \frac{R}{\sqrt{2}\sigma_0} + \frac{R}{2\sigma_0\sqrt{2\pi}} \int_{-1}^{1} \exp\left(-\frac{R^2 x^2}{2\sigma_0^2}\right) \operatorname{erfc} \frac{R\phi(1-x^2)^{\frac{1}{2}}}{\sqrt{2}\sigma_1}\, dx. \tag{36}$$

Also, since P_+ must be monotonic increasing with $|\dot{R}|$, the largest probability of error for any fixed R and ϕ occurs when $\dot{R}$ has its largest possible absolute value. These deductions are of aid in selecting the data sequences which have most and least probabilities of error.

It is shown in Appendix A that P_l can be written in the equivalent form

$$P_l = \frac{1}{\pi} \int_0^{\pi/2} \exp\left[-\frac{R^2\phi^2/(2\sigma_1^2)}{1 + \left(\frac{\sigma_0^2\phi^2}{\sigma_1^2} - 1\right)\cos^2\theta}\right] d\theta. \tag{37}$$

It is also shown that when $\phi < (\sigma_1/\sigma_0)$, the limiting form for large signal-to-noise ratio — i.e., R large compared with σ_0 — is given by

$$P_l \sim \frac{\sigma_1}{R\phi\sqrt{2\pi}} \left(\frac{\sigma_0^2\phi^2}{\sigma_1^2} - 1\right)^{-\frac{1}{2}} \exp\left(-\frac{R^2}{2\sigma_0^2}\right). \tag{38}$$

When $\phi > (\sigma_1/\sigma_0)$, the limiting form becomes

$$P_l \sim \frac{\sigma_1}{R\phi\sqrt{2\pi}} \left(1 - \frac{\sigma_0^2\phi^2}{\sigma_1^2}\right)^{-\frac{1}{2}} \exp\left(-\frac{R^2\phi^2}{2\sigma_1^2}\right). \tag{39}$$

When $\phi = \sigma_1/\sigma_0$, we have the exact result

$$P_l = \frac{1}{2} \exp\left(-\frac{R^2}{2\sigma_0^2}\right). \tag{40}$$

The general equation for error probability (34) can conveniently be expressed in terms of the following three parameters

$$\rho^2 = \frac{R^2}{2\sigma_0^2} \tag{41}$$

$$a^2 = \frac{\sigma_0^{\,2} \dot{\phi}^2}{\sigma_1^{\,2}} \tag{42}$$

$$b^2 = \frac{\dot{R}^2}{2\sigma_1^{\,2}}. \tag{43}$$

Equation (34) then becomes

$$P_+ = \frac{1}{2} \operatorname{erfc} \rho + \frac{\rho}{2\sqrt{\pi}} \int_{-1}^{1} e^{-\rho^2 x^2} \operatorname{erfc} [a\rho(1 - x^2)^{\frac{1}{2}} - bx]\, dx. \tag{44}$$

Evaluation of this equation in terms of the three parameters ρ, a, and b gives the error probability for any of the FM systems considered.

IV. ERROR PROBABILITY VS SIGNAL-TO-NOISE RATIO

In analog systems the performance is often expressed in terms of signal-to-noise ratio in the receiver output. In the case of audio and video signals, where subjective judgments determine the requirements, the signal-to-noise ratio furnishes a good criterion. In the case of data signals, however, performance is judged in terms of errors made, and the errors cannot be predicted from the signal-to-noise ratio alone. The error rate depends in general on the distribution of the noise values. Furthermore, in good systems the errors are rare and hence are associated with infrequent noise conditions. The central part of the noise distribution is of less importance than the tails.

We illustrate the difference between a straight signal-to-noise ratio analysis and a direct error probability calculation in FM by a simple example. Consider the case of a long sequence of mark signals leading to a constant signal frequency $\omega_c + \omega_d$. The signal wave can then be written in the form

$$\begin{aligned} V(t) &= A \cos (\omega_c + \omega_d)t \\ &= A \cos \omega_d t \cos \omega_c t - A \sin \omega_d t \sin \omega_c t. \end{aligned} \tag{45}$$

Comparing with (5) and noting that we are omitting the arbitrary phase angle θ, which is of trivial interest, we make the identifications

$$P(t) = A \cos \omega_d t \qquad Q(t) = A \sin \omega_d t. \tag{46}$$

Then, by differentiation

$$P'(t) = -\omega_d A \sin \omega_d t \qquad Q'(t) = \omega_d A \cos \omega_d t. \tag{47}$$

If a sample is taken at $t = 0$

$$P = A \qquad \dot{P} = 0 \qquad Q = 0 \qquad \dot{Q} = \omega_d A. \tag{48}$$

Then from (24) the error $\dot{\psi} - \omega_d$ in the detected frequency deviation because of additive Gaussian noise is

$$\nu = \dot{\psi} - \omega_d = \frac{(A + x)(\omega_d A + \dot{y}) - y\dot{x}}{(A + x)^2 + y^2} - \omega_d . \tag{49}$$

In a signal-to-noise ratio calculation for the case in which the signal amplitude is usually much larger than the noise on the line, (49) would be written in the form

$$\nu = \frac{\omega_d(1 + x/A) + \dot{y}/A + (x\dot{y} - y\dot{x})/A^2}{(1 + x/A)^2 + (y/A)^2} - \omega_d . \tag{50}$$

If we then assume that A is large compared with x, y, $\dot{x}$, and $\dot{y}$, we retain only first-order terms in small quantities and construct the following approximate result, valid most of the time

$$\begin{aligned} \nu &\approx \omega_d(1 + x/A) + \dot{y}/A - \omega_d(1 + 2x/A) \\ &= (\dot{y} - \omega_d x)/A. \end{aligned} \tag{51}$$

The approximate spectral density of the frequency deviation error is then

$$\begin{aligned} w_\nu(\omega) &\approx [w_{\dot{y}} + \omega_d^2 w_x(\omega)]/A^2 \\ &= 2(\omega^2 + \omega_d^2) w_v(\omega_c + \omega)/A^2. \end{aligned} \tag{52}$$

The approximate mean-square value of error can now be found by integrating the spectral density function $w_\nu(\omega)$ over all frequencies. However, we cannot obtain the probability of error from this value because we do not know the distribution function. A nonlinear operation has been performed on a Gaussian process, and the result must be nongaussian. In this case Rice[2] has shown that the central part of the frequency error distribution is approximately Gaussian. His argument does not apply to the tail. When the signal exceeds the noise most of the time, it is only the tails of the distribution which are important in determining the probability that an error is made in distinguishing between mark and space frequencies.

Since there is no intersymbol interference in our example the exact expression for probability of error is given by (37) with $R = A$ and $\phi = \omega_d$. It can be seen from the limiting forms for large signal-to-noise ratio, (38) through (40), that the Gaussian approximation from (52) cannot approach the correct result. The result obtained from (52) must contain both the original and differentiated noise spectra in the argument of the exponential part of the approximation at large signal-to-noise ratios. In (38) and (39) the exponential depends on either σ_0 or σ_1 but not both.

As another example of the difference between inferences from signal-to-noise ratio and error probability, it is interesting to consider the case of differentially detected binary phase modulation. In this system the polarity of the present carrier wave is compared with the polarity one bit ago. The binary message is read as 1 for a phase reversal and 0 for no phase change. By intuitive reasoning one could easily conclude that there would be a 3-db penalty relative to synchronous detection with a noise-free period. Certainly, in the differential case noise is added to both waves under comparison, and the bit interval is usually long enough to make the two noise samples substantially independent of each other. Signal-to-noise ratio analysis supports the intuitive argument when the average noise power is small relative to the average signal power. A direct calculation of error probability, however, exposes the fallacy and reminds us sharply that the noise is not small compared with the signal when errors occur. If we focus attention on the large noise peaks which cause error, we can see that the simultaneous combination of disturbances on both waves does not imply the same probability of disaster as would follow from concentration of all the noise on one wave.

The differential binary PM problem can in fact be solved as a simple special case of the analysis we have developed for FM. The input wave to the detector can be written as

$$V_r(t) = [P(t) + x(t)] \cos \omega_c t - y(t) \sin \omega_c t. \tag{53}$$

The detector operates by multiplying $V_r(t)$ and $V_r(t - T)$, selecting the low-frequency components of the product, and sampling the output at intervals T apart. If we assume $\omega_c T$ is a multiple of 2π and identify quantities evaluated at $t - T$ by the subscript d, the binary decisions are based on the sign of the wave

$$V_a(t) = (P + x)(P_d + x_d) + yy_d\,. \tag{54}$$

When the correct binary decision is 0, the signs of P and P_d are the same, and an error occurs if the sampled value V_a is negative. When the correct binary decision is 1, the signs of P and P_d are opposite, and an error occurs if the sampled value of V_a is positive. The two cases are symmetric and an analysis of either suffices. For the case of the symbol 0, $P = P_d$, while for the case of 1, $P = -P_d$.

In calculating the signal-to-noise ratio for the case of a symbol 0, we would write

$$V_a = P\left(P + x + x_d + \frac{xx_d + yy_d}{P}\right). \tag{55}$$

Then if P is large compared with x, x_d, y, and y_d, we approach a condition in which the decisions are based on the sign of $P + x + x_d$. If x and x_d are independent, the sum $x + x_d$ represents samples from random noise with twice as much average power as the samples of either x or x_d alone. This tempting argument leads to the 3-db rule.

In a direct calculation of error probability, we recognize that the influence of xx_d and yy_d cannot be ignored at the tails of the noise distribution where the errors occur. In particular, if x and x_d are both very negative, tending to cause an error in a symbol 0, the value of xx_d is large and positive, tending to prevent the threatened damage.

To find the error probability, we compare (54) with (26), and note that we have a special case of the previous solution if we make the following identification

$$\begin{aligned} z \equiv V_a \qquad x_1 &\equiv P + x \qquad & \dot{y}_1 &\equiv P_d + x_d \\ y_1 &\equiv y & \dot{x}_1 &\equiv -y_d . \end{aligned} \tag{56}$$

The remainder of the solution proceeds as before if x, y, x_d, and y_d are independent Gaussian variables. The independence is guaranteed if the second-order correlation functions vanish at lag time T. One difference between this case and the earlier one is that the variables x, y, x_d, and y_d all have the same variance. This specialization can be made in the earlier work by setting $\sigma_0 = \sigma_1 = \sigma$. By comparing with (25), we further note that we can now set $Q = \dot{P} = 0$, $\dot{Q} = P_d = P$. Hence we also have $R = P$ and $\dot{R} = 0$. Corresponding to ϕ we insert the value which V_a/R^2 assumes in the absence of noise, namely $\phi \equiv P^2/P^2 = 1$. In terms of (41), (42), and (43) we then have

$$\rho^2 = \frac{P^2}{2\sigma^2} \qquad a^2 = 1 \qquad b^2 = 0. \tag{57}$$

Hence the answer is given by (40), namely

$$P_+ = P_- = \tfrac{1}{2}e^{-\rho^2}. \tag{58}$$

In the ideal case, a bandwidth f_0 is sufficient to send signals by binary PM at a rate f_0 bits per second without intersymbol interference. This allows for upper and lower sidebands with widths $f_0/2$. If the spectral density of the noise is ν_0 watts/cps, it follows that $\sigma^2 = \nu_0 f_0$. Then M, the ratio of average signal power to the average noise power in a band of width equal to the bit rate, is equal to the ratio of $P^2/2$ to $\nu_0 f_0$ and hence $M = \rho^2$. The formula for error probability is thus found to agree with the one given by Lawton.[3] Average signal power 0.9 db greater than the coherent case is required for an error probability of 10^{-4}. The difference in performance between the differential and purely coherent cases approaches zero at very high signal-to-noise ratios.

V. SUNDE'S BAND-LIMITED FM SYSTEM WITHOUT INTERSYMBOL INTERFERENCE

E. D. Sunde[4] has described a binary FM system in which the intersymbol interference in the absence of noise can be made to vanish at the sampling instants, even when the bandwidth is limited to an extent comparable with that used in AM transmission. The method is remarkable in that a type of result similar to that given by Nyquist[5] for AM systems is obtained for all sequences in spite of the nonlinear FM detection process which invalidates the principle of superposition. The performance falls a little short of the corresponding AM case, in that some dependence on the message appears when noise is added.

Fig. 2 shows a diagram of Sunde's method. The binary message is sent by switching between two oscillators. The difference between the oscillator frequencies must be locked to the bit rate, and the oscillators must be so phased that the frequency transitions are accomplished with continuous phase. The combination of sending filter, line, and receiving filter modify the switched output to produce a spectrum at the input to the frequency detector with even symmetry about the midband and with Nyquist's vestigial symmetry about the marking and spacing fre-

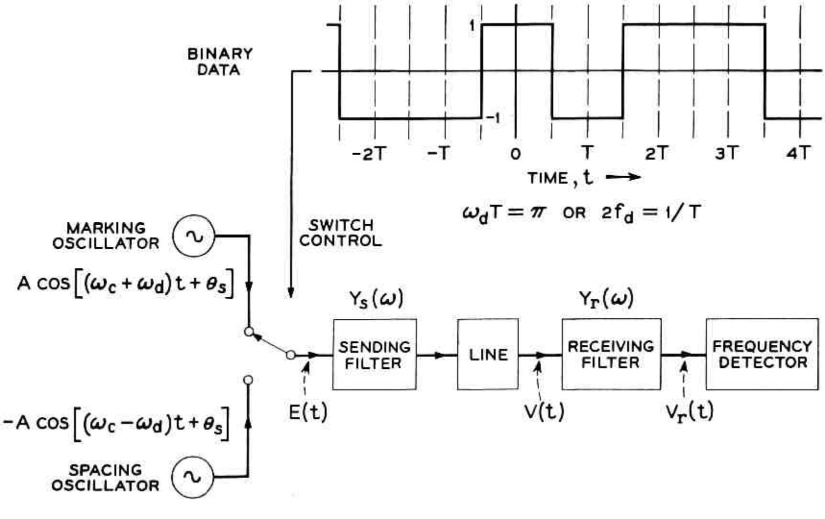


Fig. 2 — Sunde's band-limited binary FM system.

quencies. The latter must be high enough relative to the bit rate to prevent appreciable lower sideband foldover.

The output of the switch is represented by

$$E(t) = \frac{A}{2}[1 - s(t)] \cos [(\omega_c - \omega_d)t + \theta_s] + \frac{A}{2}[1 + s(t)] \cos [(\omega_c + \omega_d)t + \theta_m]. \tag{59}$$

In (59) A represents the amplitude of the output and must be the same for each oscillator. The switching function $s(t)$ represents the baseband data wave of (1). When $s(t) = -1$, the first term has amplitude A and the second term vanishes. When $s(t) = +1$, the first term vanishes and the second has amplitude A. The center of the band is the frequency ω_c and the total frequency shift is $2\omega_d$. For minimum bandwidth the angular signaling frequency $\omega_0 = 2\pi/T$ must be equal to $2\omega_d$. One of the two phase angles θ_s and θ_m can be arbitrary, but the two angles must differ by 180 degrees. Under these restrictions, the value of $E(t)$ can be written as

$$E(t) = A \sin \omega_d t \sin (\omega_c t + \theta_s) - As(t) \cos \omega_d t \cos (\omega_c t + \theta_s). \tag{60}$$

Sunde requires that the input wave to the frequency detector can be written in the form

$$V_r(t) = A \sin \omega_d t \sin (\omega_c t + \theta_r) - As_1(t) \cos (\omega_c t + \theta_r) \tag{61}$$

where $s_1(t)$ represents the data sequence with $g(t)$ replaced by $g_1(t)$. The latter must be a pulse which gives no intersymbol interference when the data rate is $1/T$. That is,

$$s_1(t) = \sum_{n=-\infty}^{\infty} (-)^n b_n g_1(t - nT) \tag{62}$$

and $g_1(t)$ assumes the value unity at $t = 0$ and has nulls at all instants differing from $t = 0$ by multiples of T. In mathematical notation

$$g_1[(m - n)T] = \delta_{mn} \tag{63}$$

and

$$s_1(mT) = (-)^m b_m. \tag{64}$$

The requirement as actually stated by Sunde differs from (61) in that his analysis is based on a switching function which assumes the values 1 and 0 at the sampling instant rather than 1 and -1. The two expressions for the requirement can be shown to be equivalent. Equation

(61) has the advantage that the function $s_1(t)$ has the average value zero for a random data sequence with equal probability of the two binary symbols. This fact enables an easy separation of the spectral density of $V_r(t)$ into line spectra contributed by the first term of (61) and a continuous spectral density function for the second part.

Incidentally, it is clear from (61) that all the signal information is contained in the second term, and that the first term can be regarded as a pair of pilot tones at the marking and spacing frequencies $\omega_c \pm \omega_d$. The sole function of these pilot tones is to enable an FM detector to recover the message. The information carrying part of $V_r(t)$ can equally well be regarded as double-sideband suppressed-carrier binary AM or binary phase modulation, with the carrier frequency placed at ω_c. The ideal way of detecting such signals is by multiplication with a coherent carrier wave, which must be transmitted as part of the data wave in some way. Detection of $V_r(t)$ as FM has a practical advantage in that there is no carrier recovery problem; the wave is ready for the frequency detector with no further processing. The penalty for transmitting pure sine waves is a waste of signal power. As will be shown quantitatively later, such waste results in an unfavorable comparison with more nearly ideal systems.

To show that the stipulated conditions are sufficient to suppress intersymbol interference in the detected frequency of $V_r(t)$, we identify $P(t)$ and $Q(t)$ of (5) with the applicable terms of (61) as follows

$$P(t) = -As_1(t) \tag{65}$$

$$Q(t) = -A \sin \omega_d t. \tag{66}$$

We then calculate

$$P'(t) = -As_1'(t) \tag{67}$$

$$Q'(t) = -\omega_d A \cos \omega_d t. \tag{68}$$

If we take frequency samples at $t = mT$ we find that since $\omega_d T = \pi$

$$\begin{aligned} P(mT) &= (-)^{m+1} b_m \\ P'(mT) &= -As_1'(mT) \\ Q(mT) &= 0 \\ Q'(mT) &= (-)^{m+1} \omega_d A. \end{aligned} \tag{69}$$

Hence in (23), evaluated at $t = mT$

$$\dot{\phi} = \dot{Q}/P = \omega_d/b_m = b_m \omega_d . \tag{70}$$

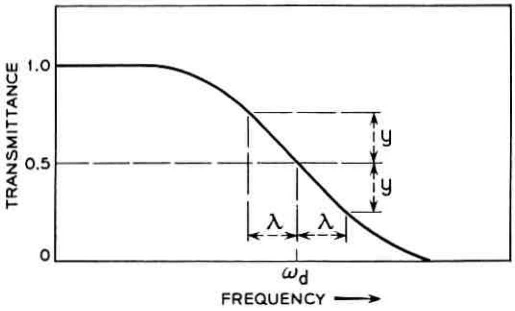


Fig. 3 — Nyquist's condition of vestigial symmetry.

The value of the instantaneous frequency deviation at the *m*th sampling point is, therefore, equal to ω_d if $s(mT) = 1$ and equal to $-\omega_d$ if $s(mT) = -1$. Freedom from intersymbol interference is thus obtained if (64) is satisfied.

As shown by Nyquist, a sufficient condition for obtaining (64) is that the standard pulse $g_1(t)$ is the impulse response of a network with transmittance $G_1(\omega)$ of the form shown in Fig. 3, described mathematically by

$$G_1(\pm\omega_d - \lambda) + G_1(\pm\omega_d + \lambda) = 2G_1(\omega_d) = T \qquad 0 < \lambda < \omega_d . \quad (71)$$

We say that a function satisfying (71) has vestigial symmetry about frequency ω_d because it has the type of symmetry called for in a vestigial sideband filter with the carrier at ω_d. We can think of the response at a frequency exceeding ω_d by an amount λ as exactly compensating the deficiency in the response at the frequency less than ω_d by the same amount λ. The ideal low-pass filter is a limiting special case occurring when the transmittance vanishes for $|\omega| > \omega_d$. The amplitude can be associated with linear phase shift, which changes only the origin of time. Unnecessary complication is avoided by carrying through the calculations with zero phase shift.

The conditions imposed on the filters and line to transform (60) to (61) can be expressed in terms of the Fourier transforms of $g(t)\cos\omega_d t$ and $g_1(t)$, which we represent respectively by $C(\omega)$ and $G_1(\omega)$. Both $C(\omega)$ and $G_1(\omega)$ are purely real and are given by

$$\begin{aligned} C(\omega) &= \int_{-\infty}^{\infty} g(t)\cos\omega_d t \cos\omega t \, dt \\ &= [G(\omega - \omega_d) + G(\omega + \omega_d)]/2 \end{aligned} \qquad (72)$$

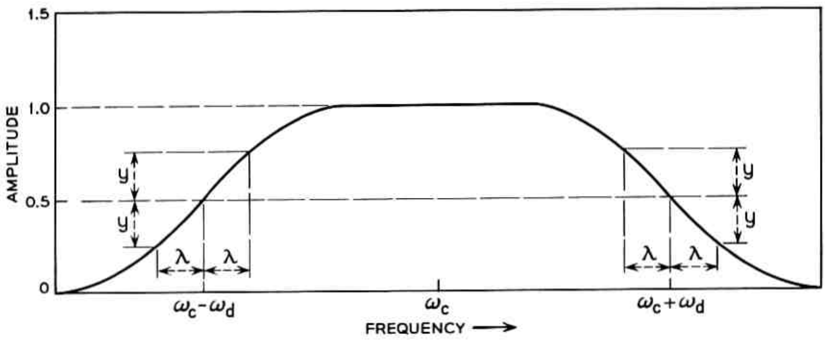


Fig. 4 — Spectrum at input to detector in Sunde's FM system.

$$G_1(\omega) = \int_{-\infty}^{\infty} g_1(t) \cos \omega t \, dt. \tag{73}$$

The result, obtained by multiplying $\cos(\omega_c t + \theta_s)$ by $g(t) \cos \omega_d t$ or $g_1(t)$, is to place upper and lower sidebands on the frequencies $\pm\omega_c$, as shown in Fig. 4, with spectra equal to $C(\omega - \omega_c)/2$ and $G_1(\omega - \omega_c)/2$ respectively on ω_c. The required transmittance function for the combination of sending filter, line, and receiving filter is then

$$Y(\omega) = \frac{G_1(\omega - \omega_c)}{C(\omega - \omega_c)}. \tag{74}$$

This function transforms the second term of (60) to the second term of (61). It is also necessary for the first term of (60) to remain unchanged. The first term can be written as the difference of sine waves of frequencies $\omega_c - \omega_d$ and $\omega_c + \omega_d$. These components will be unchanged by the operation $Y(\omega)$ if

$$C(\pm\omega_d) = G_1(\pm\omega_d) \quad \text{or} \quad Y(\omega_c \pm \omega_d) = 1. \tag{75}$$

It can readily be seen that the condition (71) required on $G_1(\omega)$ translates to the same condition for $G_1(u)$ where $u = \omega - \omega_c$.

The relations can be made clearer by working out an example. Suppose the switching is rectangular and there is no lost time between contacts. The function $g(t)$ is then defined by

$$g(t) = \begin{matrix} 1, & -T/2 < t < T/2 \\ 0, & |t| > T/2. \end{matrix} \tag{76}$$

Let the received signal $V_r(t)$ have a full raised cosine spectrum centered at ω_c, with vestigial symmetry about $\omega_c + \omega_d$ and $\omega_c - \omega_d$. We then write

$$G_1(u) = \begin{cases} T\left(1 + \cos \dfrac{\pi u}{2\omega_d}\right) \Big/ 2 & |u| \leqq 2\omega_d \\ 0 & |u| > 2\omega_d . \end{cases} \tag{77}$$

We calculate

$$C(\omega) = 2 \int_0^{T/2} \cos \omega_d t \cos \omega t \, dt = \frac{2\omega_d \cos (\omega T/2)}{\omega_d^2 - \omega^2} \tag{78}$$

$$Y(\omega) = \frac{\pi(\omega_d^2 - u^2)\left(1 + \cos \dfrac{\pi u}{2\omega_d}\right)}{4\omega_d^2 \cos \dfrac{\pi u}{2\omega_d}} \qquad u = \omega_c - \omega. \tag{79}$$

This function satisfies the required condition that $Y(\omega_c \pm \omega_d) = 1$.

In practice it is difficult to control two oscillators with the necessary precision to meet Sunde's requirements. One method of realizing the system approximately is to begin with two high-frequency crystal-controlled oscillators of frequencies $n(\omega_c - \omega_d)$ and $n(\omega_c + \omega_d)$, where n is a large integer. The phases of the two oscillators are not under control and are assumed to be θ_1 and θ_2, respectively. Frequency step-down circuits are introduced after each oscillator to give outputs of frequency $\omega_c - \omega_d$ and $\omega_c + \omega_d$ with respective phases θ_1/n and θ_2/n. By multiplying these two outputs and selecting the low-frequency component as shown in Fig. 5, we obtain a wave of frequency $2\omega_d$ and phase $(\theta_2 - \theta_1)/n$. This wave can be used to control the timing of the binary input symbols. For the switched marking and spacing frequency sources we use the stepped-down component of frequency $\omega_c - \omega_d$ directly and the component of frequency $\omega_c + \omega_d$ with reversed polarity. The required frequency and phase relations are then satisfied

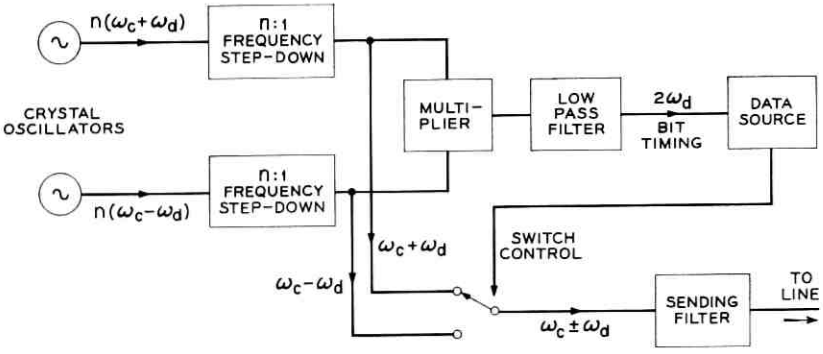


Fig. 5 — Practical realization of Sunde's system.

except for a slow drift in the time scale caused by the lack of perfect stability in the original oscillators.

To calculate the probability of error when Gaussian noise is added to Sunde's FM signal, we identify the values of $P(mT)$, $P'(mT)$, $Q(mT)$, and $Q'(mT)$ of (69) with P, $\dot{P}$, Q, and $\dot{Q}$ respectively. The general expression for the probability of error, (34), is expressed in terms of R and $\dot{R}$. We calculate

$$R = (P^2 + Q^2)^{\frac{1}{2}} = A \tag{80}$$

$$\begin{aligned} R'(t) &= \frac{d}{dt}[P^2(t) + Q^2(t)]^{\frac{1}{2}} \\ &= [P(t)P'(t) + Q(t)Q'(t)]/R(t) \end{aligned} \tag{81}$$

$$\dot{R} = R'(mT) = (P\dot{P} + Q\dot{Q})/R = (-)^{m+1}Ab_m s_1'(mT). \tag{82}$$

From (62)

$$s_1'(mT) = \sum_{n=-\infty}^{\infty} (-)^n b_n g_1'[(m-n)T]. \tag{83}$$

From (73) we verify

$$\begin{aligned} g_1(rT) &= \frac{1}{\pi}\int_0^{2\omega_d} G_1(\omega)\cos(\omega rT)\,d\omega \\ &= \frac{1}{\pi}\int_0^{\omega_d} G_1(\omega_d - \omega)\cos[rT(\omega_d - \omega)]\,d\omega \\ &\qquad + \frac{1}{\pi}\int_0^{\omega_d} G_1(\omega_d + \omega)\cos[rT(\omega_d + \omega)]\,d\omega \\ &= \frac{2}{\pi}G_1(\omega_d)\cos r\pi\int_0^{\omega_d}\cos r\omega T\,d\omega = \delta_{r0}\,. \end{aligned} \tag{84}$$

This checks our previous requirements expressed by (63) and (64). By differentiating (73) and substituting $t = rT$, we find

$$g_1'(rT) = -\frac{1}{\pi}\int_0^{2\omega_d} \omega G_1(\omega)\sin\omega rT\,d\omega. \tag{85}$$

The value of this integral in general is not zero except when $r = 0$. It appears, therefore, that at any sampling instant $t = mT$ the value of $\dot{R}$ depends on all the values of b_n in the sequence except b_m.

For further progress we take a specific example, namely the full raised cosine spectrum for $G_1(\omega)$. We set

$$G_1(\omega) = T\left(1 + \cos\frac{\pi\omega}{2\omega_d}\right)\Big/2 \qquad |\,\omega\,| \leqq 2\omega_d\,. \tag{86}$$

Then

$$\begin{aligned} g_1'(rT) &= -\frac{T}{2\pi}\int_0^{2\omega_d}\omega\left(1 + \cos\frac{\pi\omega}{2\omega_d}\right)\sin\omega rT\,d\omega \\ &= \frac{f_0}{r(1 - 4r^2)} \qquad r \neq 0. \end{aligned} \tag{87}$$

From (85), we noted that $g_1'(0) = 0$. The value of $\dot{R}$ can now be found from (82), thus

$$\begin{aligned} \dot{R} &= (-)^{m+1}\, b_m f_0 A\left[\sum_{n=-\infty}^{m-1} + \sum_{n=m+1}^{\infty}\right]\frac{(-)^n b_n}{(m - n)[1 - 4(m - n)^2]} \\ &= -b_m f_0 A \sum_{n=1}^{\infty} (-)^n\,\frac{b_{m+n} - b_{m-n}}{n(4n^2 - 1)}\,. \end{aligned} \tag{88}$$

We observe from our previous study of the integral defining the probability of error that for fixed R the most vulnerable sequence is the one which has the largest absolute value of $\dot{R}$. The least vulnerable sequence is the one for which $\dot{R} = 0$, and this can be obtained by setting $b_{m+n} = b_{m-n}$ for all n. The maximum absolute value of $\dot{R}$ occurs when b_{m+n} and b_{m-n} have opposite signs and the signs are reversed when n changes by unity. The resulting value of $|\,\dot{R}\,|$ is[6]

$$\begin{aligned} \dot{R}_m &= 2f_0 A \sum_{n=1}^{\infty}\frac{1}{n(4n^2 - 1)} \\ &= 2f_0 A\ (\log_e 4 - 1) = 0.7726\, f_0 A. \end{aligned} \tag{89}$$

The upper and lower bounds for the error probability are found by substituting $\dot{R}_m$ and 0 respectively for $\dot{R}$ in (34). By (80) the value of R is constant and equal to A. From (70), $\phi = b_m\omega_d$. It is important to note that while the intersymbol interference is suppressed in the absence of noise the error probability with noise present does depend on the signal sequence. This occurs because frequency detection is a nonlinear process, and the effect of noise cannot be found by merely adding a noise wave to the detected frequency output.

