

THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XLIV

DECEMBER 1965

NUMBER 10

Copyright © 1965, American Telephone and Telegraph Company

Measurements of Electromagnetic Back-scattering from Known, Rough Surfaces*

By JACQUES RENAU and JAMES A. COLLINSON

(Manuscript received August 20, 1965)

We have measured the cross section for backscattering of laser beams from rough aluminum surfaces and a magnesium oxide slab. These surfaces were specially prepared and their statistical properties were measured. The laser wavelengths were $\lambda = 0.63, 1.15, \text{ and } 3.39\mu$, and both parallel and perpendicular polarizations were used. The angle of incidence was varied from 0° to 89° . In these experiments the ratio of the surface rms height h (from the mean surface) to the wavelength λ is larger than $1/4$; for such surfaces the cross section for backscatter at normal incidence is inversely proportional to the square of the rms surface slope, h/l , and is independent of wavelength. At large angles of incidence the cross section increases with increasing slope and also with increasing h/λ , approaching an upper limit which appears to be predicted by a Lambert scattering law. The angular dependence of the cross section differs for the two polarizations; at grazing incidence the cross section is larger for the parallel polarization. The published characteristics of the angular dependence of the cross section of microwave backscattering from the sea and the moon are in remarkable agreement with the backscattering cross section obtained from the various randomly-rough laboratory prepared surfaces at all angles of incidence. Comparison of the laboratory results with published moon data yields for the moon surface an rms height of 40 ± 10 cm, a mean correlation distance or mean scale size of 2.8 ± 0.7 meters,

* A summary of this paper was presented at the URSI Commission 2 April 1965 meeting in Washington, D.C.

an rms slope of $8^\circ \pm 4^\circ$, and a dielectric constant ϵ at microwave frequencies of 1.9 ± 0.3 gigacycles.

I. INTRODUCTION

Measurements of backscattering of electromagnetic (EM) waves from rough surfaces have been performed in the past using microwaves scattered from moon, sea and terrain surfaces. Among others, we refer to the experimental work of Davies and Macfarlane,¹ Grant and Yaplee,² Wiltse et al³ and Evans and Pettengill.⁴ More recently there has also been interest in backscattering of EM waves from rough overdense plasma surfaces, particularly near grazing angles of incidence.

In order to relate the measurement of backscattering of EM waves from randomly-rough surfaces to the characteristics of the scattering surface, the statistical properties of such surfaces must be independently measured. The bulk of EM wave scatter measurements from randomly-rough surfaces with gentle slopes consists of those obtained from the sea. Although efforts have been made to specify the state of the sea by means of the prevailing winds at the time of the experiment, the statistical parameters of the rough sea surfaces and of the moon have been matters of conjecture based on many untested assumptions. On the other hand, a direct statistical study of the surface irregularities of the moon is impossible, at present, and very difficult for the sea. For these reasons, the use of randomly-rough surfaces specially prepared for a scatter experiment was desirable and furthermore, was necessary to test the range of validity of the available rough-surface scattering theories. Surface preparation for such scatter experiments at microwave frequencies is a formidable task because of the following requirements: the mean height correlation distance (or scale size) l of the surface should be much larger than the wave length λ to correspond to most cases of physical interest, the beam diameter d must be much larger than l in order to make the scattering area a representative member of the statistical ensemble, and the largest dimension L of the scattering surface must be much larger than d so that all of the beam is intercepted even near grazing incidence. Thus in order to perform a meaningful experiment one requires that $L \gg d \gg l \gg \lambda$. At optical frequencies, however, where $\lambda \cong 1$ micron, the above requirements are satisfied when $l =$ tens of microns, $d =$ few mm, and $L =$ tens of cm, and the scaled down experiment can be performed conveniently in the laboratory with compact prepared surfaces. Lasers provide the beam power, directionality, spatial coherence and monochromaticity needed for successful measurements of backscattering at optical frequencies. Moreover, a He-Ne gas laser can be operated at a wavelength of 0.6328 (hereafter referred to as 0.63) 1.15, or 3.39 microns

simply by changing the cavity mirrors, so that wavelength dependence of backscattering is easily measured.

Reported in this paper are the results obtained from a laser backscatter experiment using prepared and statistically studied randomly-rough metallic surfaces at angles of incidence varying continuously from 0° (normal) to 89° , for parallel and perpendicular polarizations and for differing wavelengths. (Parallel and perpendicular polarization refer to the orientation of the electric field with respect to the plane of incidence. Workers describing the sea data use a different nomenclature. Our parallel polarization is equivalent to their vertical polarization and our perpendicular polarization is equivalent to their horizontal polarization.) We will describe first the preparation of the surfaces and the measurement of their statistical properties. The latter are the relevant correlation functions, height distributions, the rms height, and the mean height correlation distance. We will then give the arrangement for the backscatter measurements and the experimental results.

II. PREPARATION OF THE SURFACES AND THEIR STATISTICAL PROPERTIES

Ten by 22 cm flat aluminum surfaces were blasted at various pressures with hard alumina grits. One of the surfaces was blasted with sizes distributed from very small up to approximately 200 microns, another with sizes distributed up to approximately 100 microns, and a third surface was blasted with steel spheres of sizes distributed from very small up to 44 microns. In order to remove sharp edges caused by the blasting, we experimented with various methods of polishing, such as chemical polishing, electroplating, and electropolishing. Electropolishing was used since this process removed material more from the protrusions than from the valleys. Fig. 1 is a perspective view of one surface. The 3-inch ruler next to the surface in Fig. 1 is used for comparison of dimensions. In Fig. 2 we show the microscopic views at normal incidence of all three rough surfaces.

Contour traces or profiles of the surface irregularities were obtained at various locations both within and without the area illuminated by the laser beam, after the backscatter measurements were completed in order to avoid any possible damage to the surfaces. For each location, the tracing was done over a length of 2 cm, a distance which was found to be much larger than the mean correlation length of any of the surfaces.

The traces were obtained with a stylus which ran along the surface in a manner analogous to a phonograph pickup. The stylus pressure was kept low enough so that there was no detectable distortion of the surface. The radius of the stylus of 13 micron was measured by a Bausch and Lomb optical comparator. A short sample of these traces for each of the

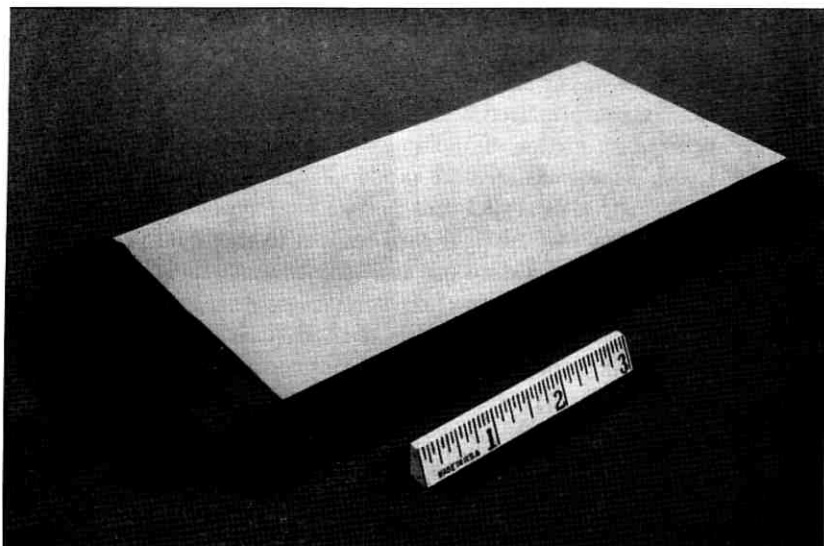
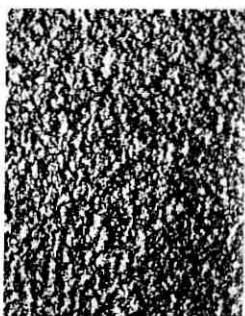


Fig. 1 — View of one of the aluminum surfaces.

aluminum surfaces is shown in Fig. 3. The estimated effective radius of the stylus at a depth of 1 micron in the surface is approximately 5 microns; at a depth of 2 microns it is 7.5 microns and at a depth of 13 microns or more the radius of the stylus should approach the constant



(a) SURFACE NO. 1
BLASTED WITH NO. 60
QUARTZ GRIT AND
ELECTROPOLISHED
MAGNIFICATION 42X



(b) SURFACE NO. 2
BLASTED WITH 150 MESH
 Al_2O_3 AND ELECTRO-
POLISHED
MAGNIFICATION 42X



(c) SURFACE NO. 3
BLASTED WITH STAINLESS
STEEL BALLS AND
ELECTROPOLISHED
MAGNIFICATION 42X

Fig. 2 — Microscopic views of the rough surfaces.

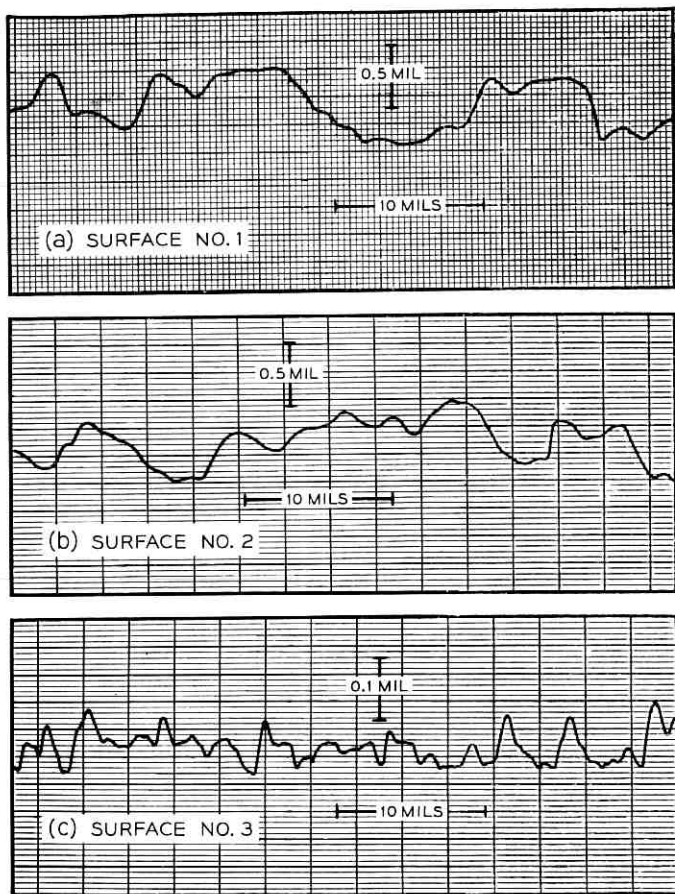


Fig. 3 — Portions of the profile traces of the rough surfaces.

value of 13 microns. Since the mean height correlation distance (mean scale size) obtained for each of the surfaces was much larger than the effective stylus radius, we believe the measured values truly represent the mean correlation distances l . Photo-micrographs of the cross section of the surface irregularities shown in Fig. 4 do not indicate any unusual irregularities which the stylus would be incapable of measuring.

These contour traces were used to obtain the normalized autocorrelation functions of the surface heights for each of the surfaces. The height autocorrelation function $\tau(\xi)$ was obtained from the contour traces of the surface heights $Z(\rho)$ by calculating for each 2-cm long trace

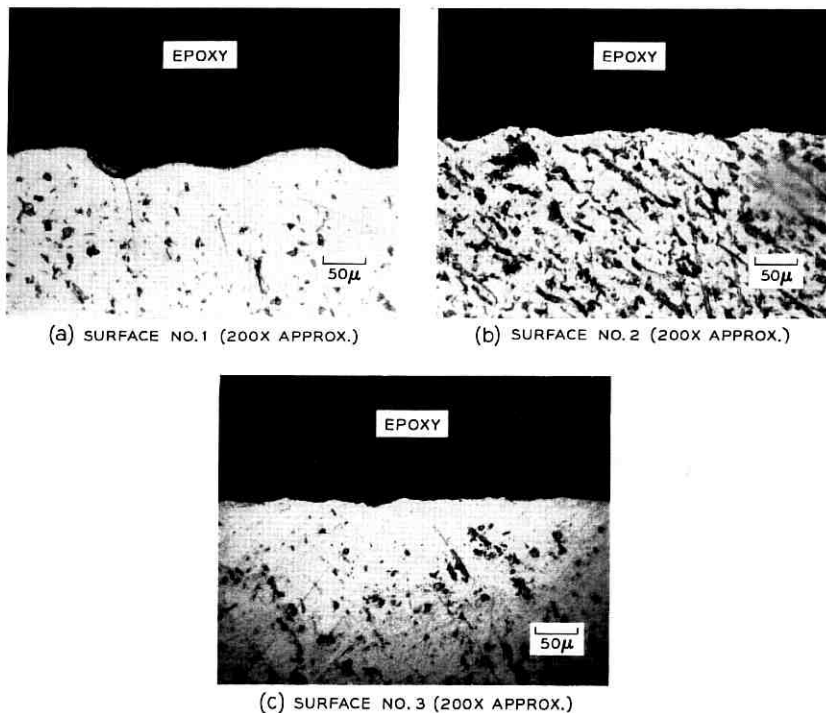


Fig. 4 — Photo-micrographs of the cross section of the rough surfaces.

$$\tau(\xi) = \frac{\frac{1}{3200} \sum_{i=1}^{3200} Z(\rho_i)Z(\rho_i + \xi)}{\frac{1}{3200} \sum_{i=1}^{3200} Z(\rho_i)Z(\rho_i)}$$

where $\rho_i = (x_i^2 + y_i^2)^{1/2}$ represents the position of i th element of height on the average surface and the separation distance ξ was successively increased in integral multiples of 4 microns. Since several 2-cm profiles were obtained at various parts of each surface, a collection of normalized autocorrelation functions for that surface resulted, and these are shown superimposed in Fig. 5 for surfaces No. 1, 2, and 3.

It may be remarked that the analytical function

$$\tau(\xi) = \sin \left[\frac{\pi}{2} \exp \left(- \frac{2}{\pi} \frac{|\xi|}{l} \right) \right],$$

used to fit the experimental correlation data of turbulent media by Corrsin and Kistler⁵ and to the studies of wakes from bodies in high

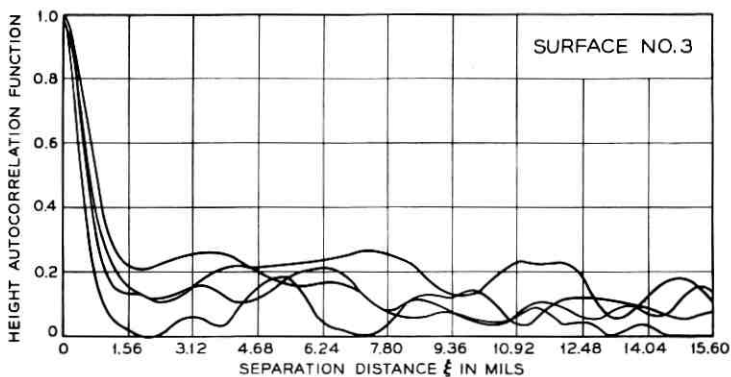
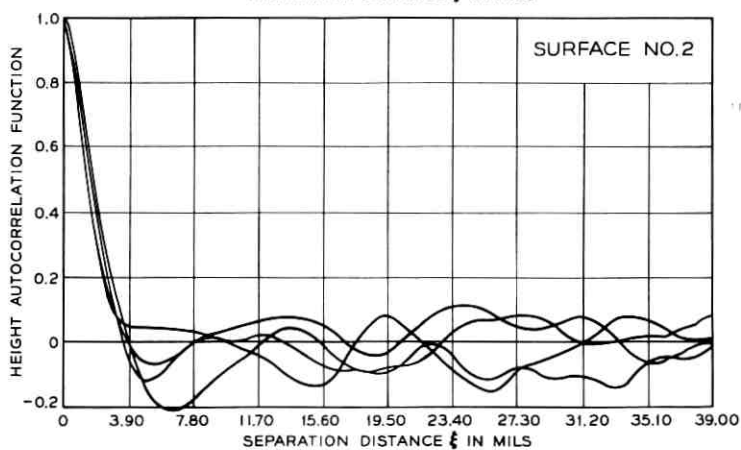
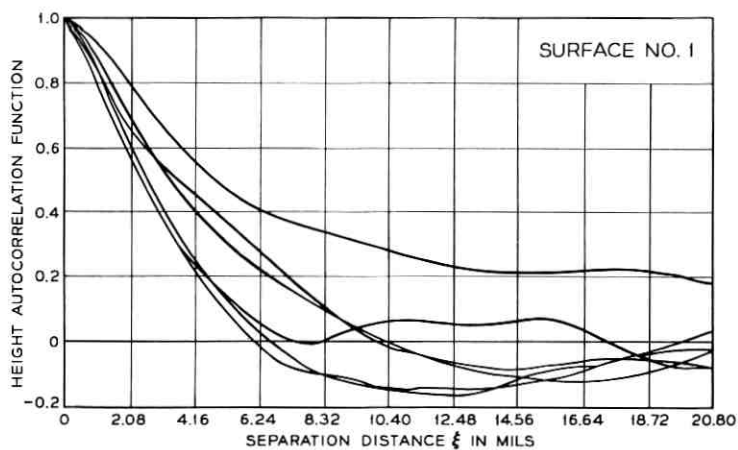


Fig. 5— Normalized height autocorrelation function (with zero mean) of rough surfaces No. 1, 2, and 3. (Multiple curves represent sampling of various locations of each surface).

speed flight by Schapker,⁶ also fits the experimentally obtained autocorrelation functions in our Fig. 5 over a wide range of separation distances ξ . However,

$$\tau(\xi) = \exp \left[- \frac{\left(\frac{\xi}{l}\right)^2}{2 \left(1 + \frac{|\xi|}{2l}\right)} \right]$$

has equivalent properties as the sine function, and fits the data almost as well, with a very slight change of mean scale size.

The experimental normalized autocorrelation functions, at their half value, have been used to estimate the mean height correlation distance l of the surfaces. The rms height from the zero mean height for each of the surfaces was, by definition, the square root of the unnormalized autocorrelation function at $\xi = 0$. The results of the studies for each of the aluminum surfaces as well as an additional surface (MgO Slab) that will be discussed later are given in Table I. We estimate a standard deviation of the order of $\frac{1}{2}$ of the values of h and l shown above. Since the results of Table I are independent of direction taken on the surface, we conclude the surface irregularities are statistically isotropic.

Also from the traces, the height distributions obtained at least at five different locations for each surface were averaged and are shown in histogram form in Fig. 6 for surfaces 1, 2, and 3. Because of the average shape of the height distribution it seems one may conclude that the statistical characteristics of the surface irregularities although not strictly Gaussian, may be approximated by a Gaussian distribution. We now describe the scatter experiments performed with these surfaces.

III. EXPERIMENTAL ARRANGEMENT

The arrangement of the experiment is shown in Fig. 7. The He-Ne laser tube had an overall length of 120 cm, a discharge length of 100 cm, and a tube bore of 5 mm. By using the proper mirrors, the laser was made to oscillate at 0.63μ , 1.15μ , or 3.39μ . The laser beam was plane polarized with an orientation determined by the Brewster-angle windows of the

TABLE I

	Al Surfaces			MgO Slab
	1	2	3	
Surface number	1	2	3	
rms height h from the mean height	7μ	3μ	1μ	25μ
Mean correlation distance l	50μ	26μ	10μ	90μ
rms slope h/l	$1/7.1$	$1/8.6$	$1/10$	$1/3.6$

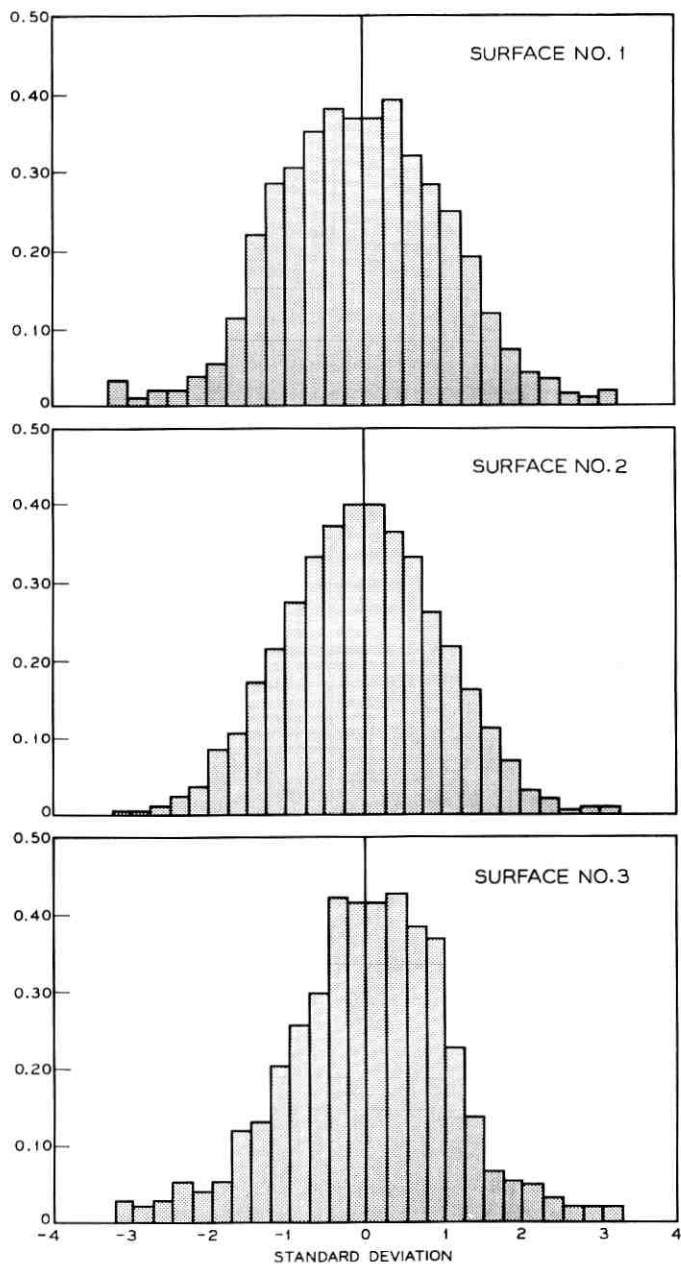


Fig. 6 — Average distribution of surface heights for surfaces No. 1, 2, and 3. (Overlapping histograms representing samples at various portions of each surface were used in averaging.)

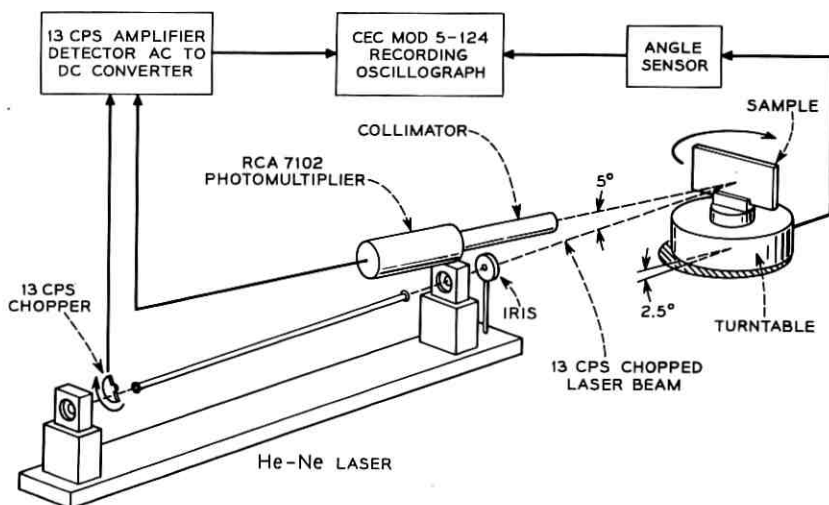


Fig. 7 — Diagram of experimental arrangement.

discharge tube. The diameter of the laser beam was about 4 mm, and it passed through an iris of approximately 4-mm diameter. The beam power was about 10 mw. The beam was incident upon one of the surfaces mounted upon an electronically controlled turntable which carried a protractor. The distance from the front laser mirror to the scattering surface was one meter. At wavelengths of 0.63μ and 1.15μ the relative backscattered energy was measured at all angles of incidence using RCA 7265 and 7102 photomultiplier tubes. The absolute level of backscattering at normal incidence was measured at all three wavelengths by focusing the backscattered energy onto a calibrated thermopile with a quartz lens. The thermopile output power was read with a Keithley Model 149 microvoltmeter.

The turntable was tilted by 2.5° so as to reflect specularly into the center of the detector. The constant speed turntable rotated the sample surface in the path of the incident beam, thus varying continuously the angle of incidence. The angular position of the turntable was calibrated and electronically recorded on an oscillograph. In order to minimize the noise level, the beam was chopped at a rate of 13 cps. The detected signal was amplified by a 13-cps amplifier ac-dc converter and fed simultaneously with the angle information into the oscillograph. The walls of the dark room where the experiment was performed, as well as all protruding objects, were covered with highly absorbent black material. At grazing incidence and with the most weakly backscattering surface, the signal was approximately 10 db above the background noise for both

0.63 μ and 1.15 μ wavelengths. The power input-output linearity of the over-all system at $\lambda = 0.63\mu$ and 1.15 μ were checked by the use of calibrated Kodak neutral filters.

In Fig. 8, we show a retrace of the raw data for one run as observed on the oscillograph. The largest observed fluctuations were much smaller than the width of the line of the redrawn curves. When the gain of the amplifier was increased to give suitable response for the weaker signal at larger angles of incidence the amplified signals did not saturate the equipment over the range of angles for which the data are shown.

IV. RELATION BETWEEN THE MEASURED BACKSCATTERED POWER AND THE BACKSCATTERING CROSS SECTION

It will be helpful in interpreting the data that follow to show the relationship between the backscattered power $P_B(\psi, f)$ incident upon the photomultiplier tube and the radar backscattering cross section $\sigma_B(\psi, f)$

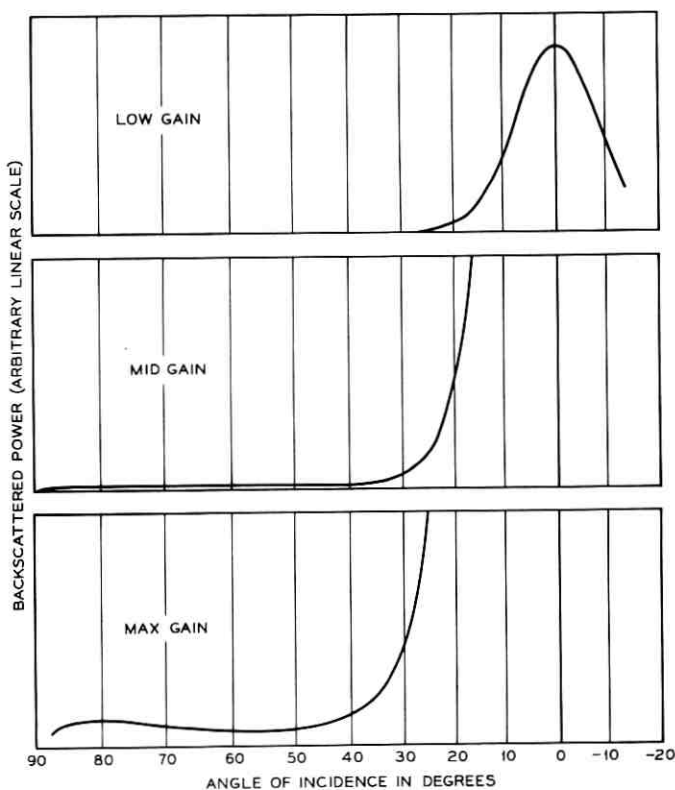


Fig. 8 — Trace of oscillograph raw data.

where f is the frequency of the incident beam, and ψ is the angle of incidence of the beam at the surface. The subscript B denotes the case of backscattering. By definition

$$P_B(\psi, f) = \frac{1}{2\eta} \langle \mathbf{E}_{SB}(\psi, f, R) \cdot \mathbf{E}_{SB}^*(\psi, f, R) \rangle A_{\text{rec}}$$

where $\langle \rangle$ denotes the ensemble average of the stochastic backscattered electric field $\mathbf{E}_{SB}(R)$ observed at the receiver located at a distance R from the scattering surface, η is the characteristic impedance of free space and A_{rec} is the energy sensitive area of the photomultiplier tube. Let the power incident upon the target at frequency f be denoted by $P_i(f)$, then

$$P_i(f) = \frac{1}{2\eta} |\mathbf{E}_i|^2 A_0,$$

where $|\mathbf{E}_i|$ is the amplitude of the incident electric field, and A_0 is the area of the collimated incident beam and is equal to the area illuminated on the surface at normal incidence. Therefore,

$$\frac{P_B(\psi, f)}{P_i(f)} = \frac{\langle \mathbf{E}_{SB}(\psi, f, R) \mathbf{E}_{SB}^*(\psi, f, R) \rangle A_{\text{rec}}}{|\mathbf{E}_i|^2 A_0}. \quad (1)$$

Let the numerator and the denominator be multiplied by $4\pi R^2$. For R much larger than any other dimension of interest in the experiment the radar cross section is defined as

$$\sigma(\psi, f) = \frac{4\pi R^2 \langle \mathbf{E}_s(\psi, f, R) \mathbf{E}_s^*(\psi, f, R) \rangle}{|\mathbf{E}_i|^2}$$

and in particular for the case of backscattering

$$\sigma_B(\psi, f) = \frac{4\pi R^2 \langle \mathbf{E}_{SB}(\psi, f, R) \mathbf{E}_{SB}^*(\psi, f, R) \rangle}{|\mathbf{E}_i|^2}. \quad (2)$$

Using (2), (1) may be rewritten as follows:

$$\frac{P_B(\psi, f)}{P_i(f)} = \frac{\sigma_B(\psi, f)}{A_0} \frac{\Omega}{4\pi} \quad (3)$$

where Ω , the solid angle, is defined as

$$\Omega \equiv \frac{A_{\text{rec}}}{R^2}.$$

We shall therefore use interchangeably the words normalized backscat-

tered power and normalized cross section, since

$$\frac{\frac{P_B(\psi, f)}{P_i(f)}}{\frac{P_B(0, f)}{P_i(f)}} = \frac{P_B(\psi, f)}{P_B(0, f)} = \frac{\frac{\sigma_B(\psi, f)}{A_0}}{\frac{\sigma_B(0, f)}{A_0}} = \frac{\sigma_B(\psi, f)}{\sigma_B(0, f)}. \quad (4)$$

In the following pages we shall neglect to use the subscript B since all our results are only for the case of backscattering. For convenience in notation we shall also denote the cross section $\sigma(\psi, f)$ by $\sigma(\psi)$.

V. BACKSCATTERING RESULTS

We will examine first the behavior of the absolute backscattering cross section per unit area at normal incidence $\sigma(0)/A_0$ from each of the surfaces. At normal incidence the data were taken with a photomultiplier ($\lambda = 0.63, 1.15\mu$) as well as with a thermopile detector ($\lambda = 0.63, 1.15,$ and 3.39μ) as described earlier. We then will give the cross section measured with a photomultiplier detector for continuously varying angles of incidence. In Figs. 9 to 15, the symbols E_{\parallel} and E_{\perp} refer, respectively, to the parallel and perpendicular orientation of the electric field to the plane of incidence.

5.1 Absolute Cross Section at Normal Incidence

Table II displays $P_B(0, f)/P_i(f)$, the ratio of backscattered power to incident power at normal incidence. Referring to (3), the values shown in Table II are also the backscattering radar cross section per unit area at normal incidence times 10^{-4} ($R \approx 100$ cm, $A_{\text{rec}} \approx 13$ cm² or $\Omega/4\pi \approx 10^{-4}$). No measurable difference was found in the results when using 0.63μ or 1.15μ . To obtain the radar backscatter cross section per unit area at normal incidence, $\sigma(0)/A_0$, one adds 40 db to the above values. Had we backscattered from a smooth flat metallic surface, the backscattering cross section per unit area, at normal incidence, would then be calculated on the basis of $[\sigma(0)/A_0]_{\text{smooth}} = 4\pi A_0/\lambda^2$ which for a beam of 2-mm radius and $\lambda \approx 1\mu$ is equal to 82 db. It is interesting to compare this result with those obtained from the rough surfaces at normal incidence.

It is evident that, within experimental error, the cross section, at normal incidence, is independent of wavelength, and that it increases as the rms slope of the surface irregularities decreases. A more quantitative conclusion is given in the section on Discussions and Conclusions. Meas-

TABLE II—RATIO OF BACKSCATTERED POWER TO INCIDENT POWER
 $P_B(0, f)/P_i(f)$ in DB

	$\lambda = 0.63\mu$ $\lambda = 1.15\mu$	$\lambda = 3.39\mu$
Surface No. 1 (roughest)	-25.5 ± 0.5	-25.0 ± 0.5
Surface No. 2	-24.0 ± 0.5	-23.5 ± 0.5
Surface No. 3 (smoothest)	-23.0 ± 0.5	-22.5 ± 0.5
MgO slab	-32.0 ± 0.5	-32.0 ± 0.5

urements were made with both polarizations, and at normal incidence no dependence on polarization was found implying statistical isotropy of the surface irregularities.

5.2 Normalized Cross Section vs Angle of Incidence

The angular dependence of backscattering cross section is most conveniently studied if the data are normalized to unity at normal incidence. This was done for all measurements of backscattering made with continuously changing angle of incidence. These data include dependence on polarization, wavelength, and surface characteristics. In all of the measurements of the angular dependence of cross section, the error did not exceed ± 0.5 db.

The results of the normalized backscattering cross section $\sigma(\psi)/\sigma(0)$

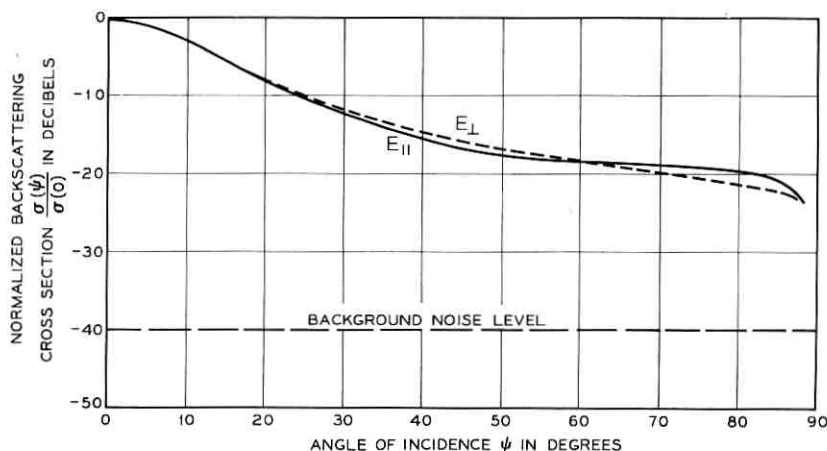


Fig. 9—Normalized backscattered power of laser beam versus angle of incidence, for both polarization, from rough Al surface No. 1, $h \approx 7\mu$, $l \approx 50\mu$, $\lambda \approx 0.63\mu$.

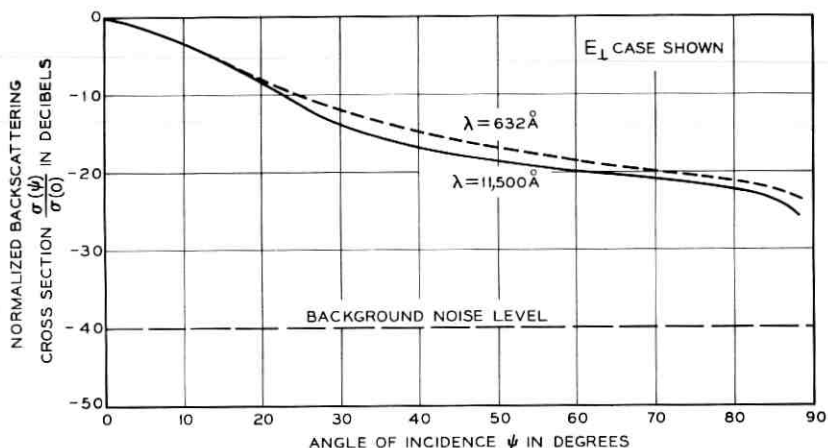


Fig. 10 — Normalized backscattered power of laser beam versus angle of incidence for perpendicular polarization from rough surface No. 1 for wavelengths $\lambda = 0.63\mu$ and 1.15μ .

versus angle of incidence ψ for surface No. 1 (roughest aluminum surface) is shown in Fig. 9. The laser wavelength λ was 0.63μ and the results for both the parallel and perpendicular polarizations are included.

Fig. 10 presents a comparison of the results for two wavelengths, 0.63μ and 1.15μ . The perpendicular polarization was used in both cases. We notice that the cross section decreases as the wavelength increases.

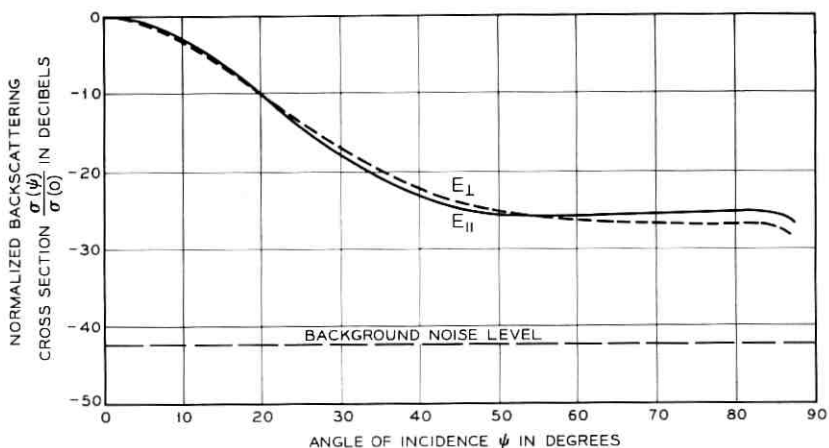


Fig. 11 — Normalized backscattered power of laser beam versus angle of incidence, for both polarization, from rough surface No. 2; $h = 3\mu$, $l = 26$, $\lambda = 0.63\mu$.

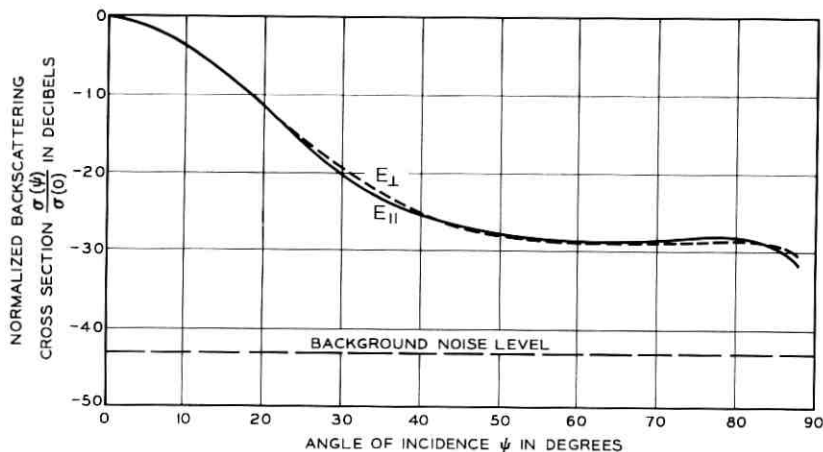


Fig. 12 — Normalized backscattered power of laser beam versus angle of incidence, for both polarization, from rough surface No. 3; $h \approx 1\mu$, $l \approx 10\mu$, $\lambda \approx 0.63\mu$.

The $\sigma(\psi)/\sigma(0)$ versus ψ for surface No. 2 is shown in Fig. 11, which includes curves obtained for both polarizations. The wavelength was 0.63μ .

The normalized cross section $\sigma(\psi)/\sigma(0)$ versus the angle of incidence ψ for surface No. 3 (the least rough of the Al surfaces) is shown in

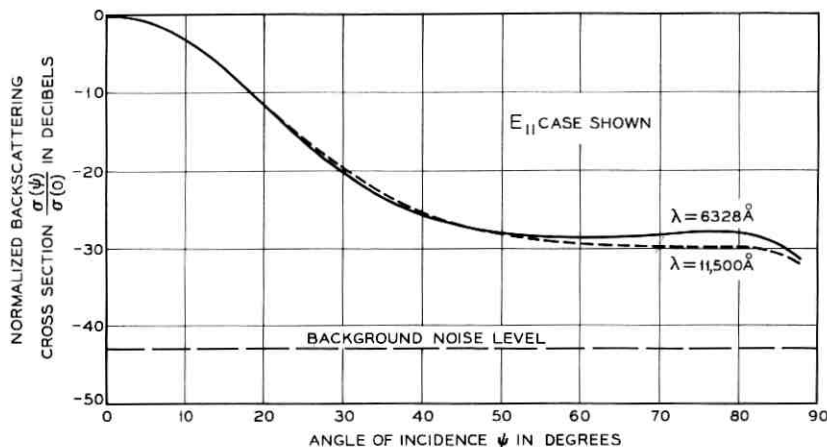


Fig. 13 — Normalized backscattered power of laser beam versus angle of incidence, for both polarization, from rough surface No. 3; $h \approx 1\mu$, $l \approx 10\mu$ and for wavelengths $\lambda \approx 0.63\mu$ and 1.15μ .

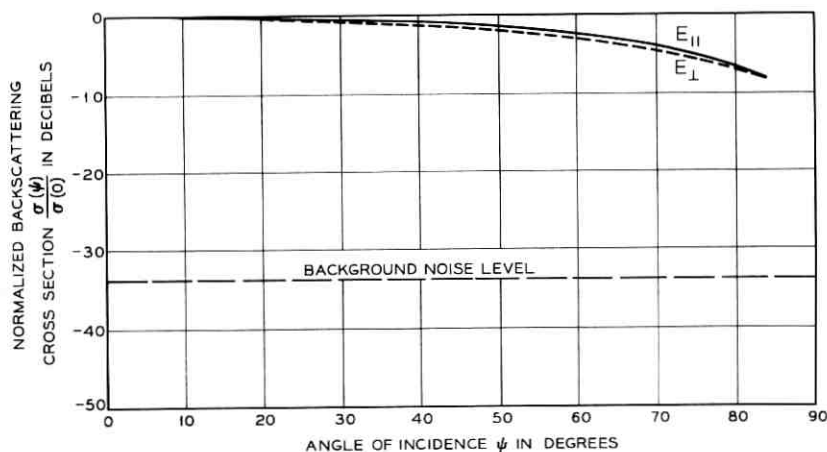


Fig. 14 — Normalized backscattered power of laser beam versus angle of incidence, for both polarization, from the MgO oxide surface; $h = 25\mu$, $l = 90\mu$, $\lambda = 0.63\mu$.

Fig. 12 where the variations with change of polarization are included. The wavelength was 0.63μ . The differences due to the change in polarization seem to be within experimental error, but repeated measurements consistently gave similar differences. The comparison of results at two wavelengths, 0.63μ and 1.15μ , is made in Fig. 13. Again the cross section decreases as the wavelength increases.

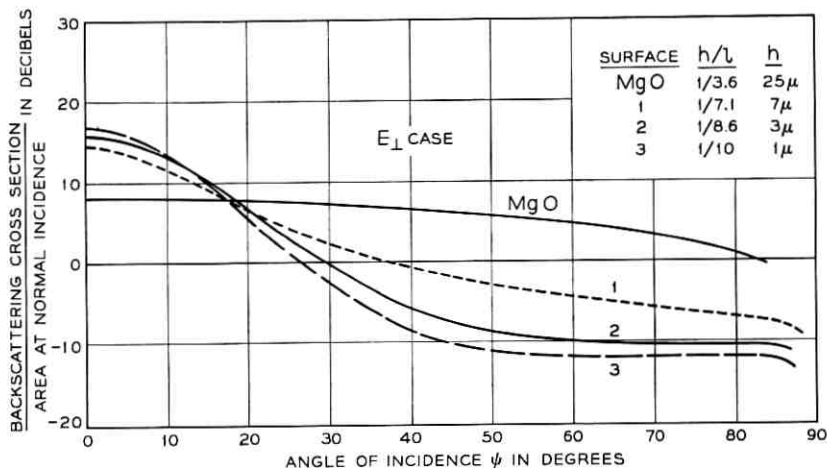


Fig. 15 — Absolute backscattering cross section versus angle of incidence, for the perpendicular polarization from the four rough surfaces; $\lambda = 0.63\mu$.

5.3 Magnesium Oxide as a Very Rough Scattering Surface

For purposes of comparison, measurements were made of the normalized cross section versus angle of incidence using a slab of magnesium oxide. Microscopic inspection of the MgO powder showed that 90 per cent of the grains of the magnesium oxide were in the range of 20 and 30 μ . We thus estimate an rms height of about 25 μ for the slab. The mean height correlation size l is estimated to be about 90 μ . The normalized backscattering cross section versus ψ is shown in Fig. 14 for both polarizations. The wavelength was 0.63 μ . This surface, which has a very large ratio of h/λ and $h/l \approx 1/3.6$, is a cosine law scatterer from 0° to approximately 80°. Therefore, the returned echoes are almost constant over a large range of angles of incidence.

5.4 Comparison of Data

The data of all the surfaces were replotted in Fig. 15 to compare surfaces 1, 2, and 3, and that of the MgO slab using the perpendicular polarization and a wavelength of 0.63 μ . We note that since the total reflection coefficient of MgO, similar to Al, is near unity at $\lambda = 0.63$ (K. W. Wecht et al, Bell Laboratories Tech. report MM-63-1153-11 August 19, 1963), comparison of the scattering properties of all the surfaces of Fig. 15 is appropriate.

We see that at large angles of incidence (about 50 to 80°) the back cross section increases not only with increasing slope but also with increasing rms height, approaching an angular distribution which is proportional to the cosine of the angle of incidence hereafter loosely referred to as a Lambert scattering law. Moreover as seen from Fig. 15, at grazing incidence a Lambert law scatterer appears to yield the upper limit for the backscattering cross section from rough surfaces. Fig. 15 leads us to conclude also that at normal incidence, such a surface has the least scattering cross section compared to that obtained from any other rough flat surface.

VI. CONCLUSIONS

A careful analysis of the various results of our experiments suggests that the following conclusions may be drawn:

(i) Results for the rough metallic surfaces display some common features. The cross section decreases with a negative curvature as the angle of incidence increases, then changes curvature and decreases very gradually for a large range of angles, and finally at grazing angles decreases once more. The last effect is at least partly due to shadowing

produced by the irregularities so that the area illuminated at grazing angles is reduced. If this interpretation is correct, then the experiments indicate that the shadowing effect is not as drastic as thought by some investigators on the basis of qualitative arguments.

Since the above characteristics of backscattering were observed from rough surfaces for which $h/\lambda \gtrsim \frac{1}{4}$ (including the observations at $\lambda = 3.39\mu$ at normal incidence) and at angles of incidence up to 89° , then, using m as a measure of roughness, a surface with irregularities may be considered rough whenever

$$m \equiv \frac{4\pi h}{\lambda} \cos \lambda \gtrsim \frac{1}{10}.$$

Further experiments may show that the lower limit on m can be somewhat smaller than $\frac{1}{10}$, but we feel it will not be very much smaller.

(ii) The backscattering radar cross section per unit area at normal incidence from statistically isotropic rough surfaces with $m \gtrsim \frac{1}{10}$ is independent of wavelength and of polarization. The latter is expected because of the statistically isotropic surfaces. The 10 per cent discrepancy observed in comparing $\lambda = 0.63\mu$ and $\lambda = 3.39\mu$ is perhaps due to the fact that at $\lambda = 0.63\mu$ the target is in the near field whereas at $\lambda = 3.39\mu$ the target is almost in the far field (the coherent laser aperture is taken as 2 mm).

(iii) The backscattering radar cross section per unit area at normal incidence is dependent on the rms slope of the surface and is relatively independent of the rms height. Empirically the slope dependence is well approximated as $\sigma(0)/A_0 = K/(\text{Slope})^2$, where the value of K is approximately $\frac{1}{2}$.

(iv) At large angles of incidence (from about 50 to 80°) no simple power law relationship between $\sigma(\psi)/A_0$ and h/l fits the data of Fig. 15. However, $\sigma(\psi)/A_0$ appears to increase (Figs. 10, 13, and 15) with increasing slope h/l , as well as with the rms height h .

(v) For $m \gg \frac{1}{10}$ and slope of $1/3.6$, the normalized backscatter cross section $\sigma(\psi)/\sigma(0)$ varies as the cosine of the angle of incidence from 0° to nearly 80° . For rough surfaces with larger slope h/l , one would expect that the cosine dependence (Lambert's law) would extend to 90° . In this case, conservation of energy requires that $\sigma(0)/A_0 = 4$. Consequently, when

$$\frac{1}{2} \frac{1}{(\text{slope})^2} \leq 4$$

the empirical law given in (iii) above should be replaced by the constant value of $\sigma(0)/A_0 = 4$. Thus,

$$\begin{aligned}\sigma(\psi) &= 4A_0 \cos \psi \\ &= 4A \cos^2 \psi\end{aligned}\quad \text{when } \begin{cases} m \gg \frac{1}{10} \\ \frac{h}{l} \gtrsim \frac{1}{2} \end{cases}$$

where $A = A_0 \sec \psi$ is the surface area illuminated.

(vi) At large angles of incidence, the backscattering cross section decreases as the wavelength increases. If we assume that the cross section is proportional to $\lambda^{-\gamma}$, then the experiments show that γ is less than 1.20. Our current experiments are directed at determining this wavelength dependence more precisely by the use of wavelengths whose ratio is much larger than the present value of $\lambda_1/\lambda_2 \approx 1.8$.

(vii) The difference in backscattering radar cross sections for the two polarizations increases as the angle of incidence is increased. At grazing incidence the cross section is larger for the parallel polarization.

(viii) At present we find no theoretical derivation which would account for the observed cross sections and the wavelength dependence over the entire range of angles of incidence.*

VII. COMPARISON WITH SEA AND MOON DATA

The angular dependence of backscattering discussed in (i) of the Conclusions is in good agreement with published *sea* and *moon* backscattering measurements for all angles of incidence. This agreement indicates that rough sea surfaces and the moon surface are truly randomly-rough at microwave frequencies. The sea backscattering data of Davies and Macfarlane¹ and Katz,⁷ obtained with different conditions of the sea surface, are plotted in Fig. 16 for comparison with our results. Workers reporting sea data customarily normalize the radar cross section of the sea $\sigma(\psi)$ by the area illuminated, A . This area is equivalent to our $A_0 \sec \psi$ in the range of angles where shadowing effects are negligible. Therefore, when comparing the results of our experiments with those for the sea, we note that our $\sigma(\psi)/A_0$ times $\cos \psi$ is their $\sigma(\psi)/A$ (denoted by σ° in their papers). Thus, when comparing quantitatively our $\{\sigma(\psi)/\sigma(0)\}$ with $\{\sigma^\circ(\psi)/\sigma^\circ(0)\}$ for the sea, one should add $10 \log_{10} \cos \psi$ to our $\sigma(\psi)/\sigma(0)$ expressed in db. The latter quantity may be obtained directly from Fig. 14 since the MgO surface is a cosine law scatterer up to approximately 80° .

Moon reflection data were reported by Evans and Pettengill⁴ for 3.6-cm and 68-cm wavelengths. They gave power backscattered as a function

* Although a recent theoretical study⁸ finds an expression which reasonably predicts the angular dependence of the cross section, the wavelength dependence at normal incidence is at marked variance with our experimental results (see next section).

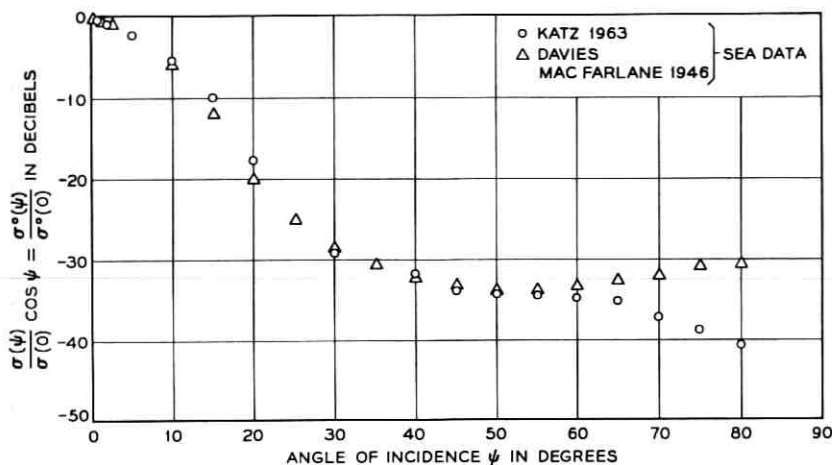


Fig. 16 — Characteristics of sea backscattering data for two differing sea surface conditions.

of delay time τ . We have redrawn their curves giving angle of incidence as the abscissa in our Fig. 17 to allow direct comparison with our results. Their equation (8) has been used which relates delay time for the backscattered power to angle of incidence ψ , i.e., $\cos \psi = 1 - (\tau/11.6)$ where τ is in milliseconds. The moon backscattering data at $\lambda = 0.86$ cm by Lynn et al⁹ has also been plotted. Superimposed on the moon data is the

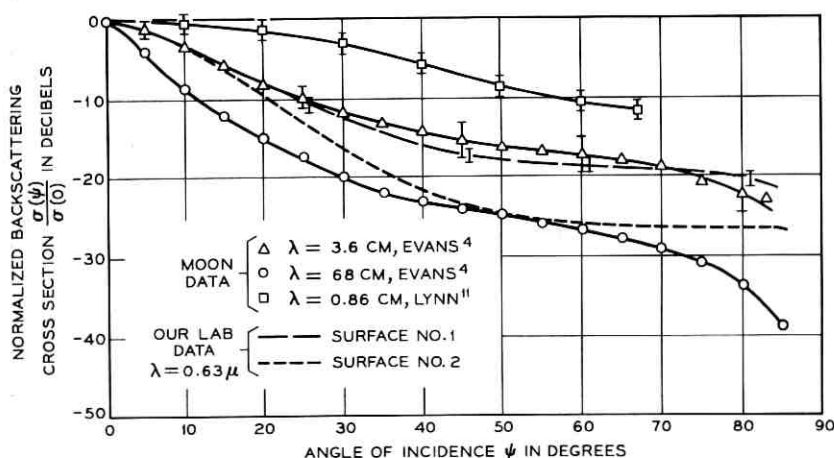


Fig. 17 — Comparison of moon and laboratory backscattering data. (Errors in our laboratory data ± 0.5 db all angles. Errors¹⁰ in moon data ± 1 db for $\psi \lesssim 30$, ± 2 db for $\psi \gtrsim 30$.)

laboratory result obtained from surface No. 1 ($h/\lambda \approx 7/0.63 \approx 11$ and $h/l \approx 1/7.1$) with $\lambda = 0.63\mu$. The curve drawn is the average of both polarizations (shown previously in Fig. 9), since the moon result is an average of both polarizations. Our result from surface No. 2 ($h/\lambda \approx 4.76$ and $h/l \approx 1/8.6$) is also shown. It can be seen that the moon measurements at $\lambda = 3.6$ cm correspond closely to those which one would expect from a randomly-rough surface with parameters similar to our surface No. 1. We have assumed that the variation in the reflection coefficient of the moon with angle of incidence is less than the errors involved in the moon backscattering measurements. This correspondence permits us to estimate the values of the statistical parameters of the moon surface. Thus $h/3.6$ cm ≈ 11 and $h/l \approx 1/7.1$, and hence the parameters of the moon are: rms height $h = 40 \pm 10$ cm (with respect to the mean surface), the mean scale size $l = 2.8 \pm 0.7$ meters, and the rms slope angle is $8^\circ \pm 4^\circ$. Any larger variation in these parameters would cause the laboratory scatter data to fall outside the limits of the error bars of the moon data. These results lead us to conclude that the number of small irregularities on the moon's surface is far greater than the number of big irregularities, such as the lunar mountains.

Using the value for the radar backscattering cross section per unit area for the moon published by Safran,¹¹ it is possible to calculate an approximate value for the dielectric constant, ϵ , of the moon at microwave frequencies. The radar cross section per unit area, measured at normal incidence from a metallic surface, at the laboratory, can be compared with that of the moon's surface. The adjustment needed to obtain agreement between the two results can then be attributed to departures from unity in the reflectivity of the moon. Specifically we can choose a model in which the dielectric impedance mismatch is the sole source of reflectivity. Thus at normal incidence, we write

$$\left(\frac{\sigma(0)}{A_0}\right)_{\text{moon}} = \left(\frac{\sigma(0)}{A_0}\right)_{\text{metal}} Q^2$$

where Q is the amplitude reflection coefficient. Since the mean scale size of the irregularities is much larger than the impinging wavelength of $\lambda = 3.6$ cm, we may set $Q = (\sqrt{\epsilon} - 1)/(\sqrt{\epsilon} + 1)$, where ϵ is the dielectric constant of the moon's surface. Safran¹¹ has given for $[\sigma(0)/A_0]_{\text{moon}}$ a value of -1.7 db ± 1.5 db. The laboratory result is $[\sigma(0)/A_0]_{\text{metal}} = 14.5$ db ± 0.5 db for surface No. 1. Using these values, we obtain $\epsilon = 1.9 \pm 0.3$ for the dielectric constant of the moon at microwave frequencies. This value of ϵ is close to that for certain porous materials such as loosely packed sand⁴ and pumice^{4,12} (porous volcanic glass).

Since we have discussed the close correspondence of the characteristics of the rough sea, the moon, and the laboratory-prepared random surfaces, it is well to note that observers have obtained a range of differing empirical results on the wavelength dependence of the backscattering cross section of the sea. All find a relation of the form $\sigma \propto \lambda^{-\gamma}$, but the value of γ ranges from 0 to 4 (Skolink,¹³ p. 531). The Naval Research Laboratory¹³ data at grazing angles indicates $\gamma \approx 1$. Since (i) and (ii) of the Conclusions apply, the moon data at $\lambda = 3.6$ cm and 68 cm reproduced in Fig. 17 show also $\gamma \approx 1$ at grazing incidence. This is consistent with our sixth conclusion given above. A recent theoretical study⁸ predicts [(7) and (18) of Ref. 8] that the backscattering cross section per unit area at normal incidence is proportional to λ^2 . Our results do not confirm these predictions.

VIII. ACKNOWLEDGMENT

We are indebted to E. Krauth for assistance in the preparation of the rough aluminum surfaces, to J. A. Giordmaine for providing the MgO slab, to C. Sandahl's group for contour traces of the surfaces, to P. Rosenthal's group at Cornell Aeronautical Laboratories for programming the statistical analysis, and to J. Quackenbush for assistance in setting up the experiment and for taking the scatter data. We also acknowledge the encouragement and support given by N. Levine and wish to thank I. Jacobs, H. G. Cooper and W. E. Danielson for helpful discussions.

REFERENCES

1. Davies, H. and Macfarlane, G. G., Radar Echoes from the Sea Surface at Centimetre Wavelengths, Proc. Phys. Soc., 58, 1948, pp. 717-729.
2. Grant, C. R. and Yaplee, B. S., Backscattering from Water and Land at Centimeter and Millimeter Wavelengths, Proc. IRE, 45, July, 1957, pp. 976-983.
3. Wiltse, J. C., Schlesinger, S. P., and Johnson, C. M., Backscattering Characteristics of the Sea in the Region from 10 to 50 kmc, Proc. IRE, 45, February, 1957, pp. 220-228.
4. Evans, J. V. and Pettengill, G. H., The Scattering Behavior of the Moon at Wavelengths of 3.6, 68, and 784 Centimeters, J. Geophys. Res., 68, January 15, 1963, pp. 423-447.
5. Corrsin, S. and Kistler, A. L., Free-Stream Boundaries of Turbulent Flows, NACA Report 1244, 1955.
6. Schapker, R. L., Turbulence Front Statistics of Wakes from Bodies in High-Speed Flight, AVCO Everett Research Note 530, 1965.
7. Katz, I., Radar Reflectivity of the Earth's Surface, The Johns Hopkins University, Appl. Phys. Lab. Tech. Digest, January-February, 1963.
8. Beckmann, P., Radar Backscatter from the Moon, J. Geophys. Res., 70, May 15, 1965, pp. 2345-2349.
9. Lynn, V. L., Sohigan, M. D., and Crocker, E. A., Radar Observations of the Moon at a Wavelength of 8.6 Millimeters, J. Geophys. Res., 69, 1964, pp. 781-783.

10. Evans, J. V., private communication.
11. Safran, H., Backscattering Properties of Moon and Earth at X-Band, AIAA Journal, January, 1964, pp. 100-101.
12. Brunshwig, M., et al, Estimation of the Physical Constants of the Lunar Surface, The University of Michigan, Report No. 3544-1-F, November, 1960. (Table IV, p. 16 and Figure 6, p. 20.)
13. Skolnik, M. I., *Introduction to Radar Systems*, McGraw-Hill Book Co., Inc., 1962, p. 531.

The Effects of Digital Errors on PCM Transmission of Compandored Speech

By I. DOSTIS

(Manuscript received August 10, 1965)

The recent interest in the use of pulse code modulation (PCM) for the transmission of speech has made it desirable to determine the effects of digital line errors on certain codes. Improved performance for low-level talkers has been generally obtained by the use of instantaneous compressors and expandors (compandors) in order to avoid using a large number of digits in the PCM coder-decoder (codec). A comparison of the effects of digital errors for a logarithmically compandored system transmitting speech is made on the basis of mean square distortion power and the distribution of error magnitude. Results are obtained for the binary, Gray (reflected binary), and folded binary codes. The results indicate that the binary and Gray codes have a "click" as well as a "noise" component of distortion while the folded binary code produces only noise at low-talker levels. "Clicks" have been defined as errors which have amplitudes greater than half the full range amplitude; the remaining errors are considered "noise". For the folded binary code, the most significant digit gives polarity information; the remaining digits represent the signal magnitude in binary code.

I. INTRODUCTION

A comparison of the effects of digital errors for a logarithmically compandored^{1,2} system transmitting speech by PCM is made on the basis of mean square error power and the distribution of the error magnitude. A block diagram of the system considered in the analysis is shown in Fig. 1. The calculations are based on the assumption of at most one error per 8-digit word. The results obtained for binary and Gray codes indicate that there is a "click" as well as "noise" component of distortion at low talker levels. "Clicks" have been defined as errors which have amplitudes greater than half the full range, the remaining error

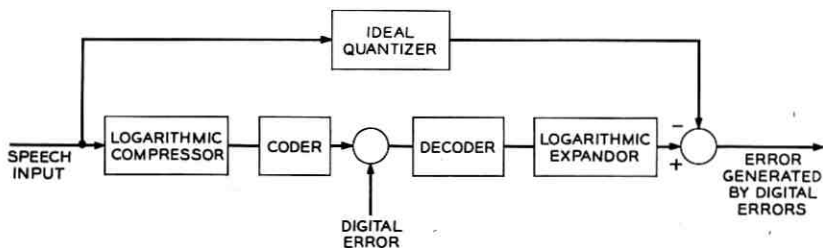


Fig. 1 — A block diagram of the system under consideration.

amplitudes are considered "noise". A folded binary code (see Fig. 2) essentially produces only the noise component at low talker levels. Low talker levels were considered to be 50 db below full load sine wave (50 db BFLSW) for all the cases calculated.

The results obtained for the folded binary code indicate that this code does not have a click problem at low talker levels when compared to the binary and the Gray codes. A question is raised, however, concerning the subjective effect at low levels of the noise components of distortion. If the noise is subjectively equivalent to quantizing error, an error rate of 10^{-6} will cause a 3-db impairment in the system performance. This occurs for speaker levels between 30 and 50 db BFLSW in a system using a logarithmic compression factor^{1,2} (μ) of 100 and the folded binary code.

The conclusions obtained from the computed results for logarithmic companding are:

- (i) The folded binary code produces fewer clicks than either the Gray or binary code at low speaker levels.
- (ii) The Gray, binary, and folded binary click and noise components are comparable at higher speaker levels.

The combined subjective effect of noise and clicks occurring simultaneously has not been evaluated.

II. RESULTS

The results obtained for the probability of a given code word occurring are plotted in Fig. 3 for logarithmic compression factors^{1,2} of 50, 100, and 200. The input signal was assumed to be Laplace distributed.^{3,4,5} The figure indicates that compression has a major effect on the probability of a given code word occurring. The tendency of the compressor is to assure a more uniform distribution of the signal across the lower code levels for low and average level talkers. One notes that for a μ of

11111	11111	10000
11110	11110	10001
11101	11101	10011
11100	11100	10010
11011	11011	10110
11010	11010	10111
11001	11001	10101
11000	11000	10100
10111	10111	11100
10110	10110	11101
10101	10101	11111
10100	10100	11110
10011	10011	11010
10010	10010	11011
10001	10001	11001
10000	10000	11000
00000	01111	01000
00001	01110	01001
00010	01101	01011
00011	01100	01010
00100	01011	01110
00101	01010	01111
00110	01001	01101
00111	01000	01100
01000	00111	00100
01001	00110	00101
01010	00101	00110
01011	00100	00111
01100	00011	00010
01101	00010	00011
01110	00001	00001
01111	00000	00000
FOLDED BINARY CODE	BINARY CODE	GRAY CODE

Fig. 2 — 5-digit codes.

100 the average talker (30 db BFLSW) has a nearly uniform distribution over the range $-\frac{1}{4}$ to $+\frac{1}{4}$ full range.

The mean square value of the distortion caused by digital line errors is compared to 8-digit quantizing error power^{1,2} for a 10^{-6} error rate, a 2-volt peak-to-peak input amplitude and compression factors (μ) of 50, 100, and 200 in Figs. 4 and 5. Altering the error rate simply results in a scale change for the line error distortion power. The results plotted in Fig. 4 include all error amplitudes whereas errors greater than half the full range amplitude (clicks) have been removed for the $\mu = 100$ case in Fig. 5. Preliminary results of subjective tests indicate that clicks are less objectionable than quantizing noise. The figures indicate the folded binary code yields superior click performance at low talker levels compared to the Gray or binary codes. The noise or non-click component for the folded binary is greater than that produced by the Gray code at low levels. For an error rate of 10^{-6} and 8-digit quantizing, the line error distortion power and quantizing error power are equal for the folded binary code in the 30 db to 50 db BFLSW range. If the line error distortion is subjectively equivalent to quantizing error, the system per-

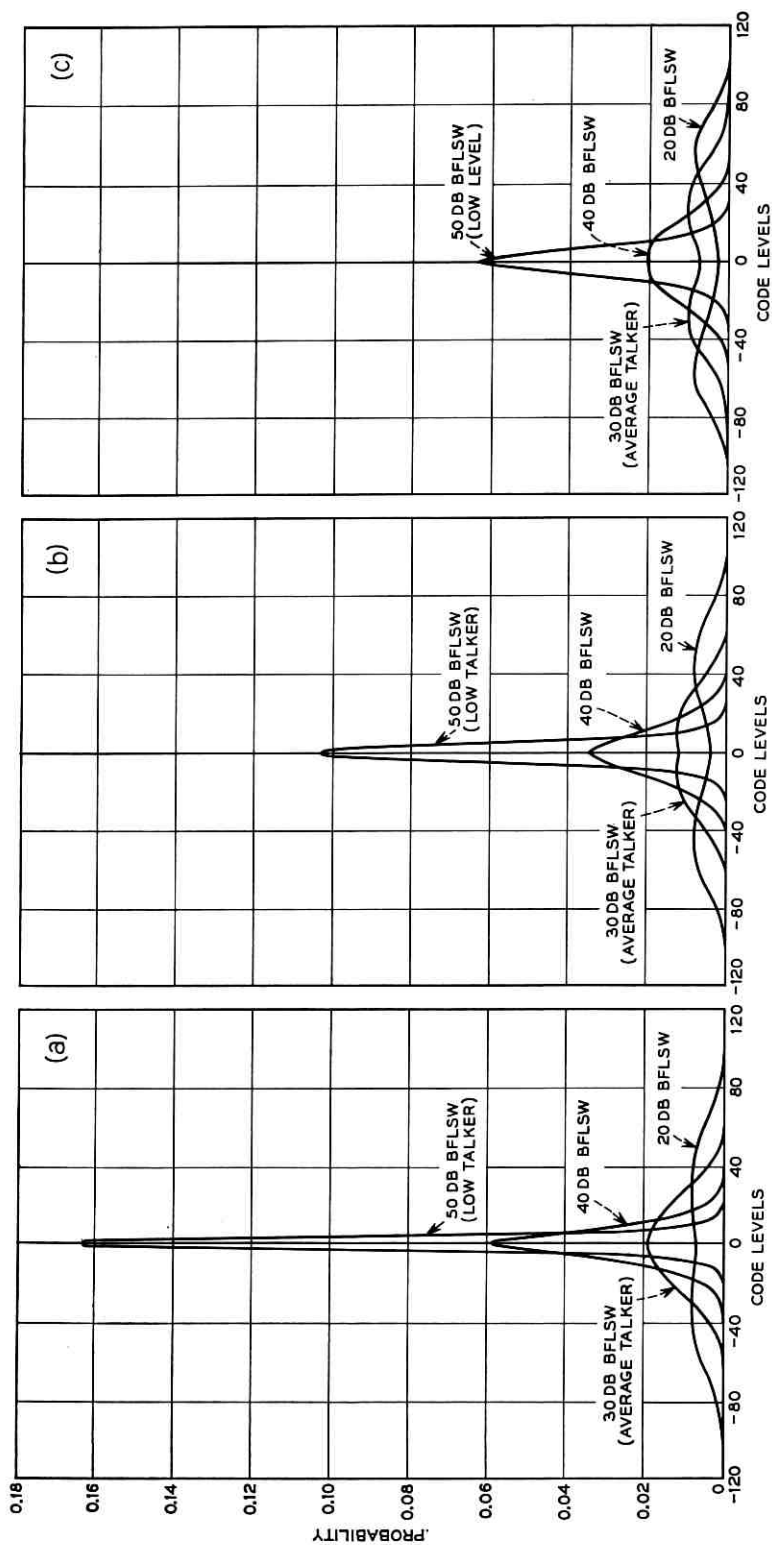


Fig. 3 — Word probability—8-digit coder Laplace distribution. (a) $\mu = 50$ logarithmic compression. (b) $\mu = 100$ logarithmic compression. (c) $\mu = 200$ logarithmic compression.

