THE BELL SYSTEM
TECHNICAL JOURNAL

Volume 48 March 1969 Number 3

Frequency Sampling Filters— Hilbert Transformers and Resonators

By R. E. BOGNER

(Manuscript received November 6, 1968)

We first briefly review the principles of frequency sampling filters. We also show that the "conventional" frequency sampling filter can be modified simply to give an output which is the Hilbert transform of the original output. Both the original and transformed outputs are made available by the use of the simple complex number resonator described. The relationship between this system and filtering by Fourier transforming is shown.

I. INTRODUCTION

Frequency sampling filters are filters whose frequency responses are synthesized as the sum of elemental frequency responses of the form (Fig. 1a)[1]

$$V_k(f) = A_k \frac{\sin [\pi(f - f_k)/f_o]}{\pi(f - f_k)/f_o} e^{-i2\pi f\tau} + A_k \frac{\sin \pi(f + f_k)/f_o}{\pi(f + f_k)/f_o} e^{-i2\pi f\tau} \quad (1)$$

where

$V_k(f)$ is the transfer function of the kth response;

A_k is a constant multiplier, the value of the amplitude response at frequency f_k;

f is frequency in hertz;

f_k is the kth sampling frequency $= kf_o$;

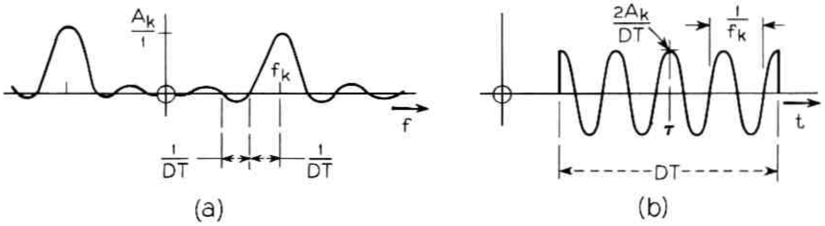


Fig. 1 — (a) Elemental frequency response contribution; (b) Elemental time response contribution.

f_o is the frequency interval between samples, that is, $f_o = f_{k+1} - f_k$, $f_o = 1/DT$, D = delay in samples;

τ is the group delay, a constant for all the responses.

Because of the constant group delay, the amplitude versus frequency response, $|V(f)|$, of the sum is given by

$$| V(f) | = \sum_k A_k \left[\frac{\sin \pi(f - f_k)/f_o}{\pi(f - f_k)/f_o} + \frac{\sin \pi(f + f_k)/f_o}{\pi(f + f_k)/f_o} \right] \cdots . \qquad (2)$$

By choice of the A_k, suitable amplitude responses for many applications may be specified. These will be bandlimited functions of frequency.

The elemental time responses, $v_k(t)$ (Fig. 1b) are convenient to realize by digital methods. They are truncated cosine waves.

Figure 2 shows a comb filter, whose impulses occur DT seconds apart, followed by a resonator, whose impulse response is a cosine wave of frequency an integral multiple of $1/DT$. The overall impulse response is the sum of the cosine responses to the two impulses; this is zero before the positive impulse, a cosine from then until DT seconds later, and thereafter zero, when the two cosines cancel.

A complete frequency sampling filter is shown in the left of Fig. 3. Usually the resonators have been programmed as conventional second

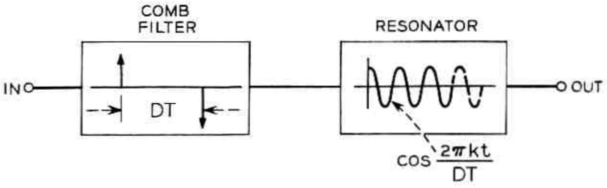


Fig. 2 — Comb filter followed by cosine resonator.

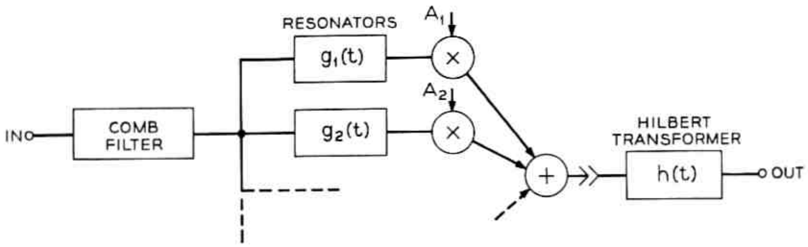


Fig. 3 — Frequency sampling filter, followed by Hilbert transformer.

order systems, with slight damping to ensure stability under conditions of error in the resonator coefficients.

II. USE AS HILBERT TRANSFORMER

A frequency sampling filter may be readily adapted to give an output which is the Hilbert transform of that of the filter described above. Consider the sampling filter (Fig. 3) followed by a Hilbert transformer, $h(t)$. This is equivalent to the system of Fig. 4, where the one Hilbert transformer has been replaced by one at the output of each elemental filter. Now, in the original frequency sampling filter, the kth resonator has an impulse response, for time sampled systems

$$g_k(nT) = \cos \omega_k(nT),\ n = 0,\ 1,\ 2,\ \cdots$$

where T is the sampling interval. The Hilbert transformed version of this is approximately

$$\hat{g}_k(nT) = \sin \omega_k(nT).$$

The approximation is discussed in Appendix A. Thus to make a system equivalent to the original frequency sampling filter plus Hilbert transformer, we need only replace the resonators by ones with impulse responses $\sin \omega_k t$. This could be done by use of modified second order delay resonators; but the system of Fig. 5 is more convenient programwise and is helpful conceptually. This system has the z transform system function

$$\frac{W(z)}{U(z)} = G(z) = \frac{1}{1 - z^{-1} \exp [(\alpha + j\omega)T]} \tag{3}$$

and corresponding impulse response

$$g(nT) = e^{\alpha nT} e^{j\omega nT},\ n = 0,\ 1,\ \cdots . \tag{4}$$

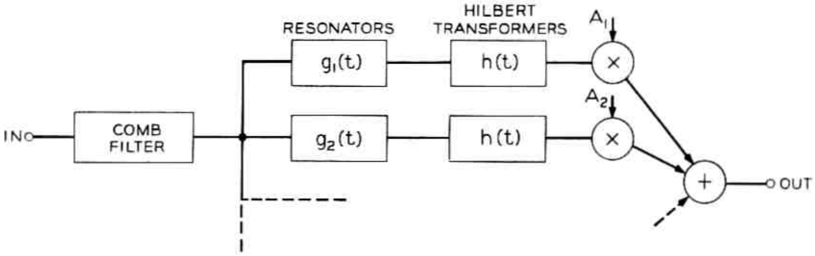


Fig. 4 — Frequency sampling filter with separate Hilbert transformers.

For $\alpha = 0$, the real and imaginary parts are cos ωnT and sin ωnT. A small negative value of α would be used for stability.

The frequency sampling filter then has the form of Fig. 4, with each channel containing one complex number resonator instead of the resonator plus Hilbert transformer. The output at each sampling time is a complex number, whose real part corresponds to the output of a conventional frequency sampling filter, and whose imaginary part is an approximation to the Hilbert transform of the real part.

In Appendix A, the analysis of the approximation results in the following observations:

(*i*) The Hilbert transformer cannot handle signals with frequencies tending to zero.

(*ii*) For signals with low-frequency components, care is necessary in specifying the frequency samples to ensure that the negative-frequency tail of the positive-frequency response component is of small amplitude.

(*iii*) The errors are in the amplitude and not phase characteristics.

The system is capable of filtering a complex input, $u + jv$ without modification of the resonators.

III. RELATION TO DISCRETE FOURIER TRANSFORM

Consider $\alpha = 0$. The response of the kth resonator at time nT, $n = 0, 1, 2, \cdots$, to a unit pulse at time mT is exp $[j\omega_k(n - m)T]$. Hence the response at time nT to a signal $s(mT)$, $m = \cdots, -1,0, 1, 2, \cdots$ is:

$$x_k(nT) + jy_k(nT) = \sum_{m=-\infty}^{n} s(mT) \exp [j\omega_k(n - m)T]$$

$$= \exp (j\omega_k nT) \sum_{m=-\infty}^{n} s(mT) \exp (-j\omega_k mT). \qquad (5)$$

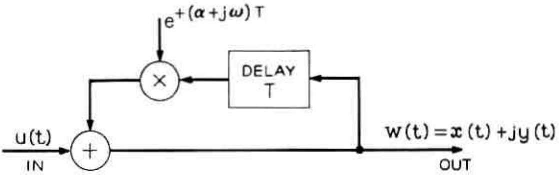


Fig. 5 — Complex number resonator.

When the comb filter precedes the resonator, the effect of its negative impulse, occurring DT seconds after the positive impulse is to add the second term of (6):

$$\begin{aligned} x_k(nT) + jy_k(nT) &= \exp(j\omega_k nT) \sum_{m=-\infty}^{n} s(mT) \exp(-j\omega_k mT) \\ &\quad - \exp(j\omega_k nT) \sum_{m=-\infty}^{n} s(m - D)T \exp(-j\omega_k mT) \\ &= \exp(j\omega_k nT)\Bigg[\sum_{m=-\infty}^{n} s(mT) \exp(-j\omega_k mT) \\ &\quad - \sum_{m=-\infty}^{n-D} s(mT) \exp(-j\omega_k mT) \exp(-j\omega_k DT) \Bigg] \cdot \qquad (6) \end{aligned}$$

But DT is an integral multiple of the period $2\pi/\omega_k$ as mentioned in Section I; thus $\exp(-j\omega_k DT) = 1$. Hence

$$x_k(nT) + jy_k(nT) = \exp(j\omega_k nT) \sum_{m=n-D+1}^{n} s(mT) \exp(-j\omega_k mT). \qquad (7)$$

This expression may be recognized as an oscillation $\exp(j\omega_k nT)$ whose coefficient is the value at frequency ω_k of the Discrete Fourier Transform (DFT) of $s(mT)$, computed over the last D samples. The output of the frequency sampling filter, taking into account the weights A_k, is

$$\begin{aligned} x(nT) + jy(nT) &= \sum_k A_k[x_k(nT) + jy_k(nT)] \\ &\sum_k \exp(j\omega_k nT) A_k \sum_{m=n-D+1}^{n} s(mT) \exp(-j\omega_k mT). \qquad (8) \end{aligned}$$

This is the Fourier synthesis (inverse DFT) of the frequency function

$$A_k \sum_{m=n-D+1}^{n} s(mT) \exp(-j\omega_k mT), \qquad k = 1, 2, \cdots, \qquad (9)$$

which may be regarded as the product of the running DFT of $s(mT)$ and a DFT whose values at frequencies ω_k are the A_k.

Frequency sampling filtering is thus equivalent to filtering by Fourier transforming, multiplying by a filter frequency function, and inverse transforming.

The filter frequency function (A_k, $k = 1, 2, \ldots$) has, so far, been considered real. There is no reason why the A_k should not be complex, permitting the filter to have an arbitrary phase characteristic. The complex values of the A_k may be specified in cartesian or polar form, the latter being more convenient for amplitude-phase specification.

Another way of looking at the resonator output is obtained by rearranging (7):

$$x_k(nT) + jy_k(nT) = \sum_{m=-(D-1)}^{0} s[(m+n)T] \exp(-j\omega_k mT). \qquad (10)$$

This may be recognized as the DFT of the last D values of $s(mT)$, shifted in time so that the latest occurs at time $mT = 0$.

IV. CONCLUSION

The use of complex number resonators in a frequency sampling filter provides a Hilbert transformed output as well as the conventional filtered output. The system can readily accept a complex time function as input, and has a very simple flow chart. The output is equivalent to that obtained by the use of Fourier transforms to perform filtering in the frequency domain.

A sampling filter subroutine using the ideas presented has been written in Fortran IV. It has been used for filtering and Hilbert transforming speech signals in a number of tasks.

V. ACKNOWLEDGMENT

Thanks are due to C. H. Coker, L. R. Rabiner and R. W. Schafer for many helpful discussions. A recursive generation of the DFT is given in Ref. 2.

APPENDIX A

Errors in the Hilbert Transformer

A cosine wave, truncated in time, is the basis of the frequency sampling filters. A correspondingly truncated sine wave has been used as an approximation to the Hilbert transform of the cosine. The errors in this approximation will be analyzed by comparing the

Fourier transform of the truncated sine wave with that of the true Hilbert transform of the cosine. The analysis is for continuous (that is, nonsampled) sines and cosines.

The truncated cosine response is taken to be

$$h_c(t) = \cos\frac{2\pi Nt}{T}, \qquad -\frac{T}{2} \leqq t \leqq \frac{T}{2}$$
$$= 0, \qquad \text{elsewhere.}$$

The F transform of $h_c(t)$ is

$$H_c(f) = \frac{T}{2}\left[\frac{\sin \pi T\left(f - \frac{N}{T}\right)}{\pi T\left(f - \frac{N}{T}\right)} + \frac{\sin \pi T\left(F + \frac{N}{T}\right)}{\pi T\left(f + \frac{N}{T}\right)}\right] \tag{11}$$

$$= H_{c1}(f) + H_{c2}(f), \quad \text{respectively.} \tag{12}$$

$H_c(f)$ may be separated further into main responses and "tails" (Fig. 6):

$$H_c(f) = H_{c1+}(f) + H_{c1-}(f) + H_{c2+}(f) + H_{c2}(f) \tag{13}$$

where

$$H_{c1+} = H_{c1}, \quad f > 0; \quad \frac{H_{c1}(0)}{2}, \quad f = 0; \quad 0, \quad f < 0$$

$$H_{c1-} = 0, \quad f > 0; \quad \frac{H_{c1}(0)}{2}, \quad f = 0; \quad H_{c1}, \quad f < 0$$

$$H_{c2+} = H_{c2}, \quad f > 0; \quad \frac{H_{c2}(0)}{2}, \quad f = 0; \quad 0, \quad f < 0$$

$$H_{c2-} = 0, \quad f > 0; \quad \frac{H_{c2}(0)}{2}, \quad f = 0; \quad H_{c2}, \quad f < 0.$$

The F transform of the Hilbert transform $[\hat{h}_c(t)]$ of $h_c(t)$ is then

$$\hat{H}_c(f) = -j \operatorname{sgn}(f) H_c(f) \tag{14}$$

$$= -jH_{c1+}(f) + jH_{c1-}(f) - jH_{c2+}(f) + jH_{c2-}(f). \tag{15}$$

The truncated sine response is taken to be

$$h_s(t) = \sin\frac{2\pi Nt}{T}, \qquad -\frac{T}{2} \leqq t \leqq \frac{T}{2}$$
$$= 0, \qquad \text{elsewhere.}$$

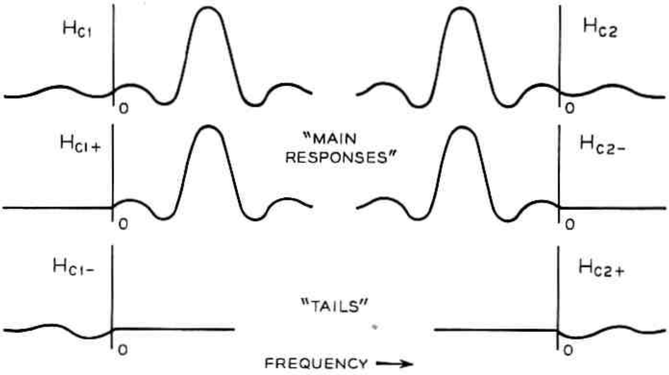


Fig. 6 — Components of elemental frequency response.

The F transform of $h_s(t)$ is

$$H_s(f) = \frac{T}{2}\left[-j\frac{\sin \pi T\left(f - \frac{N}{T}\right)}{\pi T\left(f - \frac{N}{T}\right)} + j\frac{\sin \pi T\left(f + \frac{N}{T}\right)}{\pi T\left(f + \frac{N}{T}\right)}\right] \tag{16}$$

which by comparison with (11), (12), (13) is seen to be

$$\begin{aligned} H_s(f) &= -jH_{c1}(f) + jH_{c2}(f) \\ &= -jH_{c1+}(f) - jH_{c1-}(f) + jH_{c2+}(f) + jH_{c2-}(f). \end{aligned} \tag{17}$$

Then from (15) and (17):

$$H_s(f) = \hat{H}_c(f) - 2jH_{c1-}(f) + 2jH_{c2+}(f). \tag{18}$$

The error in approximating $\hat{H}_c(f)$ by $H_s(f)$ is thus attributable to the tails $H_{c1-}(f)$ and $H_{c2+}(f)$, which are small for $N \gg 1$. From the definitions (11), (12), (13), it follows that these tails are related:

$$H_{c1-}(-f) = H_{c2+}(f). \tag{19}$$

In a complete frequency sampling filter, the transforms corresponding to all the time responses are to be added. Errors in the "Hilbert transformed" output, y, as compared with the straight filtered output, x, are determined by the resultant tails; these tails may be of small amplitude if suitable values are chosen for the frequency samples.

Just what criterion of smallness should be applied depends on the application. Some general observations may be made, however:

(*i*) The Hilbert transformer cannot be useful to zero frequency because a zero frequency sample has tails equal to the main responses, and would thus contribute gross errors. This is of course consistent with the infinite duration of the impulse response $(1/t)$ of a true Hilbert transformer.

(*ii*) To transform signals with low frequency components, many frequency samples may be required to provide the sharp and continued cutoff required for tail suppression.

(*iii*) Since $H_{c1-}(-f) = H_{c2+}(f)$, it follows from (18) that the errors, associated with $H_{c1-}(-f)$ and $H_{c2+}(f)$ are directly in or out of phase with the relevant main responses. The error in the Hilbert transform is thus an amplitude and not a phase error. This result is also consistent with the observation that the approximate Hilbert transformed response to an impulse is truly odd.

APPENDIX B

Relationship between Complex Number Resonator and Conventional Second Order Resonator

While the formal transform relation between (3) and (4) is readily shown, it is satisfying to explain how the seemingly first order delay system can produce an oscillatory response. The system of Fig. 5 is described by the equation

$$x(mT) + jy(mT) = u(mT) + e^{(\alpha + j\omega)T}[x(m-1)T + jy(m-1)T] \quad (20)$$

When a pulse $u(0) = 1$, with zero before and after is applied, the first response is

$$x(0) + jy(0) = 1 + j0$$

The next response is simply the first response multiplied by $e^{(\alpha + j\omega)T}$

$$x(1T) + jy(1T) = e^{(\alpha + j\omega)T}(1 + j0);$$

there is a similar multiplication at each subsequent sampling instant, yielding the impulse response

$$x(nT) + jy(nT) = e^{n(\alpha + j\omega T)},\ n = 0, 1, 2 \cdots, \quad (21)$$

equivalent to (4).

The complex number resonator may be shown to contain a second order delay feedback, making its oscillatory response consistent with that of the more conventional second-order systems. Its equation (20)

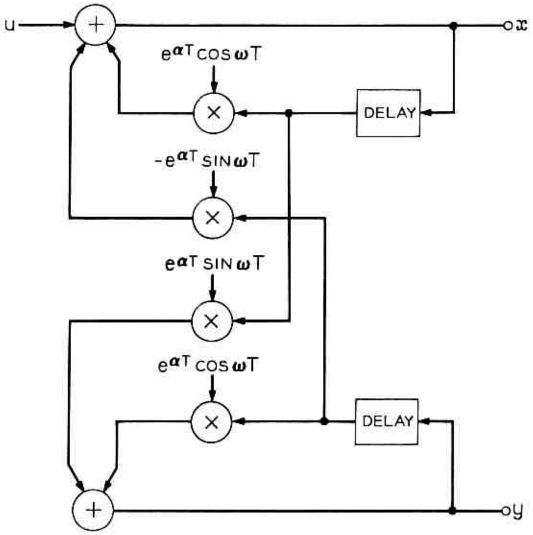

u
x
$e^{\alpha T}\cos\omega T$
DELAY
$-e^{\alpha T}\sin\omega T$
$e^{\alpha T}\sin\omega T$
$e^{\alpha T}\cos\omega T$
DELAY
y

Fig. 7 — Expanded flow chart for complex number resonator.

may be examined by equating separately real and imaginary parts:

$$x(mT) = u(mT) + (e^{\alpha T} \cos \omega T)x[(m - 1)T] - (e^{\alpha T} \sin \omega T)y[(m - 1)T] \qquad (22)$$

$$y(mT) = (e^{\alpha T} \sin \omega T)x[(m - 1)T] + (e^{\alpha T} \cos \omega T)y[(m - 1)T] \qquad (23)$$

Equations (22) and (23) may be represented by the flow chart of Fig. 7. There is, in fact, a path of delay two sampling intervals from the real output x, via y, the imaginary part of the output, back to x. Thus, y could be considered to provide the necessary memory for the second delay.

One aesthetically pleasing feature of the representation (Fig. 7) is the symmetry. If a complex input, $u + jv$ were to be filtered, then v would be found to be applied to the lower summer.

REFERENCES

1. Rader, C. M. and Gold, B., "Digital Filter Design Techniques in the Frequency Domain," Proc. IEEE, *55,* No. 2 (February 1967), pp. 149–171.
2. Halberstein, J. H., "Recursive, Complex Fourier Analysis for Real-Time Applications," Proc. IEEE Letters, *54,* No. 6 (June 1966), p. 903.

Scattering from Dielectric Mirrors

By D. GLOGE, E. L. CHINNOCK, and H. E. EARL

(Manuscript received September 9, 1968)

Most of the light scattered from high-reflectivity dielectric mirrors is radiated into directions close to the reflected beam. We measured the angular power distribution at angles between 0.01° and 1° from the beam axis by scanning with a narrow slit. From this a linear structure function is calculated for coherence lengths between 20 microns and 1 millimeter, assuming isotropic surface statistics. The corresponding power density decreases with the third power of the scattering angle. The power outside a given radius and the power density is plotted for various wavelengths and distances.