The actual spectral density of the noise facing the frequency detector is under the control of the system designer, since the selectivity of the receiving bandpass filter is not determined by the requirements thus far discussed. We have stated what the received signal spectrum at the detector input should be, but this is a resultant of signal shaping at the transmitter, the transmitting filter selectivity, and the transmittance of the line, as well as receiving filter selectivity. The latter can be varied within reasonable limits if the others are adjusted in a complementary fashion to obtain the desired output response. In evaluating the merit of different receiving filter designs it is reasonable to compare them with the same average signal power on the line. We shall also assume that the line has been equalized for unity gain and linear phase over the band so that it can be considered as a transparent link in the system.

The average signal power on the line can be computed in terms of (a) the transmittance function $Y_r(\omega)$ of the receiving filter, (b) the required function $G_1(\omega)$ representing the spectrum of the modified switching function $g_1(t)$ at the detector input, and (c) the statistics of the data sequence. Details of the calculation are given in Appendix B. An interesting consequence of the assumptions that the FM wave has continuous phase and that the frequency shift is equal to the signaling rate is the appearance of discrete components on the line at the marking and spacing frequencies even when the data sequence is random. This means there are transmitted sine waves which consume power but carry no information. An optimization procedure aimed at conserving power would very nearly suppress these components at the transmitter by balance or by sharp antiresonances and restore them to their proper relative amplitudes by complementary narrow-band resonance peaks in the response of the receiving bandpass filter. The bandwidth used to augment these frequencies at the receiver could in theory be made so small that no appreciable effect on the accepted noise would result. The system would then only have to deliver the average power associated with the continuous part of the FM spectrum.

Actually, even a partial suppression of the steady-state components on the line would destroy much of the advantage of signaling by FM. The system would become more sensitive to gain changes and overload distortion. Accurate tracking of the suppression and recovery circuits for the marking and spacing frequencies would be difficult at best and would be practically impossible over a channel with carrier frequency offset. The narrow-band recovery circuits would contribute to a sluggish start-up time. In fact, about the only remaining resemblance to FM would be the use of an FM detector. If low-level tones can actually be recovered successfully from a received wave, it would be better to use them for synchronous PM detection, which is a linear method capable of attaining ideal performance in the presence of additive Gaussian noise. It appears that Sunde's system should carry the power in the steady-state components in order to deserve the name of FM.

Standard variational procedures can be applied to find the shape of receiving filter selectivity which minimizes probability of error when the average signal power and the spectral density of added Gaussian noise on the line are specified. The solution of the optimization problem is given in Appendix B, and means are shown for completing the computation of the corresponding probabilities of error for the most and least vulnerable data sequences. In the case of white Gaussian noise on the line, the optimum receiving filter has very nearly the same cosine characteristic found by Sunde for optimum binary AM transmission. The bounds for error probability are plotted in Fig. 6 for both FM proper with no suppression of steady-state tones and the abnormal FM with marking and spacing frequencies suppressed. Also shown is the ideal curve representing what can be proved to give the best possible binary performance. The ideal curve can theoretically be obtained for example by coherent detection of binary phase modulation. Differentially detected phase modulation requires about 1 db more signal power than ideal at an error probability of 10^{-4}.

It is seen from Fig. 6 that when the suppression bands are inserted in Sunde's binary FM system, the theoretical performance is only about a half db poorer than ideal, but, as previously pointed out, this does not represent a true FM system. The more legitimate FM has error bounds from 3 to 3.5 db poorer than ideal. However, a penalty of this order of magnitude could be a fair trade in many cases for the advantages of a much simplified receiver relatively immune to many channel faults.

VI. APPLICATION TO DATA TERMINALS FOR USE ON TELEPHONE CHANNELS

We now apply our formulas to calculate error probabilities in binary FM transmission with terminals more closely resembling those actually in use on telephone channels. In the design of real-life terminals, the emphasis is placed on ruggedness and simplicity. The bit rate is not locked to the frequency deviation. The filters do not meet elaborate optimization requirements. The significant conclusion from our evaluation of error probabilities for the practical systems is that the degradation of performance compared with the ideal is actually very slight.

The probability of error as given in (44) is generally applicable to FM systems. There are three parameters, ρ, a, and b, given in (41) to (43). The first parameter ρ is a signal-to-noise ratio. It depends on the

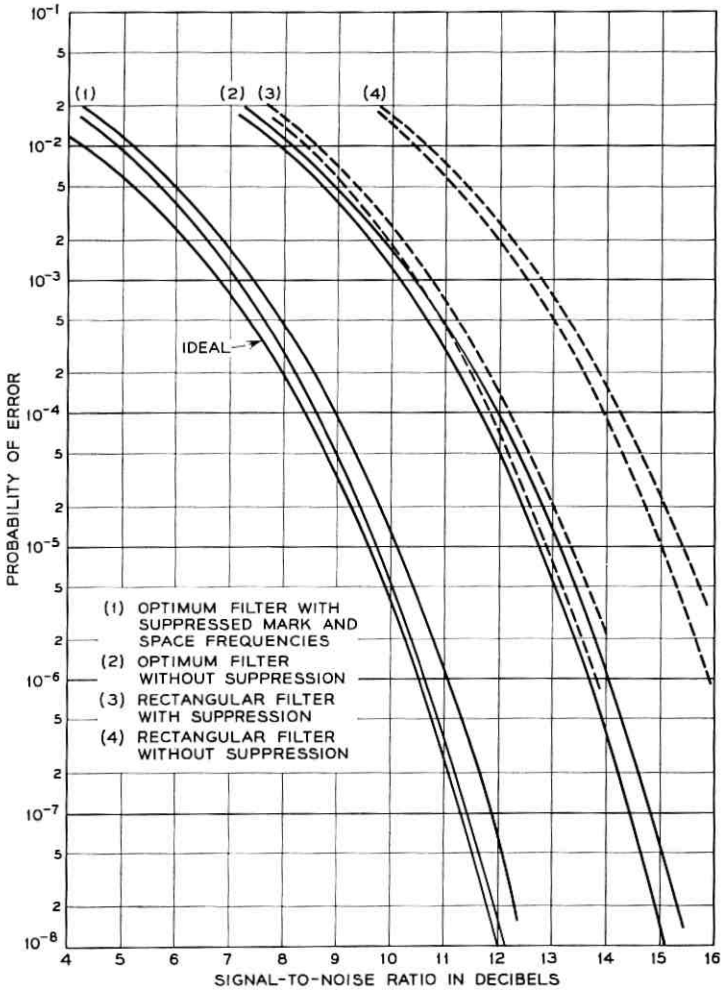


Fig. 6 — Error probabilities for Sunde's binary FM system with additive Gaussian noise. Bounds are for most and least vulnerable sequences. Noise reference is mean noise power in bandwidth equal to bit rate.

ratio of instantaneous envelope of the received signal to the rms noise voltage at the detector input. For any given front-end filter, this parameter can be expressed in terms of average signal-to-noise ratio at the input of the receiver. The parameter a depends on the ratio of instantaneous frequency displacement at the sampling time to the Gabor noise bandwidth, σ_1/σ_0, of the receiver. The third parameter b depends on the derivative of the instantaneous envelope at the sampling time. For a given channel these parameters can be computed for any particular signaling sequence. The true probability of error could conceivably be obtained by averaging over all possible sequences, but this would be a formidable task. Instead we will give bounds on the probability of error for the most and least vulnerable sequences over a finite representative set of signaling intervals.

We first consider the system in Fig. 7, which has amplitude-vs-frequency "raised cosine" type roll-off but no phase distortion. Equal filtering takes place at the transmitter and receiver. The modulator applies a pure FM wave of constant envelope to the transmitting filter. In other words, the modulator and the demodulator are ideal. The data source is composed of rectangular pulses. The frequency deviation in cps is equal to half the bit rate. These rates and deviations are characteristic of practical systems.

With the aid of a digital computer, S. Habib has calculated the parameters given in (41) to (43) for 2^{10} sequences. From these calculations we have computed an upper and a lower bound on the probability of error. These results are shown in Fig. 8. The probability of error for all other sequences will fall between the two curves labeled "best" and "worst." Superimposed on the same graph is the ideal curve, which can only be achieved with ideal phase systems and coherent detection. The FM detection is, of course, incoherent.

Our next example applies the theory to a real bandpass filter used in an operational data set. Fig. 9 shows the system considered. The curve

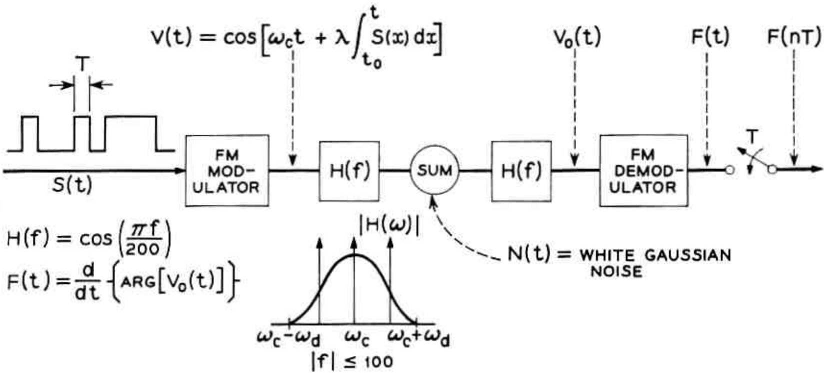

$V(t) = \cos\left[\omega_c t + \lambda \int_{t_0}^{t} S(x)\,dx\right]$
$V_0(t)$
$F(t)$
$F(nT)$
T
$S(t)$
FM MODULATOR
$H(f)$
SUM
$H(f)$
FM DEMODULATOR
T
$|H(\omega)|$
$N(t) =$ WHITE GAUSSIAN NOISE
$H(f) = \cos\left(\frac{\pi f}{200}\right)$
$F(t) = \frac{d}{dt}\left\{\text{ARG}[V_0(t)]\right\}$
$\omega_c - \omega_d$
ω_c
$\omega_c + \omega_d$
$|f| \leq 100$

Fig. 7 — Ideal FM modulator and demodulator with transmitted and received signals equally shaped by "raised cosine" type roll-off amplitude characteristics and no phase distortion.

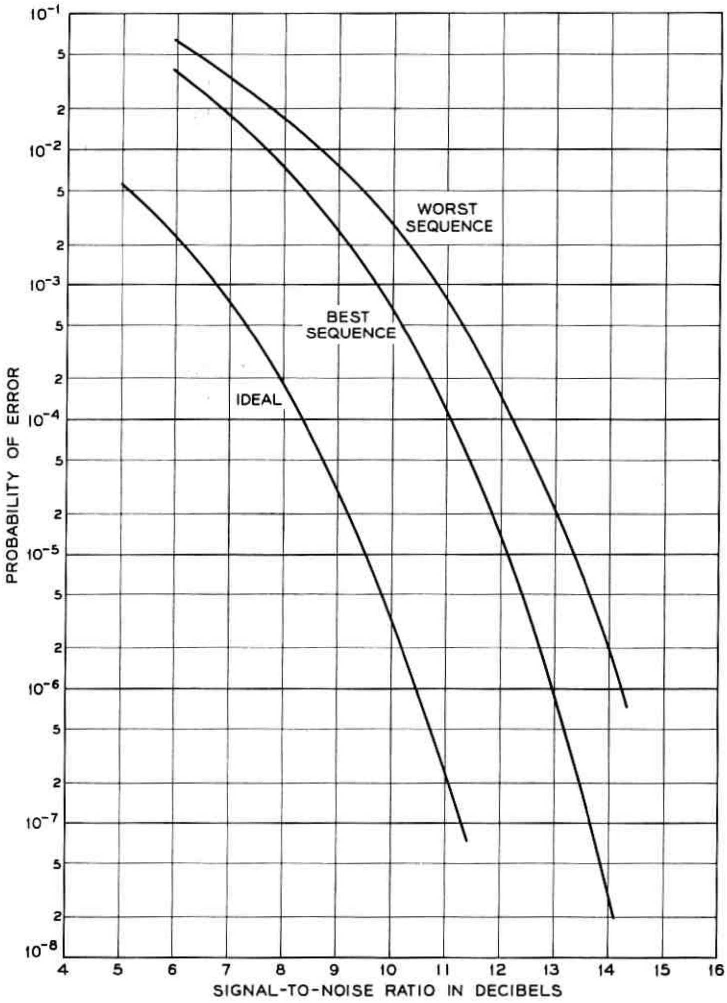


Fig. 8 — Probability of error for system depicted in Fig. 7. Noise reference is mean power in bandwidth equal to bit rate.

of loss vs frequency for the filter used is given in Fig. 10. The curve departs from the condition of symmetry about midband, and also the separation between the signal and carrier bands is not sufficient to make overlapping effects negligible. The marking and spacing frequencies were assumed to be 1200 and 2200 cps, respectively, and the signaling rate 1200 bits per second. As shown in Fig. 11, the calculated results are

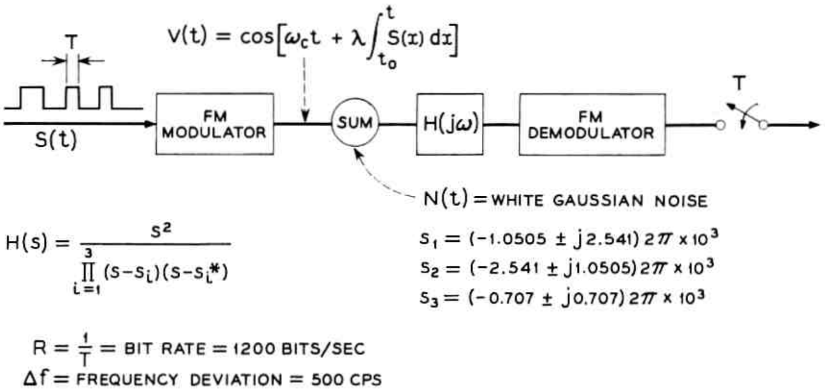

T
$v(t) = \cos\left[\omega_c t + \lambda \int_{t_0}^{t} S(\tau)\, d\tau\right]$
S(t)
FM MODULATOR
SUM
$H(j\omega)$
FM DEMODULATOR
T
N(t) = WHITE GAUSSIAN NOISE
$H(s) = \dfrac{s^2}{\prod_{i=1}^{3}(s-s_i)(s-s_i^*)}$
$s_1 = (-1.0505 \pm j2.541)\, 2\pi \times 10^3$
$s_2 = (-2.541 \pm j1.0505)\, 2\pi \times 10^3$
$s_3 = (-0.707 \pm j0.707)\, 2\pi \times 10^3$
$R = \frac{1}{T}$ = BIT RATE = 1200 BITS/SEC
Δf = FREQUENCY DEVIATION = 500 CPS

Fig. 9 — Ideal modulator and demodulator with received signal shaped by filter characteristics used in FM data set and shown in Fig. 10.

about 1 db better than the experimental results obtained with a random word generator, random noise generator, and error counter. The experimental system included an axis-crossing detector and post-detection low-pass filter, which do not correspond precisely with the theoretical model. In view of the differences cited, the agreement between calculated and experimental curves is good. The penalty suffered by the actual back-to-back channel compared with the best theoretical FM performance is between 2 and 3 db. Somewhat more optimistic estimates have been given in other published studies.[7,8] The effects of amplitude and delay-versus-frequency variation in the channel are calculable by use of the computer programs we have established.

It was shown in the previous sections that a lower bound on the probability of error occurs when the parameter b is set equal to zero. For

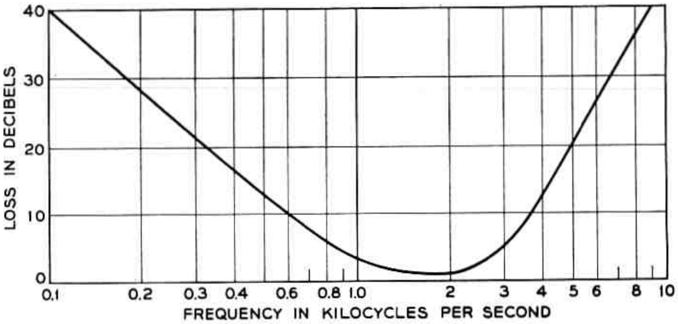


Fig. 10 — Receiver bandpass filter loss vs frequency characteristic.

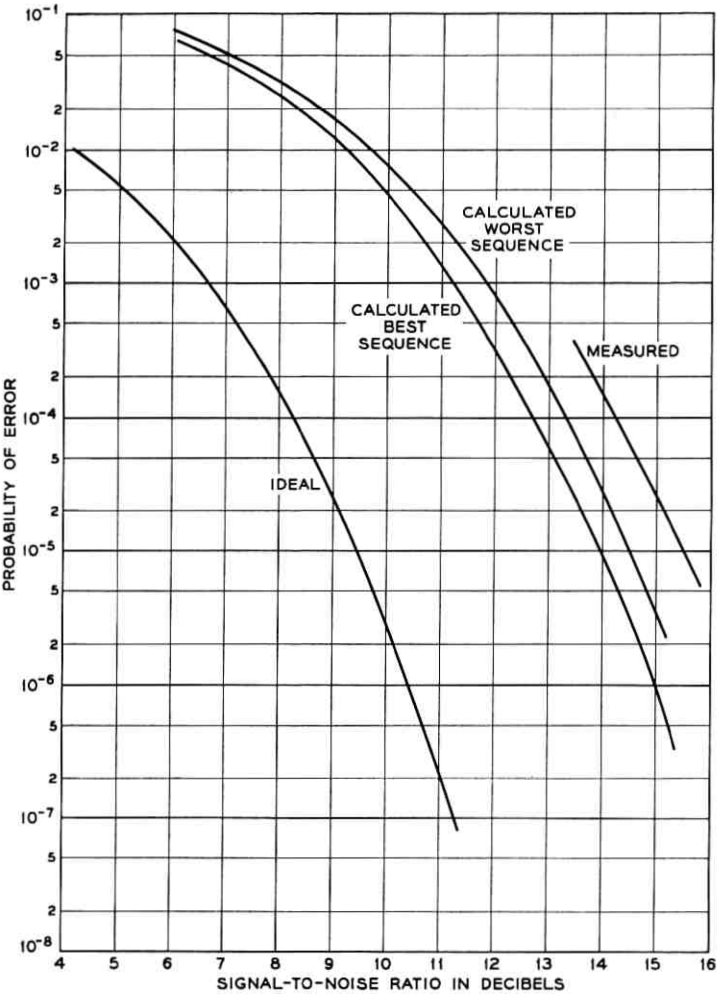


Fig. 11 — Calculated and measured error rates for system depicted in Fig. 9. Noise reference is mean noise power in bandwidth equal to bit rate.

this reason we include Fig. 12, showing a set of universal curves relating the corresponding minimum probability of error to ρ and a.

APPENDIX A

Evaluation of Integral for Error Probability

We evaluate the integral

$$P_+ = \int_0^\infty dz \int_{-\infty}^\infty \int_{-\infty}^\infty p(-z \mid x,y) q(x,y) \, dx \, dy \tag{90}$$

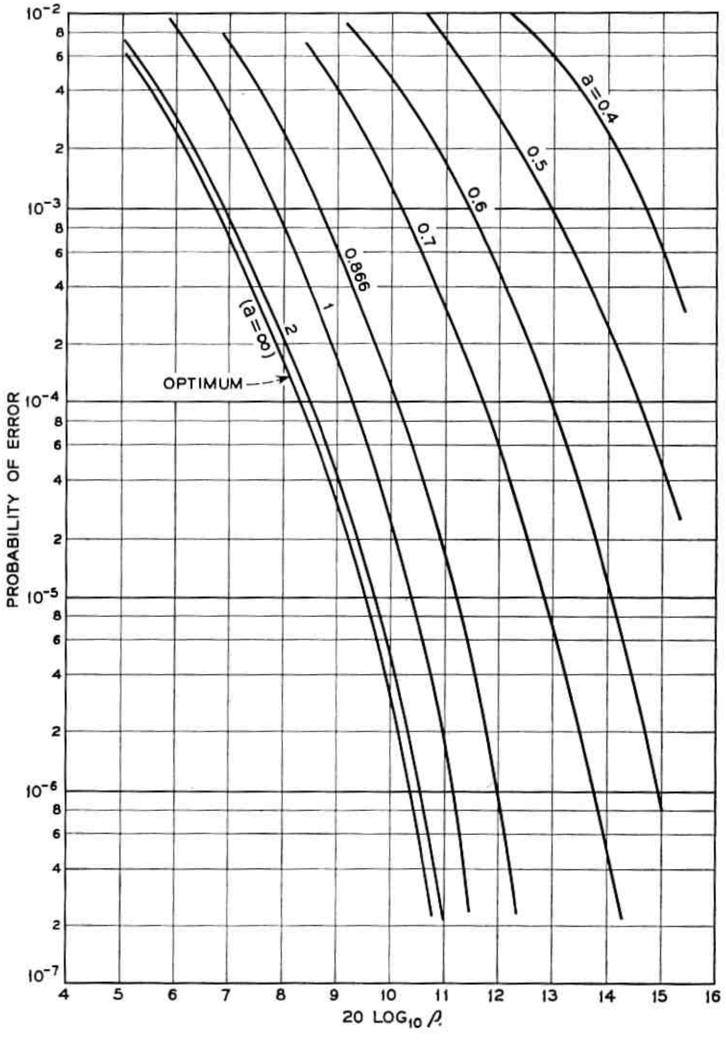


Fig. 12 — General curves of minimum probability of error vs ρ for different values of a in the range of interest.

where

$$p(-z \mid x,y) = \frac{1}{\sigma_1[2\pi(x^2 + y^2)]^{\frac{1}{2}}} \exp\left[-\frac{(z + \dot{Q}x - \dot{P}y)^2}{2\sigma_1^2(x^2 + y^2)} \right] \quad (91)$$

$$q(x,y) = \frac{1}{2\pi\sigma_0^2} \exp\left[-\frac{(x - P)^2 + (y - Q)^2}{2\sigma_0^2} \right]. \quad (92)$$

The integration with respect to z can be performed at once in terms of the error function by substituting a new variable u defined by

$$(z + \dot{Q}x - \dot{P}y)^2 = 2\sigma_1^2(x^2 + y^2)u^2. \tag{93}$$

The result is:

$$P_+ = \frac{1}{2}\frac{1}{4\pi\sigma_0^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \operatorname{erf}\frac{\dot{Q}x - \dot{P}y}{\sigma_1[2(x^2 + y^2)]^{\frac{1}{2}}} \cdot \exp\left[-\frac{(x - P)^2 + (y - Q)^2}{2\sigma_0^2}\right] dx\,dy. \tag{94}$$

We now transform to polar coordinates, setting

$$x = r\cos\theta \qquad y = r\sin\theta \qquad dx\,dy = r\,dr\,d\theta \tag{95}$$

We also let

$$P\cos\theta + Q\sin\theta = R\cos(\theta - \alpha) = R\cos\psi$$
$$\dot{Q}\cos\theta - \dot{P}\sin\theta = D\cos(\theta + \beta) = D\cos(\psi + \gamma)$$

where

$$\begin{aligned} R^2 &= P^2 + Q^2 = 2\sigma_0^2\rho^2 & \tan\alpha &= Q/P \\ D^2 &= \dot{P}^2 + \dot{Q}^2 & \tan\beta &= \dot{P}/\dot{Q} \\ \psi &= \theta - \alpha & \gamma &= \alpha + \beta. \end{aligned} \tag{96}$$

The result of the transformation is

$$1 - 2P_+ = \frac{e^{-\rho^2}}{2\pi\sigma_0^2}\int_{-\pi}^{\pi} \operatorname{erf}\frac{D\cos(\psi + \gamma)}{\sqrt{2}\,\sigma_1}\,d\psi \int_0^{\infty} \exp\left[-\frac{r^2 - 2rR\cos\psi}{2\sigma_0^2}\right] r\,dr. \tag{97}$$

The integration with respect to r can be performed by subtracting and adding the term $R\cos\psi$ to r. This enables separation of the integrand into a perfect differential and a term which can be expressed as an error function. We thereby obtain

$$1 - 2P_+ = \frac{1}{2\pi}\int_{-\pi}^{\pi} \operatorname{erf}\frac{D\cos(\psi + \gamma)}{\sqrt{2}\,\sigma_1} \left[1 + \sqrt{2\pi}\,\frac{R}{2\sigma_0}\exp\left(-\frac{R^2}{2\sigma_0^2}\sin^2\psi\right)\cos\psi\left(1 - \operatorname{erf}\frac{R\cos\psi}{\sqrt{2}\,\sigma_0}\right)\right] d\psi. \tag{98}$$

We note that both $\cos\psi$ and $\cos(\psi + \gamma)$ change sign when ψ is increased by π and that $\sin^2(\psi + \pi) = \sin^2\psi$. Furthermore, the integration in (98) is over one full period in ψ, and for every value of ψ in the left half of the period there is a corresponding value in the right half at $\psi + \pi$. Since the error function, erf z, is an odd function of z, a change in the sign of $\cos\psi$ or $\cos(\psi + \gamma)$ changes the sign of the corresponding error function in the integrand. If we multiply the first term under the integral sign by the terms within the bracket following, we see that there is only one product which does not change sign at points π apart. The integral of the other products must vanish. The integral of the one which does not change sign is twice the integral over a half period of ψ. Hence

$$1 - 2P_+ = \frac{R}{\sigma_0\sqrt{2\pi}} \int_{-\pi/2}^{\pi/2} \exp\left(-\frac{R^2}{2\sigma_0^2}\sin^2\psi\right) \cdot \cos\psi \operatorname{erf} \frac{D\cos(\psi+\gamma)}{\sqrt{2}\,\sigma_1}\,d\psi. \tag{99}$$

From (96) and (23)

$$\begin{aligned} D\cos\gamma &= D(\cos\alpha\cos\beta - \sin\alpha\sin\beta) \\ &= D\left(\frac{P}{R}\frac{\dot{Q}}{D} - \frac{Q}{R}\frac{\dot{P}}{D}\right) = \frac{P\dot{Q} - Q\dot{P}}{R} = R\phi \end{aligned} \tag{100}$$

$$\begin{aligned} D\sin\gamma &= D(\sin\alpha\cos\beta + \cos\alpha\sin\beta \\ &= D\left(\frac{Q}{R}\frac{\dot{Q}}{D} + \frac{P}{R}\frac{\dot{P}}{D}\right) = \frac{Q\dot{Q} + P\dot{P}}{R} \\ &= \frac{1}{2R}\frac{d}{dt}(R^2) = \frac{dR}{dt} = \dot{R}. \end{aligned} \tag{101}$$

Therefore

$$\begin{aligned} D\cos(\psi+\gamma) &= D\cos\gamma\cos\psi - D\sin\gamma\sin\psi \\ &= R\phi\cos\psi - \dot{R}\sin\psi. \end{aligned} \tag{102}$$

Now substituting $x = \sin\psi$ in (99) we rearrange to obtain

$$P_+ = \frac{1}{2} - \frac{\rho}{2\sqrt{\pi}} \int_{-1}^{1} e^{-\rho^2 x^2} \operatorname{erf} \frac{R\phi(1 - x^2)^{\frac{1}{2}} - \dot{R}x}{\sqrt{2}\,\sigma_1}\,dx. \tag{103}$$

Equation (34) of the main text is obtained from (103) by substituting the complementary function erfc $z = 1 -$ erf z.

The lower bound P_l on the probability of error for any fixed R and ϕ was shown in the text to be obtained by setting $\dot{R} = 0$. When this substitution is made in (103) and the definition of the error function in terms of an integral is inserted, we obtain

$$P_l = \frac{1}{2} - \frac{2\rho}{\pi} \int_0^1 e^{-\rho^2 x^2} \, dx \int_0^{a\rho(1-x^2)^{\frac{1}{2}}} e^{-z^2} \, dz. \tag{104}$$

The parameters a and ρ are defined by (41) and (42). If we substitute $\rho x = y$ the expression becomes

$$P_l = \frac{1}{2} - \frac{2}{\pi} \int_0^{\rho} \int_0^{a(\rho^2 - y^2)^{\frac{1}{2}}} e^{-y^2 - z^2} \, dy \, dz. \tag{105}$$

The region of integration in the double integral consists of the first quadrant of the ellipse

$$z^2/(a\rho)^2 + y^2/\rho^2 = 1. \tag{106}$$

After transforming to polar coordinates by setting $y = r \cos \theta$ and $z = r \sin \theta$, we can perform the integration with respect to r. The result is

$$P_l = \frac{1}{\pi} \int_0^{\pi/2} \exp \left[- \frac{a^2 \rho^2}{\sin^2 \theta + a^2 \cos^2 \theta} \right] d\theta. \tag{107}$$

This is equivalent to (37) of the main text.

The integral has a simple value when $a = 1$, which is equivalent to $\phi = \sigma_1/\sigma_0$. For this case the integrand is seen to become a constant and (40) results. This coincides with a result given for a special case by Montgomery.[9] By a change in the meaning of the parameters it also gives the error probability for differential binary phase detection as discussed in Section IV. In the general case, the limiting form of P_l for large signal-to-noise ratio can be calculated by the method of steepest descents. Saddle points occur at $\theta = 0$ and $\theta = \pi/2$. When $a > 1$, the saddle point at $\theta = 0$ determines the asymptotic form of the integral for large ρ and (38) is obtained. When $a < 1$, the saddle point at $\theta = \pi/2$ is dominant and we obtain (39).

APPENDIX B

Optimization of Receiving Filter for Sunde's FM System

Our problem is to find the receiving filter characteristic which minimizes the probability of error in Sunde's FM system when the average transmitted signal power and the spectral density of the noise on the line are specified. In terms of Fig. 13 the transmittance function for the filter is $Y_r(\omega)$ and the output of the filter is $V_r(t)$ as defined by (61), (62), (71), and (73), namely

$$V_r(t) = A \sin \omega_d t \sin (\omega_c t + \theta_r) - A s_1(t) \cos (\omega_c t + \theta_r) \tag{108}$$

$$s_1(t) = \sum_{n=-\infty}^{\infty} (-)^n b_n g_1(t - nT) \tag{109}$$

$$G_1(\pm \omega_d - \lambda) + G_1(\pm \omega_d + \lambda) = 2G_1(\omega_d) = T \qquad 0 < \lambda < \omega_d \tag{110}$$

$$G_1(\omega) = \int_{-\infty}^{\infty} g_1(t) \cos \omega t \, dt. \tag{111}$$

The input to the filter is the sum of the signal wave $V(t)$ plus the Gaussian noise wave $v_0(t)$. The wave $V(t)$ is defined as that function of time which when operated on by $Y_r(\omega)$ produces $V_r(t)$. The noise wave $v(t)$ at the input to the frequency detector has a spectral density equal to $| Y_r(\omega) |^2$ times that of $v_0(t)$.

We shall simplify our treatment by assuming a random sequence of data in which the two binary symbols are selected with equal probability. The probability is then equal to 0.5 that any particular b_n has the value $+1$ and also 0.5 that the value is -1. We regard $V_r(t)$ as a member of an ensemble of random functions with a distribution in the infinite number of independent random parameters b_n. The randomness appears entirely in the function $s_1(t)$. We can calculate the ensemble average of $s_1(t)$ at fixed t by adding the individual averages of the terms in the infinite series defining $s_1(t)$. When we do this we find that the only random variable in each term is b_n, which assumes the values ± 1 with equal likelihood and therefore has the average value zero. Hence the ensemble average of $s_1(t)$, which we shall designate by $\langle s_1(t) \rangle$, is zero for any fixed value of t. It follows that $s_1(t)$ can contain no periodic components, for the presence of any such components would give a nonzero average at some values of t. Therefore, the spectral density function of the second term in $V_r(t)$ must be a continuous function of frequency.