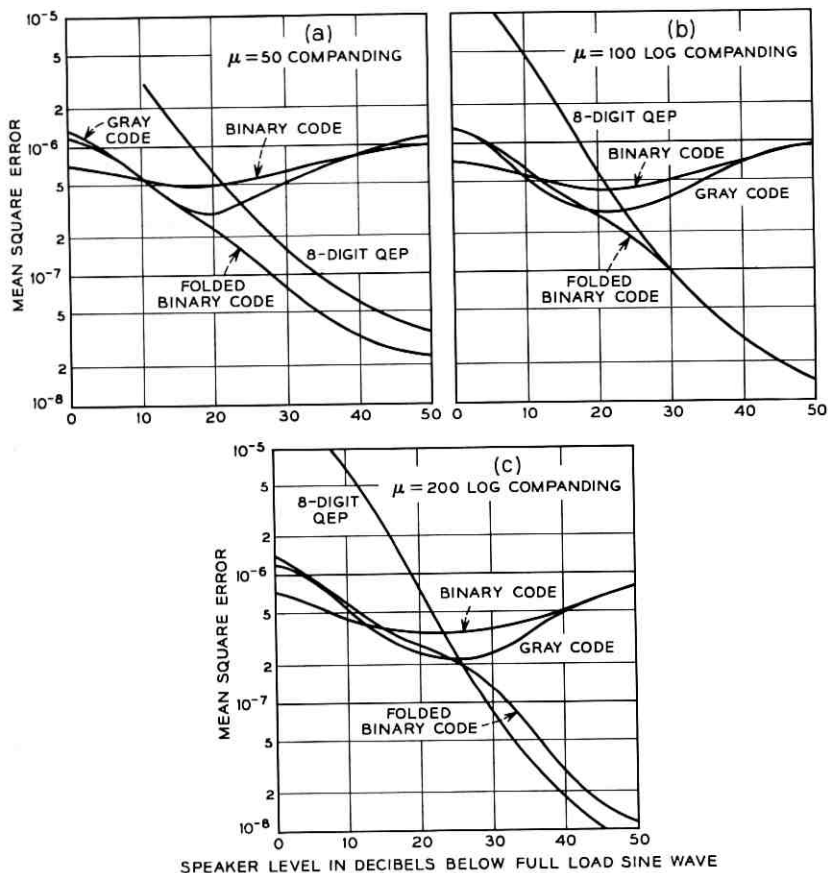


Fig. 4 — Mean square error vs speaker level—8 digit coder — 2-volt peak-to-peak full load amplitude. (a) $\mu = 50$ log companding. (b) $\mu = 100$ log companding. (c) $\mu = 200$ log companding.

formance will be degraded by 3 db. This result applies for a logarithmic compression of 100.

We conclude that the essential low-level problem is noise for the folded binary and clicks for the Gray and binary codes.

The probability of errors of a specified magnitude occurring are plotted in Figs. 6 through 9 for the binary, folded binary, and Gray codes. A clear demarcation between errors greater than and less than half the full amplitude is indicated in Figs. 6, 7, and 8 for low levels. These results led to the definition that errors greater than half the full load would

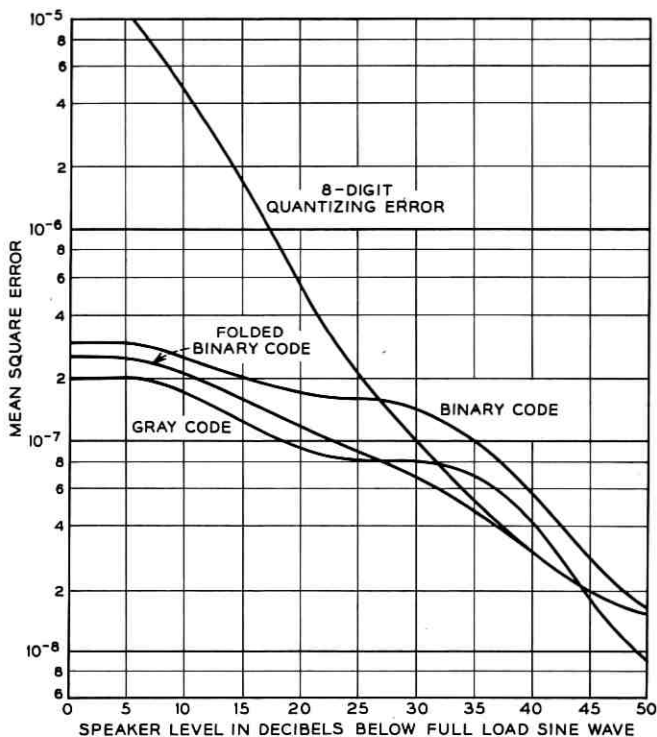


Fig. 5—8-digit codes (compression factor, $\mu = 100$ error rate 10^{-6} , clicks removed—2-volt peak-to-peak full load amplitude).

be called clicks. The dividing line between clicks and noise is not clear for average and high-level speakers.

It is apparent that the folded binary code produces virtually no clicks at low levels. This leads to the conclusion that noise rather than clicks will be the dominant low-level problem in transmission of folded binary, whereas clicks are the problem for Gray and binary.

The mean square error caused by specified digit errors for various speaker levels are plotted in Fig. 10(a), (b) and (c). The results indicate that the largest mean square error for the folded binary code is caused by digit 2 errors. The largest mean square error for Gray and straight binary depends to some extent on the speaker level. The largest error for low and average levels occurs in the 1st digit for binary and the 2nd digit for Gray.

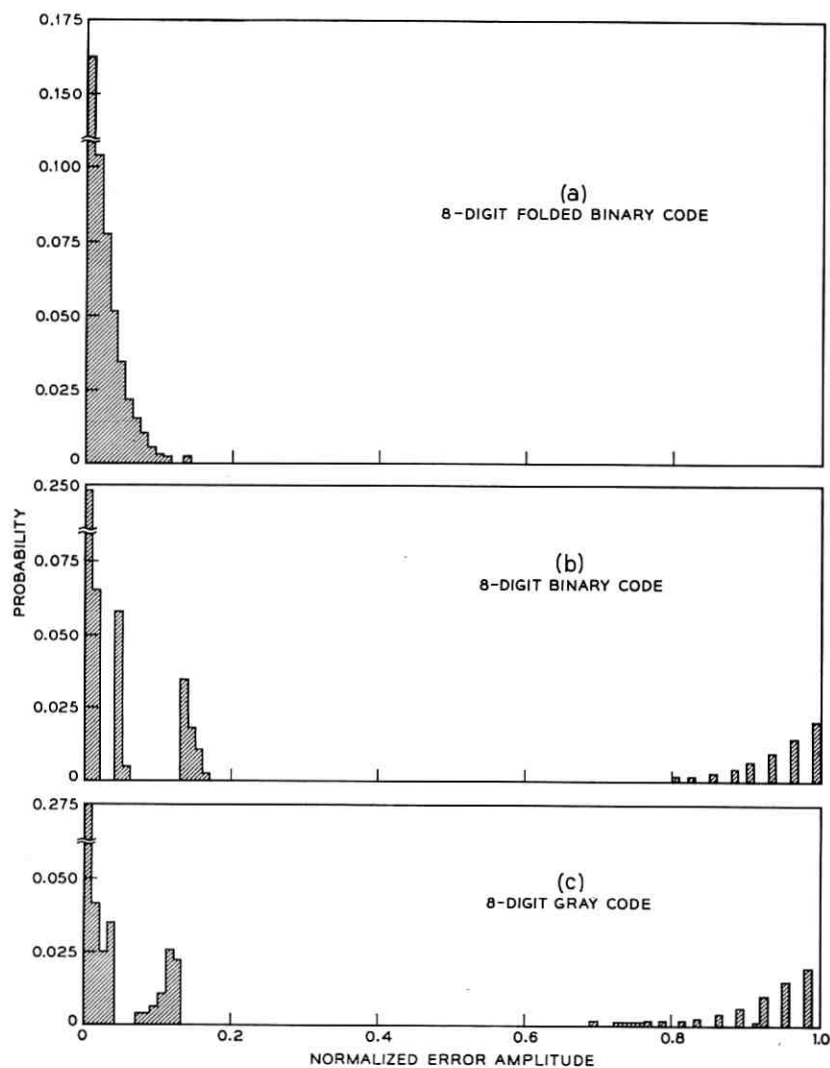


Fig. 6—Probability of error (speaker level 50 db BFLSW, $\mu = 50$ log companding, symmetric for negative amplitudes, low-level talker). (a) 8-digit folded binary code. (b) 8-digit binary code. (c) 8-digit Gray code.

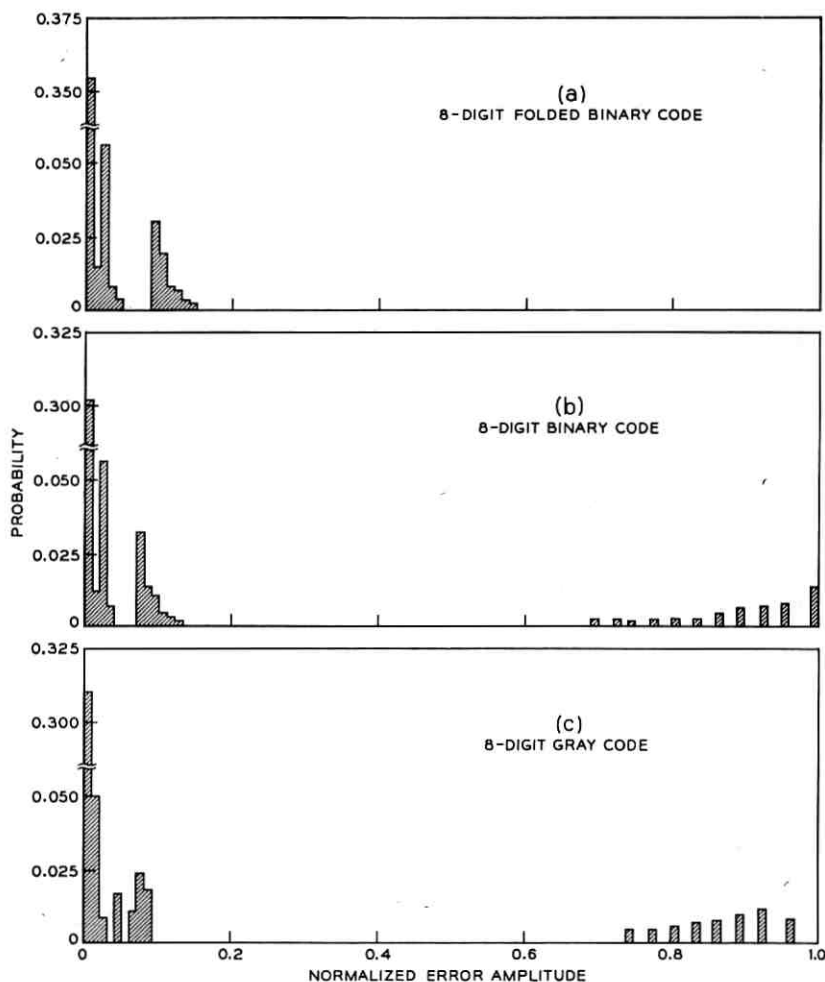


Fig. 7—Probability of error (speaker level 50 db BFLSW, $\mu = 100$ log compander, symmetric for negative amplitudes, low-level talker). (a) 8-digit folded binary code. (b) 8-digit binary code. (c) 8-digit Gray code.

III. ASSUMPTIONS AND PROCEDURES

3.1 Assumptions

The results obtained were calculated using the following assumptions:

- (i) The terminals of the transmission system including the coder and decoder are ideal.
- (ii) All digital errors are introduced in the transmission medium.

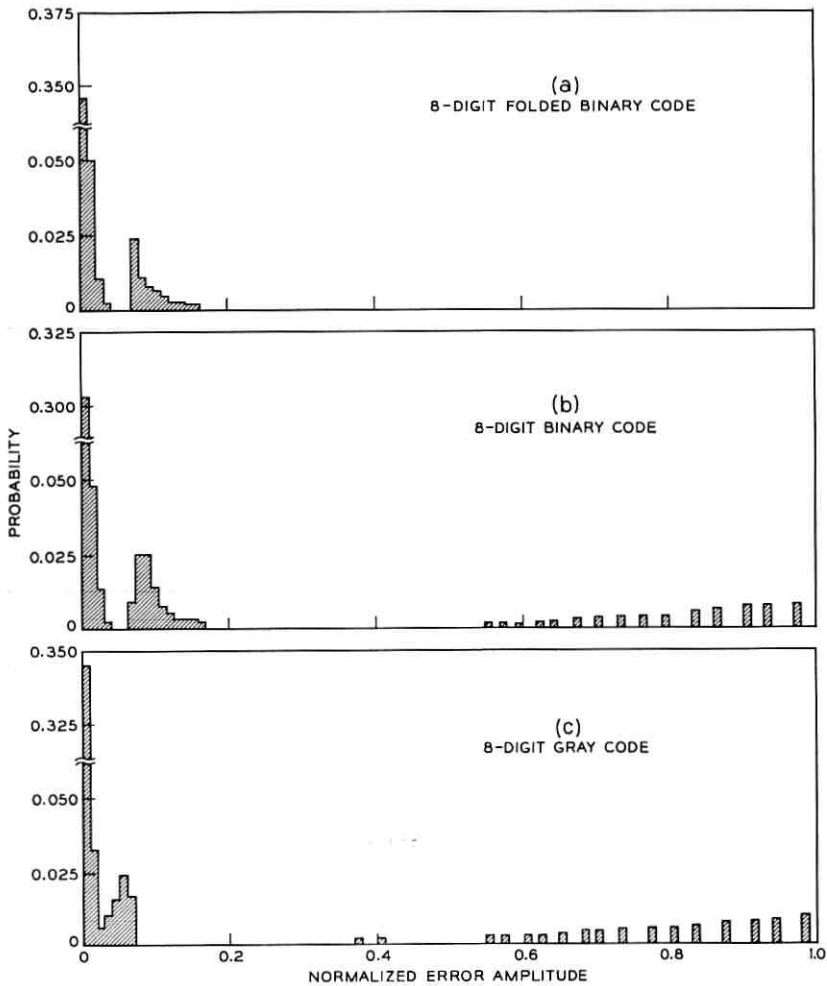


Fig. 8—Probability of error (speaker level 50 db BFLSW, $\mu = 200$ log companding, symmetric for negative amplitudes, low-level talker). (a) 8-digit folded binary code. (b) 8-digit binary code. (c) 8-digit Gray code.

(iii) No more than one digital error occurs in any given code word.

(iv) The input signal is speech and is assumed to be Laplace distributed.^{3,4,5}

(v) The mean value of the input signal corresponds to the mid-range of the coder-decoder.

(vi) Logarithmic companding^{1,2} with $\mu = 50, 100,$ and 200 .

(vii) The effects of overload have been ignored.

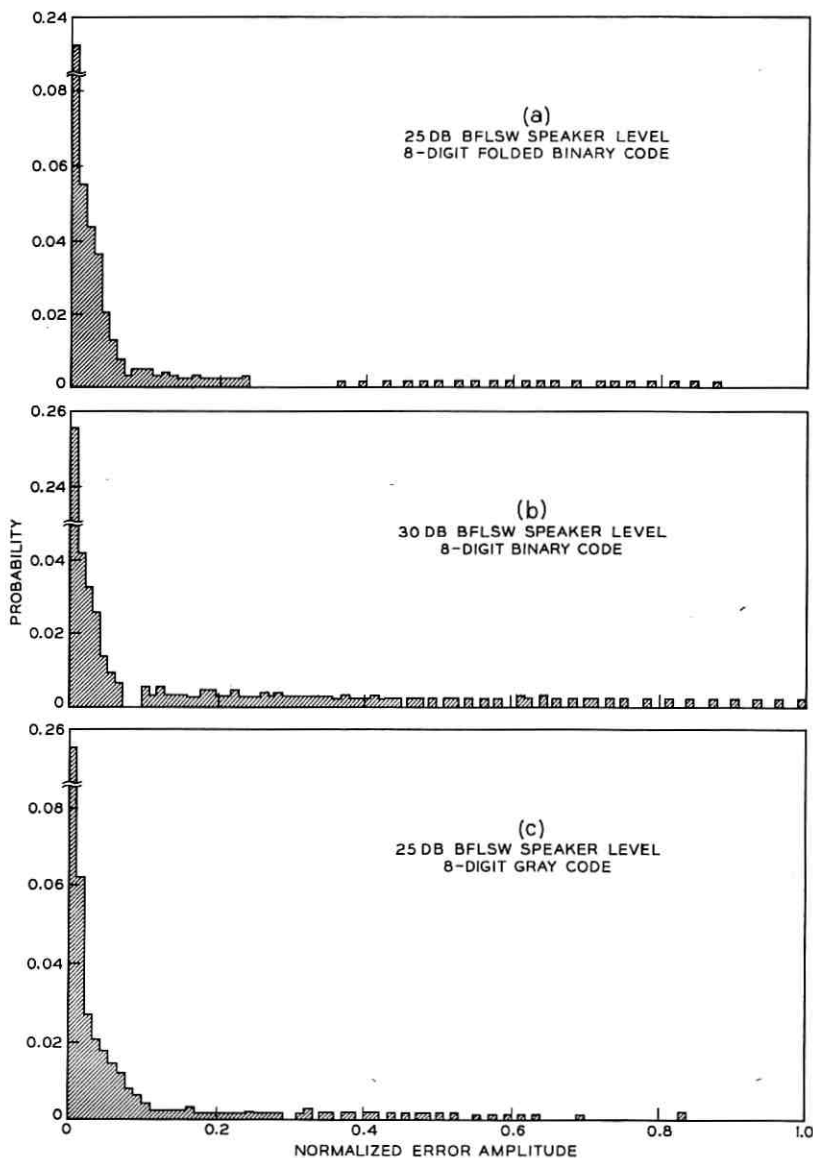


Fig. 9—Probability of error— $\mu = 100$ log compandor symmetric for negative amplitudes average talker. (a) 8-digit folded binary code (speaker level 25 db BFLSW). (b) 8-digit binary code (speaker level 30 db BFLSW). (c) 8-digit Gray code (speaker level 25 db BFLSW).

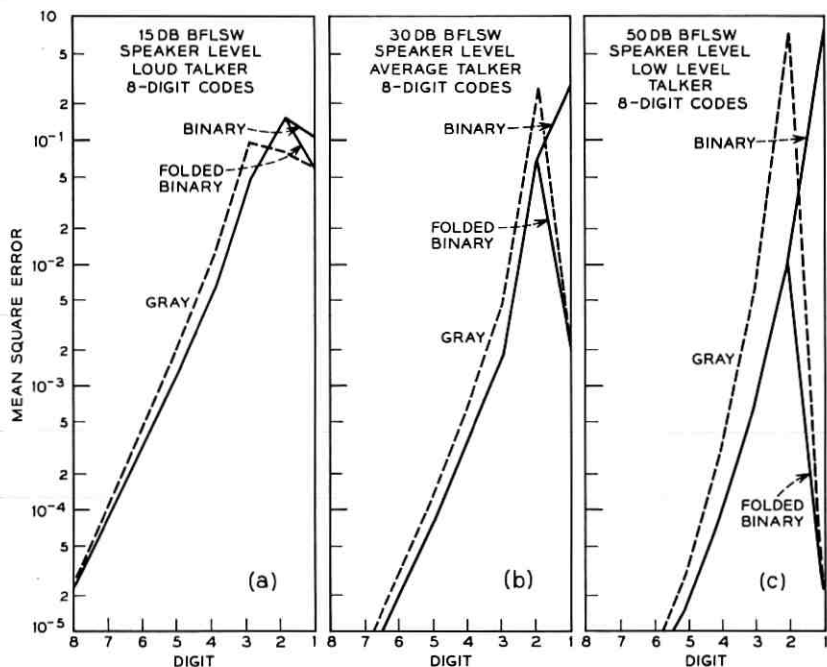


Fig. 10—Mean square error vs digit—2-volt peak-to-peak signal—(a) 8-digit codes (speaker level 15 db BFLSW—louder talker). (b) 8-digit codes (speaker level 30 db BFLSW—average talker). (c) 8-digit codes (speaker level 50 db BFLSW—low-level talker).

3.2 Procedure

3.2.1 Mean Square Error

A block diagram of the system assumed for the calculations performed is given in Fig. 1. The results were calculated using a probability of error of 10^{-6} . This error rate is sufficiently low to justify assumption (iii) above. The probability of a given word occurring on the line depends on the compression factor of the coder and the input statistics. The mean square error caused by digital line errors for the assumptions made is*

$$\bar{\eta}^2 = \sum_{J=1}^L \sum_{\substack{K=1 \\ K \neq 0}}^M \epsilon_{KJ}^2 P_K P_J P_c \quad (1)$$

* Formulations have been presented in the literature, different from the above, which may be more suitable for determining an optimal nonredundant code.^{6,7}

where

ϵ_{KJ} is the amplitude of the error for an error in digit J of word K

P_K is the probability of a particular code word occurring

P_J is the probability of a particular digit of a given word being chosen*

P_e is the probability of a digital error (error rate)

L is the number of digits in the coder

$M = 2^{(L-1)}$

Using the assumptions listed above, we reduce (1) to

$$\bar{\eta}^2 = P_e \sum_{J=1}^L \eta_J^2 P_J \quad (2)$$

where η_J^2 is the mean square error for a given digit and is defined in (3) as

$$\eta_J^2 \triangleq 2 \sum_{K=1}^M \epsilon_{KJ}^2 P_K. \quad (3)$$

The calculation of the values given in (2) and (3) are straightforward but quite tedious in the case of a companded system with varying speech levels. The problem was programmed for the IBM 7094 computer and the results of the computation for (2) and (3) presented in Figs. 4, 5, and 10(a), (b), and (c).

3.2.2 Word Probability†

The probability of a given word K occurring for a given speaker level was calculated directly from (4)^{2,3,4,5}

$$P_K = C \int_{B_K}^{B_{K+1}} \exp(-2C|x|) dx \quad (4)$$

where

$$C = \frac{\text{RMS VALUE FULL LOAD SINE WAVE}}{\text{RMS VALUE SIGNAL}} = \frac{E}{\sqrt{2\sigma^2}}$$

and the B_K are normalized coder levels for a specified compression characteristic. Normalization is with respect to full load.

The value of P_K given above is for a zero mean input signal. When the input signal and the coder do not have a zero mean, a shift can be performed easily. Using the assumptions given, the following result for P_K

* $P_J = 1/L$ for the case considered.

† An analytic approach is presented in Appendix A.

was obtained,

$$P_K = \frac{1}{2} \{ \exp(-2CB_K) - \exp(-2CB_{K+1}) \}. \quad (5)$$

The word probability for speaker levels of 20 db, 30 db, 40 db, and 50 db BFLSW have been plotted in Fig. 3 for $\mu = 50, 100, \text{ and } 200$.

3.2.3 Line Errors

The introduction of line errors was accomplished by complementing a given digit in a specific code word and determining the new word obtained. The results obtained obviously depend on the code being considered. A transition matrix was produced for each particular code considered, i.e., Gray, binary, and folded binary.

The rows of the transition matrix represent the originally transmitted code word and the columns the digit in error from most to least significant. The entry in column J and row K is the new word R produced by an error in digit J of word K . The transition matrix for a 3-digit binary code is given in Table I. This matrix is then used to determine the error caused by a specified digital error. The operation required to determine ϵ_{KJ} is decoding. If the output voltages corresponding to the code words K and R are denoted by D_K and D_R respectively, we obtain the following expression for ϵ_{KJ} ,

$$\epsilon_{KJ} = D_R - D_K. \quad (6)$$

One should note that the digit in error was specified as digit J in the above example.

TABLE I—TRANSITION MATRIX FOR 3-DIGIT BINARY CODE

		Digit-in-Error				
		1	2	3		
Binary Word Representation	000	Word-in-Error	0	4	2	1
	001		1	5	3	0
	010		2	6	0	3
	011		3	7	1	2
	100		4	0	6	5
	101		5	1	7	4
	110		6	2	4	7
	111		7	3	5	6

The error matrix having elements ϵ_{KJ} was then used to calculate the mean square error for a given digit, the mean square error for a given error rate with and without the "click" component and the distribution of error amplitudes for each code considered. Criteria other than mean square can be easily adapted to the computation.

IV. REMARKS AND CONCLUSIONS

The results obtained indicate the folded binary code yields better click performance for a given probability of digital error when compared to Gray and binary transmission. The effects of clicks should be virtually nonexistent at low talker levels in the folded binary code when compared to binary and Gray code. The question arises, however, concerning the subjective effect of the line error distortion when compared to quantizing error since the clicks at low levels have been suppressed. A 3-db reduction in S/N will occur for an 8-digit, $\mu = 100$, 10^{-6} error rate system transmitting folded binary for speaker levels between 30 db and 50 db BFLSW. The system performance will also be degraded by 3 db if line error distortion is subjectively equivalent to quantizing error.

Further work in evaluating the subjective effect of clicks and noise occurring simultaneously is required.

V. ACKNOWLEDGMENT

The author wishes to acknowledge the invaluable advice and aid on programming given by Miss E. G. Cheatham.

APPENDIX

Calculation of Density Function and Word Probability After Compression

A.1 Word Probability

An analytic expression for the probability density function at the output of a logarithmic compressor is derived. The density function is derived for a speech input which can be approximated by a Laplacian distribution.^{2,3,4,5} The density function used is given in (7) as

$$p(e_i) = C \exp(-2C |e_i|), \quad (7)$$

where, e_i is the input signal and C is defined by

$$C = \frac{\text{RMS VALUE OF FULL-LOAD SINE WAVE}}{\text{RMS VALUE OF SIGNAL}}.$$

Normalizing the result to full-load sine wave of unit amplitude results in a value $1/\sqrt{2\sigma_i^2}$ for C .

The characteristic of a logarithmic compressor normalized to a full-load unit sine wave is given by²

$$v_0 = \frac{\log(1 + \mu e_i)}{\log(1 + \mu)} \quad \text{for } 0 \leq e_i \leq 1, \quad (8)$$

and

$$v_0 = -\frac{\log(1 - \mu e_i)}{\log(1 + \mu)} \quad \text{for } -1 \leq e_i \leq 0. \quad (8b)$$

The output density function can be determined directly by transforming the input, i.e.,

$$p(v_0) = p(e_i) \left| \frac{de_i}{dv_0} \right|. \quad (9)$$

Performing the operation defined in (9) yields,

$$p(v_0) = \frac{C \log(1 + \mu)}{\mu} (1 + \mu)^{v_0} \exp \left[-\frac{2C}{\mu} \{(1 + \mu)^{v_0} - 1\} \right] \quad (10)$$

for $\mu > 0$ and $v_0 \geq 0$

and $p(-v_0) = p(v_0)$. The expression given in (10) is the probability density function of the compressor output.

Determining the word probability for a given code word K involves integrating the derived density between the lower and upper transition values for which this code word is emitted. For the case of a large number of digits (fine quantizing), an approximate expression can be derived. The step size for an n digit quantizer having a peak-to-peak amplitude of 2 is $1/2^{(n-1)}$. The value of the word probability $P(K)$ for word K is given by;

$$P(K) = \frac{C \log(1 + \mu)}{\mu} \int_{(K-1)/2^{n-1}}^{K/2^{n-1}} (1 + \mu)^{v_0} \cdot \exp \left[-\frac{2C}{\mu} \{(1 + \mu)^{v_0} - 1\} \right] dv_0 \quad (11)$$

$$\text{for } K \geq 1 \quad \text{and} \quad v_0 \geq 0.$$

We can approximate this result for the case of fine step sizes by

$$P(K) \cong \frac{C \log(1 + \mu)}{2^{n-1}\mu} (1 + \mu) e^{(K-1)/2^{n-1}} \cdot \exp \left[-\frac{2C}{\mu} \left\{ (1 + \mu) e^{(K-1)/2^{n-1}} - 1 \right\} \right] \quad (12)$$

for $K \geq 1$ and $v_0 \geq 0$,

or

$$P(K) \cong \frac{C \log(1 + \mu)}{2^{n-1}\mu} (1 + \mu) e^{(2K-1)/2^n} \cdot \exp \left[-\frac{2C}{\mu} \left\{ (1 + \mu) e^{(2K-1)/2^n} - 1 \right\} \right]. \quad (13)$$

For the case of $n = 8$, $\mu = 100$, $K = 1$, $C = 100$ (40 db BFLSW)

$$P(1) \cong \frac{4.61}{128} (1.002) e^{-2(0.002)} \cong \frac{4.61}{128} (1.002) \cong 0.036. \quad (14)$$

This agrees favorably with the value computed from the exact equation for this case of 0.0351.

A.2 Determination of the Most Probable Code

The most probable code level will correspond approximately to the peak(s) of the density function when the quantizing steps are small. For this condition we obtain,

$$\frac{dp(v_0)}{dv_0} = 0 = \frac{C \ln^2(1 + \mu)}{\mu} (1 + \mu)^{v_{0m}} \exp \left[-\frac{2C}{\mu} \cdot [(1 + \mu)^{v_{0m}} - 1] \right] \times \left\{ 1 - \frac{2C}{\mu} (1 + \mu)^{v_{0m}} \right\} \quad (15)$$

for $v_{0m} \geq 0$

where v_{0m} is the peak value of v_0 . Hence we have,

$$\frac{\mu}{2C} = (1 + \mu)^{v_{0m}}$$

or

$$v_{0m} = \frac{\log(\mu/2C)}{\log(1 + \mu)} \quad \text{for } v_{0m} \geq 0. \quad (16)$$

The results obtained using (16) compare favorably with those deter-

TABLE II — COMPARISON OF COMPUTED AND ANALYTICALLY OBTAINED VALUES

C (db BFLSW)		0	5.0	10.	15.	20.	25.	30.	35.	40.
v_{0m}	Theoretical	0.85	0.72	0.599	0.474	0.349	0.224	0.099	—	—
	Computer	0.85	0.72	0.60	0.47	0.35	0.22	0.10	0	0

mined on the computer as indicated in Table II. The value given for v_{0m} is the decimal fraction of full load code levels.

REFERENCES

1. Members of the Technical Staff of Bell Telephone Laboratories, *Transmission Systems for Communications*, Bell Telephone Laboratories, 1964.
2. Smith, B., Instantaneous Companding of Quantizing Signals, *B.S.T.J.*, 36, May, 1957, p. 653.
3. Davenport, W. B., An Experimental Study of Speech Wave Probability Distributions, *J. Acoust. Soc. Amer.*, 24, 1952, p. 390.
4. Purton, R. F., A Survey of Telephone Speech Signal Statistics and Their Significance in the Choice of a PCM Companding Law, *IEE*, January, 1952, p. 60.
5. Cramer, H., *Mathematical Methods of Statistics*, Princeton University Press, 1946.
6. Bado, J., Coding for Least RMS Error in Binary PCM Channels, *Wescon*, TR62-3, May, 1962.
7. Velichkin, A. I. and Grushko, I. I., Optimum Nonredundant Codes, *Bull. Acad. Sci. USSR, Tech. Sci. — Energetics and Automation*, 6, 1962.



Computer Solutions of the Traveling Salesman Problem

By SHEN LIN

(Manuscript received August 18, 1965)

Two algorithms for solving the (symmetric distance) traveling salesman problem have been programmed for a high-speed digital computer. The first produces guaranteed optimal solution for problems involving no more than 13 cities; the time required (IBM 7094 II) varies from 60 milliseconds for a 9-city problem to 1.75 seconds for a 13-city problem. The second algorithm produces precisely characterized, locally optimal solutions for large problems (up to 145 cities) in an extremely short time and is based on a general heuristic approach believed to be of general applicability to various optimization problems. The average time required to obtain a locally optimal solution is under $30n^3$ microseconds where n is the number of cities involved. Repeated runs on a problem from random initial tours result in a high probability of finding the optimal solution among the locally optimal solutions obtained. For large problems where many locally optimal solutions have to be obtained in order to be reasonably assured of having the optimal solution, an efficient reduction scheme is incorporated in the program to reduce the total computation time by a substantial amount.

I. INTRODUCTION

The traveling salesman problem may be stated as follows: "A salesman is required to visit each of the n given cities once and only once, starting from any city and returning to the original place of departure. What route, or tour, should he choose in order to minimize the total distance traveled?" Instead of distance, other notions such as time, cost, etc., can be considered as well. In this paper, we shall use the term "cost" to represent any such notion.

Mathematically, the problem may be stated in the following two equivalent ways:

- (1) Given a "cost matrix" $D = (d_{ij})$, where d_{ij} = cost of going from

city i to city j , ($i, j = 1, 2, \dots, n$), find a permutation $P = (i_1, i_2, i_3, \dots, i_n)$ of the integers from 1 through n that minimizes the quantity

$$d_{i_1 i_2} + d_{i_2 i_3} + \dots + d_{i_n i_1}.$$

(2) Given a "cost matrix" D as above, determine x_{ij} which minimizes the quantity $Q = \sum_{i,j} d_{ij} x_{ij}$ subject to

- (a) $x_{ii} = 0$
 (b) $x_{ij} = 0, 1$
 (c) $\sum_i x_{ij} = \sum_j x_{ij} = 1$

and

(d) for any subset $S = \{i_1, i_2, \dots, i_r\}$ of the integers from 1 through n ,

$$x_{i_1 i_2} + x_{i_2 i_3} + \dots + x_{i_{r-1} i_r} + x_{i_r i_1} \begin{cases} < r & \text{for } r < n \\ \leq n & \text{for } r = n. \end{cases}$$

The second version is a formulation of the traveling salesman problem as a linear program and hence the problem may be solved as such. However, the number of constraints becomes astronomical even for relatively small n . Dantzig, Fulkerson, and Johnson¹ have given a linear-programming approach to the symmetric ($d_{ij} = d_{ji}$) traveling salesman problem that considers only part of the required linear constraints and have found the technique effective in several cases.

Since we have only a finite number of possible tours to consider ($\frac{1}{2}(n-1)!$), the problem is really to obtain a reasonably efficient algorithm for finding an optimal solution. Certain algorithms employing branch and bound techniques have been tried and appear to be efficient for some problems; however, the computation time involved is unpredictable and increases very rapidly with n . Numerous authors have tried different techniques to obtain "near-optimal" solutions by a series of approximations and for specific problems were able to prove optimality of their solutions. For any conjectured optimal solution, however, the proof for optimality is dependent upon inspectional work which is usually heuristic in nature, and is certainly highly problem dependent, thus making it difficult to program for a computer.

Two algorithms for solving the (symmetric distance) traveling salesman problem have been programmed for a high-speed digital computer. The first algorithm, called k -length string optimization, is discussed in detail in Appendix A. It produces guaranteed optimal solutions for problems involving no more than 13 cities; the time required* varies

* IBM 7094 II.

from 60 milliseconds for a 9-city problem to 1.75 seconds for a 13-city problem. The algorithm is a slight modification of that given by Held and Karp.² However, we achieve a significant reduction in computation time by taking advantage of the fact that the distance matrix is symmetric. Due to the limitation on the size of the problem it can effectively handle, we find that it is not as useful as the second algorithm which we shall discuss below. The second algorithm (implemented by a copyrighted program) produces precisely characterized locally optimal solutions for large problems (up to 145 cities) in an extremely short time and is based on a general heuristic approach believed to be of general applicability to various optimization problems. The average time required per locally optimal solution is under $30n^3$ microseconds where n is the number of cities involved. Repeated runs on a problem from random initial tours result in a high probability of finding the optimal solution among the local optimum solutions obtained. For large problems where many locally optimal solutions have to be obtained in order to be reasonably assured of having the optimal solution, an efficient reduction scheme is incorporated in the program to reduce the total computation time by a substantial amount.

II. λ -OPTIMALITY

Before we describe the second algorithm, we first introduce the concept of λ -optimality. This serves to classify tours into a descending chain of classes possessing increasingly stronger necessary conditions for optimality. As we shall see later, this forms the basis for the construction of our second algorithm.

From the point of view of graph theory, we may consider the n cities as vertices of a nondirected complete graph, and the entries d_{ij} of the distance matrix real numbers assigned to links u_{ij} connecting city i to city j . A permutation $P = (i_1, i_2, \dots, i_n)$ representing a tour may be considered as a collection of n links $u_{i_1 i_2}, u_{i_2 i_3}, \dots, u_{i_n i_1}$ forming a Hamiltonian circuit, and the quantity $C = d_{i_1 i_2} + d_{i_2 i_3} + \dots + d_{i_n i_1}$ the cost associated with the tour.

For convenience, let us say a link u_{ij} is *admissible* if there is an optimal tour containing it. All other links are *inadmissible*. A set of links is said to be an *admissible set* if there exists an optimal tour containing all the links in the set. The *index* of a tour is the maximum number of links which the tour has in common with an optimal tour, i.e., the maximum number of links in the tour which form an admissible set.

We define λ -optimality of tours as follows:

Definition: A tour is said to be λ -optimal (or simply λ -opt) if it is im-

possible to obtain a tour with smaller cost by replacing any λ of its links by any other set of λ links.

We list below a few theorems concerning λ -optimality. Proofs are omitted since they are all fairly obvious. In Appendix D some interesting unsolved problems concerning λ -optimality will be discussed.

Theorem 1: Let T be a tour which is λ optimal with index k . Then either T is optimal or $k < n - \lambda$.

Theorem 2: Any tour is 1-optimal.

Theorem 3: The following properties of a tour are equivalent:

- (a) *The tour is 2-optimal.*
- (b) *The tour is optimal relative to inversion; where by inversion we mean reversing the order of a set of neighboring cities in the tour.*
- (c) *The tour does not intersect itself (in a generalized sense for non-Euclidean distance matrices).*

Theorem 4: A tour is optimal if and only if it is n -optimal.

Theorem 5: Let C_λ denote the set of all λ -optimal tours, then $C_1 \supset C_2 \supset \dots \supset C_n$. In other words, a λ -optimal tour is also λ' -optimal for $\lambda' < \lambda$.

Note that the well-known theorem, which states that an optimal tour does not intersect itself, is contained in Theorems 3, 4, and 5. ($C_n \subset C_2$).

III. THE SECOND ALGORITHM

In his paper, "A Method for Solving Traveling Salesman Problems", G. A. Croes³ applied a simple transformation, called "inversion" to transform a trial solution into another with smaller costs, iterating until no further inversions are desirable. Then he gave a method for deriving the optimal solution from the inversion free solution obtained. He pointed out, however, that the final adjustment procedures are difficult to program for a computer as they involve mostly inspectional work. For large problems, it seems doubtful whether a human being can exhaustively carry the computations through or even whether a computer program based upon those techniques will be feasible or efficient.

Putting aside the final adjustment procedures, we ask if there are other simple transformations which are stronger than the inversions. Since the inversion-free tours are just the 2-opt tours, we decided to write a computer program to produce 3-opt tours. As it turns out, the 3-opt tours are very much stronger than the inversion-free tours in the sense that (1) every 3-opt tour is inversion free, (2) the average tour

cost is considerably less, and (3) the probability of an optimal solution showing up as a 3-opt tour is significantly higher than that of 2-opt tours. Experimenting with a program producing 4-opt tours, we also find that we spend a great deal more computation time in producing 4-opt tours while not increasing noticeably the probability that it is optimum. Computational results on many problems support the claim that we have found a really efficient way of attacking the traveling salesman problem by generating as many 3-opt tours from random initial tours as we can afford time-wise and then choosing the best among the 3-opt solutions as our "conjectured solution". The merits of this heuristic approach based on probability as compared to the usual approach of using further complicated refinements to transform locally optimal solutions into global optima will be discussed later in the paper.

IV. THE GENERAL APPROACH

Since we have found that a 3-opt tour has a nontrivial probability of being optimal, we make, for a given problem, an estimate of this probability* (of success) P_s and produce from our program k 3-opt tours (not necessarily distinct) from random initial starts. We choose k so that $1 - (1 - P_s)^k$ is as close to 1 as we desire. Since the running time in obtaining each 3-opt tour is reasonable (25 to $30n^3$ microseconds where n is the number of cities), we can indeed afford the luxury of making k large. For example, a 30-city problem can reasonably be expected to be "solved" in 75 seconds with $k = 100$. At any rate, the best of the k locally optimal solutions, even though it may not actually be the best, will be close enough to the best solution as to offer a satisfactory answer for most practical problems arising in actual applications. Also, a large set of "satisfactory" locally optimal solutions may give an engineer more flexible choice of a solution that he may use satisfying further nonessential but nevertheless desirable features which may be hard to program.

When the number of cities involved is rather large, say >30 , the number of locally optimal solutions that needs to be generated in order to be reasonably assured of having the optimal solution will be very large, as is expected. Incorporated in the program, is a reduction scheme whereby information gained from an initial set of locally optimal solutions is used to reduce the size of the problem, thereby decreasing sub-

* This probability, in general, depends on the size and nature of the problem. From the statistics collected after running many problems, we shall give a heuristic estimate in Appendix C.

stantially the time involved in generating additional locally optimal solutions.

A brief description of the computer program to produce 3-opt tours is given in Appendix B. We mention here an alternate characterization of a 3-opt tour which is more graphic and which we use in our program. A tour T is said to be *optimal relative to inversion and insertion* if, for every k , no section of k consecutive cities in T , say $(i_{\alpha+1}, i_{\alpha+2}, \dots, i_{\alpha+k})$, can be removed from T and reinserted (as is, or inverted) between any two consecutive remaining cities to produce a tour of lesser cost. We prove the following:

Theorem 6: A tour T is optimal relative to inversion and insertion if and only if it is 3-optimal.

Proof: A tour T is not 3-optimal, if and only if there exists 3 links, say u_{ij} , u_{kl} , u_{mn} which may be exchanged by 3 other links say u_{im} , u_{jl} and u_{nk} , (as in Fig. 1, other possibilities are similar) to form a tour of lesser cost. From Fig. 1, we see that the section from m to l may be inserted between i and j and hence the tour is not optimal relative to inversion and insertion.

V. GENERAL DESCRIPTION OF THE METHODS USED TO PRODUCE 3-OPT TOURS

In the process of obtaining 3-opt tours from a random initial tour, the basic operation consists of determining whether any section of length k in the present tour can be inserted (as is, or inverted) between two other neighboring cities so as to decrease the tour cost. This was proved in Section IV (Theorem 6) to be equivalent to exchanging three links in

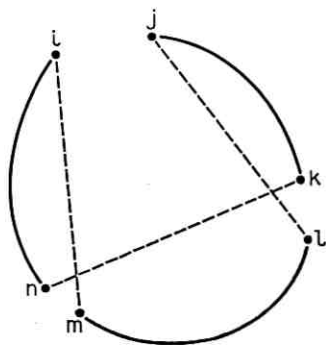


Fig. 1 — Proof of Theorem 6.

the tour for three other links. Once an improvement is found and made, we treat the resulting tour as our initial tour and iterate the process. The process terminates with a locally optimal (3-opt) tour when improvement cannot be further achieved. We call the portion of the computation from the time we made the last improvement to the verification that no further improvement can be achieved by this algorithm the "check-out period." The time involved in the "check-out period" is proportional to $\binom{n}{3}$. For different random initial tours however, the number of improvements may vary and the time it takes to find each improvement also varies. This accounts for some variation in the computation time for the individual locally optimal solutions. However, it turns out that the average computation time for a set of 10 or more cases is uniformly around $50n^3$ microseconds, which is further reduced by the techniques discussed below.

From our experiments, we find that links used in a locally optimal solution are often exchanged in and out many times in the improvement process, and this tends to increase the computation time considerably. Two methods are incorporated in the program to reduce this. First, after each improvement, the improved tour $(t_1', t_2', \dots, t_n')$ is further perturbed by a rotation (see Appendix B) so as to prevent the new links just inserted from being removed again too soon. Secondly, the following special feature making use of locally optimal tours previously obtained is used: After m 3-opt tours T_1, T_2, \dots, T_m ($m = 1, 2, \dots$) are generated, consider the set S consisting of the union of the links found in T_1, T_2, \dots, T_m . In the process of obtaining the $m + 1$ th 3-opt tour, we systematically break off 3 links $u_{t_1 t_n}, u_{t_k t_{k+1}}, u_{t_j t_{j+1}}$ in (t_1, t_2, \dots, t_n) to see if they can be replaced by 3 other links so as to form a tour of lesser cost. In the algorithm (see Appendix B for details), $u_{t_1 t_n}$ is held fixed while k goes from 1 to $n - 3$, coupled with each j going from $k + 1$ to $n - 1$. We skip this sequence of tests for possible improvements altogether if $u_{t_1 t_n} \in S$, and proceed as if all such tests for possible improvements fail. The tour (t_1, t_2, \dots, t_n) is "rotated" by the substitution $(t_n, t_1, t_2, \dots, t_{n-1}) \rightarrow (t_1, t_2, \dots, t_n)$ and the improvement process continues. When no further improvements can be made relative to this special feature, we obtain a tour which we call an "almost 3-opt tour." This almost 3-opt tour is then put through the algorithm without the special feature to obtain a final 3-opt tour.

This process may seem roundabout, but in effect it results in postponing the replacements of links which have occurred in other locally optimal tours until other replacements have been tried. Actually, an

almost 3-opt tour often has most, if not all, of its links in S so that its check-out period is almost negligible. The time required to obtain the final 3-opt tour is also usually quite small. The over-all effect is that we are able to find improvements much sooner and also the number of improvements made in reaching the locally optimal solution is significantly decreased. As a result, for a set of about 10 locally optimal solutions, we are able to reduce the total computation time by at least 40 percent.

VI. THE REDUCTION PROCEDURE

After a certain number, say r , of locally optimal solutions are obtained, consider the set I of links common to all those locally optimal solutions. Intuitively, we feel that since the r 3-opt tours are produced from randomly generated initial tours, any link in I should have a very high probability of belonging to an optimal solution when r is reasonably large. Further, for each problem certain simple features (like some obvious links connecting two cities) of the optimal solution should be reflected in a majority of 3-opt tours so that we expect the set I to be frequently nonempty. Using I , we can reduce the size of the problem as follows: A link u_{ij} is called *basic* if u_{ij} is in I , and a city i is removed if there are two basic links u_{ij} , u_{ik} incident at i . The procedure can of course remove many strings of cities at the same time. If u_{ji_1} , $u_{i_1 i_2}$, \dots , $u_{i_p j'}$ are all basic and no other basic links are connected to cities j and j' , the string of cities i_1, i_2, \dots, i_p is removed. We call cities j and j' *corresponding terminals*. If a total of t cities are removed, we then solve for 3-opt tours in the remaining $n - t$ cities. By reassigning artificial link costs to all links $u_{jj'}$, where j and j' are corresponding terminals, we make sure that j and j' will be neighboring cities in the solution to the reduced problem, and hence the string of cities between j and j' which were removed can be reinserted accordingly. This process can be iterated as many times as we please. Note that we tacitly assume an optimal tour contains the cities $j, i_1, i_2, \dots, i_p, j'$ as a substring. If this is the case, we say that the reduction is *proper*. Otherwise, the reduction is *improper* and the optimal tour will be missed in all future 3-opt tours generated in the same run. However, even if this should happen (rare if r is large or $n \leq 30$), the best tour obtained usually has a tour cost differing from the optimal tour by an extremely small amount. For large n , several independent runs should be made to guard against the possibility of an improper reduction.

When the number of cities involved is fairly large, this reduction

procedure is very effective in reducing computation time required to obtain the optimal solution as the results given in the next section show. The reduction procedure also provides a large variety of hard problems (involving a smaller number of cities) from which we can learn a great deal statistically about the characteristics of 3-opt tours. We consider those problems harder than randomly chosen problems because they retain essentially the heart of the original larger problem. The probability that a 3-opt tour generated from a reduced problem is optimal relative to the reduced problem is usually lower than the mean probability for random problems of the same size. However, in spite of the fact that improper reduction may occur, the probability that a 3-opt tour generated from a reduced problem is optimal relative to the original problem cannot be decreased. Since problem-size is reduced, more 3-opt tours can be obtained in a given time, and thus the probability of finding the optimal solution in a given amount of computation time is greatly increased.

VII. COMPUTATIONAL RESULTS

7.1 *Twenty and Fewer City Problems*

Six problems whose cities are points (x_i, y_i) generated randomly in a 100×100 square and of sizes ranging from 12 to 20 cities were tested. For these cases, 5 to 10 3-opt tours were generated per problem. It turned out that in each problem all 3-opt tours generated were identical, and the costs of the solutions obtained in all six problems were as good as, or better than, solutions obtained by other methods (3 of which are known to be optimum). It appears that randomly generated problems are easy to solve by our method.

Forty 3-opt tours were generated for the 20-city problem of G. A. Croes,³ which has a known optimal solution with cost 246. Reduction was used with $r = 8$. Successive stages of reduction reduced the number of cities from 20 to 11, 11, 11, 11, (i.e., no further reduction produced after the second round). The optimal tour appeared 13 times out of 40 and the total computation time used for the 40 3-opt tours was 3.43 seconds. This 20-city problem seems "harder to solve" than most 20-city problems we have encountered.

Many more problems with sizes around 20 cities obtained from the reduction process of larger problems were investigated. Judging from all the results, we believe that we can "solve" any 20 or fewer city problem by our method in (very conservatively) 5 to 10 seconds.

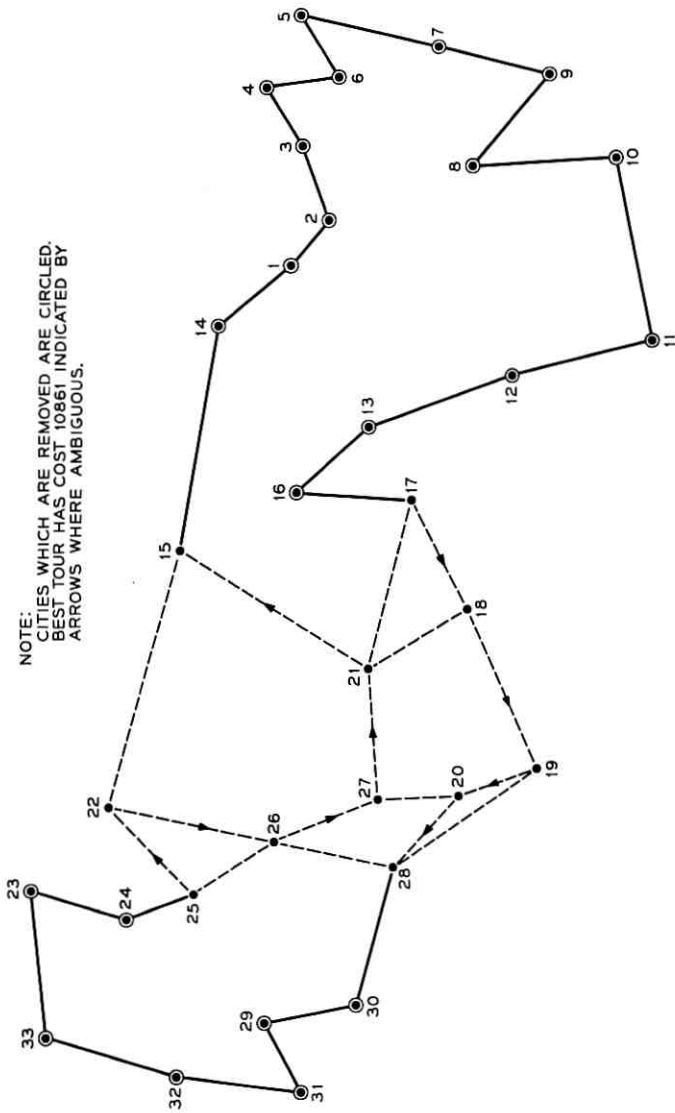


Fig. 2 — 33-City problem showing the link sets *I* and *S* after the first, second, and third stages of reduction.

7.2 *The 25-City Problem of Held and Karp*²

Forty 3-opt tours were generated with reduction in sets of 10. The reduction procedure reduced the size of the problem from 25 to 10 to 7 and 7. The optimal tour (cost 1711) appeared 26 times out of 40. Total computation time was 5.24 seconds. It is interesting to note that there are 7-city problems produced as a result of our reduction for which 3-opt tours are not necessarily optimal.

7.3 *The 33-City Problem of Karg and Thompson*⁴

This 33-city problem seems very easy to solve by our methods. Fifty 3-opt tours were generated with reduction in sets of 10. The reduction procedure reduced the size of the problem from 33 to 11, 11, 11, and 9. The optimal tour (cost 10,861) appeared 19 times out of 50 and the total computation time was 10.7 seconds. Figs. 2 and 3 illustrate some stages in the reduction process. The solid lines indicate links in the set I , and together with the broken lines, form the set S of all links found in the 3-opt tours generated.

7.4 *The 42-City Problem of Dantzig, Fulkerson and Johnson*¹

A 42-city problem was solved by Dantzig, Fulkerson and Johnson¹. The optimal tour has cost 699. Forty 3-opt tours were generated with the optimal tour appearing 11 times. Total computation time was 36.3 seconds. The successive stages of reduction and other pertinent information obtained are given in Table I. Note that d , the number of distinct 3-opt tours obtained per round, decreases, indicating that for smaller problems, there are not too many distinct 3-opt tours and hence the probability that a 3-opt tour is indeed optimal is quite large.

7.5 *The 48-City Problem of Held and Karp*²

In Ref. 2, Held and Karp obtained the "best" solution to this 48-city problem with cost = 11,470. D. W. Sweeney (private communication) later found a tour with cost = 11,461. We strongly conjecture that this is indeed the optimal tour and shall consider it as such for the purpose of our work.

Statistics collected on numerous runs indicate that a 3-opt tour has a probability $p \approx 0.05$ of being optimal, with each 3-opt tour obtained in the average computation time of 2.80 seconds.

Without the reduction procedure, we need to spend about 280 seconds to produce 100 3-opt tours if we want a probability of 0.99 of obtaining

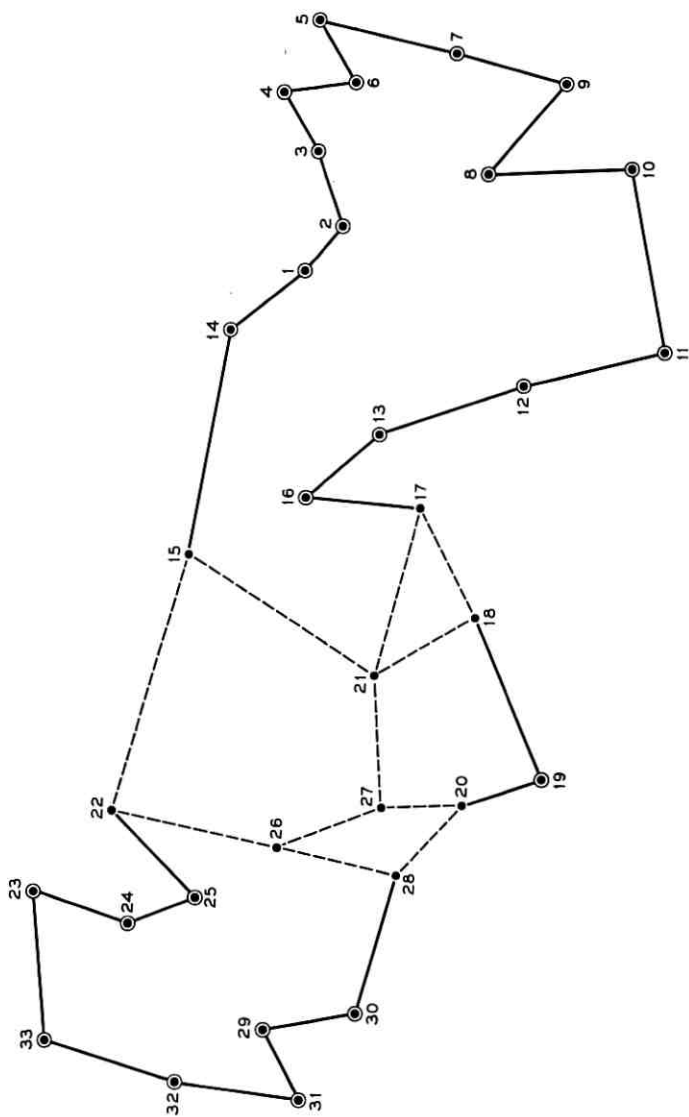


Fig. 3 — As a result of 4th round reduction cities 19 and 25 are also removed.

TABLE I—42-CITY PROBLEM SUMMARY

<i>SRR</i>	<i>n</i>	<i>r</i>	<i>d</i>	C_M	C_m	\bar{C}	<i>t</i>	<i>T</i>
1	42	10	6	734	699	713	1	20.8
2	30	10	5	713	699	704.2	4	28.3
3	24	10	5	713	699	705.9	2	32.2
4	24	10	4	713	699	703.7	4	36.3

SRR: Successive rounds of reduction.

n: Number of cities in reduced problem.

r: Number of 3-opt tours generated per round.

d: Number of distinct 3-opt tours obtained per round.

C_M : Maximum tour cost (for the current round).

C_m : Minimum tour cost (for the current round).

\bar{C} : Average tour cost (for the current round).

t: Number of occurrences of the best tour in the current round.

T: Total time of computation in seconds (accumulated).

the optimal solution. With reduction, setting $r = 10$, we obtained the optimal solution 21 times in a total computation time of 63 seconds, as shown in Table II. When d drops below $r/2$, we consider the resulting reduction too binding in evaluating the heuristic probability p . For example, from this particular run, we count 4 out of 50 instead of 21 out of 100 as the frequency of occurrences of the optimal tour.

TABLE II—A TYPICAL 48-CITY RUN

<i>SRR</i>	<i>n</i>	<i>r</i>	<i>d</i>	C_M	C_m	\bar{C}	<i>t</i>	<i>T</i>
1	48	10	10	11,887	11,470	11,666.8	1	28.4
2	37	10	8	11,989	11,474	11,622.9	1	41.4
3	34	10	8	11,716	11,461	11,592.9	1	51.0
4	27	10	5	11,704	11,461	11,566.2	1	56.1
5	20	10	5	11,556	11,461	11,511.7	2	58.3
6	18	10	5	11,556	11,470	11,527.2	1	60.0
7	17	10	2	11,556	11,461	11,508.5	5	61.4
8	10	10	2	11,556	11,461	11,480	8	62.0
9	10	10	2	11,556	11,461	11,537	2	62.5
10	10	10	2	11,556	11,461	11,537	2	63.0

7.6 The 57-City Problem of Karg and Thompson⁴

In Ref. 4, Karg and Thompson introduced a 57-city problem and found by their method tours with costs of 12,986 and 12,985. In Ref. 5, Reiter and Sherman developed a family of algorithms and found two tours which are better with costs 12,955 and 12,967.* Our program also

* The next best tour we obtained is one with cost 12,966. We believe this to be the same tour and the difference due to a discrepancy of 1 unit in the distance of one particular link used in the tour. Our distance matrix was obtained from Ref. 4.

produced the tour with cost 12,955 which we conjecture to be the optimal solution.

Statistics collected on more than 1000 3-opt tours indicate a probability ≈ 0.02 for a 3-opt tour to be optimal.

A typical run of 100 cases with reduction in sets of 10 appears in Table III.

A few highlights in the reduction process are illustrated in Figs. 4 and 5 below.

An example of an "unfortunate" run where a link not in the optimal tour is committed in the early stages of reduction is shown in Table IV. Note that improper reduction appears in round 3 and as a consequence the subsequent values of d drop sharply. For the purpose of counting the occurrences of the optimal tour, only the first 3 rounds are considered (subsequent rounds have $d < r/2$) giving us 0 out of 30 for this run. As can be seen, we do no worse than to obtain the best tour obtained by Karg and Thompson in Ref. 4. Furthermore, the computation time involved in the first round usually exceeds 40 percent of the total computation time so that even when improper reduction happens, the total computation time is still less than that for obtaining 30 3-opt tours without reduction.

7.7 A 105-City Problem

To test the effectiveness of our method, a 105-city problem was constructed from the 48-city problem and 57-city problem using the facts that $u_{30,33}$ is the largest link (cost 669) in the best tour for the 48-city problem and $u_{40,46}$ (cost 685) is the largest link in the best tour for the 57-city problem. Thus, city 30 of the 48-city problem was connected to city 40 of the 57-city problem by a link with cost 10; similarly, city 33 of the 48-city problem was connected to city 46 of the 57-city problem

TABLE III—A TYPICAL 57-CITY RUN

SRR	n	r	d	C_M	C_m	\bar{C}	t	T
1	57	10	10	13,456	12,986	13,175.5	1	50.5
2	42	10	10	13,427	12,985	13,179.5	1	68.8
3	36	10	9	13,262	12,985	13,092.6	1	80.7
4	34	10	10	13,416	12,985	13,122.3	1	91.3
5	34	10	8	13,299	12,966	13,120.1	1	101.0
6	33	10	8	13,340	12,985	13,173.4	1	110.9
7	33	10	6	13,214	12,955	13,046.5	3	119.8
8	33	10	8	13,346	12,955	13,122.3	1	129.3
9	33	10	8	13,300	12,955	13,132.0	1	138.8
10	33	10	8	13,473	12,985	13,156.6	1	149.2

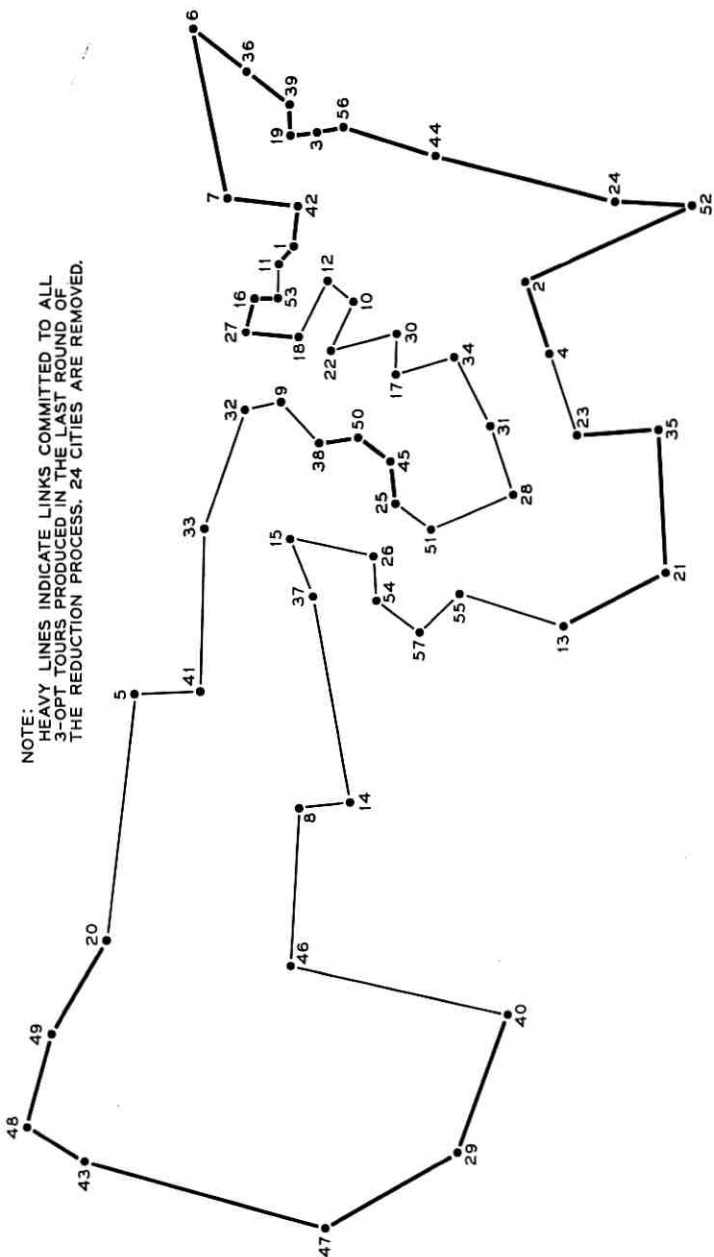


Fig. 5—The conjectured optimum tour in the 57-city problem.

TABLE IV—A 57-CITY RUN WITH IMPROPER REDUCTION

<i>SRR</i>	<i>n</i>	<i>r</i>	<i>d</i>	C_M	C_m	\bar{C}	<i>t</i>	<i>T</i>
1	57	10	10	13,741	12,993	13,430.7	1	45.7
2	45	10	10	13,416	12,986	13,123.6	1	73.4
3*	37	10	9	13,197	12,997	13,091.7	1	87.2
4	30	10	4	13,114	12,985	13,001.8	3	94.2
5	23	10	4	13,012	12,985	12,991.7	2	97.4
6	22	10	3	12,997	12,985	12,990.2	2	100.2
7	21	10	2	12,986	12,985	12,985.2	8	102.7
8	17	10	2	12,986	12,985	12,985.7	3	104.2
9	17	10	2	12,986	12,985	12,985.6	4	105.9
10	17	10	2	12,986	12,985	12,985.7	3	107.2

* Improper reduction appears in this round.

by a link of cost 10. All other links between the cities in the 48-city problem and the 57-city problem were assigned random costs varying from 686 to 750, while links between the cities in the same problem remain unchanged. We purposely made those link costs moderate compared to big links in each of the two problems in order to induce sufficient mixing when a random tour is reduced to a 3-opt tour. From the method of constructing the problem there is a tour (the conjectured best tour) for this 105-city problem for which the cost is $(12,955 + 11,461 + 20) - (669 + 685)$ or 23,082. Since the probability of obtaining the tours with costs 12,955 and 11,461 are ≈ 0.02 and 0.05, respectively, we expect that the probability of a 3-opt tour being optimal in this 105-city problem is less than 0.001. Computation time per 3-opt tour without reduction is about 35 seconds. A run of 20 3-opt tours was made with one round of reduction which reduced the number of cities from 105 to 81. Total computation time for the 20 3-opt tours was 476 seconds for an average of 23.8 seconds per local optimum. Although we did not obtain the conjectured best solution the 3-opt tour costs are surprisingly good; the worst being 24,581 and the best 23,096. An interesting fact about the 3-opt tours obtained is that besides the two short links bridging the two problems, at least one other pair of links connecting two cities in different problems appear in all the 3-opt tours obtained, indicating that this 105-city problem has a structure by itself and is not merely the conjunction of two separate problems. This is indeed what we intended it to be.

This we believe is the largest traveling salesman problem ever attempted. Reasonable estimates indicate that we may be able to solve a problem of this size with the reduction technique in 100 minutes* with a probability of success > 0.5 .

* See Appendix C.

7.8 Other Experiments

Experiments with 2-opt tours were moderately successful for smaller problems. A 20-city, 2-opt tour may be obtained in approximately 0.048 seconds with a probability of being optimum ≈ 0.06 . The decrease in time compared with 3-opt tours is by a factor of 5. The decrease in probability shows that our 3-opt procedure may still be the best. When the number of cities becomes larger, the decrease in probability is so sharp as to make runs with the 2-opt procedure undesirable.

Similar experiments with programs to produce 4-opt tours show an increase of computation time per local optimum by a factor of $0.8n$,* while the chances of obtaining the optimal solution are not noticeably increased. Practically all 3-opt tours are 4-opt and the ranges of tour costs are not noticeably improved.

VIII. CONCLUSIONS AND DISCUSSION

As mentioned earlier, the methods we used here in solving the traveling salesman problem were based upon heuristic principles believed to be of general applicability to various optimization problems. These may be roughly summarized as follows:

8.1 Probabilistic Approach vs Deterministic Approach

In dealing with problems similar to a large traveling salesman problem, where a really efficient algorithm for the best solution is unavailable, it is in general time consuming, if not entirely hopeless, to work on refinement techniques to obtain the best solution. Rather, the approach should be to develop a technique by which good locally optimal solutions can be obtained very fast, and with reasonable probability that among the locally optimal solutions, we may indeed find the best. (In actual applications, the best of a set of good locally optimal solutions, even though it may not be the best, will be close enough to the best solution as to offer a satisfactory answer for most problems.) The fact that we generate for a given traveling salesman problem a large number of 3-opt tours rather than develop means of further improving a 3-opt tour is based on this principle. The fast computation gives us the ability to generate many locally optimal solutions which, coupled with a reasonable probability that a locally optimal solution is optimal, guarantees us a very high probability of success for solving the problem.

* Although the ratio of $\binom{n}{4}$ and $\binom{n}{3}$ is $\frac{n-3}{4}$, for each 4 links removed, there are 48 ways of putting the 4 strings together, compared with 8 ways of putting 3 strings together.

8.2 *Choice of Algorithm*

Consider two algorithms A_1 and A_2 which will produce locally optimal solutions for a given problem in computation times t_1 and t_2 . Suppose A_1 is "stronger" than A_2 in the sense that A_1 produces locally optimal solutions which are optimal more frequently than A_2 , say A_1 with probability p_1 and A_2 with probability p_2 and $p_1 > p_2$. A_1 need not be preferred to A_2 if t_1 is disproportionately greater than t_2 . This can be seen as follows: for a given problem, suppose we are permitted a total computation time t , (which may be the amount of computing time we are able to buy with available funds), so that in time t , we can perform $[t/t_i] = k_i$ experiments with algorithm A_i . Then the probability that among the k_i locally optimal solutions obtained, we indeed have the optimal solution is $p_i^* = 1 - (1 - p_i)^{k_i}$. This is the quantity we should maximize. In the event that we are interested in only good approximate solutions, we should choose an algorithm A_1 over another algorithm A_2 such that for given amount of time t , the best of the k_1 locally optimal solutions relative to A_1 is better than the best of the k_2 locally optimal solutions relative to A_2 .

As an example, consider the sequence of algorithms A_λ to produce λ -opt tours with associated probabilities p_λ and computing times t_λ . For the range of the size of problem we are dealing with, say from 10 to 100 cities, we have reason to believe that p_3^* is the largest among the p_λ^* 's, as indicated by our computation results, and that the best tour produced from k 3-opt tours is at least as good if not better than the best λ -opt tours produced in comparable time using any other A_λ .