I. INTRODUCTION

The improvement of dielectric mirrors during recent years has reduced their surface scattering considerably. Nevertheless, there are applications which are limited by these small amounts of scattered light. One of them is the laser gyroscope whose locking threshold depends on the light scattered back into the direction of incidence. Measurements have been performed recently to analyze this case.[1]

Another application is the simultaneous transmission of many laser beams in an optical waveguide for communication purposes.[2] The focusers in such a guide will probably be front surface mirrors rather than lenses because, for the large apertures needed, lenses are apt to have imperfections in the bulk. Dielectric mirrors have fewer imperfections, but they still scatter some light into adjacent beams where it produces crosstalk. It was the purpose of our experiment to measure some representative mirror surfaces as a basis for later feasibility studies on multiple beam waveguides. Only the light in a narrow cone around the beam is collected by the next focuser and eventually contributes to the crosstalk. The experiment showed that, in this cone, the scattered light intensity decreases relatively fast with increasing angle.

Applying these results to more complicated problems requires a simple but adequate mathematical representation of the results. We found that the standard scattering theory, which uses a covariance function to describe the mirror surface statistics, serves this purpose very poorly.[3] On the other hand, a simple structure function can be found which is a satisfactory representation of the physical reality in the range of the measurements and is easily applicable to practical problems.

II. SCANNING THE SCATTERED POWER DISTRIBUTION

The measurements were performed with a 50-cm He-Ne laser generating a 1-milliwatt gaussian beam at 6328Å. To achieve enough sensitivity and discrimination against noise, the laser beam was chopped for signal processing in a lock-in amplifier as shown in Fig. 1.

A slit was used to scan the scattered light. This requires scanning only in one direction (while the slit averages over the perpendicular coordinate) and more signal power is collected than with a pinhole method. Because of its circular symmetry, the scattered power density can be calculated from this measurement by a simple integral transformation.

To avoid scattering from dust particles in the beam path to and from the mirror, this path was evacuated to about 4 torr. But careful comparison with measurements in unfiltered, though quiet, air showed no measureable difference.

The mirror had a radius of curvature of 24 m and a diameter of 15 cm. The beam, having a $1/e$-width of 24 mm at the mirror, was focused to 0.8 mm in the plane of the slit. The slit was 0.15 mm wide. Figure 2 shows the relative intensities normalized to the peak

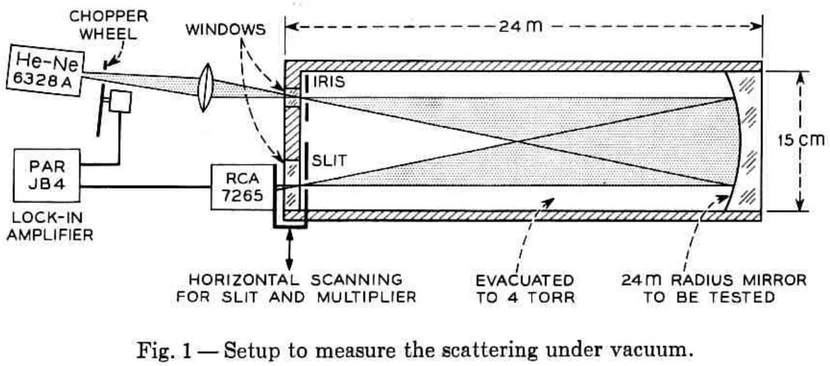


Fig. 1 — Setup to measure the scattering under vacuum.

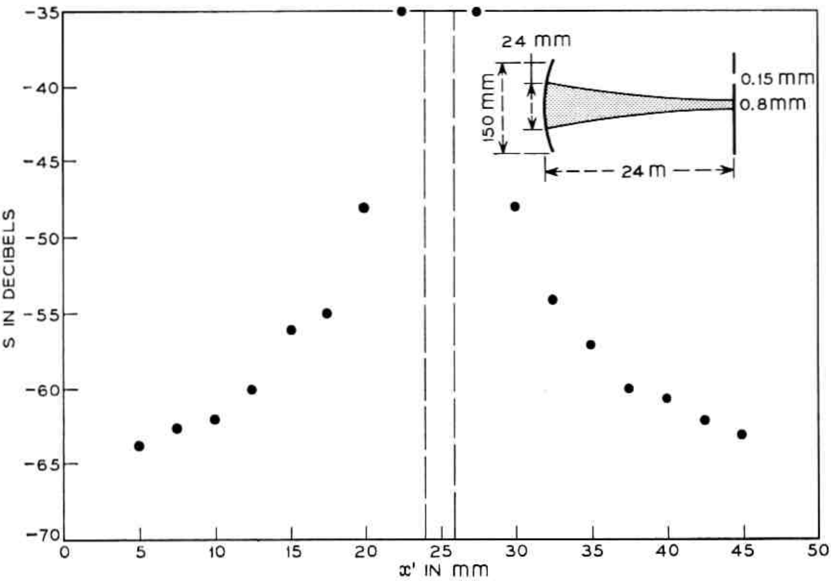


Fig. 2 — Scattered power at 24-m distance normalized to the peak power (the vertical lines are part of the coherent beam profile).

intensity in dB and plotted versus the vertical coordinate. In this logarithmic plot, the gaussian intensity profile of the coherent signal has a parabolic shape, part of which is represented by the almost vertical lines in the center of Fig. 2. If diffraction and spherical aberrations are taken into consideration, the fall-off is not quite as sharp as indicated by the parabola, but these effects were estimated to be well below the light levels measured. Therefore, we believe that surface scattering is the sole source for our results. The scanning range in Fig. 2 corresponds to angles from 0.01 to 0.1 degree.

For larger angles up to 1 degree, a 1-m set-up was used which was basically similar to the one shown in Fig. 1, but had no vacuum enclosure. The five mirrors tested in this arrangement had a radius of curvature of 1 m, a diameter of 25 mm, and the same coating as the 24-m mirror. The test beam in this set-up was 4 mm wide at the mirror and was focused to 0.2 mm at the slit. The slit had a width of 0.05 mm. Coatings from different batches showed up to 3 dB difference. Figure 3 shows average and variation of the results. Again the intensity normalized to the peak intensity is plotted in dB. The profile of the coherent beam is shown in the center.

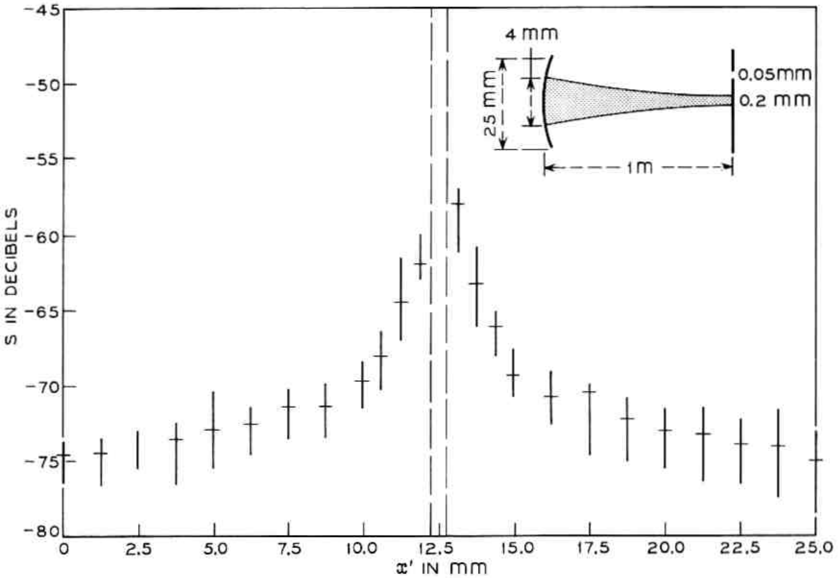


Fig. 3 — Scattered power at 1-m distance normalized to the peak power (the vertical lines are part of the coherent beam profile).

Both the 24-m mirror and the 1-m mirror were polished and coated by the same methods though by different manufacturers. They were tested to be spherical within $\lambda/10$. The reflection loss of the 24-m mirror was measured by a multiple reflection technique to be 0.135%. The mirrors were measured new without previous use, but no increase of the scattering was measured by repeated checks during the following weeks. Further lifetime studies are under way.

III. DESCRIPTION OF THE SCATTERING SURFACE

The scattering plotted in Figs. 2 and 3 originates from a slight roughness or ripple structure $\delta(X, Y)$ on the mirror surfaces which is, of course, different for different mirrors. The surfaces tested in this experiment, however, were manufactured by the same process and are therefore equivalent in a statistical sense. That implies that the average magnitude of each ripple component, that is, the "power spectrum" of $\delta(X, Y)$, is the same from mirror to mirror. The average has to be taken over an ensemble of test surfaces; however, for correlation lengths small compared to the test area, the ensemble average may be replaced by an average over the individual surface. In this case, measuring only one or a few surfaces still yields a meaningful result, though only for correlation lengths small compared to the radius w of the light beam at the mirror surface.

The "power spectrum" is closely related to the scattering profile measured by the slit method. A vertical slit at X', as in Fig. 4, collects mainly light scattered from the vertical ripple component with the spatial frequency

$$x = \frac{X'}{L\lambda} \tag{1}$$

where λ is the light wavelength and L the distance between slit plane and mirror. Therefore, apart from a constant, the scattered profile $s(X')$ of Figs. 2 and 3 agrees with the "power spectrum" $d_x(x)$ of $\delta(X, Y)$ for $Y =$ constant.[*] The quantitative relation between d_x and s is given in (38) of Appendix A and reads

$$d_x(x) = \frac{L\lambda^3}{16\pi^2 t} \operatorname{erf}\left(\frac{\pi w t}{\sqrt{2}\,\lambda L}\right) s(x). \tag{2}$$

where t is the slit width and w the $1/e$-width of the gaussian light beam at the mirror surface. A log-log plot of $d_x(x)$ is shown in Fig. 5. The points on the left hand side are taken from Fig. 2 and represent the 24-m experiment, the ones on the right hand side stem from Fig. 3 and the 1-m experiment. Since the mirrors are statistically equivalent, all these points belong to the same function. A rough approximation is attempted by the straight line in Fig. 5 which represents the function

$$d_x(x) = \frac{D}{x^2} \tag{3}$$

with

$$D = 6 \cdot 10^{-14} \text{ mm}. \tag{4}$$

The Fourier transform of d_x is the covariance of $\delta(X, Y)$ along lines $Y =$ constant.[3] It proves impossible, however, to perform this transform without knowing d_x for very small x where it increases rapidly. Accurate information about this range is unnecessary if the

[*] Strictly speaking, $s(x)$ is a two-fold convolution of $d_x(x)$ with the intensity profile of the beam and the slit aperture function; but because the latter two functions are very narrow as compared to the scattered profile, the above simplification is appropriate.

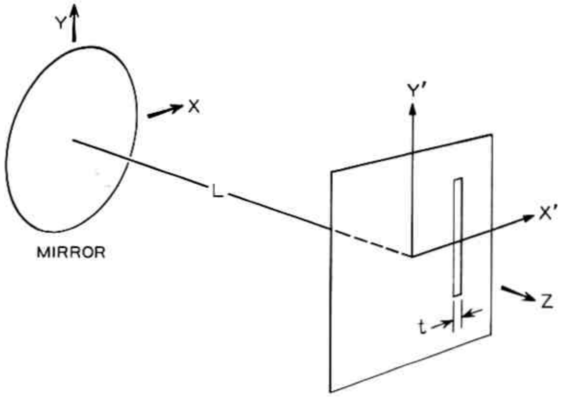


Fig. 4 — Sketch of the experiment showing the coordinate system used.

structure function

$$\Delta(X_1 - X_2, Y_1 - Y_2) = \langle[\delta(X_1, Y_1) - \delta(X_2, Y_2)]^2\rangle_{av} \tag{5}$$

is used instead. The interrelation between Δ and d_x is derived in the Appendix A and given in (39). It involves the transformation

$$\int_0^\infty d_x(x) \sin^2(\pi X x)\, dx.$$

The $\sin^2$-kernel of this transformation reduces the contribution from the zero-end of the d_x-function and $\Delta(X, 0)$ can therefore be calculated more accurately in the range of interest than the covariance.

Inserting (3) into (39) yields the functional approximation

$$\Delta(X, 0) = -\frac{\lambda^2}{8\pi^2}\left[\frac{1}{2}\frac{X^2}{w^2} + \ln\left(1 - \frac{32\pi^4}{\lambda^2} DX\right)\right], \tag{6}$$

where X has the meaning of a correlation length. In the range $X < w$, which is shown in Fig. 6, the structure function (6) is essentially a straight line given by

$$\Delta(X, 0) = 4\pi^2 DX. \tag{7}$$

This result suggests that the mean square difference between samples of δ increases proportionally to the distance at which they are taken. At the right side of Fig. 6 the quantity $(\Delta)^{1/2}$ can be read off, which indicates a direct measure of the heights of the surface irregularities as a function of their extension about the surface. This

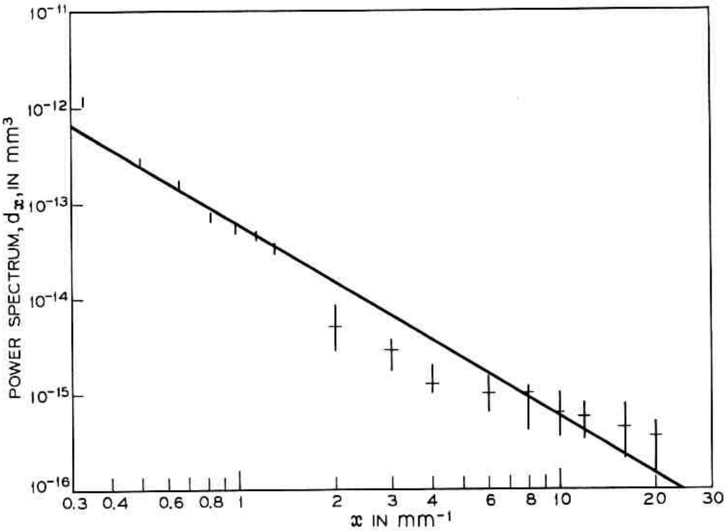


Fig. 5 — Power spectrum of the mirror surface roughness. The line represents a best-fit approximation to the measured points.

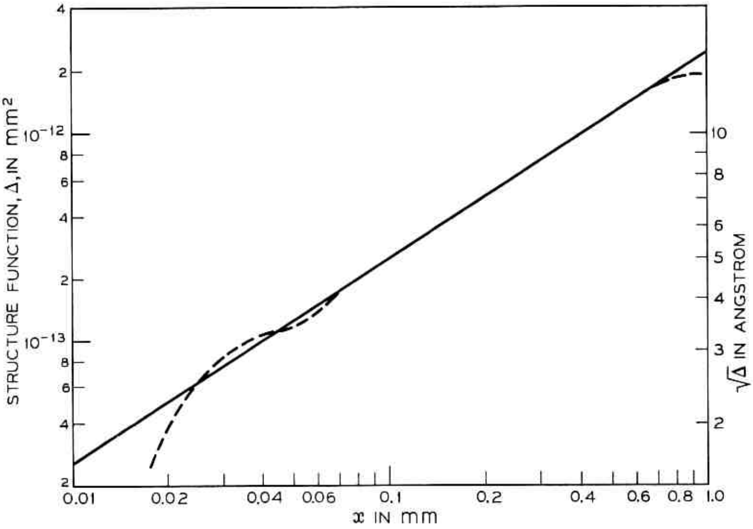


Fig. 6 — The structure function Δ calculated from the approximated power spectrum. The broken line indicates the possible uncertainty of the result.

quantity does not, at least not in the measured interval, approach a definite rms value, but increases unlimited for larger and large sampling distances. This is not necessarily in contrast to physical reality, but one has to keep in mind that, for macroscopic sampling distances, the statistics of δ are probably governed by a different process, which could mean a steeper rise as well as a leveling off for the structure function.

Unknown contributions from outside the measured interval of d_x will to some degree affect the accuracy with which (7) can be evaluated. Ruling out any poles of d_x for $x \neq 0$ (which would mean nonstatistical components), a "worst case" may be established by assuming that (3) holds only in the measured interval $x_1 < x < x_2$, everywhere else $d_x(x) = 0$. Then from (39) with $x \ll w$ and $\Delta \ll 1$

$$\Delta(X, 0) = 8D \int_{x_1}^{x_2} \frac{\sin^2(\pi X x)}{x^2}\, dx \tag{8}$$

with $x_1 = 0.3\text{mm}^{-1}$ and $x_2 = 30\text{mm}^{-1}$. The dashed line in Fig. 6 shows the evaluation of this integral. The accuracy seems satisfactory for coherence lengths between 20μ and 1 mm.

Though $\Delta(X, 0)$ describes the statistics of δ only along lines $Y =$ constant, this result can easily be generalized assuming that the mirror surface is isotropic. Then the structure function has circular symmetry and can be expressed as a function of the radius $R = (X + Y^2)^{1/2}$. This function reads

$$\Delta(R) = \Delta(R, 0) \tag{9}$$

where $\Delta(R, 0)$ is given by (7).

IV. THE DISTRIBUTION OF THE SCATTERED POWER

For most applications the actual scattered light distribution around the beam is of more immediate interest than the structure function. Of course, this light distribution not only depends on the properties of the mirror, but also on the properties of the light beam reflected off the mirror. More specifically, this light distribution is the convolution of the intensity profile of the primary beam with the "power spectrum" of the mirror irregularities. Only when the beam profile is very narrow, as in our experiment, do the scattered light distribution and the "power spectrum" become proportional functions.

In this section we evaluate this distribution for various optical wavelengths in arbitrary cross sections of the beam. If applied to problems where the width of the beam may not be neglected, the convolution of this function with the intensity profile of the beam has to be formed.

Because of the isotropy of the mirror surface, the scattering has circular symmetry. If we define a normalized radius $r = (x^2 + y^2)^{1/2}$, the scattered light distribution $p(r)$ can be calculated from (32) of Appendix A. By substituting x by r in (32), one obtains

$$S(x) = \frac{2t}{L\lambda} \int_x^\infty \frac{p(r) r \, dr}{(r^2 - x^2)^{\frac{1}{2}}}, \tag{10}$$

where r is related to the radius $R' = (X'^2 + Y'^2)^{1/2}$ by the normalization

$$r = \frac{R'}{L\lambda}. \tag{11}$$

One can solve (10) for $p(r)$ by multiplying both sides by $x dx/(x^2 - r^2)^{1/2}$ and integrating with respect to x from r to ∞.[4] After interchanging the order of integration on the right-hand side, the integral over x can be evaluated and one obtains

$$\frac{L\lambda}{2t} \int_r^\infty \frac{x S(x) \, dx}{(x^2 - r^2)^{\frac{1}{2}}} = \int_r^\infty 2\pi r p(r) \, dr. \tag{12}$$

The integral on the right represents the total power scattered outside a circle with radius r and will be called $P(r)$ in the following. Insertion of (2), (27), and (35) into (12) yields

$$P(r) = P_{\text{tot}} \frac{8\pi^2}{\lambda^2} \int_r^\infty \frac{x d_x(x) \, dx}{(x^2 - r^2)^{\frac{1}{2}}}; \tag{13}$$

and the power density $p(r)$ is finally obtained from the differentiation

$$p(r) = -\frac{1}{2\pi r} \frac{dP}{dr}. \tag{14}$$

By using (3) for d_x in (13), one obtains for the power outside the radius r

$$P(r) = \frac{4\pi^3}{\lambda^2} \frac{D}{r} P_{\text{tot}} \tag{15}$$

and the power density

$$p(r) = \frac{2\pi^2}{\lambda^2} \frac{D}{r^3} P_{\text{tot}}. \tag{16}$$

To gain information about the power density as a function of the scattering direction, one may multiply (11) by λ and find the scattering angle

$$\rho = \lambda r = \frac{R'}{L}. \tag{17}$$

The power scattered into directions deviating by more than ρ from the beam axis is obtained by inserting (17) into (15)

$$P\left(\frac{\rho}{\lambda}\right) = \frac{4\pi^3}{\lambda} \frac{D}{\rho} P_{\text{tot}} . \tag{18}$$

The derivative with respect to ρ yields the angular power density

$$p_\rho(\rho) = -\frac{1}{2\pi\rho} \frac{dP}{d\rho} = \frac{2\pi^2}{\lambda} \frac{D}{\rho^3} P_{\text{tot}} . \tag{19}$$

Equations (18) and (19) are evaluated for various wavelengths in Figs. 7 and 8. Figure 7 shows the power fraction outside ρ which decreases linearly with increasing radius. Fig. 8 shows the power fraction radiated into a given solid angle at ρ. This function decreases with the third power of ρ.

Finally, for a certain distance L, one finds the power arriving out-

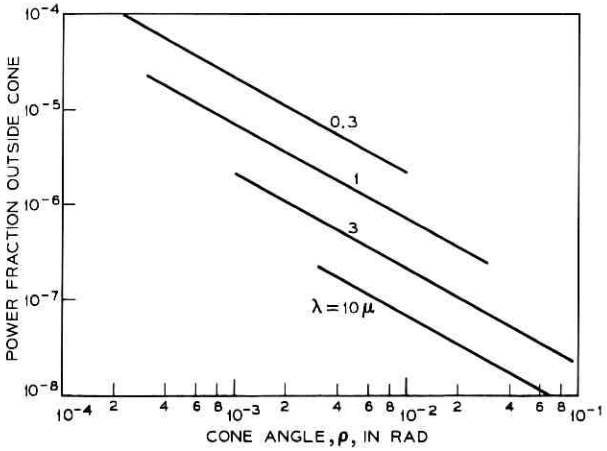


Fig. 7 — The total power fraction scattered with an angle larger than ρ off the beam axis for various wavelengths λ.