To calculate the average square of $s_1(t)$ over the ensemble, we note that $s_1(t)$ is the sum of an infinite number of independent random variables of form

$$z_n = (-)^n b_n g_1(t - nT). \tag{112}$$

The average value of each z_n is zero and the variance, or mean square minus the square of the mean, is equal to the square of $g_1(t - nT)$.

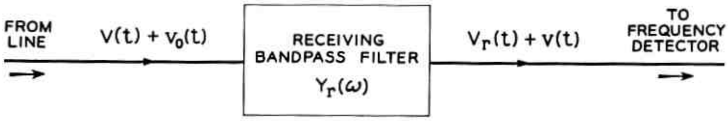


Fig. 13 — Function of receiving filter in Sunde's system.

Since the variance of the sum of independent variables is equal to the sum of the variances of the individual variables, we can write

$$\langle s_1^2(t) \rangle = \sum_{n=-\infty}^{\infty} g_1^2(t - nT). \tag{113}$$

The average in (113) is an ensemble average at fixed t. We can show that this average is periodic in t with period T by noting that

$$\begin{aligned} \langle s_1^2(t + T) \rangle &= \sum_{n=-\infty}^{\infty} g_1^2(t + T - nT) \\ &= \sum_{m=-\infty}^{\infty} g_1^2(t - mT) \\ &= \langle s_1^2(t) \rangle. \end{aligned} \tag{114}$$

Therefore the average over t can be computed by averaging over a single period from $t = 0$ to $t = T$. Hence the average over time which we shall designate by av is

$$\begin{aligned} \text{av } s_1^2(t) &= \frac{1}{T} \int_0^T \langle s_1^2(t) \rangle \, dt = \frac{1}{T} \sum_{n=-\infty}^{\infty} \int_0^T g_1^2(t - nT) \, dt \\ &= \frac{1}{T} \sum_{n=-\infty}^{\infty} \int_{-nT}^{-(n-1)T} g_1^2(\lambda) \, d\lambda = \frac{1}{T} \int_{-\infty}^{\infty} g_1^2(\lambda) \, d\lambda. \end{aligned} \tag{115}$$

By application of Parseval's theorem

$$\text{av } s_1^2(t) = \frac{1}{2\pi T} \int_{-\infty}^{\infty} G_1^2(\omega) \, d\omega. \tag{116}$$

From (116) we deduce that the spectral density of $s_1(t)$ is given by

$$w_1(\omega) = \frac{G_1^2(\omega)}{2\pi T} = \frac{\omega_d G_1^2(\omega)}{2\pi^2}. \tag{117}$$

The spectral density of $V_r(t)$ can now be easily calculated. The first term can be expressed as the sum of sine waves of amplitude $A/2$ and frequencies $\omega_c + \omega_d$ and $\omega_c - \omega_d$. The first term therefore contributes line spectra with mean square $A^2/8$ at the marking and spacing frequencies. The average square of the second term can be written

$$\text{av } [A^2 s_1^2(t) \cos^2 (\omega_c t + \theta_r)] = \frac{A^2}{2} \text{av } s_1^2(t). \tag{118}$$

The spectral components comprising $s_1(t) \cos (\omega_c t + \theta_r)$ are those of $s_1(t)$ shifted from their original positions to appear as sidebands around the frequencies $\pm\omega_c$. Hence $w_r(\omega)$, the spectral density of $V_r(t)$ with all power assigned to positive frequencies, is given by

$$w_r(\omega) = \frac{A^2}{8}\,\delta(\omega - \omega_c + \omega_d) + \frac{A^2}{8}\,\delta(\omega - \omega_c - \omega_d) + \frac{\omega_d A^2 G_1^2(\omega - \omega_c)}{4\pi^2} \qquad \omega \geqq 0. \tag{119}$$

It is convenient to let $\omega - \omega_c = u$ and write for the transmittance of the filter

$$U(u) = Y_r(\omega - \omega_c). \tag{120}$$

We shall also designate the spectral density of $V(t)$ as $w(u)$. Since the linear operator $U(u)$ can be applied individually to the components which make up (119) we must have

$$w(u) = \frac{A^2\delta(u + \omega_d)}{8\,|\,U(-\omega_d)\,|^2} + \frac{A^2\delta(u - \omega_d)}{8\,|\,U(\omega_d)\,|^2} + \frac{\omega_d A^2 G_1^2(u)}{4\pi^2\,|\,U(u)\,|^2}. \tag{121}$$

The average power on the line is proportional to W_0, the average square of $V(t)$, which is given by

$$W_0 = \int_{-2\omega_d}^{2\omega_d} w(u)\,du = \frac{A^2}{8\,|\,U(-\omega_d)\,|^2} + \frac{A^2}{8\,|\,U(\omega_d)\,|^2} + \frac{\omega_d A^2}{4\pi^2}\int_{-2\omega_d}^{2\omega_d} \frac{G_1^2(u)}{|\,U(u)\,|^2}\,du. \tag{122}$$

We make the reasonable assumption that $|\,U(u)\,|$ is an even function of u. Combined with the further assumption that the spectral density of the noise on the line is symmetrical about ω_c, this furnishes a convenient assurance of a symmetrical spectral density for the noise in the output of the receiving filter. Since $G_1(u)$ is also an even function of u, we can write (122) in the equivalent form

$$W_0 = \frac{A^2}{4X(\omega_d)} + \frac{\omega_d A^2}{2\pi^2}\int_0^{2\omega_d} \frac{G_1^2(u)}{X(u)}\,du \tag{123}$$

where

$$X(u) = |\,U(u)\,|^2. \tag{124}$$

The function $X(u)$ is to be chosen to minimize the probability of error under the constraint that W_0 is held constant. In calculating the optimum function, the signal power represented by the steady-state components can be ignored, since this power could be reduced to an arbitrarily small value by the use of narrow-band suppression techniques. The constraint on the signal power is therefore that the integral in (123) is to be held constant.

Let $N(u)$ represent the spectral density of the Gaussian noise wave $v_0(t)$ on the line. Then the spectral density of $v(t)$, the noise in the output of the receiving filter, is $X(u)N(u)$. In terms of the spectral density $w_v(\omega)$ previously defined for $v(t)$ with values symmetrically distributed between positive and negative frequencies, we have

$$X(u)N(u) = 2w_v(u + \omega_c). \tag{125}$$

The values of σ_0 and σ_1 necessary to complete the calculation of the probability of error by (34) can now be found by substituting (125) in (17) and (18) giving the results

$$\sigma_0^2 = 2 \int_0^{2\omega_d} X(u)N(u)\, du \tag{126}$$

$$\sigma_1^2 = 2 \int_0^{2\omega_d} u^2 X(u)N(u)\, du. \tag{127}$$

If we substitute (126) and (127) into the general expression for error probability, (34), and attempt to formulate a variational problem, the expressions become unmanageable. Instead, we concentrate attention on the lower bound for error probability obtained by setting $\dot{R} = 0$, (36), in which it is evident that to make the error probability as small as possible both σ_0 and σ_1 should be made as small as possible. As shown by (126) and (127), σ_0 and σ_1 are not independent. The effect of the dependence can be taken into account by performing the minimization problem in two steps. First we minimize σ_0 with both σ_1 and W_0 held constant. After this solution is obtained, we find by trial the value of σ_1 which yields the lowest minimum probability of error.

Omitting inconsequential multiplying factors, we set the variational problem as

$$\delta\left[\int_0^{2\omega_d} X(u)N(u)\, du + \lambda \int_0^{2\omega_d} u^2 X(u)N(u)\, du + \mu \int_0^{2\omega_d} \frac{|G_1(u)|^2}{X(u)}\, du\right] = 0 \tag{128}$$

where λ and μ are Lagrange multipliers and the function under variation is $X(u)$. The solution is

$$X(u) = \frac{\mu\, |G_1(u)|}{(1 + \lambda u^2)^{\frac{1}{2}}\, N^{\frac{1}{2}}(u)}. \tag{129}$$

It is straightforward to verify that this stationary value of $X(u)$ actually gives a minimum value of σ_0 and hence minimum probability of error for fixed values of σ_1 and W_0.

Substituting our partially optimized solution in (123), (126), and (127), we obtain

$$\sigma_0^{\,2} = 2\mu \int_0^{2\omega_d} \frac{|\,G_1(u)\,|\,N^{\frac{1}{2}}(u)}{(1+\lambda u^2)^{\frac{1}{2}}}\,du \tag{130}$$

$$\sigma_1^{\,2} = 2\mu \int_0^{2\omega_d} \frac{u^2\,|\,G_1(u)\,|\,N^{\frac{1}{2}}(u)}{(1+\lambda u^2)^{\frac{1}{2}}}\,du. \tag{131}$$

$$W_o - W_s = \frac{\omega_d A^2}{2\pi^2\mu} \int_0^{2\omega_d} |\,G_1(u)\,|\,N^{\frac{1}{2}}(u)(1+\lambda u^2)^{\frac{1}{2}}\,du \tag{132}$$

$$W_s = \frac{A^2 N^{\frac{1}{2}}(\omega_d)(1+\lambda\omega_d^{\,2})^{\frac{1}{2}}}{4\mu\,|\,G_1(\omega_d)\,|} \tag{133}$$

$$\frac{1}{\rho^2} = \frac{2\sigma_0^{\,2}}{A^2} = \frac{2\omega_d I_1 I_2}{\pi^2(W_0 - W_s)} \tag{134}$$

$$\frac{1}{a^2\rho^2} = \frac{2\sigma_1^{\,2}}{A^2\omega_d^{\,2}} = \frac{2I_2I_3}{\pi^2\omega_d(W_0 - W_s)} \tag{135}$$

where

$$I_1 = \int_0^{2\omega_d} \frac{|\,G_1(u)\,|\,N^{\frac{1}{2}}(u)}{(1+\lambda u^2)^{\frac{1}{2}}}\,du \tag{136}$$

$$I_2 = \int_0^{2\omega_d} |\,G_1(u)\,|\,N^{\frac{1}{2}}(u)(1+\lambda u^2)^{\frac{1}{2}}\,du \tag{137}$$

and

$$I_3 = \int_0^{2\omega_d} \frac{u^2\,|\,G_1(u)\,|\,N^{\frac{1}{2}}(u)}{(1+\lambda u^2)^{\frac{1}{2}}}\,du. \tag{138}$$

These equations furnish a straightforward procedure for calculating the optimum filter characteristic. Each assumed value of λ determines a pair of values ρ and $a\rho$ from which the corresponding upper and lower bounds for the error probability can be evaluated by computer techniques. By successive trials the best value of λ can be approximated to any desired degree and substituted in (129) to obtain the best filter selectivity function. In actual examples tried, this procedure could be shortened because the error probability turned out to be very much more sensitive to the value of ρ than to the value of $a\rho$. If this were known beforehand, we would place no constraint on σ_1 in the minimization of σ_0. This is equivalent to setting $\lambda = 0$, leading to the simpler formulas

$$\frac{1}{\rho^2} = \frac{2\omega_d}{\pi^2(W_0 - W_s)}\left[\int_0^{2\omega_d} |G_1(u)|\, N^{\frac{1}{2}}(u)\, du\right]^2 \tag{139}$$

$$\frac{1}{a^2\rho^2} = \frac{2}{\pi^2\omega_d(W_0 - W_s)}\int_0^{2\omega_d} u^2\, |G_1(u)|\, N^{\frac{1}{2}}(u)\, du \cdot \int_0^{2\omega_d} |G_1(u)|\, N^{\frac{1}{2}}(u)\, du. \tag{140}$$

By applying Schwarz' inequality to the products of integrals in (134) and (135), we verify that the case of $\lambda = 0$ gives the maximum value of ρ, but that the maximum value of $a\rho$ occurs when $\lambda = \infty$. It seems therefore that an intermediate nonzero value of λ would be best, but in the cases computed the improvement obtainable in this way turned out to be negligibly small.

As an example, consider the raised cosine signal spectrum in which $G_1(u)$ is given by (77). We also assume a white noise spectrum in which $N(u)$ is equal to a constant N_0. It is convenient to introduce as a signal-to-noise ratio the quantity M defined by

$$M = \frac{W_0}{N_0\omega_0} = \frac{W_0}{2N_0\omega_d}. \tag{141}$$

This is the ratio of average transmitted signal power to the average noise power in a band of frequencies of width equal to the bit rate. Computer results show that the case of $\lambda = 0$ is practically indistinguishable from the optimum λ. Hence we set $\lambda = 0$ and calculate for the optimum filter

$$U(u) = X^{\frac{1}{2}}(u) = \left(\frac{\mu\pi}{\omega_d N_0}\right)^{\frac{1}{2}} \cos\frac{\pi u}{4\omega_d} \qquad |u| < 2\omega_d. \tag{142}$$

This is the same cosine filter characteristic found by Sunde to be optimum for binary AM with synchronous detection. From (132) and (133) we find that with $\lambda = 0$

$$W_0 - W_s = \frac{A^2\omega_d N_0^{\frac{1}{2}}}{2\pi\mu} = W_s. \tag{143}$$

Hence

$$W_s = W_0/2 \quad \text{and} \quad W_0 - W_s = W_0/2. \tag{144}$$

From (139), (140), and (141) we then calculate

$$\rho^2 = \frac{W_0}{4\omega_d N_0} = \frac{M}{2} \tag{145}$$

$$\begin{aligned} a^2\rho^2 &= \frac{3\pi^2 W_0}{16\omega_d N_0(\pi^2 - 6)} \\ &= \frac{3\pi^2 M}{8(\pi^2 - 6)} = 0.956M. \end{aligned} \tag{146}$$

If the steady-state components were suppressed, we would set $W_s = 0$ and would then obtain $\rho^2 = M$, $a^2\rho^2 = 1.913M$. This would correspond to a 3-db shift in the direction of lower signal-to-noise ratio when the error probability curves are plotted against $10 \log_{10} M$.

The curves of Fig. 6, showing the upper and lower bounds for error probability when Sunde's FM system is optimized, were calculated by S. Habib on the digital computer. The case of a nonoptimum receiving filter is illustrated by the corresponding curves for a rectangular band defined by

$$X(u) = X_0 \qquad |u| < 2\omega_d . \tag{147}$$

For this case we compute from (126) and (127)

$$\sigma_0^2 = 2\int_0^{2\omega_d} X_0 N_0 \, du = 4\omega_d X_0 N_0 \tag{148}$$

$$\sigma_1^2 = 2\int_0^{2\omega_d} u^2 X_0 N_0 \, du = \frac{16\omega_d^3 X_0 N_0}{3}. \tag{149}$$

From (123)

$$W_0 = \frac{A^2}{4X_0} + \frac{A^2}{8\omega_d X_0}\int_0^{2\omega_d}\left(1 + \cos\frac{\pi u}{2\omega_d}\right)^2 du = \frac{5A^2}{8X_0}. \tag{150}$$

We then calculate

$$\rho^2 = 2M/5 \qquad a^2\rho^2 = 3M/10. \tag{151}$$

If the steady-state components are suppressed, the average transmitted power could be reduced to $(\frac{5}{8} - \frac{1}{4})/(\frac{5}{8}) = \frac{3}{5}$ of the previously determined value, which is a saving of 2.2 db.

REFERENCES

1. Bennett, W. R., *Electrical Noise*, McGraw-Hill Book Co., Inc., New York, 1960, pp. 234–238.
2. Rice, S. O., Statistical Properties of a Sine Wave Plus Random Noise, B.S.T.J., **27**, Jan., 1948, pp. 109–157.
3. Lawton, J. G., Theoretical Error Rates of "Differentially Coherent" Binary

and "Kineplex" Data Transmission Systems, Proc. I.R.E., **47,** Feb., 1959, pp. 333–334.
4. Sunde, E. D., Ideal Pulses Transmitted by AM and FM, B.S.T.J., **38,** Nov., 1959, pp. 1357–1426.
5. Nyquist, H., Certain Topics in Telegraph Transmission Theory, Trans. A.I.E.E., **47,** April, 1928, pp. 617–644.
6. Knopp, K., *Theory and Applications of Infinite Series*, Blackie and Son, Ltd., London, 1928, p. 269.
7. Meyerhoff, A. A. and Mazer, W. M., Optimum Binary FM Reception Using Discriminator Detection and I-F Shaping, RCA Review, **22,** Dec., 1961, pp. 698–728.
8. Smith, E. F., Attainable Error Probabilities in Demodulation of Random Binary PCM/FM Waveforms, I.R.E. Trans. on Space Elec. and Telemetry, **SET-8,** Dec., 1962, pp. 290–297.
9. Montgomery, G. F., A Comparison of Amplitude and Angle Modulation for Narrow-Band Communication of Binary-Coded Messages in Fluctuation Noise, Proc. I.R.E., **42,** Feb., 1954, pp. 447–454.

A Functional Analysis Relating Delay Variation and Intersymbol Interference in Data Transmission

By R. W. LUCKY

(Manuscript received May 17, 1963)

A relationship is derived between the delay characteristic in a data transmission system and the distortion in the form of intersymbol interference created by the delay variation. The relationship is valid for small delay and involves a sequence of linear functionals, each of which has a particular significance. In addition to applications in the analysis of specific systems, problems of a more general nature may be studied using this approach. By various manipulations on the sequence of functionals, bounds on distortion in terms of rms and peak-to-peak delay are derived. On examining the problem of delay equalization, a set of virtually distortion-free delay functions is derived and related to minimum-effort and compromise equalization. Both low-pass and bandpass systems are discussed in turn, with the same general method of analysis applying to each.

I. INTRODUCTION

In this paper we will analyze and discuss one aspect of the problems associated with the transmission of digital data through an unknown linear network. The particular aspect with which we will be concerned is the effect of delay distortion on the fidelity of the transmission. Delay distortion arises generally from nonlinearity in the phase shift with frequency of the system transmission characteristic. This nonlinearity causes different frequencies of the input waveform to arrive at the receiver at different times, thereby distorting the input waveform.

The problem of delay distortion is particularly acute in the voice telephone network. Since speech is relatively insensitive to phase, the switched telephone network has not been equalized for phase shift as well as it has been for attenuation. However, the need has now arisen to make use of this network for transmission of digital data at high speeds. The digital data receiver takes the waveforms it receives quite literally and becomes hopelessly confused by delay distortion if we try to send at too high a rate. For instance, in a voice-band channel of 3 kc Nyquist's famous result tells us that it might be possible to send 6000 independent signals (representing data symbols) per second. However, the usual rate is around 1000 symbols per second (2000 bits for quaternary systems), and higher speeds are impossible because of transmission distortion.

In subsequent sections of this paper we will be concerned with quantitative effects that delay distortion has on data transmission. This does not mean that we will analyze any particular system operating in the presence of a particularly shaped delay variation. This has been done previously by a number of authors, but notably E. Sunde.[1,2,3,4] Rather, we will be more concerned with the gross features of the relationship between delay and system performance. For example, if a particular level of performance is required, what standards may be set on delay such that this minimum performance level will be guaranteed? What shapes of delay are particularly bad or good? How well does differential delay (the difference between the maximum and minimum values of delay across the band) define performance? If a channel is equalized within a certain tolerance of delay, what level of performance can be achieved?

Inasmuch as the telephone network consists of an ensemble of transmission characteristics from which the channel is randomly chosen, these questions would seem to be more meaningful than specific performance figures for particular channels. Consider the problem of comparing data systems for transmission over the voice network. It is clear that this susceptibility to delay distortion is an important factor in such a comparison. The performance of system A will be a random variable defined over the set of possible connections we could dial, likewise system B. The analytical portion of comparing the two systems would be best shown as the probability distributions of performance for the two systems. One system could only be said to be statistically "better" than the other. For example, its average performance over the ensemble of channels might be greater.

Data are now becoming available on the delay characteristics of the voice network (e.g., Alexander, Gryb and Nast).[5] Using these data, it might be possible to analyze systems for which the delay characteristic is chosen randomly from this network, or it might be possible to synthesize systems which operate well (with high probability) over this network. The latter could be accomplished by taking advantage of certain features common to the majority of delay characteristics. One way of doing this is to use a "compromise" equalizer. Quite a question exists as to how to design such an equalizer and how much would be gained by its use. Other possibilities include the use of a bank or set of equalizers from which a best choice may be made for each call or the ultimate use of automatic equalization.

In the remainder of this paper we will show an approximate method whereby the effects of delay distortion may be easily considered in answering such questions as we have asked here. The first section will be devoted to explaining what performance criterion will be used whereby a particular channel may be judged as to goodness for data transmission. In subsequent sections the criterion suggested will be manipulated to show clearly its dependence on delay for both lowpass and bandpass systems. A summary of results obtained is presented following the section on criteria.

II. A PERFORMANCE CRITERION

What we seek in this section is a criterion which may be applied to a transmission channel to determine how good the channel is for data transmission. Obviously, such a criterion should depend upon what system we intend to use over the channel as well as upon the noise environment and input data statistics and the over-all system performance criterion (such as probability of error) that is used. A channel can't be said to be "good" or "bad" irrespective of how we intend to use it. Therefore, the only exact thing which can be done is to treat each possible system separately and derive the relationship between delay and performance separately for each.

For example, Sunde[1,2,3,4] has analyzed several common systems such as AM, PM, and FM in a noiseless environment, using the deviation of the detector output voltage from its undistorted values as a measure of performance. When the details are carried out, the system performance is given as a function of sample values of the impulse response of the over-all system. We call this impulse response $h(t)$

$$h(t) = \frac{1}{\pi} \int_0^\infty A(\omega) \cos [\omega t - \beta(\omega)] \, d\omega \tag{1}$$

where $A(\omega)$ includes signal shaping at the receiver and transmitter as well as the attenuation characteristic of the channel and $\beta(\omega)$ is the channel phase characteristic. Unfortunately, the relationship between the samples of $h(t)$ and system performance is a complicated one and it is unclear as to how the shape of the delay, $\beta'(\omega)$, affects the performance.

What we shall do is to take one specific system, amplitude modulation, and show that its performance is monotonically related to a quantity

$$D = \sum_{\substack{n=-\infty \\ n\neq 0}}^{+\infty} | h(t_0 + nT) | = \sum_{n}{}' | h_n | \tag{2}$$

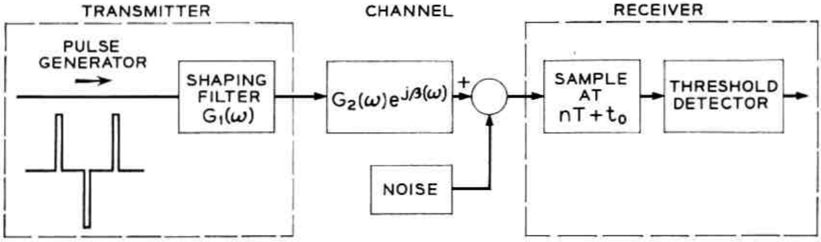
which we shall call distortion.

Now, in later sections of this paper we will see how the shape of delay affects this distortion measure. Therefore, the results given in these sections may be interpreted as *performance for AM systems*. However, we shall argue that this distortion measure may be quite plausible even though the system being used is not AM. Indeed, many common systems may have their performance related monotonically to D. This is to say that a channel which is bad for AM is probably bad for PM too. This may be considered similar to the "What's good for General Bullmoose is good for the U.S.A." proposition, but the thought may also occur that, considering the unknown nature of the noise and input data statistics, the criterion (2) may be just as good a starting point as some arbitrary definition of environment and performance measure. At any rate, we do not intend to dwell on the difficult problem of criteria here. The criterion D is monotonically related to performance for linear systems and for those systems which can be approximated as linear.

2.1 *The Performance of a Simple Baseband System*

A mathematical model of this system is shown in Fig. 1. The transmitted signal consists of amplitude-modulated waveforms whose shape is $g_1(t)$

$$s(t) = \sum_{n=-\infty}^{+\infty} a_n g_1(t - nT) \tag{3}$$



$x(t) = \sum_{n=-\infty}^{+\infty} a_n \delta(t-nT)$ $\qquad$ $s(t) = \sum_{n=-\infty}^{+\infty} a_n g_1(t-nT)$

Fig. 1 — A baseband amplitude-modulated system.

and is generated in our mathematical model (not in practice) by a train of area-modulated delta functions pulsing a filter whose impulse response is $g_1(t)$. The transfer function of the channel is $G_2(\omega)e^{j\beta(\omega)}$, so that the over-all transfer function for the impulse is $A(\omega)e^{j\beta(\omega)}$, with

$$A(\omega) = G_1(\omega)G_2(\omega).$$

In the noiseless case, the received signal $y(t)$ is

$$y(t) = \int_0^{+\infty} h(\tau) \sum_{n=-\infty}^{+\infty} a_n\delta(t - nT - \tau)\, d\tau \tag{4}$$

where $h(t)$ is the over-all system impulse response

$$h(t) = \frac{1}{\pi}\int_0^{+\infty} A(\omega) \cos [\omega t - \beta(\omega)]\, d\omega. \tag{5}$$

Equation (4) may be written

$$y(t) = \sum_{n=-\infty}^{+\infty} a_n h(t - nT). \tag{6}$$

We sample this received signal at regular intervals of T seconds starting at time t_0. At time t_0 we expect the amplitude a_0, but we actually get

$$y(t_0) = \sum_{n=-\infty}^{+\infty} a_n h(t_0 - nT) = \sum_{n=-\infty}^{+\infty} a_{-n} h(t_0 + nT) \tag{7}$$

which is a function of the history of the data sequence $\{a_n\}$. For some sequences $y(t_0)$ will be more likely to be detected wrongly than for others. The error due to intersymbol interference is

$$E = a_0 - y(t_0) = a_0[1 - h(t_0)] - \sum_{\substack{n=-\infty \\ n\neq 0}}^{+\infty} a_{-n} h(t_0 + nT). \tag{8}$$

Now, we are interested in the maximum value this error (which is frequently termed the eye opening) can assume. Assuming the maximum positive and negative values of the coefficients a_n are $\hat{a}$ and $-\hat{a}$ respectively, this maximum error is easily written as

$$E_{\max} = a_0[1 - h(t_0)] - \hat{a} \sum_n{}' \mid h(t_0 + nT) \mid. \tag{9}$$

The first term represents an amplification or attenuation of the signal by the channel, while the second term represents the worst possible effect of intersymbol interference. The prime in the summation sign means deletion of the $n = 0$ term

$$\sum_n{}' = \sum_{\substack{n=-\infty\\ n\neq 0}}^{+\infty}. \tag{10}$$

With the attenuation $h(t_0)$ close to unity* or normalized, we see that the distortion is proportional to

$$D = \sum{}' \mid h(t_0 + nT) \mid. \tag{11}$$

If the noise were additive with a unimodal distribution, the maximum probability of error over all sequences would be a monotonic function of D. For example, if the levels a_n are spaced equally between $+\hat{a}$ and $-\hat{a}$ and there are N levels, the distance between levels is

$$\mid a_n - a_{n-1} \mid = \frac{2\hat{a}}{N-1} \tag{12}$$

and the probability of making an error with Gaussian noise of mean zero and variance σ^2 is

$$\begin{aligned}\max_{\{a_n\}} \text{prob of error} = \text{prob}\Bigg(\mid \text{noise} \mid > \frac{\hat{a}}{N-1} \\ - \hat{a} \sum{}' \mid h(t_0 - nT) \mid \Bigg)\end{aligned} \tag{13}$$

$$P(e) = 1 - \frac{1}{\sigma}\,\text{erf}\left[\frac{\hat{a}}{\sqrt{2}\sigma}\left(\frac{1}{N-1} - D\right)\right] \tag{14}$$

$$\text{erf}\left(\frac{x}{\sqrt{2}}\right) = \frac{1}{\sqrt{2\pi}} \int_{-x}^{+x} e^{-t^2/2}\, dt. \tag{15}$$

The distortion D is a function of the initial delay, t_0. This sampling time is optimally chosen such that the criterion D is minimized. This best time is a functional of the delay $\beta'(\omega)$ through its influence on the impulse response. Unfortunately, it is extremely difficult to optimize t_0 even for a specific impulse response. Therefore, we shall arbitrarily choose t_0 at the peak of the impulse response. This is a very good approximation to the best possible sampling instant.

2.2 *Discussion of the Criterion D*

The distortion criterion D has been written as the sum of the absolute values of the system impulse response sampled at the symbol repetition

* This is a second-order effect for the small-delay case in which we will be interested.

rate. The zero sample of impulse response, $h_0 = h(t_0)$, is taken at the peak of the response and is deleted from the summation. We have shown that for an amplitude-modulated system this criterion is proportional to the maximum deviation in the absence of noise of the detector output voltage. The maximum is taken over all possible input symbol sequences. This performance measure is frequently termed the "eye opening," from the resemblance to an eye when the output voltage is displayed on an oscilloscope while random patterns of input symbols are transmitted.

That the criterion D is reasonable for most linear systems may be roughly shown. We recognize that $h(t)$ represents the system memory, or response from past signals. At time t_0 we are looking for symbol s_0, but unfortunately the system remembers remnants of past and future* symbols at this time. The symbols are spaced T seconds apart so that the nth past symbol is "remembered" with relative amplitude

$$| h(t_0 + nT) |.$$

It makes sense that, the larger the sum of these relative memories, the worse will be the intersymbol interference. We will show in the course of our later work how well the performance of a nonlinear system employing phase comparison detection is predicted by use of the distortion criterion D.

2.3 *A General Di:tortion Criterion*†

More generally, we would wish to send the signals chosen from a set $s_i(t)$, $i = 1, 2, \cdots, N$. Each symbol is chosen according to some probabilistic rule from this set in time sequence to form the signal

$$x(t) = \sum_{n=-\infty}^{+\infty} s_n(t - nT). \tag{16}$$

The signals $s_i(t)$ are sometimes viewed as vectors in a Hilbert space or in some finite-dimensional subspace. The effect of sending the sequence $x(t)$ through a linear network is to cause the received signal during a T-second interval to be a linear combination of the desired signal vector and all other unwanted signal vectors rotated and attenuated by the channel. For a given channel, if we consider the position of the resultant vector for all possible infinite sequences of symbols, we define regions of uncertainty in signal space surrounding the unperturbed vectors s_i.

For purposes of combating noise we are concerned with the distances

* The memory of future symbols is possible because of the time delay t_0 between input and output.

† This section is not essential to the understanding of subsequent material.

in signal space between the transmitted vectors, that is, with the numbers $\| s_i - s_j \|, i \neq j$. The greater this set of numbers is, the greater the potential noise immunity of the system. The effect of the channel is to make these distances a function of the symbol sequence. So we can say something about what the channel has done to the noise immunity by specifying the minimum protective distance $\| s_i - s_j \|, i \neq j$, with and without the channel. Thus, we might define a measure of distortion as

$$D_0 = 1 - \min_{\substack{i,j \\ i \neq j}} \left[\underset{y_i \subset R_i}{\mathrm{m{\cdot}n}} \frac{\| y_i - s_j \|}{\| s_i - s_j \|} \right]. \tag{17}$$

R_i = region of uncertainty due to intersymbol interference surrounding the symbol s_i.