8.3 *Random Improvement vs Steepest Descent*

Within the algorithm for obtaining a locally optimal solution, substantial saving in time can be achieved by not attempting to find the best improvement possible at any stage, but rather to take the first improvement that occurs. In general, the method of steepest descent tends to increase the computation time disproportionately and should not be used. Attention should be directed to finding improvements with a minimum amount of computation rather than to making the maximum improvement possible at each step.

8.4 *Restricting Search in Increasingly Smaller Domain*

When sufficient information has been gathered about the problem, ways and means should be investigated to restrict substantially the domain of search. This should be done even with the possibility that

the optimal solution may be lost in the process (of course only with small probability). This is well illustrated by our reduction procedure described in Section VII.

IX. ACKNOWLEDGMENT

The author would like to thank T. H. Crowley and A. J. Goldstein for many stimulating ideas which materially helped in producing these results.

APPENDIX A

k-Length String Optimization

In the first algorithm, called *k*-length string optimization, a dynamic programming technique is used to optimize tours (or strings) of *k* cities for $k \leq 13$.* The algorithm is a slight modification of that given in Ref. 2 by Held and Karp. However, we achieve a significant reduction in computation time by taking advantage of the fact that the distance matrix is symmetric.

Let the *k* cities be represented by the integers 1 through *k*, $X = \{2, 3, \dots, k\}$ be the set of integers from 2 to *k* and *S* be a subset of *X* consisting of *m* elements; i.e., $|S| = m$. Let $C(S, i)$ with $i \in S$ denote the minimum cost of starting from city 1 and visiting all cities *j* in *S*, terminating at city *i*. Then the quantities $C(S, i)$ can be computed recursively as follows:

$$C(\{i\}, i) = d_{1i} \quad (1)$$

$$C(S, i) = \min_{j \in S-i} [C(S - i, j) + d_{ji}]. \quad (2)$$

For example, if $S = \{3, 5, 9, 11\}$, then $C(S, 9) =$ cost of best string from 1 to 9 through 3, 5, 11

$$= \min \begin{cases} C(S - 9, 3) + d_{3,9} \\ C(S - 9, 5) + d_{5,9} \\ C(S - 9, 11) + d_{11,9} \end{cases}$$

We note here that $S - i$ is a subset of *X* such that $|S - i| = |S| - 1$, and thus the quantities $C(S - i, j)$ have all been computed a step before in the recursion scheme.

For $k = 2t + 1$, we recursively compute and store $C(S, i)$ for all

* Storage requirements in dynamic programming seriously limit the size of the problem that can be handled.

subsets S of X for $|S| = 1$ up to $|S| = t$. Then we successively compute $C(S, i)$ for $|S| = t + 1$. At this point, if we denote the complement of these S 's in X by \bar{S} , we see that $S^* = \bar{S} \cup \{i\}$ is a subset of t elements and $C(S^*, i)$ has already been computed. The cost r of the optimal tour T is therefore given by

$$r = \min_s [\min_{i \in S} [C(S, i) + C(S^*, i)]]$$

where S ranges over all subsets of X containing $t + 1$ elements. Since either S or S^* must contain say city 2, we may further restrict the range of S to only those containing the city 2.

For $k = 2t$ the procedure is similar except that we need not compute $C(S, i)$ for $|S| > t$.

The order of the cities in the tour T can now be determined. With the "middle" city i and sets S, S^* determined from the expression for r , we find a city i_1 in $S - i$ such that $C(S - i, i_1) + d_{ii_1} = C(S, i)$; then a city i_2 in $S - \{i, i_1\}$ such that $C(S - \{i, i_1\}, i_2) + d_{i_1 i_2} = C(S - i, i_1)$ and so on until S is exhausted, and similarly for the set S^* to produce the tour $T = 1, \dots, i_2, i_1, i, i_1^*, i_2^*, \dots, 1$.

This algorithm can be used, with a slightly modified distance matrix, to find the minimum string from city 1 to city k through the cities 2, 3, $\dots, k - 1$ and its associated cost $C(X, k)$. We set $d_{1k} = C(\{k\}, k) = 0$ and $C(S, k) = \infty$ for $|S| \geq 2$ in the recursive computation scheme described above. This insures that city k is next to city 1 in the tour produced, and hence by removing the link from city 1 to city k we get the best string with 1 as our initial city and k the terminal city.

Given a n -city problem, a tour T through the n cities is said to be locally optimal relative to the k -length string optimization algorithm if every ordered set of k consecutive cities in T , say $(i_{\alpha+1}, \dots, i_{\alpha+k})$, subscripts reduced modulo n , is optimal as a string from $i_{\alpha+1}$ to $i_{\alpha+k}$ going through the cities $i_{\alpha+2}, i_{\alpha+3}, \dots, i_{\alpha+k-1}$. The above program may be used to produce locally optimal tours of this type for any n -city problem without change as follows: Consider a random initial tour represented by the permutation $P = (i_1, i_2, \dots, i_n)$. We map the first k cities i_1, i_2, \dots, i_k onto 1, 2, \dots, k and use the program to find the best string from 1 to k , say 1, j_2, j_3, \dots, k . The associated permutation $P^* = (i_1, i_{j_2}, i_{j_3}, \dots, i_k, \dots, i_n)$ gives us a tour whose cost is no larger than P . When there is no gain in cost from P to P^* it means that i_1, i_2, \dots, i_k is already optimal as a string. Next, we rotate P^* by a length δ (relatively prime to n) and repeat the process until there are n consecutive rotations without a decrease in cost. Then every ordered

set of k consecutive cities in the final permutation is now optimal as a string. Experiments with this procedure indicated, however, that it is time consuming and not nearly as effective as the second algorithm which we have discussed in the paper.

APPENDIX B

Outline of Computer Program to Produce 3-Opt Tours from Random Initial Tours

Notation:

- n number of cities
- (d_{ij}) distance matrix
- r number of 3-opt tours desired before reduction
- u_{ij} link connecting city i and city j
- t_i i th city in the tour
- S the union of the set of links found in previously generated 3-opt tours
- q a program branching parameter.

1. $S = \phi$
2. Do through (14), $m = 1, 1, r$.
3. Generate a random tour (t_1, t_2, \dots, t_n) .
4. $q = 0$ if $m = 1$, otherwise $q = 1$.
5. Do through (12), count = 1, 1, n .
6. If $q = 0$, skip 7.
7. If $u_{t_1 t_n} \in S$ go to 12 (special feature).
8. Do through (11), $k = 1, 1, n - 3$.
9. Do through (11), $j = k + 1, 1, n - 1$.
10. If $d_{t_k t_{j+1}} + d_{t_1 t_j} \leq d_{t_1 t_{j+1}} + d_{t_k t_j}$ set $d = d_{t_k t_{j+1}} + d_{t_1 t_j}$ and $\alpha = 16$, otherwise set $d = d_{t_1 t_{j+1}} + d_{t_k t_j}$ and $\alpha = 18$.
11. If $d + d_{t_{k+1} t_n}$ (cost of links added) $<$ $d_{t_1 t_n} + d_{t_k t_{k+1}} + d_{t_j t_{j+1}}$ (cost of links removed) go to α (otherwise loop).
12. $(t_1, t_2, \dots, t_n) = (t_n, t_1, \dots, t_{n+1})$ (rotate).
13. If $q = 1$, set $q = 0$ and go to 5 (almost 3-opt tour obtained).
14. $S = S \cup$ links in (t_1, t_2, \dots, t_n) (3-opt tour obtained).
15. Go to reduction (see description in Section VI).
16. $(t_1, t_2, \dots, t_n) = (t_{j+2}, \dots, t_n, t_{k+1}, \dots, t_j, t_1, \dots, t_k, t_{j+1})$ (links exchanged and tour perturbed).
17. Go to 5 (treat improved tour as initial tour).
18. $(t_1, t_2, \dots, t_n) = (t_{j+2}, \dots, t_n, t_{k+1}, \dots, t_j, t_k, \dots, t_1, t_{j+1})$.
19. Go to 5.

APPENDIX C

Estimates on the Probability that a 3-Opt Tour is Optimal

From the statistics collected on running many problems, we estimate that a 3-opt tour in a 10-city problem of some difficulty has a probability of 0.5 of being optimal, and in general, this probability seems to decrease by a factor of 2 for each addition of 10 cities. We shall denote these estimates by $p(n)$ and use them as basis for our calculations. We have

$$p(n) \approx 2^{-n/10}.$$

So far, these estimates has been close for problems in the range between 30 and 60 cities and is too conservative for smaller problems, or exceptionally easy problems like the 33-city problem. For exceptionally difficult problems, we define $p^*(n) = \frac{1}{4}p(n)$ as the estimate for the "worst." Computation time $t(n)$ per 3-opt tour in an n -city problem averages $30n^3$ microseconds without reduction. With reduction, we can obtain 100 3-opt tours usually in the amount of time needed for 25 3-opt tours without reduction.

A few examples will show how to estimate time needed to "solve" a given problem. We consider a problem solved if the probability p of obtaining the optimal solution is ≥ 0.99 . It should be noted that the estimates are heuristic in nature and tend to be on the conservative side.

Example 1 — Given a 20-city problem, we have $p(20) = \frac{1}{4}$, $t(20) = 0.24$ second. In order to have

$$\left(\frac{3}{4}\right)^k \leq 0.01$$

we must have $k > 16$. Thus 4 seconds of computation should be adequate. In the "worst" case $p^*(20) = \frac{1}{16}$, $k = 72$ should be adequate. Computation time is about 18 seconds if reduction is not used and about 5 seconds if reduction is used.

Example 2 — For a 40-city problem we have $p(40) = \frac{1}{16}$, $t(20) \approx 2$ seconds. With $k = 72$, we need 144 seconds without reduction and about 40 seconds with reduction. In the "worst" case $p^*(40) = \frac{1}{64}$ and $k = 300$ is sufficient. Running the program in 3 independent runs of 100 with reduction, total computation time needed is about 2.5 minutes.

Example 3 — For a 60-city problem, we have $p(60) = \frac{1}{64}$, $t(60) \approx 6.5$ seconds. With $k = 300$ and using reduction, about 8 minutes of computa-

tion time is required to guarantee $p > 0.99$. In the "worst" case with $p^*(60) = \frac{1}{250}$, the same computation will yield $p \approx 0.7$.

Example 4 — For a 100-city problem $p(100) = \frac{1}{1024}$ and $t(100) \approx 30$ seconds. With $k = 800$ we have $p \approx 0.54$. With reduction, this can be achieved in about 100 minutes.

APPENDIX D

Two Conjectures

Using the notation of Section III, the following appears to be an interesting problem: Find the minimum number k for which $C_k = C_{k+1} = \dots = C_n$. For fairly large k it appears, at least intuitively, that a k -optimal tour should be optimal, since by Theorem 1 (Section III), if it is not optimal, then its index can be at most $n - k - 1$. Due to the intrinsic difficulties in this problem we can only state the following conjectures:

Conjecture 1 — $C_{n-1} = C_n$. That is, any tour which is $(n - 1)$ -optimal is also optimal.

Conjecture 2 — $C_k = C_{k+1} \rightarrow C_k = C_{k+1} = \dots = C_n$.

Conjecture 1 can be verified for $n \leq 6$. Also an $(n - 1)$ -optimal tour which is not optimal must have index 0, hence all of its links are inadmissible. Furthermore, it must also be n -length string optimal. The existence of such a tour seems to be extremely unlikely. Work on Conjecture 1 has led to the following interesting problem in graph theory, which is equivalent to Conjecture 1, and yet involves no concept of any distance matrix.

Problem: Suppose we are given a graph with n vertices and $2n$ links which can be partitioned into 2 sets of n links, each of which form a Hamiltonian circuit. Does there exist another partition with the same property?

If the answer to the above problem is always in the affirmative, then we can prove Conjecture 1 in the following way. Consider a graph consisting of an optimal tour T and an $(n - 1)$ -optimal tour T^* which is not optimal. Since T^* has index 0, T , and T^* have no link in common and thus serve as a partition into 2 sets of n links, each of which form a Hamiltonian circuit. Let the other partition with the same property be tours T_1 and T_2 . Since T_1 and T_2 uses the same set of $2n$ links of T and T^* ,

$$C(T_1) + C(T_2) = C(T) + C(T^*) < 2C(T^*).$$

Hence one of the tours, say T_1 must have cost $C(T_1) < C(T^*)$. But T_1 and T^* must have at least one link in common; hence T^* cannot be $(n - 1)$ -optimal.

On the other hand, suppose there is a graph with n vertices and $2n$ links which can be partitioned into 2 sets of n links A and B , each of which form a Hamiltonian circuit and that no other partition with the same property is possible. Let the n vertices represent n cities. We construct a distance matrix as follows: let each link in A be assigned a distance d_a ; each of the $n - 1$ links of B be assigned a distance d_b and the remaining link in B , d . Let all other links in the complete graph of n nodes have distance $> \max [n \cdot d_a, (n - 1)d_b + d]$. Suppose $nd_a < (n - 1)d_b + d$. Then it is clear that the set of n links in A form the optimal tour while the set of n links in B form an $(n - 1)$ -optimal tour. Furthermore, we can make $d_a > d_b$ so that the $(n - 1)$ -optimal tour which is not optimal contains $n - 1$ smallest links, and by making d large, we can also make the "next best" tour as poor as possible compared with the optimal tour.

Conjecture 2 is obviously true for $k = n - 1$. We prove it is true for $k = 1$. By hypothesis, $C_1 = C_2$, that is, no tour crosses itself. Suppose there is a tour T which is not optimum, then we may transform T into the optimal solution by a sequence of transpositions of immediate neighbors, each step being equivalent to an inversion. Since the cost must finally reduce to the cost of the optimal tour, at some point we must have a tour with the property that transposing 2 immediate neighbors reduces the cost, giving us a crossover situation contradicting the hypothesis.

REFERENCES

1. Dantzig, G. B., Fulkerson, D. R., and Johnson, S. M., Solution of a Large-Scale Traveling Salesman Problem, *Oper. Res.*, 2, 1954, pp. 393-410.
2. Held, M. and Karp, R. M., A Dynamic Programming Approach to Sequencing Problems, *J. Soc. Ind. Appl. Math.*, 10, No. 1, March, 1962, pp. 196-210.
3. Croes, G. A., A Method for Solving Traveling Salesman Problems, *Oper. Res.*, 5,
4. Karg, R. L. and Thompson, G. L., A Heuristic Approach to Solving Traveling Salesman Problems, *Manage. Sci.*, 10, No. 2, January, 1964, pp. 225-247.
5. Sherman, G. and Reiter, S., *Discrete Optimizing*, Inst. Quant. Res. Econ. and Manag., 37, Purdue University, 1963.
6. Flood, M. M., The Traveling Salesman Problem, *Oper. Res.*, 4, 1956, pp. 61-75.

On Definitions of Congestion in Communication Networks

By ERIC WOLMAN

(Manuscript received August 18, 1965)

Beginning with the familiar ideas of time-congestion and call-congestion for a full-access trunk-group, this paper considers the relations between various measures of congestion for networks of more general structure. The discussion is based on a simple heuristic model which makes the definitions of time- and call-congestion directly comparable. This model leads to a two-dimensional classification of measures of congestion, which should be useful in treating types of systems not mentioned here. Three important papers in the literature are analyzed in terms of the proposed classification

I. INTRODUCTION

The concepts of *time-congestion* and *call-congestion* have been in common use since the early days of telephone traffic theory. Both ideas are quite simple when applied to a single set of devices used in telephony, such as a full-access trunk-group. But applications of these ideas in the theory of networks of more general structure, composed of elements arranged both in series and in parallel, have not always been consistent, either internally or with each other. This paper describes an attempt to resolve some of the difficulties, which may have arisen because the simplicity of many "classical" models renders unnecessary some distinctions which are important in the general case.

This section describes the nomenclature used below and the assumptions on which the remainder of the paper rests. The following section treats a useful model in the context of a full-access trunk-group. In the third section, similar ideas are applied to the theory of general communication networks. The fourth section contains a brief discussion of some of the switching literature, in the light of Section III. Some conclusions appear in the fifth section. In order to save space, I propose the abbreviations "CC" for call-congestion and "TC" for time-congestion.

I use the following terminology. A single element of a system carrying traffic is *busy* or *idle*. This choice is binary and tells whether or not

the element can honor a request for service. When a set of elements is arranged so as to form all or part of a communication network, the events {path} and {no path} correspond respectively to the existence and non-existence of a chain of idle elements connecting two specified points of the network. A *call-attempt*, which is a request for an idle path connecting two points of a network, is *blocked* if no such path exists. Blocking is a binary concept. "Congestion" refers to a non-vanishing *probability of blocking*, which takes values on the interval $[0,1]$.

Analysis of congestion refers to a specified portion of a communication system. This may be one group of trunks, considered in isolation; it may be the switching network connecting incoming to outgoing trunks in a tandem office; or it may be an entire system. When congestion exists, some call-attempts do, or could, fail. A call-attempt must be made by a *source* and must be directed to a *destination*; sources and destinations are *terminals*. Since we may think of call-attempts as simply appearing as inputs to a particular model, a terminal really marks the boundary of the model, which may or may not include the entities originally responsible for call-attempts.

But for stated exceptions, this paper should be read with the understanding that blocked calls are cleared, i.e., that call-attempts which cannot be served immediately are dismissed and have no effect on the system. With blocked calls delayed, there may be more than one relevant notion of blocking. One may ask for the probability of positive delay or of delay exceeding a fixed amount. If only finite queues are possible, one may want to know the probability of entering the queue or of overflow from the queue. The ideas expressed below can easily be extended to such systems, but for simplicity they are introduced here in the context of loss operation (blocked calls cleared), with or without retrials.

The traffic that a source or class of sources offers to a system is described by means of a random process which specifies the instants at which call-attempts occur. The parameters of such processes (with deterministic traffic as a special case) can vary with time, from source to source, or with the states of their sources, of the network, or of other sources. When an expected number of call-attempts in an interval of time is divided by the length of the interval, the quotient is an average *calling-rate*. The present method of elucidating congestion allows, and takes into account, dependence of a calling rate upon its source and upon the state of its source and of the network; but it is simplest, with one exception treated explicitly, to require independence among different sources. There remains only the question of time-dependence.

Let us assume that the stochastic process which describes the operation

of the traffic system of interest is stationary. The method proposed below is based on subdividing a typical interval of time according to the states of various elements of the system. I write as if various quantities were *defined* as ratios of lengths of sub-intervals of an interval of finite length T . For every kind of time-congestion (TC), the definition is actually the limit, under appropriate conditions, of such a ratio as $T \rightarrow \infty$. For the sake of shortening complicated statements, this important distinction will not be mentioned in Sections II-IV.

Similarly the definition of call-congestion (CC), which is what many authors mean by "probability of blocking", is often for the stationary case given as the limit as $T \rightarrow \infty$ of the ratio of the number of *blocked* call-attempts (in $[0, T)$) to the number of *all* attempts (in the same interval). I propose to describe CC and TC in comparable terms by equating call-congestion in a typical case with such a ratio as

$$\frac{\begin{aligned} &[(\text{expected length of blocked time in } T \text{ when attempts are possible}) \\ &\cdot (\text{expected calling-rate when attempts are possible but blocked})] \\ \div &[(\text{expected length of time in } T \text{ when attempts are possible}) \cdot (\text{expected} \\ &\text{calling-rate when attempts are possible})]. \end{aligned}}$$

This procedure, of replacing an expected number of calls by a product of an expected length of time and a conditional expected calling-rate, is valid in cases for which a fraction such as that displayed here has a limit as $T \rightarrow \infty$ which agrees with the corresponding true CC as defined above. The procedure is certainly valid when the traffic system can be described by a stationary Markov chain, as is the case for many systems representable by models with finitely or infinitely many sources, blocked calls cleared or delayed, etc. But so far as I am aware, it is not now known either exactly what systems can be so described in a tractable way, or for what other systems the desired agreement holds. Thus further discussion of the applicability of this approach is deferred until Section V. The additional problems, especially in connection with ratios such as that displayed above, that are encountered in *measuring* the various kinds of congestion form a separate topic and are not discussed here.

The preceding paragraphs make it clear that this paper does not encompass traffic that varies in time (except possibly where we need only finite-time-average values of congestion), although such traffic can be quite simply described. By resorting to ensemble averages — that is, to the conceptual experiment of running a traffic system over and over again from time 0 to time t with statistically identical inputs — it is quite possible to define instantaneous values of congestion from various points of view, although it is difficult to motivate a distinction between

CC and TC. However, such generality would again serve only as a distraction from the main issue. I therefore adopt the assumption of stationarity: For present purposes this is a steady-state (equilibrium) theory.

The reader must see now that this introduction is really a sop to Cerberus. It records the framework of ideas in which the following remarks are couched, but very informally and on the assumption that the reader is thoroughly familiar with the concepts and terminology of traffic theory. I hope the loss of precision inherent in such a heuristic approach is outweighed by the gain in simplicity.

II. CONGESTION IN THE SIMPLEST SITUATION

Let us first consider a single full-access trunk-group \mathcal{G} . Its relevant properties are represented in Fig. 1. It consists of a number of channels. Subscribers at one end communicate with those at the other over a shared communication system which for our purposes consists of \mathcal{G} alone. (We do not discuss communication between two subscribers at one end of \mathcal{G} .) Two particular subscribers are labeled "A" and "B", and we define the *pair* P as the pair (A, B) . Traditional notions of blocking and congestion in this situation refer to the state of \mathcal{G} . Of course in defining CC from A 's point of view we have to distinguish blocked states in which A is busy from those in which A is idle; but the fundamental concept is that of "all trunks of \mathcal{G} busy". This is because the idea is simple and makes sense: Either some trunks are idle or they are all busy; and when all are busy, any subscriber either has a trunk or cannot get one.

Now let us represent a typical period of time T for \mathcal{G} as the sum of periods x , y , u , and v , drawn as intervals in Fig. 2. The label "no path" for intervals x and y means that all trunks of \mathcal{G} are busy, and "path"

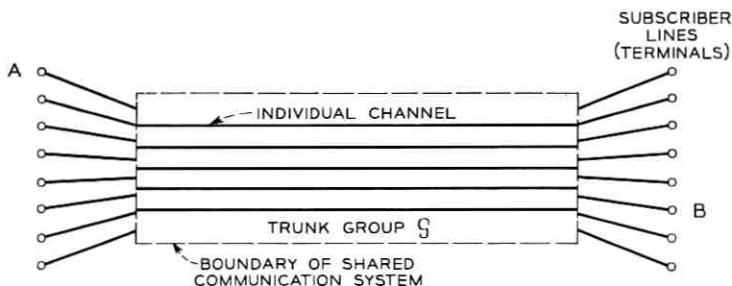


Fig. 1 — Full-access trunk-group.

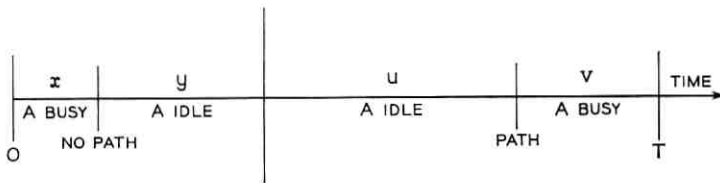


Fig. 2 — Division of time for \mathcal{G} .

means that \mathcal{G} has at least one idle trunk. With these conventions, it is clear that the standard definition gives

$$TC = \frac{x + y}{x + y + u + v} \quad (\text{Fig. 2}).$$

(In each such expression, the number in parentheses specifies the figure in which the symbols are defined.)

The intervals x and y in which there is “no path” are lumped together in the formula for TC , as are u and v , because the wire chief’s point of view ignores the state of terminal A . “The wire chief’s point of view” is a traditional phrase which expresses the possibility of viewing the possible paths between A and B with the interests of the shared communication system rather than of the terminal-pair in mind. This viewpoint reflects a concern for the network itself and its ability to establish connections between pairs of terminals, and is therefore symmetric with respect to A and B . In particular, congestion between A and B is naturally seen as the fraction of time during which no path composed of idle elements exists. The complementary point of view, emphasizing states of the source, is known as that of the “particular subscriber” (a phrase from which I shall often omit the first word). This concept of “point of view” is useful, but requires considerably more discussion in the next section.

Suppose that A ’s calling rate varies, and is on the average $r(A)$ times as large when there is no path through \mathcal{G} as when there is a path. Of course $r(A)$ can represent either an instantaneous effect or one depending on A ’s history. (Note that $r(A)$ incorporates the effects of, but is by no means simply related to, a change in A ’s calling rate triggered by an unsuccessful attempt and persevering until a successful attempt. If $r(A)$ has such a cause, then probably $r(A) > 1$ with human callers; but an automatic calling system might easily be designed to have $r(A) < 1$ in order to reduce the load on the switching equipment caused by unsuccessful attempts.) If A can initiate a call only when idle, then the

customary notion of call-congestion, the proportion of A 's attempts that are unsuccessful, is that

$$CC = \frac{r(A)y}{r(A)y + u} \quad (\text{Fig. 2}).$$

If subscribers are not all alike, this formula may apply only to A , as suggested by the traditional association of CC with the particular subscriber's point of view.

If indeed A attempts calls only when idle, it may be totally uninteresting to him to consider intervals of time during which he is busy. But A may still want to distinguish between, on the one hand, the fraction of his call-attempts which fail and, on the other hand, the fraction of that time in which he *can* initiate calls during which attempts must fail. Thus it is natural to define the *modified* (or *conditional*) time-congestion as

$$MTC = \frac{y}{y + u} \quad (\text{Fig. 2}).$$

Fig. 2 shows clearly that MTC , like CC , measures congestion from the subscriber's point of view.

What relations hold among these three measures of congestion in this situation? First, it is obvious that CC and MTC agree if and only if $r(A) = 1$, and that $CC < MTC$ if $r(A) < 1$ and *vice versa*. Second, suppose that a relation R holds between the modified and ordinary time-congestions, where R is either $<$, $=$, or $>$. Multiplying both sides of the formula $MTC R TC$ by $(y + u)(x + y + u + v)$, we get $y(x + y + u + v) R (x + y)(y + u)$. Subtraction of $(xy + y^2 + yu)$ from both sides yields the result that $MTC R TC$ if and only if $yv R xu$. Whenever $uy \neq 0$, it is helpful to rewrite this condition as

$$MTC R TC \quad \text{if and only if} \quad \frac{v}{u} R \frac{x}{y}.$$

(In certain degenerate cases $uy = 0$. For example, if \mathcal{G} contains c trunks and there are c terminals at each end, $y = 0$ because A cannot be idle when there is no path through \mathcal{G} .) In the finite-source model analyzed by Engset [Ref. 1, pp. 250-1], where $r(A) = 1$, the events $\{A \text{ busy}\}$ and $\{\text{no path}\}$ are positively correlated, so that $(v/u) < (x/y)$. Therefore $MTC < TC$, and this in turn implies the well-known fact that $CC < TC$. But notice that no inequality between MTC and TC inheres in the definitions; it is not inconceivable that in some useful model $\{A \text{ busy}\}$ and $\{\text{no path}\}$ could be negatively correlated events, which would make $(v/u) > (x/y)$ and so reverse the previous inequality.

We care most about the relation between call- and time-congestion. The previous paragraph covers this for the special case in which $r(A) = 1$. The procedure applied above, of cross-multiplication, cancellation, and division by uy , shows that in general

$$CC \ R \ TC \quad \text{if and only if} \quad r(A) \left(1 + \frac{v}{u}\right) \ R \ \left(1 + \frac{x}{y}\right).$$

Unless both CC and TC agree with MTC, we see that they are not very likely to agree with each other. Also, the most natural assumptions about a traffic system in the absence of information to the contrary would be, first, that $r(A) \geq 1$, and second, that $(v/u) \leq (x/y)$ because of a positive correlation between the events {no path} and {A busy}. Unless put in quantitative form, these assumptions lead only to the conclusion that MTC is likely to be smaller than both time- and call-congestion.

For this situation we have defined three types of congestion. The two kinds of time-congestion would agree if the probability of blocking within the shared communication system were independent of the state of the relevant source. CC agrees with MTC when a calling rate is unaffected by the availability of desired paths. CC and TC agree if both agree with MTC, or otherwise when $r(A)[1 + (v/u)] = 1 + (x/y)$. Since these three measures of congestion can differ, all are of interest.

III. CONGESTION IN NETWORKS

A communication network \mathfrak{N} is sketched in Fig. 3; this one happens to be a 3-stage switching network. Before defining anything, we note one fundamental difference from the simple example in Section II. It makes sense to emphasize the dichotomy described by "all trunks of \mathfrak{G} busy" and " \mathfrak{G} has some idle trunks". The isolated, full-access trunk-

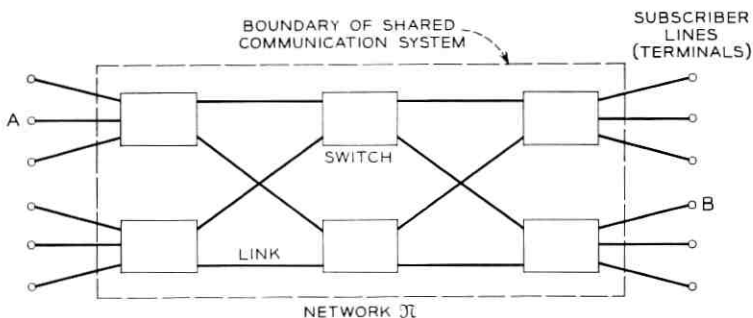


Fig. 3 — Connecting network.

group exemplifies the more general case of congestion in any *one* link, composed of interchangeable (fully accessible) channels, of a communication network. For present purposes we can completely describe such a link at a given instant by saying whether it does or does not contain at least one idle channel. But, except in special cases of little interest, no such concept as “ \mathfrak{X} is busy” can be defined in a useful way. Even the wire chief does not say, “My network is busy”; he says, “No paths exist between the following pairs of customers”. In a general network, whether it be a switching network, a network of trunk groups, or a combination of both to form a complete communication system, the notion of blocking applies only to specific pairs $P = (A, B)$ of terminals. This is why so many investigators immediately fix their attention on the subgraph $g(P)$ of \mathfrak{X} corresponding to the various routes over which A and B could be connected. Fig. 4 shows one conventional way of representing $g(P)$ for the situation of Fig. 3. The nodes are switches and the branches, links.

Let us take $\{P \text{ idle}\}$ to be the event that both A and B are idle, and $\{P \text{ busy}\}$ the event that either A or B or both are busy; that is, $\{P \text{ busy}\} = \{P \text{ not idle}\}$. Then one possible way of subdividing a typical period of time for \mathfrak{X} is shown in Fig. 5. Here the conditions “path” and “no path” refer to the subgraph $g(P)$, so that the entire subdivision of the period T is of interest only to the terminal-pair P . (It will of course relate also to other pairs, if any, whose behavior is in every respect statistically identical with that of P .) Let us call each quantity that is most naturally defined by dividing time as in Fig. 5, *congestion as measured for (or by) the particular pair*. The nature and significance of this convention are considered below.

(One important feature of the representation of Fig. 4 is that no branch of $g(P)$ is uniquely associated with one subscriber. Thus, as seen by the pair P , $g(P)$ is the entire shared communication system, just as \mathfrak{G} was in the example of Section II; though of course the *behavior* of $g(P)$ is affected by traffic passing through the remainder of \mathfrak{X} . With this viewpoint $v \neq 0$ in Fig. 5; for it is quite possible for a path between A and B to exist in $g(P)$ when P is busy, because the terminal nodes of $g(P)$

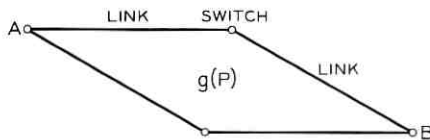


Fig. 4—Subgraph $g(P)$ of \mathfrak{X} .

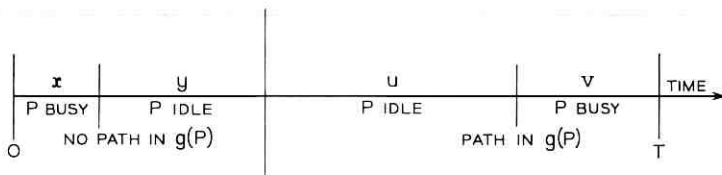


Fig. 5 — Possible division of time for \mathfrak{X} .

are the switches to which A and B have access. I have gathered in conversation that certain traffic theorists have a mental picture of \mathfrak{X} like the one in Fig. 6, in which $g(P)$ includes a branch for each subscriber. The only effect of this change, when it exists, is to make the events $\{P \text{ busy}\}$ and $\{\text{path}\}$ incompatible. In this case $v = 0$ in Fig. 5. The distinction between the conventions of Figs. 4 and 6 is conceptually quite important, because Fig. 6 by itself can suggest the *omission* of

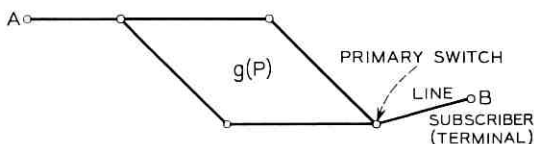


Fig. 6 — Extended subgraph.

v from all our formulae. But the distinction does not affect the character of the results discussed below, and so we assume the model of Fig. 4 without further comment.)

Returning to Fig. 5, it seems natural to define the time-congestion as

$$TC_2 = \frac{x + y}{x + y + u + v} \quad (\text{Fig. 5}),$$

where the subscript "2" refers to measurement for the pair P . This definition is clearly analogous to that of TC in Section II: For TC_2 , being a function of $x + y$ and of $u + v$, is based only on what P sees as its *shared* communication system; the states of the terminals are not taken into account. On the other hand, the intervals $x + y$ and $u + v$ relate to $g(P)$ alone rather than to all of \mathfrak{X} . For these reasons we may call TC_2 "the wire chief's definition for the pair P ", indicating that the *point of view* is the wire chief's while the *measurement* is for (or by) the particular pair.

Fig. 5 leads to natural definitions of MTC and CC (again as measured for the pair), but only if we think of calls as being attempted between A

and B only when P is idle. The usefulness of this convention, which is discussed below, depends on the physical arrangements of the network. It is hard to think of "attempts" in any *other* way in a connecting network with gas-tube crosspoints, in which one tries to set up a connection between two points by establishing a potential-difference between them; if idle paths exist, the gas tubes along one of them will break down and so connect the two points. Some forms of common control tend to lead toward similar ideas; for example, it may be useful to think of a marker as "attempting" to find a route between two devices only if both devices are idle. Here we define the modified time-congestion as

$$\text{MTC}_2 = \frac{y}{y + u} \quad (\text{Fig. 5}),$$

and the call-congestion as

$$\text{CC}_2 = \frac{r(P)y}{r(P)y + u} \quad (\text{Fig. 5}).$$

As in Section II we may distinguish between TC_2 and these quantities by saying that the latter reflect the viewpoint of a pair of subscribers rather than of the wire chief, and thus by calling them "the subscribers' definitions for the pair P ".

With congestion measured for the pair in three ways, what about the network as a whole? Surely we may want to know what fraction of all calls offered to the network are blocked, or what fraction of time finds the average pair unable, for lack of paths in \mathfrak{N} , to establish a connection. These questions are easily answered by averaging the previous quantities over all pairs P . The resulting measures of congestion could be described as "for the office, exchange, network, or system" or as the "office average", etc. Because "system" seems too broad a term, and in order not to emphasize switching applications as opposed to trunking, I choose the term "network" to describe congestion averaged over all pairs. This operation yields the quantities

$$\begin{aligned} \overline{\text{TC}_2} &= \text{avg over } P \text{ of } \text{TC}_2 ; \\ \overline{\text{MTC}_2} &= \text{avg over } P \text{ of } \text{MTC}_2 ; \end{aligned}$$

and

$$\overline{\text{CC}^2} = \text{weighted avg over } P \text{ of } \text{CC}_2 ,$$

where the weighting in calculating the network-average call-congestion is proportional to the average calling-rate for each P . (Such weights may be very hard to find when $r(P) \neq 1$.) Notice that, with homogeneity of

subscribers, averaging has no effect: For example $CC_2 = \overline{CC_2}$ if all subscribers are alike.

In some networks it is possible for A to try to call B whenever A is idle, regardless of B 's state. Step-by-step switching is an obvious example, as is the Bell System voice network including all its subscriber sets. Even with common control, it is possible for a request for connection to initiate search for an idle path even if detection of a busy destination must follow establishment of a path through $g(P)$. In fact there are systems in which it is useful to imagine busy terminals as *initiating* call-attempts. In an electronic central office, for example, a path is sometimes reserved for a connection which is to be established as soon as P becomes idle. In such a system an attempt can occur in the presence of any combination of states of the terminals of a pair, and each such combination may have a different associated calling-rate. The present treatment is restricted to less formidable models. Nevertheless it is important to realize that the situations in which only idle pairs make attempts and in which any idle terminal can make attempts, are only two of a large class of situations, many of the more complicated members of which are of engineering interest.

Calls attempted by an idle source can be successful, or can fail because there is no path in $g(P)$ or because B is busy. In situations for which it is important to distinguish between these sources of failure, the model must surely allow for calls to busy terminals. We arrange this by drawing Fig. 7. In a sense it subdivides time for A with respect to B , since the event {path in $g(P)$ } relates to a particular B . Fig. 7 suggests the definitions

$$TC_{(1)} = \frac{x + y}{x + y + u + v} \quad (\text{Fig. 7}),$$

$$MTC_{(1)} = \frac{y}{y + u} \quad (\text{Fig. 7}),$$

and

$$CC_{(1)} = \frac{r(A)y}{r(A)y + u} \quad (\text{Fig. 7}),$$

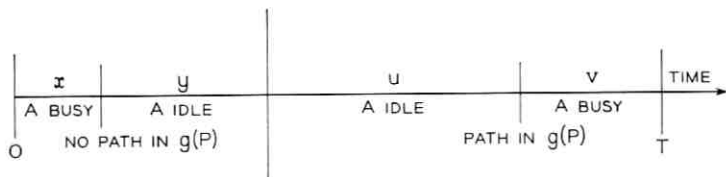


Fig. 7 — Time subdivided for A w.r.t. B .

which may be described as measures of congestion for one particular terminal with respect to another (hence the subscript "(1)"). These quantities agree as closely as possible with those (lacking subscripts) of Section II. Notice also that $TC_{(1)} \equiv TC_2$, a correspondence which does not hold for MTC or CC.

The quantities just defined measure A 's difficulties in trying to call B . We can measure A 's difficulties in making *any* calls by averaging over all B . We obtain measures of congestion for the particular terminal:

$$\begin{aligned} TC_1 &= \text{avg over } B \text{ of } TC_{(1)} ; \\ MTC_1 &= \text{avg over } B \text{ of } MTC_{(1)} ; \end{aligned}$$

and

$$CC_1 = \text{weighted avg over } B \text{ of } CC_{(1)} ,$$

where these weights are proportional to the frequencies with which A calls the various B to whom he can be connected.

One more averaging process takes us to congestion as measured for the network, for the case in which calls to busy destinations are possible. We define

$$\begin{aligned} \overline{TC}_1 &= \text{avg over } A \text{ of } TC_1 ; \\ \overline{MTC}_1 &= \text{avg over } A \text{ of } MTC_1 ; \end{aligned}$$

and

$$\overline{CC}_1 = \text{weighted avg over } A \text{ of } CC_1 ,$$

with weights according to the relative frequencies with which the different A place calls. Here, as before, the measurements before and after averaging coalesce when terminals do not differ from each other. In any case, because $TC_{(1)} \equiv TC_2$, we also know that $\overline{TC}_1 \equiv \overline{TC}_2$.

For a general network, measures of congestion have been defined for the pair and the network when attempts occur only between idle terminals, and for the terminal and the network when any idle source can attempt calls. In the latter situation, we also have measures of congestion for the particular terminal with respect to one destination. These are always useful because they lead to the particular terminal's measurements; and they have intrinsic interest in those cases, such as the control of electronic switching systems, in which it may be important to know the congestion encountered by messages from one source to each of several destinations. As for differences among the three quantities TC, MTC, and CC, the discussion in Section II on ordering relations carries over, *mutatis mutandis*, to the cases subscripted "(1)" and "2". (Notice, however, what care must be taken when dividing by uy in

repeating the calculations involving R . The generalization of the cautionary example given above is that $y = 0$ in Fig. 5 whenever the network \mathfrak{N} is non-blocking.) Similar arguments can be applied with caution to the measures obtained by averaging. It is particularly important that no simple formula for \overline{CC}_2 can be written using an " $\bar{r}(P)$ " found by taking a P -average of $r(P)$; and so on.

Here we pause to summarize the measures of congestion defined above. This is best done as in Table I, which shows how the proposed symbols and terminology are related. A more accurate description of the mode of operation associated with the number 1 in a subscript would be that attempts are made *without regard* to the state of the destination; the heading in the table is shorter and suggests the distinction that is often most important. Classification by "point of view" is omitted entirely: For as used here the phrase is merely dichotomous, the wire chief being associated with TC and the particular subscriber with MTC and CC, so that the names "time-congestion" etc. are adequate by themselves. "Measurement for (or by)" is a new classification which supplies words to go with the subscripts and bars. The word "terminal" is used in column headings instead of "subscriber" because the latter would be too reminiscent of classification by point of view.

We have not yet considered the important case in which a single source A tries to call a "destination" consisting of several terminals, such as a group of outgoing trunks. Quite different paths through the network may lead from A to various equally useful members of the "destination". The discussion of Jacobaeus's work² in Section IV covers congestion in this situation.

Before concluding that each of our apparent plethora of definitions is necessary, we must ask this question: Is there really a non-trivial difference between quantities defined for the case in which attempts occur only when P is idle, as in Fig. 5, and those for the case of Fig. 7 in which a busy terminal can be called? We answer this question in terms

TABLE I—MEASURES OF CONGESTION

Mode of Operation:		Attempts when P Idle		Attempts Possible to Busy Terminal		
As measured for:		Pair	Network	Terminal w.r.t. One Destination	Terminal	Network
Type of congestion:	Time:	TC_2	\overline{TC}_2	$TC_{(1)}$	TC_1	\overline{TC}_1
	Modified time:	MTC_2	\overline{MTC}_2	$MTC_{(1)}$	MTC_1	\overline{MTC}_1
	Call:	CC_2	\overline{CC}_2	$CC_{(1)}$	CC_1	\overline{CC}_1

of Fig. 8, which combines Figs. 5 and 7. The intervals x , y , u , and v are as in Fig. 5, but the time when P is busy is labeled according to whether A is busy or idle, and x and v are subscripted correspondingly. The most restricted definition of TC — that with subscript "2" or "(1)" — depends only on {path} or {no path} in $g(P)$, and is the same for both modes of operation. But

$$\text{MTC}_2 = \frac{y}{y + u} \quad (\text{Fig. 8})$$

and

$$\text{MTC}_{(1)} = \frac{x_i + y}{x_i + y + u + v_i} \quad (\text{Fig. 8}).$$

Cross-multiplication and cancellation as before show that

$$\text{MTC}_2 \text{ R } \text{MTC}_{(1)} \quad \text{if and only if} \quad \left(1 + \frac{v_i}{u}\right) \text{ R } \left(1 + \frac{x_i}{y}\right),$$

again assuming that $uy \neq 0$, and after adding 1 to both sides for later convenience. Although the state of B is unspecified when A is busy and P is busy, certainly B is busy when P is busy and A is idle as in the periods x_i and v_i . It follows that

$$\begin{aligned} \frac{u}{u + v_i} \quad (\text{Fig. 8}) &= \frac{\text{Pr}\{P \text{ idle} \mid \text{path}\}}{\text{Pr}\{A \text{ idle} \mid \text{path}\}} \\ &= \frac{\text{Pr}\{B \text{ idle \& } A \text{ idle} \mid \text{path}\}}{\text{Pr}\{A \text{ idle} \mid \text{path}\}} \\ &= \text{Pr}\{B \text{ idle} \mid A \text{ idle \& path}\}, \end{aligned}$$

and likewise

$$\frac{y}{x_i + y} \quad (\text{Fig. 8}) = \text{Pr}\{B \text{ idle} \mid A \text{ idle \& no path}\}.$$

These quantities are the reciprocals of those appearing in the previous

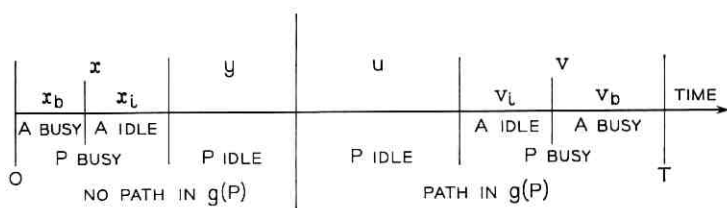


Fig. 8 — Further subdivision of time for \mathfrak{K} .

relation. Subtracting both of them from 1, we find that $MTC_2 \geq MTC_{(1)}$ if and only if

$$\Pr\{B \text{ busy} | A \text{ idle \& path}\} \geq \Pr\{B \text{ busy} | A \text{ idle \& no path}\}.$$

In some models, $\Pr\{B \text{ busy}\}$ is not affected by the existence or non-existence of a path from A when A is idle. But in most systems in which the terminal B contributes a positive fraction of the traffic carried in $g(P)$, $\Pr\{B \text{ busy}\}$ is so affected, and MTC_2 must differ from $MTC_{(1)}$.

A similar argument applies to call-congestion. Because

$$CC_2 = \frac{r(P)y}{r(P)y + u} \quad (\text{Fig. 8})$$

and

$$CC_{(1)} = \frac{r(A)(x_i + y)}{r(A)(x_i + y) + u + v_i} \quad (\text{Fig. 8}),$$

we can show that $CC_2 \geq CC_{(1)}$ when

$$\frac{1 + (v_i/u)}{1 + (x_i/y)} \geq \frac{r(A)}{r(P)},$$

assuming $r(P)uy \neq 0$. Equality in this relation requires either an unlikely coincidence or agreement of both the MTCs and the calling-rate ratios. Only these last need further comment. Whenever it is reasonable to think of the calling rate of P as being composed of calls attempted by A and by B , the ratios $r(A)$ and $r(P)$ are equal if the calling rates of A and B (to each other) respond in the same way to a "no path" condition in $g(P)$. This they may or may not do, but anyway this question is independent of whether calls attempted by P are made up equally of calls from A and from B .

Having observed a genuine difference between the "pair-attempt" and "source-attempt" models, we return to Fig. 8 to tie off one more loose end. The definition of $r(P)$ seems natural; but suppose that, when calls to busy terminals are possible, changes in A 's calling rate occur in response to the failure of attempts, with no distinction between busy-terminal and blocked-path failures. It is even possible to imagine average relative calling-rates $r_p(A)$ during {no path} and $r_b(A)$ during {path & B busy}, although relating these ratios to A 's rules of operation might be extremely difficult. Such a model would generalize the definition of call-congestion to

$$CC_{(1)} = \frac{r_p(A)(x_i + y)}{r_p(A)(x_i + y) + u + r_b(A)v_i} \quad (\text{Fig. 8}),$$

which still measures the frequency of blocking due only to {no path}. Here I drop this line of thought, which can be extended if necessary in a particular application. (The possibility of defining CC so as to illuminate particular matters of interest is discussed by S. P. Lloyd in an unpublished Bell Laboratories memorandum. He treats especially the problem of correcting the apparently high CC caused by A 's "extreme impatience" in retrying very rapidly after an attempt fails.) The main point is that the theorist may want to describe A 's behavior as conditional upon the state of the system, whereas A actually responds to the results of his call-attempts or to other *sampled* data on the system's states.

IV. QUANTITIES CALCULATED IN THE LITERATURE

The purpose of this section is to compare the foregoing ideas with discussions in several papers on connecting networks. We begin with the monumental 1950 paper by Jacobaeus.² Syski, in his book,¹ achieves wonders of condensation in his excellent summary of this paper, mostly in Section 1 of Chapter 8, "Link Systems". Although Syski's book is available to most readers, I think it will be best to quote certain passages in full from Jacobaeus. The essential remarks are in Chapter 4, "Principles of Congestion Calculation in a Link System", and Chapter 9, "Formulae for Call Congestion in a Link System", of Ref. 2. They correspond to Sections 1.1.2, 1.1.3, 1.2.1, and 1.7.2 of Syski's Chapter 8.

The portion of a system analyzed by Jacobaeus consists of two switching stages and the links joining them. Fig. 9 illustrates his nomenclature, which I use in order to facilitate comparison with his paper; American terms appear in parentheses. An "A-device" is an inlet to a primary switch (strictly, a switch in the first stage considered), and an "A-group" is the set of inlets to one switch. A "B-device" is a primary-secondary link, and a "B-column" is the set of links emanating from one primary switch. Thus a B-device connects a primary outlet to a secondary inlet. A "C-device" is an outlet from a secondary switch, and a "route" is generally a "C-column", which is the set of corresponding outlets of the secondary switches. For example the second C-column is the set of second outlets of secondary switches. A route may consist of a part of a C-column, or of more than one.

Jacobaeus calculates, under various conditions, the probability that an A-device has no access to a particular route. He does this by writing first the probability that p of the C-devices of the route are busy; then multiplying by the probability that the $m - p$ B-devices (links) leading to the idle C-devices are also busy; and finally summing over p . This

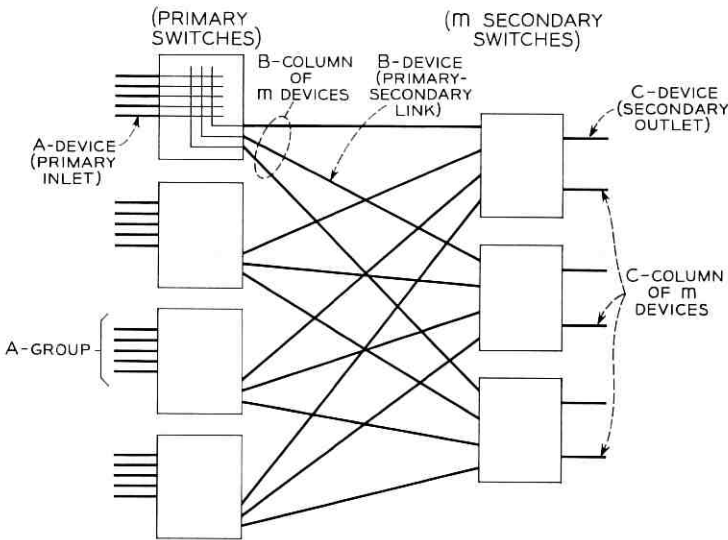


Fig. 9 — Nomenclature of Jacobaeus.

procedure yields his basic congestion-equation [Jacobaeus, p. 11, equation (1); Syski, p. 437, equation (2.1)]. The first ingredient of the product just described is a probability distribution for the number of busy C-devices in a route. The second ingredient is a distribution for the number of busy links in a specified part of a B-column. Three topics account for much of the length of Jacobaeus's paper. One is the correct choice of these two distributions in each situation studied. (It is easier to understand Jacobaeus's discussion of this topic with the aid of Syski's remarks [p. 439] on the meanings of the various traffic parameters used.) The second is the inaccuracy caused by assuming independence of the instantaneous link- and route-loads, as represented by multiplying the two blocking-probabilities. The third is the inaccuracy caused by approximating some of the difficult summations arising from the basic equation.

It is clear that this basic equation, as used in Chapters 5-8, 10, and 11 of Jacobaeus, yields $TC_{(1)}$, where this "congestion" is contributed by the portion of a system included in this model. We see this for the following reasons: First, the basic equation ignores the state of the calling A-device. Second, the goal of a call-attempt is not one destination but a route, consisting of several equally useful C-devices, and usually considered to be loaded with traffic from many sources according to Erlang's B-formula (first loss-formula). Third, calls are placed without

prior regard to the state of the desired route. Fourth, the equation holds for a single A-group calling a single route. Thus $TC_{(1)}$ is measured for the particular terminals which share one primary switch, with respect to the class of destinations attainable via one route (which may be a trunk group). This measurement is as particular as possible, given that almost any switching system has *some* symmetries among its terminals.

Application of the basic equation in such a way as to account for finite-source effects appears to yield $CC_{(1)}$. However, since the case in which $r(A) \neq 1$ is never mentioned, this quantity cannot here be distinguished from $MTC_{(1)}$.

These views are partially confirmed by Jacobaeus on p. 23 of his Chapter 9:

"The congestion formulae derived in the preceding chapters have all referred to the usual congestion concept, time congestion. The quantity used to express the congestion is the fraction of time during which the loading in the switching system is such that a new call cannot be switched. It is also possible to consider the call congestion, the fraction of the total number of calls which are blocked owing to a shortage of means of connection. In a full availability group loaded with random traffic the call congestion is equal to the time congestion, because the probability of a new call is always the same independent of the conditions ruling in the group. The condition for this is that there should be an infinite number of traffic sources, which must be the case if the traffic is to be truly random.

"In the majority of the connection systems treated above, the number of traffic sources in each A-group has been limited. . . . It may therefore be expected that the call congestion will have a smaller value than the time congestion."

There is further discussion [pp. 24, 38] of the finite size of an A-group, mostly with respect to the question of independence of sources. Jacobaeus points out that A-devices are truly independent only when they are "direct traffic-producers: subscribers' lines. . . . Otherwise the A-devices are secondary traffic sources, that is they derive their traffic from a large number of traffic sources by way of a concentrating device." We see that the "particular terminal" for which measurements are made may in fact be a "source" within the shared communication system, depending on what portion of the system is selected for study by means of this two-stage model. (Actually Jacobaeus discusses extensions of his method to more stages.)

We note in passing that some of the distributions substituted into the basic equation apply only when blocked calls are held (i.e., remain in the system for a length of time which does not depend on when or whether the desired path becomes available). Jacobaeus discusses this point [pp. 35, 38]. Relevant formulae appear in his Chapter 7; see also equations (2.15-17) of Syski [p. 442]. This situation is represented by the possibility, because of the presence of concentrating primary

switches, that an A-device can be busy even when all B-devices for the same A-group are occupied by other sources. This cannot happen in the model of C. Y. Lee, to which we turn next, since he writes [Ref. 3, p. 1300, footnote] that "we restrict our attention to networks consisting of switches of non-blocking type".

The ideas of Jacobaeus form the closest possible analogue in the general case to the simple approach of Section II. In Lee's paper we expect to find measurements for the pair rather than for the particular terminal, since he writes in his Introduction [Ref. 3, p. 1288], "... methods for calculating ... in a gas tube network are ... given ..." (see above, p. 2280). Our expectation is confirmed; for, although the links of Lee's subgraphs $g(P)$ are directed, his methods of calculation are symmetric and do not distinguish between sources and destinations. (See especially his Section 4.3, in particular the subsection on end-matching [p. 1311] in which a "voltage is applied to both ends of the network simultaneously (across a single input and a single output)".) I believe that Lee's use of directed links serves merely to specify, out of the geometrically possible ones, the allowed paths for a call in $g(P)$. But Lee does not discuss the pair-vs-terminal issue. In applying his model to line-switched traffic in a trunking network, one would naturally calculate congestion for one terminal, possibly with respect to another. This point is not important, for, as we see below, it cannot really be decided.

The essential attribute of Lee's model is that it ignores the states of terminals. He shows a four-stage switching network together with a subgraph $g(P)$ [Ref. 3, Figs. 3.2-3]; each possible path in the latter consists of just three branches. Such a model corresponds to our Fig. 4 rather than Fig. 6. The quantity found directly by Lee is " $P(i,j)$ the probability of all paths from input i to output j busy" [p. 1300]. This quantity corresponds to our TC_2 . The question of pair or terminal discussed above is moot because, as we recall from p. 2282 above, TC_2 agrees with $TC_{(1)}$; their measurements diverge only for MTC and CC.

Lee introduces the important idea of blocking-probability for a network as an average of such probabilities over all terminal-pairs [p. 1300, equation (3.1)]. This, the quantity of major interest to Lee in parts of his paper, is of course our \overline{TC}_2 . Furthermore, Lee sometimes assumes complete interchangeability of terminals [Section 3, p. 1300]: "In this section, we assume $P(i,j)$ to be independent of i and j ..." In this case, $TC_2 = \overline{TC}_2$ for every terminal-pair.

Lee's methods apply directly only to the (approximate) calculation of TC. His results are therefore consonant with those of Jacobaeus on TC, although these results are approached by the one on behalf of the pair

and by the other on behalf of the source. It also seems clear from Lee's Section 4, in which he applies his methods to problems of retrials and connection time and discusses the results, that his formulae are meant to be taken as yielding CC_2 . I think this intention constitutes an implicit assumption of what Jacobaeus (as quoted above) called "truly random traffic"; that is, that $r(P) = 1$ and that there are "infinitely many" sources.

The paper by Grantges and Sinowitz⁴ is important because it extends the methods of the earlier literature and launches a practical attack on the problems of independence, network size, and computational complexity from which all researches in this field have suffered. They say [Ref. 4, p. 969] that "the nodes of the graph represent network switches and the directed branches of the graph represent network links". Their Figs. 1 and 2(a) [p. 970] show that, as in my Fig. 4, no branch of a subgraph is associated with any single terminal. Their "program is based on a simplified mathematical model of switching networks developed by C. Y. Lee" [Ref. 4, p. 967]. They assume [p. 969] stationary traffic, non-blocking switches, and complete equivalence among terminals. As discussed above, this last assumption makes network averages agree with measurements for one pair or terminal. I write here of the quantity that is calculated *directly*, and therefore ignore the symmetry that also yields measures of congestion for the whole network.

The previous paragraph shows that the mental picture of Grantges and Sinowitz is compatible with that embodied in this paper. Let us now examine the first paragraph beginning on p. 976 of Ref. 4. The predecessor of that paragraph discusses congestion in the simple context of a full-access trunk-group. The paragraph in question describes the attitude taken by the authors toward the problem of more general networks. They say that ". . . the measure of most concern to the network designer is call congestion. Now if it is assumed that calls originate completely independent of the state of the network, time congestion will equal call congestion. Such an assumption is unjustified if calls cannot originate from busy lines, since time congestion conventionally includes busy line periods while call congestion excludes them." The last clause means "excludes from both numerator and denominator of the ratio", as is clear from all our formulae for TC and CC. The second sentence just quoted must mean that TC and CC would have the same formula; and indeed our expression naturally defining CC would be the same as that for the corresponding TC if call-attempts occurred (as they do in Erlang's infinite-source model) without regard to the state of the network. This possibility is ruled out in the third sentence, which

simply observes that the formulae for the two quantities differ. Of course, as our Section III shows, it is *because* calls often cannot originate from busy lines that we conventionally exclude busy-line periods in defining CC. Grantges and Sinowitz continue, "It is, however, reasonable to assume that idle pairs of terminals originate calls at a constant rate independent of the state of the network." (This sentence restricts us to measurements for the particular pair: that is, to TC_2 , MTC_2 , and CC_2 , with the proviso that the network averages are included by symmetry.) "In particular, there must be no change in the calling rate after a blocked call." (In other words $r(P) = 1$.) "If, under this assumption, time congestion is modified to include only periods in which both lines are idle, it will be equal to call congestion." (Here they recognize that $MTC_2 \equiv CC_2$ when $r(P) = 1$.) "Actually, even if the foregoing assumption is not met, the time between calls is likely to be much longer than the time taken by the network to return to equilibrium, so that, again, the modified time congestion will be close to the call congestion." (That is, MTC_2 is near CC_2 when $r(P)$ is near unity.)

In the second paragraph beginning on p. 976, Grantges and Sinowitz go on to say, "With a suitable choice of branch occupancies, Lee's model allows the computation of call congestion. Alternatively, the branch occupancies may be chosen so as not to reflect the requirement that only idle terminals are to be considered, thus allowing the computation of time congestion." Here "suitable" means that branch occupancies are [p. 977, footnote] "chosen to reflect the requirement that the input-output terminals j,k are idle by (usually) subtracting the load contributed by the terminals j,k from the assumed carried link loads". This confirms the claim that the pair's measurement is calculated, since otherwise one would subtract the load contributed by one terminal alone.

It appears from all this that the NEASIM program of Grantges and Sinowitz was designed to find (approximately) either TC_2 or, after correction of branch occupancies, MTC_2 , the latter quantity being supposed very close to CC_2 on the ground that $r(P)$ is very near 1. This situation agrees with that of Lee's paper, though the latter does not mention the correction procedure needed to find CC_2 . Neither paper mentions the possibility of call-attempts to busy terminals, but the correction procedure is easily modified, as mentioned above, to cover that case. Since the output of the NEASIM program [Ref. 4, p. 975] is a functional relationship between sets of branch occupancies and probabilities that no path exists through $g(P)$, no decision is actually built into the program as to which measure of congestion is calculated. The

user makes this choice separately by relating terminal loads to branch occupancies, a process which may have to be iterative and to rely on routing assumptions in order to ensure consistency.

V. DISCUSSION

Before summarizing the implications of this approach we must reconsider the question of applicability mentioned in Section I. It is not really profitable to try to characterize exactly (beyond the need for stationarity) models in which the mathematical limits required by the present method, with its use of conditional calling-rates and finite T , exist and agree in pairs. Instead we should look upon the arguments leading to Table I as exemplifying a point of view which is, if flexibly applied, relevant to a wide range of traffic systems.

Section IV shows that the thirteen independent entries of Table I — (recall that TC_2 and \overline{TC}_2 agree with $TC_{(1)}$ and \overline{TC}_1 respectively) — suffice, as they stand, for interpreting a variety of investigations in the literature. What can be said about this approach in cases of doubt, or when it is clear that a system requires new and more appropriate definitions? We need only to remember that such sketches as appear in Figs. 2, 5, 7, and 8 are intended merely to represent in a manageable way certain *conceptually* simple processes of measurement. For TC or for MTC such a process requires two clocks, one or both of them controlled by switches whose states reflect the defining states of the network under study. The measurement of CC requires instead two counters, which count respectively blocked attempts and all attempts of the desired categories.

For illustration consider a source A which, after every blocked attempt, makes retrials at a steadily increasing rate until the desired connection is set up. Application of one of our formulae for CC would require knowledge of $r(A)$, a quantity more difficult to evaluate in this case than call-congestion itself. But it is simple in principle just to count attempts and blocked attempts, which define CC directly. What is not so clear is whether this definition is useful. If periods during which a call-attempt of A would be blocked tend to be rare but long-lived, a very few unsuccessful first-attempts can lead to a high value of CC. When the cost incurred through failure to deliver a message increases rapidly with time, as does the retrial rate, this measure of congestion may be appropriate. Under other circumstances it may be preferable to count only first attempts and blocked first-attempts, or to count blocked retrials with a weighting factor that decreases with retrial

number. The conceptual framework proposed here should not obscure the fact that measures of call-congestion can always be understood by reference to the idea of counting call-attempts; and similarly for time-congestion and the measurement of time intervals. The hard part in practice is to choose a definition whose behavior accurately reflects the usefulness of a system's performance.

Furthermore, conceptual simplicity does not guarantee utility. Direct measurement of MTC_2 requires a clock that knows when both A and B are idle *and* when there is no idle path in $g(P)$. Especially in a large network, this is likely to be impossible or impractical. One must often *estimate* congestion indirectly and in finite time, rather than measure it. This brings up such matters as sampling bias and efficiency, cost of instrumentation, and so on. Thus in choosing a definition of congestion one must consider not only its *purpose*, as discussed above, but also the cost of, and attainable accuracy in, *measuring* (or estimating) the quantity defined.

These interesting problems are not covered in this paper, which treats only the *concept* of congestion. I think we can attribute much of the confusion which has characterized this subject to the natural tendency of authors to treat highly symmetric models. Neither the row nor the column structure of Table I is significant when traffic is truly random and all terminals are alike. But the principal conclusion of this paper is that the distinctions embodied in Table I should be kept in mind. In other words it is always useful, if only as a precaution, to ask in what mode a system operates, for what entity congestion is to be measured, and what type of congestion is of interest.

A subsidiary conclusion is that there is no sharp distinction between trunking networks and switching networks, but rather a range of salient properties characteristic of various kinds of communication networks. Instead the basic division is between general networks, as treated in Section III, and single links of networks (links composed of one or more channels), for which the simpler discussion of Section II is adequate.

VI. ACKNOWLEDGMENTS

Although responsibility for the peculiarities of this discussion is mine alone, I am indebted to many of my colleagues at Bell Laboratories for their help in clarifying my ideas about congestion theory. Among them I am especially grateful to V. E. Beneš, A. Descloux, and H. O. Pollak for stimulating discussions, and to R. F. Grantges, W. S. Hayward, and John Riordan for helpful comments on earlier versions of this paper.

REFERENCES

1. Syski, R., *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, Edinburgh and London, 1960.
2. Jacobaeus, C., A Study on Congestion in Link Systems, Ericsson Tech., No. 48, 1950, pp. 1-68.
3. Lee, C. Y., Analysis of Switching Networks, B.S.T.J., 34, November, 1955, pp. 1287-1315.
4. Grantges, R. F. and Sinowitz, N. R., NEASIM: A General-Purpose Computer Simulation Program for Load-Loss Analysis of Multistage Central Office Switching Networks, B.S.T.J., 43, May, 1964, pp. 965-1004.

100A Protection Switching System

By H. D. GRIFFITHS and J. NEDELKA

(Manuscript received July 12, 1965)

The 100A Protection Switching System is intended to ensure continuity of service in TD Microwave Radio Relay Systems by transferring service from a failed working channel to a standby protection channel. Switching is done at IF on a switching section basis. A switching section may contain up to ten radio repeater stations, with the average section containing three to four repeaters. A fully equipped 100A system provides means to replace any one of ten regular channels with either of two protection channels. Two regular channels may be switched simultaneously to the two protection channels. The circuit interruption due to a fade initiated or a maintenance switch is less than 10 microseconds. The interruption due to an equipment failure is less than 35 milliseconds, a time short enough to prevent false operations in the telephone switching plant. The control information necessary to sequence the operation of the switches at the transmitting and receiving ends of the switching section consists of coded voice-frequency tones. These may be transmitted over any reliable voice-frequency facility having suitable transmission and delay characteristics. All of the active devices in the system are solid state. The system requires sources of both +24 and -24 volt power.

I. INTRODUCTION

The high reliability demanded of microwave radio relay systems requires that protection be provided against system outages that occur as a result of fading or equipment failures. This protection is obtained by frequency diversity through the use of standby protection channels and an automatic protection switching system. Each radio relay route is divided into a number of switching sections, each of which may contain up to ten radio repeaters. One or two of the total number of radio channels in the switching section are designated as protection channels. The automatic protection switching system uses the protection channels to replace failed regular channels by operat-

ing IF switches at the transmitting and receiving ends of the switching section. By manual operation of the IF switches, the protection channels may also serve as alternate facilities during an emergency or while maintenance is being performed on a regular channel.

The original protection switching system developed for use with the long haul TD-2 system has a capability of protecting any one of the five regular working channels with one protection channel.¹ It was first placed in service in 1953 and since that time has given consistently good performance. About 1960 the TD-2 system was expanded to 12 channels through the use of channels placed interstitially with the original 6, and many routes are now equipped with both the regular and the interstitial channels. A second one-for-five switching system is, therefore, required to provide protection for these interstitial channels. However, a much better over-all protection switching arrangement results if the regular and the interstitial channels are combined into a two protection and ten regular working channel system, i.e., a two-for-ten system. By selecting the protection channels with maximum frequency separation, the chance of having at least one protection channel available during periods of heavy fading is greatly increased. Also, while one of the protection channels is being used to carry service during the maintenance of a regular channel, the second protection channel is available to guard against a failure of one of the remaining regular channels or the first protection channel. The 100A Protection Switching System was developed to fill this need of a two-for-ten system. The 100A system will operate with both the existing TD-2 system and the TD-3 system.

II. SYSTEM CONSIDERATIONS

Since the 100A system is intended for use with existing installations of TD-2 as well as new routes, its design must be compatible with existing system arrangements. Therefore, like the earlier one-for-five system, switching is done at IF on a switching section basis, and all of the monitoring, switching, and control equipment is confined to the transmitting and receiving ends of the switching section. However, since on some new routes, the radio system initially may have only one working channel, the 100A system must also accommodate radio systems with as few as one protection and one regular channel.

The IF switch at the transmitting end of the switching section bridges the regular channel to the protection channel. The IF switch at the receiving end transfers the channel output from the regular channel

to the protection channel. The two IF switches at the ends of the section must therefore operate in proper time sequence to minimize service interruptions. It also follows that if a regular channel fails suddenly, time will be required to recognize that the failure has taken place and time will be required to pass the control information between the transmitting and receiving ends to operate the IF switches in their proper sequence. The total interruption time for an equipment failure has a lower limit determined by the length of the switching section and the type of facility used to pass the control information. In any case, the interruption will be long enough to cause a hit in data transmission. For the 100A system, a limit of 35 milliseconds has been set. Typically, most switching sections have equipment failure interruption times of less than 30 milliseconds.

Circuit interruptions due to equipment type failures occur less frequently than those caused by fading or maintenance switching. The maximum fading rate experienced on working systems seldom exceeds 100 db per second. In about 35 milliseconds (time to complete a switch) the channel has faded at this maximum rate an additional 3.5 db. Thus, in the case of a switch which is fade initiated, the receiving end transfer is made between a slightly degraded regular channel and a good protection channel. In the case of a maintenance switch, the transfer will be made between two good channels. In both cases the interruption time is dependent only on the transfer time of the receiving end switch. Through the use of diode type switches this interruption time is held to less than 10 microseconds, a time sufficiently short to prevent a hit for moderate speed data transmission, if the transmission times of the radio channels are equalized.²

Channel quality at any time is determined by the signal-to-noise ratio of the recovered baseband signal. This can be monitored in a number of ways. In the TH Radio System, the received carrier power at each radio repeater is monitored through the automatic gain control circuits. When the carrier power falls to a predetermined value resulting in a definite signal-to-noise ratio, a relay removes a tone on an auxiliary radio channel which also passes through the repeater station. The absence of this tone at the receiving end of the auxiliary channel is the signal for the switching system to initiate a switch. In the one-for-five system for TD-2, channel quality is determined by monitoring the baseband noise in a slot at about 9 mc, a frequency well above the top message frequency. A sufficient increase in the noise results in a switch request. This method allows all of the

monitoring equipment to be concentrated at the receiving end of the switching section and removes the need to connect the switch controls into the auxiliary radio channel at each repeater. However, the baseband response over the switching section must be good to 9 mc to ensure a reasonably constant relationship between noise at 9 mc and noise at the lower baseband frequencies. Carrier is also monitored at the receiving end of the switching section in TD-2. This is done to protect against an equipment failure which might occur in the last radio repeater at a point where a noise initiated switch would not be called for. The TD-2 approach was adopted for the 100A system for the following reasons: (i) it has worked well with the earlier system, (ii) no auxiliary channel is available, and (iii) considerable modification would be required to each TD-2 radio repeater to provide the monitoring facilities.