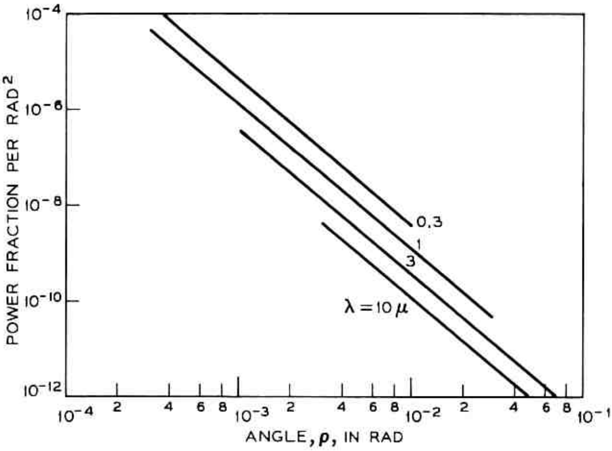


Fig. 8 — The angular power density as a function of the scattering angle ρ for various wavelengths λ.

side a circle of radius R' to be

$$P\left(\frac{R'}{L\lambda}\right) = 4\pi^3 \frac{DL}{\lambda R'} P_{\text{tot}} \tag{20}$$

and the derivative with respect to R' yields

$$p_{R'}(R') = -\frac{1}{2\pi R'}\frac{dP}{dR'} = 2\pi^2 \frac{DL}{\lambda R'^3} P_{\text{tot}}, \tag{21}$$

where $p_{R'}$ is the scattered power density at a distance L. Equation (20) is plotted for various L in Fig. 9. Figure 10 shows the power density versus the radius which decreases with the third power of R'. It is interesting that the power density at a fixed radius increases proportional to the distance from the scatterer.

Of course, equations (15) through (21) hold only for $r < 0.3$ mm^{-1}, the lower limit of the interval measured, and are based on the assumption that (3) is valid for $r > 30$ mm^{-1}. However, as d_r is small in the latter region, a possible error introduced by this assumption should not be significant in the range of interest.

V. CONCLUSIONS

The small angle scattering was measured for very highly reflecting dielectric mirrors. A reasonable functional approximation for the

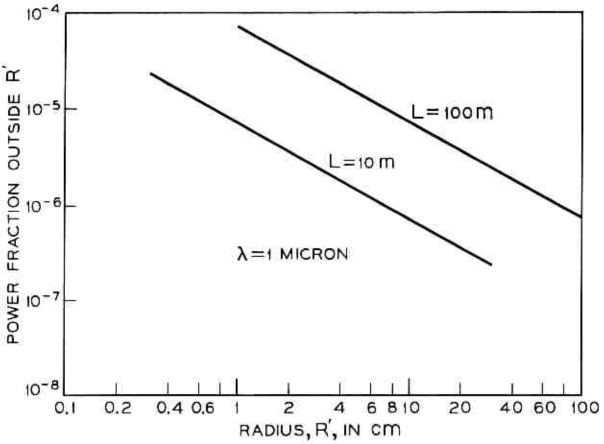


Fig. 9 — The total power fraction scattered outside a circle with radius R' at various distances L.

measurements leads to a linear structure function for coherence lengths between 20 microns and 1 mm. The rms difference between surface deviations found at two points 1 mm apart is 30 Angstroms and decreases with the root of the distance for points closer together. At an angle of 0.1° the scattered power density per cm² is 10^{-6} of the total power. It decreases proportional to the third power of the angle

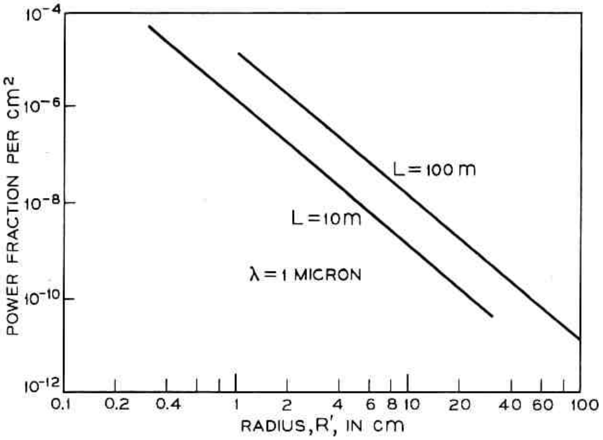


Fig. 10 — The scattered power density as a function of the radius R' at various distances L.

and linearly with the light wavelength. No considerable differences were found for mirrors polished and coated by two different manufacturers which were using the same processes and chemicals.

VI. ACKNOWLEDGMENTS

We are thankful for stimulating discussions and fruitful suggestions by H. E. Rowe and E. A. J. Marcatili.

APPENDIX

The Structure Function

Consider a coherent Gaussian beam of wavelength λ to be reflected off the mirror in Fig. 4 and focused on a plane with a slit. Assume that unperturbed phase fronts emanating from the mirror were spherical with a field distribution

$$E(X, Y) = E_0 \exp\left[-(X^2 + Y^2)/w^2\right]. \tag{22}$$

The mirror diameter may be considered sufficiently large compared to the $1/e$-width w so that the field at the mirror edge may be neglected. In this case, the field in the focal plane is

$$f(x, y) = \iint_{-\infty}^{+\infty} e^{-i2\pi(Xx+Yy)} E(X, Y)\, dX\, dY \tag{23}$$

where

$$x = \frac{X'}{L\lambda} \quad \text{and} \quad y = \frac{Y'}{L\lambda} \tag{24}$$

are the normalized coordinates in the focal plane (see Fig. 4). The solution of (23) is

$$f(x, y) = E_0 w^2 \pi \exp\left[-\pi^2 w^2 (x^2 + y^2)\right]. \tag{25}$$

The total power

$$P_{\text{tot}} = \frac{\pi w^2}{2} E_0^2 \tag{26}$$

can be calculated by integrating (22) or (25).

Assume that the slit in the focal plane is long enough to collect all the power in y-direction and has a width t in x-direction. Then for $X' = 0$ the signal received is

$$S(0) = P_{\text{tot}} \operatorname{erf}\left(\frac{\pi w t}{\sqrt{2}\, L\lambda}\right) \tag{27}$$

where erf denotes the error function. This is the peak signal measured with no scattering present. However, for the case in question where the scattered power is less than a percent, (27) can also be used for the peak center signal of the scattering profile.

The scattering under consideration originates from a slight roughness $\delta(X, Y)$ of the mirror surface which gives rise to a phase variation

$$\varphi(X, Y) = \frac{4\pi}{\lambda}\,\delta(X, Y) \tag{28}$$

on the otherwise perfect phase front. The term $\delta(X, Y)$ is assumed to be a gaussian random process with isotropic statistics. Its structure function is given by (5).

It can be shown that for gaussian statistics[5]

$$\langle \exp i[\varphi(X_0\,,\,Y_1) - \varphi(X_2\,,\,Y_2)]\rangle = \exp\left[-\frac{8\pi^2}{\lambda^2}\,\Delta\right]. \tag{29}$$

The power density in the focal plane can be calculated from (23) by introducing the phase factor $\exp[-i\varphi(X, Y)]$ and then multiplying (23) by its conjugate complex. Using (29) yields finally

$$\begin{aligned} p(x, y) = \iiiint_{-\infty}^{+\infty} & E(X_1\,,\,Y_1)E(X_2\,,\,Y_2) \\ & \cdot \exp\left[-\frac{8\pi^2}{\lambda^2}\,\Delta(X_1 - X_2\,,\,Y_1 - Y_2)\right] \\ & \cdot \exp\left[-i2\pi(X_1 - X_2)x\right] \\ & \cdot \exp\left[-i2\pi(Y_1 - Y_2)y\right]\,dX\;\,dX_2\,dY_1\,dY_2\,. \end{aligned} \tag{30}$$

After a standard coordinate transformation, this becomes

$$\begin{aligned} p(x, y) = P_{\mathrm{tot}} \iint_{-\infty}^{+\infty} & \exp\left[-(X^2 + Y^2)/2w^2\right] \\ & \cdot \exp\left[-\frac{8\pi^2}{\lambda^2}\,\Delta(X, Y)\right] \exp\left[-i2\pi(Xx + Yy)\right]\,dX\,dY. \end{aligned} \tag{31}$$

To measure the relatively flat power distribution outside the coherent beam, one may average over the slit width t and consequently the signal measured in this region is approximately

$$S(x) = \frac{t}{L\lambda} \int_{-\infty}^{+\infty} p(x, y)\,dy. \tag{32}$$

Because of (31), this becomes

$$S(x) = \frac{t}{L\lambda} P_{\text{tot}} \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}\frac{X^2}{w_4^2} - \frac{8\pi^2}{\lambda^2}\Delta(X, 0)\right] e^{-i2\pi Xx}\, dX. \tag{33}$$

For experimental convenience the normalized signal

$$s(x) = \frac{S(x)}{S(0)} \tag{34}$$

was measured and plotted in Figs. 2 and 3. From (27) and (33) one finds

$$s(x) = \frac{t}{L\lambda}\left[\operatorname{erf}\left(\frac{\pi w t}{\sqrt{2}\, L\lambda}\right)\right]^{-1} \cdot \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}\frac{X^2}{w^2} - \frac{8\pi^2}{\lambda^2}\Delta(X, 0)\right] e^{-i2\pi Xx}\, dX. \tag{35}$$

Inverting the Fourier transformation in (35) yields

$$\exp\left[-\frac{1}{2}\frac{X^2}{w^2} - \frac{8\pi^2}{\lambda^2}\Delta(X, 0)\right] = \frac{L\lambda}{t}\operatorname{erf}\left(\frac{\pi w t}{\sqrt{2}\, L\lambda}\right)\int_{-\infty}^{+\infty} s(x) e^{i2\pi Xx}\, dx. \tag{36}$$

The evaluation of this integral is problematic for small x where the measurements are impeded by the coherent beam, but where $s(x)$ is large and contributes significantly to the dc component of $\Delta(X, 0)$. To overcome this difficulty, the identity $\Delta(0, 0) = 0$ can be used which is based on the definition (5) of the structure function. Incorporating this identity (36) can be rewritten in the form

$$1 - \exp\left[-\frac{1}{2}\frac{X^2}{w^2} - \frac{8\pi^2}{\lambda^2}\Delta(X, 0)\right] = \frac{L\lambda}{t}\operatorname{erf}\left(\frac{\pi w t}{\sqrt{2}\, \lambda L}\right)\int_{-\infty}^{+\infty} s(x)[1 - e^{i2\pi Xx}]\, dx. \tag{37}$$

The function

$$d_x(x) = \frac{L\lambda^3}{16\pi^2 t}\operatorname{erf}\left(\frac{\pi w t}{\sqrt{2}\, \lambda L}\right) s(x) \tag{38}$$

may be interpreted as the "power spectrum" of $\delta(X, Y)$ for $Y =$ constant. Consequently, x has the meaning of a spatial frequency related to the Fourier components of δ along lines $Y =$ constant. d_x

as well as s are even functions of x. Therefore, the solution for $\Delta(X, 0)$ can be written in the form

$$\Delta(X, 0) = \frac{\lambda^2}{8\pi^2} \left\{ -\frac{1}{2} \frac{X^2}{w^2} - \ln \left[1 - \frac{64\pi^2}{\lambda^2} \int_0^\infty d_r(x) \sin^2 (\pi X x) \, dx \right] \right\}. \tag{39}$$

REFERENCES

1. Beatey, R., "Light Scattering by Laser Mirrors," Appl. Opt., *6,* No. 5 (May 1967), pp. 831–835.
2. Gloge, D., and Weiner, D., "The Capacity of Multiple Beam Transmission Systems and Optical Delay Lines," B.S.T.J., *47,* No. 10 (December 1968), pp. 2095–2109.
3. Beckmann, P. and Spizzichino, A., *The Scattering of Electromagnetic Waves from Rough Surfaces,* New York: Pergamon Press, 1963.
4. Harker, K. J., "Use of Scanning Slits for Obtaining the Current Distribution in Electron Beams," J. Appl. Phys., *28,* No. 11 (November 1957), pp. 1354–1357.
5. O'Neill, E. L., *Introduction to Statistical Optics.* New York: Addison-Wesley Publishing Co., 1963, p. 100.

Apparent Increase in Noise Level When Television Pictures Are Frame-Repeated

By F. W. MOUNTS and D. E. PEARSON

(Manuscript received October 16, 1968)

Subjective measurements were made of the apparent increase in noise level which occurs when television pictures are frame-repeated. We show that in all cases of practical interest this increase is small (less than 3 dB), that it is dependent on the type of scanning (greater increases with line-sequential than with line-interlaced scanning), and that it is relatively independent of the picture signal-to-noise ratio. At smaller numbers of repetitions—the region which shows most promise for practical schemes of bandwidth saving—the increase in apparent noise level with increased frame-repetition is most rapid.

I. INTRODUCTION

Seemingly attractive schemes for compressing the bandwidth of television signals sometimes render the signal highly sensitive to noise. As a result, the signal-to-noise ratio requirement for the channel becomes extremely large.[1] If this requirement is not met, the errors caused by the noise degrade the picture to an intolerable degree. Thus, noise is a great obstacle to the success of bandwidth-saving schemes, and the authors of any such schemes should always take care to check the noise-sensitivity of their compressed signals.

We report in this paper on some measurements we have made of the noise-vulnerability of picture signals in a frame-repeated television system. In a previous paper concerned with the possibilities of bandwidth-saving by frame-repetition or frame-replenishment, Brainard, Mounts, and Prasada made the observation that with increasing numbers of repeated television frames there appears to be an increase in the picture noise level.[2] The noise pattern is "frozen" for the repetition period; this tends to make it more visible to the eye. Any source noise, such as that created by delta modulation or pulse code modulation (PCM), or any channel noise, such as random gaussian noise, has its effective power raised by the process of frame-repetition. We attempted to measure this apparent increase in noise power in quantitative terms to determine what improvements in modulation methods or channel noise levels would be required if television pictures were frame-repeated.

II. EXPERIMENTAL EQUIPMENT

The equipment used to produce the frame-repeated pictures has been fully described in Ref. 3. For the experiments reported in this paper random noise was added to the video signal prior to frame-repetition. An automatically-timed switch was arranged to present, alternately, on a single display monitor, frame-repeated and non-frame-repeated (standard) versions of the same picture. Controlled amounts of noise could be added to the frame-repeated picture by the experimenter and to the standard comparison picture by the subject. This arrangement permitted the subject to carry out a visual match of the levels of noise in the frame-repeated and standard pictures. A block diagram is given in Fig. 4; a more detailed description of the apparatus is given in Appendix A.

III. EXPERIMENTAL METHOD

The method we used to measure the apparent increase in noise level in a frame-repeated picture was to ask 24 subjects to view in succession frame-repeated and standard versions of the same picture. With controlled amounts of noise added to the frame-repeated picture by the experimenter, the subjects were required to adjust the noise level in the standard picture until the noise level in the two pictures appeared to be the same. The difference in the actual or measured noise levels was taken as the apparent increase in noise level. Judgments were obtained for various numbers of repeated frames, for both interlaced and sequentially-scanned pictures, and for several values of signal-to-noise ratio. The study was restricted to band-limited white gaussian noise. A single still picture (Fig. 1) was used in all the trials as a background against which the noise was viewed. Details of the test conditions and subject instructions are given in Appendix B.

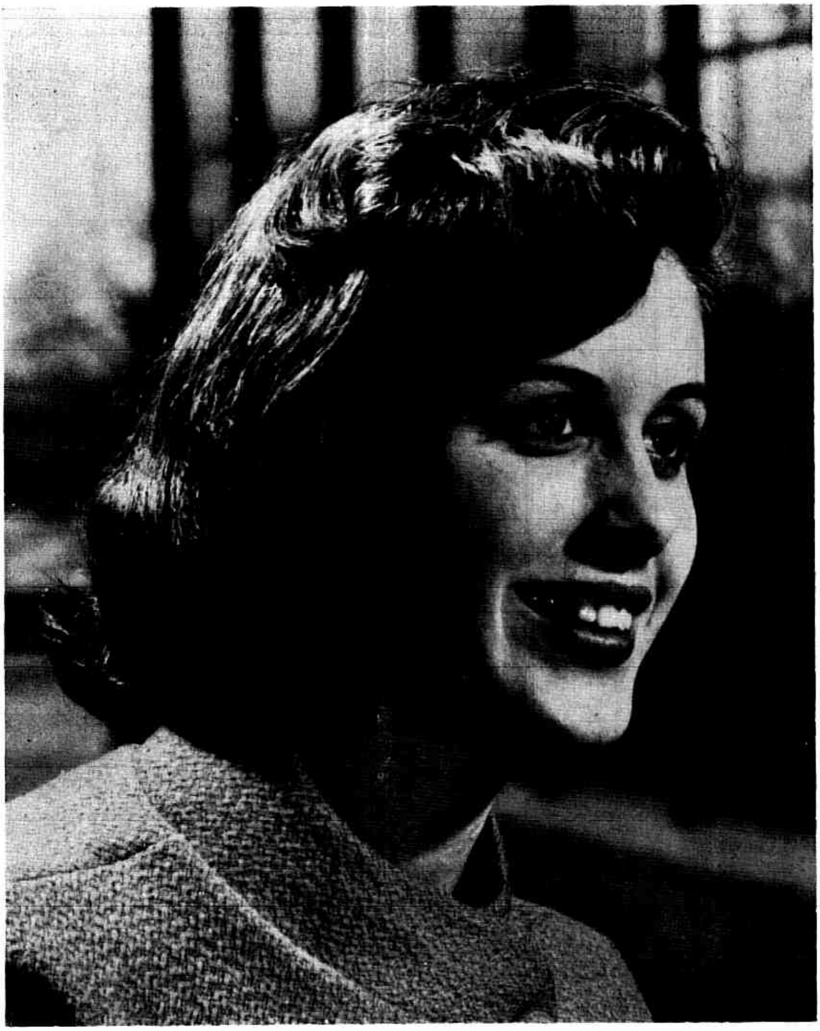


Fig. 1 — Test picture used in the experiment.

IV. RESULTS

Averaged results for the 24 subjects are shown in graphical form in Figs. 2 and 3. In both sets of graphs the average apparent increase in noise level is plotted as a function of the repetition ratio. With sequential scanning (Fig. 2) the time taken to scan a complete frame

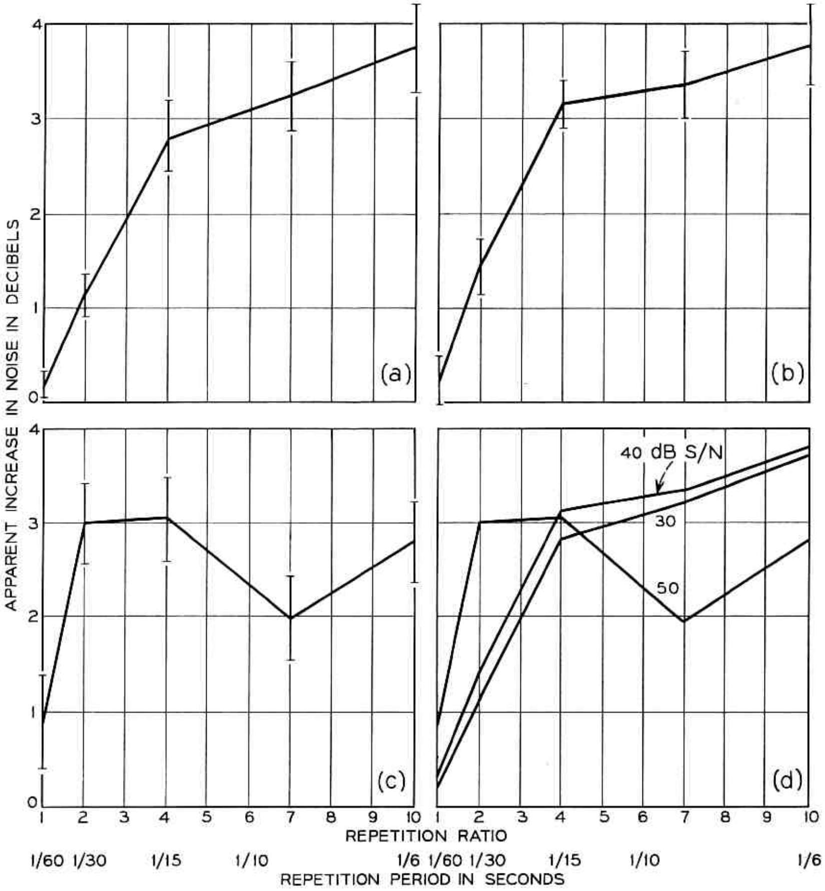


Fig. 2 — Sequential scanning: (a) 30 dB signal-to-noise ratio; (b) 40 dB signal-to-noise ratio; (c) 50 dB signal-to-noise ratio; (d) combined results.

was 1/60 second, whereas with interlaced scanning (Fig. 3) it was 1/30 second. This fact often leads to semantic confusion when comparing the two cases; what we have termed 2:1 frame-repetition of an interlaced picture is sometimes loosely referred to as 4:1 frame-repetition by virtue of the fact that the time period during which the picture is repeated is the same as that for 4:1 sequentially-scanned pictures. To emphasize our usage we have indicated the repetition period in seconds along the horizontal axis in all graphs.

Separate curves are plotted for each of the three signal-to-noise ratios used in both sequential- and interlaced-scan conditions. These signal-to-noise ratios are the true or measured ratios and refer to the frame-repeated picture, not the standard picture used for a comparison; thus, at higher numbers of repetitions the apparent signal-to-noise ratios are less than the stated figures by an amount equal to the ordinate of the curve. No attempt has been made to fit a smooth curve to the measured points as there was no way of knowing what type of curve to fit. Instead, the points have been connected by straight-line sections.