This criterion is illustrated in Fig. 2. The way the criterion was formulated did not take into effect the receiver characteristics, but rather evaluated the "loss of detectability" to an ideal maximum likelihood receiver owing to intersymbol interference. Notice also that this criterion

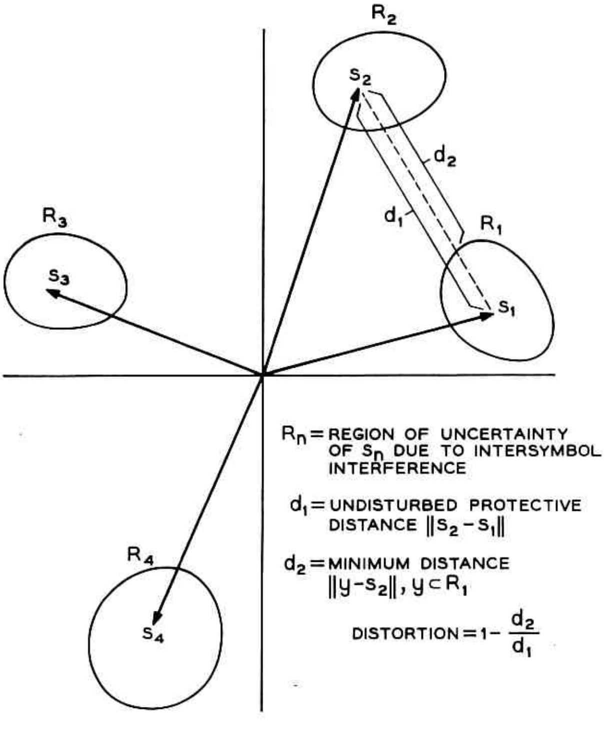

R₂
S₂
d₂
d₁
R₁
S₁
R₃
S₃
R₄
S₄
Rₙ = REGION OF UNCERTAINTY OF Sₙ DUE TO INTERSYMBOL INTERFERENCE
d₁ = UNDISTURBED PROTECTIVE DISTANCE ‖S₂ − S₁‖
d₂ = MINIMUM DISTANCE ‖y − S₂‖, y ⊂ R₁
DISTORTION = 1 − d₂/d₁

Fig. 2 — A general distortion criterion.

is dependent upon the system to the extent that it is a function of the set of possible signals, $s_i(t)$. As was previously stated, it is impossible to eliminate system dependence from a criterion and maintain usefulness for all conditions.

When this measure is applied to the AM system of Fig. 1, the result is the criterion D previously expressed in (2).

III. SUMMARY OF RESULTS

The problem now is to explore the functional dependence of the distortion, D, upon the delay characteristic, $\beta'(\omega)$. As we have previously shown, the distortion D is a measure related to the eye opening for most linear systems. Specifically, for an N-level AM system we have

$$I_{N\text{-level AM}} = 1 - (N - 1)D. \tag{18}$$

However, by ignoring second-order effects the criterion may be applied to some nonlinear systems. In examples used subsequently in the text, results are obtained for a four-phase data system using phase comparison detection. These results are in close agreement with published data on this system. The eye opening for this system is approximately

$$I_{4\text{-phase}} \approx 1 - D. \tag{19}$$

Similar expressions may be obtained for other systems.

3.1 *The Fundamental Equation*

In Section 4.1 an approximation of small delay is made. The validity of this approximation is explored in a later section, where it is shown to hold for all delay such that the peak delay is limited to 1.1 pulse intervals. For most delay curves the range is wider than this figure, however, and accuracy is generally maintained when $D \leqq 0.6$.

The fundamental equation obtained with the aid of this approximation relates the distortion to the delay variation through a sequence of linear functionals

$$D = \frac{2}{\pi} \sum_{n=1}^{\infty} |\,(\beta', f_n)\,| \tag{20}$$

where

$$(\beta', f_n) = \int_0^w \beta'(\omega) f_n(\omega)\, d\omega. \tag{21}$$

Each linear functional yields the intersymbol interference from a particular range of symbols. For example, the adjacent symbol interference is

$$| h_1 | + | h_{-1} | = \frac{2}{\pi} | (\beta', f_1) |. \tag{22}$$

For real delay curves only the first few terms of (20) are usually significant.

The linear functionals (β', f_n) are defined by a sequence of functions $\{f_n\}$, independent of delay, obtained from the amplitude shaping of the system, $A(\omega)$, by the following operation

$$f_n(\omega) = \int_\omega^w [a_n x - \sin nxT] \, A(x) \, dx \tag{23}$$

$$a_n = \frac{\int_0^w \omega A(\omega) \sin n\omega T \, d\omega}{\int_0^w \omega^2 A(\omega) \, d\omega}. \tag{24}$$

3.2 *Application*

Examples are given of the use of (20) in the analysis of system performance when raised cosine amplitude shaping is employed. By reformulating the equation to

$$D = \frac{2}{\pi} \max_{\{\epsilon_n\}} \left(\beta', \sum_{n=1}^{\infty} \epsilon_n f_n \right) \tag{25}$$

$$\epsilon_n = \pm 1$$

bounds on distortion in terms of delay may be derived. We find

$$D \leqq 1.15 \times (\text{rms delay}) \tag{26}$$

$$D \leqq 0.412 \times (\text{peak-to-peak delay}) \tag{27}$$

with delay normalized so that the bandwidth $w = \pi$. The delay curves which achieve equality in the bounds (26) and (27) are illustrated. Other bounds are considered.

In computer simulations, experimental testing, and analysis it is frequently necessary to consider only finite-length input sequences resulting in an effective truncation of the system memory. The possible error in such results is examined and bounded.

The effect of changing the input symbol rate upon distortion is analyzed for a particular example where a binary system is compared with a quaternary system operating at half speed (thus having the same information rate). Depending upon the particular delay function, either system may perform better than the other. However, it is shown that the quaternary system is ultimately the more sensitive to delay variation.

Problems connected with delay equalization are approached with the aid of (20). In the equalization of a specific delay to achieve zero distortion, it is necessary and sufficient that the resultant delay variation be orthogonal to all f_n. For raised cosine shaping there are an infinite number of nonconstant delay curves which have this property. An orthonormal basis for this space of distortionless delay is derived, and examples of projections yielding minimum effort equalization are given. [All curves so derived have, of course, zero distortion only to the order of approximation involved in (20)]. Optimum compromise equalization to match an ensemble of delay variations is also considered.

3.3 *Bandpass Modifications*

For bandpass systems the criterion D is reformulated using the sum of samples of the envelope of the impulse response

$$D = \sum{}' P(t_0 + nT) \tag{28}$$

$$h(t) = P(t) \cos [\omega_c t - \psi(t)] \tag{29}$$

$$\omega_c = \text{carrier or reference frequency.}$$

A fundamental equation for bandpass systems analogous to (20) is derived involving quadrature components

$$D = \sqrt{D_r^2 + D_q^2}. \tag{30}$$

The distortion component D_r results from even components (for symmetrical amplitude shaping) of delay variation, and the component D_q results from odd components of delay variation. Each may be treated as in the low-pass analysis by a sequence of linear functionals

$$D_r = \frac{2}{\pi} \sum_{n=1}^{\infty} | (\varphi', f_{rn}) | \tag{31}$$

$$D_q = \frac{2}{\pi} \sum_{n=1}^{\infty} | (\varphi', f_{qn}) | \tag{32}$$

where $\varphi'(\omega)$ is the bandpass delay and the sequences $\{f_{rn}\}$ of even functions and $\{f_{qn}\}$ of odd functions are derived from the amplitude shaping by operations similar to (23).

All results obtained for low-pass systems may also be obtained for bandpass systems. For example, it is shown that for raised cosine amplitude shaping

$$\begin{aligned} D &\leqq 1.337 \times \text{(rms delay)} \\ D &\leqq 0.467 \times \text{(peak-to-peak delay)}. \end{aligned} \tag{33}$$

Thus the bandpass system is slightly more sensitive to delay distortion than its baseband equivalent.

IV. THE RELATIONSHIP OF DISTORTION TO DELAY FOR LOW-PASS SYSTEMS

4.1 *Derivation of a Sequence of Linear Functionals Relating Distortion, D, to Delay, $\beta'(\omega)$*

$$D = \sum_n{}' \mid h(t_0 + nT) \mid = \sum_n{}' \epsilon_n h(t_0 + nT) \tag{34}$$

$$\epsilon_n = \begin{cases} +1 & h(t_0 + nT) \geqq 0 \\ -1 & h(t_0 + nT) < 0 \end{cases} \tag{35}$$

$$D = \sum_n{}' \epsilon_n \int_0^w A(\omega) \cos [\omega(t_0 + nT) - \beta(\omega)] \, d\omega \tag{36}$$

$$D = \frac{1}{\pi} \sum_n{}' [\epsilon_n C_n - \epsilon_n S_n] \tag{37}$$

$$C_n = \int_0^w A(\omega) \cos n\omega T \cos [\omega t_0 - \beta(\omega)] \, d\omega \tag{38}$$

$$S_n = \int_0^w A(\omega) \sin n\omega T \sin [\omega t_0 - \beta(\omega)] \, d\omega. \tag{39}$$

Equations (34) to (39) are self-explanatory reformulations of the criterion D. Equation (37) is summed over all n, $-\infty$ to $+\infty$, except $n = 0$. Because of the obvious symmetry properties of C_n and S_n, namely $C_n = C_{-n}$ and $S_n = -S_{-n}$, we can rewrite (37) as a sum over positive integers only

$$D = \frac{1}{\pi} \sum_{n=1}^{\infty} [C_n(\epsilon_n + \epsilon_{-n}) - S_n(\epsilon_n - \epsilon_{-n})]. \tag{40}$$

Since $\epsilon_n = \pm 1$, one of the pair $(\epsilon_n + \epsilon_{-n})$ and $(\epsilon_n - \epsilon_{-n})$ is zero and the other must be ± 2. Therefore the criterion becomes

$$D = \frac{2}{\pi} \sum_{n=1}^{\infty} \max \left(\mid C_n \mid , \mid S_n \mid \right). \tag{41}$$

What we are doing here is evaluating terms of the sum D of (34) two at a time. The integral C_n represents $[h(t_0 + nT) + h(t_0 - nT)]$, while the integral S_n represents $[h(t_0 + nT) - h(t_0 - nT)]$. Since what we want is $| h(t_0 + nT) | + | h(t_0 - nT) |$, this is equivalent to $| C_n |$ if these two terms are of the same sign and is $| S_n |$ if they are of opposite sign. We shall now argue that, for small delay, $| S_n | > | C_n |$ and (41) may be summed over the S_n terms alone.

Specifically, what we mean by small delay is that the sine and cosine of $[\omega t_0 - \beta(\omega)]$ may be approximated by the first terms of their expansions. Later we will investigate the conditions under which this approximation is valid. Making these approximations in (38) and (39) gives

$$C_n = \int_0^w A(\omega) \cos n\omega T \, d\omega = 0 \tag{42}$$

(since we must assume that $A(\omega)$ is such that transmission is perfect in the absence of delay, distortion; i.e., $h(nT) = 0, n \neq 0$).

$$S_n = \int_0^w [\omega t_0 - \beta(\omega)] A(\omega) \sin n\omega T \, d\omega. \tag{43}$$

Thus we see that, to this order of approximation, terms linear in

$[\omega t_0 - \beta(\omega)]$, $| S_n | > | C_n |$,* and

$$D = \frac{2}{\pi} \sum_{n=1}^{\infty} | S_n |. \tag{44}$$

Now, as we have previously explained, the initial delay t_0 is ideally chosen so as to minimize D for a given $h(t)$. Unfortunately this is impossible to do analytically. We recognize that for zero delay distortion $t_0 = 0$ and that the presence of delay increases t_0. As a good approximation to the ideal sampling time, we are using t_0 as the time of the peak value of the impulse response $h(t)$. An additional, and extremely important, consideration in this choice is that the approximation of $[\omega t_0 - \beta(\omega)]$ small has been made. To choose t_0 at the peak of the impulse response results in the smallest possible values for the function $[\omega t_0 - \beta(\omega)]$. We shall see this more clearly later on.

* The second term in the cosine expansion is

$$- \int_0^w \frac{1}{2} [\omega t_0 - \beta(\omega)]^2 A(\omega) \cos n\omega T \, d\omega.$$

Since $[\omega t_0 - \beta(\omega)]^2 < | \omega t_0 - \beta(\omega) |$ we would generally expect $| C_n |$ to be smaller than $| S_n |$. However, this is not necessarily true; e.g., $[\omega t_0 - \beta(\omega)]$ may be orthogonal to $\sin n\omega T$ on the $[0,w]$ interval.

We now solve for the time t_0 as a functional of $\beta(\omega)$ using the equation $h'(t_0) = 0$

$$h'(t_0) = 0 = \frac{-1}{\pi} \int_0^w \omega A(\omega) \sin [\omega t_0 - \beta(\omega)]\, d\omega \tag{45}$$

$$\int_0^w [\omega t_0 - \beta(\omega)]\omega A(\omega)\, d\omega \approx 0 \tag{46}$$

$$t_0 = \frac{\int_0^w \omega A(\omega)\beta(\omega)\, d\omega}{\int_0^w \omega^2 A(\omega)\, d\omega}. \tag{47}$$

Notice that t_0 is a linear functional of $\beta(\omega)$ and observe that consequently S_n, (43), is linear in $\beta(\omega)$. Therefore, according to the theorem of Riesz,[6] S_n is expressible in the succinct form

$$S_n = \int_0^w f_n(\omega)\beta'(\omega)\, d\omega \tag{48}$$

since $\beta(\omega)$ is linear in $\beta'(\omega)$. We now proceed to put S_n into the form of (48).

Combining (48) and (43), we write S_n as

$$S_n = \int_0^w \beta(\omega)[a_n\omega - \sin n\omega T]A(\omega)\, d\omega \tag{49}$$

where a_n does not depend on β, i.e.

$$a_n = \frac{\int_0^w \omega A(\omega) \sin n\omega T\, d\omega}{\int_0^w \omega^2 A(\omega)\, d\omega}.$$

Integrate (49) by parts to yield

$$S_n = \beta(\omega)g_n(\omega)\Big|_0^w - \int_0^w \beta'(\omega)g_n(\omega)\, d\omega \tag{50}$$

$$g_n(\omega) = \int_0^\omega [a_n x - \sin nxT]A(x)\, dx \tag{51}$$

which may finally be manipulated to give

$$S_n = \int_0^w f_n(\omega)\beta'(\omega)\, d\omega \tag{52}$$

where

$$f_n(\omega) = \int_\omega^w [a_n x - \sin nxT] A(x)\, dx. \tag{53}$$

We have now written the distortion D as

$$D = \frac{2}{\pi} \sum_{n=1}^{\infty} | S_n | \tag{54}$$

where each term S_n is a linear functional of delay and represents the distortion arising from intersymbol interference from symbols $\pm n$ symbols away. Obviously the terms S_n become quite insignificant for large n. For many delay curves, the principal interference is from adjacent symbols and only S_1 is of major importance. We shall demonstrate this when $A(\omega)$ is the commonly used raised cosine shaping and the delay is parabolic.

Using the Schwarz inequality in (52) gives a useful bound on S_n.

$$| S_n | \leqq \| f_n \| \; \| \beta' \| \tag{55}$$

$$\begin{aligned} \| f_n \| &= \sqrt{\int_0^w f_n^2(\omega)\, d\omega} \\ \| \beta' \| &= \sqrt{\int_0^w [\beta'(\omega)]^2\, d\omega} = \text{rms delay} \times \sqrt{w}. \end{aligned} \tag{56}$$

Thus we can see how fast the successive terms in (54) must approach zero. The total distortion is of course bounded by

$$D \leqq \frac{2}{\pi} \| \beta' \| \sum_{n=1}^{\infty} \| f_n \|. \tag{57}$$

The norm $\| f_n \|$ may be thought of as the sensitivity of a system to intersymbol interference at a distance of $\pm n$ symbols. The greater $\| f_n \|$, the more sensitive the system is to delay distortion.

The effect of the shape of the delay curve is clearly illustrated in (52), which is abbreviated

$$S_n = (\beta', f_n) \tag{58}$$

and represents the inner (or scalar, or dot) product of the functions (vectors) f_n and β'. S_n is less than or equal to the product of the lengths of the two vectors, which is what the Schwarz inequality in (55) says, and the equality occurs when the delay $\beta'(\omega)$ has the same shape as $f_n(\omega)$.

While expression (57) represents an upper bound on the distortion as a function of rms delay, this bound is generally not realizable. In fact, this bound is generally useless, since the sum of the norms $\| f_n \|$ frequently diverges. This does not mean the distortion can diverge, since this would require the delay to simultaneously have appreciable components in the direction of each of the vectors f_n. In the two examples we will study, it is shown that this divergence is possible in one case, where all the f_n are approximately in the same direction, whereas it is impossible in the other, where the f_n are nearly orthogonal.

To find a least upper bound on distortion as a function of rms delay we write

$$D = \frac{2}{\pi} \sum_{n=1}^{\infty} \epsilon_n S_n \qquad \epsilon_n = \begin{cases} +1 & S_n \geqq 0 \\ -1 & S_n < 0 \end{cases} \tag{59}$$

$$D = \frac{2}{\pi} \sum_{n=1}^{\infty} \epsilon_n (\beta', f_n) = \frac{2}{\pi} \left(\beta', \sum_{n=1}^{\infty} \epsilon_n f_n \right). \tag{60}$$

The distortion D is thus a scalar product of delay and some combination of the functions f_n. The sequence of sign coefficients $\{\epsilon_n\}$ is chosen so as to maximize this scalar product. The resulting value of D is the same as forming the sum of absolute values of the individual scalar products (β', f_n).

Using the Schwarz inequality in (60) we obtain

$$D \leqq \frac{2\sqrt{w}}{\pi} \times (\text{rms delay}) \times \max_{\{\epsilon_n\}} \left\| \sum_{n=1}^{\infty} \epsilon_n f_n \right\| \tag{61}$$

$$D \leqq \frac{2\sqrt{w}}{\pi} \times (\text{rms delay}) \times \max_{\{\epsilon_n\}} \left[\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \epsilon_n \epsilon_m (f_n, f_m) \right]^{\frac{1}{2}}. \tag{62}$$

Expression (62) is the least upper bound on distortion for a given value of rms delay, and the equality is obtained when

$$\beta'_{\text{worst}} = \sum_{n=1}^{\infty} \epsilon_n f_n . \tag{63}$$

The inequality (62) may be used to define the over-all *sensitivity* of a system to delay distortion

$$D \leqq (\text{rms delay}) \times (\text{sensitivity}) \tag{64}$$

$$\text{sensitivity} = \frac{2\sqrt{w}}{\pi} \max_{\{\epsilon_n\}} \left[\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \epsilon_n \epsilon_m (f_n, f_m) \right]^{\frac{1}{2}}. \tag{65}$$

The sensitivity is equal to $2\sqrt{w}/\pi$ times the length of the longest vector which can be obtained by summing the vectors $\pm f_n$.

We also have the obvious bounds

$$\frac{2\sqrt{w}}{\pi}\left[\sum_{n=1}^{\infty}\|f_n\|^2\right]^{\frac{1}{2}} \leqq \text{sensitivity} \leqq \frac{2\sqrt{w}}{\pi}\left[\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}|(f_n,f_m)|\right]^{\frac{1}{2}}. \tag{66}$$

4.2 *Example — Perfect Low-Pass System Operating at the Nyquist Rate*

As our first example we choose the "ideal" system, a perfect low-pass channel operating at a rate of $2W$ symbols per second

$$\begin{aligned} A(\omega) &= 1 \qquad 0 \leqq \omega \leqq \pi \\ A(\omega) &= 0 \qquad \pi \leqq \omega \\ T &= 1 \\ w &= \pi. \end{aligned} \tag{67}$$

We now evaluate the function $f_n(\omega)$, using these values

$$f_n(\omega) = \int_\omega^w [a_n x - \sin nxT]\, dx \tag{68}$$

$$a_n = \frac{\displaystyle\int_0^\pi \omega \sin n\omega\, d\omega}{\displaystyle\int_0^\pi \omega^2\, d\omega} = \frac{-3}{n\pi^2}(-1)^n \tag{69}$$

$$\begin{aligned} f_n(\omega) &= \int_\omega^\pi [a_n x - \sin nx]\, dx \\ &= \frac{(-1)^n}{n}\left[\frac{3}{2\pi^2}\omega^2 - \frac{(-1)^n \cos n\omega}{n} - \frac{1}{2}\right]. \end{aligned} \tag{70}$$

If the delay is constant, say $\beta'(\omega) = c$, we have

$$S_n = (\beta', f_n) = c\int_0^\pi f_n(\omega)\, d\omega \tag{71}$$

$$S_n = \frac{1}{n}\left(\frac{\pi^3}{3}\right)\left(\frac{3}{2\pi^2}\right)(-1)^n - \frac{1}{n}\frac{(-1)^n}{2}\pi = 0. \tag{72}$$

Thus there is no distortion when the delay is constant. Of course we already knew this, but it serves as a useful check on the method.

Looking now at the system sensitivity, we compute the following products

$$\begin{aligned}(f_n, f_m) &= \frac{1}{nm}\left[\frac{\pi}{5} - \frac{3}{\pi}\left(\frac{1}{m^3} + \frac{1}{n^3}\right)\right] \qquad n \neq m \\ &= \frac{1}{n^2}\left[\frac{\pi}{5} + \frac{\pi}{2n^2} - \frac{6}{\pi n^3}\right] \qquad n = m.\end{aligned} \tag{73}$$

Choose all the coefficients ϵ_n as positive and it is immediately seen that

$$\text{sensitivity} \geqq \frac{2}{\sqrt{\pi}}\left[\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}(f_n, f_m)\right]^{\frac{1}{2}} \to \infty \tag{74}$$

Therefore the perfect low-pass channel is infinitely sensitive to delay distortion, so that the smallest increment of delay can result in divergence of the eye picture. This result is not entirely unexpected, and is a good reason for not using "perfect" channels even if they were physically realizable.

4.3 *The Raised Cosine System*

4.3.1 *Derivation and Discussion of the Functions $f_n(\omega)$*

Now, instead of the flat amplitude shaping of the previous example which was so sensitive to delay distortion, we use a more gradual cutoff. The raised cosine shaping is the shaping most frequently used in practice, since it retains the proper zero crossings at the Nyquist rate and, as we will show, is less sensitive to delay distortion. Of course the penalty one pays for this protection against delay distortion is a doubling of the bandwidth for a given symbol rate as compared to the flat shaping previously discussed.

For the raised cosine shaping we have

$$\begin{aligned}A(\omega) &= \cos\omega + 1 \qquad 0 \leqq \omega \leqq \pi \\ A(\omega) &= 0 \qquad \omega \geqq \pi \\ T &= 2 \qquad w = \pi\end{aligned} \tag{75}$$

$$a_n = \frac{\int_0^{\pi} \omega \sin 2n\omega(\cos\omega + 1)\, d\omega}{\int_0^{\pi} \omega^2(\cos\omega + 1)\, d\omega} \tag{76}$$

$$a_n = \frac{1}{\left(\frac{\pi^2}{3} - 2\right) 2n(4n^2 - 1)}. \tag{77}$$

Notice that a_n falls off as $1/n^3$ as contrasted with the previous example where $a_n \sim 1/n$.

$$f_n(\omega) = \int_\omega^\pi [a_n x - \sin 2nx](\cos x + 1)\,dx \tag{78}$$

$$\begin{aligned} f_n(\omega) = {} & a_n \cos \omega + \frac{1}{2(2n+1)} \cos (2n-1)\omega \\ & + \frac{1}{2n} \cos 2n\omega + \frac{1}{2(2n+1)} \cos (2n+1)\omega \\ & + a_n\omega \sin \omega + \frac{a_n\omega^2}{2} - a_n \left[1 + \frac{\pi^2}{6}\right]. \end{aligned} \tag{79}$$

The first few functions $f_1(\omega)$, $f_2(\omega)$, and $f_3(\omega)$ are shown in Fig. 3. Observe that $f_1(\omega)$, representing adjacent symbol interference, has the greatest energy of these functions. Its shape in crude terms might be described as one cycle of cosine exponentially attenuated. The next function, $f_2(\omega)$, consists of about 2 cycles of cosine with less exponential attenuation, and this trend continues for the higher-order functions. The effect of delay distortion on the raised cosine system can be visualized with the aid of these functions. About the worst form of delay consists of one cycle of delay looking like $f_1(\omega)$. When the residual delay consists of a large number of ripples, say n cycles, then its distortion is not so great and comes largely from intersymbol interference at a distance of n symbols. When the delay is a slowly varying function of ω,

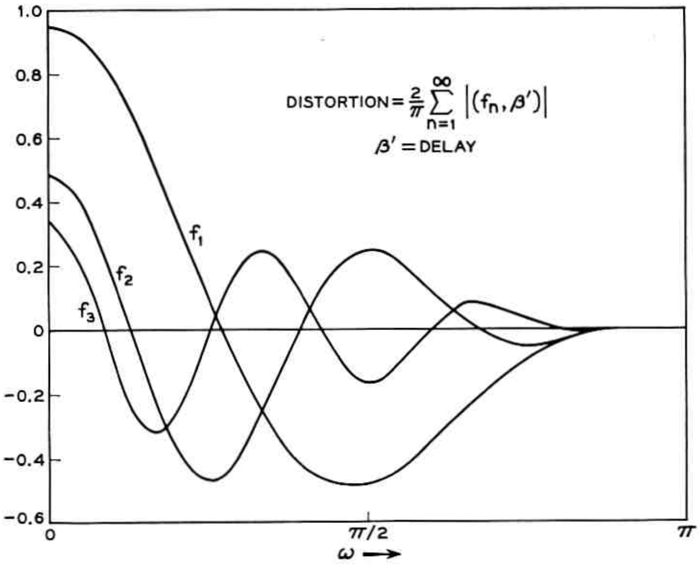


Fig. 3 — The functions $f_n(\omega)$ for raised cosine shaping.

the higher-order terms $S_n = (f_n, \beta')$ $n > 1$ become insignificant, and only adjacent symbol interference is of importance.

4.3.2 *Use of the Functions $f_n(\omega)$ in Computing Distortion*

All the functions $f_n(\omega)$ integrate to zero over the $[0,\pi]$ interval, so there is no distortion when the delay is constant. Now suppose we have parabolic delay

$$\beta'(\omega) = k\omega^2. \tag{80}$$

This is the general shape of delay to be expected in an unequalized voice channel.

Carrying out the relevant integrations gives

$$\begin{aligned} S_1 &= \int_0^\pi k\omega^2 f_1(\omega)\, d\omega = -0.411\pi k \\ S_2 &= 0.025\pi k \\ S_n &\approx \frac{0.282\pi k}{n^3}. \end{aligned} \tag{81}$$

In Ref. 2, Sunde computes the impulse response of a raised cosine network with parabolic delay distortion. In terms of the parameter m, the maximum delay in pulse intervals used in this reference, we find

$$k = 2m/\pi^2. \tag{82}$$

Using a value of $m = 2$ we read from Sunde's curve*

$$|h_1| + |h_{-1}| = 0.31 \tag{83}$$

while from (81) we have

$$|h_1| + |h_{-1}| = \frac{2}{\pi}|S_1| = 0.33. \tag{84}$$

The agreement is good, although the value of delay $m = 2$ is somewhat outside the range where the approximations are entirely valid.

To compute the distortion arising from parabolic delay we form the sum

$$D = \frac{2}{\pi}\sum_{n=1}^{\infty}|S_n| = 0.912k. \tag{85}$$

* We occasionally abbreviate $h(t_0 + nT)$ as simply h_n.

Notice that some 90 per cent of this distortion is due to the term S_1 (adjacent symbol interference). For this general shape of delay the term $S_1 = (\beta',f_1)$ would seem to be sufficiently indicative of system performance.

As another example of the computation of the effect of delay, we consider delay of the form

$$\beta' = a \cos \nu\omega. \tag{86}$$

This cosinusoidal delay is of the type frequently encountered as residual delay after partial equalization or in wider band systems. Depending on the number of cycles of delay across the band ν, only one or two of the terms S_n are of importance. These are the terms $n \approx \nu$. These various products (β',f_n) are shown as a function of ν in Fig. 4. Each product (β',f_n) peaks for ν a little less than n cycles and is very small elsewhere.

Rappeport[7] has studied the effect of this type of delay on the 4-phase data set using phase comparison detection. This study was effected using a digital computer simulation of the system. Rappeport plots curves of eye opening versus the number of cycles of delay in the passband, ν. Since the cosine is symmetrical, our low-pass results carry over directly to the passband in this case. The 4-phase system is essentially nonlinear because of the multiplication in the detection process. An exact expression relating eye opening to impulse response is not derivable for this system. An approximate expression for the eye opening is

$$I = 1 - D. \tag{87}$$

This expression neglects terms involving products such as $h_n h_m$. When the first four curves in Fig. 4 are summed to form D, the curve relating eye opening to delay frequency may be drawn as shown in Fig. 5. This curve is compared with the curve computed by Rappeport using $a = 0.5$ in each case, and it is seen that the general agreement is quite good except for somewhat more oscillation in the latter than in the former.

The exact eye opening for the 4-phase system depends not only on D, but on the relative magnitudes and signs of the samples h_n which sum to form D.