Reliability is a major consideration in the design of a switching system. Circuits in the 100A system are very conservatively designed and use only solid-state devices. Alarms and indications are provided to call attention to a system malfunction or a prolonged switching operation. However, since many of the circuits in the 100A system are trigger circuits, i.e., one transistor of a pair is on while the second is carrying no current, an actual operating test is the only effective way of checking for a quiet failure. To check for these quiet failures, the 100A system includes an automatic test circuit (referred to as an exerciser) which once daily, or on request, makes and restores switches from all regular channels to each of the protection channels. If a switch operation is not completed within a predetermined time, the channel at fault is identified and an alarm is initiated.

The continuity of transmission through the switching section, however, depends not only on the switching system to carrying out normal switching activities, but also making the right decision in the face of abnormal conditions. One such abnormal condition would be a failure of the control line connecting the receiving and transmitting end equipment. The failure could include an open circuit condition, high noise, or interfering tones. Unless precautions are taken, all of these conditions could result in false transmitting end switch operations since the logic and master control point is at the receiving end and the transmitting end is a slave unit. In the 100A system each switch order consists of a two-tone code combination. These tone combinations are decoded by the transmitting logic. If the tone combination is a valid one and remains uninterrupted for a predeter-

mined interval, then and only then, will a switch operation proceed. To guard against a high random noise level from simulating a valid order, noise level sensing circuits are used on the tone control line. These circuits prevent a switch action when the line is noisy.

Safeguards must also be built into the receiving end equipment to guard against a control line failure. Each switch order that is originated must be timed. If the logic at the receiving end of the system does not obtain verification that a switch operation is completed within the timing interval, then a new switch order to the second protection channel must be originated. If the second protection channel is not available, an immediate alarm must be given to the operating personnel. Similarly, if an order to take down a switch is not carried out in the required timing interval, the receiving end switch must be held operated and maintenance personnel must be warned. Therefore, the receiving end equipment must time and check the sequence of all its signal inputs. Any abnormal sequence of events, whether due to the control line or other system failures, must be processed, identified and translated to a warning to the maintenance personnel.

III. OVER-ALL SYSTEM OPERATION

A simplified block diagram showing the important interconnections of a typical one-way switching section using the 100A system is shown in Fig. 1. Transmission is from main station P to main station Q, over the ten regular channels designated A through J, and the protection channels designated X and Y. At the auxiliary station S, one or more of the regular and both protection channels are branched to provide a side leg TV drop facility. The switching equipment is confined to the two main stations P and Q and the auxiliary station S. All of the switching in the section is done at the 70-mc IF by means of fast acting diode switches which are shown symbolically as single-pole, double-throw switches. The automatic switching is initiated and controlled by the receiving end equipment at station Q. Switching at the transmitting and auxiliary stations is coordinated with the receiving end by means of tones transmitted over the voice-frequency facility which may be independent of the radio route. One voice-frequency line is required for each protection channel. A similar arrangement is required to provide protection switching for the opposite direction of transmission in the same switching section.

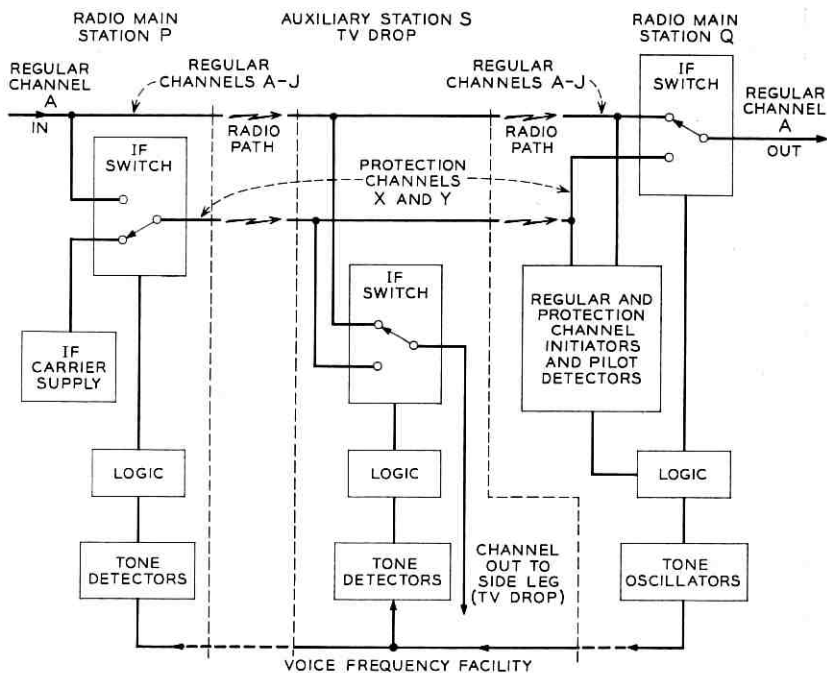


Fig. 1—Switching section—simplified IF block diagram.

All the radio channels are monitored at IF at the receiving end of the switching section by initiators bridged to each of the channels. These initiators continuously measure the IF carrier level and the channel noise. If either should become unsatisfactory, a switch request is made if a regular channel is involved or a switch inhibiting order is originated if a protection channel is involved. If an initiator for a regular channel makes a switch request, the receiving end logic, in conjunction with the protection channel initiators and associated circuits, determines if a protection channel is available for service. If a protection channel is available, an order is sent by means of a coded group of order tones over the voice-frequency control line to the transmitting end. The transmitting end equipment decodes the order and bridges the failed regular channel to the selected protection channel. The auxiliary switching station also receives the order but no immediate action is taken.

The transmitting end bridge is made when the IF switch operates

to provide transmission on both a regular and a protection channel simultaneously. Operation of the switch also removes the output of the IF carrier supply from the protection channel. The IF carrier supply provides the idle protection channel with IF carrier and a 7-mc pilot tone. The absence of the pilot tone, which is detected by the pilot detector associated with the protection channel initiator at the receiving end of the system, notifies the receiving end logic that the bridge at the transmitting end has been made. A receive end transfer is then made from the regular to the protection channel. After the transfer is made, a guard tone, which is normally present on the voice-frequency line, is removed. The absence of the guard tone tells the auxiliary switching station to complete the switch from the defective regular channel to the protection channel and provides an additional safeguard against accidental removal of the transmitting end bridge.

When the regular channel again becomes good, the receiving and transmitting end switches are restored. First, the receiving end switch is restored to its normal state. Then simultaneously, the switch order tones are removed and the guard tone is restored. This results in the removal of the bridge at the transmitting end, and the restoral of the transfer at the auxiliary switching station.

IV. SIMPLIFIED SWITCHING SYSTEM

A one-for-one protection switching system which does not require a voice-frequency control line is also provided as part of the 100A system. In this arrangement, a permanent bridge is made between the two channels at the transmitting end of the switching section. Switching between the regular and protection channels at the receiving end of the section is under the control of the two monitoring initiators and the receiving end logic. Switching and restoring is made without the delay caused by the transmission of the switching tones of the regular system. The simplified system may be expanded into a full ten regular two protection system by the addition of the voice-frequency facility and appropriate plug-in units.

V. RECEIVING END

Fig. 2 is a block diagram of the receiving end of the switching system. The regular and protection channels are taken from the IF interconnecting circuits through IF amplifiers to the IF receiving

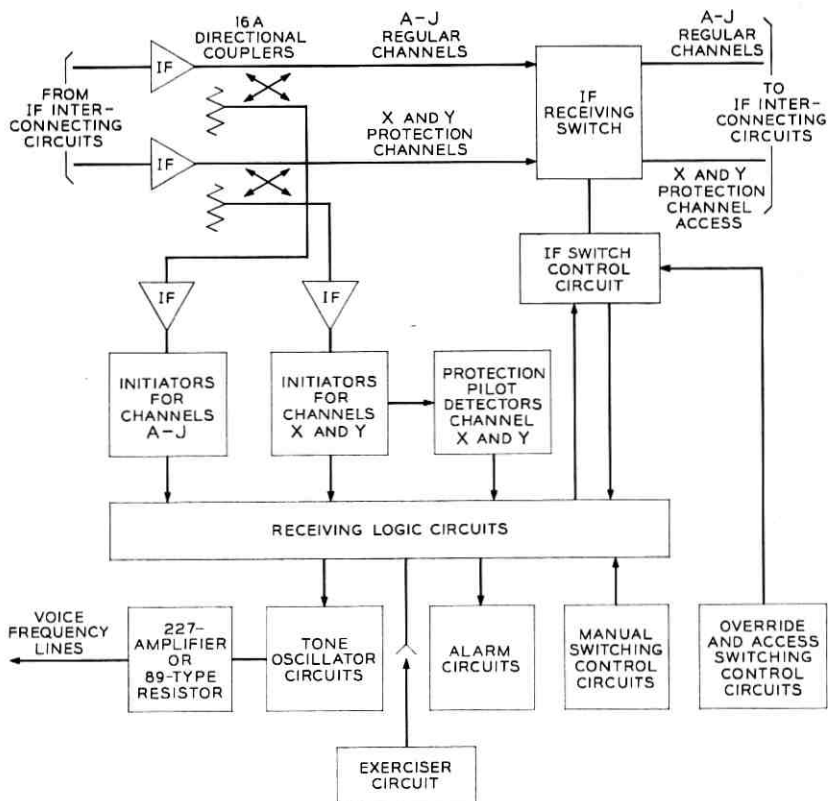


Fig. 2—Receiving end block diagram.

switch. This switch acts as a transfer type switch allowing either the regular or the protection channel to be connected to the switch output. Also, provision is made in the switch to connect the output of the protection channel directly to external circuits, thus allowing access to the protection channel should it be required for emergency service.

The regular and protection channels are also connected to the initiators through directional couplers and level restoring IF amplifiers. The initiators monitor the channels for noise and carrier. The regular and protection channel initiators are identical, but are used at different sensitivities. The regular channel initiators are set to indicate that a channel is bad on noise corresponding to a 35-db fade in a nominal TD-2 radio repeater section. The protection channel

initiators are set to indicate that a channel is bad on noise corresponding to a 33-db fade. Both regular and protection channel initiators indicate a bad channel on an 18-db drop in carrier. The difference in initiator channel noise settings helps to prevent unnecessary switching on previous section failures. When a previous section failure occurs, the regular channel initiators in succeeding sections may also indicate that the channel is bad and request protection. After the head end bridge is made in a succeeding section, the apparent simultaneous failure of both the regular and protection initiators is interpreted by the receiving logic as a previous section failure and no receiving end switch is ordered. It is important, therefore, that the protection channel initiator operate rapidly when a head end bridge is made to a noisy channel; this is obtained by setting the protection channel initiator to operate at a smaller depth of fade than a regular channel initiator.

Associated with each protection channel initiator is a pilot detector. This circuit monitors the 7-mc pilot tone normally present on the idle protection channel. The output signals from all the initiators and the pilot detectors are fed directly to the receiving logic circuit. These signals are used by the receiving logic to determine the operations of the switching system. The receiving logic controls the receiving IF switch through the IF switch control circuits. The switch control circuits in turn notify the receiving logic whether the receiving switch is operated or nonoperated. The receiving logic also controls the operation of the transmitting and auxiliary station switches by turning on the appropriate tone oscillators to produce the required order tone combinations.

The system may be controlled by the operating personnel directly through the manual switch control circuits or the override switch control circuits. The manual switch control circuits are located at the receiving end of the system only and control system operation by giving simulated channel good or bad indications to the receiving end logic. The override switch control circuits are located at both the transmitting and receiving ends of the system. They operate the IF switches directly by dc control voltages and override any of the normal logic and control. Manual switching is used to carry out all the normal maintenance switching. Override switching is used only when there is a voice-frequency control line malfunction or when the logic is being maintained.

The manual switch controls are on a per channel basis. To make a manual switch, the regular channel is made to look bad to the receiving logic independent of its actual condition. In addition to mak-

ing and restoring switches, the manual switch control circuits may be used to inhibit switching to a channel by "locking" it out. "Lock-out" of a regular channel is accomplished by making it look permanently good to the receiving logic. Lock-out of a protection channel is accomplished by making it look permanently bad to the receiving logic. When a channel is locked out, no switching alarms are originated for that channel and, effectively, it is not a part of the switching system.

Like the manual switch controls, the override switch controls are on a per channel basis with the exception of a common control, the *status quo*. No override switch operation can be made until this control is operated. Operation of the status quo control at the receiving end of the system maintains the control voltages being applied to the IF switch control circuits by the receiving logic, disconnects the IF switch control circuits from the logic, and removes all tones from the voice-frequency line. Thus while *status quo* is operated, no automatic or manual switches can be made. Its operation, however, does not disturb any switches in effect. Subsequent operation of the override switch controls can make or restore switches from any regular channel to either protection channel, regardless of initial conditions. When it is desired to restore the system to automatic operation, the procedure depends on whether any override switches are in effect. If no override switches are in effect, it is only necessary to restore the *status quo* control. If override switches are in effect, then care must be taken to ensure that the same switches are being ordered by the receiving end logic. This can be obtained by use of the manual switch controls.

Override switching at the transmitting and receiving ends of the switching section are similar but completely independent. Therefore, to avoid circuit interruptions, override switch operations at the transmitting and receiving ends of the switching section must be properly coordinated by the operators.

The access switch control circuit provides a means of operating the IF switch so that a protection channel may be connected directly to the station IF interconnecting circuits for special uses. A protection channel must first be locked out through the manual switch controls before the access switch control circuits will function. Since a protection channel carrying service cannot be locked out, this interlocking circuitry prevents a service interruption caused by an inadvertent operation of the access controls on a busy protection channel. Operation of the access control also connects the initiator for the

protection channel to the service failure alarm circuit. Thus, an immediate alarm is given should the protection channel fail while connected through the "access" terminals. The access controls at the receiving and transmitting ends are independent. Therefore, to complete the IF circuit through the switching section, access controls at the transmitting end of the system must also be operated.

The receiving logic initiates a number of alarms to warn maintenance personnel of abnormal situations. The most important of these is the service fail alarm. It is an immediate alarm and is given whenever:

- (i) a regular channel requests a switch and a switch is not completed
- (ii) a protection channel fails when it is protecting a regular channel and a second protection channel is not available
- (iii) a protection channel fails while it is being used for special message or television service, i.e., protection channel access switches have been operated
- (iv) the pilot returns to a protection channel when it is protecting a regular channel and a second protection channel is not available. The return of the pilot to the protection channel is an indication that the bridge between the regular and protection channel has been taken down, and the protection channel is no longer carrying service.

Other alarms are:

- (i) a prolonged switch request alarm which is initiated if a regular channel has failed for longer than approximately 45 seconds
- (ii) a prolonged protection channel failure alarm which is initiated if an idle protection channel fails due to excessive noise or loss of carrier, or the pilot disappears from the channel. The alarm is given after the trouble condition has been in effect for approximately 45 seconds
- (iii) a switch release fail alarm which is initiated immediately if a switch cannot be restored.

The exerciser circuit performs a test routine once daily or on request. This test routine consists of checking the protection channel initiators and the normal automatic switching operations of the system, i.e., switching and restoring each regular channel to each protection channel. Initiation of the test switching operation is made by simulating noise and carrier failures in the regular and protection channel initiators. If a test sequence is not completed satisfactorily, an alarm is given and the exerciser disengages. Thus, the exerciser is able to minimize the possibility of a recurring system malfunction

due to an undetected equipment failure which otherwise might exist for a long time.

VI. TRANSMITTING END STATION

Fig. 3 is a block diagram of the transmitting end IF switch and control circuits. Each regular channel connects through the low-loss arm of a 16A directional coupler. The high-loss arm of this coupler is connected to the transmitting IF switch. Operation of the IF switch will bridge a regular channel to one of the two protection channels. When a bridge is made, the protection carrier supply is disconnected from the protection channel. Access to the protection channels for special service is also gained through the IF switch. IF amplifiers are used in each of the protection channels to make up for the 20-db coupling loss of the 16A directional coupler and the loss of the IF switch.

Operation of the transmitting end IF switch is controlled by voice-frequency tone combinations generated at the receiving end and

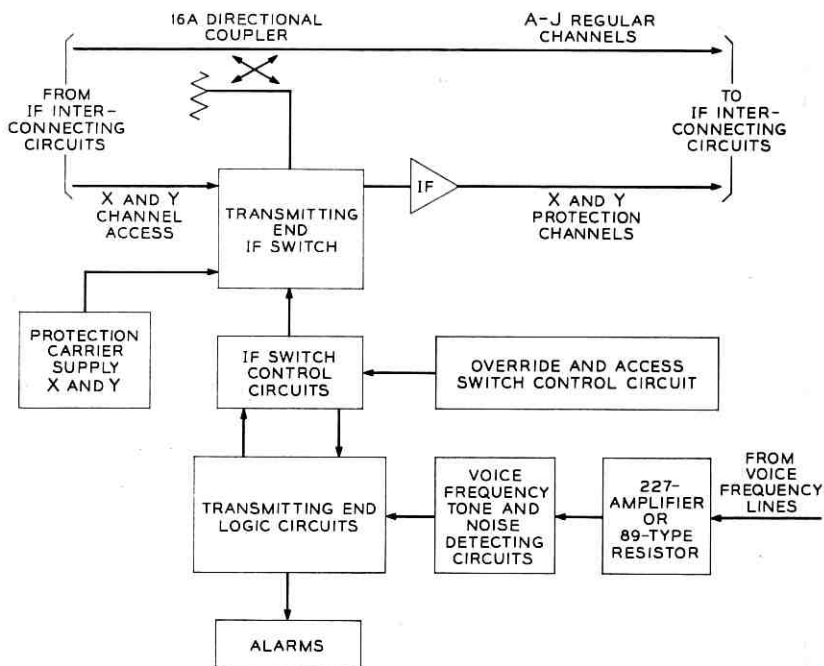


Fig. 3—Transmitting end block diagram.

transmitted over voice-frequency lines, one for each protection channel. The tone combinations used on each voice-frequency line are identical. Six tones are used. One tone is used as a guard tone and is normally present when no switches are in effect. The other five tones are used in two-tone combinations to provide ten individual channel switch orders.

Detecting circuits determine the presence or absence of tones by voltage-sensitive circuits and frequency-selective networks. The outputs of the detectors are connected directly to the transmitting end logic where the order is decoded. A valid switch order consists of the combination of two and only two order tones and the presence of the guard tone. After the switch operation is completed, the guard tone is removed. A valid order for a switch restoral consists of the removal of the two order tones and the reinsertion of the guard tone. If the order is a valid one, the transmitting end logic operates the IF switch through the IF switch control circuits. These switch control circuits, which are identical to those at the receiving end, in turn, inform the transmitting logic of the completion of the bridging operation.

Also located at the transmitting end of the switching section are the noise detecting circuits. These circuits, one for each voice-frequency line, monitor the noise in an unused band. If the noise becomes excessive, the noise detectors send an inhibiting voltage to the transmitting end logic. Operation of the noise detector will prevent the system from either making or removing a bridge to the protection channel associated with the noisy voice-frequency line.

The transmitting end logic originates a number of alarms to indicate system malfunctions. Three of these alarms are associated with the voice-frequency line equipment:

- (i) line failure alarm initiated by complete loss of tone
- (ii) noise detector alarm initiated by an operation of the noise detector
- (iii) invalid code alarm initiated by the receipt of an incorrect tone combination.

The remaining alarm which is associated with the IF switching is the transmitting prolonged bridge alarm. This alarm is initiated when a bridge has been in effect for longer than approximately 45 seconds.

VII. AUXILIARY STATION

The auxiliary switching station or TV drop switching equipment is a hybrid arrangement of the transmitting and receiving end equip-

ments. The IF switch is a receiving type transfer switch and the logic circuits are almost identical to those at the transmitting end. The voice-frequency tone and noise detecting circuits are identical to those used at the transmitting end. Override switching controls are omitted. However, should it be necessary, the same function can be performed by IF patch cords between conveniently located jacks.

VIII. IF SWITCHES

The IF switches used at the receiving, transmitting, and auxiliary stations are almost identical both electrically and mechanically. Fig. 4 is a block schematic of the switch as connected for the receiving end of a switching section.

The basic switching unit used in the IF switch is the 8-type gate. The 8-type gate is essentially a single-pole, single-throw diode switch. In its ON state the gate has a through loss of less than 1.5 db. In its OFF state the gate has a through loss in excess of 85 db. Each of the regular channels is associated with a group of three gates. As seen in Fig. 4, the regular channel is connected directly to one of the gates. The other two gates are connected to networks associated with the two protection channels. Only one of the three gates is ON at a time. The outputs of the three gates are connected to a common output by means of a 499A network. This network maintains a good 75-ohm transmission path from the ON gate to the channel output.

Each protection channel connects directly to the common input of a "one-by-twelve" 4051A network. One 4051A network is used to serve each protection channel. Ten of the outputs of the network connect directly to associated gates of the regular channels. Of the two remaining outputs, one connects through a gate to a termination, the other through a gate to the protection channel access out jack. Only one of the twelve gates associated with each network may be ON at a time. Thus, the protection channel may be connected to one of the regular channel outputs, if it is being used for protection, or to the protection channel access if it is being used for special message service or to a termination if it is idle.

When a system is not fully equipped, i.e., not all 12 channels are used in the system, 509A terminations which simulate the impedance of an OFF gate are used at the unused ports of the 4051A and 499A networks. Like the 499A network, the 4051A network is a band-pass network which absorbs the stray impedances of 11 OFF gates while preserving a good 75-ohm transmission path from its input port to the ON gate.

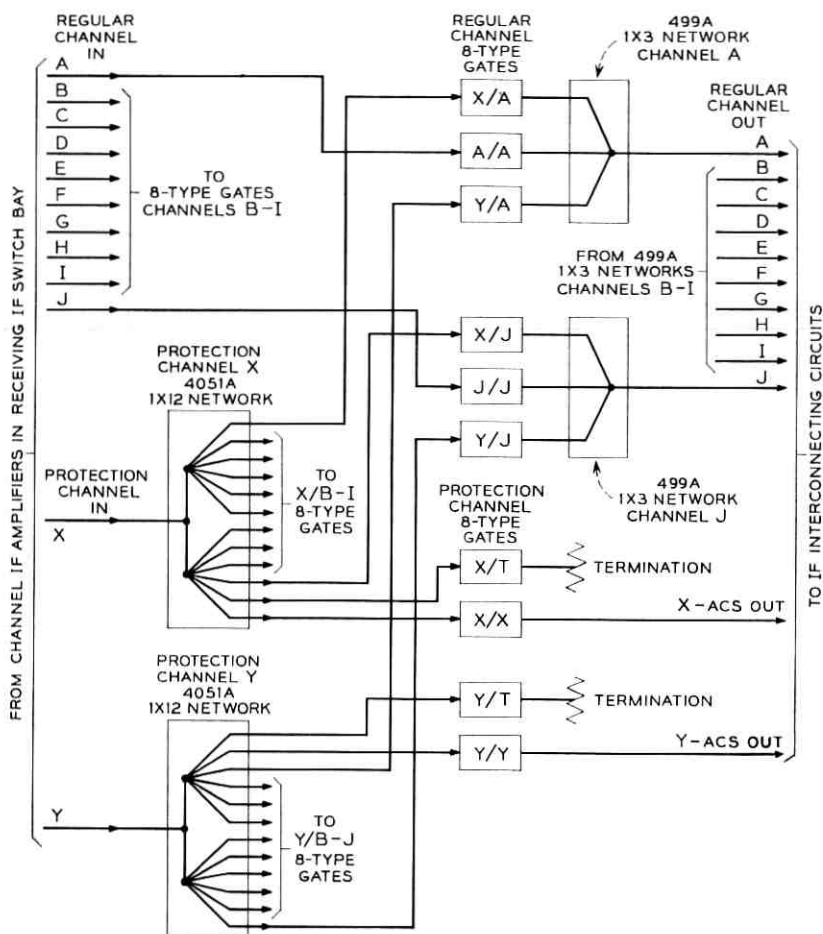


Fig. 4—IF switch—receiving end.

The 8-type gate, 499A network, and 509A terminations were originally designed for use in the TH protection switching system. They are described in an earlier B.S.T.J. article describing that system.² The 4051A network, however, was specially developed for the 100A system. The electrical basis of its construction is simple and is an extension of the design of the 499A (1-by-3) network. However, in order to preserve symmetry, control the stray reactances, and provide a transmission deviation of less than 0.1 db over the 60- to 80-mc band, the mechanical arrangement of the components becomes much

more critical. Fig. 5 is a photograph showing the internal construction of the network.

The IF switch at the transmitting end, which is shown in Fig. 6, is almost identical to that used at the receiving end. However, the direction of transmission through the switch is reversed. The high-loss arm of the directional couplers of the regular channel connect directly to the common port of the 499A networks. The common port of the 4051A network connects to the IF amplifiers of the protection channels.

Fig. 7 is a typical transmission characteristic for an IF switch. A photograph of a completely equipped IF switch is shown in Fig. 8.

IX. INITIATOR AND PILOT DETECTOR

A block diagram of the initiator and the pilot detector is shown in Fig. 9. The initiator is essentially an FM receiver with carrier and noise level sensing circuits. The pilot detector, which is used with the initiator on protection channels only, consists of a narrow-band 7-mc amplifier and level sensing circuit.

The input signal to the initiator is a sample of the 70-mc output of the final radio receiver of the switching section. The amplifier-limiter in the initiator suppresses any amplitude modulation present on the signal before applying it to the discriminator. The amplifier-limiter also provides additional automatic gain control to compensate for low limit gain IF amplifiers in the final TD radio receiver. Associated with

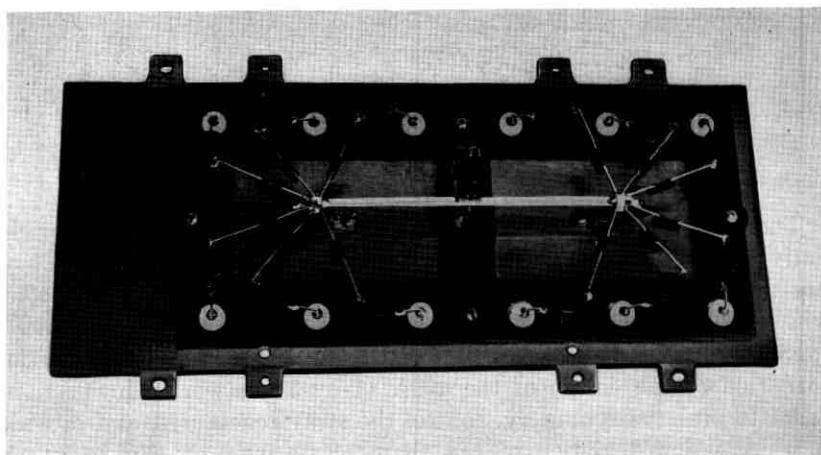


Fig. 5 — 4051A network.

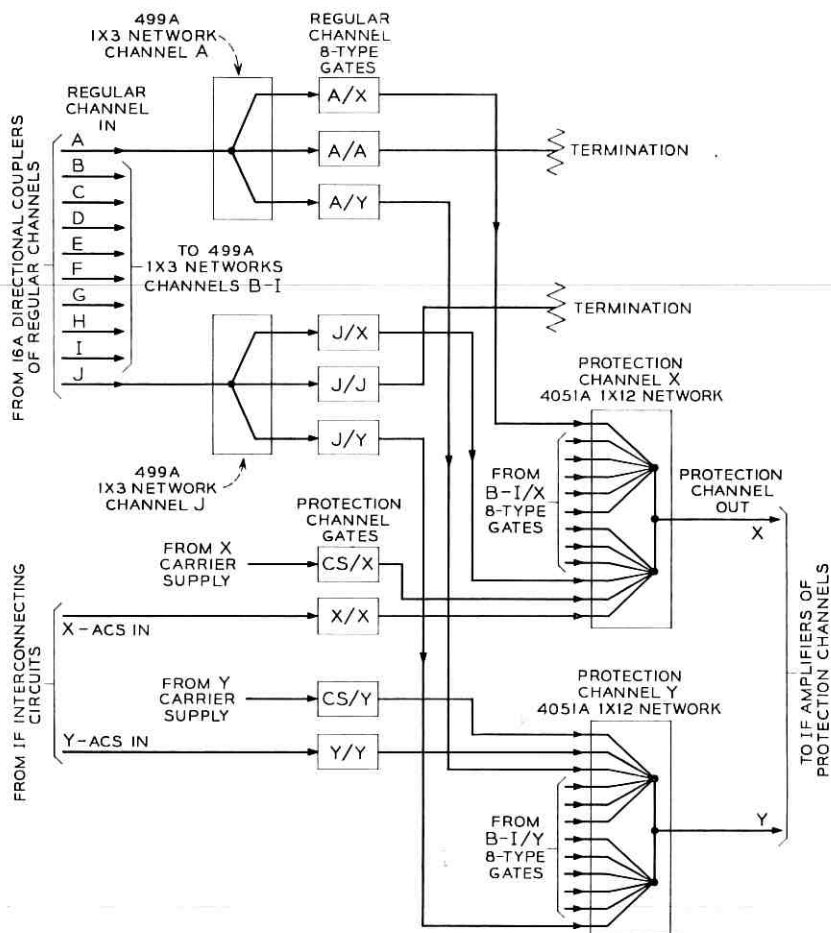


Fig. 6 — IF switch — transmitting end.

the amplifier is a narrow-band carrier detector whose output depends on the 70-mc input IF carrier power. The bandwidth of the carrier detector is so chosen that random noise will not simulate the presence of carrier.

The demodulated output of the discriminator is applied to both the pilot detector, which will be discussed later, and the noise amplifier detector portion of the initiator. The bandwidth of the noise amplifier is limited to about 100 kc at 9 mc by two similar band-pass filters. The output of the noise amplifier is applied to a noise detector. Its

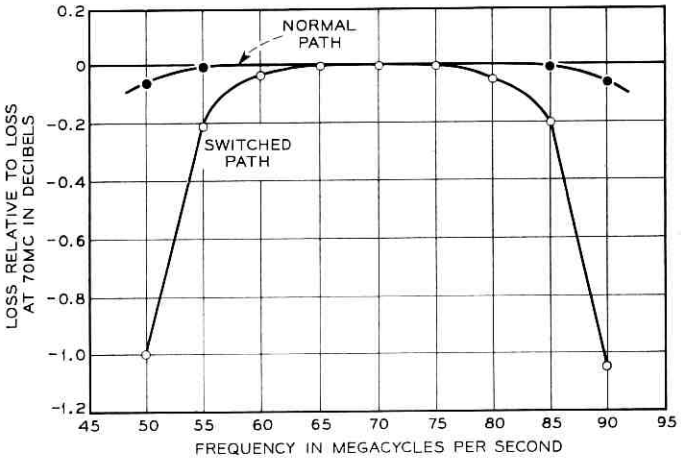


Fig. 7—Typical transmission characteristic of the IF switch.

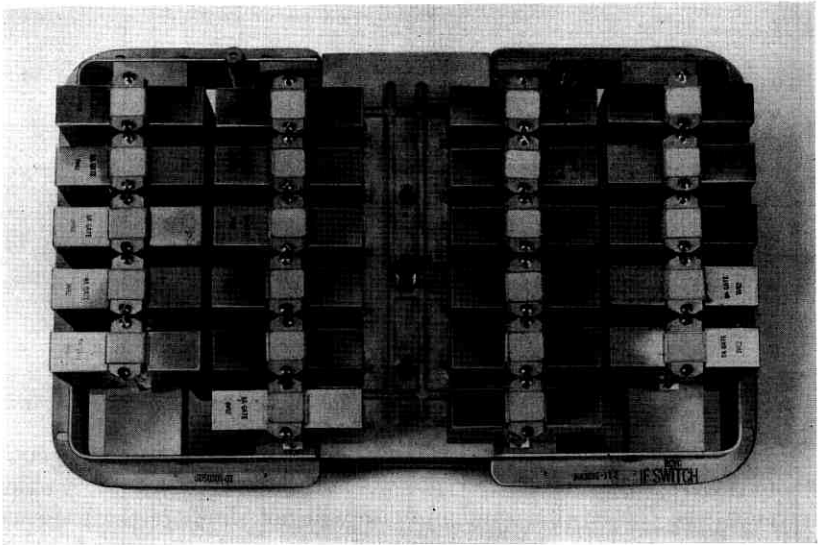


Fig. 8—Typical IF switch.

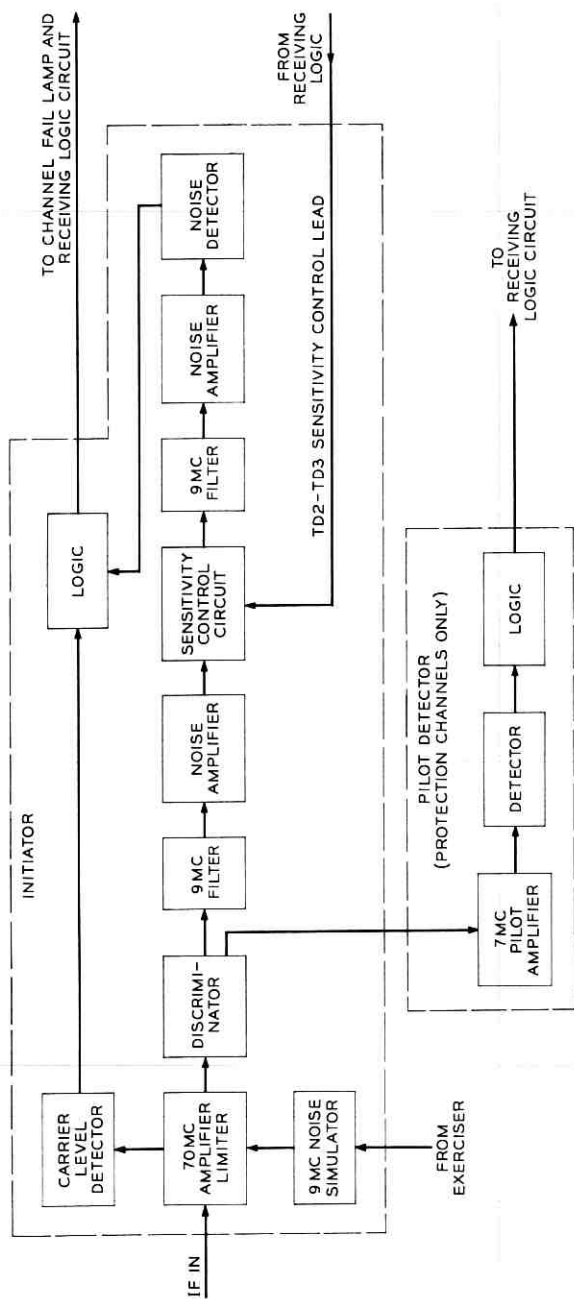


Fig. 9—Initiator and pilot detector block diagram.

output, proportional to the noise at 9 mc on the radio channel, is applied to the initiator logic circuit. The initiator logic circuit provides a channel status signal to the receiving end logic circuit which is +8 volts when the carrier power is adequate and the noise level sufficiently low and 0 volt when the carrier power is low or the noise is high. Its output is also used to provide a channel fail lamp indication.

The amount of noise (at 9 mc) required to obtain a 0-volt output from the logic circuit is adjusted by means of the sensitivity control circuit. Since the initiator is intended for use with both the TD-2 and the TD-3 systems, separate adjustments and a TD-2 — TD-3 switch are provided. The sensitivity may also be changed automatically by the receiving logic. This is required when TD-2 and TD-3 radio channels are used in the same switching section. Since TD-3 will have a wider baseband spectrum and thus will carry more message channels, the protection channels must be comprised of TD-3 equipment. The sensitivity switch of the protection channel initiator must therefore be set to the TD-3 position since the TD-3 system has a lower thermal noise than TD-2. The lower thermal noise results from a lower receiver noise figure and higher transmitted output power. Should the protection channel be required to protect one of the TD-2 radio channels, the TD-2 sensitivity will be selected automatically. The necessary control voltage is provided by the receiving end logic when the switch is ordered.

The initiator also contains a 9-mc oscillator which is normally not operating. On a signal from the exerciser, the oscillator is turned on to simulate a noisy channel during a system test cycle.

Fig. 10 shows an initiator assembly. There are three parts containing the following circuitry. The right side contains the IF amplifier-limiter and discriminator; the left side contains the selective noise amplifier and detector; and the section in the middle contains the logic circuitry.

The protection pilot detector is also connected to the discriminator of the initiator, but is contained in a separate unit. It consists of a narrow-band (approximately 200 kc) 7-mc amplifier followed by a detector and logic circuit.

Table I gives a summary of the important operating parameters of the initiator and pilot detector.

X. IF AMPLIFIERS

The 100A system uses a solid-state IF amplifier design having a maximum gain of about 26 db. A manual gain control provides a

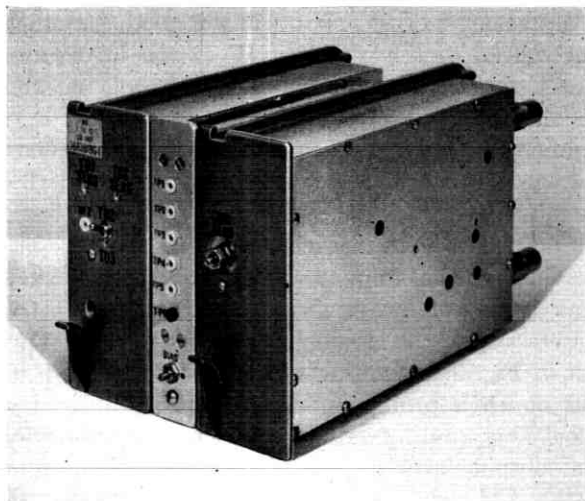


Fig. 10—Initiator unit front end.

TABLE I—INITIATOR AND PILOT DETECTOR CHARACTERISTICS

Initiator Noise Detector	
Center frequency	9 mc
Bandwidth	140 ke at 3-db down points
Switch point (channel bad)	adjustable over 20- to 40-db fade
	TD-2 and TD-3
	TD-2 position range: 30-50 db C/N*
	TD-3 position range: 45-65 db C/N*
Restore (channel good)	approx. 5 db from switch point
Initiator Carrier Detector	
Center frequency	70 mc
Bandwidth (nominal)	4 mc at 3-db down points
Switch point (carrier absent)	-3 dbm at initiator input†
Restore point (carrier present)	approx. 1 db above switch point
Pilot Detector	
Center frequency	7 mc
Bandwidth (nominal)	220 ke at 3-db down points
Nominal received pilot level	corresponds to 1.2-mc peak frequency deviation
Pilot absent point	14-db drop from normal pilot level

* Carrier-to-noise in 140-ke band.

† Corresponds to an 18-db drop in carrier level at the initiator 16-type directional coupler monitor. The IF amplifier used to compensate for coupler loss normally operates in compression.

gain range of at least 6 db. As shown in Fig. 2, an amplifier is used on each channel ahead of the receiving IF switch to ensure a fixed IF receiving level. An amplifier is used ahead of each initiator circuit to make up for the bridging loss of the couplers.

The amplifier has 7 grounded-base stages using 15-type transistors. Each interstage consists of a low-pass filter network which incorporates a broadband transformer.³ The gain-frequency response of the amplifier is adjusted by means of variable resistances in two of the interstages. Gain of the amplifier is controlled by a variable resistance network in one of the interstages. A typical gain-frequency response characteristic of the IF amplifier is shown in Fig. 11.

As shown in Fig. 3, a limiting-type amplifier is used at the transmitting end on each protection channel to stabilize the protection channel pilot. This limiting amplifier is a nine-stage device providing a maximum output level of -2 dbm, and a limiting circuit that suppresses the AM sidebands by greater than 20 db. The nominal input level is -26 dbm. The output level is adjustable from -7 dbm to -2 dbm. The gain-frequency response is similar to that of the amplifier previously described.

XI. PROTECTION CARRIER SUPPLY

The protection carrier supply provides two frequencies differing by 7 mc, one at 70 mc and the other at 63 mc at a much lower level. By passing these signals through the limiter amplifier at the transmitting end, an FM signal is generated which is detected later by the initiator and its associated pilot detector at the receiving end.

A block diagram of the protection carrier supply is shown in Fig. 12. This unit is a plug-in assembly similar to the other IF units.

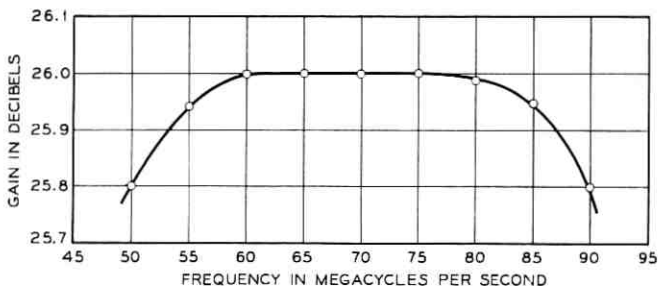


Fig. 11—IF amplifier typical gain-frequency characteristic.

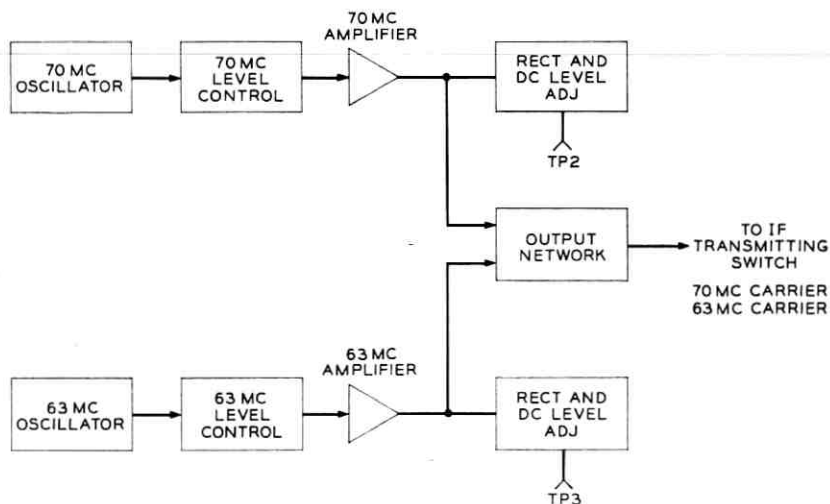


Fig. 12 — Protection carrier supply block diagram.

XII. RECEIVING LOGIC

The major portion of the logic and control circuits for the system is concentrated in the receiving logic. The receiving logic is made up of transistor resistor logic circuits, flip-flop storage or memory circuits, delay circuits, cross-connecting diode matrix circuits and gate circuits. These basic circuits are similar in design and operation to those used in the TH protection switching system.² For ease in servicing and maintenance, the circuit components are housed in plug-in units. There are twenty-eight types of plug-in units. A total of one hundred and ten plug-in units are required for a completely equipped receiving logic circuit serving a two-for-ten switching system.

XIII. ESTABLISHING A SWITCH

When a regular channel initiator indicates that a regular channel is not usable, its 0-volt output signal results in the operation of a status flip-flop memory circuit in the receiving logic. If a protection channel is available, a switch assignment is made by the receiving logic to that channel. The assignment circuits are so arranged that each regular channel prefers to be assigned to a particular protection channel. However, if the preferred channel is not available, the assignment will be made to the other protection channel. In the following description,

it will be assumed that regular channel A has failed and has been assigned to protection channel X. At the time this assignment is made, a 50-millisecond switch initiation timing circuit is energized. For the duration of this 50-millisecond timing interval, a gate circuit opens the connection between the regular channel initiator and the status memory circuit. Also, for the duration of the interval, the protection channel X is marked "good" to those logic circuits involved in a normal A to X switch. Thus, once started, the switch is allowed time to complete. Also, at the time the assignment is made, an order is sent to encoding circuits to turn on two order-tone oscillators in the voice-frequency equipment. The arrival of the tones at the transmitting end of the system causes the transmitting logic to order an A to X bridge.

When the bridge is made at the transmitting end, the pilot on the X protection channel is removed. The loss of the protection pilot is detected by the X channel pilot detector at the receiving end. If the "no pilot" indication and the original A to X assignment order are both present, an order is sent to the receiving switch control circuits which, in turn, operate the receiving end IF switch to transfer service from channel A to channel X. When the switch control circuit operates, it generates a receiving switch verification signal which causes the X protection channel voice-frequency guard-tone oscillator to turn off. The removal of the guard tone provides an additional lock on the transmitting end bridge. The receiving switch verification signal also causes an A to X switch register and an X switch register to operate. These registers indicate the number of completed A to X switches and also the total number of switches to protection channel X.

The switch from A to X is now completely established. When the 50-millisecond time interval has elapsed, the regular channel initiator is reconnected to the channel A status memory circuit in the logic and the protection channel "mark good" voltage is removed. If channel A is still requesting a switch, the A to X switch is maintained by the receiving logic. The switching operation is now complete.

XIV. RELEASING A SWITCH

When the regular channel A recovers while switched to protection channel X, the output from the channel A initiator returns to its normal "good" voltage condition, the flip-flop memory circuit at the input to the receiving logic returns to its normal "no switch request" state, the switch assignment A to X is cancelled, and the receiving end switch control circuits are ordered to restore the receiving IF switch.

When the assignment is released, a 50-millisecond switch release timing circuit is started. For the duration of this 50-millisecond interval, the channel A initiator is disconnected from the input memory circuit of the receiving logic and the protection channel is marked "bad" to the logic circuits involved in a normal A to X switch. Thus, the switch is allowed time to release completely before the regular channel A can initiate a second switch request or before one of the other regular channels can initiate a switch to channel X.

When the switch control circuits operate, a signal is sent back to the receiving logic to verify that the receiving end switch has been restored to normal. If this verification signal and the X assignment release signal are both present, logic signals are generated which return all circuits to their normal "no switch request" state, turn off the two order-tone oscillators and turn on the guard-tone oscillator. This constitutes an order to release the transmitting end bridge and the auxiliary station switch.

When the transmitting end bridge has been released, the pilot is returned to channel X and the pilot detector at the receiving end so informs the receiving logic.

XV. FAILURE TO INITIATE A SWITCH

If a switch to a protection channel is initiated but for some reason the switch is not completed during the 50-millisecond switch initiation timing interval, the receiving logic will initiate a switch to the other protection channel. If the switch to the second protection channel also cannot be completed, a double switch initiation failure occurs. This results in the regular channel being automatically locked out, i.e., its switch request is not honored. However, at ten-second intervals, the lock out on the regular channel is removed and the receiving logic will again attempt to complete the switch. The process continues until the regular channel is locked out manually or the switch is completed. A more detailed description follows.

When an A to X switch is not completed at the end of the 50-millisecond timing interval, a 65-millisecond "switch initiation failure X" timing circuit is energized. For the duration of the 65-millisecond interval, channel X is marked bad to the logic and forces the assignment circuits to make an A to Y assignment. The A to X order is cancelled and an A to Y switch started in the normal manner. If, however, the A to Y switch is not completed at the end of the 50-millisecond switch initiation timing interval, then a 65-millisecond

“switch initiation failure Y” timing circuit is energized. For the duration of its 65-millisecond interval, channel Y is marked bad. At this point, both protection channels are marked bad due to the overlap of the switch initiation failure intervals. The combination of the regular channel A status memory still requesting an assignment, and both the X and Y channels marked bad to the logic results in a “lock-up” in the assignment circuits. This lock-up prevents A from being assigned to either protection channel after the end of the switch initiation failure intervals. The locked-up state remains until an unlocking pulse is applied. This is generated by a circuit which provides a pulse once every ten seconds.

XVI. PREVIOUS SECTION FAILURES

When a regular channel fails or becomes excessively noisy, it may also appear failed in succeeding switching sections. The regular channel initiators will request a switch and the receiving end logic will order the head end bridge to be made. After the bridge is made, both the regular and protection channel will appear to be bad if the failure is in a preceding section. This apparent simultaneous failure is an indication to the receiving logic that the failure could be a previous section failure. The sequence followed by the logic under these circumstances depends upon the duration of the regular channel failure. If the channel recovers after the transmitting end is bridged, but before the end of the 50-millisecond switch initiation timing interval, the receiving logic performs the sequence for a temporary previous section failure. If the failure exists for more than 50 milliseconds, the receiving logic performs the sequence for a permanent previous section failure. In neither case is the receiving switch operated. However, one or both the protection channels will not be available for use while the receiving logic is carrying out these sequences.

16.1 *Temporary Previous Section Failure*

When the head end bridge is made, as the result of the previous section failure, the pilot detector will indicate removal of the pilot and the protection channel initiator will indicate a bad protection channel. Such a combination of signals in the receiving logic prevents the generation of a “no pilot” signal and the receiving end switch is not made. However, as soon as the regular channel recovers in the previous section, both the regular and protection channel initiators will indicate good channels. If the failure is of a tem-

porary nature so that the channel recovers within the 50-millisecond switch initiation timing interval, the receiving end switch would complete were it not for an additional inhibiting circuit not normally used in the ordinary switching sequence. The inhibiting circuit used includes delay and sequence circuits. The delay is applied to the generation of the no pilot signal so that it cannot appear until about 3 milliseconds after the indication from the protection channel initiator is changed from bad to good. This ensures that the regular channel initiator output will always indicate "channel good" before the no pilot signal is generated. The output of the regular channel initiator and the no pilot signal are then combined in a sequence circuit. This sequence circuit provides a receiving switch inhibiting voltage when the regular channel becomes good before the no pilot signal is generated. Thus, a receiving end switch is prevented.

16.2 *Permanent Previous Section Failure*

A previous section failure, which lasts longer than 50 milliseconds, is interpreted by the logic as a failure to initiate a switch and the previously described sequence for failure to initiate a switch is carried out.

XVII. FAILURE OF PROTECTION CHANNEL

If the protection channel fails while not in use, it is not available for an assignment. However, if the protection channel fails while in use, the sequence followed by the receiving logic depends on whether it fails due to loss of carrier or excessive noise.

If the protection channel fails while in use due to loss of carrier, which will occur with equipment type failures, service will be transferred to the second protection channel. If the second protection channel is not available however, no action is taken. If the second protection channel is available, loss of the carrier results in the generation of an order to the assignment circuits to drop the original assignment and to make a new assignment to the second protection channel. The A to X receiving end switch is first released and then simultaneously the order to remove the A to X bridge and the order to establish the A to Y bridge is sent to the transmitting end of the system. Finally, when the transmitting end bridge is made and the protection pilot is removed, the new A to Y receiving end switch is made. It should be noted that during the switching sequence, the regular channel A was carrying service from the time the A to X receiving end switch was

dropped until the A to Y receiving end switch was established. The degree of service interruption suffered will therefore depend upon the state of the regular channel. The maximum interruption time for the transfer from X to Y is the time taken to put up a normal switch which is approximately 30 milliseconds.

The receiving logic takes no action on a protection channel in use which fails due to noise. This design choice was made because of the undesirability of exchanging a slightly noisy channel for the circuit interruption involved in making the transfer to the second protection channel.

The return of the pilot to a protection channel while a switch is established to it is treated by the receiving logic as if the protection channel had failed due to loss of carrier, i.e., a transfer is made to the second protection channel. Return of the pilot may be caused by an equipment failure in the switching system or by the removal of the transmitting end bridge due to a false order from the voice-frequency facility. This false order, for instance, could be generated by the opening up of the voice-frequency line and the insertion of a test tone at the guard-tone frequency.

XVIII. TRANSMITTING END LOGIC

The transmitting end logic establishes and takes down bridges from regular to protection channels in response to signals received over the voice-frequency lines from the receiving end logic. A fully equipped transmitting end logic consists of a total of 19 plug-in units. The 19 plug-in units include five types of circuits, each of which is made up of combinations of gates, delay, and memory circuits. These individual circuits are similar to those used in the receiving end logic.

The transmitting end logic is made up of two similar groups of equipment, one controlling bridges to protection channel X, the other to protection channel Y. Each part receives its bridge orders from the tone detectors of the voice-frequency line associated with its respective protection channel. Both parts connect to the switch control circuits which, in turn, through the IF switches, set up the bridges between the regular and protection channels.

18.1 *Establishing a Bridge*

The outputs of the tone detectors connect directly to decoding and code validating circuits. An order for a bridge, for example from A to X, requires the presence of two order tones and the presence of the

guard tone on the voice-frequency line serving channel X. The presence of this combination of tones results in an order at the output of the decoding circuits for an A to X bridge. However, the valid code circuit delays the order from proceeding to the switch control circuit for approximately 6 milliseconds. If the order remains without interruption for the full period, the valid code circuit allows it to proceed, otherwise an invalid code alarm is initiated and a bridge is not ordered.

When the A to X bridge is made by the IF switch, the channel X carrier supply is disconnected from the X protection channel and the IF switch control circuit sends a bridge verification signal back to the transmitting end logic. On receipt of the verification signal, a 30-millisecond timer is started in the transmitting end logic. The output of this timer circuit is coupled to the valid code circuit which, in turn, prevents any new order from proceeding to the switch control circuit during the timing interval, thus temporarily locking up the bridge. The removal of the pilot tone of carrier supply is an indication to the receiving end logic that the bridge has been made and the receiving logic then removes the X guard tone. When the absence of the X guard tone is detected at the transmitting end, an inhibiting voltage is generated, which locks up the bridge.


XIX. RELEASING A BRIDGE

Once a bridge is established, the two order tones must be removed and the guard tone must be restored in order to release the bridge. When the order tones are removed, the decoding circuits generate the order to release the bridge. However, as in the case of establishing a bridge, the valid code circuit prevents the order from proceeding for 6 milliseconds. At the same time, the return of the guard tone is detected by the bridge lock-up circuits. If the order to take down the bridge remains for the 6 milliseconds and the guard tone is still present, the order is allowed to pass to the IF switch control circuits. The IF switch then takes down the bridge and returns the carrier supply to the protection channel. Once the bridge is ordered released, the switch control sends a verification voltage to the transmitting logic and all circuits are returned to their normal state.

XX. VOICE-FREQUENCY EQUIPMENT

Table II lists the frequencies and the order-tone codes used to control the transmitting end bridges and the auxiliary station switch-

TABLE II—FREQUENCY AND ORDER TONE CODES
Tone Codes Required when Requesting a Switch Order

Noise Detector	Switch to X or Y	Code Tones					Guard Tone
	A	P	P	O	O	O	P
	B	P	O	P	O	O	P
	C	P	O	O	P	O	P
	D	P	O	O	O	P	P
	E	O	P	P	O	O	P
	F	O	P	O	P	O	P
	G	O	P	O	O	P	P
	H	O	O	P	P	O	P
	I	O	O	P	O	P	P
	J	O	O	O	P	P	P
	No order	O	O	O	O	O	P
	Freq. cps 900 1400		1615	1785	1955	2125	2295

P designates presence of tone; O designates absence of tone.

ing. Two identical sets of six voice-frequency tones are used, one set for each of the voice-frequency lines. As discussed previously, one voice-frequency line is associated with each protection channel. When all of the radio circuits are normal, one of the tones, the guard tone, is present on each line. A two out of five selection of the remaining tones, the order tones, permits switch orders to be sent for any one of the ten working channels. The use of two tone combinations minimizes the possibility of setting up false bridges due to noise or interfering tones. However, false switch orders can occur if the noise becomes high enough in spite of the coding arrangement. For this reason, voice-frequency noise detecting circuits are used. These noise detecting circuits monitor the noise in a 500-cps band centered at 1150 cps. When the noise becomes excessive, the noise detector generates an inhibiting voltage. This voltage is applied to the logic circuits in such a way as to prevent any bridge or switch orders, false or real, from being acted upon at the transmitting or auxiliary station, respectively. At the same time an alarm is given. Thus, the noise detecting circuits guard the switching system against false signaling orders due to high noise on the voice-frequency line by maintaining the system in *status quo*. The noise detector inhibiting action is particularly valuable when a radio system is used to provide the voice-frequency circuits.

20.1 Control Tone Source

Six tone oscillators are required for each protection channel. Each tone oscillator circuit consists of a transistor oscillator, which is oper-

ated continuously, followed by a diode switch and a transistor amplifier. The switch which connects the output of the oscillator to the amplifier is under the control of the receiving logic. The output of the amplifier is connected by a resistive network to the other five circuits and to the voice-frequency line. A block diagram of the control tone source is shown in Fig. 13.

20.2 Tone and Noise Detecting Circuits

A block diagram of the tone and noise detecting circuits is shown in Fig. 14. A low-pass filter with a cutoff frequency above the highest tone frequency used provides protection against interference from any high frequency generated in the office where the detectors are located. Each tone detector is preceded by a voice-frequency amplifier and a band-pass filter which selects the desired tone. The outputs of the tone detectors are applied to the transmitting or auxiliary station logic circuits. The noise detecting circuit is similar to that used for tone detecting, except for the wider filter bandwidth. This wider bandwidth makes the response time of the noise detector less than that of the tone detectors. For this reason the noise detector will operate before tone detectors when a pulse of random noise is applied to the circuit.

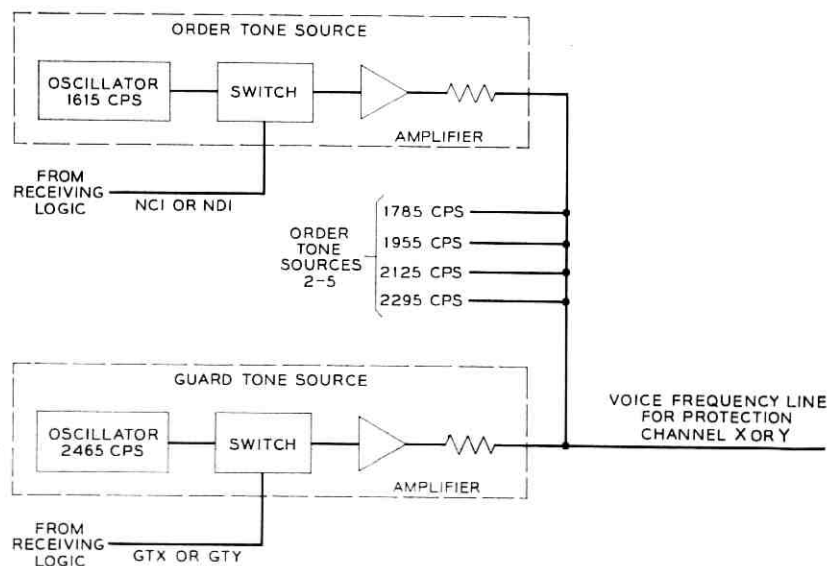


Fig. 13—Control tone source block diagram.

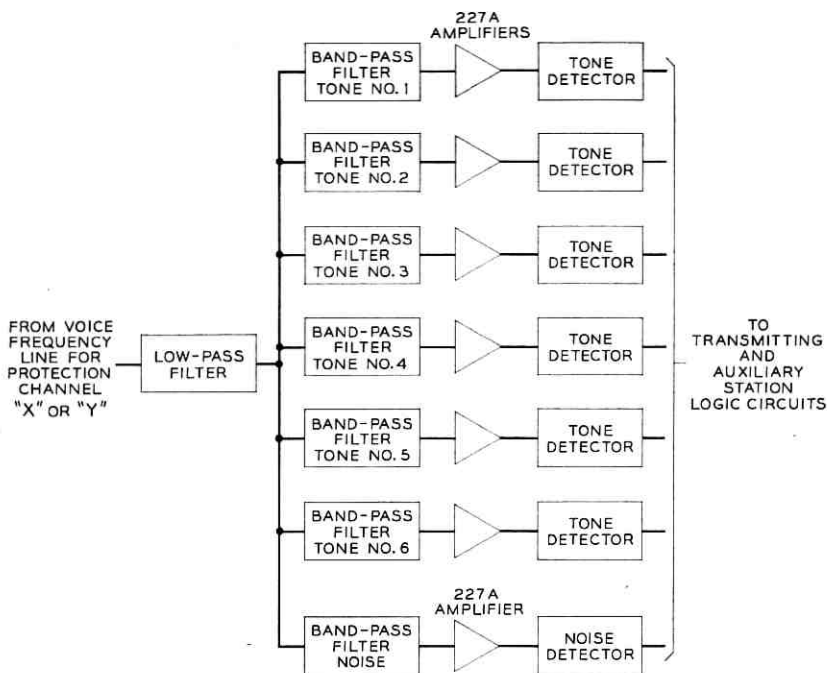


Fig. 14 — Tone and noise detecting circuits block diagram.

XXI. EXERCISER

The exerciser circuit performs test functions in the 100A Protection Switching System. A successful completion of its test routine indicates that when the initiator calls for a switch or a release that a satisfactory switch completion and switch release is made. An alarm is given by the exerciser if the test routine is not completed satisfactorily.

The test routine consists first of a check that the initiators on the protection channels are operating and indicating that the channels are good. A simulated carrier failure is then made on the first regular channel by applying a ground to the input stages of the initiator for that channel. By momentarily marking the "Y" protection channel bad in the receiving logic, the switching system is forced to provide protection by switching to the "X" protection channel. If evidence is provided by a verification signal from the receiving logic to the exerciser that the transmitting bridge and receiving switch were completed in the proper order, the switch is ordered to release. If the switch is released (also in the right sequence), the exerciser advances

to the next channel. A similar procedure is followed for the remaining regular channels. Upon completion of this sequence, the exerciser then checks switches and switch releases to protection channel Y by simulating noisy regular channels. This is performed by turning on the 9-mc oscillator in the initiator of each regular channel in turn. The switches are forced to "Y" by momentarily marking the "X" channel bad in the receiving logic. An alarm is given if any switch or release cannot be made within a predetermined time interval or if a wrong sequence of operation occurs. For example, an alarm is given if a verification of a receiving end switch is received before a head end bridge is made. A visual indication shows which working channel and which protection channel are involved in the failure. The time taken for a test routine, if no source of trouble is encountered, is less than 5 seconds for a fully equipped system. The exerciser is arranged so that if a request for protection is made by a regular channel during the test routine, the exerciser will disengage in about 1 millisecond to permit normal operation of the system.

In addition to providing a test routine, the exerciser may be used to simulate a permanent failure of any one of the regular channel initiators on carrier or noise and so force a switch to protection. The exerciser may also be used to provide simulated channel failures repetitively. Both of these features are useful for system maintenance and troubleshooting.

The exerciser does not completely check out all of the operations the system is capable of performing. It will not check the procedure for previous section failures, transfer of a switch from one protection channel to the other should the first protection channel fail, etc. Tests to check these operations are performed during normal routine maintenance.

The exerciser uses a timing clock to start the automatic test routine. The clock may be set for one operation during a twenty-four hour period. The time selected for the automatic operation should be one at which minimum automatic switching is usually experienced. The exerciser assembly and its associated key and lamp mounting are shown in Figs. 15(a) and 15(b).

XXII. DESCRIPTION OF BAY EQUIPMENT

The protection switching equipment is mounted on bays which are completely shop wired and tested for ten regular and two protection channels. These bays may be partially equipped with plug-in units by the shop as specified on order by the customer. At the receiving end,

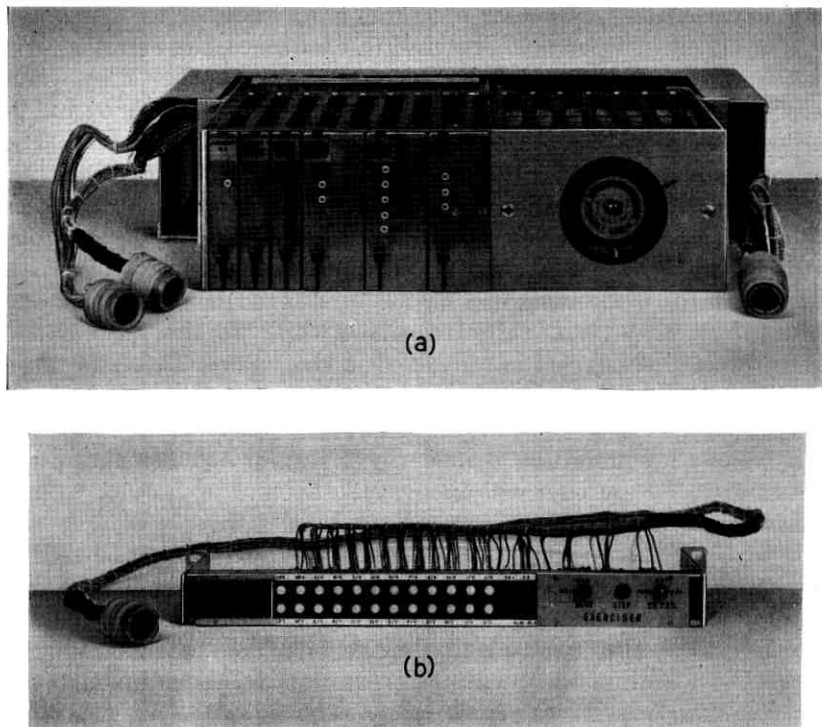


Fig. 15—Exerciser: (a) assembly, (b) key and lamp panel.

two bays are required. The first is the receiving IF switch bay as shown in Fig. 16. This is a 9'-0" duct type bay arranged to mount the IF switch and associated equipment together with initiators for the ten regular channels and two protection channels. The exerciser and its associated key and lamp panel are also mounted in this bay. The +24 volt and -24 volt supplies are obtained from a fuse panel which is mounted near the top of the bay. The fuse panel also distributes -20 volts which is a regulated voltage for operation of the IF plug-in units. The -20 volts is obtained from two pairs of voltage regulators mounted on this bay. An automatic transfer unit and a manual switch unit are provided with the regulators for maintaining a reliable regulated supply at all times. The second receiving end bay is the receiving control bay as shown in Fig. 17. This bay mounts the complete two-by-ten receiving end logic with its associated manual controls and tone oscillator facilities together with the voice-frequency line equipment associated with the transmission of these tones. Another fuse panel is

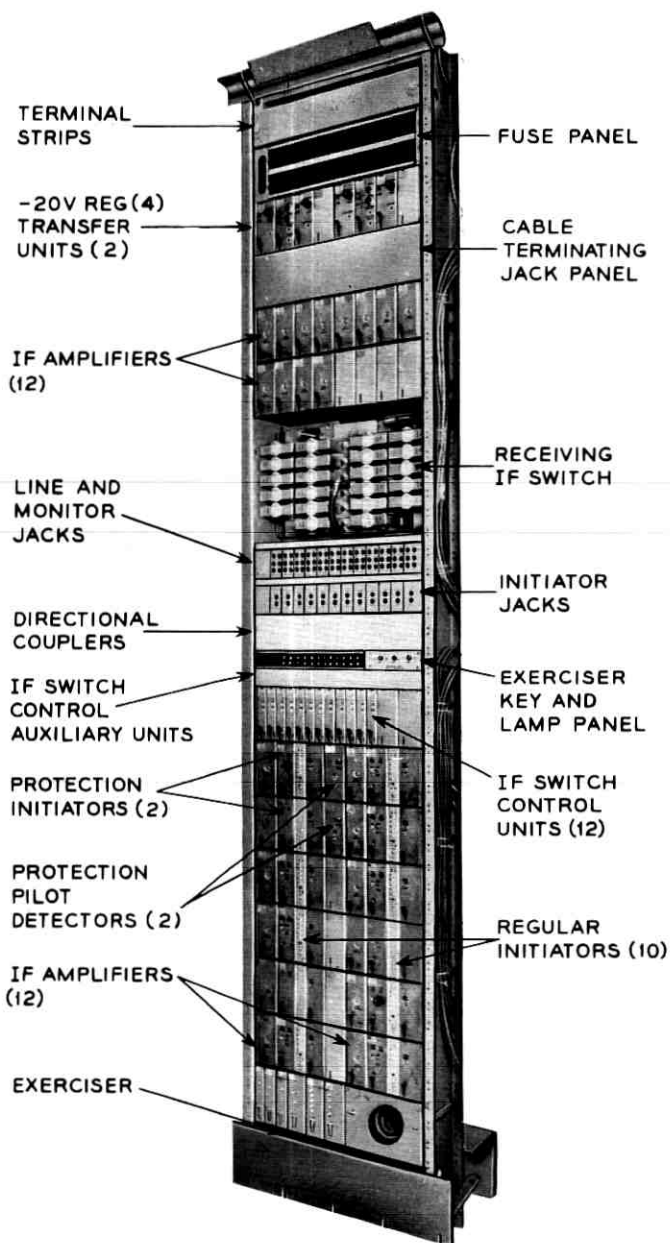


Fig. 16 — Receiving IF switch bay.