Several subjects experienced difficulty in adjusting for subjective

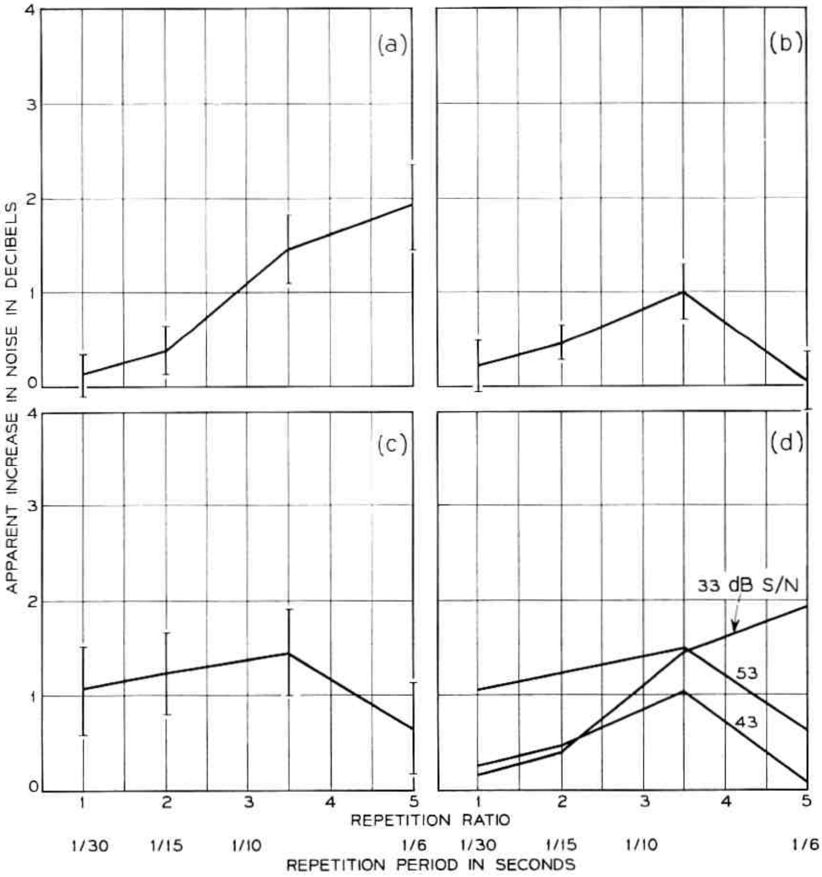


Fig. 3 — Interlaced scanning: (a) 33 dB signal-to-noise ratio; (b) 43 dB signal-to-noise ratio; (c) 53 dB signal-to-noise ratio; (d) combined results.

equality when the signal-to-noise ratio was 50 dB in the sequential case and 53 dB in the interlaced case. At these levels the noise is very near the threshold of visibility, and even allowing for the increase in noisiness with frame-repetition, the noise is barely perceptible. Furthermore, at these signal-to-noise ratios small alterations in noise level are much less noticeable than at 30 dB or 40 dB signal-to-noise ratios. The subjects discovered that for a 50 dB signal-to-noise ratio they could rotate their attenuator control through a number of steps without affecting the relative appearance of the two pictures. Two subjects maintained that altering the attenuator over its full range made no perceptible difference to the picture at 50 dB, and in consequence, when presented with a comparison at this noise level, merely reiterated their setting for the previous presentation. The variability of the 50 dB settings, as compared with the 30 dB and 40 dB settings, reflects these difficulties. In Figs. 2a, b, and c, and 3a, b, and c the standard deviations σ of the plotted means are shown. These were calculated according to the formula

$$\sigma = \left[\frac{1}{24}\sum_{i=1}^{24}(X_i - \langle X\rangle_{\text{av}})^2\right]^{\frac{1}{2}}$$

where the X_i are the 24 results whose mean $\langle X\rangle_{\text{av}}$ is plotted in the graph. The vertical lines about each plotted point extend to $\pm\sigma$. Figure 2d is a superimposition of Figs. 2a, b, and c for comparison purposes. Similarly Fig. 3d is a superimposition of Figs. 3a, b, and c.

V. DISCUSSION

A point of interest about the results is the obvious difference between the graphs for interlaced and sequentially-scanned pictures. The apparent increases in noise level are substantially larger in the sequential case (Fig. 2) than in the interlaced case (Fig. 3). For example, consider four presentations of each picture. From Fig. 2d the average increase in noise level for sequential scanning is about 3 dB, while from Fig. 3d the increase is seen to be a little over 1 dB. If this comparison is deemed to be unfair because the period of repetition is not the same in each case, then consider the 4:1 frame-repeated sequential case against the 2:1 frame-repeated interlaced case. Again, the difference is substantial.

A partial explanation of this difference may be given in terms of the time period of repetition. The scanning of a single frame takes twice as long in the interlaced case as it does in the sequential case. Thus, in standard or non-frame-repeated interlaced pictures every noise sample or element on the screen is seen for twice as long (1/30 second) before replenishment by a different noise sample as it is in sequentially-scanned pictures (1/60 second).* With a 2:1 frame-repeated sequentially-scanned picture, noise samples are replenished by different samples every other frame, that is, also at 1/30 second intervals. Therefore, crudely speaking, a standard interlaced picture is a 2:1 frame-repeated sequentially-scanned picture with a different order of line presentation. Hence, from Fig. 2 we would expect about 1.5 dB difference in the apparent noise level between standard interlaced and standard sequential pictures having the same added noise pattern, with the sequential picture having the higher apparent signal-to-noise ratio.† We have observed with our system that, by taking a picture with a fixed amount of added noise and changing the read-out method from sequential to interlaced with no frame-repetition, there appears indeed to be a slight increase in noise level.‡ This observation is only that of the authors, however, and we have yet to confirm it under properly-controlled conditions with a larger sample of subjects. Until this experiment is performed it should not be assumed, from Fig. 2 and 3, that a 4:1 frame-repeated sequential picture looks noisier than a 2:1 frame-repeated interlaced picture, each having the same signal-to-noise ratio.

If an ordinary interlaced picture can be equated to a 2:1 frame-repeated sequential picture, it follows that the upper portion of Fig. 2d, above a horizontal line drawn through 1.5 dB, should correspond to Fig. 3d. It can be seen that this correspondence is by no means exact, although in the case of the 30 dB signal-to-noise ratio curve it is quite close. The 50 dB curves are unreliable because of the previously-mentioned difficulties of adjustment at low noise levels, so that the 40 dB curve represents the main discrepancy and obstacle to accepting the correspondence. In Fig. 2d the 40 dB curve closely follows that of the 30 dB curve (an analysis of variance showed no significant difference between the plotted points) while in Fig. 3d the 43 dB curve diverges from the 33 dB curve and dips down to zero at higher repetitions. We carefully examined this phenomenon and

* The decay time of the phosphor may be considerably less than 1/30 second, but the sample is seen for a longer period owing to the persistence of vision.

† Having the same added noise pattern implies a 3 dB difference in their signal-to-noise ratios (see Appendix A). The interlaced picture has the higher actual S/N.

‡ Roughly estimated at between 1 and 3 dB.

conclude that it may well be explicable in terms of an interference or masking effect due to the interline flicker. At high levels of noise the masking is inoperative, but at lower levels it becomes predominant and acts to reduce the apparent noisiness. It also has a greater effect on the large-grain slow-moving noise produced by the higher repetitions. These conclusions are derived from personal observation and are tentative, but they do explain the shape of the Fig. 3 curves to some extent. A further point which should be clearly borne in mind in evaluating the curves is that the difference limen for random noise is probably at least 1 dB at 30 and 40 dB signal-to-noise ratios and greater at 50 dB signal-to-noise ratio; the increases in apparent noise level and variations in the apparent increase which are being considered are therefore quite small and, to the average person, frequently indiscernible.

Another point of interest in the graphs (more evident in Fig. 2 than Fig. 3) is the rapid rise in apparent noise level at small numbers of repetitions followed by a general flattening out or saturation above 1/15–1/10 second (66–100 milliseconds). This corresponds roughly to the integration period or critical duration of the eye.[4] Below the critical duration, the eye sums "frozen" noise frames and sees increasing granularity with increasing frame-repetition. Above the critical duration the granularity stays constant, but the apparent spatial movement of the noise becomes slightly more noticeable with larger numbers of repetitions. It is unfortunate that in the region which shows most promise from the point of view of useful band-compression without noticeable picture deterioration (2:1, 3:1, or 4:1 sequential and 2:1 interlaced) the increase in noise level is most rapid.

Notice finally, that subjects exhibited a slight bias in preference in the experiment toward the standard comparison picture. This can be seen in Figs. 2 and 3 by the positive intercept in apparent noise level increase at the 1:1 repetition ratio in all the graphs. This bias may have been due to slight differences in the brightness and contrast of the frame-repeated and standard pictures, as well as to difficulties in measuring signal-to-noise ratio to an accuracy of less than 1 dB.

VI. CONCLUSIONS

The apparent increase in noise level due to frame-repeating is fairly small: between 1 and 3 dB in the range of repetition ratios which are likely to be of practical interest for bandwidth-savings (up to 4:1 with sequential-scanning, 2:1 with interlacing).

With sequentially-scanned pictures the increase in apparent noise level is greater than with interlaced pictures, but interlaced pictures appear to be noisier to start with. Without further experimentation it is not possible to say with certainty whether a frame-repeated sequential picture is noisier than a frame-repeated interlaced picture, when each has the same signal-to-noise ratio.

The rate of increase in apparent noise level is greatest in the region which shows most promise from the point of view of useful bandwidth-saving without picture deterioration (up to 4:1 frame-repetition with sequential scanning, 2:1 interlaced).

The apparent increase in noise level is largely independent of signal-to-noise ratio, with one possible exception: low-level noise in interlaced pictures appears to be masked by interline flicker at the higher numbers of repetitions.

VII. ACKNOWLEDGMENT

The authors gratefully acknowledge the assistance of J. B. Pestrichelli in setting up and carrying out the experiments.

APPENDIX A

Details of the Experimental Apparatus

A block diagram of the experimental equipment is shown in Fig. 4. A 60 frame per second, 160-line, sequentially-scanned video signal was derived from a vidicon camera. To this signal random gaussian noise at one of two levels was added, dependent on the setting of switch S_1. In position A, corresponding to frame-repetition (see linked switch S_2), the noise level was completely and solely under the control of the experimenter (attenuator I), and in practice was set such that the signal-to-noise ratio (measured as the peak-white to black-level signal voltage divided by the rms noise voltage) was, in the case of line-sequential scanning, either 30, 40 or 50 dB at the display monitors. For line-interlaced scanning the levels of added noise were not changed, giving values of signal-to-noise ratios 3 dB higher at the display monitors, that is, 33, 43 and 53 dB. This was because the different manner of readout from the frame store with interlacing effectively reduced both signal and noise bandwidths by a factor of 2. In the B position the noise level was determined in part by the experimenter (attenuator II) and in part by the subject. The experimenter would, in practice, set attenuator II to 8 dB less than attenua-

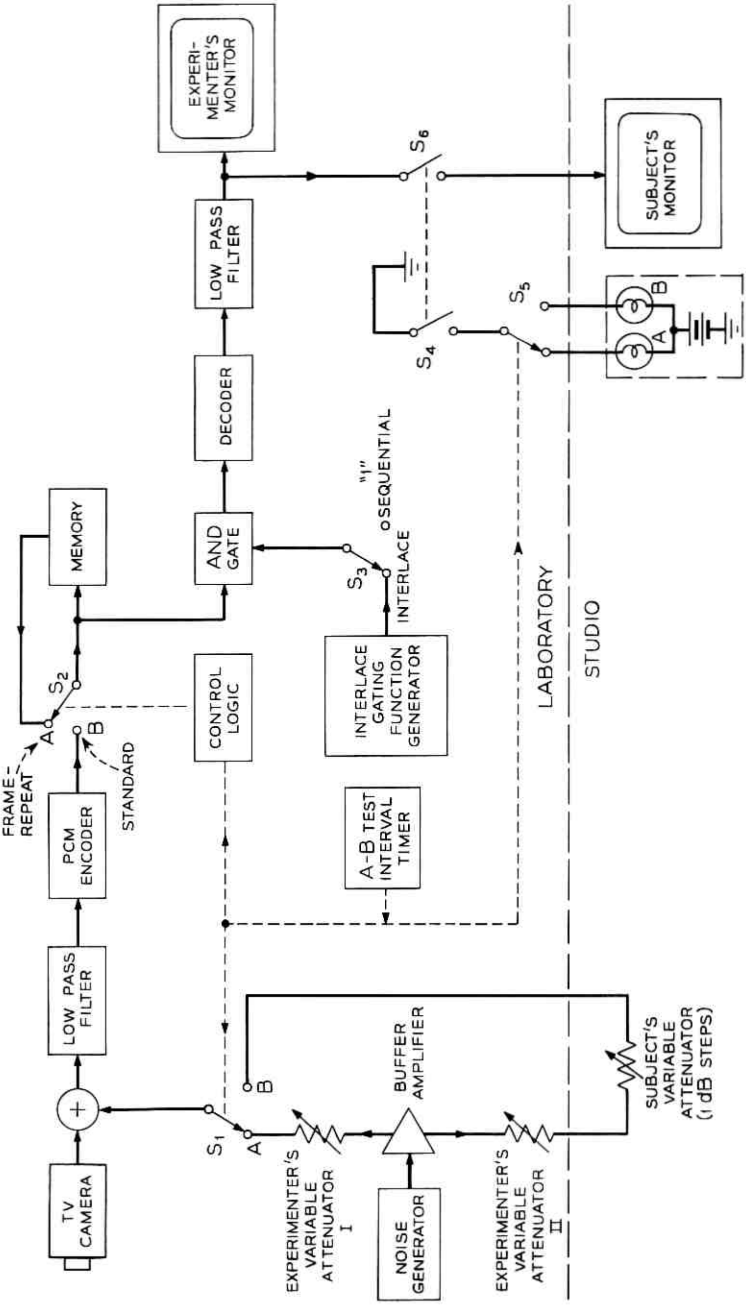


Fig. 4 — Block diagram of the experimental equipment.

tor I, that is, to give a 22, 32 or 42 dB signal-to-noise ratio (sequential scanning). The subject's attenuator was adjustable in steps of 1 dB over a range of 10 dB, so that, given, for example, a 32 dB setting on the experimenter's attenuator II and 40 dB on attenuator I, the subject could vary the signal-to-noise ratio over the range 32-42 dB. This proved to be an adequate adjustment for the subject to match *A* and *B* noise levels in almost all cases.

Subsequent to the addition of noise, the video signal was passed through a low-pass filter of bandwidth 768 kHz to the frame-repeating equipment.* This equipment, consisting of a PCM encoder, memory, decoder, and control logic, has been fully described in Ref. 3. With switch S_2 in position *B* no frame-repetition occurred and the picture seen on the subject's and the experimenter's monitor screens was a standard 60 frame per second picture with added noise. With the switch in position *A*, however, every *n*th frame was stored in the memory and read out to the display monitors *n* times in succession. The experimenter was able to select any *n* in the range 1–11. By means of switch S_3 (independently controlled and not linked to any of the other switches) the experimenter could choose to display the contents of the memory in either line-sequential or line-interlaced fashion. In both cases readout from the memory was line-sequential, the interlacing being produced by subsequently blanking every alternate line of the readout (the blanking signal was delayed by one line period in alternate readout scans to give the interlacing effect). The subjective effect of this type of interlacing is exactly equivalent to conventional interlacing when a still picture is used. For example, consider a single stored frame of a plain white picture to which noise has been added. In both the conventional interlaced readout and the sequential alternate-line readout, the noise patterns as seen on the display monitor will be identical. Viewers will not appreciate that, in the alternate-line blanking case, the lines are scanned at twice the rate with pauses in between lines. With conventional interlaced readout the noise bandwidth and the noise power are halved, and the peak-signal to rms noise ratio increased by 3 dB. A corresponding 3 dB increase in signal-to-noise ratio has, therefore, been assumed for the alternate-line blanking method used in this instance.

In both interlaced and sequential cases the memory readout was displayed on a sequentially-scanned monitor. Linked to S_3, but not shown, was an arrangement of attenuators in the video path to the

* The characteristics of this filter are fully described in Ref. 3.

display monitors such that the screen brightness and contrast was unchanged in switching from sequential to interlaced pictures. This ensured that interlaced and sequentially-scanned pictures were seen under exactly similar conditions.

Linked switches S_1, S_2, and S_5 were controlled by an automatic timer which repeatedly switched between A and B conditions, each condition being presented for 2½ seconds. To avoid visible transients these switches were arranged to operate only during the frame flyback interval. Switch S_5 controlled two lights labeled A and B, visible to the subject, in order to indicate to him which presentation he was currently viewing. Linked switches S_4 and S_6 enabled the experimenter to set up a condition on his own monitor before presentation to the subject.

APPENDIX B

Test Conditions

The viewing distance for each subject was approximately 25 inches, the picture size being 5 inches × 5 inches. Screen highlight and lowlight luminances were maintained at 60 foot-lamberts (206 cd/m^2) and 3 foot-lamberts (10 cd/m^2) respectively. Ambient illuminance was approximately 5 foot-candles (54 lm/m^2).

Noise level matching was carried out by the method of adjustment.[5] Subjects were introduced to the method in the following way. On arrival for their test, an A-B pair was presented to them (the A and B presentations occurring successively on the same screen for 2½ seconds each) in which the B presentation was noticeably noisier than the A. It was then demonstrated that by adjustment of the step attenuator (see Fig. 4) it was possible to lower the noise level in B until it matched the noise level in A. Subjects were invited to try the matching for themselves, and in all cases, with very little practice, succeeded in mastering the technique. It was explained that a number of similar pairs would be presented to them, and that for each pair they were required to adjust the attenuator until the noise levels in the A and B pictures were the same. An unlimited time was allowed for the adjustment, most subjects taking about one minute, with a few taking as much as two minutes. When the pictures were matched to their satisfaction the subjects reported the attenuator setting to the experimenter. The quantity: Attenuator I setting − (Subject's attenuator setting + attenuator II setting) was taken as the apparent increase in noise level due to frame-repeating.

All of the presentations were made with the same still picture, a portrait of a girl (Fig. 1). Had a moving picture been used, the subject's judgment would have been confounded by the motion break-up with frame-repetition. By using a still picture, the only visible difference between the two pictures was in respect to noise level.

Based on the results of preliminary experiments, repetition ratios of 1:1, 2:1, 4:1, 7:1, and 10:1 (60, 30, 15, 8.6, and 6 new frames per second respectively) were chosen to cover the range of interest for sequentially-scanned pictures. For interlaced pictures 1:1, 2:1, 3.5:1, and 5:1 (30, 15, 8.6 and 6 new frames per second respectively) were used. The 1:1 repetition here was identical to a standard or non-repeated picture, and was included as a check on the validity and accuracy of the *A*-*B* comparisons.

The 27 *A*-*B* presentations, consisting of all combinations of the 3 signal-to-noise ratios, the two methods of scanning and the various frame-repetition ratios (5 for sequential, 4 for interlaced scanning) were presented to subjects in random order, the order being different for each subject. Subjects made only one match per *A*-*B* pair. Twenty-four subjects were tested and the mean increases in apparent noise level, together with the standard deviation between subjects, were calculated for each of the 27 conditions.

REFERENCES

1. Gouriet, G. G., "Bandwidth Compression of a Television Signal," Proc. IEE, *104,* part B, No. 15 (May 1957), pp. 265–272.
2. Brainard, R. C., Mounts, F. W., and Prasada, B., "Low-Resolution TV: Subjective Effects of Frame Repetition and Picture Replenishment," B.S.T.J., *46,* No. 1 (January 1967), pp. 261–271.
3. Mounts, F. W., "Low-Resolution TV: An Experimental Digital System for Evaluating Bandwidth-Reduction Techniques," B.S.T.J., *46,* No. 1 (January 1967), pp. 167–198.
4. Graham, C. H., *Vision and Visual Perception,* New York: John Wiley & Sons, Inc., 1965, pp. 209–211.
5. Guilford, J. P., *Psychometric Methods,* New York: McGraw-Hill, 1954, p. 86.

The Generation and Accumulation of Timing Noise in PCM Systems— An Experimental and Theoretical Study

By J. M. MANLEY

(Manuscript received September 30, 1967)

Three sources of timing noise in a self-timed regenerative PCM repeater, namely, tank circuit mistuning, amplitude to phase conversion, and pulse shape, were studied both experimentally and theoretically. We discuss how these noises accumulate and combine along a chain of repeaters.

The theoretical work is from the viewpoint of frequency analysis which leads easily to the spectrum of the timing noise. We first give a simple form of this theory applicable in a number of cases, and then a more general form useful in other cases, which shows the approximations and limitations of the simple theory.

We found that the spectrum of timing noise caused by tank circuit mistuning has no energy at zero frequency and because of this fact, timing noise from this source does not build up indefinitely along a chain of repeaters but soon reaches a limit. On the other hand, the spectrum of timing noise caused by amplitude to phase conversion does have energy at zero frequency; thus, timing noise from this source increases indefinitely along a repeater chain. Some of the timing noise is attributable to pulse shape alone and in some cases may include a very low frequency part. This latter comes about through the small energy near the harmonics of the pulse rate in the tuned circuit response and the aliasing of this energy down to very low frequencies by the sampling process used in measuring the phase deviation or in generating the retiming pulses.

I. INTRODUCTION AND SUMMARY

A considerable amount of work has been done and results published on the subject of timing noise in pulse code modulation (PCM) systems.[1–11] The material in this paper comes from work which began as an experimental investigation. This led to some successful simple theories and generalizations along somewhat different lines from those followed previously.