4.3.3 *Sensitivity and Bounds on Distortion for Raised Cosine Shaping*

The sensitivity of the raised cosine system to delay distortion may be calculated from consideration of the functions f_n. For $n > 4$ the terms involving a_n become approximately negligible and f_n consists of the terms

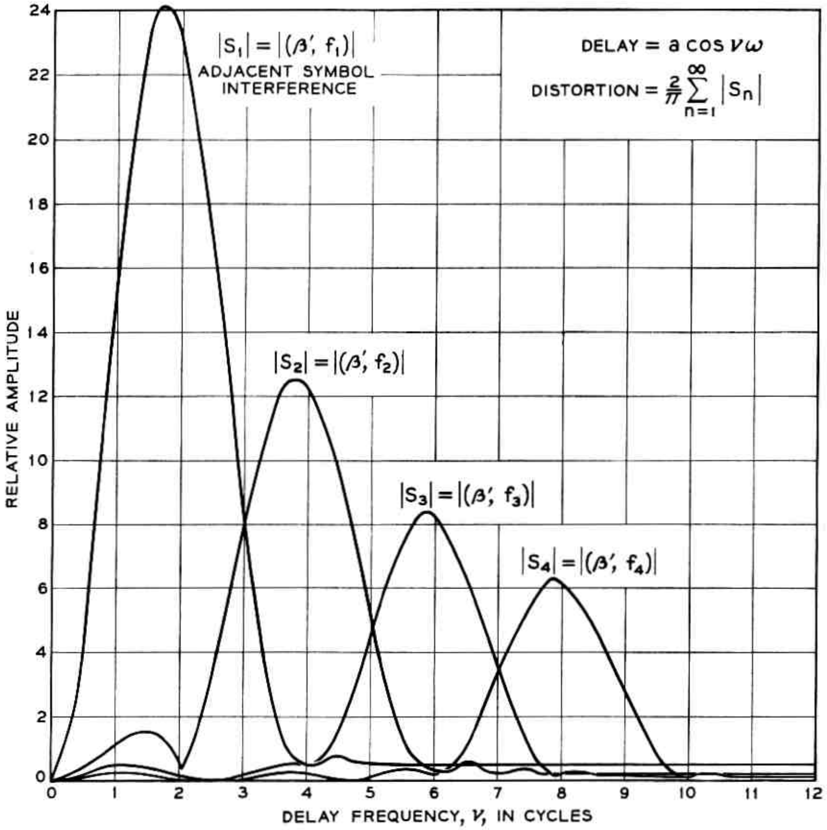


Fig. 4 — Various components of distortion for cosine delay.

$$f_n \approx \frac{1}{2(2n-1)} \cos (2n-1)\omega + \frac{1}{2n} \cos 2n\omega + \frac{1}{2(2n-1)} \cos (2n+1)\omega. \tag{88}$$

Thus f_n becomes approximately orthogonal to all f_m except for $m = n$ and $m = n \pm 1$. There is an overlap between f_n and f_{n+1} in the term $1/2(2n + 1) \cos (2n + 1)\omega$ and similarly an overlap between f_n and f_{n-1} in the term $1/2(2n - 1) \cos (2n - 1)\omega$. Obviously, to construct the sequence $\{\epsilon_n f_n\}$ of greatest energy we choose the signs ϵ_n such that all these shared cosine terms (the other terms are orthogonal) add in phase and thus reinforce each other. Thus it seems reasonable that $\epsilon_n \equiv +1$ for $n \geqq M$.

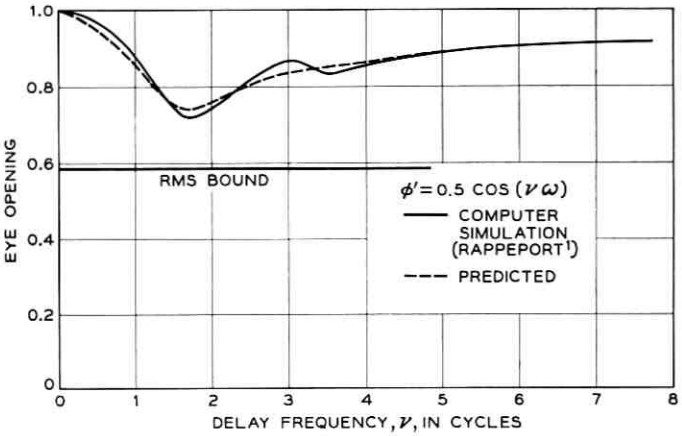


Fig. 5 — Eye opening for 4-phase data set for cosinusoidal delay.

By an exhaustive search on a digital computer of the effect of the first twelve coefficients ϵ_n, it was found that the maximum energy combination occurred for all $\epsilon_n = +1$ except for $\epsilon_1 = \epsilon_2 = \epsilon_3 = -1$. We designate this combination $f_{\text{mec}}(\omega)$ (*m*aximum *e*nergy *c*ombination). This function has the worst shape a delay can assume for a given rms value, and the sensitivity of the system is proportional to the norm of this function

$$f_{\text{mec}}(\omega) = \sum_{n=1}^{\infty} f_n(\omega) - 2[f_1(\omega) + f_2(\omega) + f_3(\omega)]. \tag{89}$$

We first perform the infinite summation involved here

$$\begin{aligned}\sum_{n=1}^{\infty} f_n(\omega) &= \left[\cos \omega + \omega \sin \omega + \frac{\omega^2}{2} - \left(1 + \frac{\pi^2}{6}\right)\right] \sum_{n=1}^{\infty} a_n \\ &\quad + \sum_{n=1}^{\infty} \left[\frac{1}{2(2n-1)} \cos (2n-1)\omega \right. \\ &\quad \left. + \frac{1}{2n} \cos 2n\omega + \frac{1}{2(2n+1)} \cos (2n+1)\omega\right].\end{aligned} \tag{90}$$

Both sums can be put in closed form. The first sum is

$$\sum_{n=1}^{\infty} a_n = \frac{3}{2(\pi^2 - 6)} \sum_{n=1}^{\infty} \frac{1}{n(4n^2 - 1)} = 0.14973 \tag{91}$$

while the last sum in (90) may be recognized as simply

$$\sum_{n=1}^{\infty} \frac{1}{n} \cos n\omega - \frac{1}{2} \cos \omega \tag{92}$$

which converges to

$$-\log \left| 2 \sin \frac{\omega}{2} \right| - \frac{1}{2} \cos \omega. \tag{93}$$

So that we finally find

$$\begin{aligned} f_{\text{mec}}(\omega) &= 0.14973 \left[\omega \sin \omega + \frac{\omega^2}{2} - 2.64493 \right] - 0.35027 \cos \omega \\ &\quad - \log \left| 2 \sin \frac{\omega}{2} \right| - 2f_1(\omega) - 2f_2(\omega) - 2f_3(\omega). \end{aligned} \tag{94}$$

This function is shown in Fig. 6. A delay curve of this shape has maximum detrimental effect on the raised cosine shaped system. The norm of $f_{\text{mec}}(\omega)$ was computed numerically to be

$$\| f_{\text{mec}} \| = 1.02 \tag{95}$$

and so the sensitivity of the raised cosine system is

$$\text{sensitivity} = \| f_{\text{mec}} \| \frac{2\sqrt{w}}{\pi} = 1.15 \tag{96}$$

$$D \leqq 1.15 \times (\text{rms delay}) \tag{97}$$

In the previous paragraphs we investigated the effect of cosinusoidal delay. For this shape and for an amplitude $a = 0.5$, the bound (97) gives $D \leqq 1.15 \times 0.5 \times 0.707 = 0.407$, so the eye opening $(1 - D)$ must be greater than or equal to 0.593. This value is shown on Fig. 5 along with the curves representing actual and computed performance for the cosine delay. At the lowest dips in these curves the distortion is about $\frac{2}{3}$ of the bound (97). The distortion computed for the parabolic delay distortion, however, is only about $\frac{1}{4}$ of its corresponding bound. As might be anticipated, the parabolic shape is a relatively weak form of delay distortion.

It is also of interest to compute bounds on samples of the impulse response.

$$| h_n | + | h_{-n} | = \frac{2}{\pi} | (\beta', f_n) | \leqq \frac{2}{\sqrt{\pi}} \| f_n \| \times (\text{rms delay}) \tag{98}$$

$$| h_1 | + | h_{-1} | \leqq 0.829 \times (\text{rms delay}) \tag{99}$$

$$| h_2 | + | h_{-2} | \leqq 0.443 \times (\text{rms delay}) \tag{100}$$

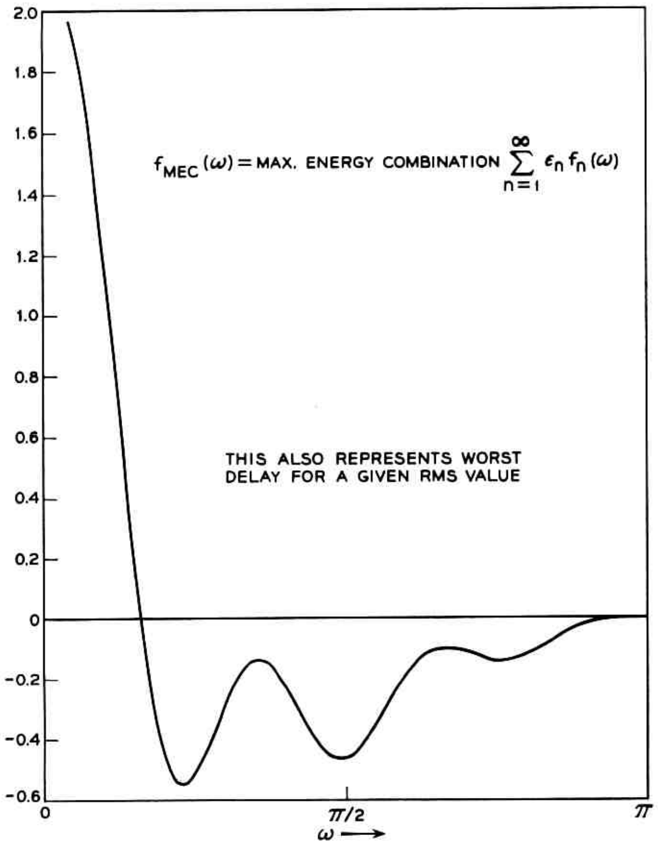


Fig. 6 — RMS bound on distortion; $D = 1.15 \times$ (rms delay).

$$| h_3 | + | h_{-3} | \leqq 0.289 \times \text{(rms delay)}. \tag{101}$$

For large n we have asymptotically

$$| h_n | + | h_{-n} | \leqq \frac{\sqrt{3}}{2n} \times \text{(rms delay)}. \tag{102}$$

Thus the individual samples of the impulse response are only bounded inversely proportionally to their distance from the peak time of the impulse response. In each case the maximum value is obtained when the delay is some constant times $f_n(\omega)$. However, the sum of all these terms can never exceed 1.15 × (rms delay), which paradoxically is less than the sum of the attainable bounds on only the first two terms.

Since adding or subtracting any constant delay does not affect the distortion D, the bounds given here are most effectively used by first subtracting the mean value of delay and dealing only with the variational component.

We can find similar bounds in terms of peak-to-peak constraints on the delay. The "differential" delay is frequently taken as the difference between the maximum and minimum values of delay across the band. We shall now find the shape of delay which maximizes distortion for a given peak-to-peak constraint and the corresponding value of distortion. From (60) we have

$$D = \frac{2}{\pi} \max_{\{\epsilon_n\}} \left(\beta', \sum_{n=1}^{\infty} \epsilon_n f_n \right). \tag{103}$$

If $\beta'(\omega)$ is peak-limited then the maximum value of (103) is obtained when $\beta'(\omega)$ is chosen as $+\beta'_{\max}$ when $\sum_{n=1}^{\infty} \epsilon_n f_n$ is positive and $-\beta'_{\max}$ when $\sum_{n=1}^{\infty} \epsilon_n f_n$ is negative. The resulting distortion is

$$D_{\max} = \frac{2\beta'_{\max}}{\pi} \max_{\{\epsilon_n\}} \int_0^w \left| \sum_{n=1}^{\infty} \epsilon_n f_n(\omega) \right| d\omega. \tag{104}$$

The problem reduces to finding the combination of signs $\{\epsilon_n\}$ such that the absolute integral of $\sum_{n=1}^{\infty} \epsilon_n f_n(\omega)$ is maximized. We call this maximizing combination $f_{\text{mai}}(\omega)$ (*m*aximum *a*bsolute *i*ntegral). By trial and error on a digital computer, the following sequence of signs for raised cosine shaping was found

$$\epsilon_n = +1 \quad \text{except} \quad \epsilon_1 = \epsilon_4 = -1 \tag{105}$$

$$f_{\text{mai}}(\omega) = \sum_{n=1}^{\infty} f_n(\omega) - 2[f_1(\omega) + f_4(\omega)]. \tag{106}$$

This function is shown in Fig. 7. The worst peak-to-peak delay is positive when $f_{\text{mai}}(\omega)$ is positive and negative when $f_{\text{mai}}(\omega)$ is negative. This worst delay curve is also shown in this figure. It is rectangular in shape with the single axis crossing at $\omega = 0.32\pi$. The integral of $|f_{\text{mai}}(\omega)|$ was computed numerically, so that from (104) we have

$$D \leqq \frac{2}{\pi} \times 1.293 \times \frac{1}{2} \text{ (peak-to-peak delay)} \tag{107}$$

$$D \leqq 0.412 \times \text{(peak-to-peak delay)}. \tag{108}$$

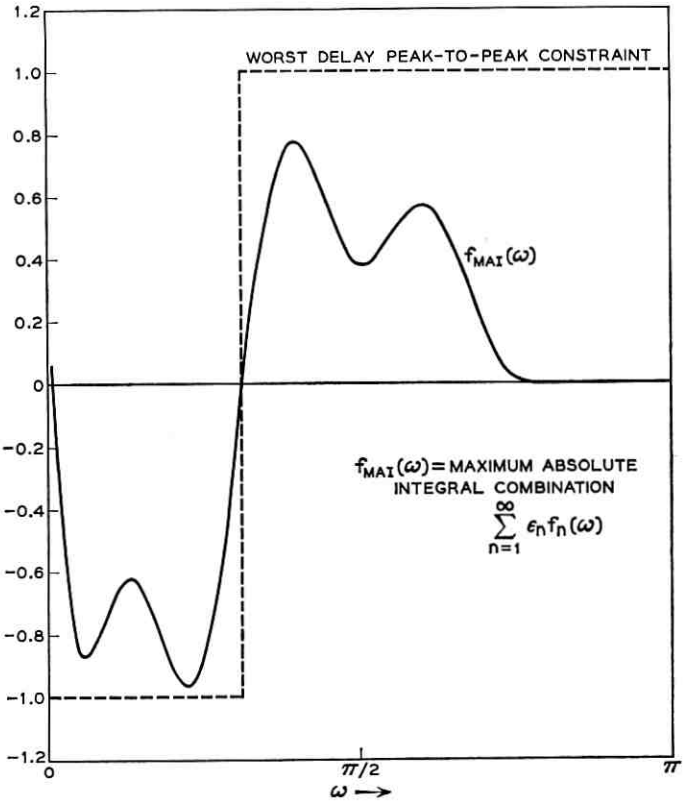


Fig. 7 — Peak bound on distortion; $D = 0.412 \times$ (peak-to-peak delay).

For example, the peak-to-peak delay of the cosine delay example was 1.0 and so the bound on D is 0.412 for this particular class of delay waveforms.

4.3.4 *The Effect of Increasing the Period T on Distortion*

It is possible to decrease the distortion due to intersymbol interference by sending symbols at a slower rate. If the same information rate is to be retained, the amount of information a given symbol conveys must be proportionally increased. With more symbols to be distinguished at the receiver, the smaller amount of distortion may be even more troublesome than before the rate was diminished, so there is a question as to whether or not the system performance for a given information rate can be improved by sending at a slower rate.

We consider a binary AM system. We previously found that the distortion from the normalized quiescent values of +1 and −1 is limited by the inequality

$$D \leqq 1.15 \times \text{rms delay}. \tag{109}$$

Suppose that we now send at half speed and, in order to maintain a constant information rate, change from a binary system to quaternary. First we compute a new value for sensitivity using the period $T = 4$.

Obviously the same functions $f_n(\omega)$ that we computed before still apply, with the change that we now use $f_{2n}(\omega)$ instead of $f_n(\omega)$. Now there is no overlap in the successive cosine terms in $f_n(\omega)$ and $f_{n\pm1}(\omega)$ and the terms $f_n(\omega), f_m(\omega)$, $n \neq m$ are very nearly orthogonal.

Assuming orthogonality we can easily compute the sensitivity from (65)

$$\text{sensitivity} = \frac{2}{\sqrt{\pi}} \left[\sum_{n=1}^{\infty} \| f_{2n} \|^2 \right]^{\frac{1}{2}} \tag{110}$$

$$\text{sensitivity} = 0.628$$

$$D(\text{half speed}) \leqq 0.628 \times (\text{rms delay}). \tag{111}$$

However, in an n-level AM system the amount of distortion necessary to cause an error is

$$D(\text{error}) = \frac{1}{n-1}. \tag{112}$$

The eye opening is defined as unity minus the ratio of distortion to the amount necessary to cause an error

$$I_{n \text{ level AM}} = 1 - (n-1)D. \tag{113}$$

Therefore we have for the same information rate

$$I_{\text{binary}} \geqq 1 - 1.15 \times (\text{rms delay}) \tag{114}$$

$$I_{\text{quaternary}} \geqq 1 - 1.884 \times (\text{rms delay}). \tag{115}$$

It is seen that the system has been made more susceptible to delay distortion by sending at a slower speed with proportionally more information per symbol. However, for any particular delay curve either system may perform better than the other. In comparing the two systems the statistics of the particular ensemble of delays to be encountered should be taken into consideration. Lacking any such statistics, the binary system is the obvious "minimax" choice in that it has less sensitivity to delay than the quaternary system.

4.3.5 *Zero-Distortion Delay Functions*

We have derived upper bounds on distortion as a function of rms and peak-to-peak delay. Now we might ask for some corresponding lower bounds. Since distortion is a positive quantity, we might wonder if it is possible to have zero distortion for some nonconstant delay curves. The distortion has been shown to be the sum of absolute values of certain linear functionals (β', f_n). Thus, to achieve zero distortion each of these functionals must be identically zero. In other words, the delay $\beta'(\omega)$ must be orthogonal to each of the functions f_n. This leads naturally to consideration of the completeness of the infinite sequence of functions $\{f_n\}$. If this sequence is complete in the space $L^2(0,\pi)$ then there exist no delay functions in this space for which the distortion is identically zero (see, for example, Ref. 6).

Actually, we have already demonstrated that constant functions are orthogonal to all f_n, so trivially the sequence is not complete. However, we are not particularly interested in constant-delay functions, so we might as well append a constant function to the sequence $\{f_n\}$ and consider the augmented sequence. That this sequence is not complete either may be easily proved by finding a function which is orthogonal to all f_n. For this purpose we write

$$f_n(\omega) = a_n\mu(\omega) + \frac{1}{2(n-1)} \cos (2n - 1)\omega + \frac{1}{2n} \cos 2n\omega + \frac{1}{2(n+1)} \cos (2n + 1)\omega \tag{116}$$

$$\mu(\omega) = \cos \omega + \omega \sin \omega + \frac{\omega^2}{2} - \left(1 + \frac{\pi^2}{6}\right). \tag{117}$$

Now observe that a function of the form

$$\psi_n(\omega) = \cos 2n\omega + b_1^n \cos (2n + 1)\omega + b_2^n \cos (n + 2)\omega + b_3^n \cos (2n + 3)\omega + b_4^n \cos (2n + 4)\omega \tag{118}$$

can be made orthogonal to all $f_n(\omega)$ by proper choice of the coefficients b^n. The four simultaneous conditions on these coefficients are

$$\begin{aligned} (\psi_n , \mu) &= 0 \\ (\psi_n , f_n) &= 0 \\ (\psi_n , f_{n-1}) &= 0 \\ (\psi_n , f_{n+1}) &= 0. \end{aligned} \tag{119}$$

These four conditions insure the orthogonality of ψ_n and f_m, since for $m > n + 1$ and $m < n - 1$ there is no overlap in the cosine terms and we made ψ_n orthogonal to μ, which constitutes the remaining portion of f_m.

Inserting (116), (117) and (118) into the four simultaneous equations (119) yields the result

$$b_1^n = \frac{4n^2 - 1}{n(2n - 1)} \tag{120}$$

$$b_2^n = \frac{-6(n + 1)}{n(2n - 1)} \tag{121}$$

$$b_3^n = \frac{(2n + 3)(2n + 5)}{n(2n - 1)} \tag{122}$$

$$b_4^n = \frac{-(n + 2)(2n + 5)}{n(2n - 1)}. \tag{123}$$

Some special considerations come in when solving for $\psi_0(\omega)$, which is a little different from the others. We merely quote the result here

$$\psi_0(\omega) = \cos \omega - 6 \cos 2\omega + 15 \cos 3\omega - 10 \cos 4\omega. \tag{124}$$

Thus, we have derived an infinite sequence of functions all of which are orthogonal to $\{f_n\}$ and therefore are distortionless. What we would really like to do, however, is to find all the functions which are distortionless in the space $L^2(0,\pi)$. We designate the subspace consisting of all distortionless functions as G. We call the linear manifold spanned by the sequence $\{f_n\}$ the distortion subspace, F. Each delay function in $L^2(0,\pi)$ can be expressed as the sum of two orthogonal functions $f \subset F$ which causes distortion and $g \subset G$ which is distortionless.

$$L^2(0,\pi) = F \oplus G. \tag{125}$$

We can form a sequence of orthonormal basis functions for F by orthonormalizing the sequence $\{f_n\}$. Unfortunately the sequence $\{\psi_n\}$ is not complete in G. For the purposes of analysis a sequence of approximate basis functions for G may be derived by the following procedure.

(*i*) Approximate f_n to the desired accuracy by

$$f_n = \sum_{m=1}^{M} a_m^{(n)} \cos m\omega. \tag{126}$$

(*ii*) Orthonormalize the functions f_n; $n = 1, 2, \cdots, (M - 1)/2$ giving $(M - 1)/2$ orthonormal basis functions in the M-dimensional space of the approximation.

TABLE I—AN APPROXIMATE SET OF ORTHONORMAL BASIS FUNCTIONS FOR THE SPACE G OF DISTORTIONLESS FUNCTIONS

$$\left(\sqrt{\frac{\pi}{2}}\, g_n = a_1 \cos \omega + a_2 \cos 2\omega + a_3 \cos 3\omega + \cdots + a_{11} \cos 11\, \omega\right)$$

	g_1	g_2	g_3	g_4	g_5	g_6
a_1	0.8478					
a_2	−0.5053	−0.2887	−0.0016	−0.0004	−0.0001	
a_3	−0.1042	0.8087				
a_4	0.1136	−0.4837	−0.3296			
a_5	0.0435	−0.1483	0.8233			
a_6	−0.0016	0.0766	−0.4322	−0.3540		
a_7	0.0044	0.0251	−0.1445	0.8259		
a_8	0.0092	−0.0131	0.0713	−0.4078	−0.3687	
a_9	0.0066	−0.0049	0.0252	−0.1443	0.8295	
a_{10}	0.0047	0.0018	−0.0116	0.0664	−0.3819	−0.4138
a_{11}	0.0021	0.0008	−0.0053	0.0302	−0.1736	0.9104

(*iii*) Derive the missing $(M + 1)/2$ orthonormal basis functions in this M-dimensional space. These are approximately g_n ; $n = 1, 2, \cdots, (M + 1)/2$.

(*iv*) $g_0 = 1$.

The reason this procedure works well is that each successive function f_n adds strong components of $\cos 2n\omega$ and $\cos (2n + 1)\omega$ as the sequence $\{f_n\}$ is orthonormalized. The g_n functions "interleave" to form a Fourier series. For $M = 11$ the six g functions thus generated are given in Table I.

Now any linear combination of the functions $g_n(\omega)$ has zero distortion. In particular, for any given delay curve we can find the closest distortion-free curve. This would indicate the minimum amount of equalization necessary to eliminate intersymbol interference. This nearest distortion-free function may be found by taking the projection of the particular delay $\beta'(\omega)$ onto the subspace G

$$P_G(\beta') = \sum_{n=1}^{\infty} (\beta', g_n) g_n \,. \tag{127}$$

In Figs. 8 and 9 an example of the use of (127) is shown. In Fig. 8 we consider a cosine delay, $\beta' = \cos 3\omega$, which we have previously considered (see Fig. 4). The projection of this delay on G using (127) is also shown in Fig. 8 and their corresponding impulse responses are shown in Fig. 9. Notice that the samples h_n of the impulse response of the uncorrected delay are poor. There is a peak in the response between h_1 and h_2 which we expect from our previous considerations in Section II. The corrected

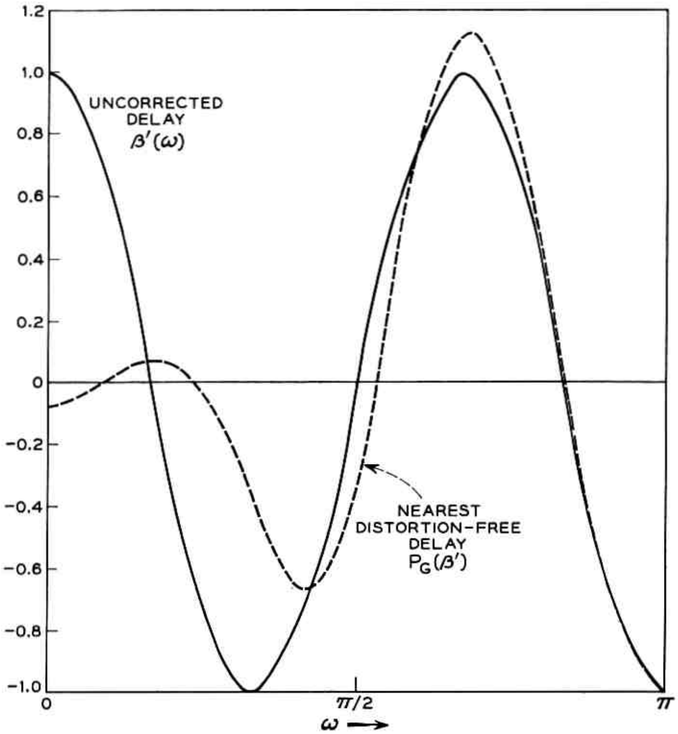


Fig. 8 — An example delay function and the nearest distortion-free delay.

delay seems to be a good fit in the interval $[\pi/2,\pi]$, and its response is well behaved, as is evidenced by the impulse response shown in Fig. 9.

Thus, we have apparently found an infinite set of delay functions corresponding to a particular amplitude characteristic such that the impulse responses satisfy the Nyquist criterion of regularly spaced zero crossings. Note that this was not possible for the case of flat amplitude characteristic mentioned in Section 4.2, since the set $\{f_n\}$ for this shaping is complete in the system bandwidth. For the raised cosine shaping we use more bandwidth for the same rate of transmission and consequently have more leeway in selection of good delay characteristics.

Actually the responses $h(t)$ corresponding to delays in G need not go exactly through zero at time $t_0 + nT$, but only approach zero to the order of approximation employed in our original assumptions. Since ordinarily* the approximation is good to terms cubic in $[\omega t_0 - \beta(\omega)]$, the

* So long as $S_n \geqq C_n$; see Section 4.1.

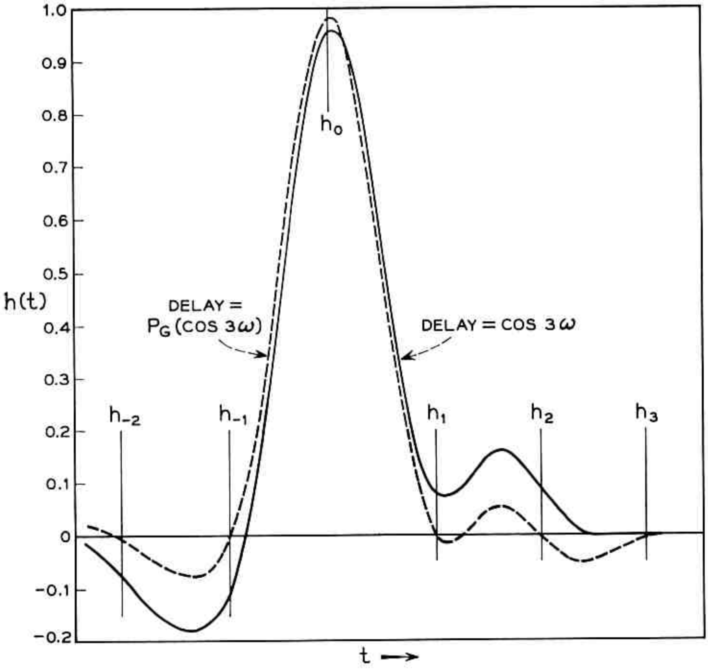


Fig. 9 — Impulse responses of uncorrected and corrected channels.

difference between $h(t_0 + nT)$ and zero becomes insignificant for small delays.

4.3.6 *Equalization*

There are several alternatives available when dealing with delay distortion. One alternative is the use of automatic equalization, whereby channel characteristics are measured and automatically equalized at the transmitter or receiver. Another alternative is the use of compromise equalization, in which a fixed network or a choice of fixed networks is designed to provide for average correction over a particular range of channels. Finally, one can do nothing to the system while amusing oneself with calculations of degradations in performance. We always assume that the particular channel to be used for transmission is chosen randomly for each call, so that it is not economically feasible to design an equalizer for each channel to be used.

For a fixed delay characteristic, the last section shows that there are an infinite number of all-pass networks which will provide near perfect equalization. The network with least rms delay has delay $\beta_c' = -\beta' + P_G(\beta')$, and in general any function $g(\omega) \subset G$ may be added to this delay so long as the approximation remains valid. Thus, the particular function easiest to realize physically may be chosen from this class of functions.

Now suppose we desire to design a delay equalizer to work over a certain class B of delay functions. For each call, the delay $\beta'(\omega)$ is to be chosen randomly from this ensemble. The optimum compromise equalizer $\beta_c'(\omega)$ is to be chosen such that the average distortion over the ensemble B is minimized. Since the delays $\beta'(\omega)$ (random) and $\beta_c'(\omega)$ (fixed) simply add, we have for the resulting distortion

$$D = \frac{2}{\pi} \sum_{n=1}^{\infty} \mid (\beta' + \beta_c', f_n) \mid \tag{128}$$

$$D = \frac{2}{\pi} \sum_{n=1}^{\infty} \mid (\beta', f_n) + (\beta_c', f_n) \mid . \tag{129}$$

The expected value of D averaged over the ensemble B is written

$$E[D] = \frac{2}{\pi} \sum_{n=1}^{\infty} E[\mid (\beta', f_n) + (\beta_c', f_n) \mid]. \tag{130}$$

This is the expression to be minimized by choice of β_c'. Knowing the statistics of the ensemble of channels we can derive the joint distribution of the variables $S_n = (\beta', f_n)$ and the marginal distributions $p(S_n)$. In terms of the latter distributions we have

$$E[D] = \frac{2}{\pi} \sum_{n=1}^{\infty} \int_{-\infty}^{+\infty} \mid S_n + (\beta_c', f_n) \mid p(S_n) dS_n . \tag{131}$$

Each term of the summation is positive, and it is possible to specify independently each component (β_c', f_n) of β_c'.

Therefore, we simply choose each component (β_c', f_n) so as to minimize the corresponding term of the summation (131). Each integral may be written

$$\begin{aligned} I_n = &\int_{-\infty}^{-(\beta_c', f_n)} [-S_n - (\beta_c', f_n)] p(S_n) dS_n \\ &+ \int_{-(\beta_c', f_n)}^{\infty} [S_n + (\beta_c', f_n)] p(S_n) dS_n . \end{aligned} \tag{132}$$

Differentiation with respect to (β_c', f_n) yields the stationary point (β_c', f_n) chosen such that

$$\int_{-\infty}^{-(\beta_c',f_n)} p(S_n)dS_n = \frac{1}{2}. \tag{133}$$

Therefore, each component (β_c',f_n) of the compromise delay is optimally chosen such that it is the *negative of the median value of* S_n. The compromise delay β_c' is not uniquely specified by these components; only its projection onto the space F has been determined. As before, any function $g(\omega) \subset G$ may be added without affecting the optimality of the resulting equalizer.