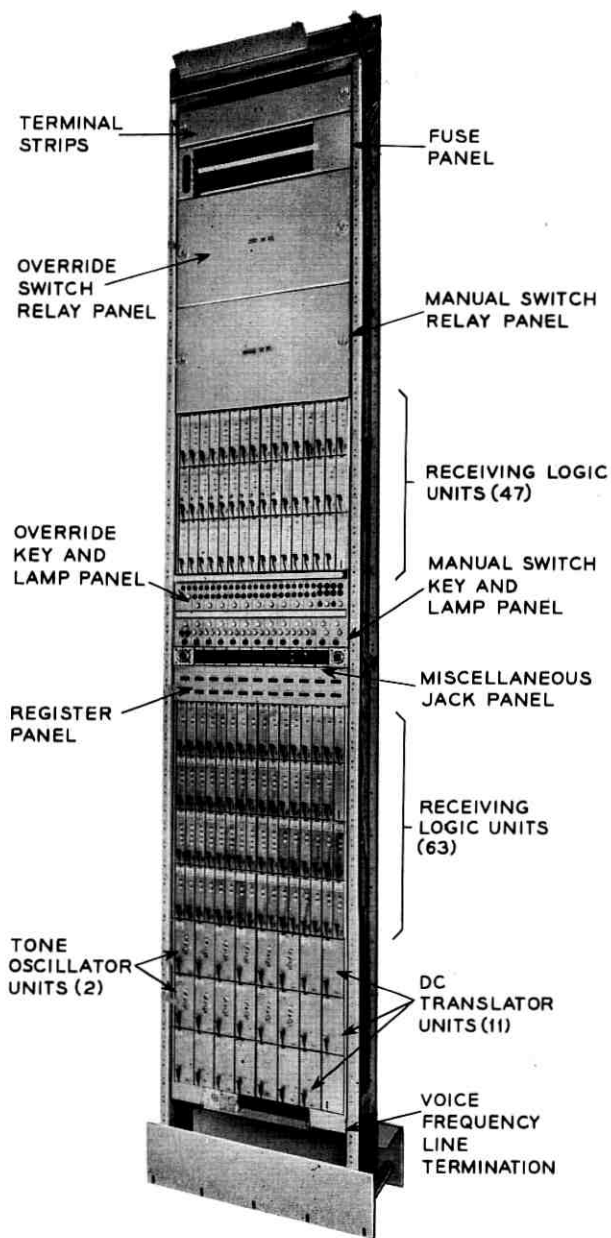


Fig. 17 — Receiving control bay.

provided for the distribution of the -24 volts and $+24$ volts required for the equipment in this bay. An override switch relay panel and a manual switch relay panel and their associated key and lamp panels are used to perform by manual operation the functions that are performed automatically by the logic circuitry. A register panel is used to record the number of switches that are made to the two protection channels. Dc translator plug-in units are used to provide sufficient current for lamp indications of alarms or failures detected by the transistorized plug-in units. A miscellaneous jack mounting is used to mount the alarm lamps and voice-frequency line jacks. Two test connectors are also provided on this mounting which provide a means of coupling the 100A test console to the receiving end logic circuitry for performing routine tests.

At the transmitting end, the transmitting IF switch and control bay shown in Fig. 18 mounts the IF transmitting switch and its associated IF switch control equipment. The plug-in units in this bay are the transmitting end logic, carrier supply, IF amplifiers, tone detectors, noise detectors, and dc translators. The voice-frequency line equipment for receiving the guard tones over the VF lines is also mounted here. An override switch relay panel and associated key and lamp panel are also provided for manual operation of the transmitting logic circuitry. A miscellaneous jack mounting contains the alarm lamps, voice-frequency line jacks and a test connector for performing routine circuitry tests.

The auxiliary station bay equipment is very similar to the arrangement for the transmitting IF switch and control bay with some minor differences. No override switch relay panel is furnished and the logic circuit uses an enabling and lock-up plug-in unit instead of a bridge lock-up unit. An auxiliary IF switch and control bay is required for each direction of transmission. Routine tests at an auxiliary station are fairly simple and are performed with a small plug-in test unit that checks out the logic functions.

XIII. DESCRIPTION OF UNITS AND PANELS

The major units in the 100A system are arranged as plug-in units. These units are primarily the ones containing active circuits that may require the most maintenance. In the event of a circuit failure, a failed unit can be readily replaced and thus reduce circuit outage time to a minimum. Other units are required that do not require as much maintenance and, therefore, can be permanently wired into the over-

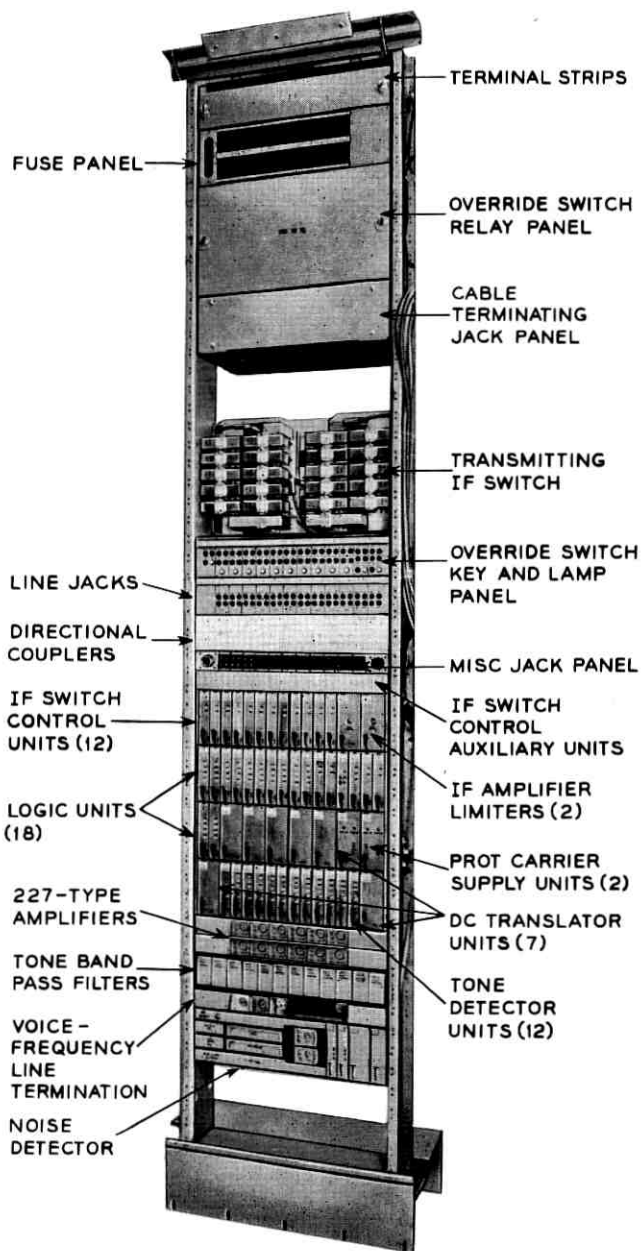


Fig. 18—Transmitting IF switch and control bay.

all bay equipment. These units or panels are secured to the bay uprights and become part of the basic structure.

The 100A protection switching system plug-in units are of two types. The first is the open unshielded type which contains primarily dc circuits for logic, alarms, dc regulators, and transfer units as well as low-frequency tone oscillators and detectors. The second is the shielded type which contains all IF circuitry such as that required for amplifiers, carrier supplies, and other high-frequency circuits associated with the initiators.

Other units required in the bay arrangements are mountings for the plug-in units, VF line terminating equipment panels, fuse panels, jack and lamp panels, and relay panels for the override and manual circuitry.

XXIV. SWITCHING TEST SET

The switching test set which is used at the transmitting and receiving ends of the section incorporates most of the test equipment required for checking the operation of the 100A system and for testing its component parts. The equipment included in the test set is listed below.

- (i) IF power meter
- (ii) IF unit test panel
- (iii) electronic counter with time interval counter
- (iv) control unit test panel
- (v) audio oscillator
- (vi) electronic voltmeter
- (vii) bay logic test panel
- (viii) power supplies
- (ix) dc voltmeter.

All units comprising the test set are mounted in the rolling console shown in Fig. 19. In addition to the test console, a high-frequency oscilloscope is required to perform certain tests.

24.1 *IF Testing*

The test set provides means to power the initiators, pilot detectors, carrier supply, and the IF amplifiers. The carrier supply oscillator frequencies may be checked with the counter, and the levels measured with the power meter. Initiator and pilot detector trip and restore points may also be checked. Transmission and return loss measurements of IF units are made using separate IF test sets generally available in the radio station while the IF units are being powered from the test set.

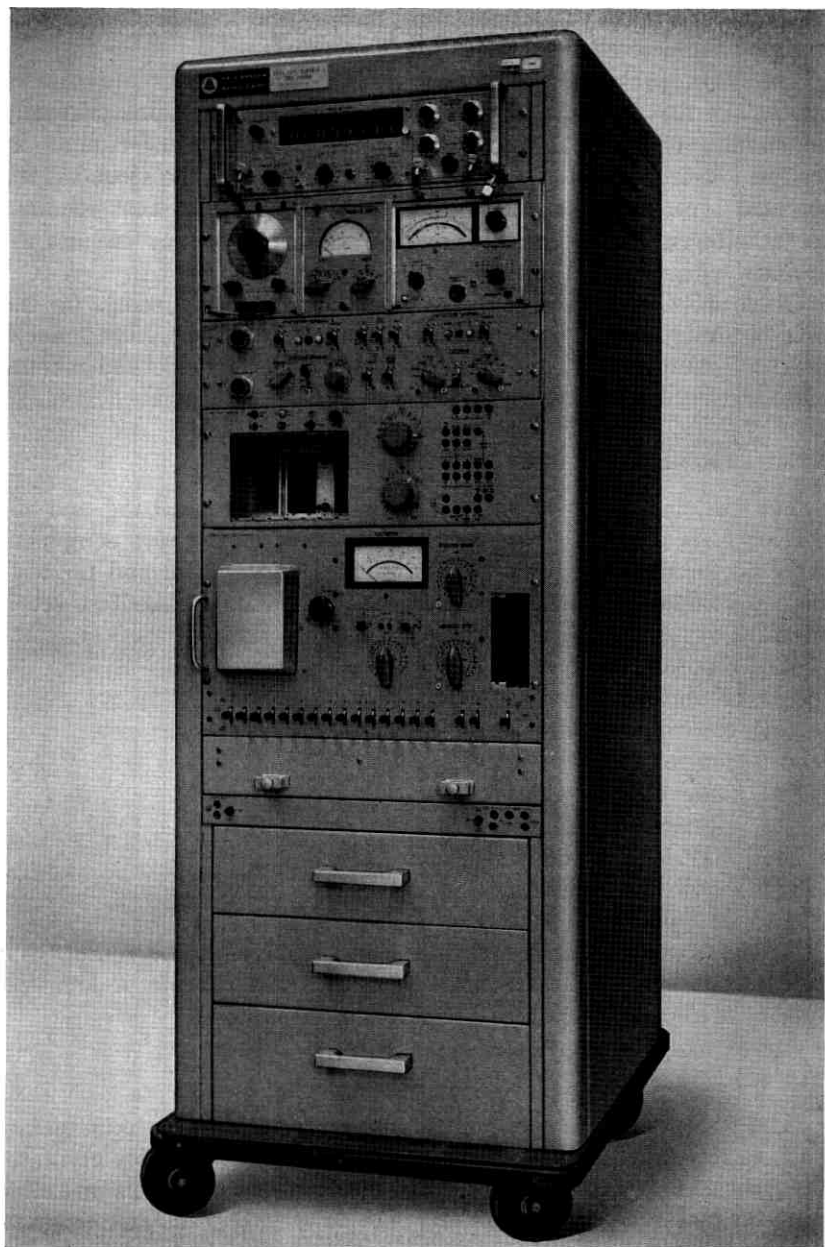


Fig. 19 — Test set console.

24.2 Control Unit Test Panel

The control unit test panel provides means to test all of the logic and control type units used in the 100A system. The unit under test is first inserted in a test slot. The proper input and output voltages for the unit under test are then set up automatically by a card reader and a perforated test card. Test keys are provided to change the test voltages in accordance with the required test sequences. Response of the unit may then be measured in terms of dc voltage, frequency or time interval by using the instruments in the test set.

24.3 Bay Logic Panel

The bay logic panel is used to check the operation of the transmitting and receiving end logic circuits. It is connected to the bay under test by means of a cable and test connectors mounted on the bay. The panel simulates the opposite end of the system to the bay under test, i.e., the transmitting end when the receiving end equipment is being tested and vice versa. By means of the operating controls, channel failures may be simulated on a manual or repetitive basis. The resulting operation of the logic may then be followed by means of the lamp display and the test oscilloscope. The switching system is always placed in *status quo* before the bay logic panel is connected, hence no actual system switching takes place during test.

At the auxiliary station, the testing of the logic circuitry is performed with a small plug-in unit which contains a set of switches for checking each channel individually. This arrangement is adequate due to the relative simplicity of the equipment.

24.4 Voice-Frequency Tests

The audio oscillator is used to provide tones to check the voice-frequency tone detectors, and the electronic voltmeter to measure received tone levels.

24.5 Test Instructions

A complete set of test instructions for each of the units, both logic and IF, are included in a file drawer in the test console.

XXV. POWER PLANTS

The 100A equipment requires sources of both +24 volts and -24 volts with battery reserve. Since the 100A system uses transistor-

resistor logic circuits, the power source must be free of relay transients and excessive noise. In some stations -24 volt plants with adequate capacity are installed. These may be used providing the feed to the 100A is adequately filtered. Where -24 volts is not available, standard power plants may be used. A small $+24$ volt plant was developed specifically for the 100A application.

REFERENCES

1. Weber, I., Evans, H. W., and Pullis, G. A., Protection of Service in the TD-2 Radio Relay System by Automatic Channel Switching, B.S.T.J., 34, May, 1955, pp. 473-510.
2. Giger, A. J., and Low, F. K., The Automatic Protection Switching System of TH Radio, B.S.T.J., 40, Nov., 1961, pp. 1665-1715.
3. Ruthroff, C. L., Some Broadband Transformers, Proc. IRE, 47, Aug., 1959.

Maximally Reliable Exponential Prediction Equations for Data-Rate-Limited Tracking Servomechanisms*

By H. D. HELMS

(Manuscript received July 8, 1965)

Certain radars, sonars, and other sampling instruments periodically measure a variable but can measure accurately only if prediction equations provide the instrument with an accurate prediction of the next value of this variable. For these instruments, it is appropriate to define reliability as the probability that the error in this prediction does not exceed some limit. In choosing the form and parameters of the prediction equations, it is reasonable to attempt to maximize this reliability.

Assumptions that the prediction equations utilize linear error measurements, are recursive, and provide least-squares smoothing with an exponential weighting function establish a realistic basis for calculating the reliability. The first through the third orders of these equations predict the variable as reliably as possible in the presence of large initial errors in the variable and its velocity, provided that the smoothing interval of the prediction equations is sufficiently short. The dynamic error component of the prediction error of these equations is proportional to a smoothed version of the q th time-derivative of the variable, where q is the order of the prediction equations. The assumption that the measurement errors are uncorrelated and stationary makes it possible to calculate the standard deviation of the random component of the prediction error.

On the assumption that the random component of the prediction error has a normal (or Gaussian) probability distribution, there exists a safety factor which is monotonically related to the reliability. The choice of the smoothing interval of the prediction equations which maximizes this safety factor can be found, which in turn permits the optimum safety factor to be calculated. The ratio of pairs of these optimum safety factors determines which order

* This paper reflects a study supported by the Army's Nike-X Project Office, Redstone Arsenal, Alabama.

of prediction equations gives the greatest reliability in the "worst case" situation in which the first unestimated time-derivative of the variable assumes its largest possible value. Graphs containing the foregoing results make it convenient to examine the tradeoffs between the reliability, the time between measurements, and other parameters. A numerical example is given.

TABLE OF CONTENTS

	<i>page</i>
I. INTRODUCTION.....	2338
II. DETERMINING THE FORM OF THE PREDICTION EQUATIONS (PRIOR TO MAXIMIZING RELIABILITY).....	2341
2.1 <i>First-Order Prediction Equations</i>	2343
2.2 <i>Second-Order Prediction Equations</i>	2343
2.3 <i>Third-Order Prediction Equations</i>	2344
III. COMPARISON OF THE EXPONENTIAL SMOOTHING CRITERION WITH OTHER OPTIMIZATION CRITERIA.....	2345
IV. DYNAMIC ERROR COMPONENT CALCULATIONS.....	2346
V. RANDOM ERROR COMPONENT CALCULATIONS.....	2348
VI. CHOOSING THE SMOOTHING INTERVAL FOR MAXIMUM RELIABILITY.....	2352
VII. OPTIMIZING THE ORDER OF THE PREDICTION EQUATIONS.....	2354
VIII. REDUCTION OF THE SAFETY FACTOR AND THE RELIABILITY DUE TO USING A NON-OPTIMUM SMOOTHING INTERVAL.....	2355
IX. CHOOSING THE TIME T BETWEEN MEASUREMENTS.....	2355
X. CONCLUSION.....	2356
XI. ACKNOWLEDGMENTS.....	2357
APPENDIX A.....	2357
APPENDIX B.....	2360
REFERENCES.....	2362

I. INTRODUCTION

Certain instruments attempt to determine the value of a variable by measuring the difference between it and a prediction of it at regular intervals of time. Examples of these instruments include echo-ranging radars and sonars. These measurements enter a computer which immediately substitutes the measurements into equations which predict the next value of the variable, thereby closing the loop illustrated in Fig. 1. This prediction must be supplied to the instrument because it is assumed that the instrument can measure sufficiently accurately only if it knows approximately where to look for the next value of the variable. Conversely, large errors in this prediction are assumed to blind or otherwise confuse the instrument.

More precisely, it is assumed that positive limits L and L' on the prediction error E are given, where E is defined as the difference between the prediction \hat{x} of the variable and its true value x . It is assumed that if E stays within the interval $-L' \leq E \leq L$, the instrument almost never grossly mismeasures the variable. However, even when E stays within this interval, it is assumed that the instrument slightly mis-

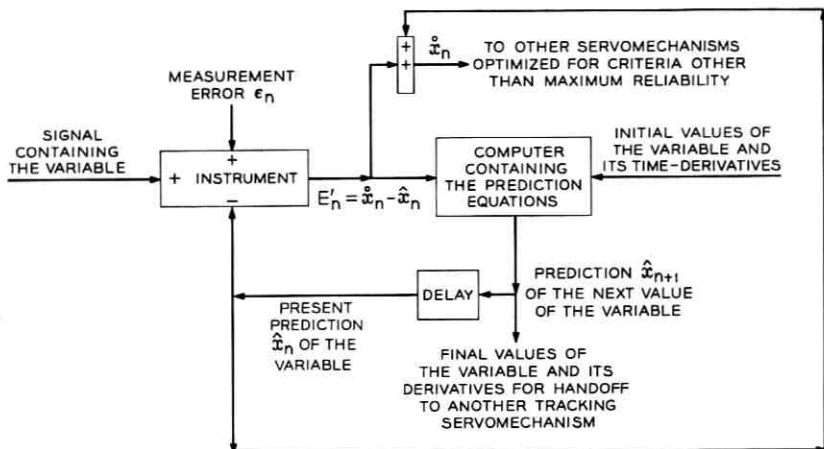


Fig. 1 — Tracking servomechanism.

measures the variable by contributing an additive random error ϵ to the measurements E' of the prediction error. The quantities L , L' , E , \hat{x} , x , ϵ , and E' are illustrated in Fig. 2. Because E' enters the prediction equations, \hat{x} contains random error caused by ϵ . These random errors help make it possible for E to exceed L or to become less than $-L'$. An example of the situation described in this paragraph is given in Ref. 1.

Reliability is defined herein as the probability that E stays within the interval $-L' \leq E \leq L$. Designers usually seem to feel that there is a need to maximize reliability defined in this manner. It must be admitted that different designers always seem to choose slightly different values of L and L' , but it almost always turns out that the results in this paper are not affected significantly by slight differences in these values.

After selecting a relatively simple class of prediction equations which

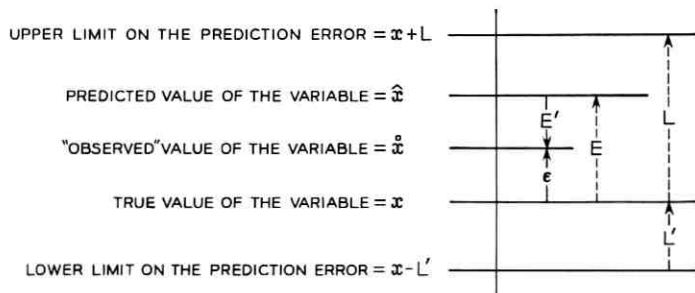


Fig. 2 — Terminology.

assume that the variable is represented in a Taylor's series, this paper attempts to explain how to choose the order and the smoothing interval of these prediction equations. The criterion used in choosing these parameters is that of maximizing reliability at the instant when the variable x is exhibiting that behavior (termed "worst case") which tends to minimize reliability. Using the same class of prediction equations (in a form less convenient for computation than the equations of this paper) and some of the assumptions employed herein, R. G. Brown³ proposed the same objective as that of this paper, but the ill-defined character of Brown's time-series (e.g., the history of an inventory) prevented him from explicitly determining an optimum order and smoothing interval.

Reliability can be maximized by maximizing a quantity which is monotonically related to it. As discussed at the end of this section, it is assumed that the probability that $E < -L'$ can be neglected in comparison with the probability that $E > L$. Assuming also that E is normally distributed [Chapter 7 of Ref. 7], the reliability is a monotonically increasing function of the safety factor

$$\lambda = \frac{L - W}{\sigma_R} \quad (1)$$

where W is defined as the mean of E , so that $W \equiv \bar{E}$. Also, R is defined as $R \equiv E - W$ (so that $\bar{R} \equiv 0$), and σ_R denotes the standard deviation of R .

It is arbitrarily assumed that the prediction equations are linear difference equations and that the mean value $\bar{\epsilon}_n$ of each of the measurement errors is zero. In consequence, the following simplifications are evident: (i) W , called the dynamic error, is caused solely by the inability of the prediction equations to predict more complicated time-histories of the variable than the equations are designed to predict, and (ii) R , called the random error component of the prediction error, is caused solely by the measurement error. In consequence of the first simplification and of the fact that λ in (1) decreases as the dynamic error W increases, the "worst case" behavior of the variable x is that behavior which maximizes W . This maximum value of W is hereafter referred to as reaching its "worst case" value W_c .

The "worst case" value W_c can be calculated with the aid of a theorem (given later) which implies that W is essentially proportional to the q th time-derivative of the variable; $q = 1, 2,$ or 3 is the order of the prediction equations used in this paper. Consequently, it is necessary to know the "worst case" bounds (i.e., the greatest lower bound or the least upper bound) on the first, second, or third time-derivatives (here-

after referred to, respectively, as the velocity v , acceleration a , or jerk j) of the variable x . These "worst case" bounds often can be deduced from constraints imposed by physics or physiology.

It is assumed that the random error component R has a normal (or Gaussian) probability distribution. This assumption is always satisfied if the measurement errors have normal probability distributions. [However, if the measurement errors are not normally distributed, the averaging effect of the prediction equations often makes the random error component distribution approximately normal (as suggested by the Central Limit Theorem)].⁷ Each measurement error ϵ_n is assumed to occur in a random manner independently of every other measurement error, and all ϵ_n are assumed to have equal standard deviations. This standard deviation (or equivalently this rms value) need not be known prior to selecting the parameters of the prediction equations, but it must be known prior to calculating the reliability expected from the selected prediction equations.

Just prior to (1), it was assumed that the probability that $E < -L'$ is negligibly small compared to the probability that $E > L$. This assumption is equivalent to the statement that, measured in units of size equal to σ_R , W_c is much closer to L than to $-L'$. Alternatively, if W_c is much closer to $-L'$ than to L , the right side of (1) is replaced by $(L' + W_c)/\sigma_R$; this replacement does not change the arguments of this paper significantly. In the seemingly uncommon situation in which, after applying the results of Sections VI and VII, it turns out that W_c is not much closer to one of the limits than it is to the other, the methods of this paper do not maximize the reliability even though λ is maximum. The methods used in Section VI also assume that W_c has the same sign as the limit to which it is much closer (i.e., W_c positive if L is much closer, or W_c negative if $-L'$ is much closer); this assumption seems very likely to be satisfied.

II. DETERMINING THE FORM OF THE PREDICTION EQUATIONS (PRIOR TO MAXIMIZING RELIABILITY)

It is pertinent to list some reasonable objectives which the prediction equations should attain.

To make it possible to achieve a level of reliability which is satisfactory at all times, provided that the reliability is large enough at some time, it is sufficient if W_c and σ_R (and therefore also λ) do not vary as time passes. If the standard deviation of the measurement errors and the behavior of the parameter do not vary with time, this constancy of

W_c and σ_R can be achieved by using prediction equations whose form and parameters do not change as time passes. Such prediction equations effectively smooth over a constant interval of preceding measurements. (Prediction equations of this sort may exhibit an initial period, whose duration equals the smoothing interval, during which W and σ_R vary. In practice, reliability usually is not too small during this initial period, because the "worst case" behavior of the variable usually does not occur at this time.)

The tradeoffs involved in maximizing the reliability of the prediction equations should be as obvious as possible. One method of achieving this is to make the prediction equations contain only a single independent parameter, such as the length of the interval over which preceding measurements are smoothed.

The prediction equations should be easy to initialize and should readily provide information by which the track can be handed off to other prediction equations. These objectives suggest that the prediction equations should explicitly estimate the time-derivatives of the variable.

The prediction equations should occupy as little storage space in the computer as possible. This objective suggests the use of recursive equations, which require storing only the most recent values of all quantities appearing in these equations. The need for keeping computation time to a minimum implies that the prediction equations should be few and easy to calculate.

It appears that the foregoing objectives can be attained by arbitrarily assuming that the variable is represented by a Taylor's series whose coefficients satisfy the "exponential smoothing" criterion. This criterion makes the expected (or mean) values of the sum of the exponentially weighted squares of the measured prediction error as small as possible. (A source of confusion may exist because the definition of the term "exponential smoothing" in this paper is more specific than another apparently obsolete definition of this term, which a few people have used to denote *any* set of recursive equations having constant coefficients.) Stating the exponential smoothing assumption algebraically, the quantity

$$\sum_{i=0}^{\infty} E'_{n-i}{}^2 K^i$$

[where E' denotes the measurements of prediction error, where

$$K^i = \left(\frac{N-1}{N+1} \right)^i$$

is the exponential weighting coefficient (with $N \geq 1$ being called the

smoothing interval), and n denotes the time of the latest prediction] represents the sum of the exponentially weighted squares of E' . This sum is to be minimized by differentiating it separately with respect to each of the coefficients of the Taylor's series representing the variable, setting the differentials equal to zero, and solving the resulting set of equations simultaneously to obtain the optimum values of the Taylor's series coefficients. These steps are explained by Levine.²

Recursive prediction equations which satisfy this requirement can be obtained from Levine² by setting his weighting coefficient ω_i equal to K^{-i} , taking the limit as n approaches infinity, and substituting the results into his equations (15) through (17) or else (54) and (56). Similar prediction equations satisfying the exponential smoothing criterion appear in Refs. 3 or 11. Using Levine's notation, the prediction equations for first, second, and third-order smoothing (corresponding to a Taylor's series containing one, two, or three coefficients) can be rearranged to minimize the amount of data-storage and computation time required. The rearranged prediction equations are given in the next three subsections.

2.1 First-Order Prediction Equations

$$\hat{x}_{n+1} = \hat{x}_n + \alpha_1 E_n' \quad (2)$$

$$E_n' = \hat{x}_n - \hat{x}_n \quad (3)$$

$$\alpha_1 = 1 - K \quad (4)$$

$$K = \frac{N - 1}{N + 1}, \quad N \geq 1, \quad (5)$$

where N is a constant, where \hat{x} is the predicted value of the parameter, and where \hat{x} denotes the "observed" value of the variable, which can be calculated by adding \hat{x} to the measured prediction error E' . (The subscript 1 in α_1 is different from Levine's subscript in that now the subscript denotes the order of the smoothing equations instead of denoting the time of the measurement being processed by Levine's equations.)

2.2 Second-Order Prediction Equations

$$\hat{u}_{n+1} = \hat{u}_n + \beta_2 E_n' \quad (6)$$

$$\hat{x}_{n+1} = \hat{x}_n + \hat{u}_{n+1} + \alpha_2 E_n' \quad (7)$$

$$\beta_2 = (1 - K)^2 \quad (8)$$

$$\alpha_2 = 1 - K^2 \quad (9)$$

$$\hat{u}_n = \hat{v}_n T, \quad (10)$$

where K and E' are defined in (3) and (5), where \hat{v} is the predicted velocity, and where T is the constant interval of time between the measurements of E' . Equation (6) is supposed to be computed immediately before (7) is computed. As a result, \hat{u} and \hat{x} can be stored in a single word of computer memory each. A similar statement can be made about \hat{s} , \hat{u} , and \hat{x} in the next three equations.

2.3 Third-Order Prediction Equations

$$\hat{s}_{n+1} = \hat{s}_n + \gamma_3 E_n' \quad (11)$$

$$\hat{u}_{n+1} = \hat{u}_n + 2\hat{s}_{n+1} + \beta_3 E_n' \quad (12)$$

$$\hat{x}_{n+1} = \hat{x}_n + \hat{u}_{n+1} - \hat{s}_{n+1} + \alpha_3 E_n' \quad (13)$$

$$\gamma_3 = \frac{1}{2}(1 - K)^3 \quad (14)$$

$$\beta_3 = \frac{3}{2}(1 - K^2)(1 - K) \quad (15)$$

$$\alpha_3 = 1 - K^3 \quad (16)$$

$$\hat{s}_n = \frac{1}{2}\hat{a}_n T^2, \quad (17)$$

where \hat{u} , K , and E' are defined in (10), (5), and (3), and where \hat{a} is the predicted acceleration.

For convenience, a graph of the coefficients calculated in (4), (8), and (9), and (14) through (16) is presented in Fig. 3. Equations (2), (6), and (7), or (11) through (13) are assumed to be initialized by using *a priori* estimates of the variable and (if required) its velocity and acceleration to calculate the initial value of \hat{x} , \hat{u} , and \hat{s} with the aid of (10) and (17). Conversely, (2), (6), and (7), or (11) through (13) together with (10) and (17) provide estimates of \hat{x} , \hat{v} , and \hat{a} which can be used to transfer (or hand off) the track to another set of prediction equations by providing initial predictions for this new set. [Of course, \hat{a} is available only from (11) and \hat{v} from (12) or (6).] These predictions can be extended to any future time merely by using a Taylor's series. For example,

$$\hat{x}_{n+(t/T)} = \hat{x}_n + \left(\frac{t}{T}\right)\hat{u}_n + \left(\frac{t}{T}\right)^2\hat{s}_n. \quad (18)$$

Alternatively, if it is desired to predict \hat{x} , \hat{u} , and \hat{s} at any integral multiple of T in the future, it suffices merely to calculate (2), (6), and (7), or (11) through (13) in appropriate number of times but with the coefficients α , β , γ set equal to zero.

The single independent parameter N in (5) is identified as the smooth-

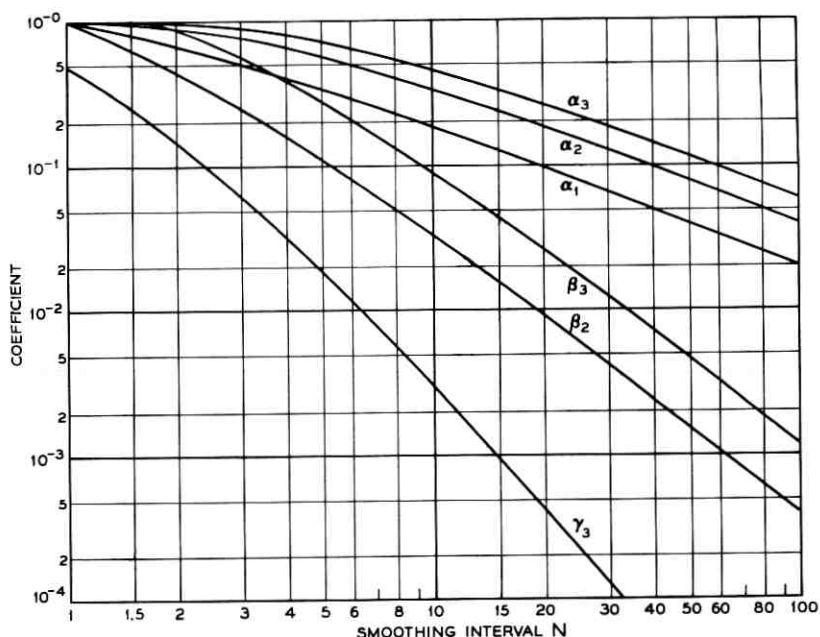


Fig. 3 — Coefficients of exponential smoothing equations.

ing interval because (2), (6), and (7), and (11) through (13) have nearly the same amounts of dynamic and random error components as polynomial smoothing equations (Ref. 4) designed to smooth over the last N samples. Additionally, K is approximated closely by $\exp(-2/N)$ for all $N \geq 1.5$, thereby making N be approximately the exponential decay constant of \sqrt{K} (which in Levine's paper² is set equal to the reciprocal of the standard deviation).

III. COMPARISON OF THE EXPONENTIAL SMOOTHING CRITERION WITH OTHER OPTIMIZATION CRITERIA

A different set of prediction equations designed specifically for maximum reliability in acquiring the track, under the assumptions that the standard deviations of the errors in the initial values of \hat{x} and \hat{u} (supplied from some external source of information) are much larger than L and that the standard deviation of the measurement error is much smaller than L , would predict the second value of \hat{x} by adding the *a priori* estimate of \hat{u} to the first value of \hat{x} . This procedure permits the initial velocity to be in error by as much as L/T without E ever exceeding L .

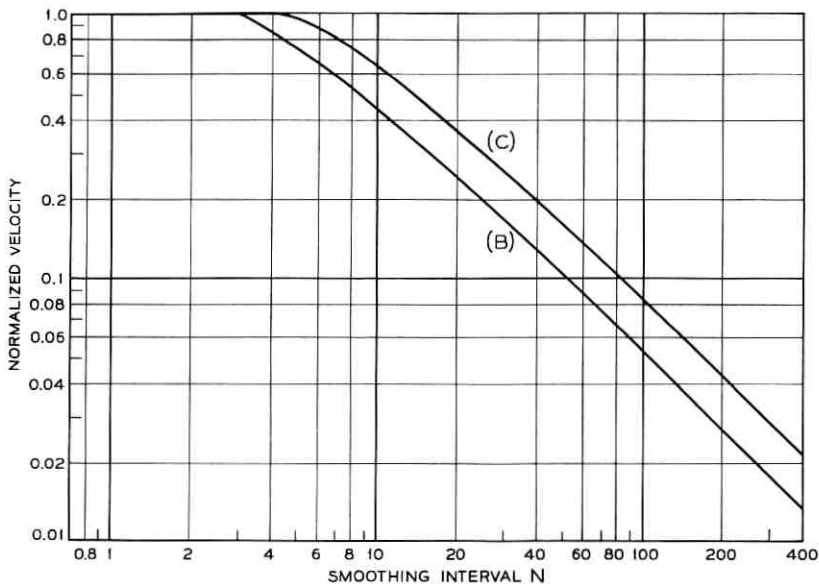


Fig. 4 — Maximum permissible initial error in velocity for (6) and (7) and (11) through (13) divided by L/T ; (b) second order; (c) third order.

A comparison between the effectiveness of (6) and (7) or (11) through (13) and the above prediction equations is shown in Fig. 4, which shows the interval of permissible initial velocities of (6) and (7) or (11) through (13) divided by L/T . (Simulations reveal that the region of permissible initial positions and velocities approximates a horizontal rectangle on the prediction error phase plane, whose position error axis is intersected by the sides of the rectangle at $\pm L$.) Fig. 4 indicates that if N is small enough, (6) and (7) and (11) through (13) perform as reliably as the above set of prediction equations because the ratios are unity.

Evaluating the extent to which (6) and (7) satisfy another optimization criterion, Benedict and Bordner⁵ state that critically damped equations equivalent to (6) and (7) give virtually the same accuracy as the set of recursive second-order smoothing equations which minimize a weighted sum of the random and the dynamic errors.

IV. DYNAMIC ERROR COMPONENT CALCULATIONS

The following theorem, which is illustrated in Fig. 5, provides a basis for calculating the dynamic error.

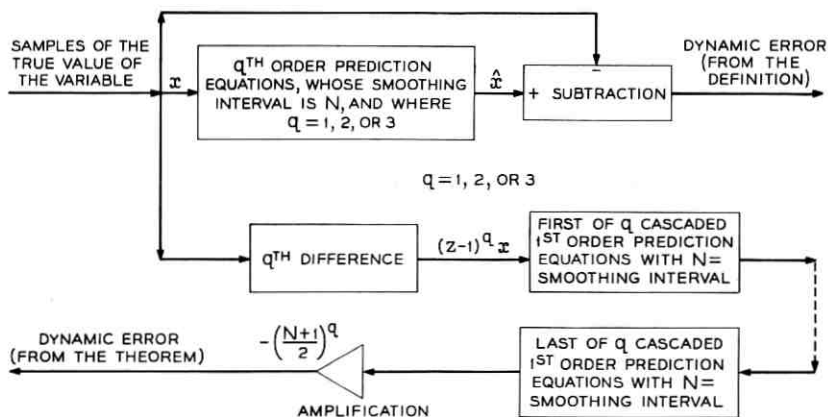


Fig. 5 — Illustration of the theorem.

Theorem: Let $q = 1, 2, \text{ or } 3$ denote the order of the prediction equations, and let the q th difference of the variable x be defined as $(z - 1)^q x$, where z is the advance operator (which denotes taking the next sample of the time series on which it is operating, as in the example $(z - 1)x_n = x_{n+1} - x_n$). Then, the dynamic error component equals the product of $-[(N + 1)/2]^q$ and the final output of q identical first-order prediction equations having the same smoothing interval as the original prediction equations and forming a series connection in which the output of one equation is the input to the next equation, with the initial input being the q th difference of the variable. A proof of the theorem is given in Appendix A.

A natural and intuitively satisfying interpretation of this theorem is obtained from the fact that the q th difference of the variable closely approximates the q th derivative of the variable multiplied by T^q . The q cascaded first-order filters, acting like simple low-pass resistance-capacitance filters, tend to attenuate all rapid fluctuations in the q th derivative and to delay the change in the dynamic error caused by a nonzero $q + 1$ th derivative by $q(N + 1)/2$ samples. Thus, to estimate the worst case dynamic error of the q th order prediction equations, it suffices simply to multiply the maximum value of the variable's q th derivative (averaged over a total smoothing interval of qN samples) by $-T^q[(N + 1)/2]^q$, where $q = 1, 2, \text{ or } 3$.

To complete this treatment of dynamic error, the dynamic error component of \hat{u} or \hat{s} can be defined as \hat{u} or \hat{s} minus the first difference or one half of the second difference of the variable. [In computing the dynamic error of u for the third-order prediction equations, it was necessary to

define the true velocity operator as $\frac{1}{2}(z - 1/z)$, instead of $(z - 1)$.] It is possible to use the methods of Appendix A to prove theorems about these dynamic errors; these theorems can be interpreted in nearly the same way as in the previous paragraph. These results, along with those of the previous paragraph, are presented in Fig. 6. The dynamic error coefficients are to be multiplied by $-vT$ in the first-order case, $-aT^2$ in the second-order case, and $-jT^3$ in the third-order case. The delays through the various connections of first-order filters (each of which introduces a delay of $NT/2$) are contained in Table I. The dynamic error coefficients plotted in Fig. 6 are listed in Table II.

V. RANDOM ERROR COMPONENT CALCULATIONS

Fig. 7 shows the coefficients of the standard deviation (or rms value) of the random error component of \hat{x} , \hat{u} , and \hat{s} , and Table III lists the asymptotic behaviors of these coefficients for large values of N . (Fig. 7

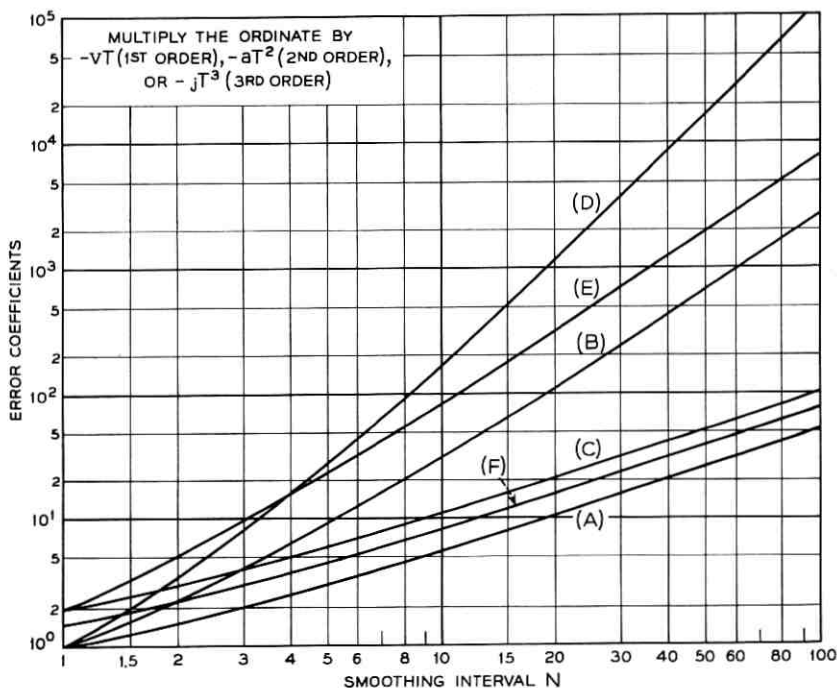


Fig. 6 — Dynamic error — first order, (a) \hat{x} ; second order, (b) \hat{x} , (c) \hat{u} ; third order, (d) \hat{x} , (e) \hat{u} , (f) \hat{s} .

TABLE I — DELAYS OF DYNAMIC ERROR FOR q TH ORDER PREDICTION EQUATIONS

	$q = 1$	2	3
\hat{x}	1	2	3
\hat{u}	—	$\frac{3}{2}$	$\frac{8}{3}$
\hat{s}	—	—	2

Multiply by $NT/2$

was obtained by numerically calculating the square root of the sums of squares of the quantities estimated by (2), (6), and (7) and (11) through (13) in response to $\hat{x}_0 = 1, \hat{x}_i = 0, i > 0$ [as in Ref. 5: (10) and Appendix I]. The actual standard deviations can be obtained by multiplying the coefficients shown in Fig. 7 and Table III by the standard deviation σ_ϵ of the measurement error.

Fig. 8 gives the correlation coefficients ρ (defined as the covariance divided by the product of the standard deviations of the quantities appearing in the covariance calculations), and Table IV lists their asymptotic values as N becomes very large. These correlation coefficients together with the standard deviation coefficients can be used to calculate the standard deviation of a prediction extended to any future time, because taking the expected value of the square of both sides of (18) gives

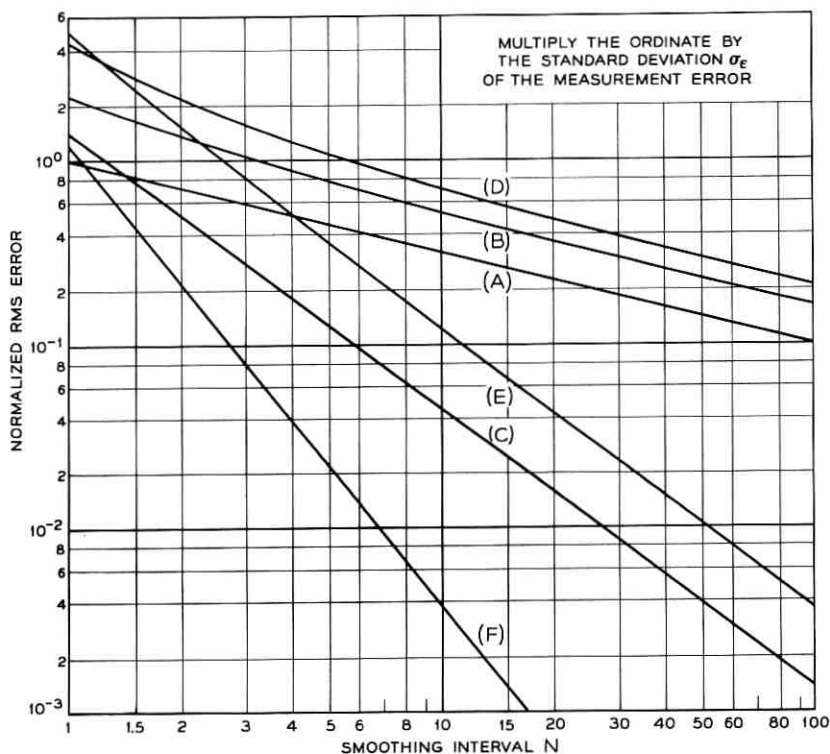
$$\begin{aligned} \sigma_{\hat{x}_{n+(t/T)}} &= [(\hat{x}_{n+(t/T)})^2]^{\frac{1}{2}} \\ &= \left\{ \sigma_{\hat{x}}^2 + \frac{t}{T} 2\sigma_{\hat{x}\hat{u}}\rho_{\hat{x}\hat{u}} + \left(\frac{t}{T}\right)^2 (\sigma_{\hat{u}}^2 + 2\sigma_{\hat{x}\hat{s}}\rho_{\hat{x}\hat{s}}) \right. \\ &\quad \left. + \left(\frac{t}{T}\right)^3 2\sigma_{\hat{u}\hat{s}}\rho_{\hat{u}\hat{s}} + \left(\frac{t}{T}\right)^4 \sigma_{\hat{s}}^2 \right\}^{\frac{1}{2}} \end{aligned} \quad (19)$$

TABLE II — DYNAMIC ERROR COEFFICIENTS

	$q = 1$	2	3
\hat{x}	$D_1 = \frac{N+1}{2}$	$D_2 = \frac{(N+1)^2}{4}$	$D_3 = \frac{(N+1)^3}{8}$
\hat{u}	—	$N+1$	$\frac{(N+1)^2}{4} \left[3 - \frac{2}{N+1} + \frac{2}{(N+1)^2} \right]$
\hat{s}	—	—	$\frac{3(N+1)}{4}$

TABLE III — ASYMPTOTIC BEHAVIORS OF THE STANDARD DEVIATION COEFFICIENTS

	$q = 1$	2	3
\hat{x}	$S_1 = \frac{1}{\sqrt{N}}$	$S_2 = \frac{\sqrt{2.5}}{\sqrt{N}} = \frac{1.58}{\sqrt{N}}$	$S_3 = \frac{\sqrt{4.125}}{\sqrt{N}} = \frac{2.03}{\sqrt{N}}$
\hat{u}	—	$\frac{\sqrt{2}}{N^{3/2}} = \frac{1.414}{N^{3/2}}$	$\frac{\sqrt{14}}{N^{3/2}} = \frac{3.74}{N^{3/2}}$
\hat{s}	—	—	$\frac{\sqrt{1.5}}{N^{5/2}} = \frac{1.224}{N^{5/2}}$

Fig. 7 — Standard deviations — first order, (a) \hat{x} ; second order, (b) \hat{x} , (c) \hat{u} ; third order, (d) \hat{x} , (e) \hat{u} , (f) \hat{s} .

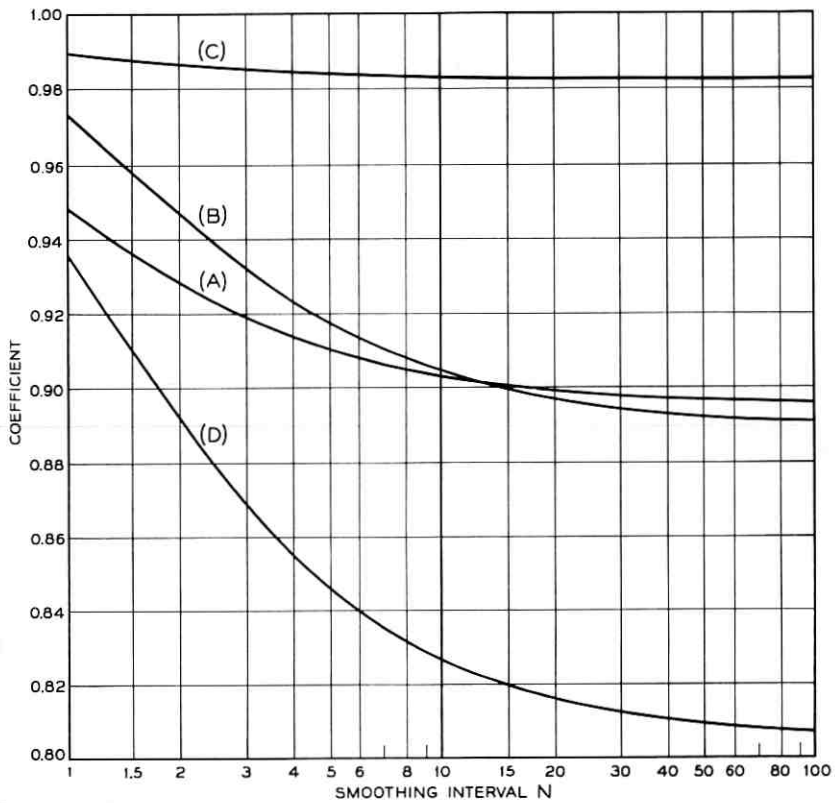


Fig. 8 — Correlation coefficients — second order, (a) \hat{x}, \hat{u} ; third order, (b) \hat{x}, \hat{u} (c) \hat{u}, \hat{s} , (d) \hat{x}, \hat{s} .

TABLE IV — ASYMPTOTIC VALUES OF THE CORRELATION COEFFICIENTS ρ

	$q = 2$	3
\hat{x}, \hat{u}	$\frac{2}{\sqrt{5}} = 0.895$	$\frac{27}{2\sqrt{231}} = 0.889$
\hat{x}, \hat{s}	—	$\frac{8}{\sqrt{99}} = 0.804$
\hat{u}, \hat{s}	—	$\frac{9}{2\sqrt{21}} = 0.983$

VI. CHOOSING THE SMOOTHING INTERVAL FOR MAXIMUM RELIABILITY

Substituting the expressions for W and σ_R shown as graphs on Figs. 6 and 7 into (1) to obtain the "worst case" safety factor produces, with v_c , a_c , and j_c denoting the "worst-case" velocity, acceleration, and jerk,

$$\lambda_1 = \frac{\left| \frac{L}{v_c T} \right| - D_1}{S_1} \cdot \frac{|v_c T|}{\sigma_\epsilon},$$

or

$$\lambda_2 = \frac{\left| \frac{L}{a_c T^2} \right| - D_2}{S_2} \cdot \frac{|a_c T^2|}{\sigma_\epsilon}, \quad (20)$$

or

$$\lambda_3 = \frac{\left| \frac{L}{j_c T^3} \right| - D_3}{S_3} \cdot \frac{|j_c T^3|}{\sigma_\epsilon},$$

where D and S , respectively, denote the dynamic and random error coefficients of the predicted value of the variable, and where the subscripts denote the order q of the prediction equations. To carry out the maximization by choosing N to maximize λ , it suffices to consider only the first factor in each of the three expressions in (20), because only D and S are functions of N . Although this statement is always true, λ does not completely determine the reliability if the dynamic error remains at its "worst case" value W_c for more than approximately N samples, because E would have two or more independent opportunities to exceed L . However, it is believed that this additional source of unreliability can be disregarded when choosing N to maximize the reliability, because as N varies the reliability depends on λ much more strongly than it depends on this additional source.

The first factors in (20) can be written as

$$\lambda_1' = \frac{P_1 - D_1}{S_1}, \quad \text{or } \lambda_2' = \frac{P_2 - D_2}{S_2}, \quad \text{or } \lambda_3' = \frac{P_3 - D_3}{S_3}, \quad (21)$$

where

$$P_1 = \left| \frac{L}{v_c T} \right|, \quad \text{or } P_2 = \left| \frac{L}{a_c T^2} \right|, \quad \text{or } P_3 = \left| \frac{L}{j_c T^3} \right|. \quad (22)$$

The values of N which maximize λ' are shown in Fig. 9 as a function of

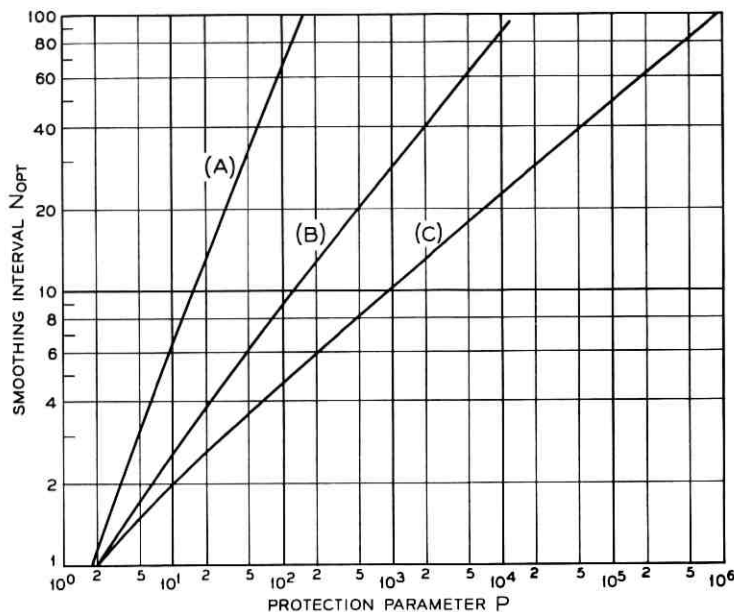


Fig. 9 — Optimum smoothing intervals versus protection parameters — (a) first order, (b) second order, (c) third order.

P , and the corresponding maximum values of λ' are shown in Fig. 10. Asymptotic formulas for these optimum values of N and λ' for large values of P can be obtained with the aid of Tables II and III as

$$\begin{aligned} N_{1_{\text{opt}}} &= 2/3 P_1 \\ N_{2_{\text{opt}}} &= 0.903 \sqrt{P_2} \end{aligned} \quad (23)$$

$$\begin{aligned} N_{3_{\text{opt}}} &= 1.045 \sqrt[3]{P_3} \\ \lambda_{1_{\text{opt}}} &' = 0.544 (P_1)^{3/2} \\ \lambda_{2_{\text{opt}}} &' = 0.478 (P_2)^{5/4} \end{aligned} \quad (24)$$

$$\lambda_{3_{\text{opt}}} &' = 0.432 (P_3)^{7/6}.$$

These formulas provide very close approximations to Figs. 9 and 10 for $N_{\text{opt}} > 2.5$ approximately.

An important relationship obtained by substituting (22) into (23) is that, for all values of N greater than approximately 2.5, the effective smoothing time $N_{\text{opt}} T$ equals a constant (which depends only on L and the "worst case" value of the q th time derivative of the variable).

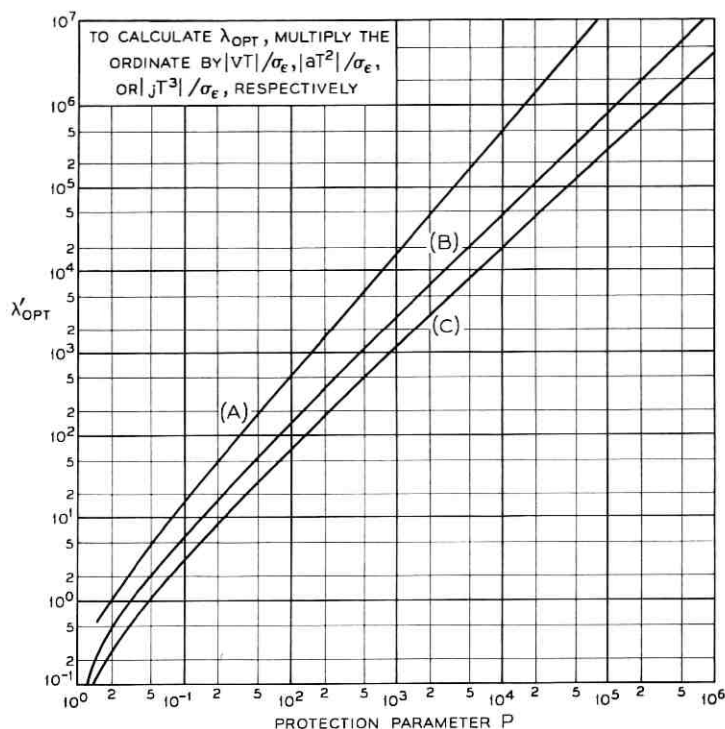


Fig. 10 — Optimum safety factors versus protection parameters — (a) first order, (b) second order, (c) third order.

A significant conclusion may be drawn from the fact that the function λ' in (21) is independent of the standard deviation of the measurement error σ_ϵ . Consequently, a single value of N (i.e., N_{opt}) simultaneously maximizes the reliability for all σ_ϵ .

VII. OPTIMIZING THE ORDER OF THE PREDICTION EQUATIONS

The order of the prediction equations can always be chosen to maximize the reliability by using the order which gives the largest value of the safety factor λ defined by (20) and shown in Fig. 10. A more general comparison between λ_1 , λ_2 , and λ_3 can be obtained for sufficiently large values of optimum smoothing intervals (i.e., $N_{\text{opt}} > 2.5$ approximately) by using the asymptotic formulas (24) to obtain the ratios, on the assumption that $L' = L$,

$$\frac{\lambda_1}{\lambda_2} = \frac{1.14 |L|^{1/4} |a_c|^{1/4}}{|v_c|^{1/2}}, \quad (25)$$

where v_c and a_c are the "worst case" velocity and acceleration, and

$$\frac{\lambda_2}{\lambda_3} = \frac{1.11 |L|^{1/12} |j_c|^{1/6}}{|a_c|^{1/4}}, \quad (26)$$

where j_c is the "worst case" jerk.

If (25) is greater than unity, then the optimum first-order prediction equations [which are optimum in the sense that they use the values N_{opt} given asymptotically by (23)] track more reliably than the optimum second-order prediction equations. Likewise, if (26) is greater than unity, then the optimum second-order prediction equations track more reliably than the optimum third-order prediction equations. The converse holds also.

It is significant that, if the time T between measurements is not so large that N is too small for (25) and (26) to hold, these equations indicate that the optimum order q_{opt} of the smoothing equations is optimum for all T (and σ_e).

Examples illustrating these methods are given in Appendix B.

VIII. REDUCTION OF THE SAFETY FACTOR AND THE RELIABILITY DUE TO USING A NON-OPTIMUM SMOOTHING INTERVAL

Ratios of the actual safety factor obtained by using an arbitrary value of N to the optimum safety factor obtained by using N_{opt} from (23) can be obtained for large values of P by using the values of λ_{opt} from (24) and using Tables II and III in conjunction with (21). A graph of these ratios is given in Fig. 11 as a function of N/N_{opt} . This graph indicates that it is worse to use a value of N larger than N_{opt} than it is to use a value proportionately smaller than N_{opt} .

The reduction in reliability caused by accepting a safety factor smaller than optimum can be calculated from tables of the normal distribution function [pp. 966-72 of Ref. 8]. For example, if $\lambda = 3$, the unreliability (i.e., unity minus the reliability) is only 0.135 percent, but the unreliability rises to 0.35, 0.82, 1.79, or 6.68 percent if λ decreases by 10, 20, 30, or 50 percent, respectively.

IX. CHOOSING THE TIME T BETWEEN MEASUREMENTS

The assumptions of this paper are not sufficient to determine an optimum value of T , but the methods of this paper do permit describ-

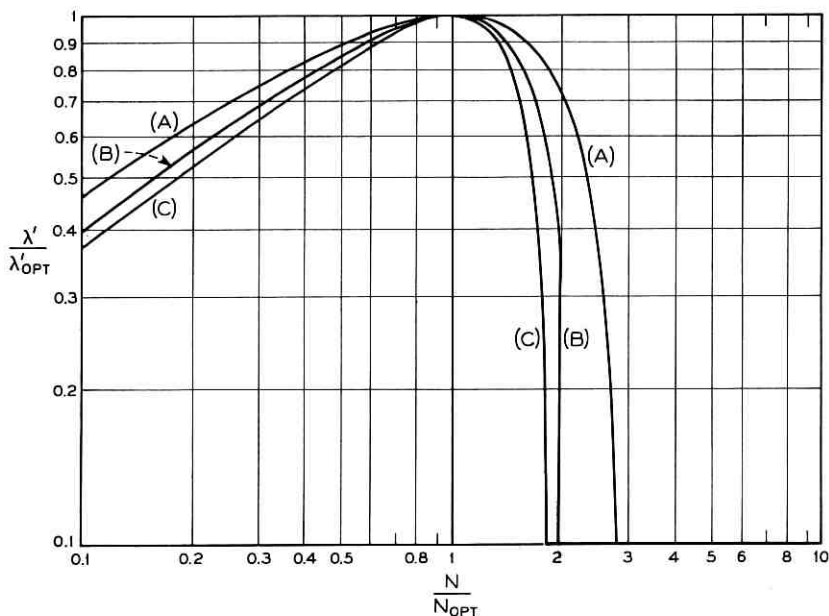


Fig. 11 — Relative safety factor versus the ratio of actual decay constant to the optimum decay constant.

ing the effects of various choices of T , if it is assumed that the optimum value of N is used and that N_{opt} is large enough (i.e., $N_{opt} > 2.5$ approximately) so that $N_{opt} T$ equals a constant.

One effect is that as T decreases (and N_{opt} increases), the maximum initial velocity error (for which E barely keeps from exceeding L) increases because it equals L/T until the appropriate break point shown in Fig. 4 is reached, beyond which point the maximum tolerable initial velocity error remains constant as T decreases.

Another effect is that as T increases, the standard deviations of \hat{x} , $\hat{\theta}$, and \hat{a} [the latter two quantities being related to \hat{u} and \hat{s} by (10) and (17)] increase proportionately to \sqrt{T} . In consequence, an increase of T causes the safety factor to decrease as $1/\sqrt{T}$, because σ_R is proportional to \sqrt{T} , and because W is independent of T (due to the assumption that N_{opt} is large).

X. CONCLUSION

Except for the previous section, the objective of this paper has been the combining of limits on the maximum tolerable prediction error L

and the "worst case" time-derivatives v_c , a_c , or j_c with a time T between measurements, for the purpose of determining the optimal order q_{opt} and smoothing interval N_{opt} of the prediction equations. The results of this paper make it possible to perform parameter-variation studies in which q_{opt} and N_{opt} are assumed to be used, and in which tradeoffs between the following quantities are examined: (i) L , (ii) T , (iii) the standard deviation σ_ϵ of the measurement error, (iv) the reliability (defined as the probability that $-L' \leq E \leq L$ at the time of the "worst case" value v_c , a_c , or j_c), and (v) the "worst case" value v_c , a_c , or j_c .

XI. ACKNOWLEDGMENTS

A. R. Eckler provided sponsorship and guidance. R. T. Piotrowski and Miss A. T. Seery programmed the digital computer to prepare the graphs, which are plotted directly on microfilm with the aid of J. F. Kaiser's subroutines. A. J. Pellecchia (of Western Electric) and G. A. Jones (now of Pacific Gas and Electric Company) assisted. W. L. Nelson, F. A. Russell, and R. A. McDonald suggested useful revisions. The author is grateful to all of the above.

APPENDIX A

Proof of Dynamic Error Theorem

The first-order prediction (2) can be written in the operator notation as

$$\hat{x} = \frac{\alpha_1}{z + \alpha_1 - 1} \dot{x} \quad (27)$$

where z is the advance operator for which $z(x_n) = x_{n+1}$. Calculating the dynamic error component of the prediction error E is accomplished by subtracting the true value x of the variable from \hat{x} and then by setting the observed value \hat{x} of the variable equal to its true value x , because setting the (random) measurement error $\epsilon = \hat{x} - x$ equal to zero removes the random error component from E . Thus, the dynamic error component of E equals

$$\begin{aligned} \hat{x} - x &= \frac{\alpha_1}{z + \alpha_1 - 1} \dot{x} - x = \left[\frac{\alpha_1}{z + \alpha_1 - 1} - 1 \right] x \\ &= \frac{1}{\alpha_1} \cdot \left(\frac{\alpha_1}{z + \alpha_1 - 1} \right) [-(z - 1)x]. \end{aligned} \quad (28)$$

According to (4) and (5), $1/\alpha_1 = (N + 1)/2$, so that the dynamic error equals

$$\frac{N + 1}{2} \left(\frac{\alpha_1}{z + \alpha_1 - 1} \right) [-(z - 1)x]. \quad (29)$$

Equation (27) shows that (29) denotes passing the negative of the first difference of x through a first-order prediction equation which is identical to the first-order prediction equation being considered.

The second-order equations (6) and (7) can be written in operator notation and simultaneously solved for \hat{x} to give

$$\hat{x} = \frac{[\beta_2 z + \alpha_2(z - 1)]\hat{x}}{\beta_2 z + \alpha_2(z - 1) + (z - 1)^2}. \quad (30)$$

Computing the dynamic error by subtracting x from both sides and setting $\hat{x} = x$ as before makes (30) become

$$\hat{x} - x = \frac{-(z - 1)^2 x}{\beta_2 z + \alpha_2(z - 1) + (z - 1)^2}. \quad (31)$$

Using (8) and (9) and then (4) and (5) gives

$$\begin{aligned} \hat{x} - x &= \frac{-(z - 1)^2 x}{(z - K)^2} \\ &= \frac{[-(z - 1)^2 x]}{\alpha_1^2} \left[\frac{\alpha_1}{z - 1 + \alpha_1} \right]^2 \\ &= \left(\frac{N + 1}{2} \right)^2 \left(\frac{\alpha_1}{z - 1 + \alpha_1} \right)^2 [-(z - 1)^2 x]. \end{aligned} \quad (32)$$

Equation (32) represents two cascaded first-order prediction equations, whose smoothing interval N is the same as the smoothing interval of the original second-order prediction equations, operating on the negative of the second difference of the variable, as stated in the theorem.

The third-order prediction equations (11) through (13) can be written in operator notation and solved simultaneously to give

$$\hat{x} = \left[\frac{\alpha_3(z - 1)^2 + \beta_3 z(z - 1) + \gamma_3 z(z + 1)}{(z - 1)^3 + \alpha_3(z - 1)^2 + \beta_3 z(z - 1) + \gamma_3 z(z + 1)} \right] \hat{x}, \quad (33)$$

so that the dynamic error equals

$$\hat{x} - x = \frac{[-(z - 1)^3 x]}{(z - 1)^3 + \alpha_3(z - 1)^2 + \beta_3 z(z - 1) + \gamma_3 z(z + 1)}. \quad (34)$$

Using (14) through (16) gives

$$\begin{aligned} \dots x &= \frac{[-(z-1)^3 x]}{(z-K)^3} \\ &= \left(\frac{N+1}{2}\right)^3 \left(\frac{\alpha_1}{z-1+\alpha_1}\right)^3 [-(z-1)^3 x], \end{aligned} \quad (35)$$

where the smoothing interval N of the three cascaded first-order prediction equations is the same smoothing interval as that which appears in the original third-order prediction equations, as stated in the theorem.*

The abridgment of this theorem obtained by omitting the q cascaded first-order prediction equations {i.e., the dynamic error is approximated by multiplying the *unsmoothed* q th difference by $-[(N+1)/2]^q$ } has approximated the dynamic error very accurately in simulations of the slowdown of ballistic devices re-entering the earth's atmosphere.

As an application of the theorem, the maximum dynamic error component for the second-order prediction equations (6) and (7) for an initial error u_0' in \hat{u} can be approximated by performing the steps listed in the theorem. Thus, the second difference of $x = u_0' n$ for $n \geq 0$ and $x = 0$ for $x < 0$ is an impulse of height u_0' . This impulse enters two cascaded first-order prediction equations, whose combined impulse response {obtained by convolving two exponentials together and multiplying by $[(N+1)/2]^2$ } is approximately equal to $u_0' n \exp(-2n/N)$. The maximum value of this impulse response occurs at $n = N/2$, so that the maximum value of the dynamic error component of the prediction error E equals approximately $u_0'(N/2e)$.

Similarly, it is possible to calculate the dynamic error due to an initial error s_0' in \hat{s} in the third-order equations (11) through (13) since the third difference of $x = s_0' n^2$ for $n \geq 0$ and $x = 0$ for $n < 0$ is two successive impulses of height equal s_0' . These two impulses can be regarded as having the effect of a single impulse of height $2s_0'$ if the smoothing interval N is sufficiently large compared to two. The impulse response of three cascaded first-order prediction equations is

$$(n^2/2) \exp[-2(n-1)/(N+1)],$$

* For p th order exponential smoothing (as opposed to prediction) equations in which p is any integer, it is possible to prove that the dynamic error defined as $\bar{x} - x$ is equal to $[(N-1)/2]^p$ times the negative of the p th difference {defined as $[1 - (1/z)]^p$ } and passed through p first-order exponential smoothing equations of the same smoothing interval. (This proof starts with (7) of Ref. 10 and uses the lemma

$$(z-1)S\hat{x} = \frac{\alpha_1 z}{1-\alpha_1} [e - S]\hat{x},$$

where S is an operator denoting a first-order exponential smoothing equation.)

after multiplying by the $[(N + 1)/2]^3$ factor. Thus, the dynamic error component of the prediction error E can be approximated by $s_0' n^2 \exp [-2(n - 1)/(N + 1)]$. The maximum of this approximation occurs at $n = N + 1$, so that the maximum dynamic error equals approximately $s_0' (N + 1)^2 \exp [-2N/(N + 1)]$. (This approximation is never more than 12 percent larger than the true value of the dynamic error for any $N \geq 7$. The expression $s_0' (N + 1)^2 \exp (-2)$ provides an approximation which is always smaller than the true value of the dynamic error.)

To conclude these examples of the use of the theorem, the dynamic error components of the prediction error E for the third-order prediction equations (11) through (13) caused by an initial error u_0' or x_0' in \hat{u} or \hat{x} can be approximated respectively by the first or second derivative with respect to n of $\frac{1}{2}n^2 \exp (-2n/N)$ multiplied by u_0' or x_0' .

APPENDIX B

Examples

It is necessary to illustrate how the assumptions and results discussed in the text apply to an actual instrument. Thus, echo-ranging radars have measurement errors which occur independently of the measurement error on any other pulse. The reason for this independence is that the interval T between measurements is always much larger than the reciprocal of the bandwidth of the instrument, because the time for the echo to return from the target always greatly exceeds the duration of the echo.

In monopulse radars measuring the range and angles of a target, thermal noise originating in the receiver often results in σ_e being equal to the product of the following two quantities: (i) a large fraction of the pulse-width or beamwidth, and (ii) the reciprocal of the square root of the ratio of the peak signal power to the average noise power [Chapter 10 of Ref. 6]. If quantization errors are present because the instrument measures the prediction error digitally, σ_e can usually be calculated by taking the square root of the sum of squares of the standard deviations of the thermal and quantization errors.