It appears to be impossible for a regenerative repeater to perfectly restore a train of signal pulses to its original form because of the difficulty of obtaining a perfect timing source at a repeater location remote from the transmitter. The most widely used simple method of obtaining a timing wave is to pass the incoming pulse train, or some modification of it, through a narrowband resonant tank circuit, tuned as nearly as possible to the pulse rate. Since the tuning of the tank is unlikely to coincide with the pulse rate, since the bandwidth of this selective circuit is finite, and for other reasons, the derived timing wave is not perfect. Through this imperfect timing source, a certain amount of timing noise is added at each repeater to that already present in the incoming signal train.

Because this noise arises from imperfections in the system, it may be considered to be analogous to the modulation interference noise in amplitude systems caused by small departures from linearity in various components. Thus, if the narrowband tank circuit could be centered exactly on the pulse rate and kept there, and if the pulse generating circuits were always triggered exactly at a zero crossing of the timing wave, and if nonlinearity were not required to generate the pulse rate, then major sources of timing noise at a regenerative repeater would not exist.

As pointed out by W. R. Bennett and others, the principal effects of timing noise are two:[1]

(*i*) At any one repeater, the phase of the timing wave may be displaced in an irregular way from the proper place for optimum gating of signal pulses so that, at best, the tolerance of the system to noise is reduced and, at worst, errors in recognition of pulses or spaces are made.

(*ii*) Even if the sequence of pulses and spaces arrives at the receiver with no errors, the decoded signal samples will be irregularly spaced, thus introducing into the signal circuits a distortion which has the frequency of the deviation. The seriousness of this effect depends on its magnitude and the character of the signal. This effect is analyzed by W. R. Bennett in Ref. 1.

A program of measurements for studying the properties of this timing noise, and how it accumulates along a chain of regenerative repeaters, was begun because of the difficulties which had been encountered in trying to calculate the noise. It was planned to isolate the sources of timing noise and so consider each one separately and then in combination. The work described here is not concerned with the effects of the noise on the system.

Measurements were not made on chains of varying numbers of real repeaters. Instead the chain is simulated by one real special repeater and a multitrack tape recorder as indicated by Fig. 1. While the previous repeater output is being reproduced from two tracks of the recorder and used as input for the real repeater, the new output is being recorded on two other tracks. One of each pair of tracks is used for the pulse train and the other for timing information. Because no recorder has steady enough speed, the timing wave cannot be recorded directly. Instead, the phase deviations are detected and these are recorded. During playback, the timing wave is reconstructed with a phase modulator. The recording is sufficiently long so that statistical fluctuations are well smoothed. For most of the work the pulse train consisted of random unipolar pulses at a 1 kHz rate.

The first results obtained were on the noise caused by mistuning of the timing tank in one repeater and then two in tandem. Study of these results led to the development of a simple theory for the gen-

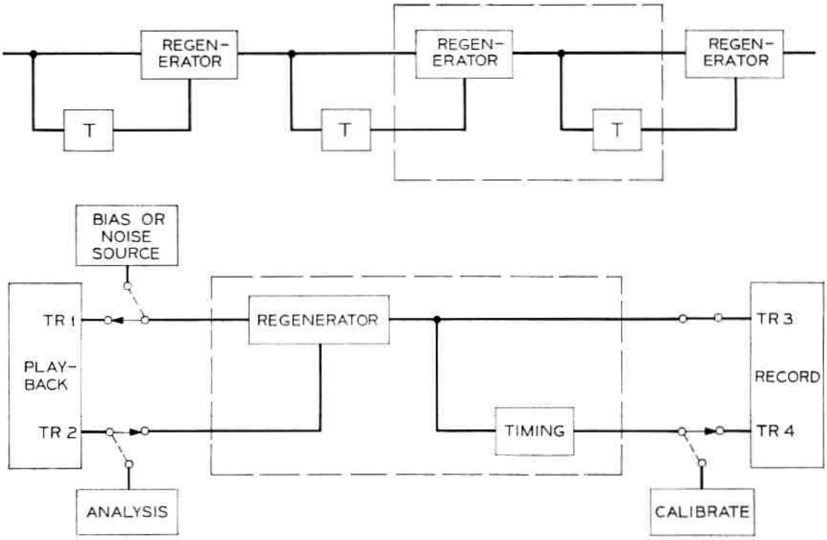


Fig. 1 — Simulation of regenerative repeater chain.

eration of the timing noise and its accumulation along a chain of repeaters. Subsequent work demonstrated that this theory may be used to calculate the noise satisfactorily, not only for longer chains, but for amplitude to phase conversion sources of timing noise as well. Very good agreement was obtained between the noise calculated from this theory and that measured.

A brief summary of these results obtained when narrow rectangular pulses are used follows:

The spectra of timing noise at each of the repeaters in a chain of six, all mistuned alike, are shown in Fig. 2. This noise is designated

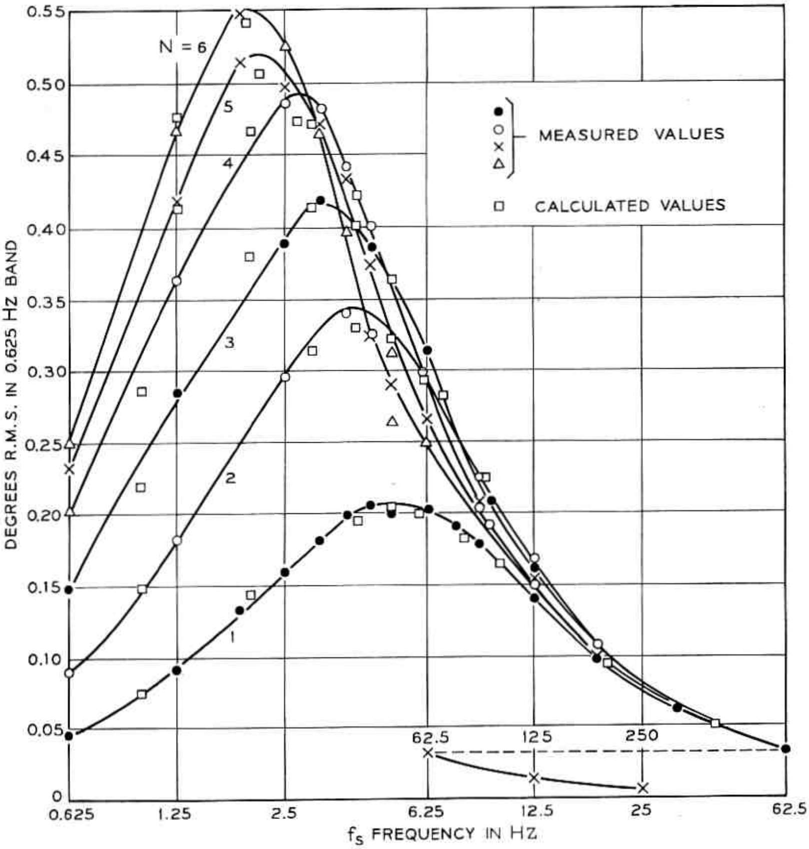


Fig. 2 — Spectra of timing noise caused by 0.1 percent mistuning of timing tank, $Q = 100$. Random rectangular pulses, 10 percent duty factor. N = number of like repeaters in chain.

Type A; the most important characteristic of these spectra is that there is no energy at zero frequency. Because of this, the peak of the spectrum at the end of a chain never becomes larger than four times the peak at the first repeater, no matter how long the chain is. The root mean square (rms) value of the total noise increases to about twice the amount at the first repeater as we go along a very long chain. If, as is most likely, some of the repeaters in the chain are mistuned in the opposite direction or to a smaller degree, the noise at the end is smaller than that above. Thus it is seen that mistuning of the timing tank is not a factor in the accumulation of large amounts of timing noise in a long chain of regenerative repeaters.

The situation is different, though, if we have a pulse generator whose trigger point is offset from a zero crossing. The timing noise in this case is a direct consequence of the amplitude variation of the timing wave, which variations have a spectrum with nonzero value at zero frequency. As shown in Fig. 3, this causes the rms value of very low frequency timing noise to increase linearly at successive repeaters in a chain having equally offset triggers in each. This noise is designated type B. The total noise at the end of the chain increases without limit as the number of repeaters increases. The total amount varies inversely as the Q of the tank circuits.

It was demonstrated that the theory applies also when both mistuning and amplitude-to-phase conversion are present simultaneously.

Spectra of timing noise caused by pulse shape alone are shown for several particular shapes in Fig. 4. While the total noise for the wider pulses is fairly large here, because it is spread over a wide frequency band, the magnitude of the undesirable very low frequency components is quite small. For example, the total noise for the asymmetrical overlapping pulses is only one-fifth the amount per repeater measured in the T1 system.[8] The results of this investigation indicate that some form of amplitude-to-phase conversion is probably the greatest source of very low-frequency timing noise.

The idea that, for the propagation of phase deviations, the chain of regenerative repeaters resembles a chain of tandem low-pass resistance-capacitance (RC) filters follows from considering the phase deviation to be a modulation of a carrier wave at the pulse rate. Experiments verified this idea which had been suggested earlier.[8]

A brief outline of the simple theory and method of calculation will now be given; a more detailed description will be made in Sections 2.1—2.5. Consider the spectrum of the incoming pulse train,

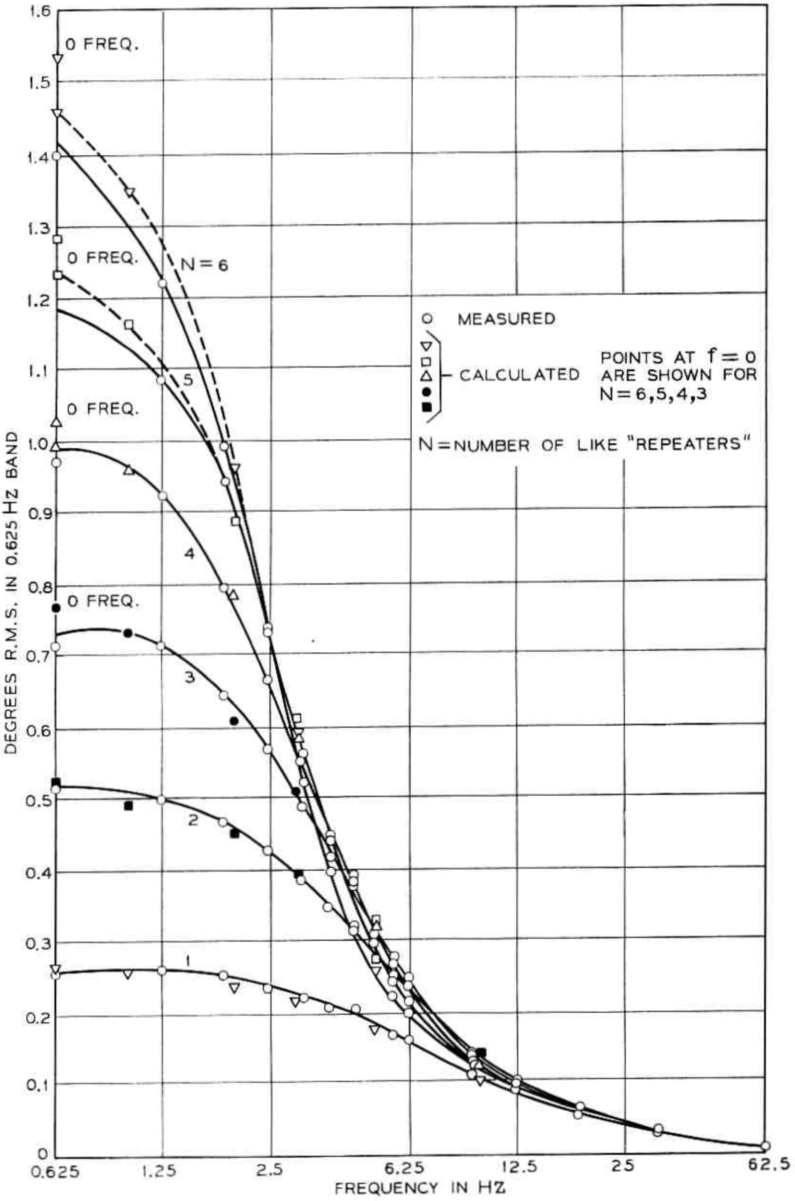


Fig. 3 — Spectra of timing noise caused by timing wave amplitude variations and trigger circuit offset from zero. Tank tuned to pulse rate. Random rectangular pulses, 10 percent duty factor. N = number of like repeaters in chain.

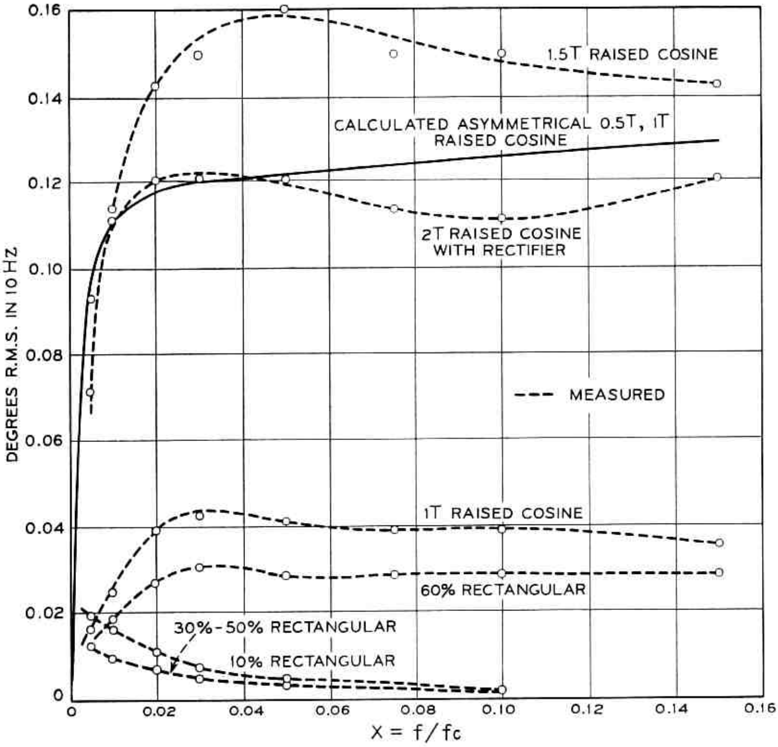


Fig. 4 — Timing noise spectra caused by various pulse shapes.

which will contribute very little timing noise itself if the pulses are narrow.* This spectrum consists of discrete components at harmonics of the pulse rate and a broadband of a special kind of noise.[1] Next consider this broadband of noise to be divided into small evenly spaced bands, each small band replaced by a single frequency component having the same power as the small band. Taking these side-frequencies in pairs about the pulse rate, we may think of the spectrum of the pulse train in the vicinity of the pulse rate as a carrier wave amplitude modulated by a number of small components.

The tuned circuit by which the pulse rate fundamental is selected from this spectrum to provide a timing wave, also admits some of the noise side frequencies which are still symmetrical if the tuned circuit is centered exactly on the pulse rate. When this tank circuit is

* The meaning of this is brought out in Sections 4.3, 2.6.5, and 2.6.6.

detuned from the pulse rate, the side-frequency pairs in the tank response are no longer symmetrical about the carrier. Sometimes the dissymmetry is described by saying that a quadrature carrier term has been introduced. The dissymmetry in amplitude, or phase, or both, is equivalent to phase modulation of the timing wave. This phase modulation of the timing wave is transferred to the outgoing pulse train in the regeneration process.

Next, assume that these asymmetrical side frequencies (which are another description of the above phase modulation), after being attenuated and phase shifted in transmission through the narrowband tank of the second repeater, would add directly to the corresponding ones newly generated by the detuning of the second tank. Phase deviation calculated from this assumption agrees very closely with that measured, not only after two repeaters, but after many have been traversed.

If the amplitude modulation at the tank output, corresponding to the symmetrical components, is not entirely removed by a limiter, that remaining may cause further phase modulation. For example, if a pulse generating circuit is supposed to trigger at a zero crossing of the timing wave but actually triggers a few degrees away from zero, amplitude variations of the timing wave will be converted to phase variations. The magnitude of these phase variations and their increase along a chain of repeaters may be successfully calculated using the same methods described above.

The simple theory applicable in a number of cases, and which leads easily to the spectrum of timing noise, is inadequate for pulse shapes other than narrow ones. Here the pulse train itself is now a source of timing noise. For example, if the pulses have a finite width, a small additional amount of noise (type B) can arise, although not in all cases. The limitations of the simple theory and how it fits into a more general theory is discussed in detail in Section 2.6.

The more general theory, described in Section 2.6, is also developed from a frequency viewpoint and is based on analysis by S. O. Rice to whom I am indebted for this work. With it the amount of timing noise in the situations of the previous paragraph were calculated. Also, this theory was used to calculate the timing noise for raised cosine pulses two time slots wide, hence with large enough overlapping so that a non-linear device is required in order to derive the pulse rate fundamental. It was found that in this case, the timing noise is Type A (that is, it does not build up in a long chain of repeaters) with a small qualification discussed in Section 2.6.6.

Another source of type B noise (which builds up indefinitely) is in any low-frequency distortion of the pulse spectrum if this is followed by a nonlinear operation of any kind.

II. THEORY OF PHASE NOISE GENERATION AND ACCUMULATION

As mentioned in the Section I, the simple theory is described first, with the more general one and its relation to the simple one being discussed in Section 2.6. The principal area in which the simple theory is satisfactory is that in which the pulses are narrow. In this case the pulse train itself causes very little timing noise, and so other sources may be considered separately.

2.1 *Spectrum of Narrow Pulse Train*

It is assumed that the message pulses are represented by a random train of narrow, rectangular, unipolar pulses. By random pulse train is meant one having regularly spaced pulse positions which are filled or not at random. Although most of the work was done for an average pulse density of one-half, the result would not be appreciably different except in magnitude, if this parameter differed somewhat from the value one-half.

The spectrum of this train has been calculated by W. R. Bennett.[1] Part of the spectrum is a series of harmonics, the fundamental of which is the pulse rate, and the magnitudes of which are determined by the shape of the individual unit pulses. The spectrum of the other part has the same shape as the envelope of harmonics, and is continuous and therefore is a noise. Bennett points out that while this noise is like thermal noise in some respects, that is, for example, in the proper frequency band the two sound alike; nevertheless it has a phase structure which thermal noise does not.

If the original pulses are rectangular of height V_o and duration τ and occur at regular intervals $T = 1/f_c$ with a probability of 1/2, Bennett's calculation shows that the mean square value of the fundamenal term at $f = f_c$ is

$$A_i^2/2 = (V_o^2/2\pi^2) \sin^2 \pi\tau/T \tag{1}$$

and that the mean square value (in a band B Hz wide) of the noise part is

$$W(f) = B(V_o^2/2\pi^2)(f_c/f^2) \sin^2 (\pi\tau f). \tag{2}$$

From a somewhat different point of view, the spectrum of this random train of narrow pulses may be calculated by first considering the train to consist of repeated blocks of random pulses, each block being N pulse periods T in length. The Fourier series representation of this train consists of harmonics of the pulse rate f_c plus single frequency "noise" components spaced f_c/N apart. The harmonics have nonzero average values. While the noise components have zero average values, their average powers are nonzero. If N is made to approach infinity, this spectrum approaches that calculated by Bennett. The representation by finite components means, in effect, that a band of noise f_c/N Hz wide is replaced by a discrete term having the same mean square value. If S and A_i are the rms amplitudes of one of the noise components and the pulse rate fundamental, respectively, then, as shown in (113), Section 2.6.5,

$$S^2/A_i^2 = \frac{(Bf_c)\sin^2(\pi f\tau)}{f^2\sin^2(\pi\tau f_c)}\,, \tag{3}$$

where f_c/N has been replaced by B. This agrees with (1) and (2) from Bennett's calculation. In the vicinity of the fundamental, we have

$$S/A_i \approx (B/f_c)^{\frac{1}{2}}.$$

The spectrum of this representation of the pulse train is shown in Fig. 5.

Near the pulse rate component, the noise amplitudes on both sides of it are nearly equal because the pulses are narrow. For example, when $\tau/T = 0.1$, values of S for components 2.5 percent above and below f_c differ by about 0.2 percent. Hence an approximate representation of this region of the spectrum is

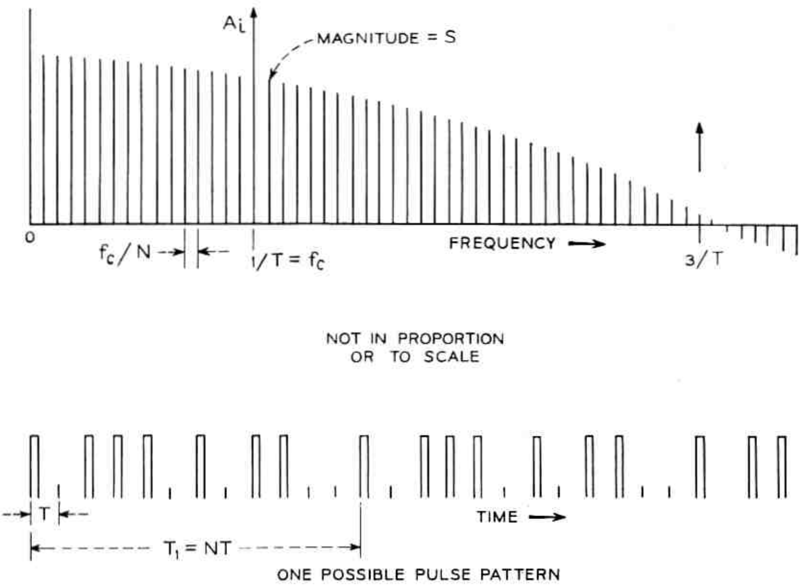
$$E_i = A_i[1 + \sum a_k \cos(2\pi kf_ct/N)] \cos 2\pi f_ct \tag{4}$$

where

$$a_k = 2S/A_i \approx 2(B/f_c)^{\frac{1}{2}} \tag{5}$$

which describes the input as an amplitude modulated carrier. A more accurate representation would include a modulated quadrature term to account for the slight dissymmetry of side frequencies.