As a somewhat trivial example, suppose we desire to equalize a set of channels bounded by the narrow "ribbon" of width 2Δ.

$$\beta_0'(\omega) - \Delta \leqq \beta'(\omega) \leqq \beta_0'(\omega) + \Delta. \tag{134}$$

Now suppose that each of the scalar products (β',f_n) is equally likely to be greater or less than (β_0',f_n). In this event the completely trivial solution is to use $\beta_c' = -\beta_0' + g$. In particular, the smallest rms function of this sort is $\beta_c' = -\beta_0' + P_G(\beta_0')$. The residual delay in using this equalizer is bounded by $\pm\Delta$ across the band (plus a harmless g function), and our peak-to-peak distortion bound derived previously may be used to give

$$D \text{ residual} \leqq 0.824\Delta. \tag{135}$$

4.3.7 *The Range of the Approximation*

The key approximation made in the analysis thus far has been that $[\omega t_0 - \beta(\omega)]$ is small enough to use

$$\sin\,[\omega t_0 - \beta(\omega)] \approx [\omega t_0 - \beta(\omega)] \tag{136}$$

where t_0 is chosen to be the time of the peak value of the impulse response. We will now briefly examine the range of delay for which this approximation is valid.

By setting $h'(t_0) = 0$, we were able to derive an expression relating t_0 and the phase shift, $\beta(\omega)$. This equation, (47), was

$$t_0 = \frac{\int_0^w \omega A(\omega)\beta(\omega)\, d\omega}{\int_0^w \omega^2 A(\omega)\, d\omega} \tag{137}$$

and this was the value used for t_0 in (136).

Now we will demonstrate that the resulting function $[\omega t_0 - \beta(\omega)]$ is the *error in a least-squares straight-line fit to* $\beta(\omega)$. Hence, consider fitting a straight line $y = c\omega$ to the phase curve $\beta(\omega)$. The integral square error in the fit weighted, as all our integral expressions are, by the amplitude shaping $A(\omega)$ is

$$\text{integral squared error} = \int_0^w [c\omega - \beta(\omega)]^2 A(\omega)\, d\omega. \qquad (138)$$

By minimizing the error (138) with respect to c and using (137) we obtain $c_{\min} = t_0$.* Thus, the use of the peak time of the impulse response for t_0 results in the smallest values of $[\omega t_0 - \beta(\omega)]$ in a mean-square sense, which was an assertion previously made in connection with the choice of t_0. Also, we see that the time t_0 is the slope of a best fit straight line to the phase $\beta(\omega)$. This state of affairs is depicted in Fig. 10.

In order to be assured of, say, 10 per cent accuracy in the use of approximation (136), we might guarantee that $[\omega t_0 - \beta(\omega)] \leqq \pi/4$. Thus, the phase should not deviate from a straight line by more than $\pi/4$ radians. (Of course, these are *sufficient* but *not necessary* conditions for 10 per cent accuracy.) Now we ask what limits we may put on *peak delay* such that the *phase* will meet this condition no matter what the exact shape of the delay happens to be. Remember that $\beta(0) = 0$ and $\beta'(\omega) \geqq 0$ for physical realizability.

This problem is best suited for the semi-mathematical method called "common sense." Since the delay is to be peak limited between $\beta'_{\max}$ and 0, we allow only use of these two extreme values in finding the shape of delay such that the deviation of its integral (phase) from a straight line is maximized. Furthermore, it is evident that only one transition from $\beta' = \beta'_{\max}$ to $\beta' = 0$ should be used in the interval $[0,w]$. The reader may convince himself that more transitions in delay would result in a better straight-line fit to the phase. Therefore, the shape of the phase which for a given peak delay results in the poorest use of the approximation (136) is specified except for the transition point ω_0. This situation is shown in Fig. 11.

For raised cosine amplitude shaping, the error near the edge of the band is very lightly weighted, and so we take the error at $\omega = \omega_0$ as the largest important error. This error is equal to $\omega_0 t_0$ and is to be maximized with respect to the transition point ω_0.

$$E = \omega_0 t_0 = \frac{\omega_0}{m} \int_0^w \omega A(\omega) \beta(\omega)\, d\omega \qquad (139)$$

* This expression with $c = t_0$ may be used as a useful distortion measure relating performance and phase shift. For an AM system operating at rate $2W$ symbols/second this may be shown to be proportional to the mean-square estimation error at the receiver due to intersymbol interference.

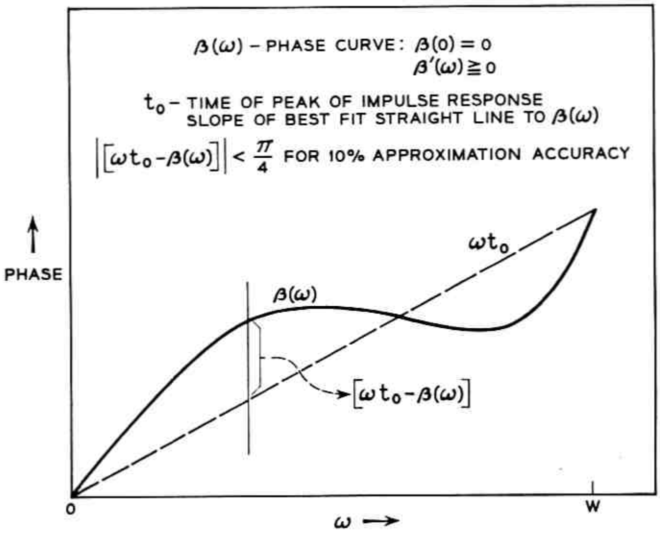


Fig. 10 — Factors involved in the approximation.

where m is a constant

$$m = \int_0^w \omega^2 A(\omega)\, d\omega. \tag{140}$$

Using the curve $\beta(\omega)$ shown in Fig. 11 gives

$$E = \frac{\omega_0}{m} \int_{\omega_0}^{w} \omega \beta'_{\max}(\omega - \omega_0) A(\omega)\, d\omega \tag{141}$$

$$\frac{dE}{d\omega_0} = \frac{\beta'_{\max}}{m} \left\{ \int_{\omega_0}^{w} \omega^2 A(\omega)\, d\omega - 2\omega_0 \int_{\omega_0}^{w} \omega A(\omega)\, d\omega \right\}. \tag{142}$$

For the raised cosine shaping the maximum point of (141) may be found by a solution of the resulting transcendental equation when $dE/d\omega_0 = 0$ and $A(\omega) = \cos \omega + 1$ are used in (142). This procedure yields

$$\omega_0 = 0.255\pi. \tag{143}$$

Notice that the corresponding delay curve is very similar to the worst delay from the standpoint of distortion, which is shown in Fig. 7. For this choice of ω_0 we may evaluate the maximum error from (141).

$$E_{\max} = 0.37\beta'_{\max}. \tag{144}$$

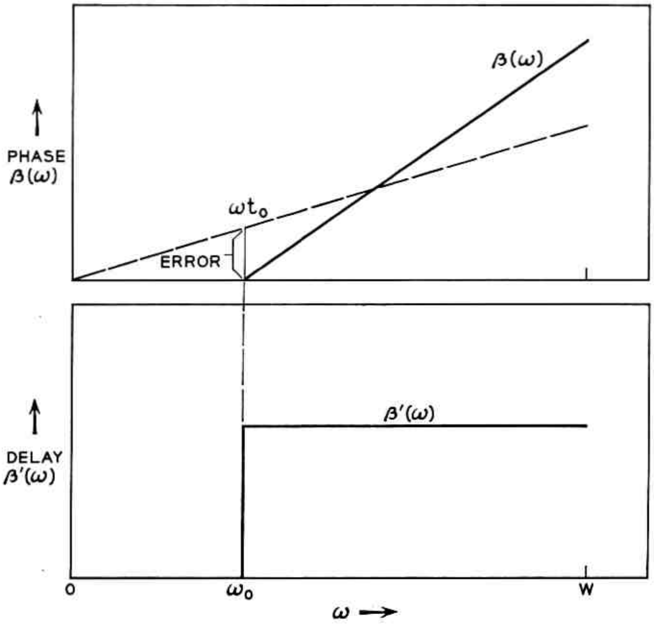


Fig. 11 — Most unfavorable (approximation poorest) delay curve for a given peak value.

For this error to be less than $\pi/4$ for an assured 10 per cent accuracy, we finally arrive at

$$\beta'_{\max} \leqq 2.12. \tag{145}$$

Thus, as long as the delay variation does not exceed 2.12 seconds across the band we are assured of at least 10 per cent accuracy in results regardless of the actual shape of the delay. This is, of course, normalized for the choice of $\omega = \pi$ and $T = 2$ seconds, so that 2.12 seconds of delay variation corresponds to 1.06 pulse intervals.

When the delay differs from the worst form we have just derived, the approximation holds for greater ranges. For example, we found that for the parabolic delay discussed in Section 4.2.2 the approximation was accurate to within 10 per cent in spite of a total delay variation of 4 seconds across the band. The important consideration is that the distortion is quite appreciable before the approximation breaks down. From the peak distortion bound derived previously, we find that the distortion corresponding to a variation of 2.12 seconds may be as large as 0.873. There is a definite connection between the value of the distortion and the goodness of the approximation through the validity of (138) as a distortion measure. Thus, we would expect that the techniques would be accurate in nearly all cases of, roughly, $D \leqq 0.6$. For this value of distortion, the eye in a binary system is more than half closed, and the channel may be unsuitable for higher alphabet size transmission.

4.3.8 *Channel Memory Truncation Error*

Theoretically, the response from a band-limited network lasts to infinity, so that in calculating distortion an infinite number of terms must be used for the criterion D

$$D = \sum_{\substack{n=-\infty \\ n \neq 0}}^{+\infty} | h_n |. \tag{146}$$

In many analysis problems and particularly in experimental work and computer simulations it is necessary to neglect the channel memory for $| t | > NT$ seconds. In experimental runs and computer simulations this corresponds to using all possible patterns of length $2N + 1$ symbols. The problem arises as to how much of an error can be made in computing the distortion D using a finite number of terms.

Assume that the terms $| h_n |$ for $| n | > N$ are to be neglected in the summation. The error in computing D is

$$E(N) = \sum_{\substack{n=-\infty \\ n \neq 0}}^{+\infty} | h_n | - \sum_{\substack{n=-N \\ n \neq 0}}^{+N} | h_n | \tag{147}$$

$$E(N) = \sum_{n=-\infty}^{-N-1} | h_n | + \sum_{n=N-1}^{\infty} | h_n |. \tag{148}$$

As we have previously shown, this expression is approximately equal to

$$E(N) = \frac{2}{\pi} \sum_{n=N+1}^{\infty} | (f_n , \beta') | \tag{149}$$

and thus the error can be bounded using the Schwarz inequality on the maximum energy combination of the functions f_n , $n = N + 1, \cdots , \infty$. For $N > 3$ we may as well neglect the terms in f_n involving the fast-vanishing constant a_n , leaving only the three cosine terms. Obviously, the maximum energy combination of these is all terms adding in-phase

$$E(N) \leqq \frac{2}{\sqrt{\pi}} \times (\text{rms delay}) \times \| \xi_N \| \tag{150}$$

$$\xi_N(\omega) = \sum_{n=N+1}^{\infty} f_n(\omega) \tag{151}$$

$$\begin{aligned}\xi_N(\omega) \approx \sum_{n=N+1}^{\infty} \Bigg[& \frac{1}{2(2n-1)} \cos (2n - 1)\omega + \frac{1}{2n} \cos 2n\omega \\ & + \frac{1}{2(2n+1)} \cos (2n + 1)\omega \Bigg]\end{aligned} \tag{152}$$

$$\xi_N(\omega) = \sum_{n=2N+1}^{\infty} \frac{1}{n} \cos n\omega - \frac{1}{2(2N+1)} \cos (2N + 1)\omega \tag{153}$$

$$\| \xi_N \|^2 = \frac{\pi}{2} \sum_{n=2N+1}^{\infty} \frac{1}{n^2} - \frac{3\pi}{8(2N+1)^2}. \tag{154}$$

We finally write the maximum error as

$$E(N) \leqq e_N \times (\text{rms delay}) \tag{155}$$

$$e_N = \sqrt{\frac{\pi^2}{3} - \frac{3}{2(2N+1)^2} - 2\sum_{n=1}^{2N} \frac{1}{n^2}}. \tag{156}$$

As shown in Fig. 12, this bound drops rapidly for N small and then levels out to a very slow descent, so that some 20 per cent of the original distortion bound can still remain after consideration of 16 pulse intervals on each side of the peak. This rather negative result tells us only that there exist mathematical delay functions that have considerable distor-

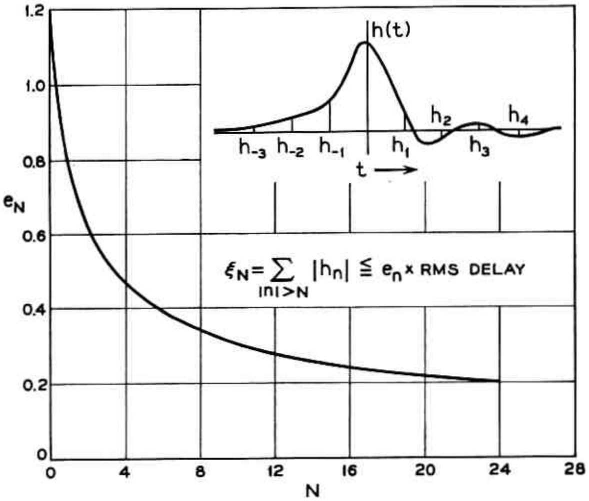


Fig. 12 — An upper bound on distortion arising from symbols at a distance greater than N.

tion at great distances from the reference (peak response) time. These functions, however, are high-frequency waveforms which would not ordinarily be encountered.

Consider a computer simulation in which 40 samples across the band $[0,\pi]$ are used to specify the delay. The highest delay frequency which need be considered is then 20 cycles of delay across the $[0,\pi]$ band. Consequently, only the functions f_n for $n \leqq 20$ will contribute distortion components. In order to test distortion over this interval it would be necessary to test 2^{40} sequences of binary symbols, which is, of course, quite unreasonable. Therefore, the effect of intersymbol interference is usually only measured to the extent of, say, four symbols in either direction.

We can find the maximum error now by computing the norm of the sum of the functions $f_n(\omega)$, $n = 5, 6, \cdots, 20$

$$\xi_{5,20}(\omega) = \sum_{n=11}^{40} \frac{1}{n} \cos n\omega - \frac{1}{22} \cos 11\omega \tag{157}$$

$$\| \xi_{5,20} \|^2 = \frac{\pi}{2} \left\{ \sum_{n=11}^{40} \frac{1}{n^2} - \frac{3}{22^2} \right\}. \tag{158}$$

The sums involved in the expressions are conveniently computed using a Euler-Maclaurin expansion for the integral of $1/x^2$. This gives

$$E_{5,20} \leqq 0.352 \times \text{rms delay} \tag{159}$$

which is still a considerable error, even though the delay is bandwidth limited.

In all our computations of distortion bounds we have been using the maximum energy combination of the functions $\pm f_n(\omega)$. For each particular delay it will be one of the combinations of functions $\pm f_n(\omega)$ which defines the distortion functional, not necessary the maximum energy combination. However, it is interesting to note that the combination with *least* energy would only result in a factor of $\sqrt{2}$ in the bounds calculated.

V. THE RELATIONSHIP OF DISTORTION TO DELAY FOR BANDPASS SYSTEMS

5.1 *Derivation of the Sequence of Functionals Involved for Bandpass Systems*

In dealing with bandpass systems, the system impulse response is most conveniently given in terms of its envelope and phase with respect to a carrier or other reference frequency within the bandwidth of the system

$$h(t) = P(t) \cos [\omega_c t - \psi(t)]. \tag{160}$$

Alternately, the response may be written in terms of in-phase and quadrature components at the carrier frequency

$$h(t) = R(t) \cos \omega_c t - Q(t) \sin \omega_c t \tag{161}$$

$$P(t) = \sqrt{R^2(t) + Q^2(t)}. \tag{162}$$

As discussed by Sunde,[1] the in-phase and quadrature components of the impulse response may be related to the amplitude and phase characteristics of the channel's frequency domain description by a simple transformation of the defining Fourier integral. This transformation to passband coordinates gives

$$R(t) = \frac{1}{\pi} \int_{-\omega_c}^{\infty} \mathcal{A}(\omega) \cos [\omega t - \varphi(\omega)] \, d\omega \tag{163}$$

$$Q(t) = \frac{1}{\pi} \int_{-\omega_c}^{\infty} \mathcal{A}(\omega) \sin [\omega t - \varphi(\omega)] \, d\omega. \tag{164}$$

In this section we use $\mathcal{A}(\omega)$ and $\varphi(\omega)$ for amplitude and phase instead of $A(\omega)$ and $\beta(\omega)$ since these functions are now defined with respect to the carrier frequency, ω_c. That is

$$\mathcal{A}(\omega) = A(\omega_c + \omega) \tag{165}$$

$$\varphi(\omega) = \beta(\omega_c + \omega). \tag{166}$$

Now we will work under the hypothesis that a suitable criterion of distortion for bandpass systems is

$$D = \sum_{\substack{n=-\infty \\ n \neq 0}}^{+\infty} P(nT + t_0) = \sum_n{}' P_n . \tag{167}$$

This criterion is similar to the low-pass criterion, except we now assume that the receiver makes use of the envelope properties of the impulse response.

In terms of samples of the quadrature components we have

$$D = \sum_n{}' \sqrt{R_n^2 + Q_n^2}. \tag{168}$$

Unfortunately this is a fairly hopelessly nonlinear criterion to work with, so we shall make the judicious approximation shown in Fig. 13. Here we take the distortion D as the length of the vector formed by summing all vectors of the form

$$P_n \angle \tan^{-1} \frac{| Q_n |}{| R_n |}.$$

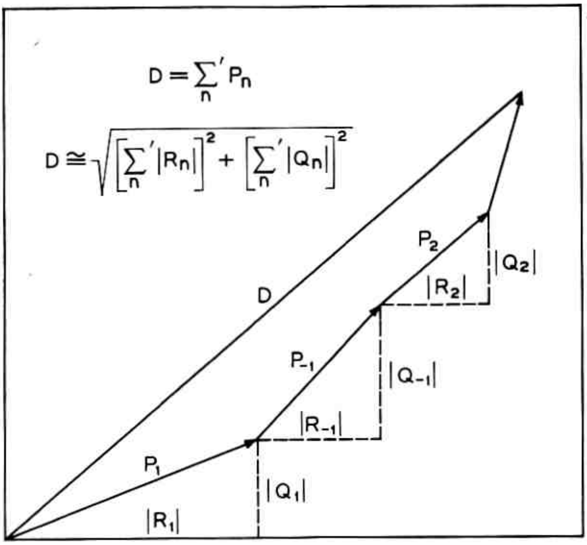


Fig. 13 — The approximate distortion criterion for bandpass systems.

Thus we have

$$D \approx \sqrt{D_r^2 + D_q^2} \tag{169}$$

$$D_r = \sum_n{}' \mid R_n \mid \tag{170}$$

$$D_q = \sum_n{}' \mid Q_n \mid . \tag{171}$$

Now, the component distortions D_r and D_q are of the same form as the low-pass distortion treated previously. We take D_q as an example in what follows

$$\begin{aligned} Q_n \pm Q_{-n} = \frac{1}{\pi} \int_{-\omega_c}^{\infty} \mathfrak{A}(\omega) \{ &\sin [\omega(t_0 + nT) - \varphi(\omega)] \\ &\pm \sin [\omega(t_0 - nT) - \varphi(\omega)] \} \, d\omega \end{aligned} \tag{172}$$

$$Q_n + Q_{-n} = \frac{2}{\pi} \int_{-\omega_c}^{\infty} \mathfrak{A}(\omega) \sin [\omega t_0 - \varphi(\omega)] \cos n\omega T \, d\omega \tag{173}$$

$$Q_n - Q_{-n} = \frac{2}{\pi} \int_{-\omega_c}^{\infty} \mathfrak{A}(\omega) \cos [\omega t_0 - \varphi(\omega)] \sin n\omega T \, d\omega. \tag{174}$$

Using the approximation $[\omega t_0 - \varphi(\omega)]$ small gives

$$Q_n + Q_{-n} \approx \frac{2}{\pi} \int_{-\omega_c}^{\infty} \mathcal{A}(\omega)[\omega t_0 - \varphi(\omega)] \cos n\omega T \, d\omega \tag{175}$$

$$Q_n - Q_{-n} \approx 0 \tag{176}$$

$$| Q_n | + | Q_{-n} | \approx \frac{2}{\pi} \left| \int_{-\omega_c}^{\infty} \mathcal{A}(\omega)[\omega t_0 - \varphi(\omega)] \cos n\omega T \, d\omega \right|. \tag{177}$$

Notice that in (177) the quantity $(Q_n - Q_{-n})$ need not be identically zero as in the approximation (176), but should only be smaller in absolute value than the quantity $(Q_n + Q_{-n})$.

We find the time of the peak value of the impulse response, t_0, by requiring $R'(t_0) = 0$. The quadrature component goes through zero at $= 0$ and is small enough at t_0 to be neglected

$$R'(t_0) = 0 = \frac{-1}{\pi} \int_{-\omega_c}^{\infty} \omega\mathcal{A}(\omega) \sin [\omega t_0 - \varphi(\omega)] \, d\omega \tag{178}$$

$$t_0 \approx \frac{\displaystyle\int_{-\omega_c}^{\infty} \omega\mathcal{A}(\omega)\varphi(\omega) \, d\omega}{\displaystyle\int_{-\omega_c}^{\infty} \omega^2\mathcal{A}(\omega) \, d\omega}. \tag{179}$$

This is incidental to the development of the quantity $| Q_n | + | Q_{-n} |$, because for symmetric shaping of $\mathcal{A}(\omega)$ the terms involving t_0 in (177) integrate to zero. This shows that the antisymmetric portion of the delay $\varphi'(\omega)$ does not have a first-order influence on t_0. However, in solving for $| R_n | + | R_{-n} |$ the equations do involve t_0 in first-order terms.

To maintain notational continuity with low-pass results as much as possible we designate

$$S_{qn} = - \int_{-\omega_c}^{\infty} \mathcal{A}(\omega)[\omega t_0 - \varphi(\omega)] \cos n\omega T \, d\omega \tag{180}$$

$$| Q_n | + | Q_{-n} | = \frac{2}{\pi} | S_{qn} |. \tag{181}$$

For $\mathcal{A}(\omega)$ symmetrical about zero (the carrier frequency) (180) becomes

$$S_{qn} = \int_{-\omega_c}^{\infty} \varphi(\omega)\mathcal{A}(\omega) \cos n\omega T \, d\omega. \tag{182}$$

We integrate by parts and make the arbitrary assignment of zero phase shift at the reference frequency to obtain*

$$S_{qn} = \int_{-w/2}^{+w/2} f_{qn}(\omega)\varphi'(\omega)\,d\omega = (f_{qn}, \varphi') \tag{183}$$

$$f_{qn}(\omega) = \int_0^\omega \mathfrak{A}(x) \cos nxT\,dx. \tag{184}$$

A similar development holds for the terms $|R_n| + |R_{-n}|$, and the resulting expressions are exactly the same as the low-pass equations except they are translated to the reference and are defined for both negative and positive deviations from this reference

$$|R_n| + |R_{-n}| = \frac{2}{\pi}|S_{rn}| \tag{185}$$

$$S_{rn} = \int_{-w/2}^{+w/2} f_{rn}(\omega)\varphi'(\omega)\,d\omega = (f_{rn}, \varphi') \tag{186}$$

$$f_{rn}(\omega) = \int_\omega^{w/2} [\alpha_n x - \sin nxT]\mathfrak{A}(x)\,dx \tag{187}$$

$$\alpha_n = \frac{\displaystyle\int_{-w/2}^{w/2} \omega\mathfrak{A}(\omega) \sin n\omega T\,d\omega}{\displaystyle\int_{-w/2}^{w/2} \omega^2\mathfrak{A}(\omega)\,d\omega}. \tag{188}$$

To briefly summarize, we have written the distortion in a bandpass system as the length of a vector whose two quadrature components are D_r and D_q

$$D = \sqrt{D_r^2 + D_q^2}. \tag{189}$$

Each of these components is the sum of the absolute values of a sequence of linear functionals of delay

$$D_r = \frac{2}{\pi}\sum_{n=1}^{\infty} |(f_{rn}, \varphi')| \tag{190}$$

$$D_q = \frac{2}{\pi}\sum_{n=1}^{\infty} |(f_{qn}, \varphi')|. \tag{191}$$

The functions f_{rn} and f_{qn} are of course independent of delay and are obtained from the amplitude characteristics by operations (187) and (184).

We are working with symmetric amplitude characteristics, and consequently it may be seen that

* Another choice of reference phase may easily be made here.

$$f_{qn}(\omega) = -f_{qn}(-\omega) \tag{192}$$

$$f_{rn}(\omega) = f_{rn}(-\omega). \tag{193}$$

The quadrature functions f_{qn} are odd functions of frequency and the in-phase functions f_{rn} are even functions. The two parts into which the distortion was divided therefore arise separately from the odd and even portions of the delay. For example, if the delay is an even function, $D_q = 0$ and the only distortion is D_r. Since D_r is defined identically except for a translation as the low-pass distortion D, we have the necessary result that the system may be treated as low-pass with identical results in the event of even delay. Obviously also

$$(f_{rn}, f_{qm}) = 0 \qquad \text{all } n \text{ and } m. \tag{194}$$

The delay function $\varphi'(\omega)$ may be divided into its even and odd components, $\varphi_r'(\omega)$ and $\varphi_q'(\omega)$, and the analysis of the distortion properties of each of these components proceeds exactly as in the low-pass analysis. For the even component of delay, we use the functionals defined by the sequence $\{f_{rn}\}$ and for the odd components we use the sequence $\{f_{qn}\}$. The two distortions D_r and D_q are then added root-sum-square.

5.2 *The Raised Cosine System*

5.2.1 *The Functions $f_{rn}(\omega)$ and $f_{qn}(\omega)$*

We now consider the use of an amplitude shaping of the raised cosine form

$$\mathcal{A}(\omega) = \tfrac{1}{2}(1 + \cos\omega) \qquad -\pi \leqq \omega \leqq +\pi. \tag{195}$$

By substitution into equations (188), (187) and (184) the following results are obtained

$$\alpha_n = \frac{1}{\left(\frac{\pi^2}{3} - 2\right) 2n(4n^2 - 1)} \tag{196}$$

$$\begin{aligned} f_{rn}(\omega) = \frac{\alpha_n}{2}\left[\cos\omega + \omega\sin\omega + \frac{\omega^2}{2} - \left(1 + \frac{\pi^2}{6}\right)\right] \\ + \frac{1}{4(2n-1)}\cos(2n-1)\omega + \frac{1}{4n}\cos 2n\omega \\ + \frac{1}{4(2n-1)}\cos(2n+1)\omega \end{aligned} \tag{197}$$

$$f_{qn}(\omega) = \frac{1}{4(2n-1)} \sin (2n-1)\omega + \frac{1}{4n} \sin 2n\omega + \frac{1}{4(2n+1)} \sin (2n+1)\omega. \tag{198}$$

The first few pairs of these functions are shown in Fig. 14.

5.2.2 *Sample Distortion Calculations*

Having assumed any particular shape of delay curve, one may easily compute the resultant distortion to the desired accuracy by computing a number of the linear functionals. Although the primary use of these functionals is in understanding the effect of shape in delay on the distortion and in ascertaining bounds and other factors in this relationship, it is quite necessary that when confronted with the reality of an actual system the mathematics be able to predict specific results.

First, we consider a check on the mathematical methods and approxi-

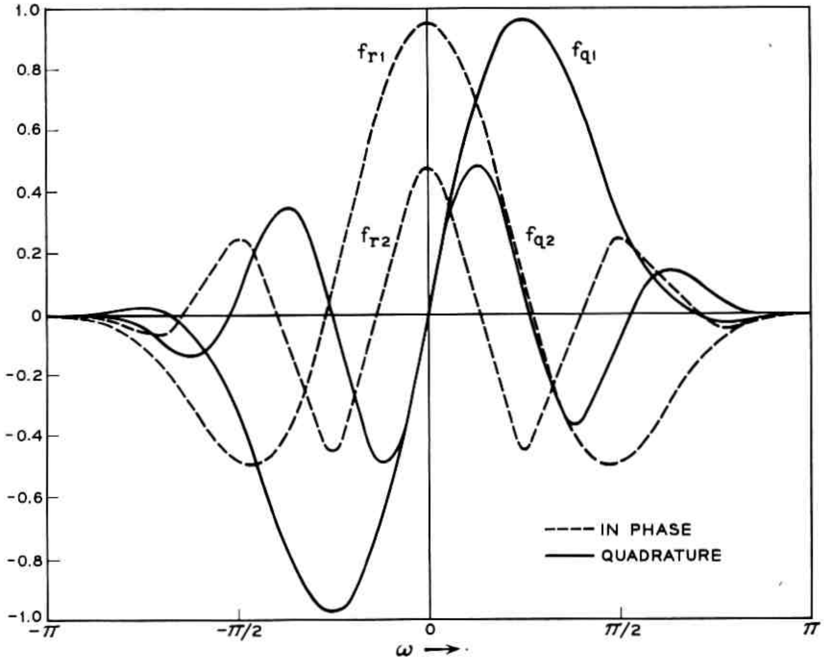


Fig. 14 — Some of the functions f_{rn} and f_{qn} for raised cosine shaping.

mations used. Consider the effect of linear delay on the raised cosine response

$$\varphi'(\omega) = k\omega. \tag{199}$$

Since this delay function is odd, the products (f_{rn}, φ') vanish and the only distortion is D_q. For adjacent symbol interference we have

$$P_1 + P_{-1} = \frac{2}{\pi} \mid S_{q1} \mid = \frac{2}{\pi} \mid (f_{q1}, k\omega) \mid \tag{200}$$

$$(f_{q1}, k\omega) = \int_{-\pi}^{+\pi} k\omega f_{q1}(\omega)\, d\omega = \frac{11\pi}{36}\, k \tag{201}$$

$$P_1 + P_{-1} = 11k/18. \tag{202}$$

For a specific example we take $k = 2/\pi$ which gives

$$P_1 + P_{-1} = 0.389 \qquad \text{(predicted)} \tag{203}$$

while from Sunde[2] the computed impulse response for this value of slope is

$$P_1 + P_{-1} = 0.387 \qquad \text{(computed)}. \tag{204}$$

The agreement here is probably better than should ordinarily be expected.