At the end of Section VI of the text, it is stated that a single value of the smoothing interval N simultaneously maximizes the reliability for all σ_e . In radar terminology, this value of N (i.e., N_{opt}) is optimum for any ratio of signal power to noise power. Likewise, Section VII concludes with a result which can be interpreted as stating that the optimum value q_{opt} of the order of the prediction equations is optimum for any data rate and any ratio of signal power to noise power.

The use of some of the equations and graphs presented previously is exemplified by the following: A radar emitting a 100 foot-wide pulse every 0.27 seconds is tracking the range of a descending spaceship whose maximum velocity is $-35,000$ ft/sec, whose maximum acceleration is 320 ft/(sec)², and whose maximum jerk is simulated to be 8.3 ft/(sec)³. (These maxima occur at different times during the re-entry of the spaceship into the atmosphere.) It is discovered that the radar can measure the prediction error sufficiently accurately only if the prediction error $|E|$ is smaller than half the pulsewidth, so that $L = 50$ feet.* The standard deviation σ_ϵ of the measurement error ϵ is found to be approximately equal to $L/\sqrt{S/N}$, where S/N denotes the ratio of the instantaneously maximum signal power to the average noise power in the output of the radar's IF strip.

Equation (25) reveals that second-order smoothing is more reliable than first-order smoothing if $L < 2.26 \times 10^6$ feet, and (26) reveals that third-order smoothing is more reliable than second if $L < 4.36 \times 10^5$ feet. In consequence of assuming L to be only 50 feet, third-order smoothing should be used.

Using (22) to calculate P_3 gives $P_3 = 306$. Fig. 9 indicates that $N_{\text{opt}} = 7$, and Fig. 10 indicates that $\lambda' = 275$. Computing λ by multiplying λ' by $|j_c T^3|/\sigma_\epsilon$ gives $\lambda = 0.9 \sqrt{S/N}$. If $S/N = 16 = 12$ db, $\lambda = 3.6$ and the "worst-case" reliability is 99.98 percent, according to pp. 966-972 of Ref. 8. Similarly, if $S/N = 4 = 6$ db, $\lambda = 1.8$, and the "worst-case" reliability is 96.4 percent. The values of α_3 , β_3 , and γ_3 corresponding to $N_{\text{opt}} = 7$ can be calculated with the aid of (14) through (16) or looked up on Fig. 3 as $\alpha_3 = 0.578$, $\beta_3 = 0.164$, and $\gamma_3 = 0.0078$, or, in exact octal fractions, $\alpha_3 = (0.45)_8$, $\beta_3 = (0.124)_8$, $\gamma_3 = (0.004)_8$.

The "worst case" dynamic error component W_c can be calculated with

* Because of this assumption that $L' = L$, it is also assumed that the dynamic error $|W_c|$ is large enough and σ_R is small enough so that

$$\frac{L + |W_c|}{\sigma_R} \gg \frac{L - |W_c|}{\sigma_R} = \lambda,$$

thereby making the reliability be significantly affected only by changes in the value of λ . This assertion can be verified by calculating

$$\frac{L + |W_c|}{\sigma_R} \quad \text{and} \quad \frac{L - |W_c|}{\sigma_R}$$

with the aid of the values of W_c and σ_R given in the last paragraph of this appendix and using the tables of the normal distribution function (pp. 966-972 of Ref. 8) to compare the effects of changes in

$$\frac{L + |W_c|}{\sigma_R} \quad \text{and} \quad \frac{L - |W_c|}{\sigma_R}.$$

the aid of Fig. 6 as 64 times $j_c T^3$, or 10.5 feet. The standard deviation σ_R of the random component can be calculated with the aid of Fig. 7 as $0.88 \sigma_\epsilon = 44' / \sqrt{S/N}$. If $S/N = 4 = 6$ db, $\sigma_R = 22$ feet and if $S/N = 16 = 12$ db, $\sigma_R = 11$ feet. As a check on the consistency of the results, (1) gives values of λ which are identical to those calculated in the previous paragraph.

REFERENCES

1. Helms, H. D., Random Walk Theory Probabilities That a Radar Servo Will Fail to Lock On, IEEE Transactions on Automatic Control, AC-9, No. 3, July, 1964, pp. 288-289.
2. Levine, N., Increasing the Flexibility of Recursive Least Squares Data Smoothing, B.S.T.J., 40, May, 1961, pp. 821-840. Also Bell Telephone Laboratories Monograph, 3892.
3. Brown, R. G., *Smoothing, Forecasting, and Prediction of Discrete Time Series*, Prentice-Hall, Englewood Cliffs, N. J., 1963.
4. Nesline, F. William, Jr., Polynomial Filtering of Signals, Proc. Conv. Mil. Elec., 1961, pp. 531-542 (Tables VI and VIII).
5. Benedict, T. R. and Bordner, G. W., Synthesis of an Optimal Set of Radar Track-While-Scan Smoothing Equations, IRE Trans. on Automatic Control, AC-7, July, 1962, pp. 27-32.
6. Skolnik, Merrill I., *Introduction to Radar Systems*, McGraw-Hill Book Co., New York, 1962.
7. Feller, William, *An Introduction to Probability Theory and Its Applications*, 2nd ed., John Wiley & Sons, Inc., New York.
8. Abramowitz, Milton and Stegun, Irene C., *Handbook of Mathematical Functions*, Nat. Bur. of Stds., App. Math. Series No. 55, U. S. Dept. of Commerce, U. S. Govt. Printing Office, Washington, D. C., June, 1964.
9. Whittaker, E. T. and Robinson, G., *The Calculus of Observations*, 4th ed., Blackie & Son Ltd., London, 1944, p. 308.
10. Brown, R. A. and Meyer, R. F., The Fundamental Theorem of Exponential Smoothing, Operations Research, 9, 1961, pp. 673-687.
11. Morrison, Norman, Recursive Numerical Filtering with Exponentially Decaying Weights, to be published.

The Joint Optimization of Transmitted Signal and Receiving Filter for Data Transmission Systems

By J. W. SMITH

(Manuscript received June 23, 1965)

The optimum signal and receiving filters for a data transmission system with a fixed channel and detection process are found. The criteria of optimality are: (i) the minimization of mean square error between the input multilevel data signal and the output to the decision threshold and (ii) the minimization of noise power output with no intersymbol interference. Explicit frequency characteristics are obtained for the separate and joint signal and receiver optimization problems. For the joint problem it is seen that some freedom exists in assigning phases to the transmitter and receiver. In addition, explicit equations are obtained for the output signal-to-noise ratio for the problems considered. The optimization procedures are carried out in detail in some examples. In comparing binary and correlative multilevel signalling (e.g., duobinary), it is seen that there may be channels for which the optimum multilevel system is superior to the optimum binary system.

I. INTRODUCTION

In most instances a data system designer is faced with the problem of transmitting through a noisy channel over which he has no control. In other words, the designer must concern himself with transmitter and receiver terminals which reduce the disturbing effects of the channel. Nyquist's classic paper¹ considered over-all system designs which eliminate intersymbol interference. Others (see, for example, Ref. 2) have also reduced the effects of noise by designing the transmitter and receiver to minimize the noise output for a given Nyquist characteristic and an ideal channel. Tufts' recent work³⁻⁶ has recognized that transmission without intersymbol interference may not be the most desirable. He has considered the problem of optimizing transmitted signal waveforms or receiver filters under the criterion of minimizing the mean square error (thus minimizing the joint contribution of noise and inter-

symbol interference). His attempt at joint optimization of the transmitter and receiver was successful only with the added condition that the transmitted waveform be time limited to one bit interval. This is an important case, but the unsolved joint optimization problem without this constraint is also important. The joint optimization solution would, in effect, provide a performance bound for a given channel.

This paper extends Tufts' results by solving the joint optimization problem (mean square error criterion) for a time-invariant transmitter and receiver subject only to an average power constraint on the transmitted waveform. Explicit equations are obtained for the signal, receiving filter, and the output signal-to-noise ratio for both the separate and joint optimization. In addition, the optimum signal and receiver are found for an arbitrary channel subject to the constraint of no intersymbol interference. The output signal-to-noise ratio for this case is obtained and is found to be only slightly less than the optimum for usual noise levels. The results show clearly the importance of the ratio of the noise spectral density to the square of the channel amplitude characteristic on the design techniques and the performance bound. The phase characteristics of the channel are irrelevant to the bound and there is some freedom in assigning phase characteristics to the transmitter and receiver in the joint optimization problems.

An important reason for the solution of the joint optimization problem is in the notation used. The frequency domain is broken into disjoint intervals and the characteristic within each interval is considered as a separate function. The equations to be minimized are easily stated and the constraints (such as no intersymbol interference) are compactly and completely represented by this notation. Finally, the joint solution is easily obtained with the notation used here whereas with the standard approaches,⁶ the solution is obscured.

II. GENERAL CONSIDERATIONS

Fig. 1 illustrates the general linear, time-invariant, noisy (zero mean) multilevel transmission system considered. One may assume that the information is contained in a random sequence of impulses (of weight

$$a_l, a_l = \{-2M, \dots, 0, \dots, 2M - 2, 2M\}$$

or

$$\{-2M - 1, \dots, -1, 1, \dots, 2M + 1\}$$

and spaced T seconds apart) at the input of the system. Thus, a signal, $s(t)$, having an amplitude of a_l is transmitted every T seconds. The

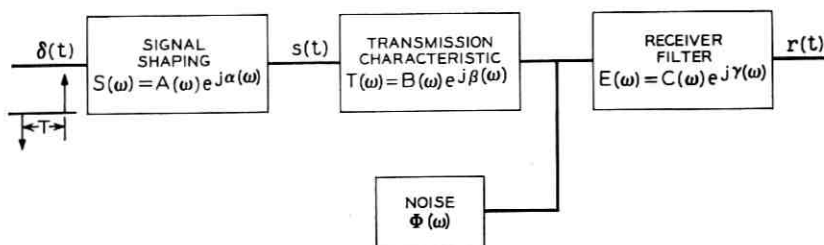


Fig. 1—General digital transmission system.

noiseless system output, $r(t)$, has the Fourier transform

$$R(\omega) = S(\omega)T(\omega)E(\omega) \quad (1)$$

A sequence of input signals

$$\sum_l a_l s(t - lT)$$

(all sums from $-\infty$ to ∞ unless otherwise noted) will, in general, produce a sequence of overlapping output pulses

$$\sum_l a_l r(t - lT).$$

The total output at the sampling time (taken at $t = 0$ without loss of generality) is

$$v = a_0 r(0) + \sum_{l \neq 0} a_l r(-lT) + n(0) \quad (2a)$$

or

$$v = a_0 r_0 + \sum_{l \neq 0} a_l r_{-l} + n_0 \quad (2b)$$

where r_l , $l \neq 0$ represents the intersymbol interference and r_0 is the output value at the main sample point.

In order to assess the performance of such a data system, it is necessary to choose a criterion of quality. One useful measure which Tufts³⁻⁶ has used extensively is the mean square error. This paper considers a slightly different criterion by using a normalized mean square error

$$E\{[\text{normalized input sample} - \text{output sample}]^2\}$$

or

$$E\{[a_0 r_0 - (a_0 r_0 + \sum_{l \neq 0} a_l r_{-l} + n_0)]^2\}$$

where a simple change of amplitude between input and output has been eliminated as a factor. Then

$$\begin{aligned} M.S.E. &= E\{[a_0 r_0 - (a_0 r_0 + \sum_{l \neq 0} a_l r_{-l} + n_0)]^2\} \\ &= E\{n_0^2\} + E\{2n_0 \sum_{l \neq 0} a_l r_{-l}\} + E\{(\sum_{l \neq 0} a_l r_{-l})^2\} \end{aligned} \quad (3a)$$

$$= E\{n_0^2\} + E\{\sum_{i, j \neq 0} a_i a_j r_{-i} r_{-j}\} \quad (3b)$$

$$= \sigma^2 + \overline{a_j^2} \sum_{l \neq 0} r_l^2 + 2 \sum_{m=1}^{\infty} \overline{a_j a_{j+m}} \sum_{l \neq 0, l+m} r_l r_{l+m} \quad (3c)$$

where

$$\sigma^2 = E\{n_0^2\}.$$

This expression cannot be an absolute measure of quality but should be related to the output level r_0 . Much of the remainder of the paper will be concerned with the design of the transmitted signal and the receiver filter (both separately and jointly) to minimize the mean square error given by (3c) for a given value of r_0 . In addition, this expression will also be minimized subject to the constraint that the intersymbol interference be zero, i.e.,

$$\sum_{l \neq 0} r_l^2 = 0.$$

It is particularly useful for our purposes to work with the frequency characteristics of the system. Appendix A gives the details of the transformation of the mean square error expression from a time domain to a frequency domain function. In terms of the system component transforms

Transmitted signal	$S(\omega) = A(\omega)e^{j\alpha(\omega)}$
Channel characteristic	$T(\omega) = B(\omega)e^{j\beta(\omega)}$
Receiver filter	$E(\omega) = C(\omega)e^{j\gamma(\omega)}$
Noise spectral density	$= \Phi(\omega)$

and the shorthand notation $A[u + (2n\pi/T)] = A_n(u)$ etc., the mean square error may be written

$$\begin{aligned}
M.S.E. = \int_{-\pi/T}^{\pi/T} \left\{ \sum_n C_n^2(u) \Phi_n(u) \right. \\
+ \frac{2\pi}{T} \left[\sum_n A_n(u) B_n(u) C_n(u) \cos [\alpha_n(u) + \beta_n(u) \right. \\
+ \left. \left. \gamma_n(u) \right] - \frac{r_0 T}{2\pi} \right]^2 F(u) \\
+ \frac{2\pi}{T} \left[\sum_n A_n(u) B_n(u) C_n(u) \sin [\alpha_n(u) + \beta_n(u) \right. \\
+ \left. \left. \gamma_n(u) \right] \right]^2 F(u) \left. \right\} du
\end{aligned} \quad (4)$$

where

$$F(u) = \bar{a}_j^2 + 2 \sum_{m=1}^{\infty} \bar{a}_j \bar{a}_{j+m} \cos umT. \quad (5)$$

In (4) the frequency range has been split into disjoint bands of width $2\pi/T$ and the characteristic within each band is treated as a separate function {e.g., $A_n(u) = A[u + (2n\pi/T)]$ is $A(\omega)$ for $(2n-1)\pi/T < \omega < (2n+1)\pi/T$ }. For simplicity, the argument, u , will not be made explicit (except for the dependence of the Lagrange multipliers) throughout the rest of the development.

The last two terms on the right side of (4) represent the intersymbol distortion. By simultaneously making these terms zero, one has the necessary conditions for transmission without intersymbol interference

$$\sum_n A_n B_n C_n \cos (\alpha_n + \beta_n + \gamma_n) = \frac{r_0 T}{2\pi} \quad (6)$$

$$\sum_n A_n B_n C_n \sin (\alpha_n + \beta_n + \gamma_n) = 0 \quad (7)$$

$$-\frac{\pi}{T} \leq u \leq \frac{\pi}{T}$$

which have been discussed in a previous paper.⁷

In the remainder of the paper, the optimum signal and equalizer (receiver filter) will be determined (separately in Section III and jointly in Section IV) under the constraint that the average signal power is limited and the output level is fixed. That is, the mean square error

$$\begin{aligned}
\int_{-\pi/T}^{\pi/T} \left\{ \sum_n C_n^2 \Phi_n + \frac{2\pi F}{T} \left(\sum_n A_n B_n C_n \cos (\alpha_n + \beta_n + \gamma_n) - \frac{r_0 T}{2\pi} \right)^2 \right. \\
\left. + \frac{2\pi F}{T} \left(\sum_n A_n B_n C_n \sin (\alpha_n + \beta_n + \gamma_n) \right)^2 \right\} du
\end{aligned}$$

will be minimized subject to

$$\int_{-\pi/T}^{\pi/T} \sum_n A_n B_n C_n \cos(\alpha_n + \beta_n + \gamma_n) du = r_0 \quad (8)$$

and the average power constraint which may be written⁸ (assuming $\overline{a_j} = 0$)

$$P = \int_{-\infty}^{\infty} F(\omega) A^2(\omega) d\omega = \int_{-\pi/T}^{\pi/T} F \sum_n A_n^2 du. \quad (9)$$

Note that this differs from Tufts' development⁵ in that he considers a fixed pulse energy as the constraint.

In addition, the suboptimum signal and equalizer which eliminate intersymbol interference will be found by minimizing

$$\int_{-\pi/T}^{\pi/T} \sum_n C_n^2 \Phi_n du$$

subject to the constraints given by (6), (7), and (9).

Note that (6) contains the constraint on r_0 , given by (8).

III. RESULTS OF THE SEPARATE OPTIMIZATION OF SIGNAL AND RECEIVER FILTER

3.1 Simultaneous Reduction of Noise and Intersymbol Distortion

Appendix B shows the details of the minimization of the *M.S.E.* given by (4) subject to the constraints of (8) and (9). For a specified receiver filter the optimum signal characteristics are given by

$$\sin(\alpha_k + \beta_k + \gamma_k) = 0 \quad (10)$$

and

$$A_k = \frac{T(2r_0 - \lambda_2/F)}{4\pi} \frac{B_k C_k}{\sum_k B_k^2 C_k^2 + \frac{\lambda_1 T}{2\pi}} \quad (11)$$

where it is assumed that $F \neq 0$ at any point. The constants λ_1 and λ_2 may be determined by using the constraining equations.

For a fixed signal, the optimum receiver characteristics are given by

$$\sin(\alpha_k + \beta_k + \gamma_k) = 0 \quad (10)$$

and

$$C_k = \frac{T(2r_0 - \lambda_2/F)}{4\pi} \frac{A_k B_k F / \Phi_k}{F \sum_k \frac{A_k^2 B_k^2}{\Phi_k} + \frac{T}{2\pi}} \quad (12a)$$

The coefficient is found in Appendix B to be

$$\frac{T(2r_0 - \lambda_2/F)}{4\pi} = \frac{\frac{r_0 T}{2\pi} \left[\int_{-\pi/T}^{\pi/T} \frac{\sum_k \frac{A_k^2 B_k^2}{\Phi_k} du}{F \sum_k \frac{A_k^2 B_k^2}{\Phi_k} + \frac{T}{2\pi}} + \frac{1}{F} \int_{-\pi/T}^{\pi/T} \frac{\frac{T}{2\pi} du}{F \sum_k \frac{A_k^2 B_k^2}{\Phi_k} + \frac{T}{2\pi}} \right]}{\int_{-\pi/T}^{\pi/T} \frac{\sum_k \frac{A_k^2 B_k^2}{\Phi_k} du}{F \sum_k \frac{A_k^2 B_k^2}{\Phi_k} + \frac{T}{2\pi}}} \quad (12b)$$

The ratio of the signal output to the square root of the mean square error is

$$\frac{r_0}{\sqrt{M.S.E.}} = \frac{2\pi}{T} \left[\frac{\int_{-\pi/T}^{\pi/T} \frac{F du}{\Delta} \int_{-\pi/T}^{\pi/T} \frac{\sum_k \frac{A_k^2 B_k^2}{\Phi_k} du}{\Delta} + \frac{T}{2\pi} \left(\int_{-\pi/T}^{\pi/T} \frac{du}{\Delta} \right)^2}{\int_{-\pi/T}^{\pi/T} \frac{\sum_k \frac{A_k^2 B_k^2}{\Phi_k} du}{\Delta}} \right]^{-1} \quad (13a)$$

where

$$\Delta = F \sum_k \frac{A_k^2 B_k^2}{\Phi_k} + \frac{T}{2\pi}.$$

This function may be approximated by

$$\frac{r_0}{\sqrt{M.S.E.}} \approx \frac{2\pi}{T} \left[\int_{-\pi/T}^{\pi/T} \frac{du}{\sum_k \frac{A_k^2 B_k^2}{\Phi_k}} - \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{du}{F \left(\sum_k \frac{A_k^2 B_k^2}{\Phi_k} \right)^2} \right]^{-1} \quad (13b)$$

for large transmitted power to noise power ratios

$$\frac{2\pi}{T} F \sum_k \frac{A_k^2 B_k^2}{\Phi_k} \gg 1.$$

In each case, the variable phase should be adjusted to produce zero phase (with any fixed delay removed) across the band

$$(i.e., \alpha_k = -\beta_k - \gamma_k).$$

The amplitude characteristics may be considered to be the product

of a term which depends upon u and k and one that only depends upon u , e.g.,

$$C_k(u) = [B_k(u)A_k(u)/\Phi_k(u)]q(u) \quad (14)$$

where

$$q\left(\omega + \frac{2k\pi}{T}\right) = q\left(\omega + \frac{2l\pi}{T}\right)$$

for all k, l . The function $q(u)$ may then be considered as the output of a tapped delay line

$$\sum_p q_p e^{-j\omega p T}$$

since this function satisfies the condition on $q(u)$. Thus, the solutions may be interpreted as linear filters matched to the signal and channel followed by a tapped delay line, e.g.,

$$C(\omega) = [B(\omega)A(\omega)/\Phi(\omega)]q(\omega). \quad (15)$$

3.2 Minimization of Noise Output Power with No Intersymbol Distortion

The minimization of

$$\int_{-\pi/T}^{\pi/T} \sum_n C_n^2 \Phi_n du + \lambda_3(u) \sum_n A_n B_n C_n \cos(\alpha_n + \beta_n + \gamma_n) \\ + \lambda_4(u) \sum_n A_n B_n C_n \sin(\alpha_n + \beta_n + \gamma_n) + \lambda_1 \int_{-\pi/T}^{\pi/T} F \sum A_n^2 du$$

subject to the constraints of (6, 7, 9) is similar to the previous case and will not be presented in detail. For a fixed receiver filter, the optimum signal characteristics are given by

$$\sin(\alpha_k + \beta_k + \gamma_k) = 0 \quad (10)$$

and

$$A_k = \frac{-\lambda_3(u)}{2\lambda_1 F} B_k C_k. \quad (16a)$$

The coefficient is easily found and

$$A_k = \frac{r_0 T}{2\pi} \frac{B_k C_k}{\sum_k B_k^2 C_k^2}. \quad (16b)$$

For a fixed signal, the optimum receiver characteristics are given by

$$\sin(\alpha_k + \beta_k + \gamma_k) = 0 \quad (10)$$

and

$$C_k = \frac{-\lambda_3(u)}{2} \frac{A_k B_k}{\Phi_k}. \quad (17a)$$

Again, the coefficient is easily found and

$$C_k = \frac{r_0 T}{2\pi} \frac{\frac{A_k B_k}{\Phi_k}}{\sum_k \frac{A_k^2 B_k^2}{\Phi_k}}. \quad (17b)$$

The ratio of signal output to the noise output is

$$\frac{r_0}{\sqrt{\sigma^2}} = \frac{2\pi}{T} \left[\int_{-\pi/T}^{\pi/T} \frac{du}{\sum_k \frac{A_k^2 B_k^2}{\Phi_k}} \right]^{-1}. \quad (18)$$

This function is seen to be identical to the first term of (13b). The difference between (18) and (13b) is small under the condition that

$$\frac{2\pi}{T} F \sum_k \frac{A_k^2 B_k^2}{\Phi_k} \gg 1$$

and one does not sacrifice much in designing for no intersymbol interference. In this case, there is no significant advantage in using the true optimum design with large signal-to-noise ratio systems.

The interpretation of the above results is the same as for the previous minimization problem.

IV. JOINT OPTIMIZATION OF SIGNAL AND RECEIVER FILTER

4.1 Simultaneous Reduction of Noise and Intersymbol Interference

The over-all optimization requires the minimization of (4) with respect to both the signal and receiver functions A_k , C_k , α_k and γ_k . From the minimization details in Appendix B, this requires the solution of

$$\sin(\alpha_k + \beta_k + \gamma_k) = 0 \quad \text{for all } k, \quad (10)$$

$$A_k = \frac{-2\pi}{\lambda_1 T} B_k C_k \left\{ \sum_n A_n B_n C_n - \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} \right) \right\}, \quad (19)$$

and

$$C_k = \frac{-2\pi}{T} \frac{A_k B_k F}{\Phi_k} \left\{ \sum_n A_n B_n C_n - \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} \right) \right\} \quad \text{for all } k \quad (20)$$

where it is assumed $F \neq 0$. In some correlation schemes F may be zero at certain points and it is easily seen that $A_k = C_k = 0$ at these points. For (19) and (20) to have nontrivial solutions it is necessary that the determinant

$$1 - \left(\frac{2\pi}{T} \right)^2 \frac{B_k^2 F}{\lambda_1 \Phi_k} \left\{ \sum_n A_n B_n C_n - \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} \right) \right\}^2$$

vanish. This requires

$$\sum_n A_n B_n C_n - \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} \right) = \pm \frac{T}{2\pi} \frac{\lambda_1^{\frac{1}{2}} \Phi_k^{\frac{1}{2}}}{B_k F^{\frac{1}{2}}} \quad \text{for all } k. \quad (21)$$

In general (unless $\Phi_k^{\frac{1}{2}}/B_k$ is a constant with k) for any u the equation can only be satisfied for one k . This means that for any u there is only one value of k for which there is a nontrivial solution and for the other values of k the solutions vanish.

The question now becomes one of choosing, for each u , the proper A_k and C_k which leads to the true minimization. Combining (19) and (20)

$$A_k = \frac{\Phi_k^{\frac{1}{2}}}{\lambda_1^{\frac{1}{2}} F^{\frac{1}{2}}} C_k \quad (22)$$

and using

$$A_k B_k C_k - \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} \right) = \frac{-T}{4\pi} \frac{2\lambda_1^{\frac{1}{2}} \Phi_k^{\frac{1}{2}}}{B_k F^{\frac{1}{2}}} \quad (21)$$

one gets

$$A_k^2 = \frac{T}{4\pi} \left\{ 2r_0 - \frac{\lambda_2}{F} - \frac{2\lambda_1^{\frac{1}{2}} \Phi_k^{\frac{1}{2}}}{B_k F^{\frac{1}{2}}} \right\} \frac{\Phi_k^{\frac{1}{2}}}{\lambda_1^{\frac{1}{2}} B_k F^{\frac{1}{2}}} \quad (23)$$

$$C_k^2 = \frac{T}{4\pi} \left\{ 2r_0 - \frac{\lambda_2}{F} - \frac{2\lambda_1^{\frac{1}{2}} \Phi_k^{\frac{1}{2}}}{B_k F^{\frac{1}{2}}} \right\} \frac{\lambda_1^{\frac{1}{2}} F^{\frac{1}{2}}}{\Phi_k^{\frac{1}{2}} B_k} \quad (24)$$

Equations (23) and (24) represent the nonzero solutions. The best k (frequency region) for any u is the one which minimizes the function

$$\int_{-\pi/T}^{\pi/T} \left[\sum_k C_k^2(u) \Phi_k(u) + \frac{2\pi F}{T} \left\{ \sum_k A_k(u) B_k(u) C_k(u) - \frac{r_0 T}{2\pi} \right\}^2 \right] du$$

where the dependence on u has been reinserted for clarity. For every u there is at most one C_k and A_k with a nonzero value so the function may be written

$$\sum_k \int_{L_k} \left[C_k^2(u) \Phi_k(u) + \frac{2\pi}{T} F(u) \left\{ A_k(u) B_k(u) C_k(u) - \frac{r_0 T}{2\pi} \right\}^2 \right] du$$

where L_k is the subinterval within the interval

$$[(2k - 1) (\pi/T), \quad (2k + 1) (\pi/T)]$$

over which $A_k(u)$ and $C_k(u)$ have a value. It is necessary for the true optimum that L_k be chosen to minimize the above expression. The true minimization may not require the use of the total interval $-\pi/T < u < \pi/T$, i.e., it may be possible to decrease noise output faster than inter-symbol distortion is increased by using a smaller interval. Because this situation only arises when $\Phi_k^{1/2}/B_k$ in the optimum L_k regions have large variation, and is cumbersome to consider in detail, it will be assumed that nonzero A_k and C_k will exist over $-\pi/T < u < \pi/T$ (unless $F = 0$). Using (21, 23, 24) the above function may be written

$$\begin{aligned} \sum_k \int_{L_k} \left[\frac{T}{4\pi} \left\{ 2r_0 - \frac{\lambda_2}{F(u)} - \frac{2\lambda_1^{1/2} \Phi_k^{1/2}(u)}{B_k(u) F^{1/2}(u)} \right\} \frac{\lambda_1^{1/2} \Phi_k^{1/2}(u) F^{1/2}(u)}{B_k(u)} \right. \\ \left. + \frac{T}{4\pi} \frac{F(u)}{2} \left\{ -\frac{\lambda_2}{F(u)} - \frac{2\lambda_1^{1/2} \Phi_k^{1/2}(u)}{B_k(u) F^{1/2}(u)} \right\}^2 \right] du. \end{aligned}$$

The constants may be evaluated by using (21, 23, 24) and the constraining equations. Thus,

$$r_0 = \int_{-\pi/T}^{\pi/T} \sum_k A_k B_k C_k du \quad (8)$$

$$= \sum_k \int_{L_k} \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} - \frac{2\lambda_1^{1/2} \Phi_k^{1/2}}{B_k F^{1/2}} \right) du \quad (25a)$$

$$= r_0 - \frac{\lambda_2 T}{4\pi} \sum_k \int_{L_k} \frac{du}{F} - \frac{T \lambda_1^{1/2}}{2\pi} \sum_k \int_{L_k} \frac{\Phi_k^{1/2}}{B_k} du \quad (25b)$$

since the total range of all L_k is the interval on u ($-\pi/T, \pi/T$) and

$$\frac{\lambda_2}{\lambda_1^{1/2}} = -2 \sum_k \int_{L_k} \frac{\Phi_k^{1/2}}{B_k} du / \sum_k \int_{L_k} \frac{du}{F}. \quad (26)$$

Using the average power condition

$$P = \int_{-\pi/T}^{\pi/T} F \sum_k A_k^2 du \quad (9)$$

$$= \sum_k \int_{L_k} \frac{T}{4\pi} \left\{ 2r_0 - \frac{\lambda_2}{F} - \frac{2\lambda_1^{1/2} \Phi_k^{1/2}}{B_k F^{1/2}} \right\} \frac{\Phi_k^{1/2} F^{1/2}}{\lambda_1^{1/2} B_k} \quad (27a)$$

$$\frac{r_0 T}{2\pi \lambda_1^{\frac{1}{2}}} \sum_k \int_{L_k} \frac{F^{\frac{1}{2}} \Phi_k^{\frac{1}{2}}}{B_k} du = P + \frac{T}{2\pi} \sum_k \int_{L_k} \frac{\Phi_k}{B_k^2} du + \frac{T}{4\pi} \frac{\lambda_2}{\lambda_1^{\frac{1}{2}}} \sum_k \int_{L_k} \frac{\Phi_k^{\frac{1}{2}}}{B_k F^{\frac{1}{2}}} du. \quad (27b)$$

Combining (27b) and (26) one gets

$$\lambda_1^{\frac{1}{2}} = \frac{r_0 T}{2\pi} \sum_k \int_{L_k} \frac{F^{\frac{1}{2}} \Phi_k^{\frac{1}{2}}}{B_k} du \left[P + \frac{T}{2\pi} \sum_k \int_{L_k} \frac{\Phi_k}{B_k^2} du - \frac{T}{2\pi} \frac{\left(\sum_k \int_{L_k} \frac{\Phi_k^{\frac{1}{2}}}{B_k F^{\frac{1}{2}}} du \right)^2}{\sum_k \int_{L_k} \frac{du}{F}} \right]^{-1}. \quad (28)$$

Finally, combining (21-23, 25 and 27)

$$M.S.E. = \frac{\left(\frac{r_0 T}{2\pi} \sum_k \int_{L_k} \frac{F^{\frac{1}{2}} \Phi_k^{\frac{1}{2}}}{B_k} du \right)^2}{\left[P + \frac{T}{2\pi} \sum_k \int_{L_k} \frac{\Phi_k}{B_k^2} du - \frac{T}{2\pi} \frac{\left(\sum_k \int_{L_k} \frac{\Phi_k^{\frac{1}{2}}}{B_k F^{\frac{1}{2}}} du \right)^2}{\sum_k \int_{L_k} \frac{du}{F}} \right]}. \quad (29)$$

Equation (29) represents the mean square error in the transmission and L_k must be chosen to minimize this expression. This is rather complicated but to a close approximation, the minimization of

$$\sum_k \int_{L_k} \frac{F^{\frac{1}{2}} \Phi_k^{\frac{1}{2}}}{B_k} du$$

corresponds to the minimization of (29). The minimum mean square error results by choosing for each value of u

$$\left(-\frac{\pi}{T} \leq u \leq \frac{\pi}{T} \right)$$

the k for which

$$\frac{\Phi_k^{\frac{1}{2}}(u) F^{\frac{1}{2}}(u)}{B_k(u)}$$

is a minimum. The interval L_k is that region of u for which

$$\frac{\Phi_k^{\frac{1}{2}}(u)F^{\frac{1}{2}}(u)}{B_k(u)}$$

is a smaller than any other

$$\frac{\Phi_j^{\frac{1}{2}}(u)F^{\frac{1}{2}}(u)}{B_j(u)}.$$

Since $F(u)$ is independent of j , L_k may also be defined as the region for which $\Phi_k^{\frac{1}{2}}(u)/B_k(u)$ is smaller than any other $\Phi_j^{\frac{1}{2}}(u)/B_j(u)$. Thus, the choice of L_k to minimize

$$\sum_k \int_{L_k} \frac{\Phi_k^{\frac{1}{2}}}{B_k} du$$

also corresponds to the minimization of the *M.S.E.* and the regions in which power is transmitted are independent of the correlation properties of the data.

Once the L_k are chosen, the resulting characteristics are

$$A_k^2 = \left[\frac{P + \frac{T}{2\pi} \sum_k \int_{L_k} \frac{\Phi_k}{B_k^2} du - \frac{T}{2\pi} \left(\sum_k \int_{L_k} \frac{\Phi_k^{\frac{1}{2}}}{B_k F^{\frac{1}{2}}} du \right)^2 / \sum_k \int_{L_k} \frac{du}{F}}{\sum_k \int_{L_k} \frac{\Phi_k^{\frac{1}{2}} F^{\frac{1}{2}}}{B_k} du} \right] \quad (30)$$

$$\times \frac{\Phi_k^{\frac{1}{2}}}{B_k F^{\frac{1}{2}}} + \left[\frac{T}{2\pi} \sum_k \int_{L_k} \frac{\Phi_k^{\frac{1}{2}}}{B_k F^{\frac{1}{2}}} du / \sum_k \int_{L_k} \frac{du}{F} \right] \frac{\Phi_k^{\frac{1}{2}}}{B_k F^{\frac{1}{2}}} - \frac{T}{2\pi} \frac{\Phi_k}{B_k^2 F}$$

and

$$C_k^2 = \frac{\lambda_1 F}{\Phi_k} A_k^2 \quad (22)$$

where λ_1 is given by (28). For (30) to be a solution requires that $A_k^2 \geq 0$ or

$$\frac{\Phi_k^{\frac{1}{2}}}{B_k F^{\frac{1}{2}}} < K_1 + \frac{K_2}{F} \quad (31)$$

for the regions defined by the L_k . Taking into account the difference in problems considered, this equation corresponds to Tuft's condition for a solution [Ref. 5, (28)]. Equation (31) states that the solution must be limited to regions without sharp peaks in the $\Phi_k^{\frac{1}{2}}/B_k F^{\frac{1}{2}}$ characteristic. The phases, α_k and γ_k , may be arbitrarily assigned to A_k and C_k subject to the condition of (10). The ratio of the signal to the square root of the mean square distortion is

$$\frac{r_0}{\sqrt{M.S.E.}} = \frac{2\pi}{T} \times \frac{\left[P + \frac{T}{2\pi} \sum_k \int_{L_k} \frac{\Phi_k}{B_k^2} du - \frac{T}{2\pi} \left(\sum_k \int_{L_k} \frac{\Phi_k^{\frac{1}{2}}}{B_k F^{\frac{1}{2}}} du \right)^2 / \sum_k \int_{L_k} \frac{du}{F} \right]^{\frac{1}{2}}}{\sum_k \int_{L_k} \frac{F^{\frac{1}{2}} \Phi_k^{\frac{1}{2}}}{B_k} du}. \quad (32)$$

It is difficult to make a general comparison of this result and (13) because of the arbitrary signal, A_k , which appears in (13). A comparison for a specific case is found in the examples of Section V.

4.2 Minimization of Noise Output with No Intersymbol Interference

From the previous results, the minimization of the noise output with respect to both the signal and receiver functions, A_k , C_k , α_k and γ_k requires the simultaneous solution of

$$\sin(\alpha_k + \beta_k + \gamma_k) = 0 \quad \text{for all } k, \quad (10)$$

$$A_k = \frac{-\lambda_3(u)}{2\lambda_1 F} B_k C_k \quad (16a)$$

and

$$C_k = \frac{-\lambda_3(u)}{2} \frac{A_k B_k}{\Phi_k} \quad \text{for all } k \quad (17a)$$

where again nonzero F is assumed (if $F = 0$, $C_k = 0$, $A_k = \infty$). A non-trivial solution of (16a) and (17a) requires that the determinant

$$1 - \frac{\lambda_3^2(u)}{4\lambda_1 F} \frac{B_k^2}{\Phi_k}$$

vanish. This requires

$$\lambda_3(u) = \pm \frac{2\lambda_1^{\frac{1}{2}} \Phi_k^{\frac{1}{2}} F^{\frac{1}{2}}}{B_k}. \quad (33)$$

Again, for a particular u , the right hand side of the equation depends upon k and can only be satisfied for one value of k . For the other values of k , (16a) and (17a) have trivial solutions. At each value of u there must be at least one nontrivial A_k and C_k because of the condition

$$\sum_k A_k B_k C_k = \frac{r_0 T}{2\pi} \quad -\frac{\pi}{T} \leq u \leq \frac{\pi}{T}. \quad (6)$$

Thus, for each value of u there is one and only one nonzero A_k and C_k and the question becomes one of finding which k leads to the true optimum.

Using the same techniques as shown in the previous section, the noise output is given by

$$\sum_k \int_{L_k} C_k^2(u) \Phi_k(u) du = \left(\frac{r_0 T}{2\pi} \right)^2 \frac{1}{P} \left\{ \sum_k \int_{L_k} \frac{F^{\frac{1}{2}}(u) \Phi_k^{\frac{1}{2}}(u)}{B_k(u)} du \right\}^2. \quad (34)$$

This expression is minimized by choosing for each value of u , the k for which

$$\frac{\Phi_k^{\frac{1}{2}}(u)}{B_k(u)} \quad \left(\text{or } \frac{\Phi_k^{\frac{1}{2}}(u)}{B_k(u)} F^{\frac{1}{2}}(u) \right)$$

is a minimum. The interval L_k is that region of u for which

$$\frac{\Phi_k^{\frac{1}{2}}(u)}{B_k(u)}$$

is smaller than any other

$$\frac{\Phi_j^{\frac{1}{2}}(u)}{B_j(u)}.$$

The resulting characteristics are easily found to be

$$A_k = P^{\frac{1}{2}} \left[\sum_k \int_{L_k} \frac{F^{\frac{1}{2}} \Phi_k^{\frac{1}{2}}}{B_k} du \right]^{-\frac{1}{2}} \frac{\Phi_k^{\frac{1}{2}}}{B_k^{\frac{1}{2}} F^{\frac{1}{2}}} \quad (35)$$

and

$$C_k = \frac{r_0 T}{2\pi} P^{-\frac{1}{2}} \left[\sum_k \int_{L_k} \frac{F^{\frac{1}{2}} \Phi_k^{\frac{1}{2}}}{B_k} du \right]^{\frac{1}{2}} \frac{F^{\frac{1}{2}}}{\Phi_k^{\frac{1}{2}} B_k^{\frac{1}{2}}}. \quad (36)$$

Note here and in the previous case, the bound depends upon the ratio $\Phi_k^{\frac{1}{2}}/B_k$ but the transmitting and receiving filters required to achieve the bound depend upon Φ_k and B_k separately.

Again, the phases, α_k and γ_k , may be arbitrarily assigned to the transmitter and receiver subject to the condition of (10). The ratio of the signal output to the noise output is

$$\frac{r_0}{\sqrt{\sigma^2}} = \frac{2\pi}{T} P^{\frac{1}{2}} \left[\sum_k \int_{L_k} \frac{F^{\frac{1}{2}} \Phi_k^{\frac{1}{2}}}{B_k} du \right]^{-1}. \quad (37)$$

Comparing this equation and (32) shows the degradation in the performance bound to be expected when no intersymbol interference is allowed. This degradation is small under the usual condition of low

noise. The two equations are identical if

$$\frac{\Phi_k^{1/2}(u)}{B_k(u)} = \frac{K}{F^{1/2}(u)}$$

If this condition holds, the optimum solution is one with no intersymbol interference. This condition is especially useful with uncorrelated data where F is a constant.

V. EXAMPLES

Consider the problem of transmitting uncorrelated binary data ($a_i = -1$ or 1 , and $F = 1$) at a rate of $1/T$ bits per second over the channel whose amplitude characteristics $B(\omega)$ and noise spectral density function $\Phi(\omega)$ are shown in Fig. 2. For ease of presentation these functions are limited to the region $(-2\pi/T, 2\pi/T)$ and the phase characteristic is assumed to be zero. The preceding discussion will be illustrated by assuming a fixed signal and then finding the optimum receiver filter. The bound for the utilization of this channel will then be found by finding the joint optimum receiver and signal combination. For this purpose, the function

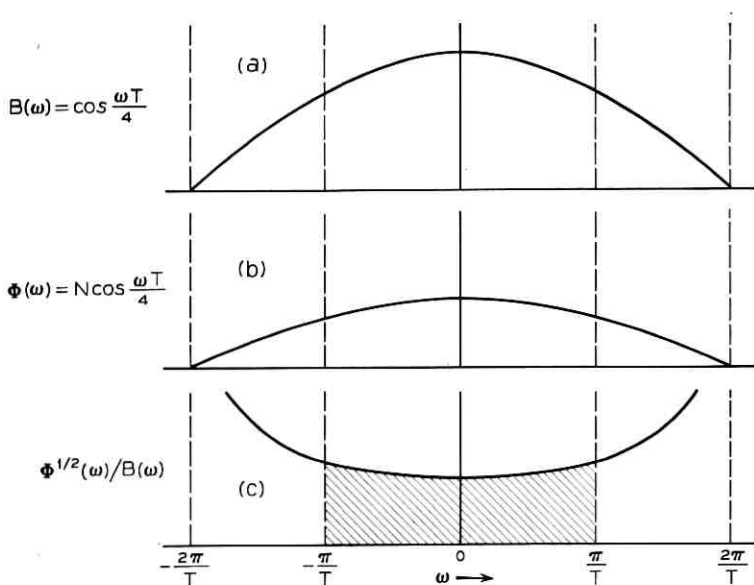


Fig. 2 — (a) Amplitude characteristic; (b) noise spectral density; (c) optimum operating region.

$$\frac{\Phi^{\frac{1}{2}}(\omega)}{B(\omega)}$$

is given in Fig. 2(c).

The various quantities of interest are given in the figure and it will be assumed that the initial choice of a signal gives a cosine roll-off and produces no intersymbol interference

$$A(\omega) = \sqrt{\frac{PT}{2\pi}} \cos \frac{\omega T}{4} \quad (38)$$

where the signal normalized to the average power restriction. The initial receiver filter is assumed to be flat. For this system

$$\frac{r_0}{\sqrt{\sigma^2}} = \sqrt{\frac{PT}{2\pi N}} \frac{\int_{-2\pi/T}^{2\pi/T} \cos^2 \frac{\omega T}{4} d\omega}{\left[\int_{-2\pi/T}^{2\pi/T} \cos \frac{\omega T}{4} d\omega \right]^{\frac{1}{2}}} = \sqrt{\frac{P\pi}{4N}} = 0.88 \sqrt{\frac{P}{N}}. \quad (39)$$

For no intersymbol interference, the optimum receiver filter is given by

$$C_k = \frac{r_0 T}{2\pi} \frac{\frac{A_k B_k}{\Phi_k}}{\sum_k \frac{A_k^2 B_k^2}{\Phi_k}} \quad (17b)$$

$$= \frac{r_0 T}{2\pi} \sqrt{\frac{2\pi}{PT}} \frac{B_k}{\cos^3 \frac{uT}{4} - \sin^3 \frac{uT}{4}} \quad -\frac{\pi}{T} \leq u \leq 0 \quad (40a)$$

$$C_k = \frac{r_0 T}{2\pi} \sqrt{\frac{2\pi}{PT}} \frac{B_k}{\cos^3 \frac{uT}{4} + \sin^3 \frac{uT}{4}} \quad 0 \leq u \leq \frac{\pi}{T} \quad (40b)$$

and

$$\frac{r_0}{\sqrt{\sigma^2}} = \frac{2\pi}{T} \left[\int_{-\pi/T}^{\pi/T} \frac{du}{\sum_k \frac{A_k^2 B_k^2}{\Phi_k}} \right]^{\frac{1}{2}} \quad (18)$$

$$= \sqrt{\frac{2\pi P}{NT}} \left[2 \int_0^{\pi/T} \frac{du}{\cos^3 \frac{uT}{4} + \sin^3 \frac{uT}{4}} \right]^{\frac{1}{2}} \quad (41a)$$

$$\approx \sqrt{\frac{2\pi P}{NT}} \left[\frac{2.4\pi}{T} \right]^{\frac{1}{2}} = 0.91 \sqrt{\frac{P}{N}}. \quad (41b)$$

For the optimum filter where intersymbol interference is allowed,

$$C_k = \frac{r_0 \sqrt{\frac{2\pi}{PT}}}{2 \int_0^{\pi/T} \frac{\cos^3 \frac{uT}{4} + \sin^3 \frac{uT}{4} du}{\cos^3 \frac{uT}{4} + \sin^3 \frac{uT}{4} + \frac{N}{P}}} \frac{B_k}{\cos^3 \frac{uT}{4} + \sin^3 \frac{uT}{4} + \frac{N}{P}} \quad (42)$$

$$0 \leq u \leq \frac{\pi}{T}$$

which will not be much different from the result with no intersymbol interference for low noise. The optimum detector with no intersymbol interference (40a,b) is shown on Fig. 3(a).

The joint optimization requires constraining $C(\omega)$ and $A(\omega)$ to the region $-\pi/T \leq u \leq \pi/T$. From Fig. 2(c) this is seen to be the region for which

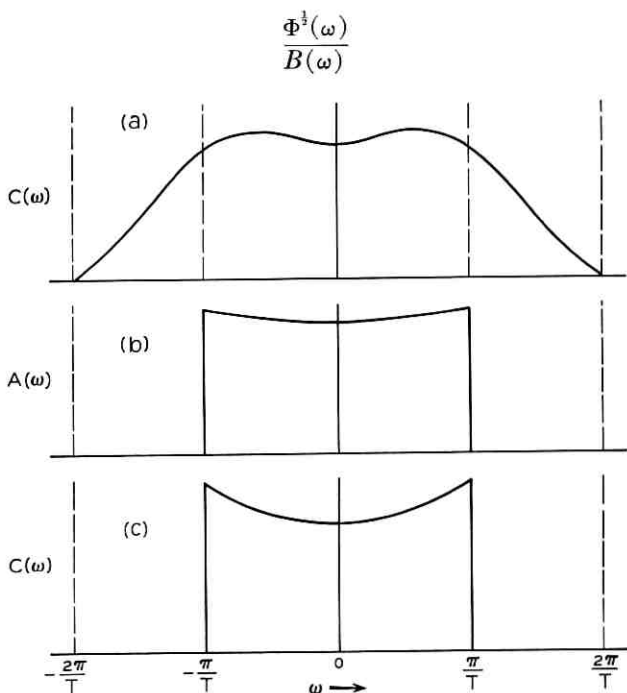


Fig. 3—(a) Optimum receiver shaping for given $A(\omega)$; (b) signal shaping for joint optimization; (c) receiver shaping for joint optimization.

is a minimum. The optimum signal and receiver for no intersymbol interference are, using (35, 36),

$$A_0 = \frac{P^{\frac{1}{2}}}{\left\{ \int_{-\pi/T}^{\pi/T} \frac{du}{\cos^{\frac{1}{2}} \frac{uT}{4}} \right\}^{\frac{1}{2}}} \frac{1}{\cos^{\frac{1}{2}} \frac{uT}{4}} \quad (43)$$

$$C_0 = \frac{r_0 T}{2\pi P^{\frac{1}{2}}} \frac{\left\{ \int_{-\pi/T}^{\pi/T} \frac{du}{\cos^{\frac{1}{2}} \frac{uT}{4}} \right\}^{\frac{1}{2}}}{\cos^{\frac{1}{2}} \frac{uT}{4}} \quad (44)$$

and are shown in Figs. 3(b) and 3(c). The resulting ratio of signal-to-noise is from (37)

$$\frac{r_0}{\sigma} = \frac{2\pi}{T} \frac{P^{\frac{1}{2}}}{N^{\frac{1}{2}}} \left[\int_{-\pi/T}^{\pi/T} \frac{du}{\cos^{\frac{1}{2}} \frac{uT}{4}} \right]^{-1} \quad (45a)$$

$$\approx 0.95 \sqrt{\frac{P}{N}}. \quad (45b)$$

The true optimum filters, with intersymbol interference, will be close to the previous results for low noise. The improvement in signal-to-noise ratio may be obtained from (45b) by replacing P by

$$P + \frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{N}{\cos \frac{uT}{4}} du - \left(\frac{T}{2\pi} \int_{-\pi/T}^{\pi/T} \frac{N^{\frac{1}{2}} du}{\cos^{\frac{1}{2}} \frac{uT}{4}} \right)^2$$

which is greater than P .

As a further illustration, consider the utilization of the channel whose amplitude characteristic $B(\omega)$ and noise spectral density function $\Phi(\omega)$ are shown in Fig. 4. The main point of this example is that the solution of the joint optimization problem is not confined to the interval $(-\pi/T, \pi/T)$. It is, as a matter of fact, split up into three regions; L_1 for $-\pi/T \leq u < -a$, L_0 for $-a < u < a$, and L_{-1} for $a < u \leq \pi/T$ as shown in Fig. 4(c). These turn out to be the regions of minimum

$$\frac{\Phi_k^{\frac{1}{2}}(u)}{B_k(u)}$$

as can be seen by the curve Fig. 4(c). The optimum receiver and signal are given in Fig. 5 for the case of no intersymbol interference.

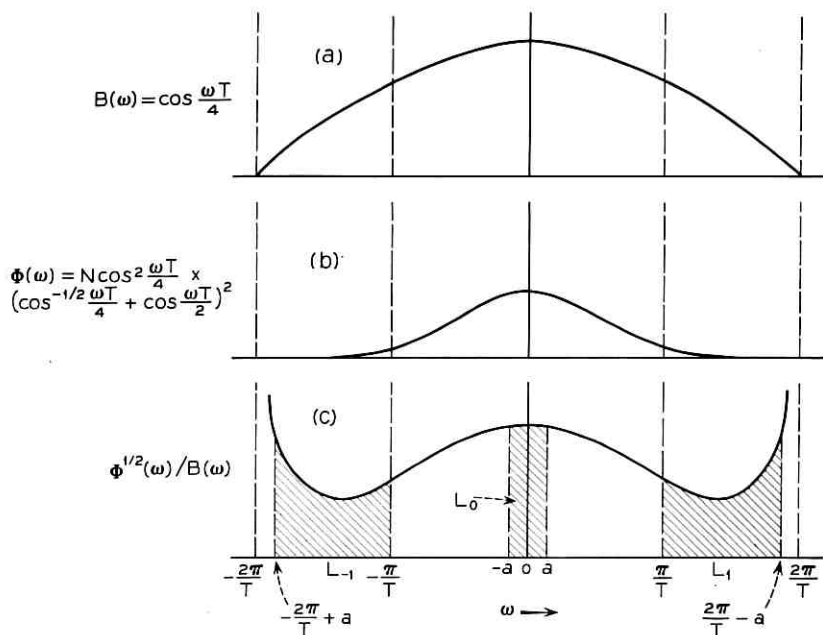


Fig. 4—(a) Amplitude characteristic; (b) noise spectral density; (c) optimum operating region.

The previous two examples have considered the problem of finding the optimum regions, L_k , for transmitting energy. Once these regions have been determined it is of interest to examine the effect of $F(u)$ (the one parameter over which the designer may exercise some control) on the performance bound. In particular, since it is generally considered that a binary system is optimum for transmitting data in fixed time slots, it is interesting to ask if some sort of correlation scheme leads to a better output signal-to-noise ratio than binary for the same average input power. In other words, using (37) [the case for no intersymbol interference] are there cases in which

$$\sum_k \int_{L_k} \frac{F^{\frac{1}{2}}(u) \Phi_k^{\frac{1}{2}}(u)}{B_k(u)} du \leq \sum_k \int_{L_k} \frac{\Phi_k^{\frac{1}{2}}(u)}{B_k(u)} du? \quad (46)$$

It should be emphasized that the signal-to-noise ratio is being examined and not the error rate. For correlative multilevel schemes which lead to a solution of (46) the error rate may be greater than or less than that of the binary system. In addition, there are certain error detecting

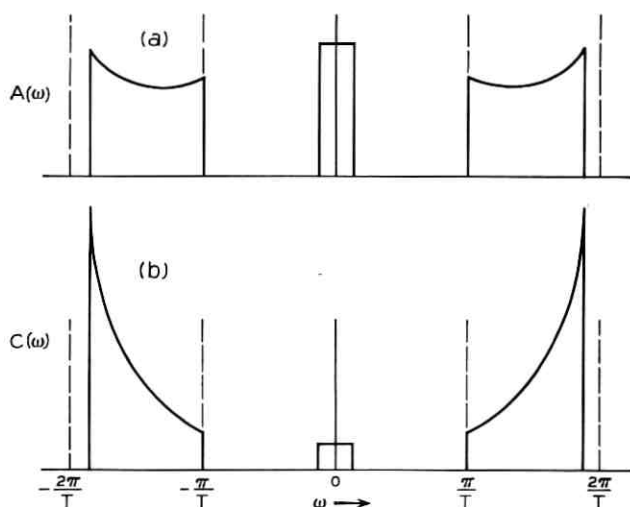


Fig. 5—(a) Signal shaping for joint optimization; (b) receiver shaping for joint optimization.

features of the correlative multilevel systems¹⁰ which make it difficult to compare its error rate to binary.

Equation (46) will be easier to consider if it is written

$$\int_{-\pi/T}^{\pi/T} \frac{F^{\frac{1}{2}}(u)\Phi^{\frac{1}{2}}(u)}{B(u)} du \leq \int_{-\pi/T}^{\pi/T} \frac{\Phi^{\frac{1}{2}}(u)}{B(u)} du \quad (47)$$

where $\Phi(u)$ and $B(u)$ are the composites of $\Phi_k(u)$ and $B_k(u)$ in the (generally) disjoint regions L_k in ω . The total interval over which the L_k 's extend is $-\pi/T < u < \pi/T$. For the case shown in Fig. 4 one would have

$$\frac{\Phi^{\frac{1}{2}}(u)}{B(u)} = \begin{cases} \Phi_1^{\frac{1}{2}}(u)/B_1(u) & -\pi/T < u \leq -a, \\ \Phi_0^{\frac{1}{2}}(u)/B_0(u) & -a \leq u \leq a, \\ \Phi_{-1}^{\frac{1}{2}}(u)/B_{-1}(u) & a \leq u < \pi/T. \end{cases}$$

As a specific example, consider correlated data which is formed by adding binary $(-1, 1)$ bits separated by T seconds (i.e., adding continuously the present bit and the immediately preceding bit). Such data (duobinary) has three levels $(-2, 0, 2)$ and

$$F(u) = 2 + 2 \cos uT.$$

It is easily shown that if

$$\frac{\Phi^{\dagger}(u)}{B(u)} = d_0 - d_1 \cos \frac{uT}{2}$$

where

$$d_0 \geq d_1 \geq \frac{8 - 2\pi}{2\pi - 4} d_0 \approx 0.75 d_0$$

(47) is satisfied. This means that for the above characteristic the optimum duobinary transmission system operates with greater output S/N than the best binary system. It is obvious that for

$$\frac{\Phi^{\dagger}(u)}{B(u)}$$

constant, (the usual case considered) binary transmission is optimum (with duobinary poorer by a factor of $\pi/4$ as pointed out by Bennett¹¹).

This example is not intended to exhaust the possibilities but to point out that for deviations from flat channels and white noise there is the possibility that signaling with correlated multilevel data is superior to binary.

VI. CONCLUDING REMARKS

The joint receiver and signal optimization has been carried out for the general data transmission system. The optimization considers simultaneously the noise and intersymbol distortion. The results may be applied to bandpass (baseband transmission, not carrier) or gradual cutoff systems as well as the sharp cutoff systems illustrated in the examples. Unfortunately, the joint solution is unrealizable, but it does provide a bound on the transmission performance with a fixed channel. Thus, with a given channel, one can do no better than (32) for the ratio of signal mean square error, and no better than (37) for the output signal-to-noise ratio with no intersymbol interference. From these equations, the importance of the ratio

$$\frac{\Phi^{\dagger}(\omega)}{B(\omega)}$$

is clearly seen. It is the only contribution to the bound other than the signal correlation function F , and the average signal power. Notice that the phase characteristic of the channel is irrelevant to the bound.

In the first example, it is seen that the original choice of signal and receiver was close to the optimum bound. Hence, there is little improvement to be gained with more complex processing. The second example illustrates a case where a more complex signal and receiver would be

theoretically useful. The signal and receiver of the first example would not effectively combat the noise in this case. The last example illustrates an advantage of correlative multilevel signaling in matching certain channel characteristics.

In the examples, the suboptimum (no intersymbol interference) solutions are presented in detail for three reasons:

(i) The solutions are of simpler form but just as illustrative as the optimum solutions.

(ii) The solutions depend only on the shape of $B(\omega)$ and $\Phi(\omega)$ and not upon the relative noise and signal powers.

(iii) The optimum and suboptimum solutions are nearly identical in the standard situation of large signal-to-noise power.

The explicit results for the suboptimum case are possible because of the explicit and complete statement of conditions for no intersymbol interference [Equations (6) and (7)].

APPENDIX A

Frequency Domain Representation of Mean Square Error

By writing the time domain samples, r_l , in terms of the Fourier transform of $r(t)$ one gets

$$r(t) = \int_{-\infty}^{\infty} R(\omega) e^{j\omega t} d\omega \quad (48a)$$

$$r_l = \int_{-\infty}^{\infty} R(\omega) e^{j\omega l T} d\omega \quad (48b)$$

$$r_l = \sum_{n=-\infty}^{\infty} \int_{(\pi/T)(2n-1)}^{(\pi/T)(2n+1)} R(\omega) e^{j\omega l T} d\omega \quad (48c)$$

$$= \sum_{n=-\infty}^{\infty} \int_{-\pi/T}^{\pi/T} R\left(u + \frac{2n\pi}{T}\right) e^{ju l T} du. \quad (48d)$$

Assuming that

$$\sum_n R\left(u + \frac{2n\pi}{T}\right) e^{ju l T}$$

is a uniformly convergent series

$$\left(|R(\omega)| \rightarrow \frac{1}{\omega^q}, \quad q \geq 2, \quad \text{as } \omega \rightarrow \infty \right)$$

$$r_l = \int_{-\pi/T}^{\pi/T} \sum_{n=-\infty}^{\infty} R\left(u + \frac{2n\pi}{T}\right) e^{ju l T} du. \quad (48e)$$

Noting that r_l is just the l th coefficient of an exponential Fourier series expansion of

$$\frac{2\pi}{T} \sum_{n=-\infty}^{\infty} R\left(u + \frac{2n\pi}{T}\right) \quad -\frac{\pi}{T} \leq u \leq \frac{\pi}{T},$$

one may write

$$\sum_{n=-\infty}^{\infty} R\left(u + \frac{2n\pi}{T}\right) = \frac{T}{2\pi} \sum_{l=-\infty}^{\infty} r_l e^{-julT} \quad (49a)$$

or

$$\sum_{n=-\infty}^{\infty} R\left(u + \frac{2n\pi}{T}\right) - \frac{Tr_0}{2\pi} = \frac{T}{2\pi} \sum_{l \neq 0} r_l e^{-julT} \quad (49b)$$

and

$$\left[\sum_{n=-\infty}^{\infty} R\left(u + \frac{2n\pi}{T}\right) - \frac{Tr_0}{2\pi} \right]^* e^{jumT} = \frac{T}{2\pi} \sum_{l \neq 0} r_l e^{julT} e^{jumT}. \quad (49c)$$

Multiplying (49b) and (49c) together and integrating one obtains, using the shorthand notation

$$R_n(u) = R\left(u + \frac{2n\pi}{T}\right),$$

$$\sum_{l \neq 0, -m} r_l r_{l+m} = \frac{2\pi}{T} \int_{-\pi/T}^{\pi/T} \left| \sum_n R_n(u) - \frac{r_0 T}{2\pi} \right|^2 \cos umT \, du. \quad (50)$$

Further,

$$\overline{a_j^2} \sum_{l \neq 0} r_l^2 + 2 \sum_{m=1}^{\infty} \overline{a_j a_{j+m}} \sum_{l \neq 0, -m} r_l r_{l+m}$$

$$= \frac{2\pi}{T} \int_{-\pi/T}^{\pi/T} \left| R_n(u) - \frac{r_0 T}{2\pi} \right|^2 F(u) \, du \quad (51)$$

where

$$F(u) = \overline{a_j^2} + 2 \sum_{m=1}^{\infty} \overline{a_j a_{j+m}} \cos umT. \quad (5)$$

The noise variance may be written

$$E\{n_0^2\} = \int_{-\infty}^{\infty} |E(\omega)|^2 \Phi(\omega) \, d\omega \quad (52a)$$

where $\Phi(\omega)$ is the channel noise power spectral density

$$E\{n_0^2\} = \sum_n \int_{\pi/T(2n-1)}^{\pi/T(2n+1)} |E(\omega)|^2 \Phi(\omega) \, d\omega \quad (52b)$$

$$= \sum_n \int_{-\pi/T}^{\pi/T} \left| E \left(u + \frac{2n\pi}{T} \right) \right|^2 \Phi \left(u + \frac{2n\pi}{T} \right) du \quad (52c)$$

$$= \int_{-\pi/T}^{\pi/T} \sum_n |E_n(u)|^2 \Phi_n(u) du \quad (52d)$$

if $\sum_n |E_n(u)|^2 \Phi_n(u)$ is a uniformly convergent series. The mean square error expression may then be written

$$M.S.E. = \int_{-\pi/T}^{\pi/T} \left\{ \sum_n |E_n(u)|^2 \Phi_n(u) + \frac{2\pi}{T} \left[\operatorname{Re} \sum_n R_n(u) - \frac{r_0 T}{2\pi} \right]^2 F(u) + \frac{2\pi}{T} \left[\operatorname{Im} \sum_n R_n(u) \right]^2 F(u) \right\} du. \quad (53)$$

In terms of the system component functions

$$S(\omega) = A(\omega)e^{j\alpha(\omega)}$$

$$T(\omega) = B(\omega)e^{j\beta(\omega)}$$

$$E(\omega) = C(\omega)e^{j\gamma(\omega)}$$

(53) may be written

$$M.S.E. = \int_{-\pi/T}^{\pi/T} \left\{ \sum_n C_n^2(u) \Phi_n(u) + \frac{2\pi}{T} \left[\sum_n A(u) B_n(u) C_n(u) \cos [\alpha_n(u) + \beta_n(u) + \gamma_n(u)] - \frac{r_0 T}{2\pi} \right]^2 F(u) + \frac{2\pi}{T} \left[\sum_n A_n(u) B_n(u) C_n(u) \sin [\alpha_n(u) + \beta_n(u) + \gamma_n(u)] \right]^2 F(u) \right\} du. \quad (4)$$

APPENDIX B

Simultaneous Reduction of Noise and Intersymbol Interference

The function

$$\int_{-\pi/T}^{\pi/T} \left[\sum_n C_n^2 \Phi_n + \frac{2\pi F}{T} \left\{ \sum_n A_n B_n C_n \cos (\alpha_n + \beta_n + \gamma_n) - \frac{r_0 T}{2\pi} \right\}^2 + \frac{2\pi F}{T} \left\{ \sum_n A_n B_n C_n \sin (\alpha_n + \beta_n + \gamma_n) \right\}^2 \right] du$$

$$\begin{aligned}
& + \lambda_1 \int_{-\pi/T}^{\pi/T} F \sum_n A_n^2 du \\
& + \lambda_2 \int_{-\pi/T}^{\pi/T} \sum_n A_n B_n C_n \cos(\alpha_n + \beta_n + \gamma_n) du
\end{aligned}$$

is to be minimized subject to the constraints

$$\int_{-\pi/T}^{\pi/T} F \sum_n A_n^2 du = P \quad (9)$$

and

$$\int_{-\pi/T}^{\pi/T} \sum_n A_n B_n C_n \cos(\alpha_n + \beta_n + \gamma_n) du = r_0. \quad (8)$$

For the optimum signal, this minimization is carried out with respect to A_k and α_k . Using the standard techniques of the calculus of variations⁹ the minimization with respect to α_k yields

$$\begin{aligned}
A_k B_k C_k \sin(\alpha_k + \beta_k + \gamma_k) & \left[\sum_n A_n B_n C_n \cos(\alpha_n + \beta_n + \gamma_n) \right. \\
& \left. - \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} \right) \right] - A_k B_k C_k \cos(\alpha_k + \beta_k + \gamma_k) \\
& \cdot \left[\sum_n A_n B_n C_n \sin(\alpha_n + \beta_n + \gamma_n) \right] = 0,
\end{aligned} \quad (54)$$

assuming that $F \neq 0$. The special case when $F = 0$ will be considered later. Summing over all k leads to

$$\left(2r_0 - \frac{\lambda_2}{F} \right) \sum_k A_k B_k C_k \sin(\alpha_k + \beta_k + \gamma_k) = 0 \quad (55)$$

since λ_2 depends upon the constraints

$$\sum_k A_k B_k C_k \sin(\alpha_k + \beta_k + \gamma_k) = 0 \quad (56)$$

and

$$\begin{aligned}
A_k B_k C_k \sin(\alpha_k + \beta_k + \gamma_k) & \left[\sum_n A_n B_n C_n \cos(\alpha_n + \beta_n + \gamma_n) \right. \\
& \left. - \frac{T}{4\pi} \left((2r_0 - \frac{\lambda_2}{F}) \right) \right] = 0
\end{aligned} \quad (57)$$

Minimizing with respect to A_k one gets

$$\begin{aligned} \frac{4\pi}{T} F[B_k C_k \cos(\alpha_k + \beta_k + \gamma_k)] & \left[\sum_n A_n B_n C_n \cos(\alpha_n + \beta_n \right. \\ & \left. + \gamma_n) - \frac{r_0 T}{2\pi} \right] + \frac{4\pi}{T} F[B_k C_k \sin(\alpha_k + \beta_k + \gamma_k)] \\ & \cdot \left[\sum_n A_n B_n C_n \sin(\alpha_n + \beta_n + \gamma_n) \right] \\ & + 2\lambda_1 F A_k + \lambda_2 B_k C_k \cos(\alpha_k + \beta_k + \gamma_k) = 0 \end{aligned} \quad (58)$$

or using (56)

$$\begin{aligned} [B_k C_k \cos(\alpha_k + \beta_k + \gamma_k)] & \left[\sum_n A_n B_n C_n \cos(\alpha_n + \beta_n + \gamma_n) \right. \\ & \left. - \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} \right) \right] + \frac{2\lambda_1 T}{4\pi} A_k = 0. \end{aligned} \quad (59)$$

Comparing (57) and (59) it is seen that a nontrivial solution requires

$$\sin(\alpha_k + \beta_k + \gamma_k) = 0 \quad (10)$$

and

$$A_k = -\frac{2\pi}{\lambda_1 T} B_k C_k \left\{ \sum_n A_n B_n C_n - \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} \right) \right\}. \quad (19)$$

Multiplying each side of the equation by $B_k C_k$ and summing one gets

$$\sum_k A_k B_k C_k = -\frac{2\pi}{\lambda_1 T} \sum_k B_k^2 C_k^2 \left\{ \sum_n A_n B_n C_n - \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} \right) \right\}. \quad (60)$$

$$\sum_k A_k B_k C_k = \frac{T}{4\pi} \frac{\left(2r_0 - \frac{\lambda_2}{F} \right) \sum_k B_k^2 C_k^2}{\sum_k B_k^2 C_k^2 + \frac{\lambda_1 T}{2\pi}} \quad (61)$$

and

$$\sum_k A_k B_k C_k - \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} \right) = -\frac{T}{4\pi} \left(\frac{\lambda_1 T}{2\pi} \right) \frac{\left(2r_0 - \frac{\lambda_2}{F} \right)}{\sum_k B_k^2 C_k^2 + \frac{\lambda_1 T}{2\pi}}. \quad (62)$$

Finally,

$$A_k = \frac{T}{4\pi} \frac{\left(2r_0 - \frac{\lambda_2}{F} \right)}{\sum_k B_k^2 C_k^2 + \frac{\lambda_1 T}{2\pi}} B_k C_k. \quad (11)$$

For the optimum receiver filter, the minimization with respect to γ_k again yields (56) and (57) and with respect to C_k leads to

$$\begin{aligned} \frac{4\pi}{T} F[A_k B_k \cos(\alpha_k + \beta_k + \gamma_k)] \left[\sum_n A_n B_n C_n \cos(\alpha_n + \beta_n + \gamma_n) \right. \\ \left. - \frac{r_0 T}{2\pi} \right] + \frac{4\pi}{T} F[A_k B_k \sin(\alpha_k + \beta_k + \gamma_k)] \\ \left[\sum_n A_n B_n C_n \sin(\alpha_n + \beta_n + \gamma_n) \right] \\ + 2C_k \Phi_k + \lambda_2 A_k B_k \cos(\alpha_k + \beta_k + \gamma_k) = 0 \end{aligned} \quad (63)$$

or using (56)

$$\begin{aligned} [A_k B_k \cos(\alpha_k + \beta_k + \gamma_k)] \left[\sum_n A_n B_n C_n \cos(\alpha_n + \beta_n + \gamma_n) \right. \\ \left. - \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} \right) \right] + \frac{2T}{4\pi} \frac{C_k \Phi_k}{F} = 0. \end{aligned} \quad (64)$$

Again, comparing this equation and (57) reveals that a nontrivial solution requires

$$\sin(\alpha_k + \beta_k + \gamma_k) = 0. \quad (10)$$

Then

$$C_k = -\frac{2\pi}{T} \frac{A_k B_k F}{\Phi_k} \left[\sum_n A_n B_n C_n - \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} \right) \right]. \quad (20)$$

Multiplying each side of the equation by $A_k B_k$ and summing one gets

$$\sum_k A_k B_k C_k = -\frac{2\pi}{T} F \sum_k \frac{A_k^2 B_k^2}{\Phi_k} \left[\sum_n A_n B_n C_n - \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} \right) \right] \quad (65)$$

$$\sum_k A_k B_k C_k = \frac{T}{4\pi} \frac{\left(2r_0 - \frac{\lambda_2}{F} \right) F \sum_k \frac{A_k^2 B_k^2}{\Phi_k}}{F \sum_k \frac{A_k^2 B_k^2}{\Phi_k} + \frac{T}{2\pi}} \quad (66)$$

and

$$\sum_k A_k B_k C_k - \frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F} \right) = -\frac{T}{4\pi} \left(\frac{T}{2\pi} \right) \frac{\left(2r_0 - \frac{\lambda_2}{F} \right)}{F \sum_k \frac{A_k^2 B_k^2}{\Phi_k} + \frac{T}{2\pi}}. \quad (67)$$

Finally,

$$C_k = \frac{T}{4\pi} \frac{(2r_0 - \lambda_2/F) A_k B_k F}{F \sum_k \frac{A_k^2 B_k^2}{\Phi_k} + \frac{T}{2\pi}} \quad (12a)$$

The coefficient, $T/4\pi(2r_0 - \lambda_2/F)$ may be determined from the constraint

$$\int_{-\pi/T}^{\pi/T} \sum_k A_k B_k C_k du = r_0. \quad (8)$$

Multiplying both sides of (12a) by $A_k B_k$, summing, and integrating one obtains

$$r_0 = \frac{T}{4\pi} \int_{-\pi/T}^{\pi/T} \left(2r_0 - \frac{\lambda_2}{F}\right) \frac{F \sum_k \frac{A_k^2 B_k^2}{\Phi_k}}{F \sum_k \frac{A_k^2 B_k^2}{\Phi_k} + \frac{T}{2\pi}} du \quad (68)$$

or

$$\frac{T}{4\pi} \left(2r_0 - \frac{\lambda_2}{F}\right) = \frac{r_0 T}{2\pi} \frac{\left[\int_{-\pi/T}^{\pi/T} \sum_k \frac{A_k^2 B_k^2}{\Phi_k} du + \frac{1}{F} \int_{-\pi/T}^{\pi/T} \frac{T}{2\pi} du \right]}{\int_{-\pi/T}^{\pi/T} \frac{\sum_k \frac{A_k^2 B_k^2}{\Phi_k}}{F \sum_k \frac{A_k^2 B_k^2}{\Phi_k} + \frac{T}{2\pi}} du} \quad (12b)$$

The mean square error

$$\int_{-\pi/T}^{\pi/T} \left[\sum_k C_k^2 \Phi_k + \frac{2\pi F}{T} \left\{ \sum_k A_k B_k C_k - \frac{r_0 T}{2\pi} \right\}^2 \right] du$$

may be found using (12a,b) to be

$$\left(\frac{r_0 T}{2\pi} \right)^2 \frac{\left[\int_{-\pi/T}^{\pi/T} \frac{F du}{\Delta} \int_{-\pi/T}^{\pi/T} \frac{\sum_k \frac{A_k^2 B_k^2}{\Phi_k} du}{\Delta} + \frac{T}{2\pi} \left(\int_{-\pi/T}^{\pi/T} \frac{du}{\Delta} \right)^2 \right]}{\int_{-\pi/T}^{\pi/T} \frac{\sum_k \frac{A_k^2 B_k^2}{\Phi_k} du}{\Delta}}$$

where

$$\Delta = F \sum_k \frac{A_k^2 B_k^2}{\Phi_k} + \frac{T}{2\pi}.$$

REFERENCES

1. Nyquist, H., Certain Topics in Telegraph Transmission Theory, AIEE Trans., 47, April, 1928, pp. 617-644.
2. Bennett, W. R. and Davey, J. R., *Data Transmission*, McGraw-Hill Book Co., New York, 1965.
3. Tufts, D. W., Matched Filters and Intersymbol Interference, Cruft Laboratory, Harvard University, Cambridge, Massachusetts, Tech. Rept. No. 345, July, 1961.
4. Tufts, D. W., Nyquist's Problem in Pulse Transmission Theory, Cruft Laboratory, Harvard University, Cambridge, Massachusetts, Tech. Rept. No. 425, Sept., 1963.
5. Tufts, D. W., Certain Results in Pulse Transmission Theory, Cruft Laboratory, Harvard University, Cambridge, Massachusetts, Tech. Rept. No. 335, Feb., 1962.
6. Tufts, D. W., Nyquist's Problem—The Joint Optimization of Transmitter and Receiver in Pulse Amplitude Modulation, Proc. IEEE, 53, March, 1965, pp. 248-259.
7. Gibby, R. A. and Smith, J. W., Some Extensions of Nyquist's Telegraph Transmission Theory, B.S.T.J., 44, Sept., 1965, pp. 1487-1510.
8. Bennett, W. R., Statistics of Regenerative Digital Transmission, B.S.T.J., 37, Nov., 1958, pp. 1501-1542.
9. Courant, R. and Hilbert, D., *Methods of Mathematical Physics, 1*, Interscience Publishers, New York, 1953.
10. Lender, A., The Duobinary Technique For High Speed Data Transmission, IEEE Trans. on Communication and Electronics, 82, May, 1963, pp. 214-218.
11. Bennett, W. R., unpublished memorandum.