It is the regular spacing of the random pulses which gives the phase structure to the noise spectrum, causes zeros of the wave E_i to appear at regular intervals $T = 1/f_c$ apart, and which makes possible the representation in (4).



NOT IN PROPORTION OR TO SCALE

ONE POSSIBLE PULSE PATTERN

Fig. 5 — Pulse pattern and spectrum.

In the simple theory, attention is given only to that part of the pulse spectrum in the vicinity of the pulse rate fundamental as indicated by the representation (4). The strength of the noise terms with respect to the fundamental is obtained by the indicated statistical averaging of the pulse train components with the result (3). The reasons why these simplifications are satisfactory are discussed in Section 2.6.

In the Section 2.2, the response of the tuned circuit to this restricted portion of the pulse spectrum, considering the noise terms to have fixed amplitudes, is calculated.

2.2 *Response of Tuned Circuit to Narrow Pulses*

In calculating the response of the tuned circuit to (4), we need to consider only one representative modulation term of frequency $q/2\pi = kf_c/N$, namely

$$E_{ik} = A_i[1 + 2(S/A_i) \cos qt] \cos \omega_c t. \tag{6}$$

The response to the two side frequencies $S \cos (\omega_c \pm q)t$ of this one term is

$$\begin{aligned} E_{os} &= S_1 \cos (\omega_c t + qt + \theta_1) + S_2 \cos (\omega_c t - qt + \theta_2) \\ &= S_1 \cos [(\omega_c t + \varphi) + qt + (\theta_1 - \varphi)] \\ &\qquad + S_2 \cos [\omega_c t + \varphi) - qt - (\varphi - \theta_2)], \qquad (7) \end{aligned}$$

where θ_1 and θ_2 and φ are the phase shifts received by the two side frequencies and the carrier respectively in going through the tuned circuit. If the tuned circuit is resonant at ω_c, the symmetry of the side frequencies about ω_c in both amplitude and phase which exists in (4) is preserved in the response. But if it is resonant at ω_o, different from ω_c, the response side frequencies are unsymmetrical as indicated in Fig. 6. In this case, they may be resolved into a pair with even symmetry and a pair with odd symmetry, or into a component in phase with the carrier and another in quadrature with the carrier. That is, we get

$$\begin{aligned} E_{os} &= A_s \cos [\omega_c t + \varphi + qt + \varphi_s] + A_s \cos [\omega_c t + \varphi - qt - \varphi_s] \\ &\qquad + A_a \cos [\omega_c t + \varphi + qt + \varphi_a] - A_a \cos [\omega_c t + \varphi - qt - \varphi_a] \\ &= 2A_s \cos (qt + \varphi_s) \cos (\omega_c t + \varphi) \\ &\qquad - 2A_a \sin (qt + \varphi_a) \sin (\omega_c t + \varphi) \qquad (8) \end{aligned}$$

where

$$\begin{aligned} 2A_s &= \{S_1^2 + S_2^2 + 2S_1S_2 \cos [(\varphi - \theta_2) - (\theta_1 - \varphi)]\}^{\frac{1}{2}} \\ 2A_a &= \{S_1^2 + S_2^2 - 2S_1S_2 \cos [(\varphi - \theta_2) - (\theta_1 - \varphi)]\}^{\frac{1}{2}} \qquad (9) \\ \tan \varphi_s &= \frac{S_1 \sin (\theta_1 - \varphi) + S_2 \sin (\varphi - \theta_2)}{S_1 \cos (\theta_1 - \varphi) + S_2 \cos (\varphi - \theta_2)} \\ \tan \varphi_a &= \frac{S_2 \sin (\varphi - \theta_2) - S_1 \sin (\theta_1 - \varphi)}{S_2 \cos (\varphi - \theta_2) - S_1 \cos (\theta_1 - \varphi)}. \end{aligned}$$

Thus the resonant tank response to the amplitude modulated wave (6) is

$$\begin{aligned} E_o &= [A_o + 2A_s \cos (qt + \varphi_s)] \cos (\omega_c t + \varphi) \\ &\qquad - [2A_n \sin (qt + \varphi_a)] \sin (\omega_c t + \varphi). \qquad (10) \end{aligned}$$

When S/A_t is small, as at present, (10) is approximately described by

$$\begin{aligned} E_o &\approx A_o[1 + 2(A_s/A_o) \cos (qt + \varphi_s)] \cos [(\omega_c t + \varphi) \\ &\qquad + 2(A_a/A_o) \sin (qt + \varphi_a)]. \qquad (11) \end{aligned}$$

The tuned circuit to which the train of random pulses is applied

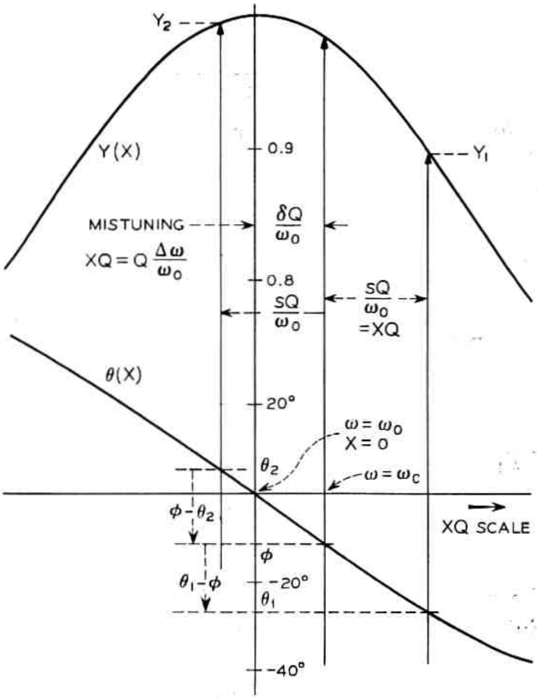


Fig. 6 — Resonant circuit characteristics.

in order that the pulse rate fundamental may be selected is described by

$$\frac{I}{E} = Y(j2\pi f) = \frac{1}{R + j\left(\omega L - \frac{1}{\omega C}\right)}. \tag{12}$$

Let

$$\begin{aligned} \omega_o^2 LC &= 1 \\ \omega_o L/R &= Q \\ \omega &= n\omega_o + \Delta\omega \\ \Delta\omega/\omega_o &= x. \end{aligned} \tag{13}$$

Then

$$Y(j2\pi f_o) = \frac{1}{R} \tag{14}$$

and

$$Y(j2\pi f)/Y(j2\pi f_o) = \frac{n + x}{n + x + jQ[(n + x)^2 - 1]}. \tag{15}$$

This general formula will be used in the treatment for wide pulses. Here we will be concerned only with frequencies in the neighborhood of f_o and so we set $n = 1$. Thus

$$Y_1(j2\pi f)/Y(j2\pi f_o) = \frac{1 + x}{1 + x + j2Qx(1 + x/2)}. \tag{16}$$

When $\Delta\omega/\omega_o = x$ is small, a satisfactory approximation to the transmission $Y_1(j2\pi f)R$ is

$$Y_1(j2\pi f)R \approx \frac{1}{1 + j2Q\ \Delta\omega/\omega_o}. \tag{17}$$

The phase of Y_1 is θ, where $\tan\theta = -2Q\Delta\omega/\omega_o$. Let

$$\begin{aligned} x_o &= \delta/\omega_o, & x_q &= q/\omega_o, \\ x_1 &= x_o + x_q, & x_2 &= x_o - x_q, \end{aligned} \tag{18}$$

where δ is the amount of tank circuit detuning from the carrier and q is the modulation frequency. It is convenient also to write

$$\begin{aligned} S_1/S &= \mid Y[j2\pi f_0(1 + x_o + x_q)] \mid R = y_1 = \cos\theta_1 \\ S_2/S &= \mid Y[j2\pi f_o(1 + x_o - x_q)] \mid R = y_2 = \cos\theta_2. \end{aligned} \tag{19}$$

The amounts of amplitude and phase modulation in the tank circuit response (11) can now be given explicitly. Writing

$$\frac{2A_s}{A_o} = \left(\frac{S}{A_i}\right)\left(\frac{A_i}{A_o}\right)\left(\frac{2A_s}{S}\right),$$

then substituting (5) for the first ratio on the right, and the upper of (9) for the third ratio, and noticing that

$$A_o/A_i = \mid Y[j2\pi f_o(1 + x_o)] \mid R,$$

we have

$$\begin{aligned} 2A_s/A_o = {} & (B/f_c)^{\frac{1}{2}}(1 + \tan^2\varphi)^{\frac{1}{2}} \\ & \cdot\{y_1^2 + y_2^2 + 2y_1y_2\cos[(\varphi - \theta_2) - (\theta_1 - \varphi)]\}^{\frac{1}{2}} \end{aligned} \tag{20}$$

Similarly,

$$\begin{aligned} 2A_a/A_o = {} & (S/A_i)(A_i/A_o)(2A_a/S) = (B/f_c)^{\frac{1}{2}}(1 + \tan^2\varphi)^{\frac{1}{2}} \\ & \cdot\{y_1^2 + y_2^2 - 2y_1y_2\cos[(\varphi - \theta_2) - (\theta_1 - \varphi)]\}^{\frac{1}{2}}. \end{aligned} \tag{21}$$

When the amount of detuning is small (0.1 percent detuning with $Q = 100$ makes $\tan \varphi = 0.2$), the expression inside the large radicals of (20) and (21) may be simplified so that

$$2A_s/A_o \approx \frac{2(B/f_c)^{\frac{1}{2}}}{[1 + (2qQ/\omega_o)^2]^{\frac{1}{2}}} \tag{22}$$

$$2A_a/A_o \approx \frac{8(B/f_c)^{\frac{1}{2}}(\delta Q/\omega_o)(qQ/\omega_o)}{1 + (2qQ/\omega_o)^2}\,, \tag{23}$$

and also $\varphi_s \approx \theta$.

The two expressions in (22) and (23) describe the amplitudes of the amplitude and phase modulation in (11) which are the responses of the resonant tank circuit to an amplitude modulated wave representing part of the incoming pulse train. Before discussing the meaning of these results for the generation of phase modulation or how it accumulates in a chain of repeaters, we will calculate the amount of phase deviation generated by another possible source within the repeater.

2.3 *Amplitude to Phase Conversion Factor of Offset Trigger*

In an ideal repeater, a perfect limiter following the resonant tank circuit would remove all the amplitude modulation from the derived timing wave (11) which would then, at one of its zero crossings, trigger a pulse generator as a part of the retiming process. But in a real repeater, the limiting would not be perfect so that some amplitude modulation remains on the timing wave; also the trigger point may have drifted away from the zero crossing. This is one way in which amplitude variations of the timing wave are converted to phase variations in a regenerative repeater. The diagram of Fig. 7 illustrates the conversion.

Referring to Fig. 7, where the triggering level has been offset by the bias b or by the angle γ_o which are related by

$$\sin \gamma_o = b/A_o\,, \tag{24}$$

it is seen that the change, $\Delta\gamma$, in triggering angle for a change, ΔA, in amplitude is

$$\Delta\gamma = -\tan \gamma_o(\Delta A/A_o). \tag{25}$$

The amplitude variation in (11) at the tank circuit output is reduced by a factor K_L in the limiter which follows, so that the input

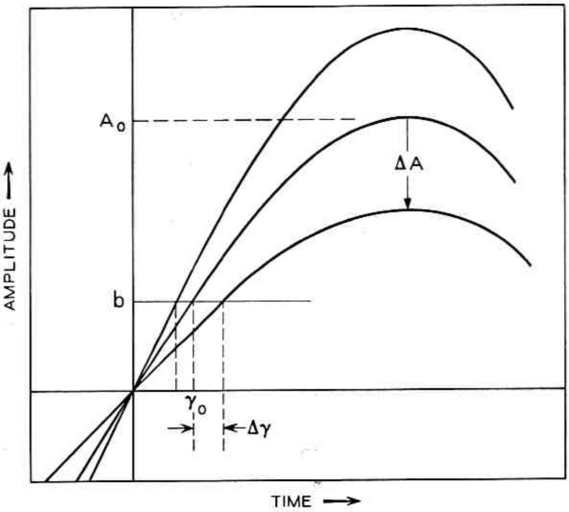


Fig. 7 — Amplitude-to-phase conversion in trigger with offset bias.

for the pulse generator is

$$A_o[1 + K_L(2A_s/A_o) \cos (qt + \varphi_s)] \cos [(\omega_c t + \varphi) + (2A_a/A_o) \sin (qt + \varphi_a)]. \qquad (26)$$

Hence the phase variation introduced by the offset trigger, using $K_L 2A_s/A_o$ from (22) for $\Delta A/A_o$, is

$$\Delta\gamma = \frac{-2(B/f_c)^{\frac{1}{2}} K_L \tan \gamma_o}{[1 + (2qQ/\omega_o)^2]^{\frac{1}{2}}} \cos (q_t + \varphi_s). \qquad (27)$$

2.4 *Application of These Results to Timing Noise*

The phase variations (referred to the carrier fundamental) of the pulse generator driven by (26) are then given by the sum of $\Delta\gamma$ from (27) and $2(A_a/A_o) \sin (qt + \varphi_a)$ from (23), that is,

$$\varphi_1 = \frac{8(B/f_c)^{\frac{1}{2}}(\delta Q/\omega_o)(qQ/\omega_o)}{1 + (2qQ/\omega_o)^2} \sin (qt + \varphi_a) - \frac{2(B/f_c)^{\frac{1}{2}} K_L \tan \gamma_o}{[1 + (2qQ/\omega_o)^2]^{\frac{1}{2}}} \cos (qt + \varphi_s) \quad \text{(radians)}. \qquad (28)$$

From the magnitudes of the two components of (28), the spectra of timing noise caused by tank circuit mistuning and by amplitude to phase conversion are determined. These are plotted in Fig. 8 with $K_L \tan \gamma_o = 1$. The phases φ_a and φ_s are plotted in Fig. 9. The first term in (28) specifies the spectrum if mistuning alone is present; the second term applies if mistuning is zero and there is amplitude to phase conversion. It is seen that the spectra in these cases are quite different at very low frequencies near $q = 0$. The phase modulation for mistuning alone is zero at zero frequency and has a maximum

$$\max\,(2A_a/A_o) = 2(B/f_c)^{\frac{1}{2}}(\delta Q/\omega_o) \tag{29}$$

at

$$2qQ/\omega_o = 1, \tag{30}$$

while that for amplitude to phase conversion alone has a maximum value of $2(B/f_c)^{1/2}K_L \tan \gamma_o$ at zero frequency. This difference between the two spectra at low frequencies has a very important effect on the accumulation of timing noise in chains of repeaters, as will be seen in Section 2.5. Second, we see that the first component depends directly on the amount of detuning, δ, and is zero for zero detuning. See Section 4.1. This emphasizes what Bennett has said about the difference between the noise spectra of random pulses and thermal noise.[1] In the latter case, phase noise would not be zero for zero detuning.

The good agreement between phase deviations calculated in this

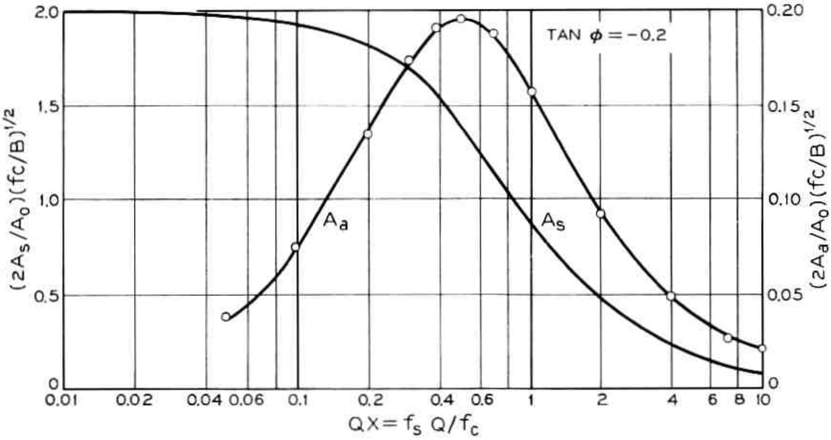


Fig. 8 — Amplitude spectra of symmetrical (A_s/A) and antisymmetrical (A_a/A) side frequencies caused by tank mistuning (calculated).

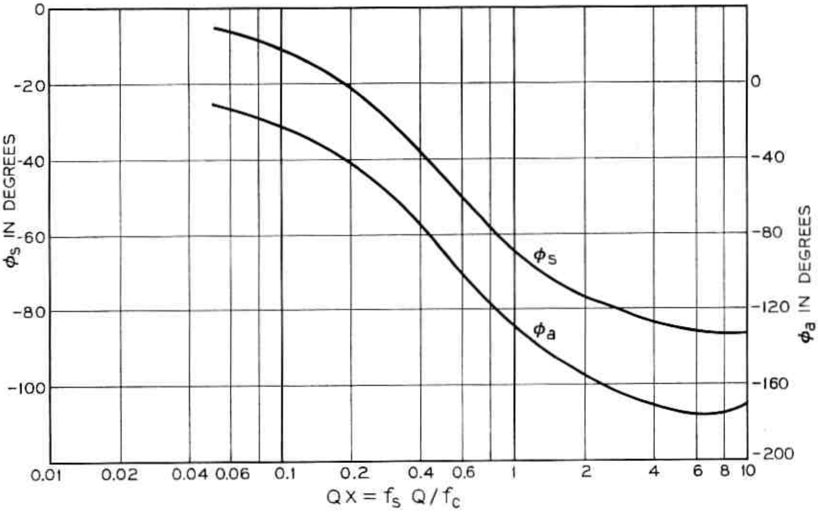


Fig. 9 — Phase spectra of symmetrical (Φ_s) and antisymmetrical (Φ_a) side frequencies caused by tank mistuning (calculated).

way, and measured values, has already been mentioned in the introduction and is shown in the curves of Fig. 2 and 3.

In one sense both of these effects are amplitude-to-phase conversion with the amplitude modulated carrier representation of part of the pulse train, as in (4), being the original amplitude variation. From this viewpoint, in both (23) and (27),

$$2(B/f_c)^{\frac{1}{2}}$$

is the applied amplitude variation; the factor

$$[1 + (2qQ/\omega_o)^2]^{\frac{1}{2}}$$

is the attenuation of the tuned circuit; while

$$-K_L \tan \gamma_o$$

is the amplitude-to-phase conversion factor in (27), and

$$\frac{4(\delta Q/\omega_o)(qQ/\omega_o)}{[1 + (2qQ/\omega_o)^2]^{\frac{1}{2}}}$$

is the corresponding conversion factor for mistuning from (23).

When both tank mistuning and trigger offset are present in the timing wave path, their combined effect may be calculated by adding the two as suggested in (28). The correctness of this has been verified by experiment as seen in Fig. 10 where both measured data and values calculated according to (28) are plotted. This result emphasizes again that both of these phase modulation effects have a common source in the special noise side frequencies about the pulse rate in the spectral representation of the random pulse train.

2.5 *Accumulation of Phase Modulation in a Chain of Repeaters*

Next consider how phase modulation accumulates in a chain of repeaters when the same amount and kind of phase modulation is generated at each repeater of the chain.

Assume that phase modulation generated in the derived timing wave at repeater 1 is

$$\varphi_1 = \Phi_1 \sin (qt + \varphi_a) \tag{31}$$

and that this is passed along unchanged by the regenerator. Then at the input to repeater 2, this is equivalent to the presence of a pair

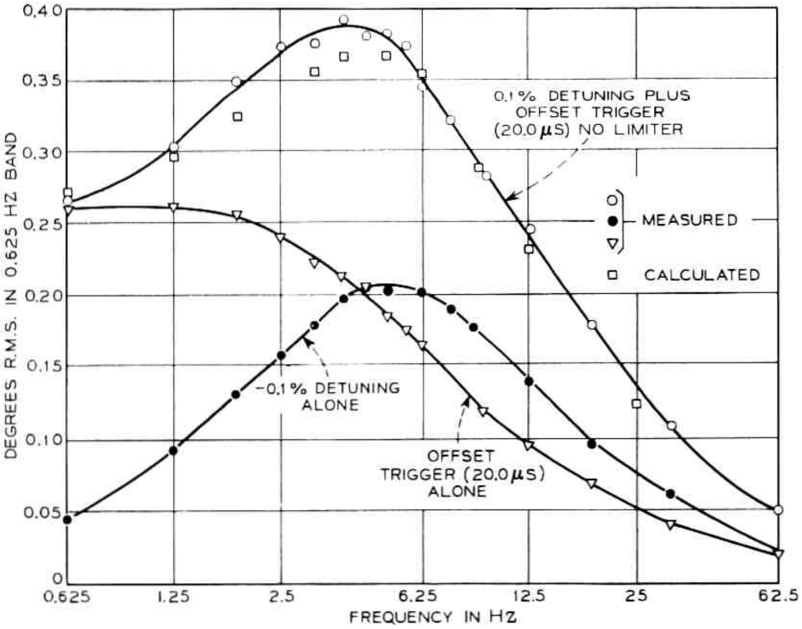


Fig. 10 — Spectrum of timing noise caused by tank detuning and trigger offset.

of antisymmetrical side frequencies

$$S_{\pm}/A = \pm(\Phi_1/2) \cos (\omega_c t + \varphi \pm qt \pm \varphi_a) \tag{32}$$

about the carrier. The transmission of these side frequencies through the tuned circuit of repeater 2 is governed by expressions (11), (20), and (22), developed above for the transmission of amplitude modulation. In using these in the present circumstance, the amplitude $\Phi_1/2$ in (32) takes the place of S/A_i in (20). The response of the tank circuit of repeater 2 to these side frequencies is then approximately

$$\pm\frac{\Phi_1/2}{[1 + (2qQ/\omega_o)^2]^{\frac{1}{2}}} \cos (\omega_c t + \varphi \pm qt \pm \varphi_a \pm \varphi_s) \tag{33}$$

if it is assumed that the pulse amplitude, and hence that of the carrier, are the same at repeater 2 as at repeater 1. This expression is independent of mistuning when the degree of mistuning is small, as we have assumed. In addition to this response we have the pair of antisymmetrical side frequencies, as expressed by (32), but now generated at repeater 2, namely

$$\pm\ (\Phi_1/2) \cos (\omega_c t + \varphi \pm qt \pm \varphi_a). \tag{34}$$

Thus at the output of the tank circuit of repeater 2, the antisymmetrical side frequencies are represented by the sum of these two, that is (34) and (33) or,

$$\pm(\Phi_1/2)\left\{\cos (\omega_c t + \varphi \pm qt \pm \varphi_a) \right.$$
$$\left. + \frac{1}{[1 + (2qQ/\omega_o)^2]^{\frac{1}{2}}} \cos (\omega_c t + \varphi \pm qt \pm \varphi_a \pm \varphi_s)\right\}.$$

Therefore the total modulation Φ_2 at the output of repeater 2 is seen to be

$$\Phi_2 = |\ \Phi_1 + \Phi_1 (\cos \theta) \exp (j\theta)\ | \tag{35}$$

where $\cos \theta$ has been substituted for

$$\frac{1}{[1 + (2qQ/\omega_o)^2]^{\frac{1}{2}}}$$

according to (19), and φ_s has been replaced by its approximation θ. Carrying through the same process for repeater 3, we find the phase modulation at repeater 3 output to be

$$\Phi_3 = \Phi_1\ |\ 1 + (\cos \theta) \exp (j\theta) + (\cos \theta)^2 \exp (j2\theta)\ |. \tag{36}$$

These show that, while the actual combining process at each repeater involves the direct addition of side frequencies, we can drop the carrier reference and consider the generation and propagation of modulation alone. And the Φ_1 generated at each repeater need not be only the component (23), but may be the more general one (28) in which mistuning and amplitude to phase conversion effects are combined.