Now we turn to predicting the performance, measured by the eye opening, of the four-phase data subset. As explained in the previous chapter, this system is inherently nonlinear and using $I = 1 - D$ as the eye aperture for this system involves a certain approximation. In particular, we will examine the performance of this system for delay of the form

$$\varphi'(\omega) = \alpha \sin \nu\omega \tag{205}$$

since there are published results for this choice of delay. Again we are dealing for the moment with an odd delay function and need only evaluate D_q. As a function of the number of delay ripples in the band, ν, the various products (f_{qn}, φ') are easily visualized, since f_{qn} consists of only three sine terms itself. The behavior of these products is very nearly like the behavior of its cosine counterpart shown previously in Fig. 4 and will not be depicted here. In Fig. 15 the distortion for $\alpha = 0.5$ calculated by summing these products is illustrated as a function of ν along with the corresponding curve from Rappeport.[7] The latter curve was obtained by use of a computer simulation of the system, and the agreement between this simulation and actual results is claimed to be

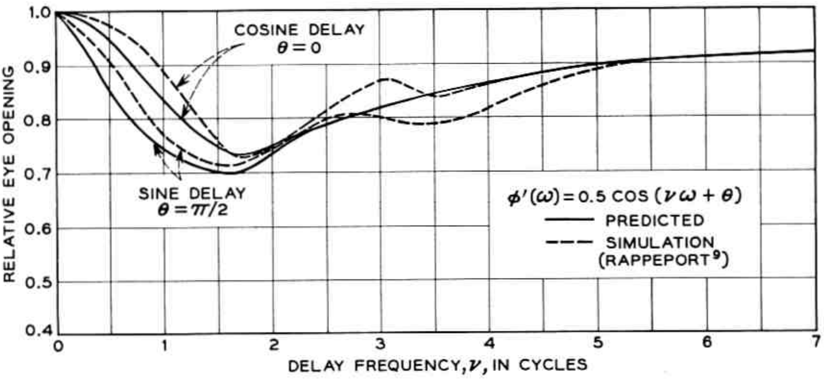


Fig. 15 — The performance of a four-phase system for sine and cosine delay.

excellent. Two other curves are also drawn in Fig. 15 representing the system performance for cosine delay variation previously shown in Fig. 5. Notice that the slight difference in rate of deterioration of performance at low frequency between sine and cosine delay is correctly predicted by the mathematical model.

To test the mathematical model with delay which has both odd and even components, we turn to published results concerning a different data system. This system is an amplitude-modulated system investigated by computer simulation by R. A. Gibby in Ref. 8. Again our criterion is not expected to hold exactly, since Gibby's binary system is an on-off system rather than bipolar. Gibby considers delay of the form

$$\varphi'(\omega) = \alpha \cos (b\omega + \theta) \tag{206}$$

and plots loci of constant eye aperture (constant distortion) on a polar diagram of delay amplitude α and phase θ for a given value of delay frequency b. The quadrature components of the distortion for the delay (206) are

$$D_r = \frac{2}{\pi} \alpha \cos \theta \sum_{n=1}^{\infty} |\, (f_{rn}, \cos b\omega) \,| \tag{207}$$

$$D_q = \frac{2}{\pi} \alpha \sin \theta \sum_{n=1}^{\infty} |\, (f_{qn}, \sin b\omega) \,| . \tag{208}$$

The integrations are performed and summed to give (for $b = 1.5$)

$$D_r = 0.508\, \alpha \cos \theta \tag{209}$$

$$D_q = 0.590\ \alpha \sin \theta \tag{210}$$

$$D^2 = 0.258\ [\alpha \cos \theta]^2 + 0.348\ [\alpha \sin \theta]^2. \tag{211}$$

Therefore, lines of constant distortion are ellipses on a polar chart of α and θ. Fig. 16 shows two of these ellipses of constant distortion with $b = 1.5$ along with the corresponding curves obtained by Gibby.[8]

Figs. 15 and 16 demonstrate that the mathematical model has provided a good description of the behavior of two diverse modulation systems under the influence of delay distortion.

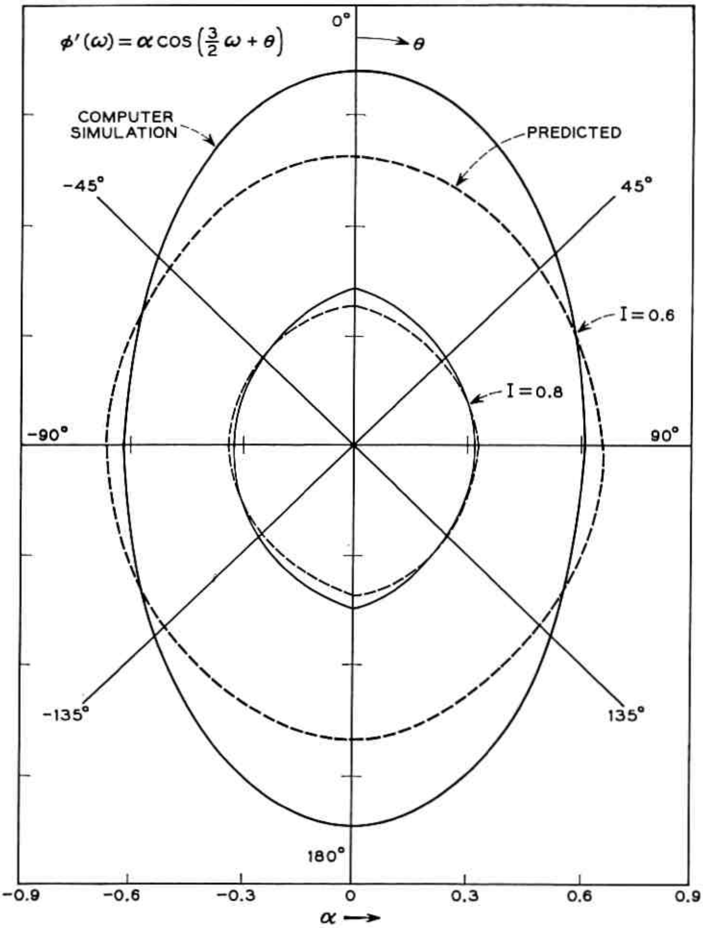


Fig. 16 — Loci of constant eye aperture for cosine delay.

5.2.3 *Sensitivity and Bounds on Distortion*

In the low-pass analysis we found the system sensitivity, which we defined as the maximum achievable distortion for one rms unit of delay, by summing the functions $\pm f_n(\omega)$ in such a fashion as to produce a combination of greatest energy. Obviously, in the bandpass case we can bound both distortion components in like fashion. The system will have a certain sensitivity to even delay and a certain sensitivity to odd delay. Any delay function can be divided into its odd and even components and these components are orthogonal. The contribution to the distortion from each is bounded by the system sensitivities to odd and even delay. Now we ask for a given rms value of delay, how should the delay energy be divided between odd and even components such that the distortion is maximized? Naturally, all the delay energy should be put into the component (odd or even) which has greatest sensitivity to delay distortion. Therefore the over-all system sensitivity is the maximum of the pair of odd and even delay sensitivities.

The sensitivity to even delay is the same as the sensitivity calculated previously for low-pass systems:

$$\text{even sensitivity} = 1.15. \tag{212}$$

For the sensitivity to odd delay, we find the maximum energy combination of the functions $\pm f_{qn}(\omega)$. This is found trivially as the sum of the functions $f_{qn}(\omega)$, since all the terms add in phase in this sum

$$f_{q\text{mec}}(\omega) = \sum_{n=1}^{\infty} f_{qn}(\omega) \tag{213}$$

$$f_{q\text{mec}}(\omega) = \frac{1}{2} \sum_{n=1}^{\infty} \frac{\sin n\omega}{n} - \frac{\sin \omega}{4} \tag{214}$$

$$f_{q\text{mec}}(\omega) = \begin{cases} \frac{1}{4}(\pi - \omega - \sin \omega) & \omega \geqq 0 \\ \frac{1}{4}(-\pi - \omega - \sin \omega) & \omega < 0. \end{cases} \tag{215}$$

This particular form of delay is the worst odd delay for a given rms value and is shown in Fig. 17. The norm of this function is

$$\| f_{q\text{mec}} \| = \frac{1}{2} \sqrt{\frac{\pi^3}{6} - \frac{3\pi}{4}} = 0.835 \tag{216}$$

and the sensitivity becomes

$$\text{odd sensitivity} = 2 \sqrt{\frac{2}{\pi}} \, \| f_{q\text{mec}} \| = 1.337. \tag{217}$$

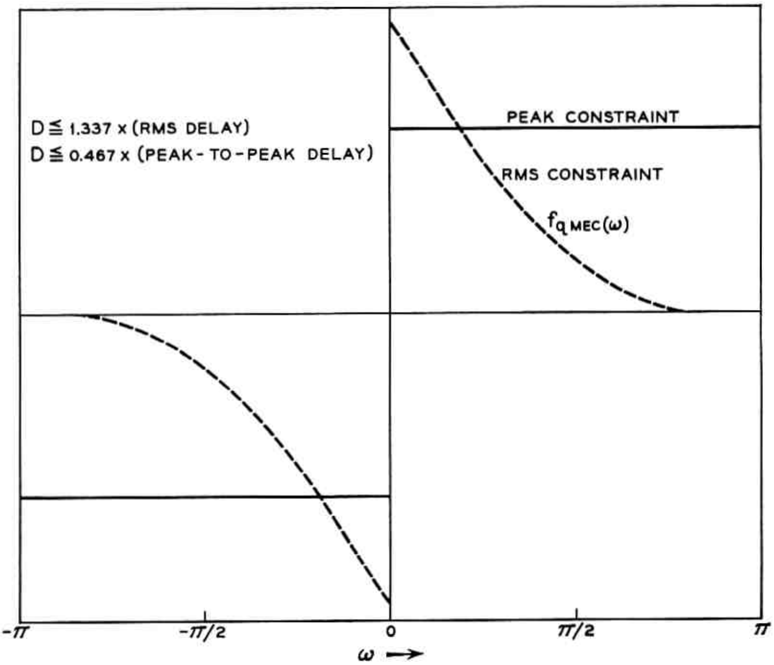


Fig. 17 — Bounds on distortion for bandpass, raised cosine systems.

Thus we see that the bandpass, raised cosine system is somewhat more sensitive to odd delay than it is to even delay. The over-all system sensitivity is therefore equal to the odd sensitivity and consequently we have

$$D \leqq 1.337 \times (\text{rms delay}). \tag{218}$$

Also, the delay curve $f_{q\text{mec}}(\omega)$ in Fig. 17 becomes the worst shape a delay can assume for a given rms value of delay.

For a peak-to-peak constraint on delay, the technique for bounding the distortion is less obvious. Clearly we can find the combinations of signs $\{\epsilon_{rn}\}$ and $\{\epsilon_{qn}\}$ such that the resulting functions

$$f_r(\omega) = \sum_{n=1}^{\infty} \epsilon_{rn} f_{rn}(\omega) \qquad \epsilon_{rn} = \pm 1 \tag{219}$$

$$f_q(\omega) = \sum_{n=1}^{\infty} \epsilon_{qn} f_{qn}(\omega) \qquad \epsilon_{qn} = \pm 1 \tag{220}$$

have maximum absolute integrals and the distortion may be as large as

TABLE II—AN APPROXIMATE SET OF ORTHONORMAL BASIS FUNCTIONS FOR THE SPACE G OF DISTORTIONLESS FUNCTIONS FOR BANDPASS, RAISED COSINE SYSTEMS

$$(\sqrt{\pi}\, g_{rn} = a_1 \cos \omega + a_2 \cos 2\omega + \cdots + a_{11} \cos 11\, \omega)^*$$
$$(\sqrt{\pi}\, g_{qn} = b_1 \sin \omega + b_2 \sin 2\omega + \cdots + b_{11} \sin 11\, \omega)$$

	g_1	g_2	g_3	g_4	g_5	g_6
b_1	0.7205					
b_2	−0.6674	−0.2715				
b_3	−0.1593	0.8145				
b_4	0.0947	−0.4841	−0.3294			
b_5	0.0288	−0.1473	0.8234			
b_6	−0.0151	0.0773	−0.4321	−0.3540		
b_7	−0.0051	0.0258	−0.1444	0.8260		
b_8	0.0025	−0.0128	0.0713	−0.4078	−0.3687	
b_9	0.0009	−0.0045	0.0252	−0.1443	0.8294	
b_{10}	−0.0004	0.0021	−0.0116	0.0664	−0.3819	−0.4138
b_{11}	−0.0002	0.0009	−0.0053	0.0302	−0.1737	0.9104

* For values of a_n see Table I.

the greater of these absolute integrals. The question is, can we do better than this by using a delay with both odd and even components?

If we were to use both odd and even components in the delay, the only acceptable strategy for maximizing distortion would be to use odd delay ($\varphi'(\omega) = -\varphi'(-\omega)$) when $| f_q(\omega) | > | f_r(\omega) |$ and even delay ($\varphi'(\omega) = \varphi'(-\omega)$) when $| f_r(\omega) | > | f_q(\omega) |$.* In addition we would have to run through all possible sequences of the signs $\{\epsilon_{rn}\}$ and $\{\epsilon_{qn}\}$. This procedure was carried out to the extent of time limitations on the IBM 7090 digital computer with the result that the best such delay has distortion less than a delay using an all odd strategy.

The maximum absolute integral of $f_q(\omega)$ is obtained by using $\epsilon_{qn} = +1$, i.e., by simply adding all the functions $f_{qn}(\omega)$. We found this function previously as $f_{q\text{mec}}(\omega)$

$$D = D_q \leqq \frac{2}{\pi} \varphi'_{\max} \int_{-\pi}^{+\pi} \left| \sum_{n=1}^{\infty} f_{qn}(\omega) \right| d\omega \tag{221}$$

$$D \leqq \frac{2}{\pi} \varphi'_{\max} \int_{0}^{+\pi} | \tfrac{1}{2}(\pi - \omega - \sin \omega)| \, d\omega \tag{222}$$

$$D \leqq \left(\frac{\pi}{2} - \frac{2}{\pi}\right) \varphi'_{\max} \tag{223}$$

* This is not exactly true, but an exact proof here does not seem to be worth the considerable effort involved.

$$D \leqq 0.467 \times (\text{peak-to-peak delay}). \qquad (224)$$

The worst peak-to-peak delay is simply positive when $f_{qmec}(\omega)$ is positive and negative when $f_{qmec}(\omega)$ is negative. This is a particularly simple delay function which is $+\varphi'_{\max}$ for $\omega \geqq 0$ and $-\varphi'_{\max}$ for $\omega < 0$. This function is also shown in Fig. 17.

5.2.4 *Zero-Distortion Delay Functions*

A space G of distortion-free delay functions for raised cosine systems may be obtained using techniques similar to the low-pass methods. After orthonormalizing the sequences $\{f_{rn}\}$ and $\{f_{qn}\}$, we find the orthonormal

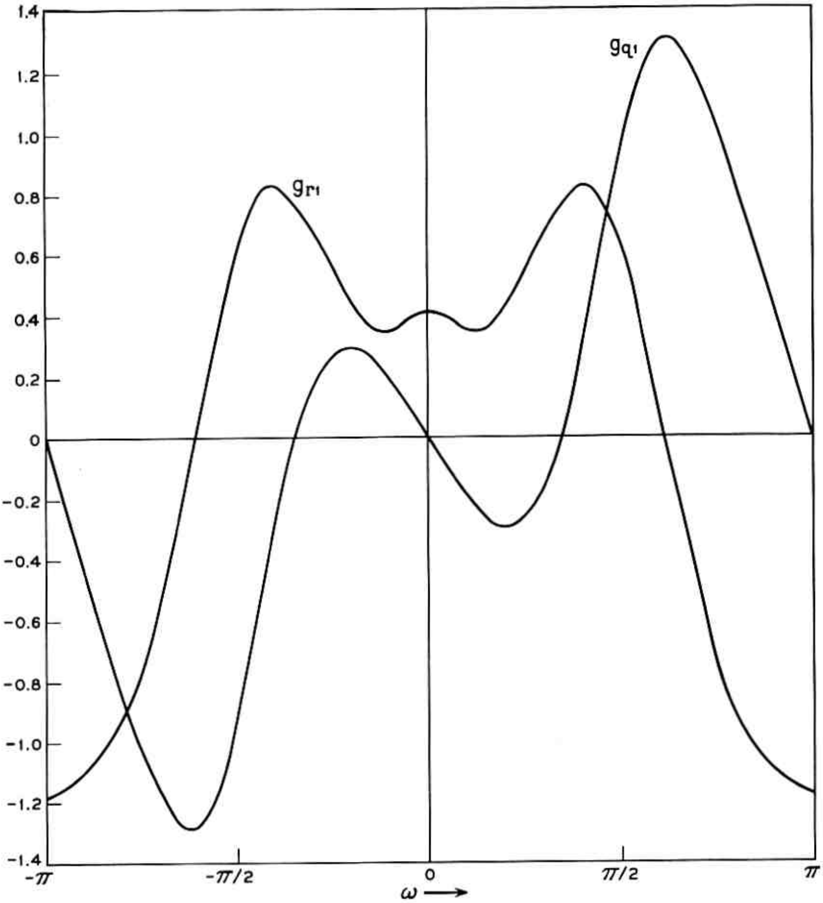


Fig. 18 — Bandpass distortionless delays.

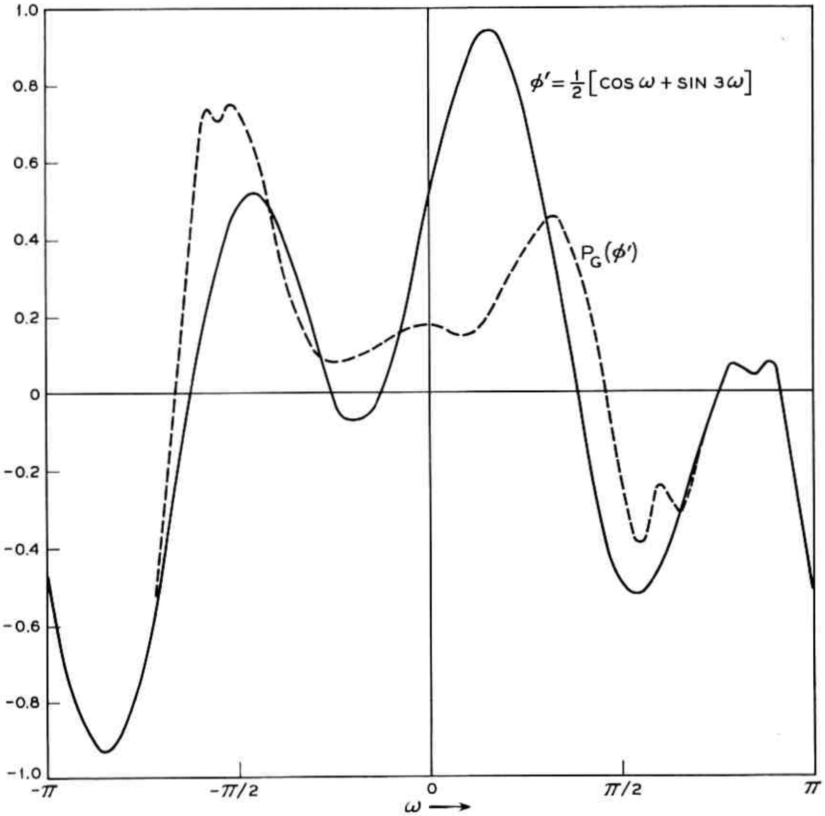


Fig. 19 — An example delay and the nearest distortion-free delay.

functions which complete these sequences in their respective subspaces of even and odd square-integrable functions. Thus we derive the sequences $\{g_{rn}\}$ and $\{g_{qn}\}$ of even and odd functions which span the distortionless space G. Any delay φ' in $L^2(-\pi, +\pi)$ can then be expanded in terms of the functions f_{rn}, f_{qn}, g_{rn}, and g_{qn} with the terms involving g functions comprising the projection of φ' upon G and yielding zero distortion and the terms involving f functions containing all the distortion content of φ'

$$\begin{aligned}\varphi' = \sum_{n=1}^{\infty} b_{rn} f_{rn} + \sum_{n=1}^{\infty} b_{qn} f_{qn} \\ \text{even} \qquad\qquad \text{odd} \\ + \sum_{n=1}^{\infty} c_{rn} g_{rn} + \sum_{n=1}^{\infty} c_{qn} g_{qn}\end{aligned} \tag{225}$$

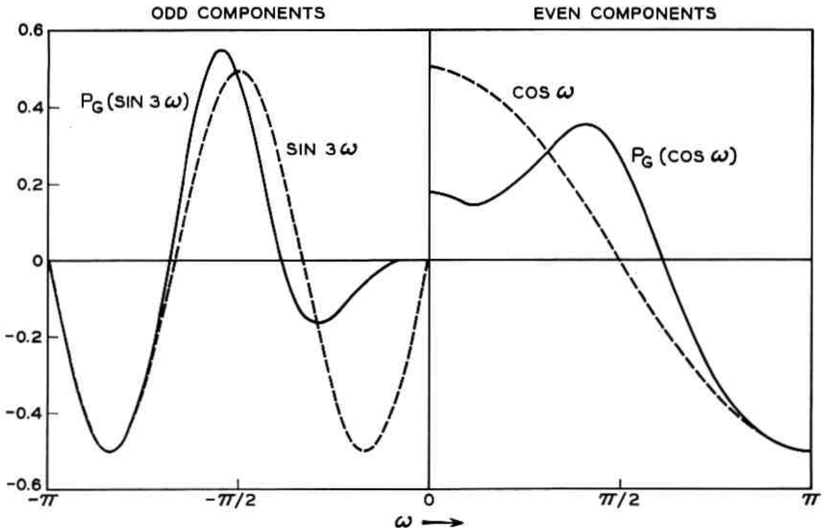


Fig. 20 — Resolution of the example delay into odd and even components.

$$P_G(\varphi')$$

$$c_{rn} = (\varphi', g_{rn}) \qquad (g_{rn}, g_{rm}) = 0 \; n \neq m \qquad (226)$$

$$c_{qn} = (\varphi', g_{qn}) \qquad\qquad = 1 \; n = m.$$

A list of the functions g_{rn} and g_{qn} obtained for raised cosine shaping is given in Table II, and the first few functions are illustrated in Fig. 18.

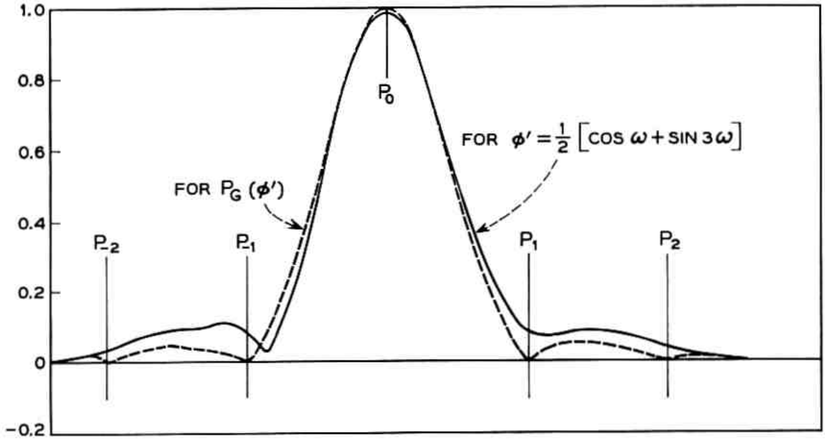


Fig. 21 — Impulse response envelopes for the corrected and uncorrected delays.

It is necessary here to reiterate the fact that functions $g \subset G$ have zero distortion only to the extent of the approximations employed in obtaining the fundamental relationship for distortion in terms of the sequence of linear functionals. As an example computation a delay $\varphi' = \frac{1}{2}[\cos \omega + \sin 3\omega]$ is shown in Fig. 19. The odd and even components of this delay and their respective projections on G are shown in Fig. 20. Combining these projection components the total projection is shown in Fig. 19 back with the original delay function. The envelopes of the impulse responses of the corrected and uncorrected delays are illustrated finally in Fig. 21. As may be seen from this figure, the correction is near perfect. The corrected envelope approaches zero at each sample point as close as the numerical integration techniques employed permit.

REFERENCES

1. Sunde, E. D., Theoretical Fundamentals of Pulse Transmission, B.S.T.J., **33,** May, 1954, pp. 721–787.
2. Sunde, E. D., unpublished work.
3. Sunde, E. D., Ideal Binary Pulse Transmission by AM and FM, B.S.T.J., **38,** Nov., 1959, p. 1357.
4. Sunde, E. D., Pulse Transmission by AM, FM, and PM in the Presence of Phase Distortion, B.S.T.J., **40,** Mar., 1961, p. 353.
5. Alexander, A. A., Gryb, R. M. and Nast, D. W., Capabilities of the Telephone Network for Data Transmission, B.S.T.J., **39,** May, 1960, p. 431.
6. Akheizer, N. I., and Glazman, I. M., *Theory of Linear Operators in Hilbert Space*, Ungar Publishing Company, New York, 1961.
7. Rappeport, M. A. and Gibby, R. A., Data Transmission over Channels with Sinusoidal Delay, General Meeting A.I.E.E., New York, Winter 1963.
8. Gibby, R. A., An Evaluation of AM Data System Performance by Computer Simulation, B.S.T.J., **39,** May, 1960, p. 675.

Contributors to This Issue

William R. Bennett, B.S. in E.E., 1925, Oregon State University; M.A. in Physics, 1928, Ph.D., 1949, Columbia University; Bell Telephone Laboratories, 1925—. His early work was concerned with low-frequency transmission over wires and cables. He later became associated with the first coaxial carrier project and made basic studies on noise and distortion in broadband amplifiers. Time division multiplex and pulse code modulation were areas of subsequent major interest. He is now Head of the Data Theory Department in the Data Communications Development Laboratory. Fellow, IEEE; member, American Physical Society, U.R.S.I., Sigma Xi, Tau Beta Pi and Eta Kappa Nu.

W. S. Brown, B.S., 1956, Yale University; Ph.D., 1961, Princeton University; Bell Telephone Laboratories, 1961—. Since joining the Laboratories Mr. Brown has been working on the theoretical and practical problems of symbolic computing. Member, Amer. Phys. Society, Association for Computing Machinery, Phi Beta Kappa, Sigma Xi, and American Association for the Advancement of Science.

C. Chapin Cutler, B.S., 1937, Worcester Polytechnic Institute, Bell Telephone Laboratories, 1937—. He has made significant contributions in the areas of microwave antennas, microwave tubes, and new radar and communication systems. As Director, Electronic Systems Research, he heads a group which has worked on communications engineering for both the Project Echo and Project Telstar satellite communications experiments. Fellow, IEEE.

C. Dragone, M.S. (Electrical Engineering), 1961, Padua University (Italy); Bell Telephone Laboratories, 1961—. During his stay in the radio research department at Holmdel, Mr. Dragone worked mainly on microwave antenna problems. At the end of the period he was working on frequency multipliers using microwave varactors.

Edwin O. Elliott, A.B., 1949, M.A., 1951, Ph.D., 1959, University of California, Berkeley; Operations Evaluation Group of MIT, 1954–1958; Stanford Research Institute, 1958–1959; Assistant Professor of Mathematics, University of Nevada, Reno, 1959–1960; Bell Telephone Laboratories, 1960—. At the Laboratories he has been engaged in mathematical analysis of error-control methods for digital data communication systems and in the application of measure-theoretic techniques in the study of stochastic processes. He has recently worked on problems in the congestion theory of traffic for a model of a store-and-forward data communications network. Member, American Mathematical Society, Operations Research Society of America, Pi Mu Epsilon, Sigma Xi and Phi Beta Kappa.

M. J. Evans, B.S.E.E., 1957, University of Utah; M.E.E., 1959, New York University; Bell Telephone Laboratories, 1957—. His work at BTL has included the analysis of ballistic missile control systems and the development of guidance equations for ballistic missiles. He has specialized in applying the BTL command guidance system to space missions. Member, Phi Kappa Phi, Eta Kappa Nu and Tau Beta Pi.

Harvey J. Fletcher, B.S., 1944, Massachusetts Institute of Technology; M.S., 1948, California Institute of Technology; Ph.D., 1953, University of Utah; Bell Telephone Laboratories, 1961–1962; Bellcomm, Inc., 1963—. At the Laboratories, he engaged in the analysis of attitude control of a two-body gravitationally oriented satellite. He has worked on the analysis of lunar trajectories at Bellcomm, Inc. Member, The Mathematical Association of America.

S. D. Hathaway, B.E.E., 1947, University of Virginia; M.S.E.E., 1950, Virginia Polytechnic Institute; M.S.E.E., 1952, University of Illinois. Bell Telephone Laboratories, 1952—. He has been engaged in systems engineering on microwave radio relay systems, including studies of the effects of rainfall on radio transmission. At present he supervises a group working on short-haul systems. Member, IEEE, Eta Kappa Nu and Tau Beta Pi.

D. C. Hogg, B.Sc., 1949, University of Western Ontario; M.Sc., 1950, and Ph.D., 1953, McGill University; Bell Telephone Laboratories, 1953—. His work has included studies of artificial dielectrics for microwaves, diffraction of microwaves, and over-the-horizon and millimeter wave propagation. He has been concerned with evaluation of sky noise, analysis of performance characteristics of microwave antennas and, most recently, propagation of optical waves. Senior member, IEEE; member, Commission 2, U.R.S.I., and Sigma Xi.

Rudolf Kompfner, Diplom. Ingenieur, Technische Hochschule, Vienna, 1933; Ph.D., Oxford, 1951; Bell Telephone Laboratories, 1951—.

Dr. Kompfner invented the traveling-wave tube while at Birmingham University during World War II. At Bell Laboratories, he has specialized in microwave electronics, work which has more recently been enlarged to include research on quantum electronics and satellites communications. Director of Electronics Research, 1955; Director of Electronics and Radio Research, 1957; Associate Executive Director, Research, Communications Systems Research Division, 1962. Fellow, IEEE, 1950; Duddell Medal, Physical Society, 1955; A.I.E.E. David Sarnoff Award, 1960; Franklin Institute Stuart Ballentine medal, 1960.

Robert W. Lucky, B.S.E.E., 1957, M.S.E.E., 1959, Ph.D., 1961, Purdue University; Bell Telephone Laboratories, 1961—. Mr. Lucky has been concerned with various analysis problems in the area of digital data communications. Member, IEEE, Sigma Xi, Tau Beta Pi and Eta Kappa Nu.

Miss K. L. McAdoo, B.A., 1956, Wilson College; Bell Telephone Laboratories, 1956—. Miss McAdoo has largely been engaged in programming aspects of simulating exchange area facilities on the computer and also speech volume studies. At present, she is involved in the means of improving exchange area transmission performance through loop impedance compensators.

John D. Musa, A.B., 1954, M.S., 1955, Dartmouth College; Bell Telephone Laboratories, 1958—. He has been involved in various types of work in military systems engineering some of which have involved application of data smoothing. He has taught an out-of-hours course in data smoothing. Currently, he is engaged in radar data processing systems engineering work.

George H. Myers, S.B., S.M., 1952. Massachusetts Institute of Technology; Eng. Sc. D., 1959, Columbia University; Bell Telephone Laboratories, 1952—. He has worked on development of analog and digital computers for automatic control including the computer for the K-5 bombing-navigation system, the TRADIC digital computer and the Terrier fire control system. Recently, he has been concerned with guidance equations for space vehicles. Senior member, IEEE; member, AIAA, Sigma Xi.

Winston L. Nelson, B.S., 1950, University of Utah; M.S., 1953, and Ph.D., 1959, Columbia University; Bell Telephone Laboratories, 1960—. He has been engaged in research in optimum control theory, particularly satellite attitude control and satellite tracking systems. He has also worked on weak-signal detection techniques employing feedback and at present is studying problems in stochastic estimation and control. Member, Sigma Xi, Society for Industrial and Applied Mathematics, and IEEE.