Some Effects of Laminar and Turbulent Flow on Breakdown in Gases*

By D. S. BUGNOLO

(Manuscript received July 12, 1965)

The conservation equation for electrons in a laminar or turbulent flow has been used to determine a criterion for breakdown. The results can be used to define a characteristic length, L_s , and time, t_s , which determine the effects of the flow on the power required to break down the gas. The theory has been compared to the experimental results of Buchsbaum and Cottingham in hydrogen with reasonable agreement in the laminar region. In the presence of turbulence, the results will depend on the velocity dependence of the turbulent diffusion coefficient for the electrons.

I. INTRODUCTION

The physics governing the breakdown of a stationary gas (no flow) has been discussed at great length by a number of authors. For some recent studies the interested reader is referred to a text by Brown¹ and papers by Buchsbaum² and by McDonald.³ The effects of gas flow on breakdown has been studied experimentally for a laminar flow of less than 100 meters per second by Skinner and Brady⁵ and for the case of laminar and turbulent flows by Buchsbaum and Cottingham.⁴

Theoretically, the effects of gas flow can be studied by noting that the transport of electrons and ions will depend on the flow as well as the laminar or turbulent diffusion. If the flow is laminar and the electron density such that the diffusion is ambipolar, then the average drift velocity of the electrons and ions are equal and given by

$$\bar{v} = -D_a (\nabla n/n) + \mathbf{V} \quad (1)$$

where D_a is the ambipolar diffusion coefficient, n , the electron density and \mathbf{V} , the flow velocity. The ambipolar diffusion coefficient may, in

* The study was supported by the U. S. Army Nike X Project Office, Redstone, Alabama.

turn, be related to the diffusion coefficient for the ions, D_i , and the temperature of the ions, T_i , and electrons, T_e , by

$$D_a = D_i \{1 + T_e/T_i\} \quad (2)$$

provided the mobility of the electrons is much larger than that of the ions.

If the flow is turbulent, then the ambipolar diffusion coefficient must be replaced by the turbulent diffusion coefficient, D_T .

The theory to follow will contain ratios of the form

$$\bar{V}/2D \quad \text{and} \quad \bar{V}^2/2D,$$

where \bar{V} is the average velocity in a turbulent flow or simply the velocity in a laminar flow. The first can be used to define an inverse length, L_s , such that

$$\bar{V}/2D \equiv L_s^{-1}, \quad (3)$$

and the second an inverse time, t_s , such that

$$\bar{V}^2/2D \equiv t_s^{-1}. \quad (4)$$

The characteristic time, t_s , and length, L_s , are a measure of the extent to which electrons are removed by the flow. These may be compared to the removal of electrons by ordinary diffusion to the walls with the characteristic time, t_D , given by

$$D/\Lambda^2 \equiv t_D^{-1}, \quad (5)$$

where Λ is the size of the container. If

$$t_s \gg t_D,$$

then it is reasonable to expect that the effects of the flow on breakdown are negligible. If, in turn,

$$t_D \gg t_s,$$

it is reasonable to expect that the effect of ordinary diffusion to the walls is negligible as compared to "sweeping" as an electron removal process. In this case, breakdown will be controlled by the flow.

II. GENERAL THEORY

The effects of gas flow on breakdown can be considered theoretically by noting that the transport of electrons, on the average, is due to the

mean flow as well as the laminar or the turbulent diffusion. In the absence of secondary emission from the walls, the appropriate conservation equation for the electrons is given by⁶

$$\frac{\partial \bar{n}}{\partial t} + \nabla \cdot (\bar{n} \bar{\mathbf{V}}) \cong (\nu_i - \nu_a) \bar{n} + \alpha \bar{n}^2 + \nabla \cdot (D \nabla \bar{n}) \quad (6)$$

where

\bar{n} = the electron density in a laminar flow or the ensemble average of the electron density (at any given time) in a turbulent flow.

$\bar{\mathbf{V}}$ = the velocity in a laminar flow or the ensemble average of the gas velocity in a turbulent flow. (Experimental conditions usually justify a time average in this case.)

ν_i = the ionization frequency

ν_a = the attachment frequency

α = the recombination coefficient

D = the diffusion coefficient for the electrons (laminar or turbulent).

Consider the geometry of Fig. 1. The originally ambient gas is flowing between two parallel planes separated by a distance d . Two grids are placed normal to the direction of flow and separated by a distance l . For this geometry, (6) reduces to,

$$\frac{\partial \bar{n}}{\partial t} + \bar{n} \nabla \cdot \bar{\mathbf{V}} + \bar{V}_y \frac{\partial \bar{n}}{\partial y} = (\nu_i - \nu_a) \bar{n} + \alpha \bar{n}^2 + D \left\{ \frac{\partial^2 \bar{n}}{\partial y^2} + \frac{\partial^2 \bar{n}}{\partial z^2} \right\}. \quad (7)$$

If it is further assumed that the distance l in the direction of flow is not excessive, such that,

$$\frac{\partial \bar{V}_y}{\partial y} = 0 \quad \text{for } 0 \leq y \leq l,$$

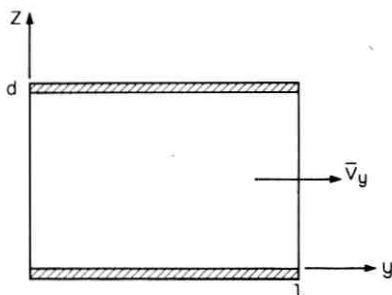


Fig. 1 — Cross section of flow geometry.

$$t_s = \frac{2}{3} \left[\frac{\langle v_i \rangle^2}{V_y^2} \right] \{1 + T_e/T_i\} \frac{1}{v_{ci}} \quad (17)$$

or by

$$t_s = \frac{2}{3} \left[\frac{\langle v_i \rangle}{V_y} \right] \{1 + T_e/T_i\} \frac{l_i}{V_y}. \quad (18)$$

As expected, both L_s and t_s approach infinity as the flow velocity, V_y , approaches zero.

The magnitude of the characteristic sweeping length, L_s , must be compared to the length of the discharge, l , in the direction of flow, y . If $L_s \gg l$, then the term $(V_y/2D)y$, may be neglected.

The magnitude of the characteristic sweeping time, t_s , must be compared to the characteristic time for electron loss by diffusion to the walls, t_D . If $t_D \gg t_s$, then diffusion to the walls may be neglected as an electron removal process. If $t_s \gg t_D$, then sweeping may be neglected. It is evident that since

$$t_s \sim V_y^{-2} \quad (19)$$

then, $t_s \gg t_D$, for low flow velocities. As expected, sweeping may be neglected as an electron removal process at low velocities.

2.2 Turbulent Flow

When the flow of the gas is turbulent, it is only reasonable to expect that the diffusion coefficient, D , will increase provided the electron density is sufficient to insure ambipolar or near ambipolar diffusion in the absence of turbulence. While this may be confusing, it is of importance to note that electrons in *free* diffusion are probably unaffected by turbulence.

As in the case of laminar flow, it is possible to define a characteristic length, L_s , and time, t_s , for the sweeping due to the mean flow. Using (12), it follows that

$$L_s^{-1} \equiv \bar{V}_y/2D_{T_e}, \quad t_s^{-1} \equiv \bar{V}_y^2/2D_{T_e}, \quad (20)$$

where D_{T_e} is the turbulent diffusion coefficient for the electrons. It can be shown that

$$D_{T_e} = D_{i_T} \{1 + T_e/T_i\}, \quad (21)$$

where D_{i_T} is the turbulent diffusion coefficient for the ions. D_{T_e} will simply be written as D_T for the remainder of this paper.

The characteristic time for electron loss by diffusion to the walls will also be modified by the turbulent flow. It follows that

$$t_{D_{mn}}^{-1} = D_T / \Lambda_{mn}^2. \quad (22)$$

III. COMPARISON WITH EXPERIMENT

While the theoretical geometry is not duplicated by any existing experiment, some verification of the theory can be obtained from the results of Buchsbaum and Cottingham⁴ in hydrogen. Their experimental set-up has been sketched in Fig. 2. The results for breakdown in hydrogen (H_2) has been plotted as a function of gas velocity in Fig. 3. The peak power required to produce breakdown varied from about 1.7 to 3.2 kw.

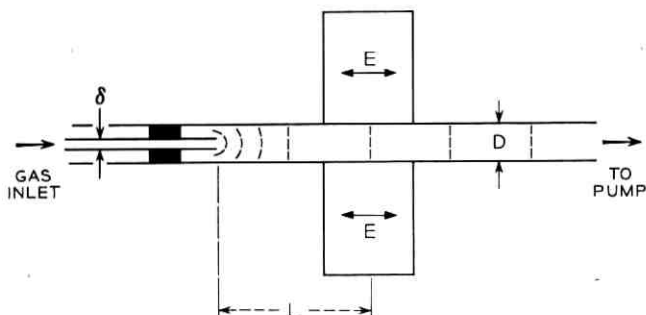


Fig. 2—Geometry of the experiment of Buchsbaum and Cottingham (Ref. 4).

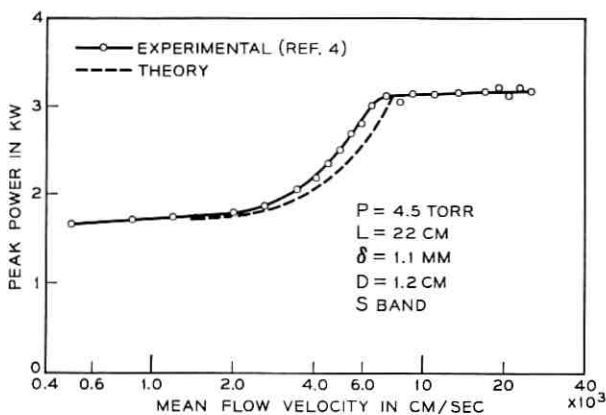


Fig. 3—Comparison of the theoretical and experimental results for peak power as a function of flow velocity in hydrogen (experimental results of Ref. 4).

These results were used to calculate the effective electric field intensity required to produce breakdown as a function of gas velocity. In this calculation, the effective frequency for electron-molecular elastic collisions was taken as²

$$\nu_c = 4.8 \times 10^{-9} p, \text{ sec}^{-1}. \quad (23)$$

An approximate criteria for breakdown can be obtained from the theory as given by (11). (See Appendix B.) It can be shown that

$$\nu_i t_{\text{on}} = \left[\frac{\bar{V}^2}{2D} + \frac{D}{\Lambda^2} \right] t_{\text{off}}, \quad (25)$$

where ν_i in H_2 has been measured and compared to dc data.² For this particular experiment,⁴ the geometry was cylindrical with

$$t_{\text{on}} = 1 \times 10^{-6} \text{ sec}, \quad t_{\text{off}} = 1 \times 10^{-3} \text{ sec},$$

and

$$\frac{1}{\Lambda^2} = \left(\frac{\pi}{3.8} \right)^2 + \left(\frac{2.405}{0.6} \right)^2 = 16.7$$

or

$$\Lambda = 0.245 \text{ cm}.$$

It should be noted that the length of the discharge in the direction of flow has been taken as 3.8 cm even though grids were not used (at $y = 0$ and $y = 3.8$ cm) in the experiment.

3.1 Laminar Flow

When the flow of the gas is laminar, the diffusion coefficient for the electrons should be independent of the velocity of the flow. The magnitude of the effective diffusion coefficient, D , can be calculated from the experimental data by noting that, for small \bar{V} ,

$$D/\Lambda^2 \gg \bar{V}^2/2D \quad \text{for} \quad \bar{V} = 500 \text{ cm sec}^{-1}. \quad (25)$$

(This is the smallest velocity for which data is available in this experiment. See Fig. 3.) This, in turn, implies that the characteristic time for electron removal by diffusion to the walls is much less than the characteristic time associated with the sweeping effect at this velocity of flow. For the laminar case, t_s is given by (17) or (18) provided the diffusion is ambipolar or nearly so.

If the electron density, during the pulse off time, is reduced to a point where the electrons are in free diffusion, then $D \rightarrow D_-$, where²

$$D_p \cong 1 \times 10^6 \text{ cm}^2 \text{ sec}^{-1} \text{ torr.}$$

As the electron density increases, D decreases. The resulting effective diffusion coefficient, D_s , has been calculated by Allis and Rose.⁷ Using their results and the geometry of this experiment, it follows that

$$D_s \cong 0.2D_- \quad \text{for } \bar{n} = 10^6 \text{ elec/cm}^3,$$

and

$$D_s \cong 0.04 D_- \quad \text{for } \bar{n} = 10^8 \text{ elec/cm}^3.$$

Returning to the question posed by (44) of Appendix B, we note that the ratio

$$\frac{\nu_i \Lambda^2}{D'} \cong 5 \quad \text{for } \bar{n} \text{ small,}$$

$$\frac{\nu_i \Lambda^2}{D'} \cong 25 \quad \text{for } \bar{n} \cong 10^6 \text{ elec/cm}^3,$$

$$\frac{\nu_i \Lambda^2}{D'} \cong 125 \quad \text{for } \bar{n} \cong 10^8 \text{ elec/cm}^3.$$

Consequently, diffusion can only be neglected during the pulse on time when the electron density is larger than 10^6 elec/cm^3 . This would appear to be the case during the $1\text{-}\mu\text{sec}$, 3-gc pulse used in this experiment.

Since the incident microwave pulse was relatively flat, (see Fig. 2 of Ref. 4), the peak power can be used to calculate the effective electric field strength and, in turn, the ionization frequency. For a flow velocity of 500 cm sec^{-1} , the peak power required to produce breakdown at the end of one microsecond was 1.7×10^3 watts. This power level corresponds to an effective field strength of about 200 volts per cm or an ionization frequency of 1.5×10^7 radians per sec. Using (24), it follows that

$$D/\Lambda^2 \cong 1.5 \times 10^4 \text{ sec}^{-1}. \quad (26)$$

This corresponds to an effective diffusion coefficient for the electrons of $920 \text{ cm}^2 \text{ sec}^{-1}$ or

$$D_p \cong 4 \times 10^3 \text{ cm}^2 \text{ sec}^{-1} - \text{ torr.} \quad (27)$$

Using this value for D , it follows that (25) is satisfied at 500 cm sec^{-1} , since

$$\bar{V}^2/2D \cong 120 \text{ sec}^{-1}.$$

The magnitude of the effective diffusion coefficient for the electrons can be compared to the ambipolar diffusion coefficient for hydrogen as

measured by Persson and Brown⁸ and to that calculated from the mobility measurements of Rose.⁹ The results of Persson and Brown,⁸ derived from afterglow data yield an ambipolar diffusion coefficient of

$$D_{ap} = 700 \pm 50 \text{ cm}^2 \text{ sec}^{-1} - \text{torr.} \quad (28)$$

This is a factor of six less than the effective diffusion coefficient as given by (27). The difference could be due to the fact that the diffusion in the afterglow was not ambipolar for all t , or to the presence of higher order modes, or to an elevated temperature for the electrons. Since the collision frequency was on the order of 10^9 radians per sec, it would appear that the plasma was rapidly reduced to thermal equilibrium during the pulse off time of one millisecond. It would appear reasonable to conclude that either or both of the first two possibilities are likely.

The mobility data of Rose⁹ yields a value for D_{ap} which is less than that measured by Persson and Brown⁸ in the afterglow.

At higher flow velocities, the effect of the flow can no longer be neglected. Equation (24) can be used together with (23, 26, 27) and Fig. 3 of Ref. 2 to calculate the peak power required to produce breakdown in one microsecond. The results have been plotted in Fig. 3 for flow velocities of less than 8000 cm/sec, together with the experimental results of Buchsbaum and Cottingham.⁴

While the agreement between theory and experiment is excellent, it should be noted that (24) is based on the condition

$$v_i \gg \frac{D'}{\Lambda^2} + \frac{\bar{V}^2}{2D'}, \quad \text{pulse on.}$$

From the results, it is evident that the condition is reasonable provided the electron density is such that the diffusion of electrons is never free, or provided any free diffusion mode is restricted to a negligible part of the pulse on time.

IV. CONCLUSIONS AND RECOMMENDATIONS

The conservation equation for the electron density in a laminar or turbulent flow (2) has been solved for the parallel plane geometry of Fig. 1 (11). Extension to other geometries is straightforward.

For cases where

$$v = v_i - v_a \gg D/\Lambda^2, \quad (29)$$

it was shown that pulsed breakdown in a laminar or turbulent gas may be controlled by the rather simple criterion,

$$\nu \times t_{\text{on}} = \left\{ \frac{\bar{V}^2}{2D} + \frac{D}{\Lambda^2} \right\} t_{\text{off}}, \quad (30)$$

where \bar{V} is the mean velocity of the flow.

The power required to produce breakdown in a turbulent gas may, or may not, be a function of the mean velocity of the flow. A velocity independent result is obtained when

$$D_T/\Lambda^2 \gg \bar{V}^2/2D_T \quad (31)$$

provided $D_T \sim \bar{V}l_e$, where l_e is the mixing length for the plasma, and provided l_e is inversely proportional to the Reynold's number of the flow.

In order to test the present theory further, future experiments should be designed such that

$$\nu_i \gg D/\Lambda \quad \text{for all } t.$$

It would be of interest to measure the electron density in the discharge at a number of positions in the direction of flow. This result could be used to check the predicted theoretical variation as well as insure the presence or absence of turbulence.

V. ACKNOWLEDGMENTS

The author would like to thank S. J. Buchsbaum, L. C. Hebel, and V. L. Granatstein for their helpful comments.

APPENDIX A

Equation (5) can be solved by the method of separation of variables. Let

$$n(y, z, t) = Y(y) \cdot Z(z) \cdot T(t). \quad (32)$$

It follows that

$$\frac{1}{T} \frac{dT}{dt} + \bar{V}_y \frac{1}{Y} \frac{dY}{dy} = (\nu_i - \nu_a) + D \left\{ \frac{1}{Y} \frac{d^2Y}{dy^2} + \frac{1}{Z} \frac{d^2Z}{dz^2} \right\}, \quad (33)$$

or

$$\frac{1}{DT} \frac{dT}{dt} = \frac{1}{D} (\nu_i - \nu_a) + \frac{1}{Y} \left\{ \frac{d^2Y}{dy^2} - \frac{\bar{V}_y}{D} \frac{dY}{dy} \right\} + \frac{1}{Z} \frac{d^2Z}{dz^2}. \quad (34)$$

Taking

$$\frac{1}{Z} \frac{d^2Z}{dz^2} = -\alpha^2$$

and using the appropriate boundary condition yields

$$Z(z) = A_m \sin \frac{m\pi}{d} z \quad (35)$$

where

$$\alpha^2 = \left(\frac{m\pi}{d}\right)^2.$$

Taking

$$\frac{1}{Y} \left\{ \frac{d^2 Y}{dy^2} - \frac{\bar{V}_y}{D} \frac{dY}{dy} \right\} = -\beta^2,$$

yields

$$\frac{d^2 Y}{dy^2} - \frac{\bar{V}_y}{D} \frac{dY}{dy} + \beta^2 Y = 0$$

with the solution $Y(y) = e^{+uy}$, where

$$u^2 - \frac{\bar{V}_y}{D} u + \beta^2 u = 0.$$

Hence,

$$u = \frac{\bar{V}_y}{2D} \pm \frac{1}{2} \sqrt{\left(\frac{\bar{V}_y}{D}\right)^2 - 4\beta^2}.$$

If the solution is to satisfy the boundary conditions at $y = 0$ and $y = l$, it follows that $Y(y)$ must be periodic in y ; hence, $4\beta^2 > (\bar{V}_y/D)^2$. Consequently,

$$u = \frac{\bar{V}_y}{2D} \pm j \sqrt{\beta^2 - \left(\frac{\bar{V}_y}{2D}\right)^2}$$

with the solution

$$Y(y) = B_n e^{(\bar{V}_y/2D)y} \sin \frac{n\pi}{l} y \quad (36)$$

where

$$\beta^2 - \left(\frac{\bar{V}_y}{2D}\right)^2 = \left(\frac{n\pi}{l}\right)^2.$$

Finally, the differential equation in t , reduces to

$$\frac{1}{T} \frac{dT}{dt} = (v_i - v_a) - D \left\{ \left(\frac{m\pi}{d}\right)^2 + \left(\frac{n\pi}{l}\right)^2 + \left(\frac{\bar{V}_y}{2D}\right)^2 \right\},$$

with the solution

$$T(t) = \exp \left\{ (\nu_i - \nu_a) - D \left[\left(\frac{m\pi}{d} \right)^2 + \left(\frac{n\pi}{l} \right)^2 + \left(\frac{\bar{V}_y}{2D} \right)^2 \right] \right\} t. \quad (37)$$

It follows that the solution of (5) satisfying the boundary conditions given in (4) is

$$n(x, y, z, t) = \epsilon^{(V_y/2D)y} \sum_m \sum_n A_{mn} \sin \frac{n\pi}{l} y \sin \frac{m\pi}{d} z \\ \times \exp \left\{ (\nu_i - \nu_a) - D \left[\left(\frac{m\pi}{d} \right)^2 + \left(\frac{n\pi}{l} \right)^2 + \left(\frac{\bar{V}_y}{2D} \right)^2 \right] \right\} t. \quad (38)$$

It is apparent at the offset that the usual method for evaluating the constant A_{mn} cannot be used for the general case since

$$\int_0^l \sin \frac{m\pi}{l} y \sin \frac{n\pi}{l} y \epsilon^{(V_y/2D)y} dy \quad (39)$$

does not vanish for $m \neq n$. In view of this, it is apparent that the usual modes *do not exist*.

Fortunately, a separation into modes is still possible, provided the electron density, at $t = 0$, is given by

$$f(x, y, z) \epsilon^{(V_y/2D)y}. \quad (40)$$

This would appear to be reasonable for the steady state, pulse on pulse off, conditions of the usual pulsed experiment.

Let us assume that the density at $t = 0$, (end of the off cycle) is given by

$$n(x, y, z, 0) = n_0 \times \epsilon^{(V_y/2D)y}. \quad (41)$$

For this case,

$$A_{mn} = n_0 \frac{4}{\pi^2} \frac{1}{mn} [1 - \cos n\pi][1 - \cos m\pi] \quad (42)$$

from which it is evident that $A_{mn} = 0$ for m or n even. It is of interest that the 1st "mode" for this case is

$$n(x, y, z, t) = n_0 \frac{16}{\pi^2} \sin \frac{\pi}{l} y \sin \frac{\pi}{d} z \epsilon^{(\bar{V}_y/2D)y} \\ \times \exp \left\{ (\nu_i - \nu_a) - D \left[\left(\frac{\pi}{d} \right)^2 + \left(\frac{\pi}{l} \right)^2 + \left(\frac{\bar{V}_y}{2D} \right)^2 \right] \right\} t. \quad (43)$$

In practice, the ratio $\bar{V}_y/2D$ at any velocity will depend on the particular gas used.

For example, in the experiment of Ref. 1,

$$(\bar{V}/2D)y < 11 \quad \text{for} \quad 0 \leq y \leq 3.5 \text{ cm}$$

in hydrogen, for $\bar{V} < 7 \times 10^3$ cm/sec, the laminar region. In the turbulent region, the ratio decreases since D increases.

APPENDIX B

The direct application of the theory to any particular experiment may be complicated by the variation of the diffusion coefficient with electron density and, in turn, time. The interested reader is referred to a recent paper by Buchsbaum and Cottingham² for a discussion of this problem. Following their example, the effects of electron attachment in H_2 will be neglected.

Let the flow conditions be such that the initial density, n_0 , is sufficient to insure ambipolar diffusion or near ambipolar diffusion of the electrons for all or nearly all time. This assumption will be justified for the particular experiment.

During the pulse on time the appropriate equation for the electron density is given by (43) with $D = D'$ (hot electrons) and $\nu_a = 0$. Let D' and \bar{V} be such that

$$\nu_i \gg \frac{D'}{\Lambda^2} + \frac{\bar{V}^2}{2D'} \quad \text{for all} \quad 0 \leq t \leq t_{on}. \quad (44)$$

This implies that

$$\nu_i \gg t_D'^{-1} \quad \text{and} \quad t_s'^{-1}$$

for all flow velocities during the pulse on time. It follows that (43) for the first "mode" reduces to

$$n(x,y,z,t) \cong n_0 \cdot \frac{16}{\pi^2} \sin \frac{\pi}{l} y \sin \frac{\pi}{d} z \epsilon^{(\bar{V}y/2D)y} \epsilon^{\nu_i t}. \quad (45)$$

This result can be applied directly to the experiment of Buchsbaum and Cottingham.⁴

In this experiment, electron density was monitored indirectly by observing the intensity of the power reflected from the discharge region. Since the electron density and, in turn, the reflection coefficient was a function of position within the discharge it is evident that the experimental results can be related to (45) by choosing (x,y,z) such that the reflection coefficient is a maximum. Hence,

$$n(t) \cong n_0^* \exp(\nu_i t), \quad \text{for} \quad 0 \leq t \leq t_{on}. \quad (46)$$

The appropriate equation for the pulse off time may be derived from (43) in a similar manner. It can be shown that

$$n(t) = n_{\max}^* \exp - \left\{ \frac{D}{\Lambda^2} + \frac{\bar{V}^2}{2D} \right\} t \quad \text{for } 0 \leq t \leq t_{\text{off}}. \quad (47)$$

Equations (46) and (47) can be solved to yield the approximate criteria for breakdown.

$$\nu_i t_{\text{on}} = \left\{ \frac{D}{\Lambda^2} + \frac{\bar{V}^2}{2D} \right\} t_{\text{off}}. \quad (48)$$

LIST OF SYMBOLS

- d = distance between the parallel planes in the z direction.
 l = distance between the grids in the direction of flow.
 l_e = effective mixing length for the electrons in a turbulent flow.
 l_i = mean free path for the ions.
 \bar{n} = electron density in a laminar flow or the ensemble average of the electron density (at a given time) in a turbulent flow.
 t_D = characteristic time for electron removal by diffusion to the walls.
 t_S = characteristic time for electron removal by the flow of the gas.
 \bar{v} = average velocity of the electrons or ions resulting from ambipolar diffusion and the mean flow.
 $\langle v_i \rangle$ = mean thermal velocity of the ions.
 D_a = coefficient of ambipolar diffusion for the electrons in a laminar flow.
 D_i = diffusion coefficient for the ions in a laminar flow.
 D_{iT} = diffusion coefficient for the ions in a turbulent flow.
 D_T = diffusion coefficient for the electrons in a turbulent flow.
 L_S = characteristic "sweeping" length for electron removal by the flow of the gas.
 R_e = Reynold's number of the flow.
 T_e = temperature of the electrons.
 T_i = temperature of the ions.
 \bar{V} = velocity of the gas in a laminar flow or the ensemble average of the gas velocity (at a given time) in a turbulent flow. Experimental conditions usually permit a time average in this case.
 α = recombination coefficient for the electrons.
 Λ_{mn} = effective diffusion distance for the mn mode.
 ν_a = attachment frequency for the electrons to H^+ and H_2^+ .
 ν_c = elastic collision frequency for the electrons.
 ν_{c_i} = elastic collision frequency for the ions.
 ν_i = ionization frequency for the electrons.

REFERENCES

1. Brown, S. C., *Basic Data of Plasma Physics*, M.I.T. Press, 1959.
2. Buchsbaum, S. J. and Cottingham, W. B., Electron Ionization Frequency in Hydrogen, *Phys. Rev.*, *130*, May, 1963, pp. 1002-1006.
3. McDonald, A. D., Gaskell, D. V., and Gitterman, H. N., Microwave Breakdown in Air, Oxygen and Nitrogen, *Phys. Rev.*, *130*, June, 1963, pp. 1841-1850.
4. Buchsbaum, S. J. and Cottingham, W. B., Diffusion in a Microwave Plasma in the Presence of a Turbulent Flow, presented at Buffalo meeting of the American Physical Society, Buffalo, New York, June 24-26, 1963. *J. Appl. Phys.*, *36*, June, 1965, pp. 2075-2078.
5. Skinner, J. G. and Brady, J. J., Effect of Gas Flow on the Microwave Dielectric Breakdown of Oxygen, *J. Appl. Phys.*, *34* (Part 1), April, 1963, pp. 975-978.
6. Bugnolo, D. S., Homogeneous Anisotropic Turbulence in a Weakly Ionized Gas, *J. Geophys. Res.*, *70*, August 1, 1965, pp. 3721-3724.
7. Allis, W. P. and Rose, D. J., Transition from Free to Ambipolar Diffusion, *Phys. Rev.*, *93*, January, 1954, pp. 84-93.
8. Persson, K. B. and Brown, S. C., Electron Loss Process in the Hydrogen Afterglow, *Phys. Rev.*, *100*, October, 1955, pp. 729-733.
9. Rose, D. J., Mobility of Hydrogen and Deuterium Positive Ions in their Parent Gas, *J. Appl. Phys.*, *31*, April, 1960, pp. 643-645.
10. Townsend, A. A., *The Structure of Turbulent Shear Flow*, Cambridge University Press, 1956.
11. Hinze, J. O., *Turbulence*, McGraw-Hill Book Company, Inc., 1959, pp. 285-286.

Transient Motion of Circular Elastic Plates Subjected to Impulsive and Moving Loads

By R. S. WEINER

(Manuscript received July 30, 1965)

Forced transient motions of peripherally supported, circular, elastic plates are analyzed according to the classical plate theory. The Green's function for the plate is developed and used to construct solutions for concentrated impulsive loadings, suddenly applied loadings, and moving pressure-wave loadings. The boundary of the plate is considered to be elastically built-in in a manner that prevents transverse edge motion and provides a restoring edge moment linearly related to edge rotation. Thus, limiting cases include a clamped plate and a simply supported plate. Numerical results are included to illustrate the influence of structural and loading parameters on the dynamic response of the plate.

I. INTRODUCTION

During the past decade considerable effort has been channeled toward increasing the capability of equipment to sustain severe nuclear weapon environments. These efforts, commonly called "hardening," employ various combinations of analytical and experimental approaches, each approach having certain difficulties and shortcomings.

Problems that are analytically tractable are usually restricted to simple and regular geometries and usually incorporate simplifying approximations as to material properties, weapon phenomenon, and separation of the combined weapon effects into independent and separate effects. One of the areas of hardening which lends itself to both analytical and experimental treatment and for which some full-scale nuclear test data are available is the response of structures to an air-blast wave; the analysis of structures subjected to air-blast pressure waves is the subject of the present article. In particular, the transient displacements of circular elastic plates subjected to impulsive and moving loads will be analyzed according to the classical plate theory. The classical, or small

deflection, plate theory does not account for internal structural damping, effects of transverse shear, or rotatory inertia. Consequently, this analysis is strictly applicable when the plate is "thin" (small thickness to radius ratio) and when the higher modes of vibration are of secondary importance;* neglecting internal structural damping results in predicted deflections that are conservative in the sense that they will be larger than corresponding deflections with damping present.

Equations are derived and numerical results are presented for several elemental loadings and for pressure loading waves of constant and decaying magnitude that sweep across the plate with uniform speed. Sweeping pressure-wave loadings occur when the blast wave approaches the plate in other than a face-on direction.

The method of analysis is to construct the Green's function for the plate and then to use principles of superposition in synthesizing solutions for the various loadings of interest.

Previous analyses of circular elastic plate vibrations can be traced to the free vibration analyses of Poisson,¹ and of Kirchhoff,² who considered axisymmetric and nonaxisymmetric vibrations respectively. More recent studies of forced vibrations of circular elastic plates include the work of Flynn,³ Sneddon,⁴ Riesman,⁵ and the present writer.⁶ Mindlin⁷ discussed the effects of rotatory inertia and transverse shear deflections on the dynamic plate equation. Elastically restrained plates were investigated by Kantham⁸ and by Reid.⁹

II. FORMULATION

2.1 *Equation of Motion*

Forced transverse motions of a homogeneous, isotropic, elastic plate of constant thickness (thickness is restricted to be small in comparison with radius) are governed by the partial differential equation

$$D\nabla^4 w(r,\theta,t) + m \frac{\partial^2 w(r,\theta,t)}{\partial t^2} = p(r,\theta,t) \quad (1)$$

under the restriction that the deflections are small in comparison with the plate thickness, and that the shear deflections, rotatory inertia, and damping can be neglected. Fig. 1 defines the coordinate system and configuration that correspond to (1)

* For a complete discussion of the limitations of the classical plate equation, the reader is referred to Ref. 7.

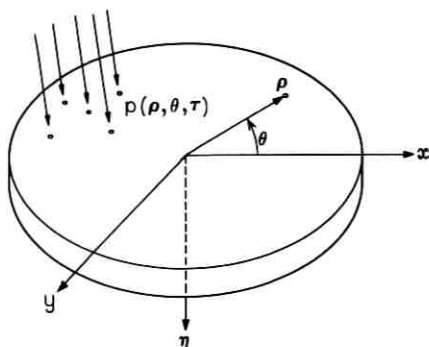


Fig. 1 — Circular plate configuration.

2.2 Boundary and Initial Conditions

A plate that is continuous at the origin (non-annular) and elastically built-in* along its periphery has the four boundary (or regularity) conditions:

$$\left. \begin{aligned}
 (a) \quad & w(0, \theta, t) \text{ must be finite} \\
 (b) \quad & \frac{\partial w(0, \theta, t)}{\partial r} \text{ must be finite} \\
 (c) \quad & w(a, \theta, t) = 0 \\
 (d) \quad & -D \left[\frac{\partial^2 w(a, \theta, t)}{\partial r^2} + \frac{\nu}{a} \frac{\partial w(a, \theta, t)}{\partial r} \right] = \beta \frac{\partial w(a, \theta, t)}{\partial r}
 \end{aligned} \right\} \quad (2)$$

Initial conditions that characterize an initially undeflected and stationary plate are

$$w(r, \theta, 0) = 0$$

and

$$\frac{\partial w(r, \theta, 0)}{\partial t} = 0. \quad (3)$$

These initial conditions are sufficiently general for the applications to

* Elastically built-in edge, as used in this article, refers to a boundary support that prevents transverse edge motion and provides a restoring edge-moment proportional to edge rotation. By properly selecting the constant of proportionality, special cases corresponding to a clamped edge and a simply-supported edge are obtained. A more detailed discussion of this boundary condition appears in Ref. 6.

be considered here; other initial conditions can be treated in a similar manner and will introduce minor changes in the equations developed herein.

III. BASIC SOLUTION

3.1 Dimensionless System of Equations

Introducing the dimensionless quantities $\rho \equiv r/a$, $\eta \equiv w/a$, $\tau \equiv t/T$, $\alpha = ma^4/DT^2$, and $\gamma \equiv \nu + (\beta a/D)$, (1) and (2) become, respectively,

$$\nabla^4 \eta(\rho, \theta, \tau) + \alpha \frac{\partial^2 \eta(\rho, \theta, \tau)}{\partial \tau^2} = \frac{a^3}{D} p(\rho, \theta, \tau) \quad (4)$$

and

$$\frac{\partial^2 \eta(1, \theta, \tau)}{\partial \rho^2} + \gamma \frac{\partial \eta(1, \theta, \tau)}{\partial \rho} = 0. \quad (5)$$

3.2 Homogeneous Equation

As a prelude to the solution of (4), the set of eigenfunctions for the plate will be determined. These eigenfunctions are obtained by starting with the homogeneous equation

$$\nabla^4 \eta(\rho, \theta, \tau) + \alpha \frac{\partial^2 \eta(\rho, \theta, \tau)}{\partial \tau^2} = 0. \quad (6)$$

Separable product solutions of such form that the angular functions and time functions possess the necessary periodicity suggest the following functions:

$$\eta(\rho, \theta, \tau) = R_{nj}(\rho) \begin{cases} \cos n\theta \\ \text{or} \\ \sin n\theta \end{cases} e^{i\omega_{nj}\tau}, \quad (7)$$

where $i = \sqrt{-1}$, and n and j are integers.

Substituting (7) into (6) produces an ordinary differential equation in ρ

$$\left(\frac{d^2}{d\rho^2} + \frac{1}{\rho} \frac{d}{d\rho} - \frac{n^2}{\rho^2} + \kappa_{nj}^2 \right) \left(\frac{d^2}{d\rho^2} + \frac{1}{\rho} \frac{d}{d\rho} - \frac{n^2}{\rho^2} - \kappa_{nj}^2 \right) R_{nj}(\rho) = 0, \quad (8)$$

where

$$\kappa_{nj}^2 = \alpha \frac{1}{2} \omega_{nj}.$$

Equation (8) and the boundary conditions (2) define radial eigenfunctions for the plate.

These radial eigenfunctions are readily found to be

$$R_{nj}(\rho) = I_n(\kappa_{nj})J_n(\kappa_{nj}\rho) - J_n(\kappa_{nj})I_n(\kappa_{nj}\rho), \quad (9)$$

where κ_{nj} is determined from the transcendental equation

$$\frac{2\kappa_{nj}}{1-\gamma} I_n(\kappa_{nj})J_n(\kappa_{nj}) = I_n(\kappa_{nj})J_{n+1}(\kappa_{nj}) + J_n(\kappa_{nj})I_{n+1}(\kappa_{nj}). \quad (10)$$

Two identifying subscripts are associated with κ_{nj} because there is a doubly-infinite set of eigenvalues and eigenfunctions. The first subscript, n , indicates the number of nodal diameters occurring in that mode of vibration and can be any positive integer (representing nonaxisymmetric modes) or zero (representing axisymmetric modes); the second subscript, j , indicates the number of nodal circles (including the boundary circle) and can be any positive integer. Eigenvalues for representative parameters are presented in Table I.

The radial eigenfunction $R_{nj}(\rho)$ as defined by (9) and (10) will be used in the next section as a building block for the development of the Green's function.

3.3 Green's Function

Next, the Green's function associated with (4) will be developed. Toward this end, consider the loading function

$$p(\rho, \theta, \tau) = \frac{1}{\rho_0} \delta(\rho - \rho_0) \delta(\theta) \delta(\tau - \tau_0). \quad (11)$$

The solution of (4) with the particular loading function (11) is defined as the Green's function and is denoted by $G(\rho, \theta, \tau; \rho_0, 0, \tau_0)$. Without loss of generality, G may be represented by the double series

$$G(\rho, \theta, \tau; \rho_0, 0, \tau_0) = \sum_{k=0}^{\infty} \sum_{j=1}^{\infty} R_{kj}(\rho) \cos k\theta g_{kj}(\tau), \quad (12)$$

where $g_{kj}(\tau)$ is an unknown function of time that must now be determined.

Introducing (12) into (4), and interchanging the order of differentiation and summation, the result is

$$\begin{aligned} \sum_k \sum_j \left(\frac{d^2}{d\rho^2} + \frac{1}{\rho} \frac{d}{d\rho} - \frac{k^2}{\rho^2} \right) R_{kj}(\rho) \cos k\theta g_{kj}(\tau) \\ + \alpha \sum_k \sum_j R_{kj}(\rho) \cos k\theta \ddot{g}_{kj}(\tau) = \frac{a^3}{D} \frac{\delta(\rho - \rho_0) \delta(\theta) \delta(\tau - \tau_0)}{\rho_0}. \end{aligned} \quad (13)$$

TABLE I—EIGENVALUES FOR CIRCULAR PLATES

 n = number of nodal diameters j = number of nodal circles

		$\gamma = 0$			$\gamma = 1.0$			
		0	1	2	3	0	1	2
1	2.1080	3.6744	5.0244	6.2931	2.4048	3.8317	5.1356	6.3802
2	5.4188	6.9380	8.3534	9.7044	5.5261	7.0156	8.4171	9.7585
3	8.5920	10.188	11.895	13.018	8.6568	10.173	11.937	13.042
4	11.747	13.285	15.013	16.214	11.792	13.324	15.047	16.238
5	14.896	16.439	18.139	19.397	14.931	16.471	18.167	19.418
6	18.043	19.590	21.270	22.569	18.071	19.616	21.294	22.588
7	21.187	22.738	24.403	25.734	21.212	22.760	24.424	25.752
8	24.331	25.884	27.538	28.895	24.352	25.904	27.557	28.911
9	27.475	29.029	30.675	32.052	27.493	29.047	30.692	32.067
10	30.618	32.174	33.812	35.207	30.635	32.190	33.828	35.220
		$\gamma = 10$			$\gamma = 10^8$			
1	2.9529	4.3001	5.5478	6.7497	3.1962	4.6109	5.9057	7.1435
2	5.9280	7.3770	8.7422	10.031	6.3064	7.7992	9.1957	10.144
3	8.9762	10.461	12.190	13.158	9.4687	10.947	12.694	13.322
4	12.057	13.566	15.265	16.374	12.567	14.102	15.813	16.668
5	15.156	16.679	18.358	19.556	15.708	17.251	18.939	19.948
6	18.266	19.798	21.463	22.721	18.850	20.398	22.069	23.184
7	21.383	22.922	24.576	25.878	21.991	23.543	25.202	26.392
8	24.506	26.050	27.695	29.029	25.133	26.687	28.336	29.582
9	27.633	29.179	30.818	32.178	28.274	29.831	31.472	32.760
10	30.762	32.311	33.943	35.324	31.416	32.974	34.609	35.930

Capitalizing on the fact that $R_{kj}(\rho)$ satisfies (8), the first differential operator in (13) can be replaced by κ_{kj}^4 and (13) becomes

$$\sum_k \sum_j [\kappa_{kj}^4 g_{kj}(\tau) + \alpha \ddot{g}_{kj}(\tau)] R_{kj}(\rho) \cos k\theta \\ = \frac{a^3 \delta(\rho - \rho_0) \delta(\theta) \delta(\tau - \tau_0)}{D \rho_0}. \quad (14)$$

Next, both sides of (14) are multiplied by

$$\rho R_{lm}(\rho) \cos l\theta d\theta d\rho,$$

and integration is performed, first with respect to θ from $0 \rightarrow 2\pi$ and then with respect to ρ from $0 \rightarrow 1$; by virtue of the orthogonalities (see Appendix), this reduces to

$$\kappa_{lm}^4 \Theta_{ll} N_{mm}^l g_{lm}(\tau) + \alpha \Theta_{ll} N_{mm}^l \ddot{g}_{lm}(\tau) = \frac{a^3}{D} R_{lm}(\rho_0) \delta(\tau - \tau_0), \quad (15)$$

where

$$\Theta_{ll} = \begin{cases} \pi & (l \neq 0), \\ 2\pi & (l = 0), \end{cases}$$

and

$$N_{mm}^l = \frac{1}{2} \{ I_l^2(\kappa_{lm}) J_{l+1}^2(\kappa_{lm}) - J_l^2(\kappa_{lm}) I_{l+1}^2(\kappa_{lm}) \} \\ - \left[\frac{1 + \gamma + 2l}{1 - \gamma} \right] I_l^2(\kappa_{lm}) J_l^2(\kappa_{lm}).$$

For the prescribed initial conditions, the solution of (15) is

$$g_{kj}(\tau) = \frac{a^3}{D} \frac{1}{\alpha} \frac{R_{kj}(\rho_0)}{\Theta_{kk} N_{jj}^k} \frac{\sin \omega_{kj}(\tau - \tau_0)}{\omega_{kj}} 1(\tau - \tau_0). \quad (16)$$

Equation (16), when used in conjunction with (12) is the Green's function for the circular elastic plate. Physically, it is the response due to a transverse point impulse applied to the plate.

3.4 Generalized Green's Function

The Green's function given by (12) was developed for the case of an impulsive loading singularity located on the line $\theta = 0$. Later, need will arise for the response due to a loading singularity located at the point $(\rho_0, -\alpha_0)$ (see Fig. 2). Employing the trigonometric identity,

$$\cos(A + B) = \cos A \cos B - \sin A \sin B,$$

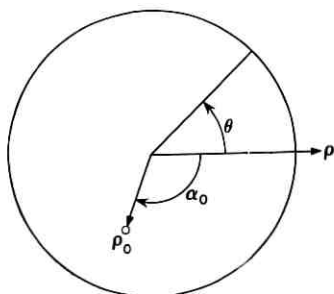


Fig. 2 — Plate coordinate system.

a trivial change of variables produces the corresponding generalized Green's function $G(\rho, \theta, \tau; \rho_0, \alpha_0, \tau_0)$

$$G(\rho, \theta, \tau; \rho_0, \alpha_0, \tau_0) = \sum_{k=0}^{\infty} \sum_{j=1}^{\infty} R_{kj}(\rho) \cos k\alpha_0 \cos k\theta g_{kj}(\tau) \quad (17)$$

$$- \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} R_{kj}(\rho) \sin k\alpha_0 \sin k\theta g_{kj}(\tau).$$

IV. APPLICATION OF GREEN'S FUNCTION

The Green's function as given by (17) will now be utilized to construct solutions for several loadings of technical interest and of practical concern.

4.1 Ring Loading

The first example is that of an impulsive concentrated ring pressure distribution given by

$$p(\rho, \theta, \tau) = p(\rho_0) \frac{\delta(\rho - \rho_0) \delta(\tau - \tau_0)}{\rho_0} \quad (18)$$

(see Fig. 3). Physically, this loading corresponds to a concentrated ring impulsive loading applied at radius ρ_0 and at time τ_0 . A differential element on the ring has a force per unit length $p(\rho_0)/\rho_0$. The resultant displacement of the plate is obtained by superposing contributions due to all of the elements on the ring. This superposition is expressed mathematically by the integral

$$\eta(\rho, \theta, \tau) = \int_0^{2\pi} G(\rho, \theta, \tau; \rho_0, \alpha_0, \tau_0) p(\rho_0) d\alpha_0, \quad (19)$$

where α_0 is defined by Fig. 3.

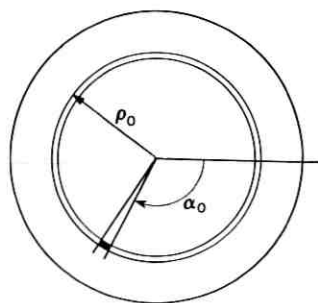


Fig. 3 — Ring loading.

The evaluation of (19) may be simplified considerably by observing that for $l \neq 0$ the integrals are all zero, leaving only the $l = 0$ terms. Accordingly, (19) simplifies to

$$\eta(\rho, \theta, \tau) = \frac{a^3 p(\rho_0)}{D\alpha} \sum_{j=1}^{\infty} \frac{R_{0j}(\rho) R_{0j}(\rho_0)}{N_{jj}^0 \omega_{0j}} \sin \omega_{0j}(\tau - \tau_0) 1(\tau - \tau_0). \quad (20)$$

Equation (20) may be recognized as the axisymmetric solution that was obtained previously,⁶ provided of course that proper account is taken for the change in nomenclature.

4.2 Concentrated Line Loading

As a second example, consider a concentrated line impulse loading specified by

$$p(x, y, \tau) = \delta(x - x_0) \delta(\tau - \tau_0), \quad (21)$$

where x and y are cartesian coordinates appropriate for this problem and defined by Fig. 4. Physically, this loading corresponds to an impulsive concentrated line loading applied at position $x = x_0$ and at time $\tau = \tau_0$.

The force per unit length acting on each element of the line loading given by (21) is

$$\delta(\tau - \tau_0).$$

Likewise, the force acting on the plate due to the Green's function singularity located at (x_0, y) is

$$\delta(\tau - \tau_0).$$

It follows that the resultant deflection due to all differential elements of the line loading is expressed by the superposition integral

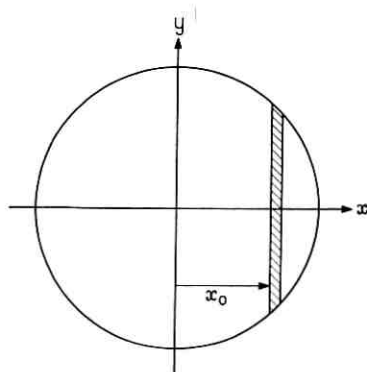


Fig. 4 — Concentrated line loading.

$$\eta(\rho, \theta, \tau; x_0, \tau_0) = \int_{y=-(1-x_0^2)^{\frac{1}{2}}}^{(1-x_0^2)^{\frac{1}{2}}} G\left(\rho, \theta, \tau; [x_0^2 + y^2]^{\frac{1}{2}}, \tan^{-1} \frac{y}{x_0}, \tau_0\right) dy. \quad (22)$$

Substituting (17) into (22) gives the following expression for the dynamic response of the plate:

$$\begin{aligned} & \eta(\rho, \theta, \tau; x_0, \tau_0) \\ &= \frac{a^3}{D\alpha} \sum_{l=0}^{\infty} \sum_{m=1}^{\infty} \left[\frac{R_{lm}(\rho) \cos l\theta}{\Theta_{ll} N_{mm}^l \omega_{lm}} \int_{-(1-x_0^2)^{\frac{1}{2}}}^{(1-x_0^2)^{\frac{1}{2}}} R_{lm}(x_0^2 + y^2)^{\frac{1}{2}} \right. \\ & \quad \cdot \cos \left[l \tan^{-1} \frac{y}{x_0} \right] dy - \frac{R_{lm}(\rho) \sin l\theta}{\Theta_{ll} N_{mm}^l \omega_{lm}} \int_{-(1-x_0^2)^{\frac{1}{2}}}^{(1-x_0^2)^{\frac{1}{2}}} R_{lm}(x_0^2 + y^2)^{\frac{1}{2}} \\ & \quad \cdot \sin \left[l \tan^{-1} \frac{y}{x_0} \right] dy \left. \right] \times \sin \omega_{lm}(\tau - \tau_0) 1(\tau - \tau_0). \end{aligned} \quad (23)$$

The integrals in (23) are not expressible in closed form; however, they can be integrated numerically. Fortunately, there is one point on the plate where the evaluation does simplify considerably, and that is at the center of the plate, i.e., $\rho = 0$. Restricting our attention to the center of the plate, only the $l = 0$ terms are nonzero, and consequently the center deflection is given by

$$\begin{aligned} \eta(0, \tau; x_0, \tau_0) &= \frac{a^3}{\pi\alpha D} \sum_{m=1}^{\infty} \frac{R_{0m}(0)}{N_{mm}^0 \omega_{0m}} \int_0^{(1-x_0^2)^{\frac{1}{2}}} R_{0m}(x_0^2 + y^2)^{\frac{1}{2}} dy \\ & \quad \times \sin \omega_{0m}(\tau - \tau_0) 1(\tau - \tau_0). \end{aligned} \quad (24)$$

The integrals in (24) are perfectly well behaved and finite, although they cannot be expressed in closed form. Certain special cases (e.g., for

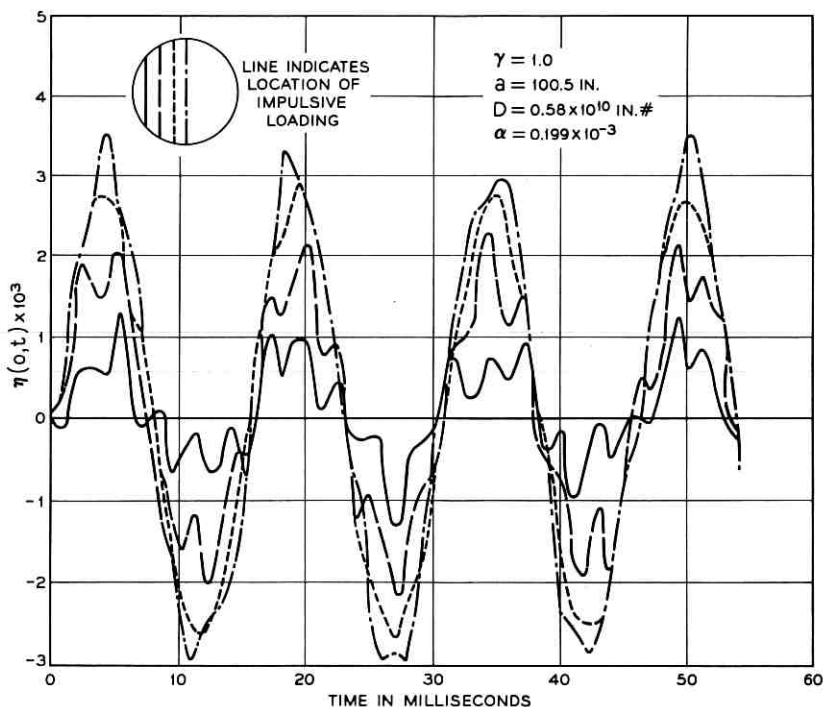


Fig. 5—Dynamic response due to impulsive line loading ($\gamma = 1.0$).

$x_0 = 0$) have been tabulated.⁴ Several other cases have been evaluated by the writer, and the results have been incorporated into the numerical examples.

4.3 Pressure Acting on a Portion of the Plate

A third example is a plate that is impulsively loaded by a uniform pressure acting on a segment of its surface and unloaded elsewhere. This loading is depicted by Fig. 7.

The center deflection may be thought of as the resultant deflection due to loadings on each line segment from

$$-1 \leq x_0 \leq x'$$

A pressure of magnitude $P_0 \delta(\tau - \tau_0)$ loads each element of width dx_0 with a line loading of magnitude $P_0 \delta(\tau - \tau_0) dx_0$. The resultant center deflection is obtained directly from the superposition integral

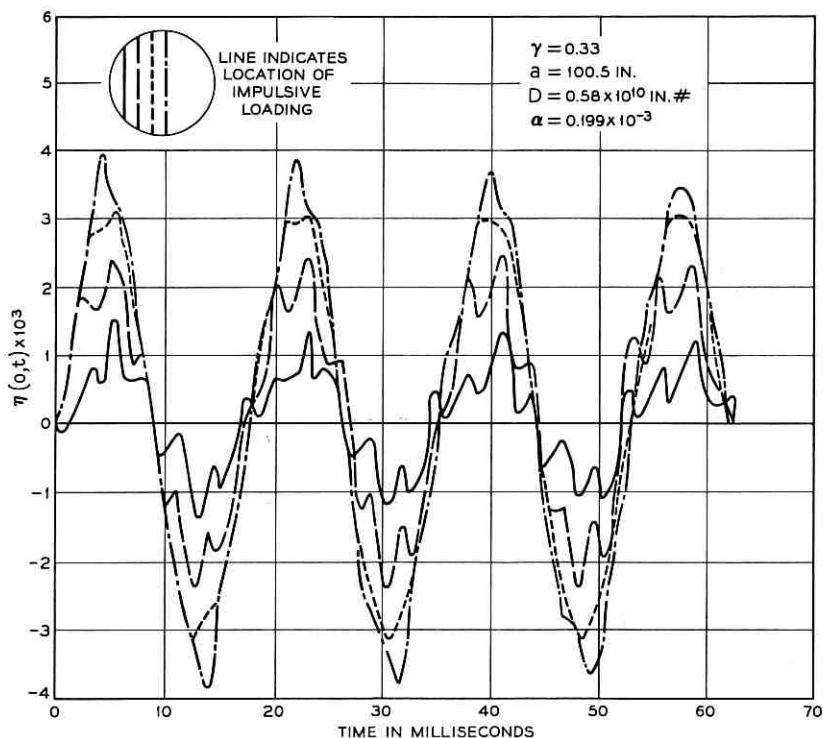
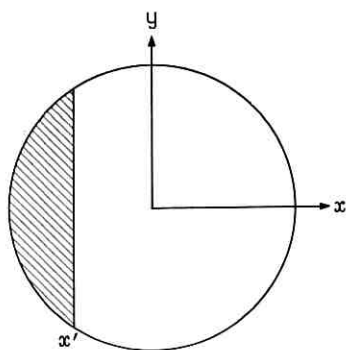
Fig. 6 — Dynamic response due to impulsive line loading ($\gamma = 0.33$).

Fig. 7 — Partially loaded plate.

$$\eta(0, \tau; x', \tau_0) = \int_{-1}^{x'} P_0 \eta(0, \tau; x_0, \tau_0) dx_0, \quad (25)$$

where $\eta(0, \tau; x_0, \tau_0)$ is defined in (24).

The integrals in (25) may be integrated numerically. One case that is readily integrable in closed form is when $x' = 1$; that case corresponds to a plate that is suddenly loaded by a uniform pressure over its entire surface. It is readily shown in that instance that (25) reduces to the previously obtained solutions presented in Ref. 6.

If a sector of the plate is loaded with a uniform pressure impulse loading, as indicated by Fig. 8, the superposition methods employed above confirm the intuitive suspicion that the center deflection is the deflection for a uniformly loaded plate multiplied by the quantity $\theta_0/2\pi$, where θ_0 is the included angle of the sector (expressed in radians).

4.4 Sweeping Pressure Wave

Next, consider the case of a circular plate loaded by a step blast wave that sweeps across its surface with constant velocity c (expressed in plate diameters per dedimensionalized time) as indicated in Fig. 9. The sector of the plate behind the wave front is loaded by pressure P_0 , whereas that portion ahead of the wave front is as yet unloaded.

It is convenient to consider the plate to be divided into strips of equal width Δx_0 , with the strips numbered successively from the left hand edge ($x_0 = -1$). This is depicted in Fig. 10.

The first strip is then loaded by a step pressure wave (unloaded until $\tau_0 = 0$) and then loaded with pressure P_0 for $\tau_0 > 0$). After an increment

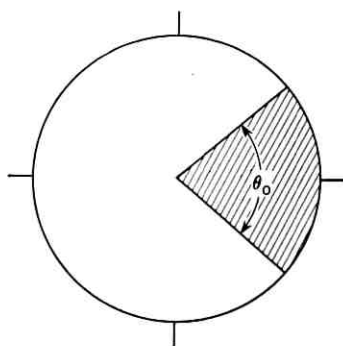
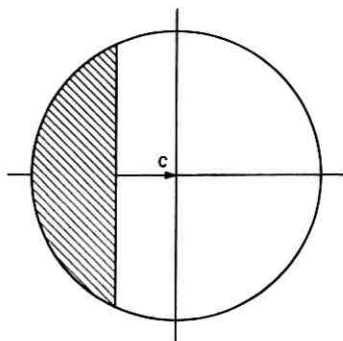


Fig. 8 — Plate loaded on sector.

Fig. 9 — Sweeping pressure wave with velocity c .

of time $\Delta\tau_0$ has elapsed, the second strip is loaded by the same step pressure wave (unloaded until $\tau_0 = \Delta\tau_0$ and then loaded with pressure P_0 for $\tau_0 > \Delta\tau_0$). After time interval $2\Delta\tau_0$, the next strip is loaded by the same step pressure loading, and so on, until eventually all strips are loaded by pressure P_0 . As the width of the strip is reduced to an infinitesimal, and correspondingly the number of strips increases to infinity, the continuously sweeping pressure wave results.

The solution for a suddenly applied *step* line loading is obtained by integrating (23) with respect to τ_0 (Duhamel integral). The suddenly applied step line loading, denoted by $\bar{\eta}(\rho, \theta, \tau; x_0, \tau_0)$, is given by

$$\bar{\eta}(\rho, \theta, \tau; x_0, \tau_0) = \int_{\tau_0}^{\tau} \eta(\rho, \theta, \tau; x_0, \tau_0) d\tau_0.$$

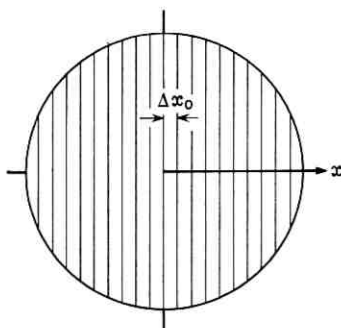


Fig. 10 — Sweeping pressure wave.

The result of this integration is

$$\begin{aligned} &\bar{\eta}(\rho, \theta, \tau; x_0, \tau_0) \\ &= \frac{a^3}{D\alpha} \sum_{l=0}^{\infty} \sum_{m=1}^{\infty} \left[\frac{R_{lm}(\rho) \cos l\theta}{\Theta_{ll} N_{mm} l \omega_{lm}^2} \int_{-(1-x_0^2)^{\frac{1}{2}}}^{(1-x_0^2)^{\frac{1}{2}}} R_{lm}(x_0^2 + y^2)^{\frac{1}{2}} \right. \\ &\quad \cdot \cos \left[l \tan^{-1} \frac{y}{x_0} \right] dy - \frac{R_{lm}(\rho) \sin l\theta}{\Theta_{ll} N_{mm} l \omega_{lm}^2} \int_{-(1-x_0^2)^{\frac{1}{2}}}^{(1-x_0^2)^{\frac{1}{2}}} R_{lm}(x_0^2 + y^2)^{\frac{1}{2}} \\ &\quad \cdot \sin \left[l \tan^{-1} \frac{y}{x_0} \right] dy \left. \right] \times [1 - \cos \omega_{lm}(\tau - \tau_0)] 1(\tau - \tau_0). \end{aligned} \tag{26}$$

In terms of the solution for the suddenly applied step line loading, the sweeping step wave solution, as obtained by superposition is

$$\begin{aligned} \eta^*(\rho, \theta, \tau; c) &= \lim_{\Delta x_0 \rightarrow 0} \sum_{n=0}^N P_0 \bar{\eta} \left(\rho, \theta, \tau; -1 + n\Delta x_0, \frac{n\Delta x_0}{c} \right) \Delta x_0, \\ &\text{for } 0 \leq \tau \leq \frac{2}{c} \end{aligned} \tag{27}$$

where $\bar{\eta}$ is given by (26), and η^* is the deflection due to the sweeping wave. In the limit, (27) is expressed by the integral

$$\eta^*(\rho, \theta, \tau; c) = P_0 \int_{y_0=-1}^{x'} \bar{\eta} \left(\rho, \theta, \tau; y_0, \frac{1 + y_0}{c} \right) dy_0 \quad 0 \leq \tau \leq \frac{2}{c}. \tag{28}$$

Although (28) cannot be evaluated in closed form, (27) is perfectly well suited for numerical evaluation (see Figs. 11, 12, 13), and the ac-

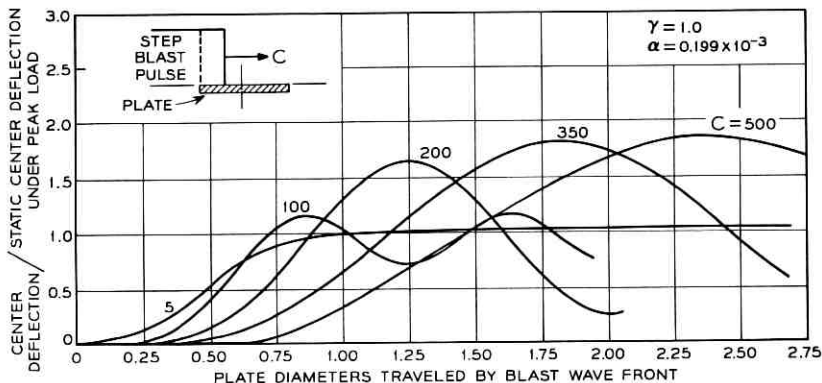


Fig. 11 — Dynamic response due to sweeping step wave ($\gamma = 1.0$).

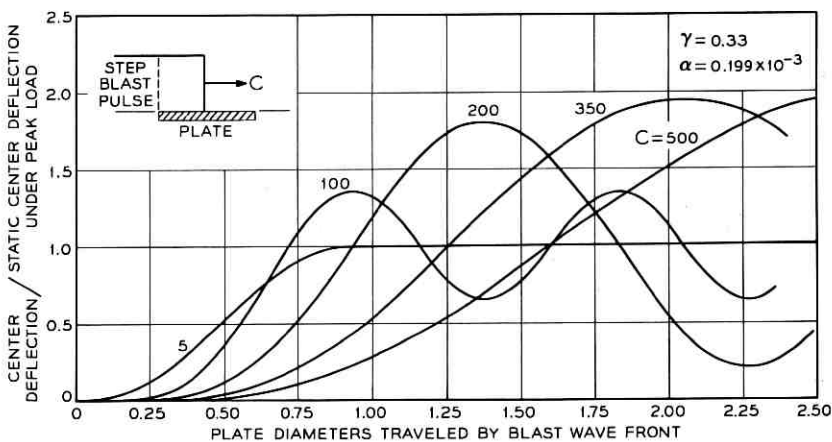


Fig. 12—Dynamic response due to sweeping step wave ($\gamma = 0.33$).

curacy obtainable is limited only by the computer time expenses associated with the numerical integrations. In the numerical evaluations included in this article, certain symmetry properties were exploited to minimize computer time requirements.