The process begun in (36) may be generalized to give the amount of phase noise at the Nth repeater in a chain of N-like repeaters. This is

$$\Phi_N = \Phi_1 \mid (Y)_N \mid = \Phi_1 \left| \sum_{n=0}^{N-1} (Y_1R)^n \right| = \Phi_1 \left| \frac{1 - (Y_1R)^N}{1 - Y_1R} \right|, \tag{37}$$

where Y_1R has been written for (16). The approximation (17) to Y_1R and the relation (19) are used so that

$$Y_1R \approx \frac{1}{1 + j2x_qQ} = (\cos\theta) \exp(j\theta), \tag{38}$$

where $x_q = q/\omega_o$ as defined in (18). In (37), $(Y)_N$ is not an admittance as Y_1 is, but is a transfer function. By direct expansion of $(Y)_N$, using (38),

$$\mid (Y)_N \mid^2 = \frac{1 + (\cos\theta)^{2N} - 2(\cos\theta)^N \cos N\theta}{\sin^2\theta}. \tag{39}$$

$(Y)_N$ may be considered a sort of phase deviation transfer function for a chain of N repeaters when Φ_1 is the phase deviation generated in each repeater. $|(Y)_N|$ is plotted for $N = 4$, 30, 100 in Figs. 11, 12, and 13.

Since Φ_1 and $(Y)_N$ are functions of the phase deviation frequency q, $\mid \Phi_N \mid$ will describe the spectrum of accumulated phase noise. Consider first the case of tank circuit mistuning alone. From (23), we have

$$\Phi_1 = K_1 \frac{4x_qQ}{1 + (2x_qQ)^2} = 2K_1 \sin\theta \cos\theta \quad \text{(radians)} \tag{40}$$

where

$$K_1 = 2(B/f_c)^{\frac{1}{2}}(\delta Q/\omega_o). \tag{41}$$

Then for mistuning,

$$\mid \Phi_N \mid^2 = 4K_1^2 \cos^2\theta[1 + (\cos\theta)^{2N} - 2\,(\cos\theta)^N \cos N\theta]. \tag{42}$$

This is a rather unwieldy expression when N is large, but it may be

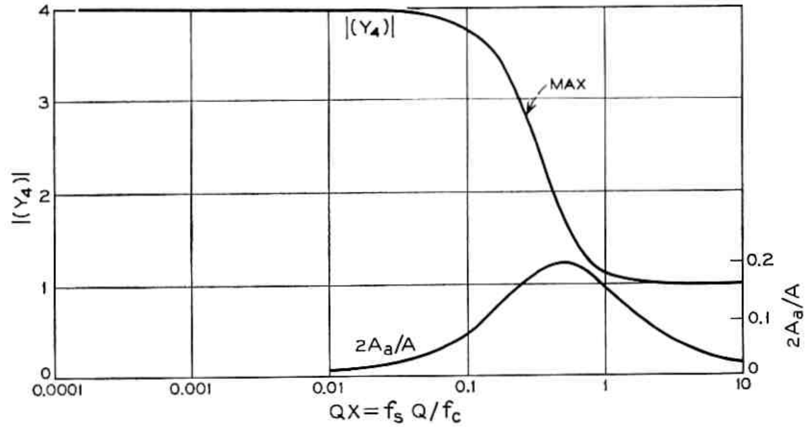


Fig. 11 — Effective transfer characteristic of four tandem timing tank circuits to modulation on pulse train.

approximated for the purpose of finding its maximum value by

$$| \Phi_N | \approx 4K_1 (\cos \theta)^{(1+N/2)} \sin (N\theta/2) \tag{43}$$

since θ is small for large N at the maximum of $| \Phi_N |$. The expression in (43) is maximum with respect to θ when

$$\tan \theta_m \tan (N\theta_m/2) = N/(N + 2). \tag{44}$$

A list of maximum values for Φ_N are given in Table I:

These are plotted in Fig. 14 along with four of the full calculated

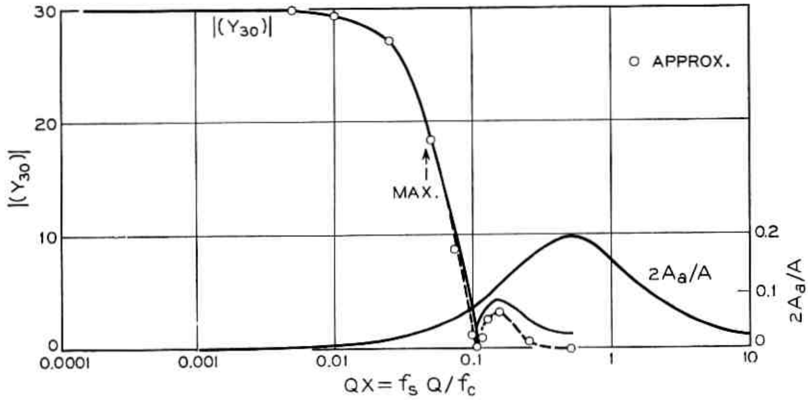


Fig. 12 — Effective transfer characteristic of thirty tandem timing tank circuits to modulation on pulse train.

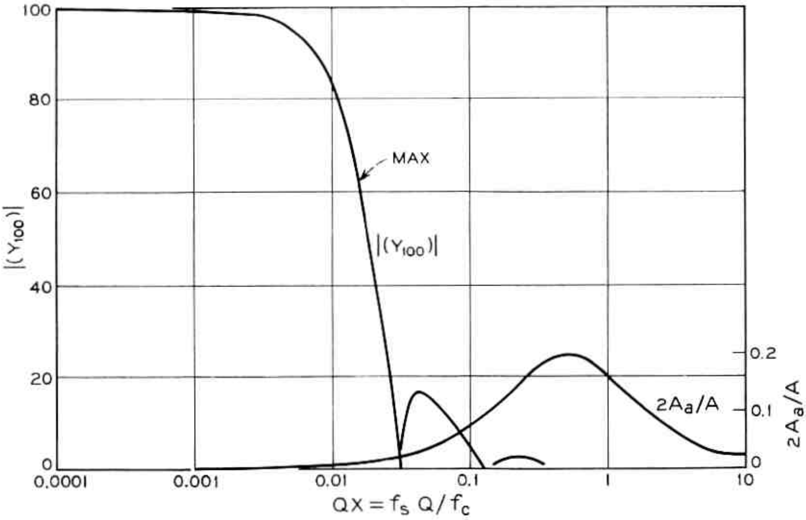


Fig. 13 — Effective transfer characteristic of one hundred tandem timing tank circuits to modulation on pulse train.

spectrum curves of Fig. 2. It is seen from (43) that $4K_1$ is the largest value Φ_N can have.

The spectrum of Φ_1 is plotted along with the $(Y)_N$ function in Figs. 11, 12, and 13 to suggest why the peak values of Φ_N do not increase indefinitely with N.

To get the total mean square "power" of Φ_N, $|\Phi_N|^2$ in (42) is integrated with respect to x from $x = 0$ to ∞ or, with respect to θ, from θ to $0 - \pi/2$. Since the expression for Φ_N in (42) gives the peak value, we have for the total mean square power P_N,

$$P_N = \frac{1}{2} \int_0^\infty [\Phi_N(\theta)]^2 \, df, \tag{45}$$

where a value of 1 Hz is used for the bandwidth B in K_1. Evaluation

TABLE I—MAXIMUM VALUES FOR Φ_m

N	θ_m	Qx_m	Φ_m	
1	$\pi/4$	0.5	K_1	exact
4	0.49	0.267	2.384 K_1	exact
4	0.463	0.249	2.290 K_1	approximate
10	0.2525	0.1290	3.14 K_1	approximate
30	0.0997	0.049	3.68 K_1	approximate
50	0.0603	0.0392	3.80 K_1	approximate
100	0.0308	0.0154	3.90 K_1	approximate

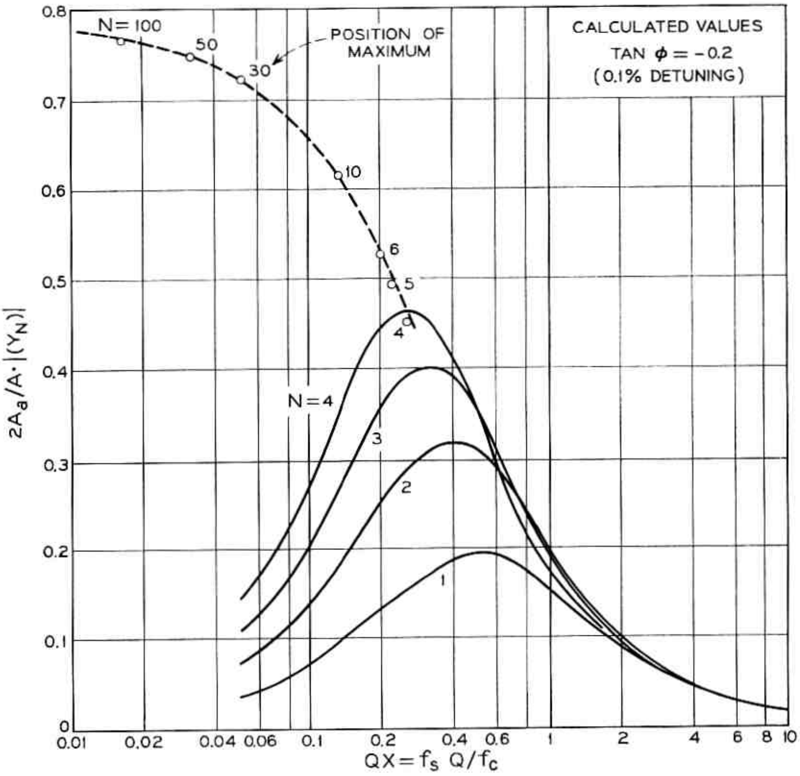


Fig. 14 — Calculated spectra of timing noise caused by tank circuit detuning.

of this integral gives

$$P_N = 2\pi Q \frac{\delta^2}{\omega_c^2} \left\{ 1 - \frac{1}{2^{N-1}} + \frac{1}{2^{2N}} \frac{(2N)!}{N!\,N!} \right\} \quad (\text{radians})^2 \tag{46}$$

ignoring the small difference between ω_o and ω_c. For $N = 1$, we have

$$P_1 = \pi Q \delta^2 / \omega_c^2 \tag{47}$$

and for $N \geqq 4$, P_N is approximated very closely by

$$P_N \approx 2\pi Q \frac{\delta^2}{\omega_c^2} \left\{ 1 - \frac{1}{2^{N-1}} + \frac{1}{(\pi N)^{\frac{1}{2}}} \right\}. \tag{48}$$

Twice the quantity in the braces is plotted in Fig. 15, where exact values from (46) are used through $N = 4$.

The expression in (47) for P_1 was derived in a different way by W. R. Bennett.[1] The expression (46) for P_N has been derived independently and differently by S. O. Rice.

When there is no tank circuit mistuning and amplitude to phase conversion is the only source of phase deviation, that generated at each repeater is given by (27), that is,

$$\begin{aligned}\Phi_1 &= \frac{2(B/f_c)^{\frac{1}{2}}K_L \tan \gamma_o}{[1 + (2qQ/\omega_o)^2]^{\frac{1}{2}}}\\ &= 2K_o \cos \theta \quad \text{(radians)}. \end{aligned} \tag{49}$$

In this case

$$| \Phi_N |^2 = 4K_o^2 \frac{\cos^2 \theta}{\sin^2 \theta} (\cos)^N [(\cos \theta)^N + (\cos \theta)^{-N} - 2 \cos N\theta]. \tag{50}$$

When $(\cos \theta)^N > 0.8$, that is for θ, and hence q, sufficiently small, the sum of the first two terms within the bracket is very close to 2. For this condition, Φ_N may be approximated by

$$\begin{aligned}| \Phi_N | &\approx 4K_o(\cos \theta)^{(1+N/2)} \frac{\sin (N\theta/2)}{\sin \theta}\\ &\approx 2K_oN[\cos (N\theta/2)](\cos \theta)^{(1+N/2)}, \end{aligned} \tag{51}$$

which shows that the phase deviation very near zero frequency increases directly with the number N of repeaters in the chain. This is in contrast with the similar result (43) for mistuning where the largest value of $| \Phi_N |$ is only four times that of $| \Phi_1 |$. $(Y)_N$ is the same for both. The difference arises because of the difference in spectrum shape of the generated phase deviation in the two cases. In mistuning, $| \Phi_1 |$ is zero at zero frequency, while in amplitude to phase conversion effects, $|\Phi_1 |$ is flat and nonzero for very low frequencies.

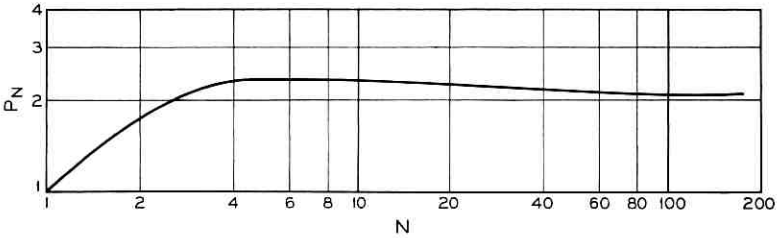


Fig. 15 — Calculated total timing noise caused by tank circuit mistuning as a function of the number N of like repeaters in a chain.

To find the total mean square "power" of Φ_N, the integral (45) is calculated, using (50) this time instead of (42). The value of this integral is given in a paper by Byrne, Karafin, and Robinson.[8] It is

$$P_N = \frac{\pi f_o}{Q} K_o^2 \left\{ N - \frac{1}{2} \frac{(2N-1)!}{4^{(N-1)}[(N-1)!]^2} \right\} \tag{52}$$

which shows that the total phase noise power increases directly with the number N of repeaters in a chain and varies inversely with the Q of the tank circuits.

In Ref. 8, the result (52) and the spectrum (50), or something closely related to it, were derived by assuming that at each repeater there was a noise source (nature and magnitude unknown) with a flat spectrum and "power" density $2K_0^2$. The process followed is similar to work first done by R. C. Chapman of Bell Telephone Laboratories. Also in the paper of Ref. 8, the results of measurements of accumulated timing noise in an experimental $T1$ system are given and the spectra of these are like those of Fig. 3.

2.6 *Spectrum and Phase Noise of Wide Pulses*

In most practical systems it is not desirable for the transmitted pulses arriving at the timing circuit input to have the narrow shape considered in the previous section of the paper. It is to be expected from the above discussions of the effects of dissymmetry between upper and lower side frequencies, that wider pulses having spectra with these characteristics would introduce phase noise in the derived timing wave even when there is no mistuning or amplitude to phase conversion. But the calculation of dissymmetry needs to be more elaborate for the wider pulses; the simple calculation used above is inadequate.

A more generally applicable theory is available from analysis by S. O. Rice who worked it out originally for pulses wide enough to spread over two time slots. How the simple theory is related to the more general one to be described in Section 2.61 will be discussed in Section 2.66. An outline of Rices' analysis follows. It is assumed that neither mistuning nor amplitude to phase conversion is present so that the phase noise calculated is caused by the pulse train alone.

2.6.1 *Fourier Series for Pulse Train*

The general theory is based on a frequency analysis also and so the pulse spectrum is calculated first. We consider the train to consist of repeated blocks of N pulse periods each, with each of the N pulse periods having a pulse or not at random. The train has a period NT and may be described by the Fourier series

$$I(t) = \sum_{m=-\infty}^{\infty} C_m \exp (j2\pi mt/NT). \tag{53}$$

If the parameter N is made to approach infinity, a random pulse train will be obtained. But a good approximation results if N is just large, say 100.

In order to consider pulses which may be as much as two pulse periods wide so that there is considerable overlapping of adjacent pulses, four auxiliary functions are necessary to specify the current $I(t)$ in any one pulse period of duration T. These functions correspond to the four possibilities of (*i*) no pulse present, (*ii*) only the leading edge of a pulse present, (*iii*) only the trailing edge of a pulse present, and (*iv*) overlapping pulses present.

This is illustrated in Fig. 16 which is a short section in time of a random train of raised cosine pulses exactly two pulse periods wide. The four possibilities mentioned above occur in that order in periods 3, 1, 2, 5 of Fig. 16. For a random train of pulses all these possibilities eventually occur in any one time slot as suggested in the composite drawing of Fig. 17a. In this a number of sections in time of Fig. 16 have been overlapped as they would be in an oscilloscope presentation. In Fig. 17a, the four possibilities in the order given above are *AC, AD, BC, BD*. Another illustration is the same wave after passing through a half-wave rectifier which begins conduction at the half amplitude level as shown in Fig. 17b. Figures 18a and b are photographs of oscilloscope patterns of real pulse trains which approximate the idealized ones of Figs. 17a and b.

Then $I(t)$ may be represented by a sequence of functions $I_n(t)$, each of which is specified by the auxiliary functions $F_2(t')$, $F_3(t')$, $F_4(t')$ and the parameter a_n as follows

$$\begin{aligned}
a_n &= 1, & a_{n-1} &= 1, & I_n(t) &= F_4(t') \\
a_n &= 1, & a_{n-1} &= 0, & I_n(t) &= F_2(t') \\
a_n &= 0, & a_{n-1} &= 1, & I_n(t) &= F_3(t') \\
a_n &= 0, & a_{n-1} &= 0, & I_n(t) &= 0,
\end{aligned} \tag{54}$$

in the time slot interval

$$(n - 1)T - \nu T \leqq t \leqq nT - \nu T$$

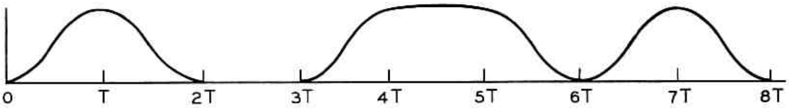


Fig. 16 — Short, time section of a random train of raised cosine pulses, each two time slots wide.

and is zero outside this interval. In this representation, $a_n = 1$ if a pulse begins in the interval under consideration and $a_n = 0$ if a pulse does not begin in it. The time scale t' used for describing the pulse train is related to the time scale t of the Fourier series by

$$t' = t - (n - 1)T + \nu T. \tag{55}$$

The time shift $(n - 1)T$ brings the pulse forms in the nth time slot back to the first one for description by the auxiliary F functions. The

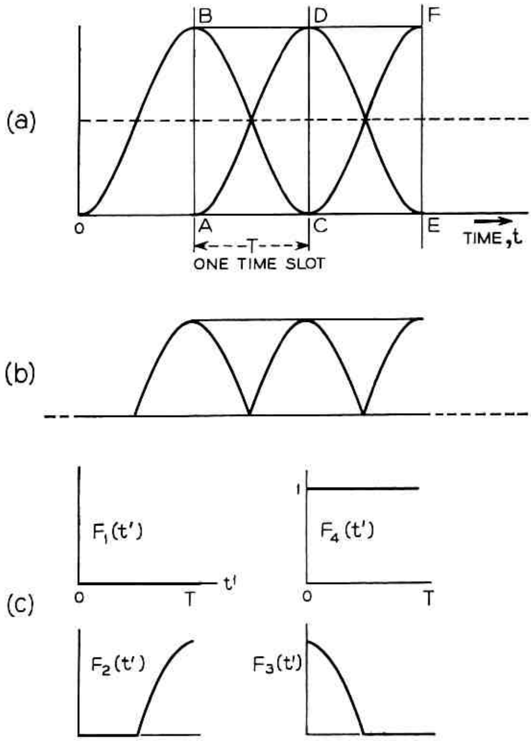


Fig. 17 — Idealized pulse waveforms: (a) Overlapping raised cosine; (b) wave in (a) applied to half-wave rectifier; (c) functions used in calculation.