E. A. Ohm, B.S., 1950, M.S., 1951, Ph.D., 1953, University of Wisconsin; Bell Telephone Laboratories, 1953—. Mr. Ohm has worked on circulators, isolators, microwave filters and channel branching networks. He has also been concerned with the measurement of sky temperature and the development of waveguide parts for ultra low-noise receiving systems. He was the assistant project engineer, and responsible for the receiving system at Bell Laboratories during the Project Echo communications experiments. At present, he is working on antenna and system problems of a satellite steerable array repeater. Member, IEEE, Sigma Xi and Tau Beta Pi.

B. Paul, B.S.E. (Mechanical Engineering), 1953, Princeton University; M.S. (Engineering Mechanics), 1954, Stanford University; Ph.D. (Engineering Mechanics), 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1961–1963; Ingersoll-Rand Research Labs., 1963—. He has worked on problems connected with stress, vibration, heating and rigid-body dynamics of passively oriented communications satellites. He has also worked on the problem of reaction forces associated with sublimation of solids into high vacuum. Member, A.S.M.E., A.I.A.A., Sigma Xi and Phi Beta Kappa.

Stephen O. Rice, B.S., 1929, D.Sc. (Hon.), 1961, Oregon State College; Graduate Studies, California Inst. of Tech., 1929–30 and 1934–35; Bell Telephone Laboratories, 1930—. In his first years at the Laboratories, Mr. Rice was concerned with nonlinear circuit theory, especially with methods of computing modulation products. Since 1935 he has served as a consultant on mathematical problems and in investigation of telephone transmission theory, including noise theory, and applications of electromagnetic theory. He was a Gordon McKay Visiting Lecturer in Applied Physics at Harvard University for the Spring, 1958, term. Fellow, IEEE.

L. Rongved, B.S. (Civil Engineering), 1950, M.S. (Civil Engineering), 1952, Ph.D. (Theoretical Mechanics), 1954, Columbia University; Bell Telephone Laboratories, 1956–1962; Bellcomm, Inc., 1962—. He has been engaged in theoretical problems in design of electron devices for high shock and vibration environments. He also worked on metal-ceramic seal problems and made several contributions to the thermal and mechanical design and testing for the Telstar satellite. He was supervisor of the mechanics exploratory group. Member, Executive Committee of Industrial and Professional Advisory Council.

DONALD D. SAGASER, B.S.E. (EE), 1948, University of Michigan; Bell Telephone Laboratories 1948—. His early work was on development of short-haul carrier systems. Assignments in development areas concerned with negative impedance repeaters, submarine cable, short haul microwave TV transmission systems and mobile radio systems preceded the project on TL radio. Mr. Sagaser is presently responsible for circuit development activities on short-haul carrier systems. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

J. SALZ, B.S.E.E., 1955, M.S.E., 1956, Ph.D., 1961, University of Florida; The Martin Company, 1958–1960; Bell Telephone Laboratories, 1961—. He first worked on the remote line concentrators for the electronic switching system. He has since engaged in theoretical studies of data transmission systems. Member, IEEE; associate member, Sigma Xi.

IRWIN W. SANDBERG, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. He has been concerned with analysis of military systems, particularly radar systems, and with synthesis and analysis of active and time-varying networks. He is currently involved in a study of the signal-theoretic properties of nonlinear systems. Member, IEEE, Eta Kappa Nu, Sigma Xi and Tau Beta Pi.

WILLIAM W. SNELL, JR., Bell Telephone Laboratories, 1955—. His early work for the radio research department centered around waveguide components for use in the 4-, 6- and 11-kmc common carrier bands: ferrite devices, microwave diode detectors and polarization couplers. He later participated in the Shotput experiments, suborbital proving tests for Project Echo. During Project Echo he operated the Crawford Hill receiving system. Presently, he is working on strip line components for a proposed communications satellite.

LEROY C. TILLOTSON, B.S.E.E., 1938, University of Idaho; M.S.E.E., 1940, University of Missouri; Bell Telephone Laboratories, 1941—. Mr. Tillotson's early work included design of filters and networks; he has since been concerned with microwave radio relay systems. From June, 1958, to July, 1959 he served as a member of technical staff of the Advanced Research Projects Agency division of the Institute for Defense Analyses. As Director, Radio Research, he is presently engaged in research on microwave and optical communications.

J. W. Timko, B.S.E.E., 1951, Rutgers University; M.S.E.E., 1952, Yale University; Bell Telephone Laboratories, 1952—. He first was engaged in the design and development of the analog computer control system for the AN/MSG-3 fire control system. From 1957–1960 he worked on the ground radar receiver and tracking circuits for the WS-107A-2 ground guidance system. From 1961 to the present, he has been supervisor of a group responsible for the preparation of guidance equations for use with the WS-107A-2 system in the guidance of launch vehicles for space satellites. Member, Tau Beta Pi, Eta Kappa Nu and Sigma Xi.

J. W. West, B.S. (Physics), 1946, City College of New York; Bell Telephone Laboratories, 1930—. His early work was in the field of electron tube design and development. He later headed groups concerned with the mechanical development of microwave tubes and semiconductors. For 10 years he was associated with the final development device activity at Bell Laboratories branches of the Western Electric Company. His department was responsible for the mechanical development, attitude control and thermal design of the Telstar satellite. At present, his department is concerned with satellite attitude control studies, engineering mechanics studies associated with device work and mechanical development of microwave tubes and parametric amplifiers. Member, AIAA, Soc. for Experimental Stress Analysis.

John A. Word, B.S., 1930, University of California (Berkeley); Bell Telephone Laboratories, 1930—. Prior to World War II he worked on the design of toll terminal room equipment. During World War II he worked on sonar, communications counter measures and microwave radio. At present, he supervises a group in the equipment design of short- and long-haul microwave radio systems. Member, Tau Beta Pi and Eta Kappa Nu; associate member, Sigma Xi; senior member, IEEE.

Er-Yung Yu, M.S. (Mechanical engineering), 1957, Washington University; Ph.D. (Engneering mechanics), 1960, Stanford University; Bell Telephone Laboratories, 1960—. Mr. Yu has been engaged in exploratory mechanics studies in problems of passive attitude control of satellites. His work has included studies of magnetic orientation of medium-altitude communications satellites and the related damping problems. He also participated in the Telstar satellite dynamics analysis and precession damper design. At present, he is working on the system design and dynamics analysis of a two-body gravitationally oriented satellite. Member, Sigma Xi and A.I.A.A.

B.S.T.J. BRIEFS

A Self-Reorganizing Synchronization Network

By J. V. SCATTAGLIA

(Manuscript received July 5, 1963)

This paper describes a method of increasing the reliability of synchronization in a network of remote clocks. A common synchronization technique embodies the master-slave relation where one particular clock is a reference for the others. The geometry of a typical network can be likened to a "tree" structure. The master clock transmits its signal, simultaneously, over several transmission links to synchronize slave oscillators at the ends. The slaves retransmit the reference signal to other slaves one link away. This process is iterated as the system expands. Each slave has only one input; hence, a network of this type can be disabled, in varying degree, by the failure of any clock or transmission link.

The probability of failure can be reduced by creating redundant paths between nodes in the network. These paths can consist of more than one link and include the intervening clocks. This allows bypassing of a disabled clock or reorganization of synchronization authority.

To implement such a system we need interrogation equipment at each clock station to decide which incoming path has the highest priority for a given situation. If the reorganization takes place automatically, we can refer to it as a "self-reorganizing synchronization network." Application of this scheme in a large "tree" network can become very complex.

A technique which provides orderly organization of complex networks was patented by G. P. Darwin and R. C. Prim (Patent No. 2,986,723). Their system, in its most general form, requires a three-part "signature" to be transmitted, along with the synchronizing signal, from each local clock.

The technique proposed here simplifies the system controls by constraining network geometry to advantageous geometrical patterns. These patterns are subject to predetermined reorganizational rules. The advantages offered are: (*i*) simpler interrogation equipment; (*ii*) adaptability to multilevel systems, i.e., subgroup, supergroup, regional, local, etc.; (*iii*) more predictable reliability; (*iv*) large networks can be designed in orderly blocks.

The specific geometrical network to be described here will be referred to as the "wheel."

Wheel Network. Fig. 1(a) illustrates a wheel configuration, e.g., a regional network. Clock M, the primary master of the wheel, would be the regional master. Peripheral units A to F are secondary or local clocks. M is likely to have more stringent requirements than A to F; hence it has only outgoing paths. A to F are identical clocks and neighbors are connected by two-way paths.

M can be connected by two-way paths to adjacent regional masters to become a peripheral clock in a larger wheel.

The set of numbers at each input to a clock denotes alternate priority values of that path. The existing value is a reflection of the transmitting clock's source of synchronization. A description indicating the input-output relations for all conditions is too lengthy for this short paper.

Examples of path structures for several conditions are illustrated in Figs. 1(b) to 1(e). The presiding master clocks and each path's priority value are indicated. Clock A has first preference in becoming master of

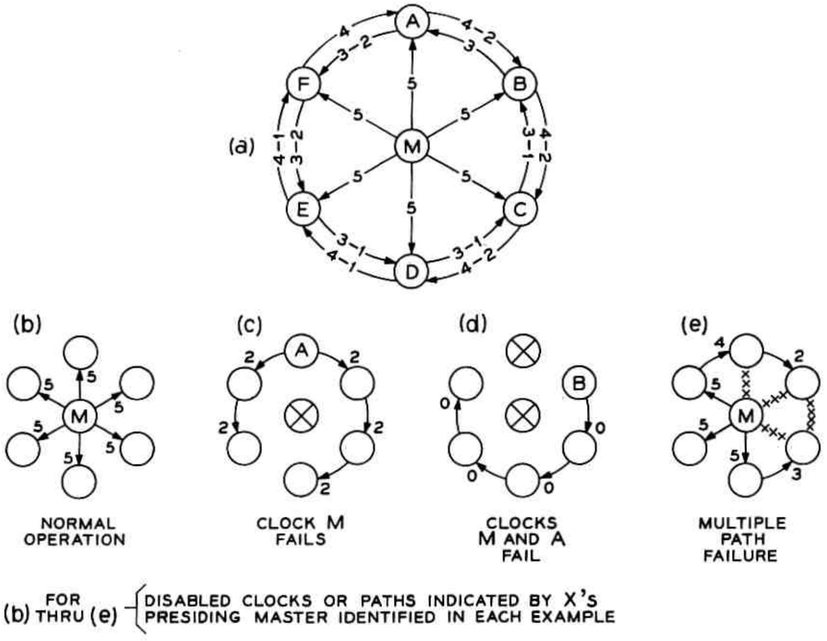


Fig. 1 — Self-reorganizing wheel network. In parts (b) through (e), the disabled clocks or paths are indicated by x's; the presiding master is indicated in each example.

the peripheral clocks upon failure of the primary master. A discontinuity in the priority ratings of signals "passing through" A prevents "closed loop" conditions in the periphery that could otherwise occur on failure of the primary master.

Interrogation Circuitry. This network is very simple to implement because there is a maximum of only two "modes" to be identified at any input. The essentials for a typical station, clock "C" for example, are shown in Fig. 2. The illustration assumes a dc transmission path. The presence of a signal on a path is indicated by a dc bias added to the synchronization signal at the transmitter. At the receiver the dc operates a particular relay. The two modes of a peripheral input are indicated by a positive or negative bias. If the primary master input is active, relay switch (RS) #5 closes contact 5A. The master synchronization signal is thus directed to the local clock's comparator. Contact 5A simultaneously disconnects all other inputs from the comparator. The remaining function of relay #5 is to close contacts 5B and 5C which will add the appropriate dc bias to the outgoing signals. If now the master input is absent, the priority chain looks for the next highest input

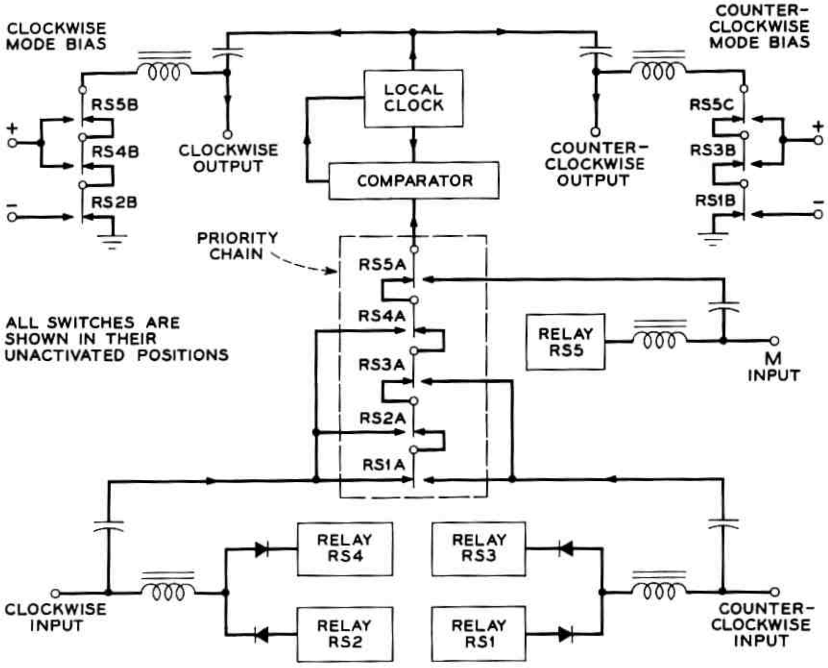


Fig. 2 — Clock station "C."

present. The clockwise rule to cover the situation shown in Fig. 1(d) is implemented by connecting the clockwise ac input to the unactivated contact on 1A. If none of the inputs is present, the local clock free runs but can still synchronize its clockwise neighbor.

Description of a dc system was for simplicity of illustration. Other indications of priority could be employed: e.g., tone modulation with tuned reed relays for ac analog systems and simple codes for digital systems.

Point-Contact Wafer Diodes for Use in the 90- to 140-Kilomegacycle Frequency Range

By W. M. SHARPLESS

(Manuscript received July 11, 1963)

In millimeter wave systems, one of the most important components is the first converter or mixer. This brief paper describes a recently developed point-contact diode of the wafer type which operates efficiently as a first converter in the frequency range 90 to 140 kilomegacycles (F-band).

Fig. 1 is a photograph of the wafer diode and its holder. The assembly is quite similar in appearance to the diode-holder combination designed for the 45- to 75-kmc range.[1] The tuning procedure is also the same. The wafer is inserted in the holder and moved transversely to the waveguide, thereby adjusting the location of the point-contact relative to the guide to effect a resistive match. (The pin at the left of the wafer slides in a chuck on the inner conductor of the coaxial low-frequency output circuit.) The wafer is then locked in position by means of the knurled clamping knob, and the reactance of the diode is tuned out with the waveguide piston at the rear of the holder.

The present design differs from the older one in the use of smaller waveguide (RG 138/u instead of RG 98/u) and, most importantly, in the addition of the milled slots on either side of the wafer which encompass the rectangular window containing the point-contact diode. When the wafer is inserted in the holder, these slots engage small guiding shoes which automatically align the window of the wafer with the waveguide sections in the holder to better than 0.0005 inch; this accuracy is essential at the extremely high frequency of operation. The method of forming

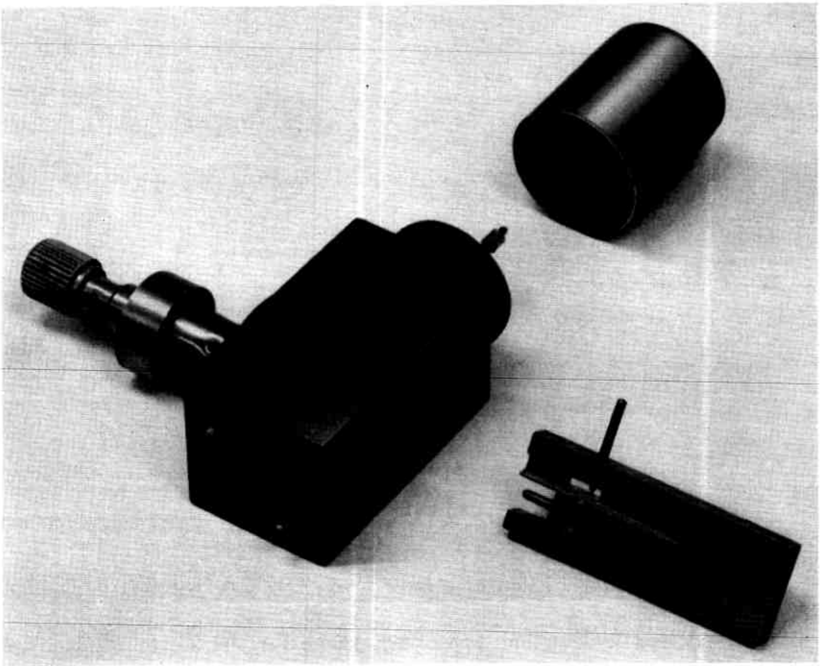


Fig. 1 — Millimeter-wave point-contact wafer diode and holder for use in the 90- to 140-kmc frequency band.

the rectifying junction is also different in the present design. The older units used boron-doped silicon which required that the units be "tapped" into adjustment. The present units use aluminum-doped silicon and do not require tapping. For very high frequency operation, tapping should be avoided if possible since it tends to increase the point-contact area.

The apparatus used to evaluate the diodes is shown in Fig. 2. It constitutes a complete double-detection measuring system. Many of the millimeter wave components shown had to be developed in order to measure the conversion loss of the diodes.

The conversion losses of several types of diodes mounted in the new wafer units are listed in Table I. The measurement consisted of determining the ratio of the millimeter-wave power input to the converter, measured by a calorimeter,[2] to the 60-mc output power measured by comparison with a known signal level obtained from a calibrated signal generator. The conversion loss quoted includes the heat losses of the waveguide input circuits of the diode as well as the losses associated with the output circuitry. The diodes were matched at 115 kmc, were

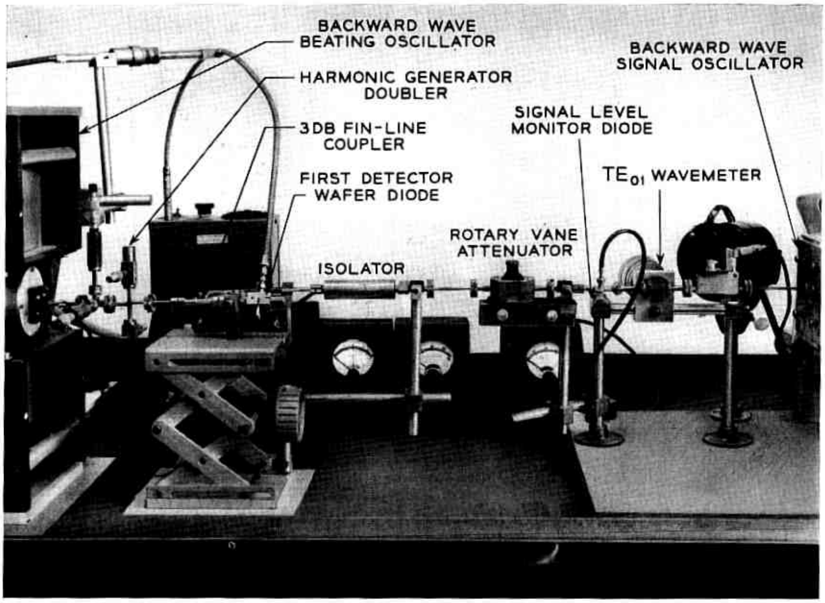

BACKWARD WAVE BEATING OSCILLATOR
BACKWARD WAVE SIGNAL OSCILLATOR
HARMONIC GENERATOR DOUBLER
SIGNAL LEVEL MONITOR DIODE
3DB FIN-LINE COUPLER
FIRST DETECTOR WAFER DIODE
TE_{01} WAVEMETER
ROTARY VANE ATTENUATOR
ISOLATOR

Fig. 2 — A double-detection measuring system for the 90- to 140-kmc frequency band.

optimumly biased, and were driven with 0.6 milliwatt of local oscillator power.

If one assumes a 12-db diode with a noise output ratio, N_R, of 2, which is 30 per cent above the value measured at 55 kmc for similar units, it may be calculated that a balanced converter followed by an IF amplifier with a 4-db noise figure will yield an over-all receiver noise figure of 18 db at 115 kmc. Noise figures very near this value were obtained in practice.

The important contributions of Messrs. E. F. Elbert and S. E. Reed are gratefully acknowledged.

TABLE I — CONVERSION LOSSES

Type of Diode	115-kmc Conversion Loss in db
Average silicon diode (25 units)	12.4
Best silicon diode	11.1
Best gallium arsenide diode	9.9
Best germanium backward diode[3]	11.5

REFERENCES

1. Sharpless, W. M., Wafer-Type Millimeter-Wave Rectifiers, B.S.T.J., **35**, November, 1956, pp. 1385–1402.
2. Sharpless, W. M., A Calorimeter For Power Measurements at Millimeter Wavelengths, IRE Trans. Microwave Theory and Techniques, MTT-**2**, September, 1954, pp. 45–47.
3. Burrus, C. A., Backward Diodes for Low Level Millimeter-Wave Detection, IEEE Trans. Microwave Theory and Techniques, MTT-**11**, September, 1963.

present. The clockwise rule to cover the situation shown in Fig. 1(d) is implemented by connecting the clockwise ac input to the unactivated contact on 1A. If none of the inputs is present, the local clock free runs but can still synchronize its clockwise neighbor.

Description of a dc system was for simplicity of illustration. Other indications of priority could be employed: e.g., tone modulation with tuned reed relays for ac analog systems and simple codes for digital systems.

Point-Contact Wafer Diodes for Use in the 90- to 140-Kilomegacycle Frequency Range

By W. M. SHARPLESS

(Manuscript received July 11, 1963)

In millimeter wave systems, one of the most important components is the first converter or mixer. This brief paper describes a recently developed point-contact diode of the wafer type which operates efficiently as a first converter in the frequency range 90 to 140 kilomegacycles (F-band).

Fig. 1 is a photograph of the wafer diode and its holder. The assembly is quite similar in appearance to the diode-holder combination designed for the 45- to 75-kmc range.[1] The tuning procedure is also the same. The wafer is inserted in the holder and moved transversely to the waveguide, thereby adjusting the location of the point-contact relative to the guide to effect a resistive match. (The pin at the left of the wafer slides in a chuck on the inner conductor of the coaxial low-frequency output circuit.) The wafer is then locked in position by means of the knurled clamping knob, and the reactance of the diode is tuned out with the waveguide piston at the rear of the holder.

The present design differs from the older one in the use of smaller waveguide (RG 138/u instead of RG 98/u) and, most importantly, in the addition of the milled slots on either side of the wafer which encompass the rectangular window containing the point-contact diode. When the wafer is inserted in the holder, these slots engage small guiding shoes which automatically align the window of the wafer with the waveguide sections in the holder to better than 0.0005 inch; this accuracy is essential at the extremely high frequency of operation. The method of forming

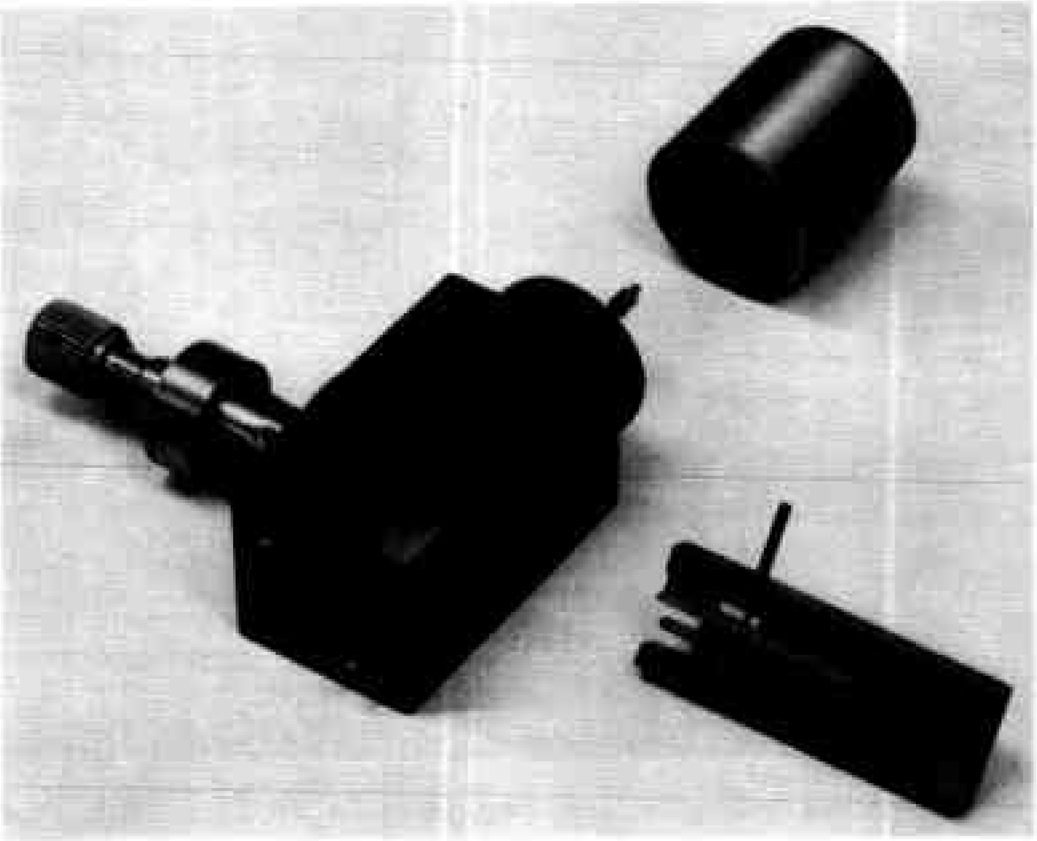


Fig. 1 — Millimeter-wave point-contact wafer diode and holder for use in the 90- to 140-kmc frequency band.

the rectifying junction is also different in the present design. The older units used boron-doped silicon which required that the units be "tapped" into adjustment. The present units use aluminum-doped silicon and do not require tapping. For very high frequency operation, tapping should be avoided if possible since it tends to increase the point-contact area.

The apparatus used to evaluate the diodes is shown in Fig. 2. It constitutes a complete double-detection measuring system. Many of the millimeter wave components shown had to be developed in order to measure the conversion loss of the diodes.

The conversion losses of several types of diodes mounted in the new wafer units are listed in Table I. The measurement consisted of determining the ratio of the millimeter-wave power input to the converter, measured by a calorimeter,[2] to the 60-mc output power measured by comparison with a known signal level obtained from a calibrated signal generator. The conversion loss quoted includes the heat losses of the waveguide input circuits of the diode as well as the losses associated with the output circuitry. The diodes were matched at 115 kmc, were

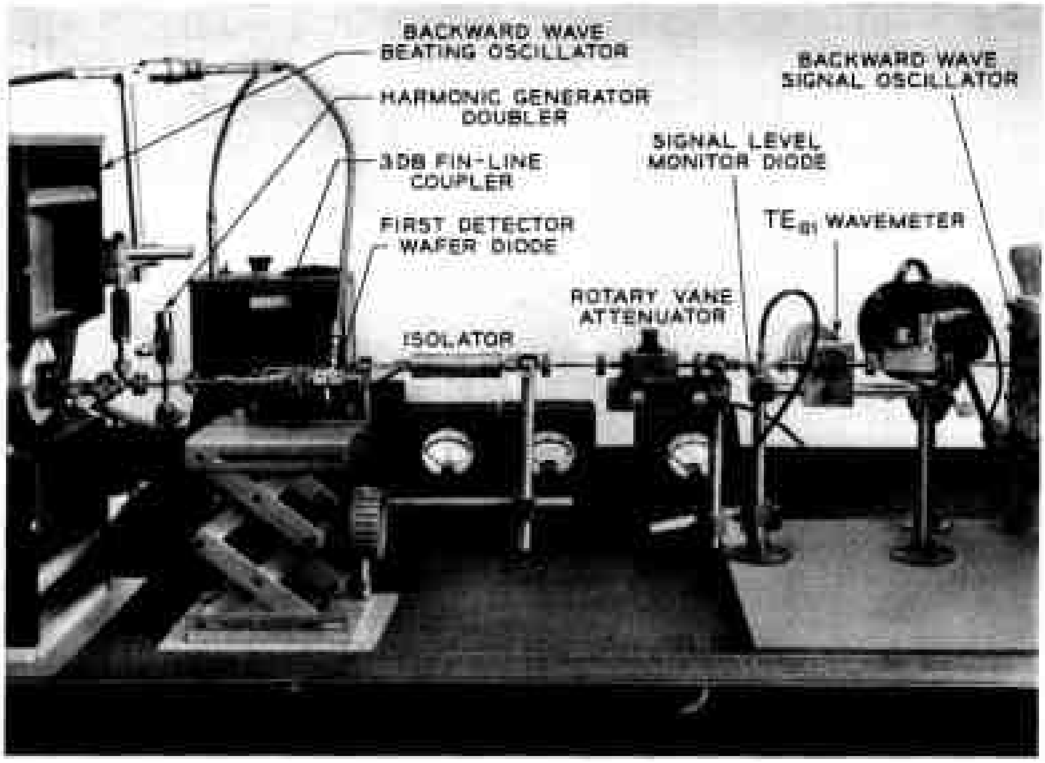

BACKWARD WAVE BEATING OSCILLATOR
HARMONIC GENERATOR DOUBLER
3DB FIN-LINE COUPLER
FIRST DETECTOR WAFER DIODE
ISOLATOR
ROTARY VANE ATTENUATOR
SIGNAL LEVEL MONITOR DIODE
TE$_{01}$ WAVEMETER
BACKWARD WAVE SIGNAL OSCILLATOR

Fig. 2 — A double-detection measuring system for the 90- to 140-kmc frequency band.

optimumly biased, and were driven with 0.6 milliwatt of local oscillator power.

If one assumes a 12-db diode with a noise output ratio, N_R, of 2, which is 30 per cent above the value measured at 55 kmc for similar units, it may be calculated that a balanced converter followed by an IF amplifier with a 4-db noise figure will yield an over-all receiver noise figure of 18 db at 115 kmc. Noise figures very near this value were obtained in practice.

The important contributions of Messrs. E. F. Elbert and S. E. Reed are gratefully acknowledged.

Table I — Conversion Losses

Type of Diode	115-kmc Conversion Loss in db
Average silicon diode (25 units)	12.4
Best silicon diode	11.1
Best gallium arsenide diode	9.9
Best germanium backward diode[3]	11.5

REFERENCES

1. Sharpless, W. M., Wafer-Type Millimeter-Wave Rectifiers, B.S.T.J., **35**, November, 1956, pp. 1385–1402.
2. Sharpless, W. M., A Calorimeter For Power Measurements at Millimeter Wavelengths, IRE Trans. Microwave Theory and Techniques, MTT-**2**, September, 1954, pp. 45–47.
3. Burrus, C. A., Backward Diodes for Low Level Millimeter-Wave Detection, IEEE Trans. Microwave Theory and Techniques, MTT-**11**, September, 1963.