4.5 Sweeping Pressure Wave with Decay

As a final example, the case of a sweeping pressure pulse will be considered where the magnitude of the pressure at any point behind the wavefront decays exponentially as a function of the distance behind the

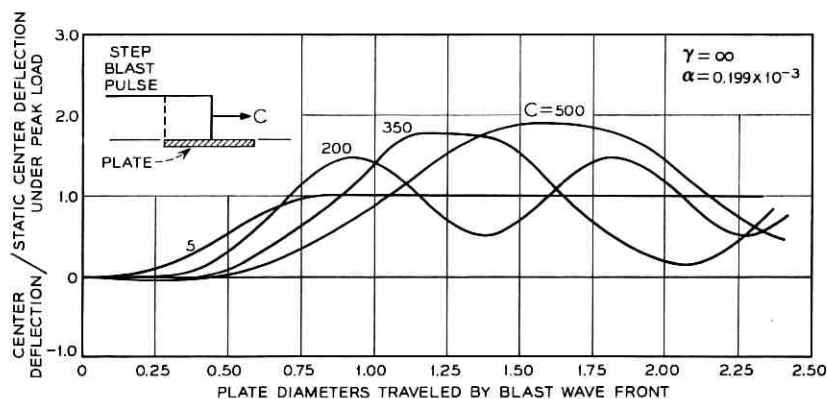


Fig. 13—Dynamic response due to sweeping step wave ($\gamma = \infty$).

wavefront. As in the previous example, the speed of the pressure wave is considered to be constant.

Referring to Fig. 14, a point at $x = x_0$ is unloaded until time $\tau = (1 + x_0)/c$, at which time the pulse arrives, loading it with pressure P_0 . Because of the decay of the loading pulse, the point at x_0 is loaded with pressure

$$P(x_0, \tau) = P_0 \exp \left\{ \frac{c}{2} \ln \delta \left[\tau - \frac{1 + x_0}{c} \right] \right\} 1 \left(\tau - \frac{1 + x_0}{c} \right). \quad (29)$$

The expressions for the deflection due to the loading given by (29) are identical to (27) and (28) with the exception that $\bar{\eta}$ must be modified. The necessary modification is to replace the factor $[1 - \cos \omega_{lm}(\tau - \tau_0)]$ that appears in (26) by the corresponding factor

$$\frac{\omega_{lm}^2}{\omega_{lm}^2 + (\ln \delta)^2} \left\{ \exp [\ln \delta (\tau - \tau_0)] - \frac{\ln \delta}{\omega_{lm}} \sin \omega_{lm}(\tau - \tau_0) - \cos \omega_{lm}(\tau - \tau_0) \right\}. \quad (30)$$

It can be seen that for the case of zero decay ($\ln \delta = 0$) (30) reduces to the step loading solution (Figs. 15, 16, 17).

V. RESULTS

Dynamic response curves are presented in this article for line impulse loadings and for sweeping waves of both constant and time-diminishing pressure pulses. Figs. 5 and 6 illustrate that a line impulsive loading applied near the diameter excites primarily the first mode of vibration; line loadings applied away from the diameter excite a larger percentage of the higher modes, although as one should expect, the magnitude of the deflection diminishes as the line loading is applied nearer to the edge.

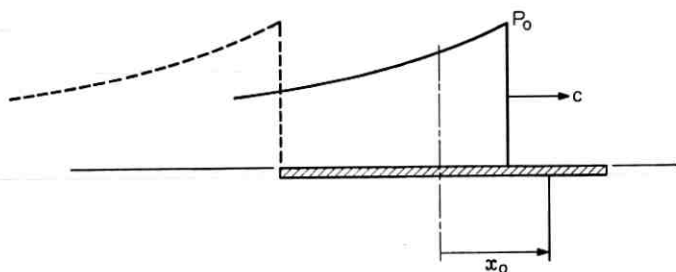


Fig. 14—Sweeping pressure wave with decay.

functions were computed. In this appendix, corresponding orthogonality properties and normalizing factors will be explored for the non-axisymmetric modes of vibration.

The eigenfunctions,

$$R_{nj}(\rho) \begin{cases} \cos n\theta \\ \text{or} \\ \sin n\theta \end{cases} = \{I_n(\kappa_{nj})J_n(\kappa_{nj}\rho) - J_n(\kappa_{nj})I_n(\kappa_{nj}\rho)\} \begin{cases} \cos n\theta \\ \text{or} \\ \sin n\theta \end{cases},$$

where $n = 0, 1, 2, \dots$ and κ_{nj} satisfies the transcendental equation

$$\frac{2\kappa_{nj}}{1-\gamma} I_n(\kappa_{nj})J_n(\kappa_{nj}) = I_n(\kappa_{nj})J_{n+1}(\kappa_{nj}) + J_n(\kappa_{nj})I_{n+1}(\kappa_{nj}),$$

are solutions to the partial differential equation (6) and satisfy the boundary conditions.

A.1 Orthogonality

To begin, recall the well known¹⁰ results

$$\int_0^{2\pi} \cos n\theta \sin m\theta d\theta = 0 \quad m, n \text{ any integers;}$$

$$\int_0^{2\pi} \cos n\theta \cos m\theta d\theta = \begin{cases} 0 & (n \neq m), \\ \pi & (n = m \neq 0), \\ 2\pi & (n = m = 0); \end{cases}$$

$$\int_0^{2\pi} \sin n\theta \sin m\theta d\theta = \begin{cases} 0 & (n \neq m), \\ \pi & (n = m). \end{cases}$$

These orthogonality relationships imply, in our case, that there is *no coupling* between modes of vibration having n nodal diameters and those modes having m nodal diameters, where $n \neq m$. Thus, one need only concern oneself with relationships such as

$$\int_0^1 \rho R_{nj}(\kappa_{nj}\rho) R_{nk}(\kappa_{nk}\rho) d\rho.$$

Therefore, consider the ordinary differential equations (8) that R_{nj} and R_{nk} must satisfy

$$\left(\frac{d^2}{d\rho^2} + \frac{1}{\rho} \frac{d}{d\rho} - \frac{n^2}{\rho^2}\right)^2 R_{nj} = \kappa_{nj}^4 R_{nj}$$

and

$$\left(\frac{d^2}{d\rho^2} + \frac{1}{\rho} \frac{d}{d\rho} - \frac{n^2}{\rho^2}\right)^2 R_{nk} = \kappa_{nk}^4 R_{nk}.$$

Multiplying the first equation by $\rho R_{nk} d\rho$, and the second by $\rho R_{nj} d\rho$, subtracting, and integrating with respect to ρ from $0 \rightarrow 1$ gives

$$\begin{aligned} (\kappa_{nj}^4 - \kappa_{nk}^4) \int_0^1 \rho R_{nj} R_{nk} d\rho &= \int_0^1 \rho R_{nk} \left(\frac{d^2}{d\rho^2} + \frac{1}{\rho} \frac{d}{d\rho} - \frac{n^2}{\rho^2}\right)^2 R_{nj} d\rho \\ &\quad - \int_0^1 \rho R_{nj} \left(\frac{d^2}{d\rho^2} + \frac{1}{\rho} \frac{d}{d\rho} - \frac{n^2}{\rho^2}\right)^2 R_{nk} d\rho. \end{aligned} \quad (32)$$

The right side of (32) is expanded, and integrated by parts. The lengthy, but routine details of this operation are suppressed in the interest of brevity; the result is

$$\begin{aligned} &\int_0^1 \rho R_{nj} R_{nk} d\rho \\ &= \frac{R_{nk} \left\{ \rho \frac{d^3 R_{nj}}{d\rho^3} + \frac{d^2 R_{nj}}{d\rho^2} - \frac{1}{\rho} \frac{d R_{nj}}{d\rho} \right\} - R_{nj} \left\{ \rho \frac{d^3 R_{nk}}{d\rho^3} + \frac{d^2 R_{nk}}{d\rho^2} - \frac{1}{\rho} \frac{d R_{nk}}{d\rho} \right\}}{(\kappa_{nj}^4 - \kappa_{nk}^4)} \Bigg|_0^1 \\ &\quad + \left\{ \rho \frac{d R_{nj}}{d\rho} \frac{d^2 R_{nk}}{d\rho^2} - \rho \frac{d R_{nk}}{d\rho} \frac{d^2 R_{nj}}{d\rho^2} \right\} + \frac{2n^2}{\rho} \left\{ \frac{d R_{nk}}{d\rho} R_{nj} - \frac{d R_{nj}}{d\rho} R_{nk} \right\} \Bigg|_0^1. \end{aligned} \quad (33)$$

For the case of $n = 0$, (33) was shown to vanish in Ref. 6 for $j \neq k$. For $n = 1, 2, \dots$ similar reasoning demonstrates that each of the terms in (33) either vanishes or the terms mutually annihilate each other, so that the eigenfunctions are orthogonal under the rather general boundary conditions given in (2).

A.2 Normalizing Factor

For $n = 0, 1, 2, \dots$, but where $j = k$, both numerator and denominator of (33) are zero. Thus, a limiting process must be employed to find the value of

$$N_{jj}^n \equiv \int_0^1 \rho R_{nj}^2(\rho) d\rho$$

from (33). Differentiating numerator and denominator of (33) with respect to κ_{nk} and then setting $\kappa_{nk} = \kappa_{nj}$, the result is

$$\begin{aligned} N_{jj}^n &= \frac{1}{2} \{ I_n^2(\kappa_{nj}) J_{n+1}^2(\kappa_{nj}) - J_n^2(\kappa_{nj}) I_{n+1}^2(\kappa_{nj}) \} \\ &\quad - \frac{1 + \gamma + 2n}{1 - \gamma} I_n^2(\kappa_{nj}) J_n^2(\kappa_{nj}). \end{aligned} \quad (34)$$

This normalizing factor is identical to the axisymmetric normalizing factor derived in Ref. 6 for the axisymmetric case ($n = 0$). An alternative way to derive (34), which was used to check this equation, is to perform the indicated integration directly.

NOMENCLATURE

a	radius of boundary of plate
c	speed at which pressure wave sweeps across plate (expressed in plate diameters per dimensionless time unit)
D	flexural rigidity of plate $\left(\equiv \frac{Eh^3}{12(1-\nu^2)} \right)$
E	Young's modulus
G	Green's function
$g_{kj}(\tau)$	function of time defined by (10)
h	plate thickness
$J_n(z), I_n(z)$	n th order Bessel function and modified Bessel function of argument z
m	mass per unit projected area of plate
N_{jj}^n	normalizing factor for radial functions, defined by (34)
$p(\rho, \theta, \tau)$	pressure loading on plate
P_o	pressure of loading pulse (constant)
$R_{nj}(\rho)$	radial eigenfunction defined by (7)
r	radial coordinate
t	time
T	time duration used to de-dimensionalize the time variables (chosen for convenience)
w	plate deflection
x, y	cartesian coordinates
α	dimensionless parameter $(\equiv ma^4/DT^2)$
β	modulus of spring on edge of plate
γ	edge-fixity parameter $\left(\equiv \nu + \frac{\beta a}{D} \right)$
δ	decay rate of sweeping pressure pulse
$\delta(z)$	Dirac delta function of argument z
η	dimensionless plate deflection $(\equiv w/a)$
θ	angular coordinate
Θ_{ll}	normalizing factor for angular normal functions
κ_{nj}	eigenvalue corresponding to mode with n nodal diameters and j nodal circles

ν	Poisson's ratio
ρ	dimensionless radial coordinate ($\equiv r/a$)
τ	dimensionless time variable ($\equiv t/T$)
ω_{nj}	dimensionless angular frequency associated with the $n-j$ mode of vibration
∇^2	Laplacian operator ($\equiv \frac{\partial^2}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2}{\partial \theta^2}$)
$1(z)$	unit step function of argument z

Differentiation with respect to time is denoted by dots.

REFERENCES

1. Poisson, S. D., Sur le Mouvement des Corps Elastiques Memoires de l'Academie Royale des Sciences de l'Institut de France, 8, 1829, pp. 357-370.
2. Kirchoff, G., Uber das Gleichgewicht und die Bewegung einer Elastischen Scheibe, Journal für die reine und angewandte Mathematik (Crelle), 40, 1850, p. 51.
3. Flynn, P. D., Elastic Response of Simple Structures to Pulse Loadings, Ballistic Research Laboratories Memorandum, Report No. 525, November, 1950.
4. Sneddon, I. N., *Fourier Transforms*, McGraw-Hill Book Company, Inc., New York, N. Y., 1st Ed., 1951, pp. 150-153.
5. Reismann, H., Forced Vibrations of a Circular Plate, Trans. ASME, J. Appl. Mech., 26, Series E, n. 4, 1959, pp. 526-27.
6. Weiner, R. S., Forced Axisymmetric Motions of Circular Elastic Plates, J. of Appl. Mech., Paper No. 65-APMW-7.
7. Mindlin, R. D., Influence of Rotatory Inertia and Shear on Flexural Motions of Isotropic Elastic Plates, J. of Appl. Mech., 18; Trans. ASME, 73, 1951, pp. 31-38.
8. Kantham, C. L., Bending and Vibration of Elastically Restrained Circular Plates, J. Franklin Institute, 265, June, 1958, pp. 483-91.
9. Reid, W. P., Free Vibrations of a Circular Plate, U. S. Naval Ordnance Laboratory Report 61-186, ASTIA No. 279140, January, 1962.
10. Churchill, R. V., *Fourier Series and Boundary Value Problems*, McGraw-Hill Book Company, Inc., New York, N. Y., 1941, pp. 53-4.

Computation of Lattice Sums: Generalization of the Ewald Method II

By WALTER J. C. GRANT

(Manuscript received August 13, 1965)

Our method of computing lattice sums¹ is carried through the final stage of numerical computation. We calculate the expansion coefficients of the crystal potential, evaluated at each inequivalent site of a number of cubic structures. Some of the structures chosen are simple, to afford comparison with previous calculations. Many, however, are too complex to be amenable to other methods, and our results are the first reported. The sequence of computer programs, timing, and accuracy are discussed.

In a previous paper¹ concerning lattice sums we discussed summation methods which hinge on two facts: First, many lattices can be decomposed into so-called "primitive" lattices. Second, a wide class of summation procedures is feasible for such primitive lattices, which are impossible or impractical for arbitrary lattices. In particular, we developed an extension of the Ewald method, following the general philosophy of Nijboer and DeWette,² Adler,³ and Barlow and Macdonald.⁴ The conditions under which a given lattice can be decomposed into primitive lattices, and the algorithm accomplishing this, have been discussed with full generality and rigor by Graham.⁵

In this paper we present a sample of numerical results obtained by our method, together with some discussion of computational techniques and computational efficiency. We have evaluated the coefficients, up to order six, for the spherical harmonic expansion of the potential due to a lattice of point charges. We have done this for every inequivalent site in a number of cubic lattices. Some of the lattices were chosen to provide comparison with other calculations; some, on the contrary, were chosen because their complexity puts them beyond the reach of other methods; others still because cubic lattices without an inversion center are interesting in a number of other connections.

We define terms and display the relevant equations, as briefly as possible. For a fuller exposition we refer to Refs. 1 and 5.

A lattice is primitive if it consists of a set of points of position \mathbf{r}_n and charge q_n such that

$$\mathbf{r}_n = \sum_{i=1}^3 n_i \mathbf{b}_i, \quad n_i = 0, \pm 1, \pm 2, \dots \quad (1a)$$

$$q_n = q_0 (-1)^{n_1+n_2+n_3}. \quad (1b)$$

For instance, NaCl is primitive, with $\mathbf{b}_1 = (\frac{1}{2}, 0, 0)$, $\mathbf{b}_2 = (0, \frac{1}{2}, 0)$, and $\mathbf{b}_3 = (0, 0, \frac{1}{2})$, where the components of the \mathbf{b} 's are given along the cube axes in units of the cube dimensions. We call the vectors \mathbf{b}_i the basis vectors of the primitive lattice. These should be distinguished from primitive translations \mathbf{c}_i , which are defined by the relation $q(\mathbf{r}) = q(\mathbf{r} + n_1 \mathbf{c}_1 + n_2 \mathbf{c}_2 + n_3 \mathbf{c}_3)$, for all \mathbf{r} , and for all integers n_1, n_2, n_3 .

An arbitrary three-dimensional lattice can be decomposed if it is of periodicity $2^N \times 2^N \times 2^N$. This condition can be loosened to periodicities $2^L \times 2^M \times 2^N$ with $L, M < N$, since by taking appropriate multiples in the direction in which the periodicity is 2^L or 2^M one can reduce this case to the previous one. The existence condition can be further loosened to include lattices composed of "interlocking" sublattices of periodicity $2^L \times 2^M \times 2^N$, provided each such sublattice is separately electrically neutral.

We consider, then, an array of periodicity $2^N \times 2^N \times 2^N$, and primitive translations $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$. This array can be decomposed into 2^{3N-1} primitive lattices. These component lattices will be grouped into $3N$ sets. All lattices within one set will be defined by the same basis vectors, but will differ in choice of origin. With respect to the origin in the original lattice, the position of the origin of a component lattice is denoted by \mathbf{R} . We define auxiliary vectors $\mathbf{a}_i = \mathbf{c}_i/2^N$. We wish to express the basis vectors \mathbf{b}_i of the primitive component lattices, as well as the origin positions \mathbf{R} , in terms of the \mathbf{a}_i . Let the columns of the matrix \mathbf{A} denote the components of \mathbf{a}_i , and likewise for \mathbf{B} and \mathbf{b}_i . Superscripts will denote sets of primitive component lattices, subscripts denote lattices within a given set. Then the required relations are

$$\mathbf{B}^{(1)} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{A} \quad (2a)$$

$$\mathbf{B}^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{bmatrix} \mathbf{A} \quad (2b)$$

$$\mathbf{B}^{(3)} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix} \mathbf{A} \quad (2c)$$

$$\mathbf{B}^{(n+3)} = 2\mathbf{B}^{(n)} \quad (2d)$$

$$\{\mathbf{R}^{(1)}\} = (0,0,0) \quad (3a)$$

$$\{\mathbf{R}^{(n+1)}\} = \{\mathbf{R}^{(n)}\} + \{\mathbf{R}^{(n)} + \mathbf{B}_1^{(n)}\} \quad (3b)$$

$$q^{(1)} = 2^{-3N/2} \quad (4a)$$

$$q^{(n+1)} = q^{(n)}\sqrt{2}. \quad (4b)$$

These definitions differ from those given in Refs. 1 and 5 by relabeling some of the indices, which has the effect that all the \mathbf{b} vectors now define right-handed bases. The primitive lattices so defined are orthonormal in the sense that

$$\sum_n q_\alpha^{(i)}(\mathbf{r}_n) q_\beta^{(j)}(\mathbf{r}_n) = \delta_{ij} \delta_{\alpha\beta} \quad (5)$$

where n runs over the sites in a unit cell, the superscript i or j denotes a set of primitive lattices, and the subscript α or β denotes a particular lattice belonging to that set. The primitive lattices do not necessarily correspond to physically real structures. If we take $\mathbf{A} = \mathbf{1}$, representative lattices belonging to the first three sets are shown in Fig. 1. There is one member in set 1, two in set 2, four in set 3. For every set we show only the lattice with origin at $\mathbf{R} = 0$. Only set 3 has a physical counterpart, and will be recognized as the CsCl structure. Further examples may be found in Ref. 5.

We now consider the expansion coefficients of the crystal potential. Just as the physical lattice is represented as the sum of primitive lattices, so a given expansion coefficient is computed as the sum of the corresponding coefficients proper to the primitive lattices. As suggested in the previous paper,¹ the coefficients arising from primitive component lattices can be computed once and for all, so that the problem for a real physical lattice, no matter how complicated, is reduced to the

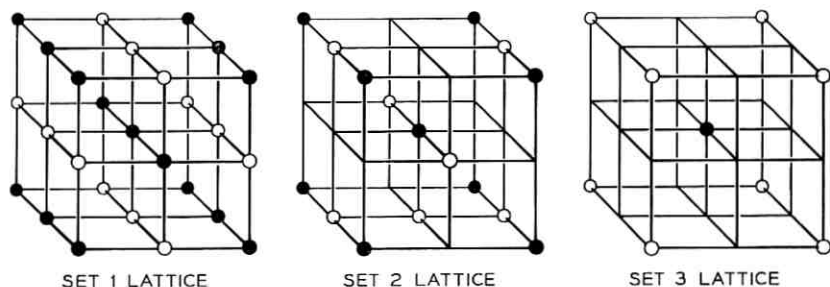


Fig. 1—Conventional unit cell representation of primitive lattices, sets 1, 2, and 3 (Translation $R = 0,0,0$).

decomposition of that lattice into primitive components. If the potential is given, as usual, by

$$V(\mathbf{x}) = \sum_{l=0}^{\infty} \sum_{m=-l}^l C_{lm} |\mathbf{x}|^l Y_{l,-m}(\theta_{\mathbf{x}}, \varphi_{\mathbf{x}}) \quad (6)$$

then the expansion coefficients may be represented as

$$C_{lm} = \sum_{\text{sets } (i)} \sum_{\text{translations } (\alpha)} C_{lm}(\mathbf{b}^{(i)} - \mathbf{R}_{\alpha}^{(i)}) \quad (7)$$

where the notation is consistent with that of (2) through (5). Shortening the argument of the primitive lattice C_{lm} to ρ , ($\rho = \mathbf{b}^{(i)} - \mathbf{R}_{\alpha}^{(i)}$), we have, for each primitive lattice,

$$C_{lm} = \frac{4\pi}{2l+1} \sum_n q_n |\boldsymbol{\rho}_n|^{-l-1} Y_{l,m}(\theta_{\boldsymbol{\rho}_n}, \varphi_{\boldsymbol{\rho}_n}) \quad (8)$$

where n runs over the sites of that particular lattice. The spherical harmonics are defined as usual:

$$Y_{l,m}(\theta, \varphi) = \left[\frac{2l+1}{4\pi} \cdot \frac{(l-|m|)!}{(l+|m|)!} \right]^{\frac{1}{2}} e^{im\varphi} P_{lm}(\theta). \quad (9)$$

For small l , the summation in (8) converges poorly or not at all. By means of the formalism discussed in Ref. 1, (8) can be transformed into the following, which converges rapidly for all l :

$$\begin{aligned}
 C_{lm} = & \frac{4\pi}{2l+1} \cdot \frac{1}{\Gamma(l+\frac{1}{2})} \\
 & \cdot \left\{ \sum_n |\boldsymbol{\rho}_n|^{l-1} \Gamma(l+\frac{1}{2}, \pi\alpha^2 |\boldsymbol{\rho}_n|^2) \right. \\
 & \cdot Y_{lm}(\theta_{\boldsymbol{\rho}_n}, \varphi_{\boldsymbol{\rho}_n}) (-1)^{n_1+n_2+n_3} \\
 & + i^l \pi^{l-\frac{1}{2}} v_c^{-1} \sum_n \exp(i2\pi \mathbf{n}_n \cdot \mathbf{R}_j^{(i)}) |\mathbf{n}_n|^{l-2} \\
 & \left. \cdot \Gamma(1, \pi\alpha^{-2} |\mathbf{n}_n|^2) \cdot Y_{lm}(\theta_{\mathbf{n}_n}, \varphi_{\mathbf{n}_n}) - \alpha \delta_l \delta_{\boldsymbol{\rho}_n} \right\}. \quad (10)
 \end{aligned}$$

Here, Γ is the incomplete Gamma function, and v_c is the cell volume $\mathbf{b}_1 \cdot \mathbf{b}_2 \times \mathbf{b}_3$. The vector \mathbf{n} in reciprocal space is defined by

$$\mathbf{n}_n = \mathbf{h}_n^{(i)} - \frac{1}{2}(\mathbf{h}_1^{(i)} + \mathbf{h}_2^{(i)} + \mathbf{h}_3^{(i)}) \quad (11)$$

where the \mathbf{h} 's are basis vectors in the reciprocal lattice,

$$\mathbf{h}_1^{(i)} = \frac{1}{v_c} (\mathbf{b}_2^{(i)} \times \mathbf{b}_3^{(i)}), \quad (12)$$

and \mathbf{h}_n is a lattice vector of the reciprocal lattice. The sums on n are to be understood as triple sums on n_1, n_2, n_3 . The scale factor α is given by

$$\alpha = v_c^{-\frac{1}{3}} \quad (13)$$

and makes the rate of convergence equal in both coordinate and reciprocal space, regardless of choice of units or size of cell. The result of the summation is independent of the choice of α , but the correct choice is crucial in practice since it can affect computation time by orders of magnitude. The last term in (10) exhibits explicitly the correction discussed in Ref. 1 for the case $\rho = 0$.

In practice, there are four computational processes, performing the following functions: (i) We generate all the primitive component lattices of periodicity $2^N \times 2^N \times 2^N$. We did this for $N = 1, 2, 3$, requiring respectively 7, 63, and 511 lattices. (ii) For each component lattice, we calculate the C_{lm} , both real and imaginary parts, with $0 \leq l \leq 6$, $0 \leq m \leq l$. (iii) We decompose a given lattice, by taking appropriate dot-products (see (5)). (iv) We reconstitute the C_{lm} 's for the given lattice by combining the results of steps (ii) and (iii). Parts (i) and (ii) are done once only for a given type of geometrical grid. Parts (iii) and (iv) are done for each charge configuration that can be placed on that grid.

The computing time is spent almost entirely on Part (ii). About an hour on the IBM 7094 was required to calculate all the C_{lm} 's for all the primitive lattices. To compensate, the application of these results to 32 physical configurations, for all 56 coefficients, took about twelve seconds.

The computation for Part (ii) is done shell-wise. By the k th shell we mean here a set of points $\mathbf{r}_n = n_1\mathbf{b}_1 + n_2\mathbf{b}_2 + n_3\mathbf{b}_3$ such that at least one n_i is equal to k in absolute value, but no n_i is greater than k in absolute value. The number of points in the k th shell is $(24k^2 + 2)$, and the number of points in all shells up to and including the k th is $(2k + 1)^3$. These shells differ from "Evjen" shells, which, unlike ours, use fractional charges, and preserve crystal symmetry and electrical neutrality for every shell. The k th shell defined above contains an excess charge of $2q(-1)^k$, if the charge $q(-1)^{n_1+n_2+n_3}$ is associated with each site. Since

the excess charge alternates in sign, and the fractional excess decreases as k^2 , convergence is not affected adversely. Summation was stopped when the contribution of a shell was less than 10^{-6} of the aggregate sum. This required, on the average, about 4 shells. Allowing for normal round-off errors, we expect an accuracy, in final results, of five significant figures.

The computed expansion coefficients are exhibited in Tables I through XIII. These tables are reproductions of computer output. Each table is headed by the chemical formula, and by the length of the unit cell, in angstroms. Next follows a list of all atomic positions in the unit cell. The crystallographic data is taken from Wyckoff.⁶ Next follows a tabular listing of the expansion coefficients, i.e., the C_{lm} of (6). Due to evolution in notation, these coefficients are called *ALM* in Tables I–XIII. Both the real and imaginary parts are given for positive m ; our definition of the spherical harmonics requires $C_{l,-m} = C_{l,m}^*$. The coefficients are listed for the expansion of the potential at every inequivalent site. For purposes of this computation, sites whose environment differs only by rotations and/or reflections are considered equivalent. All the structures are cubic, except SnF_4 . This substance is tetragonal, with $c = 7.93 \text{ \AA}$, $a = 4.048 \text{ \AA}$. Since c is so nearly twice a , we have cheated a little by stacking together four of the tetragonal cells to make an almost cubic

TABLE I

				NA-CL			A=5.640					
NA	0.	0.	0.	0.500	0.500	0.	0.500	0.	0.500	0.	0.500	0.500
CL	0.500	0.500	0.500	0.	0.	0.500	0.500	0.	0.	0.	0.500	0.

L	M	ORIGIN		NA
		ALM	REAL	
0	0	-2.19679242		0.
1	0	0.		0.
1	1	0.		0.
2	0	0.		0.
2	1	0.		0.
2	2	0.		0.
3	0	0.		0.
3	1	0.		0.
3	2	0.		0.
3	3	0.		0.
4	0	-0.02371096		0.
4	1	0.		0.
4	2	0.		0.
4	3	0.		0.
4	4	-0.01417001		0.
5	0	0.		0.
5	1	0.		0.
5	2	0.		0.
5	3	0.		0.
5	4	0.		0.
5	5	0.		0.
6	0	-0.00068597		0.
6	1	0.		0.
6	2	0.		0.
6	3	0.		0.
6	4	0.00128333		0.
6	5	0.		0.
6	6	0.		0.

TABLE II

		CS-CL			A=4.123	
CS		0.	0.	0.		
CL		0.500	0.500	0.500		
L	M	ØRIGIN		CS		
		ALM	REAL	ALM	IMAG	
0	0	-1.	74998032	0.		
1	0	0.		0.		
1	1	0.		0.		
2	0	0.		0.		
2	1	0.		0.		
2	2	0.		0.		
3	0	0.		0.		
3	1	0.		0.		
3	2	0.		0.		
3	3	0.		0.		
4	0	0.	00924632	0.		
4	1	0.		0.		
4	2	0.		0.		
4	3	0.		0.		
4	4	0.	00552573	0.		
5	0	0.		0.		
5	1	0.		0.		
5	2	0.		0.		
5	3	0.		0.		
5	4	0.		0.		
5	5	0.		0.		
6	0	-0.	00020874	0.		
6	1	0.		0.		
6	2	0.		0.		
6	3	0.		0.		
6	4	0.	00039051	0.		
6	5	0.		0.		
6	6	0.		0.		

TABLE III

		ZN-S			A=5.409								
		0.	0.	0.	0.500	0.	0.500	0.500	0.500	0.	0.	0.500	0.500
		0.250	0.250	0.250	0.250	0.750	0.750	0.750	0.250	0.750	0.750	0.250	0.250
M		ØRIGIN		ZN									
		ALM	REAL	ALM	IMAG								
0		-4.	95817298	0.									
0		0.		0.									
1		0.		0.									
0		0.		0.									
1		0.		0.									
2		0.		0.									
0		0.		0.									
1		0.		0.									
2		0.		0.	17556879								
3		0.		0.									
0		0.	04689946	0.									
1		0.		0.									
2		0.		0.									
3		0.		0.									
4		0.	02802778	0.									
0		0.		0.									
1		0.		0.									
2		0.		0.	02525550								
3		0.		0.									
4		0.		0.									
5		0.		0.									
0		-0.	00491204	0.									
1		0.		0.									
2		0.		0.									
3		0.		0.									
4		0.	00918959	0.									
5		0.		0.									
6		0.		0.									

TABLE VI

K-TA-03				A=3.988							
TA	0.	0.	0.	0.	0.500	0.	0.	0.	0.500		
B	0.500	0.	0.	0.	0.500	0.	0.	0.	0.500		
K	0.500	0.500	0.500								
M	ORIGIN			TA	ORIGIN			B	ORIGIN		K
0	ALM	REAL		ALM	REAL		ALM	REAL	ALM	REAL	ALM
0	-12.80986524	0.	0.	0.	6.17036784	0.	0.	-2.97906271	0.	0.	0.
0	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
1	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0	0.	0.	0.	0.	-0.77709828	0.	0.	0.	0.	0.	0.
1	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
2	0.	0.	0.	0.	0.95174713	0.	0.	0.	0.	0.	0.
0	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
1	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
2	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
3	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0	-0.25413066	0.	0.	0.	0.13074156	0.	0.	-0.00784040	0.	0.	0.
1	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
2	0.	0.	0.	0.	-0.18082657	0.	0.	0.	0.	0.	0.
3	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
4	-0.15187213	0.	0.	0.	0.21482505	0.	0.	-0.00468555	0.	0.	0.
0	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
1	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
2	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
3	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
4	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
5	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0	-0.01117794	0.	0.	0.	-0.02269032	0.	0.	0.00485897	0.	0.	0.
1	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
2	0.	0.	0.	0.	0.02488248	0.	0.	0.	0.	0.	0.
3	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
4	0.02091200	0.	0.	0.	-0.03023670	0.	0.	-0.00909032	0.	0.	0.
5	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
6	0.	0.	0.	0.	0.03690669	0.	0.	0.	0.	0.	0.

TABLE VII

SR-TI-03				A=3.905							
TI	0.	0.	0.	0.	0.500	0.	0.	0.	0.500		
B	0.500	0.	0.	0.	0.500	0.	0.	0.	0.500		
SR	0.500	0.500	0.500								
M	ORIGIN			TI	ORIGIN			B	ORIGIN		SR
0	ALM	REAL		ALM	REAL		ALM	REAL	ALM	REAL	ALM
0	-11.23581445	0.	0.	0.	5.86043859	0.	0.	-4.89031279	0.	0.	0.
0	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
1	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0	0.	0.	0.	0.	-0.57006907	0.	0.	0.	0.	0.	0.
1	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
2	0.	0.	0.	0.	0.69818916	0.	0.	0.	0.	0.	0.
0	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
1	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
2	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
3	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0	-0.29458212	0.	0.	0.	0.10063877	0.	0.	0.00341623	0.	0.	0.
1	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
2	0.	0.	0.	0.	-0.18383221	0.	0.	0.	0.	0.	0.
3	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
4	-0.17604650	0.	0.	0.	0.19910725	0.	0.	0.00204157	0.	0.	0.
0	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
1	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
2	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
3	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
4	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
5	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
0	-0.01265471	0.	0.	0.	-0.02098419	0.	0.	0.00532835	0.	0.	0.
1	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
2	0.	0.	0.	0.	0.02283370	0.	0.	0.	0.	0.	0.
3	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
4	0.02367478	0.	0.	0.	-0.02744369	0.	0.	-0.00996845	0.	0.	0.
5	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
6	0.	0.	0.	0.	0.03386784	0.	0.	0.	0.	0.	0.

TABLE VIII

		AG2-B3			A=4.940									
AG		0.250	0.250	0.250	0.250	0.750	0.750	0.750	0.750	0.250	0.750	0.750	0.750	0.250
B		0.500	0.500	0.	0.500	0.	0.	0.	0.	0.	0.500	0.	0.500	0.
		ORIGIN			AG		ORIGIN			B				
L	M	ALM REAL			ALM IMAG		ALM REAL			ALM IMAG				
0	0	-8.14379931			0.		5.44399786			0.				
1	0	0.			0.		0.			0.				
1	1	0.			0.		0.			0.				
2	0	0.			0.		0.28160676			0.				
2	1	0.22806948			0.22806949		0.			0.				
2	2	0.			0.22806947		0.			0.				
3	0	0.			0.		0.			0.				
3	1	0.			0.		0.			0.				
3	2	0.			0.		0.			-0.37861259				
3	3	0.			0.		0.			0.				
4	0	0.11074642			0.		-0.14234010			0.				
4	1	-0.01359074			-0.01359070		0.			0.				
4	2	0.			0.03844038		0.			0.				
4	3	-0.03595767			0.03595764		0.			0.				
4	4	0.06618364			0.		-0.12796166			0.				
5	0	0.			0.		0.			0.				
5	1	0.			0.		0.			0.				
5	2	0.			0.		0.			-0.06530282				
5	3	0.			0.		0.			0.				
5	4	0.			0.		0.			0.				
5	5	0.			0.		0.			0.				
6	0	-0.01390760			0.		0.01744156			0.				
6	1	-0.00355838			-0.00355838		0.			0.				
6	2	0.			-0.00533705		0.			0.				
6	3	-0.00168260			0.00168259		0.			0.				
6	4	0.02601873			0.		-0.01976332			0.				
6	5	-0.00459289			-0.00459289		0.			0.				
6	6	0.			-0.00284672		0.			0.				

TABLE IX

		NA-PT3-B			A=5.689									
NA		0.	0.	0.	0.500	0.500	0.500	0.500	0.500	0.250	0.750	0.	0.500	
PT		0.250	0.	0.500	0.500	0.250	0.	0.	0.500	0.250	0.750	0.	0.500	
B		0.250	0.250	0.250	0.250	0.750	0.750	0.750	0.750	0.250	0.750	0.250	0.750	
		ORIGIN			NA		ORIGIN			PT		ORIGIN		
L	M	ALM REAL			ALM IMAG		ALM REAL			ALM IMAG		ALM REAL		
0	0	-3.35616586			0.		-8.34012449			0.		5.5515836		0.
1	0	0.			0.		0.			0.		0.		0.
1	1	0.			0.		0.			0.		0.		0.
2	0	0.			0.		-0.36297223			0.		0.		0.
2	1	0.			0.		0.			0.		-0.20518058		-0.20518055
2	2	0.			0.		0.44454841			0.		-0.20518056		0.
3	0	0.			0.		0.			0.		0.		0.
3	1	0.			0.		-0.01247910			0.		-0.13190589		0.13190588
3	2	0.			0.		0.			0.		0.16684922		0.
3	3	0.			0.		-0.00966616			0.		-0.10217387		-0.10217386
4	0	0.08977990			0.		0.12715140			0.		-0.08866502		0.
4	1	0.			0.		0.13210720			0.		-0.01313988		-0.01313991
4	2	0.			0.		0.			0.		0.		0.03716524
4	3	0.			0.		0.			0.		-0.03476488		0.03476491
4	4	0.05365372			0.		-0.02387614			0.		-0.05298738		0.
5	0	0.			0.		0.			0.		0.		0.
5	1	0.			0.		-0.00066188			0.		0.01643208		-0.01643208
5	2	0.			0.		0.			0.		0.05803908		0.
5	3	0.			0.		0.			0.		-0.01777077		-0.01777077
5	4	0.			0.		0.			0.		0.		0.
5	5	0.			0.		0.00068512			0.		-0.01700882		0.01700882
6	0	-0.00708890			0.		0.00653744			0.		-0.01030961		0.
6	1	0.			0.		0.			0.		0.00545928		0.00545928
6	2	0.00276061			0.		0.00421905			0.		0.		-0.00479379
6	3	0.			0.		0.			0.		-0.00740649		0.
6	4	0.01326215			0.		-0.02455502			0.		0.01928758		0.
6	5	0.			0.		0.			0.		-0.00461697		-0.00461697
6	6	-0.00186120			0.		0.00625787			0.		0.		0.00773438

TABLE X

K-FE-82												A=7.960											
E																							
0.	0.	0.	0.	0.	0.500	0.500	0.500	0.250	0.250	0.250	0.250	0.250	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	
0.500	0.	0.500	0.	0.500	0.	0.500	0.	0.750	0.250	0.250	0.250	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	
0.	0.	0.500	0.	0.500	0.	0.500	0.	0.250	0.250	0.750	0.250	0.250	0.750	0.750	0.750	0.250	0.750	0.750	0.250	0.250	0.250	0.250	
0.500	0.	0.	0.500	0.500	0.500	0.500	0.500	0.750	0.250	0.250	0.250	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	
0.125	0.125	0.125	0.125	0.375	0.375	0.375	0.375	0.125	0.625	0.625	0.625	0.125	0.875	0.875	0.875	0.125	0.875	0.875	0.625	0.875	0.875	0.625	
0.375	0.125	0.375	0.375	0.375	0.375	0.125	0.375	0.375	0.625	0.875	0.875	0.375	0.625	0.875	0.625	0.375	0.875	0.875	0.625	0.875	0.875	0.625	
0.625	0.125	0.625	0.625	0.375	0.875	0.875	0.625	0.625	0.625	0.125	0.625	0.625	0.125	0.625	0.875	0.625	0.875	0.875	0.625	0.875	0.375	0.375	
0.875	0.125	0.875	0.875	0.875	0.375	0.625	0.625	0.875	0.625	0.375	0.875	0.875	0.375	0.875	0.875	0.875	0.875	0.875	0.875	0.875	0.125	0.125	
M	BRIGIN			FE			BRIGIN			K			BRIGIN			B							
	ALM REAL			ALM IMAG			ALM REAL			ALM IMAG			ALM REAL			ALM IMAG							
0	-9.58162487			0.			-1.95541081			0.			6.05419123			0.							
0	0.			0.			0.			0.			0.			0.							
1	0.			0.			0.			0.			0.			0.							
0	0.			0.			0.			0.			0.			0.							
1	0.			0.			0.			0.			0.			0.							
2	0.			0.			0.			0.			0.57498499			0.57498499							
0	0.			0.			0.			0.			0.			0.							
1	0.			0.			0.			0.			0.			0.							
2	0.			0.62112275			0.			-0.02205049			0.			0.							
3	0.			0.			0.			0.			0.			0.							
0	0.23410912			0.			-0.01350321			0.			-0.16853560			0.							
1	0.			0.			0.			0.			-0.06204139			-0.06204142							
2	0.			0.			0.			0.			0.			0.17547956							
3	0.			0.			0.			0.			-0.16414609			0.16414613							
4	0.13990695			0.			-0.00806973			0.			-0.10071925			0.							
0	0.			0.			0.			0.			0.			0.							
1	0.			0.			0.			0.			0.			0.							
2	0.			0.16091968			0.			-0.00173405			0.			0.							
3	0.			0.			0.			0.			0.			0.							
4	0.			0.			0.			0.			0.			0.							
5	0.			0.			0.			0.			0.			0.							
0	-0.03802649			0.			0.00061209			0.			-0.03073002			0.							
1	0.			0.			0.			0.			-0.02280463			-0.02280463							
2	0.			0.			0.			0.			-0.03775779			-0.03775779							
3	0.			0.			0.			0.			-0.01315276			0.01315277							
4	0.07114104			0.			-0.00114511			0.			-0.05749060			0.							
5	0.			0.			0.			0.			-0.03220146			-0.03220146							
6	0.			0.			0.			0.			0.			-0.01744504							

TABLE XI

SI-82												A=7.160											
I																							
0.	0.	0.	0.	0.	0.500	0.500	0.500	0.250	0.250	0.250	0.250	0.250	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	
0.500	0.	0.500	0.	0.500	0.	0.500	0.	0.750	0.250	0.250	0.250	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.750	
0.125	0.125	0.125	0.125	0.375	0.375	0.375	0.375	0.125	0.625	0.625	0.625	0.125	0.875	0.875	0.875	0.125	0.875	0.875	0.625	0.875	0.875	0.625	
0.375	0.125	0.375	0.375	0.375	0.375	0.125	0.375	0.375	0.625	0.875	0.875	0.375	0.625	0.875	0.625	0.375	0.875	0.875	0.625	0.875	0.875	0.625	
0.625	0.125	0.625	0.625	0.375	0.875	0.875	0.625	0.625	0.625	0.125	0.625	0.625	0.125	0.625	0.875	0.625	0.875	0.875	0.625	0.875	0.375	0.375	
0.875	0.125	0.875	0.875	0.875	0.375	0.625	0.625	0.875	0.625	0.375	0.875	0.875	0.375	0.875	0.875	0.875	0.875	0.875	0.875	0.875	0.125	0.125	
M	BRIGIN			SI			BRIGIN			B													
	ALM REAL			ALM IMAG			ALM REAL			ALM IMAG													
0	-12.38263190			0.			7.98499638			0.													
0	0.			0.			0.			0.													
1	0.			0.			0.			0.													
0	0.			0.			0.			0.													
1	0.			0.			1.18484712			1.18484712													
2	0.			0.			0.			1.18484716													
0	0.			0.			0.			0.													
1	0.			0.			0.			0.													
2	0.			0.89161128			0.			0.													
3	0.			0.			0.			0.													
0	0.39038371			0.			-0.39493667			0.													
1	0.			0.			-0.13557618			-0.13557624													
2	0.			0.			0.			0.38346738													
3	0.			0.			-0.35870086			0.35870098													
4	0.23329886			0.			-0.23601974			0.													
0	0.			0.			0.			0.													
1	0.			0.			0.			0.													
2	0.			0.29911921			0.			0.													
3	0.			0.			0.			0.													
4	0.			0.			0.			0.													
5	0.			0.			0.			0.													
0	-0.07994447			0.			0.08471100			0.													
1	0.			0.			-0.06395741			-0.06395740													
2	0.			0.			0.			-0.10588079													
3	0.			0.			-0.03687867			0.03687868													
4	0.14956238			0.			-0.15847975			0.													
5	0.			0.			-0.09030070			-0.09030071													
6	0.			0.			0.			-0.04892916													

TABLE XII

		AL2-MG-89						A=8.059											
AL		0.625	0.625	0.625	0.625	0.875	0.875	0.875	0.625	0.875	0.875	0.875	0.875	0.875	0.625				
		0.625	0.125	0.125	0.625	0.375	0.375	0.875	0.125	0.375	0.375	0.875	0.375	0.875	0.125				
		0.125	0.625	0.125	0.125	0.875	0.375	0.375	0.625	0.625	0.375	0.375	0.375	0.875	0.625				
		0.125	0.125	0.625	0.125	0.375	0.875	0.875	0.375	0.125	0.875	0.375	0.375	0.375	0.625				
MG		0.	0.	0.	0.	0.500	0.500	0.500	0.	0.500	0.500	0.500	0.500	0.500	0.				
		0.250	0.250	0.250	0.250	0.750	0.750	0.750	0.750	0.250	0.750	0.750	0.750	0.750	0.250				
B		0.375	0.375	0.375	0.375	0.625	0.625	0.625	0.375	0.625	0.625	0.625	0.625	0.625	0.375				
		0.875	0.875	0.875	0.875	0.625	0.625	0.625	0.875	0.625	0.625	0.625	0.625	0.625	0.875				
		0.375	0.875	0.875	0.375	0.125	0.125	0.125	0.625	0.875	0.125	0.625	0.625	0.125	0.875				
		0.875	0.375	0.375	0.875	0.125	0.125	0.125	0.625	0.375	0.125	0.625	0.625	0.125	0.375				
		0.875	0.375	0.875	0.875	0.625	0.125	0.125	0.125	0.375	0.125	0.125	0.625	0.625	0.875				
		0.375	0.875	0.375	0.375	0.625	0.125	0.625	0.125	0.875	0.625	0.125	0.125	0.375	0.375				
		0.875	0.875	0.375	0.875	0.125	0.625	0.625	0.125	0.375	0.625	0.625	0.125	0.125	0.875				
		0.375	0.375	0.875	0.375	0.125	0.625	0.625	0.125	0.375	0.625	0.625	0.125	0.125	0.875				
L	M	ORIGIN			AL			ORIGIN			MG			ORIGIN			B		
		ALM REAL			ALM IMAG			ALM REAL			ALM IMAG			ALM REAL			ALM IMAG		
0	0	-7.90614063			0.			-7.72952527			0.			6.27619386			0.		
1	0	0.			0.			0.			0.			-0.48349914			0.		
1	1	0.			0.			0.			0.			-0.34188542			-0.3418855		
2	0	0.			0.			0.			0.			0.			0.		
2	1	0.14782479			0.14782480			0.			0.			0.21011862			0.2101186		
2	2	0.			0.14782477			0.			0.			0.			0.2101186		
3	0	0.			0.			0.			0.			-0.34034213			0.		
3	1	0.			0.			0.			0.			0.14737244			0.1473722		
3	2	0.			0.			0.			0.			-0.52087519			0.		
3	3	0.			0.			0.			0.			-0.19025676			0.1902568		
4	0	-0.25165626			0.			0.22841588			0.			0.13994851			0.		
4	1	0.00384695			0.00384693			0.			0.			-0.01881876			-0.0188187		
4	2	0.			-0.01088079			0.			0.			0.			0.0532275		
4	3	0.01017814			-0.01017799			0.			0.			-0.04978986			0.0497897		
4	4	-0.15039334			0.			0.13650463			0.			0.08363526			0.		
5	0	0.			0.			0.			0.			-0.05583144			0.		
5	1	0.			0.			0.			0.			-0.03652287			-0.0365228		
5	2	0.			0.			0.			-0.14423722			0.			0.0354473		
5	3	0.			0.			0.			0.			-0.02353180			0.0235318		
5	4	0.			0.			0.			0.			-0.03039386			0.		
5	5	0.			0.			0.			0.			-0.03370390			-0.0337039		
6	0	-0.01370632			0.			-0.03512712			0.			0.02047696			0.		
6	1	-0.00176585			-0.00176584			0.			0.			-0.00740547			-0.0074054		
6	2	0.			0.00169610			0.			0.			0.			-0.01110711		
6	3	0.00206141			-0.00206143			0.			0.			-0.00350171			0.0035017		
6	4	0.02564218			0.			0.06571680			0.			-0.03830889			0.		
6	5	0.00110306			0.00110306			0.			0.			-0.00955843			-0.0095584		
6	6	0.			-0.00238906			0.			0.			0.			-0.0059244		

cell of average dimension 8 \AA . We have included SnF_4 mainly as an illustration of a case where the actual periodicity is not the same in all dimensions.

For purposes of checking, there exist simple relations between the Madelung constant and our coefficients C_{00} . The Madelung constants for NaCl , CsCl , ZnS , CaF_2 , Cu_2O hark back to an extensive tabulation by J. Sherman.⁷ (Sherman's number for Cu_2O , quoted throughout the literature, e.g., by Born and Huang,⁸ is in error. The correct value has been given by Hund.⁹ Very accurate computations for some cubic crystals have been made by Benson and Zeggeren.¹⁰ For MX compounds, the Madelung constant is related to our C_{00} by a simple multiplicative factor. This factor is $(4\pi)^{1/2}$ times the ionic charge divided by the length of the unit cell. For compounds like CaF_2 and Cu_2O , the Madelung constant is related in the same fashion to the average of the absolute values of the C_{00} 's for the various sites. For Cu_2O , the potential at each site separately has more recently been computed by Dahl.¹¹ Madelung calculations for the perovskites have been considered by Templeton,¹² Fumi and Tosi,¹³ and Cowley.¹⁴ Cowley has calculated numerically the

TABLE XIII

SN-F4				A=8.000									
0.	0.	0.	0.250	0.250	0.250	0.500	0.500	0.	0.	0.	0.750	0.	0.
0.	0.500	0.	0.250	0.750	0.750	0.	0.500	0.500	0.	0.	0.750	0.500	0.
0.	0.250	0.	0.250	0.	0.	0.	0.250	0.	0.500	0.	0.	0.250	0.500
0.500	0.250	0.	0.750	0.	0.	0.	0.750	0.	0.	0.500	0.	0.250	0.500
0.	0.750	0.	0.250	0.500	0.	0.	0.250	0.500	0.500	0.	0.	0.750	0.500
0.500	0.750	0.	0.750	0.500	0.	0.	0.750	0.500	0.500	0.	0.	0.500	0.500
0.	0.	0.250	0.	0.	0.	0.750	0.250	0.250	0.250	0.750	0.250	0.250	0.250
0.500	0.	0.250	0.500	0.	0.	0.750	0.750	0.250	0.250	0.750	0.750	0.250	0.250
0.	0.500	0.250	0.	0.	0.500	0.750	0.250	0.750	0.750	0.750	0.250	0.750	0.250
0.500	0.500	0.250	0.500	0.500	0.500	0.750	0.750	0.750	0.750	0.750	0.750	0.750	0.250

N	ORIGIN		SN	ORIGIN		F1	ORIGIN		F2
	ALM	REAL	ALM	REAL	ALM	REAL	ALM	REAL	ALM
0	-6.58065987	0.	0.	5.55493051	0.	2.55387265	0.		
0	0.	0.	0.	0.	0.	2.36476621	0.		
1	-0.52353968	0.33058744	0.	-0.27535433	0.	-0.42667971	0.		
0	-0.34731891	0.	0.	-0.93092058	0.	0.74515594	0.		
1	0.	0.	0.	0.	0.	0.	0.10740093		
2	0.	-0.14623467	0.	-0.92373976	0.	0.	0.		
0	0.	0.	0.	0.	0.	0.29895561	0.		
1	0.01702513	-0.03159145	0.	-0.01105826	0.	-0.01496998	0.		
2	0.	0.	0.	0.	0.	0.	-0.03669877		
3	0.03774419	0.03278823	0.	0.01427609	0.	0.01932615	0.		
0	-0.09091583	0.	0.	0.14195608	0.	0.16193924	0.		
1	0.	0.	0.	0.	0.	0.	-0.00244038		
2	0.	0.01162022	0.	0.11792579	0.	0.	0.		
3	0.	0.	0.	0.	0.	0.	0.00645663		
4	-0.07077464	0.	0.	0.16640348	0.	0.01216549	0.		
0	-0.00211426	0.00191500	0.	-0.00070479	0.	0.06735962	0.		
1	0.	0.	0.	0.	0.	0.00125949	0.		
2	0.	0.	0.	0.	0.	0.	-0.00241358		
3	-0.00041911	-0.00083644	0.	-0.00031195	0.	0.00196165	0.		
4	0.	0.	0.	0.	0.	0.00081620	0.		
5	0.00288825	-0.00288930	0.	-0.00042522	0.	0.00057840	0.		
0	-0.00668188	0.	0.	-0.01733700	0.	0.03332756	0.		
1	0.	0.	0.	0.	0.	0.	-0.00024364		
2	0.	-0.00081959	0.	-0.01953730	0.	0.	0.		
3	0.	0.	0.	0.	0.	0.	0.00011520		
4	0.01081367	0.	0.	-0.02548826	0.	-0.00445095	0.		
5	0.	0.	0.	0.	0.	0.	-0.00031447		
6	0.	0.00129473	0.	-0.02929148	0.	0.	0.		

"partial Madelung coefficients" as defined by Templeton, and Fumi and Tosi have shown how the potential at any site of an arbitrary perovskite structure can be obtained by linear combinations of these partial Madelung coefficients. Our numerical comparison for the perovskite C_{00} 's is therefore, essentially with Cowley's calculation. Comparison of our C_{00} 's

TABLE XIV—COMPARISON BETWEEN PRESENT COMPUTATION AND PREVIOUS WORK FOR THE COEFFICIENTS C_{00}

Substance	Site	Present Result	Previous Result	Source for Previous Result
NaCl	either	2.19679	2.1968	Sherman
CsCl	either	1.74998	1.75003	Benson, Zeggeren
ZnS	either	4.95817	4.95845	Benson, Zeggeren
CaF ₂	$\left[\begin{array}{c} \text{Ca} \\ \text{Cu} \end{array} \right] \begin{array}{c} + \\ + \end{array} \left[\begin{array}{c} \text{F} \\ \text{O} \end{array} \right]$	7.55090	7.55090	Benson, Zeggeren
Cu ₂ O			8.51809	8.5172
Cu ₂ O	Cu	3.14084	3.1406	Dahl
Cu ₂ O	O	5.37725	5.3768	Dahl
KTaO ₃	K	2.97906	2.9765	Cowley
KTaO ₃	Ta	12.80987	12.8115	Cowley
KTaO ₃	O	6.17037	6.1711	Cowley
SrTiO ₃	Sr	4.89031	4.8905	Cowley
SrTiO ₃	Ti	11.2358	11.2362	Cowley
SrTiO ₃	O	5.86044	5.8606	Cowley

with previous work is made in Table XIV. In general, some divergence appears in the fifth significant figure.

We have ourselves checked a portion of our results by using the Evjen summation method. This method is particularly useful for coefficients with higher l -values, but converges quite slowly for C_{00} . Only for CsCl and NaCl were we able, in reasonable computing time, to carry the summation to a sufficient degree of convergence for C_{00} . The Evjen results confirm our Ewald results for these substances, for all the coefficients. For the perovskites, we pushed the summation at the Ta site as far as the ninth Evjen shell (24,564 neighbors), but for C_{00} shell-to-shell fluctuations were still in the fourth significant figure. For higher l -values, the Evjen summation was able to confirm our Ewald results for all coefficients, for all sites, for both perovskites. The same is true for Cu_2O . It should be pointed out that the computation of the coefficients (exclusive of C_{00}) takes about ten minutes of 7094 time per crystal if the Evjen method is used. We attempted a check on one site of Al_2MgO_4 . The computer was stopped after 20 minutes, at which time usefully convergent results were obtained only for coefficients with $l \geq 3$.

We note that the appearance of coefficients of odd l , at any site, is associated with lack of inversion symmetry.

We believe that the computations we have discussed show that in certain applications our method of doing crystal sums offers definite advantages, both in accuracy and in time. Suitable applications are those where a number of structures must be considered that can be described by basically the same coordinate grid, especially if the number of ions per unit cell is large.

REFERENCES

1. Grant, W. J. C., Computation of Lattice Sums: Generalization of the Ewald Method, B.S.T.J., 44, March, 1965, p. 427.
2. Nijboer, B. R. A. and Dewette, F. W., Physica, 23, 1957, p. 309.
3. Adler, S., Physica, 27, 1961, p. 1193.
4. Barlow, C. A. and MacDonald, J. R., J. Chem. Phys. 40, 1964, p. 1535.
5. Graham, R. L., On the Decomposition of the Lattice-Periodic Functions, B.S.T.J., 44, July-Aug., 1965, p. 1191.
6. Wyckoff, R., *Crystal Structures*, (2nd ed.) Interscience Publishers, N. Y., 1963.
7. Sherman, J., Chem. Revs., 11, 1932, p. 93.
8. Born, M. and Huang, K., *Dynamical Theory of Crystal Lattices*, Oxford University Press.
9. Hund, F. Z., Physik., 94, 1935, p. 11.
10. Benson, G. C. and van Zeggeren, F., J. Chem. Phys., 26, 1957, p. 1083; *ibid*, 26, 1957, p. 1077.
11. Dahl, J. P., Fifty-first Quarterly Program Report, M.I.T. Solid State on Molecular Theory Group, January, 1964, p. 69.
12. Templeton, D. H., J. Chem. Phys., 23, 1955, p. 1826.
13. Fumi, F. G. and Tosi, M. P., J. Chem. Phys., 33, 1960, p. 1.
14. Cowley, R. A., Acta Cryst., 15, 1962, p. 687.

Contributors To This Issue

DIMITRI S. BUGNOLO, B.S.E.E., 1952, University of Pennsylvania; M.Eng., 1955, Yale University; Eng.Sc.D., 1960, Columbia University; Bell Telephone Laboratories, 1960-1961, 1963-1965. He has been active in various areas of electromagnetic theory and the theory of turbulence in reacting gases. Member, IEEE, American Physical Society, American Geophysical Union, Commission 2 of URSL.

J. A. COLLINSON, A.B., 1950, Oberlin College; M.S., 1951, Yale University; Ph.D., 1954, Yale University; Bell Telephone Laboratories, 1962—. He has worked on gas lasers and placed emphasis on frequency characteristics and atmospheric transmission of laser beams. Member, American Physical Society, Sigma Xi, Phi Beta Kappa.

I. DOSTIS, B.E.E., 1959, College of the City of New York; M.E.E., 1961, New York University; Western Electric Company 1958; Bell Telephone Laboratories 1959-1962; College of the City of New York 1962-1963; Bell Telephone Laboratories 1963—. Mr. Dostis first worked on waveguide filter design. He is presently engaged in the investigation of PCM system problems. Member, Eta Kappa Nu, Tau Beta Pi, IEEE.

WALTER J. C. GRANT, A.B., 1951, M.A., 1952, Boston College; B.S., 1958, Ph.D., 1962, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1962-1965; Wayne State University, 1965—. He has been engaged in theoretical study of paramagnetic and electro-optic solid-state devices. He is presently associate Professor of Physics at Wayne State University. Member, American Physical Society, Sigma Xi.

H. D. GRIFFITHS, B.Sc., 1949, University of Western Ontario; M.Sc., 1950, McGill University; Bell Telephone Laboratories, 1954—. Mr. Griffiths initially-designed broadband IF equipment for the TH Radio System. In 1957, he became responsible for group engaged in final design of the TH Radio System. Later, he was responsible for the design of the 100A Protection Switching System. He is now responsible for test equipment for the TD-3 Radio System. Member, IEEE.

HOWARD D. HELMS, B.S.E., 1956, Ph.D., 1961, Princeton University; Bell Telephone Laboratories, 1959—. Mr. Helms has considered questions arising in the interpolation of sampled data and in the design of equations for processing radar data. At present, as a consulting engineer-statistician, he is helping to devise digital computer simulations of systems for military defense. Member, Phi Beta Kappa, Sigma Xi, IEEE, AIAA, AAAS.

SHEN LIN, B.S. (summa cum laude), 1951, University of the Philippines; M.A., 1953, Ph.D., 1963, Ohio State University; Bell Telephone Laboratories, 1963—. He has worked in the field of Turing machine theory, combinatorial analysis and number theory. At present, he is working on applications of computers in various optimization and number-theoretic problems. Member, American Mathematical Society, Mathematical Association of America, SIAM, Phi Kappa Phi, Sigma Pi Sigma, Pi Mu Epsilon.

JOSEPH NEDELKA, E.E., 1937, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1922—. Mr. Nedelka initially worked on the equipment design of various telephone systems. During the war, his work involved the equipment design of military radar systems. Since that time, he has worked on the equipment design of overseas radio-telephone equipment, central office TV equipment, and N, O, L, and T carrier equipment. He is presently engaged in equipment design of long-haul radio equipment with the radio systems group.

JACQUES RENAULT, B.S., 1954, M.S., 1955, University of Chicago; Ph.D., 1960, Cornell University; Bell Telephone Laboratories, 1963—. Earlier Mr. Renault was engaged in the analysis of interaction of EM waves with ionized gases. Recently, he has been engaged in experimental and theoretical investigation of scattering of EM waves from rough surfaces. Member, American Geophysical Union, Sigma Xi (inactive).

JAMES W. SMITH, B.E.S., 1956, Dr. Eng., 1963, Johns Hopkins University; Bell Telephone Laboratories, 1963—. He has been concerned with various analysis problems in the areas of analog and digital data transmission. Member, IEEE, Tau Beta Pi, Sigma Xi, Eta Kappa Nu.

ROBERT S. WEINER, B.M.E., 1957, M.S. (Mechanics), 1959, Rensselaer Polytechnic Institute; Ph.D. (Mechanics), 1962, Northwestern University; Bell Telephone Laboratories, 1963—. He has been concerned

with thermal and structural analysis of array radar components and structures that are subjected to both normal operating environments and transient blast and shock environments. Member, ASME, Pi Tau Sigma, Tau Beta Pi, Sigma Xi.

ERIC WOLMAN, A.B., 1953, A.M., 1954, Ph.D., 1957, Harvard University; Bell Telephone Laboratories, 1957—. He was first associated with data systems engineering, and headed a group responsible for studying applications of queuing theory and coding theory. Now a member of the activities analysis research department, he is investigating the flow of traffic in data communication systems and in time-shared digital computers. He was a visiting lecturer in applied mathematics at Harvard University for the spring term of 1964. Member, AAAS, AMS, IEEE, Phi Beta Kappa, Sigma Xi, Society of Harvard Engineers and Scientists, SIAM.

B.S.T.J. BRIEFS

Holographic Image Projection through Inhomogeneous Media

By H. KOGELNIK

(Manuscript received October 18, 1965)

This brief reports on an experimental test of a hologram technique that is capable of projecting undistorted images through inhomogeneous media. Coherent laser light and the hologram order that produces a conjugate image are used. The technique may be of interest in controlling wave transmission through inhomogeneous media for communication purposes, and it allows the scrambling or encrypting of texts, drawings, and similar messages for private communication.

In common holography there are two coherent waves incident upon the hologram or recording plate.¹ These are a signal wave S , originating from an object of interest, and a reference or background wave R . A recording is made of the intensity pattern

$$I = R.R^* + S.S^* + R.*S + R.S^* \quad (1)$$

where R and S are the complex amplitude patterns of the two waves across the plate, and the asterisk denotes a complex conjugate. Illumination of the developed hologram produces waves of several orders. The order that corresponds to the third term in (1) reconstructs the original wavefront of S and produces a true image. The order corresponding to the last term in (1) produces a conjugate image. This latter hologram order is used in the present technique of image projection through inhomogeneous media.

The inhomogeneities of a lossless optical medium can be described by the spatial variation of the refractive index $n(\mathbf{r})$. For inhomogeneities of geometrical dimensions that are much greater than the wavelength the scalar wave equation

$$\nabla^2 u + k^2 n^2(\mathbf{r}) u = 0 \quad (2)$$

is a good description of wave propagation in the medium.² Here $k = 2\pi/\lambda$ is the propagation constant in free space, and u is a vector component or potential of the wave field. A wave that propagates through the medium more or less in the z direction is represented by a solution of the form

$$u = S(\mathbf{r}) \cdot \exp(-j\beta z) \quad (3)$$

where S is a slowly varying (complex) function of \mathbf{r} . This solution of the wave equation (2) has a conjugate solution

$$u^* = S^*(\mathbf{r}) \cdot \exp(j\beta z) \quad (4)$$

which corresponds to a wave that travels in the opposite direction. The ray families associated with the two conjugate solutions have the same ray paths, which they follow in opposite directions. The intensity patterns $u \cdot u^*$ of the two waves are the same.

An experiment was performed in which the wave S originating from an object was recorded on a hologram after it had passed through the inhomogeneous medium. The return wave S^* was launched by means of the conjugate-image hologram order. In the original object plane the intensity pattern of the return wave reproduces the object, which was observed.

Fig. 1 shows the experimental arrangement. The object is a transparent text on an opaque background. In recording the hologram the object was diffusely illuminated by light from a 6328 Å gas laser. The

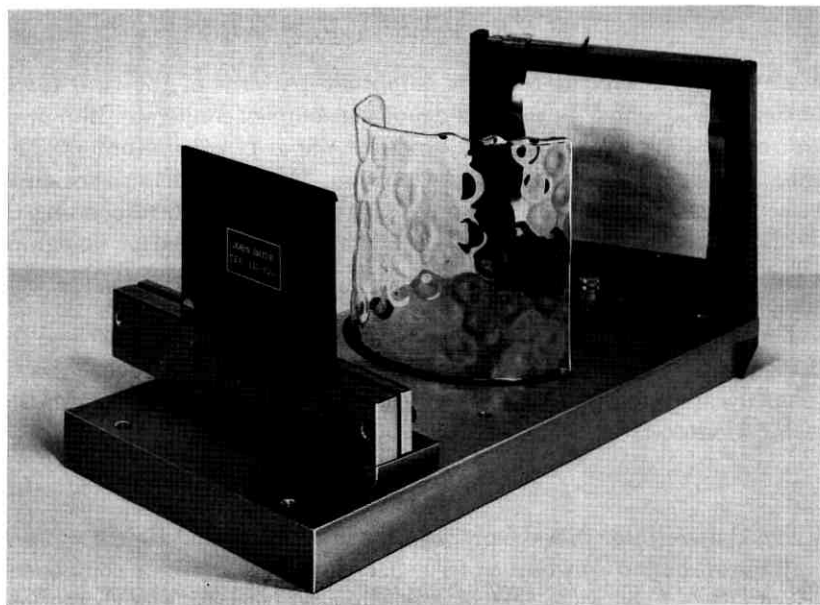


Fig. 1—Experimental arrangement with transparent object text on the left, distorting warped glass in the middle, and hologram plate on the right.

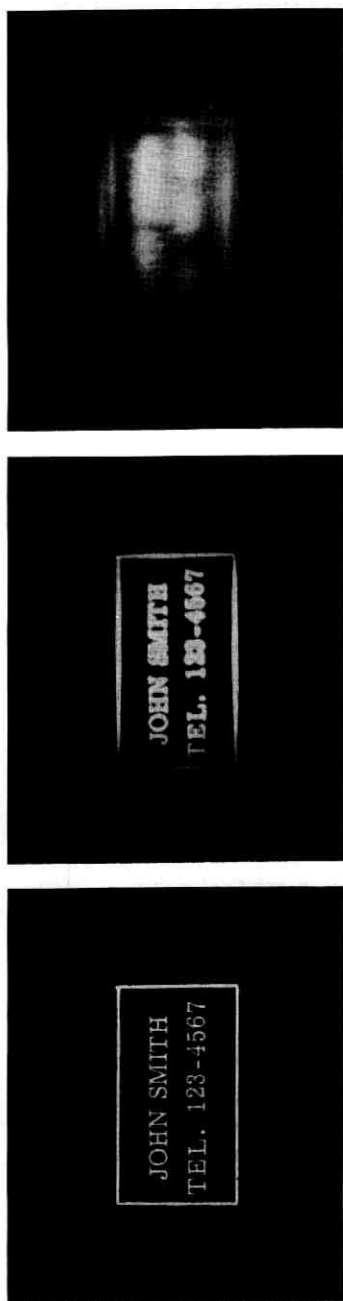
light from the object passed through a warped glass plate which represents the inhomogeneous medium. The hologram was recorded on a Kodak 649-F spectroscopic plate in the presence of a coherent plane reference wave³ incident at an angle of about 50 degrees of arc with respect to the rays from the object.

Fig. 2 gives an indication of the distortions introduced by the warped glass plate. It is a photograph of the object text taken through the glass plate.

In Fig. 3 the original object is reproduced on the left. The text is framed by a transparent rectangle which measures $\frac{3}{4}$ by $1\frac{1}{2}$ inches. A photograph of the intensity pattern of the return wave in the object plane is shown in the middle of the figure. For easier focusing, a diffusing plate was inserted in the object plane (in place of the object). The return wave was produced by illuminating the developed hologram with a plane beam of laser light antiparallel to the reference beam used in recording. Care must be taken to preserve the relative positions of the warped glass plate and the hologram plate, and the direction of the illuminating beam had to be adjusted with a tolerance of about ± 10 minutes of arc to achieve a good reconstruction of the object. The pattern produced in the



Fig. 2—Object text photographed through warped glass plate.



(a)

(b)

(c)

Fig. 3—(a) left—original text; (b) middle—reconstruction of text projected through warped glass; (c) right—scrambled reconstruction of text seen when warped glass is removed.

object plane when the same return wave is launched from the hologram with the warped glass plate removed is shown on the right in Fig. 3. The text is now scrambled and unreadable. It can be read only if the glass plate is reinserted in its original position.

REFERENCES

1. Gabor, D., Microscopy by Reconstructed Wavefronts, Proc. Roy. Soc., *A* 197, July, 1949, p. 454-487; Proc. Phys. Soc., *B* 64, June, 1951, p. 449-469.
2. Tatarski, V. I., *Wave Propagation in a Turbulent Medium*, McGraw-Hill Book Co., New York, 1961, p. 93.
3. Leith, E. N. and Upatnieks, J., Wavefront Reconstruction with Diffused Illumination and Three-Dimensional Objects, J. Opt. Soc. Am., *54*, November, 1964, p. 1295-1301.

Errata

In the October 1965 B.S.T.J., Fig. 21 on page 1600, the expression

$$P_R \simeq 8R \text{ for small } p$$

should read

$$P_R \simeq 8Rp \text{ for small } p.$$