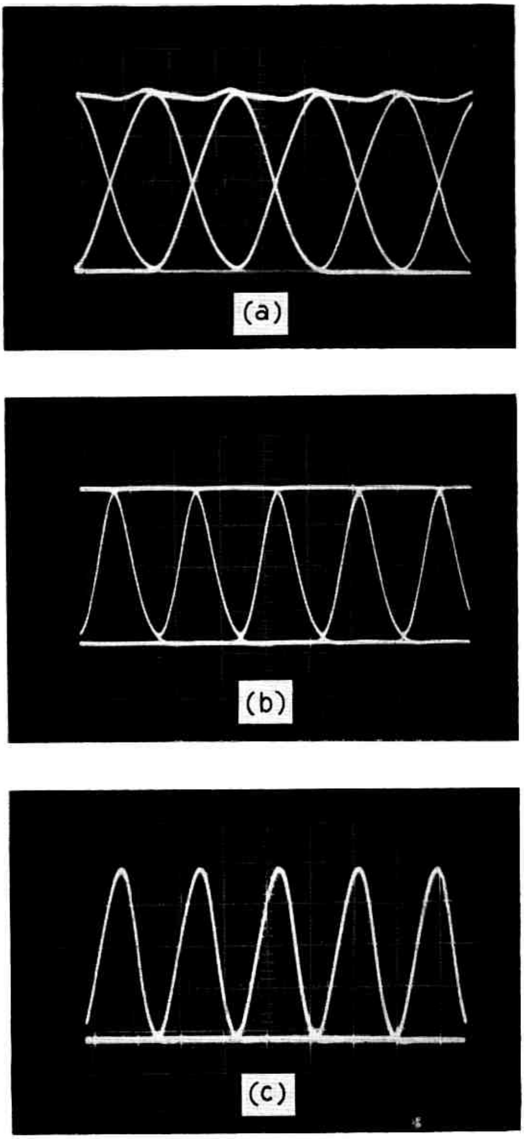


Fig. 18 — Photographs of random pulse oscilloscope traces: (a) Raised cosine pulses, two time slots wide at the base; (b) bottom half of (a) obtained with half-wave rectifier; (c) raised cosine pulses, one time slot wide at the base.

shift νT will be convenient later when it will be desirable to have the occurrence of pulses adjustable with respect to the zeros of the sinusoids in the Fourier series (53). To make $I(t)$ a random pulse train, we let $a_1, a_2, \ldots, a_n \ldots$ be independent random variables with probability of $a_n = 1$ being p and probability of $a_n = 0$ being $q = 1 - p$.

The representation just described was designed with the situation of Fig. 17 in mind; but it is valid for other pulse shapes including those which occupy just one time slot or less. The transformation of time scales is illustrated for rectangular pulses in Fig. 19.

The specifications (54) may be expressed by

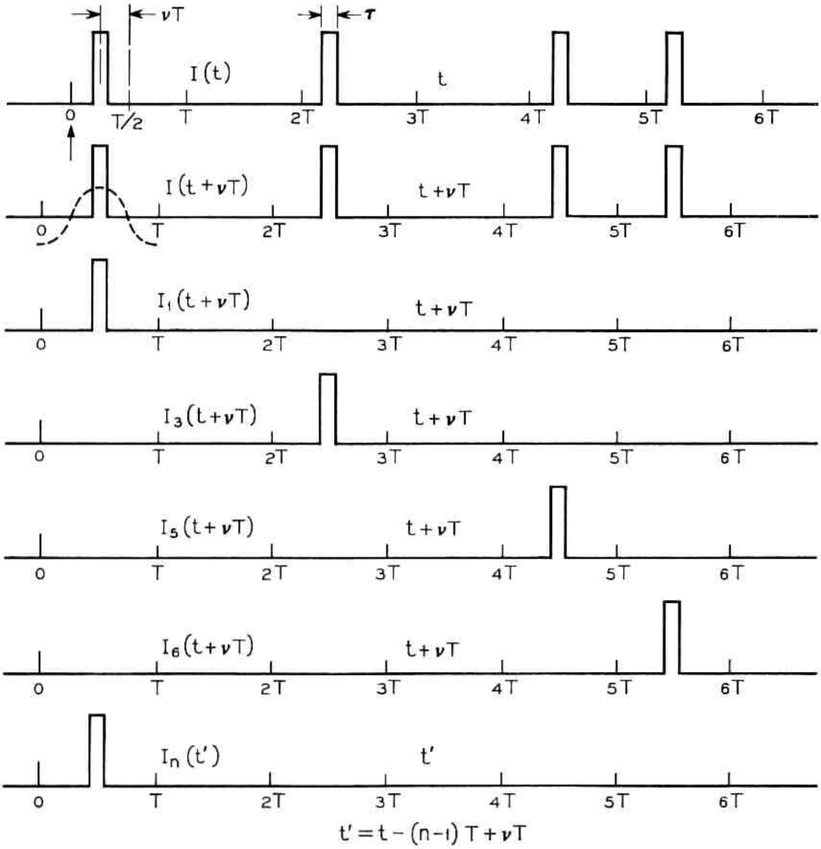


Fig. 19 — Representation of pulse train for Fourier analysis.

$$I_n(t) = a_n(1 - a_{n-1})F_2(t') + a_{n-1}(1 - a_n)F_3(t') + a_na_{n-1}F_4(t')$$

$$= a_nF_2(t') + a_{n-1}F_3(t') + a_na_{n-1}[F_4(t') - F_2(t') - F_3(t')] \quad (56)$$

and $I(t)$ is the sum of N separate $I_n(t)$.

2.6.2 *Fourier Coefficients*

The Fourier coefficients C_m in (53) are then

$$C_m = \frac{1}{NT}\int_{-\nu T}^{NT-\nu T} \exp(-j2\pi mt/NT)I(t)\,dt$$

$$= \frac{1}{NT}\sum_{n=1}^{N} \exp[-j2\pi m(n - 1 - \nu)/N]\int_0^T \exp(-j2\pi mt'/NT)$$

$$\cdot I_n[(n - 1 - \nu)T + t']\,dt' \quad (57)$$

where the integral from $-\nu T$ to $NT - \nu T$ has been written in the second form as a sum of N integrals, each over an interval T and then the transformation $t = (n - 1 - \nu)T + t'$ applied to each integral. Let

$$\beta_m = \frac{r_m}{T}\int_0^T \exp(-j2\pi mt'/NT)F_2(t')\,dt' \quad (58)$$

$$\gamma_m = \frac{r_m}{T}\int_0^T \exp(-j2\pi mt'/NT)F_3(t')\,dt' \quad (59)$$

$$\delta_m = \frac{r_m}{T}\int_0^T \exp(-j2\pi mt'/NT)[F_4(t') - F_2(t') - F_3(t')]\,dt \quad (60)$$

$$z = \exp(-j2\pi m/N), \qquad r_m = \exp(j2\pi\nu m/N). \quad (61)$$

Then

$$C_m = \frac{1}{N}\sum_{n=1}^{N} z^{n-1}(a_n\beta_m + a_{n-1}\gamma_m + a_na_{n-1}\,\delta_m), \quad (62)$$

and the average value of C_m is

$$\langle C_m\rangle_{\text{av}} = \frac{1}{N}\sum_{n=1}^{N} z^{n-1}(p\beta_m + p\gamma_m + p^2\,\delta_m). \quad (63)$$

Since

$$\sum_{n=1}^{N} z^{n-1} = 0 \qquad m \neq lN$$

$$= N \qquad m = lN,$$

where l is an integer, we have

$$\langle C_m \rangle_{\text{av}} = p(\beta_m + \gamma_m + p\ \delta_m), \qquad m = 0, \quad \pm N, \quad \pm 2N, \cdots,$$
$$\langle C_m \rangle_{\text{av}} = 0, \qquad \text{otherwise}. \tag{64}$$

The Fourier components for $m = \pm N$ are those at the fundamental frequency of the pulse rate, $1/T$. The average values of the "noise" components are all zero but the second order averages are not zero. Rices' calculation of these follows.

From the expressions for C_m and $\langle C_m \rangle_{\text{av}}$ in (62) and (63), it follows that

$$(C_m - \langle C_m \rangle_{\text{av}})(C_l - \langle C_l \rangle_{\text{av}})$$
$$= N^{-2} \sum_{n=1}^{N} \sum_{k=1}^{N} z^{n-1} \zeta^{k-1} [(a_n - p)\beta_m + (a_{n-1} - p)\gamma_m + (a_n a_{n-1} - p^2)\ \delta_m]$$
$$\cdot [(a_k - p)\beta_l + (a_{k-1} - p)\gamma_l + (a_k a_{k-1} - p^2)\ \delta_l] \tag{65}$$

where $\zeta = \exp(-j2\pi l/N)$. Expanding the last product and using the independence of the a_n's (except for $a_o = a_N$) shows that the ensemble average of (65) depends upon the averages

$$\langle (a_n - p)^2 \rangle_{\text{av}} = p - p^2 = pq$$
$$\langle (a_n - p)(a_n a_j - p^2) \rangle_{\text{av}} = p^2 - p^3 = p^2 q, \qquad j \neq n$$
$$\langle (a_n a_{n-1} - p^2)^2 \rangle_{\text{av}} = p^2 - p^4$$
$$\langle (a_n a_j - p^2)(a_n a_i - p^2) \rangle_{\text{av}} = p^3 - p^4 = p^3 q,$$
$$j \neq n, \qquad i \neq n, \qquad i \neq j. \tag{66}$$

For n fixed, the only values of k which lead to nonzero averages are $k = n$ and $k = n \pm 1$ with the understanding that for $n = 1$ the values $k = 0, 1, 2$ mean $k = N, 1, 2$, and for $n = N$ the values $k = N - 1, N, N + 1$ mean $k = N - 1, N, 1$. When $k = n$ the average value of the summand in (65) is

$$z^{n-1} \zeta^{n-1} [pq\beta_m \beta_l + pq\gamma_m \gamma_l + (p^2 - p^4)\ \delta_m\ \delta_l$$
$$+ p^2 q(\beta_m\ \delta_l + \gamma_m\ \delta_l + \delta_m \beta_l + \delta_m \gamma_l)]. \tag{67}$$

For $k = n + 1$ it is

$$z^{n-1} \zeta^{n} [pq\beta_m \gamma_l + p^2 q(\beta_m \delta_l + \delta_m \gamma_l) + p^3 q \delta_m \delta_l], \tag{68}$$

and for $k = n - 1$,

$$z^{n-1} \zeta^{n-2} [pq\gamma_m \beta_l + p^2 q(\gamma_{ml} + \delta_m \beta_l) + p^3 q \delta_m \delta_l]. \tag{69}$$

Some experimentation shows that the sum of (67), (68), and (69) can be written as

$$(z\zeta)^{n-1}[pq\{\beta_m y + \gamma_m y^{-1} + p\delta_m(y + y^{-1})\} \cdot \{\beta_l y^{-1} + \gamma_l y + p\delta_l(y^{-1} + y)\} + p^2q^2\delta_m\delta_l], \tag{70}$$

where

$$y = \zeta^{\frac{1}{2}} = \exp(-j\pi l/N). \tag{71}$$

The sum of $(z\zeta)^{n-1}$, taken from $n = 1$ to N, is 0 unless $z\zeta = 1$, that is, unless $M + l = 0, \pm N, \pm 2N, \ldots$. In this case

$$\begin{aligned} N^{-2} \sum_{n=1}^{N} (z\zeta)^{n-1} &= N^{-1}, \qquad \zeta = z^{-1} = \exp(+j2\pi m/N), \\ y = \exp(-j\pi l/N) &= \exp[-j\pi(l + m)/N] \exp(+j\pi m/N) \\ &= (-1)^{(l+m)/N} \exp(j\pi m/N), \end{aligned} \tag{72}$$

and (70) can be written as

$$[pq(-1)^{(l+m)/N} S_m S_l + p^2q^2\delta_m\delta_l] \tag{73}$$

where

$$S_m = \beta_m \exp(j\pi m/N) + \gamma_m \exp(-j\pi m/N) + 2p\ \delta_m \cos\frac{\pi m}{N}, \tag{74}$$

and S_l is defined similarly with l in place of m, and δ_m is the function defined in (60).

Collecting results shows that averaging (65) over the ensemble gives

$$\langle (C_m - \langle C_m \rangle_{\text{av}})(C_l - \langle C_l \rangle_{\text{av}}) \rangle = \begin{cases} [pq(-1)^{(m+l)/N} S_m S_l + p^2q^2\ \delta_m\ \delta_l]/N, \\ \qquad m + l = 0, \quad \pm N, \quad \pm 2N, \cdots \\ 0, \qquad \text{otherwise} \end{cases} \tag{75}$$

Replacing $C_l - \langle C_l \rangle_{\text{av}}$, ζ, β_l, γ_l, δ_l, y by their complex conjugates in expressions (65) to (71), and noting that the sum of $(z\zeta^*)^{n-1}$ is zero unless $z\zeta^* = 1$, that is, unless $m - l = 0, \pm N, \pm 2N, \cdots$, carries (75) into

$$\langle (C_m - \langle C_m \rangle_{\text{av}})(C_l - \langle C_l \rangle_{\text{av}})^* \rangle = \begin{cases} [pq(-1)^{(m-l)/N} S_m S_l^* + p^2 q^2 \, \delta_m \, \delta_l^*]/N, \\ \qquad m - l = 0, \quad \pm N, \quad \pm 2N, \cdots . \\ 0, \quad \text{otherwise} \end{cases} \tag{76}$$

In terms of the functions $F_j(t)$ we have

$$\delta_m = \frac{r_m}{T} \int_0^T \exp\,(-j2\pi mt/NT)[F_4 - F_3 - F_2]\, dt$$

$$S_m = \frac{r_m}{T} \int_0^T \exp\,(-j2\pi mt/NT)\Big\{[2pF_4 + (q - p)F_3 + (q - p)F_2] \cdot \cos \frac{\pi m}{N} + j[F_2 - F_3] \sin \frac{\pi m}{N}\Big\}\, dt. \tag{77}$$

The mean square value of the noise component of frequency m/NT is

$$\sigma_m^2 = 4\langle C_m C_m^* \rangle_{\text{av}} , \tag{78}$$

considering a positive frequency only spectrum. Thus from (76), we get

$$\sigma_m^2 = 4pq[S_m S_m^* + pq \; \delta_m \; \delta_m^*] \frac{1}{N}. \tag{79}$$

2.6.3 *Phase Modulation of Tuned Circuit Response*

To calculate the phase modulation on the recovered fundamental, we first obtain the response of the tank circuit to the train of rectified pulses, $I(t)$. This response is the approximate sine wave at the pulse rate frequency.

The response $I_o(t)$ is described by

$$I_o(t) = R_o(t) \cos\,[2\pi f_c t + \varphi(t)] \tag{80}$$

where the envelope $R_o(t)$ and the phase angle $\varphi(t)$ are slowly fluctuating functions whose rate of change is proportional to the bandwidth of the tank circuit. Considering now, only those components of $I(t)$ in the vicinity of the pulse rate $1/T$, we have

$$I_o(t) = 2\text{Re} \sum_{m \approx N} Y_m C_m \exp\,(j2\pi mt/NT). \tag{81}$$

Writing $C_m = (C_m - \langle C_m \rangle_{\text{av}}) + \langle C_m \rangle_{\text{av}}$ and noting that $\langle C_N \rangle_{\text{av}}$ is the only nonzero value of $\langle C_m \rangle_{\text{av}}$ in the summation, (81) may be rearranged* to be

* Y_N as used here is Y_m for $m = N$ and is not the $(Y)_N$ of Section 2.5.

$$\begin{aligned} I_o(t) &= 2\mathrm{Re}\{ Y_N\langle C_N\rangle_{\mathrm{av}} \exp(j2\pi t/T) \\ &\quad + \sum_{m\approx N} Y_m(C_m - \langle C_m\rangle_{\mathrm{av}}) \exp(j2\pi mt/T)\} \\ &= 2\mathrm{Re}\, Y_N\langle C_N\rangle_{\mathrm{av}} \exp(j2\pi t/T) \\ &\quad \cdot\left\{1 + \sum_{m\approx N} \frac{Y_m(C_m - \langle C_m\rangle_{\mathrm{av}}) \exp[j2\pi(m-N)t/NT]}{Y_N\langle C_N\rangle_{\mathrm{av}}}\right\}. \end{aligned} \tag{82}$$

The summation part of this is small, partly because σ_m/σ_N is small and partly because of the attenuation of the tank circuit Y_m except for frequencies near the pulse rate $1/T$. Thus if the part within the brackets is represented by $1 + A + jB$, $| A + jB |$ is small compared with unity and so

$$1 + A + jB \approx \exp[A + jB] \tag{83}$$

and

$$\begin{aligned} I_o(t) &\approx 2\mathrm{Re}\, Y_N\langle C_N\rangle_{\mathrm{av}} \exp[j(2\pi t/T) + A + jB] \\ &\approx 2 \,| Y_N\langle C_N\rangle_{\mathrm{av}} \exp(A) |\cos[2\pi t/T) + B + \arg(Y_N\langle C_N\rangle_{\mathrm{av}})]. \end{aligned} \tag{84}$$

This is the approximate sine wave at the pulse rate frequency which has been recovered from the pulse train by means of the narrowband tank circuit. Its phase modulation is

$$\begin{aligned} \varphi(t) &= \arg(Y_N\langle C_N\rangle_{\mathrm{av}}) + B \\ &= \arg(Y_N\langle C_N\rangle_{\mathrm{av}}) + \mathrm{Im} \sum_{m\approx N} b_m \exp[j2\pi(m-N)t/NT] \end{aligned} \tag{85}$$

where

$$b_m = Y_m(C_m - \langle C_m\rangle_{\mathrm{av}})/(Y_N\langle C_N\rangle_{\mathrm{av}}). \tag{86}$$

The Fourier component of $\varphi(t)$ at frequency k/NT is determined by the two side frequencies $mf_c/N = (N + k)/NT$ and $mf_o/N = (N - k)/NT$ about the pulse rate as was found in the earlier analysis. To calculate this, we use the two terms for which $m = N \pm k$ in the above sum. That is

$$\begin{aligned} \varphi_k &= \mathrm{Im}[b_{N+k} \exp(j2\pi kt/NT) + b_{N-k} \exp(-j2\pi kt/NT)] \\ &= \mathrm{Im}[(b_{N+k} - b^*_{N-k}) \exp(j2\pi kt/NT)]. \end{aligned} \tag{87}$$

The time average of the "phase power" in this component is

$$\langle\varphi_k^2\rangle_{\mathrm{av}t} = \tfrac{1}{2}\,| b_{N+k} - b^*_{N-k} |^2 \quad \text{radians}^2 \tag{88}$$

which can be written

$$\langle\varphi_k^2\rangle_{\mathrm{av}t} = \tfrac{1}{2}[b_+b_+^* - b_+b_- - b_-^*b_+^* + b_-^*b_-] \tag{89}$$

where the subscripts + and − denote $N \pm k$.

The ensemble average of the expression (89) may be computed with the help of the second order moments of the C_m's given by (75) and (76). For example

$$\begin{aligned}\langle b_+b_+^*\rangle_{\mathrm{av}} &= Y_+Y_+^*(C_{N+k} - \langle C_{N+k}\rangle_{\mathrm{av}})(C_{N+k} - \langle C_{N+k}\rangle_{\mathrm{av}})^*\rangle / \mid Y_N\langle C_N\rangle_{\mathrm{av}}\mid^2 \\ &= Y_+Y_+^*(pqS_+S_+^* + p^2q^2\ \delta_+\ \delta_+^*)/\mid Y_N\langle C_N\rangle_{\mathrm{av}}\mid^2 N \\ &= [pqU_+U_+^* + p^2q^2V_+V_+^*]/N\end{aligned} \tag{90}$$

where

$$U_+ = (Y_+/Y_N)(S_+/\langle C_N\rangle_{\mathrm{av}}) \quad V_+ = (Y_+/Y_N)(\delta_+/\langle C_N\rangle_{\mathrm{av}}) \tag{91}$$

and the subscripts + and − are used to indicate $m_\pm = N \pm k$. Also $m - l = (N + k) - (N + k) = 0$ so that the (-1) to the power $(m - l)/N$ appearing in (26) is $+1$. Similarly, in the calculation of $\langle b_+b_-\rangle$, $m + l = (N + k) + (N - k) = 2N$ and the (-1) to the power $(m + l)/N$ appearing in (75) is again $+1$, and so on.

After the ensemble average of (90) has been calculated, taking each of the four parts of (90) in turn, the terms may be combined to give one of the results we have been seeking:

$$\langle\varphi_k^2\rangle_{\mathrm{av}} = \frac{1}{2}\frac{pq}{N}\{\mid U_+ - U_-^*\mid^2 + pq \mid V_+ - V_-^*\mid^2\} \quad \text{radians}^2. \tag{92}$$

Since the Fourier components are f_c/N apart, this expression for φ_k^2 is equivalent to the "phase power" in a band f_c/N wide. The averaging has been done over both time and the ensemble.

2.6.4. *Effect of Sampling and of High Frequencies in the Response*

Next we consider a generalization of the expression (92) as derived above for the phase modulation on the fundamental pulse rate obtained by passing the random pulse train through a tuned circuit.

In deriving the result (92) for the phase modulation of the timing wave obtained from the tuned circuit response, only that part of the pulse train spectrum in the vicinity of the pulse rate was considered.

While the major part of the tuned circuit response lies in the frequency region near the pulse rate $1/T$, as assumed in the calculation beginning with (81) and ending with (92), the way in which this response is used in a regenerative repeater (or is measured in a phase detector) gives some importance to the part of the response neglected in (81).

In a regenerative repeater, retiming of the message pulses is done by very sharp pulses which are generated at each positive (or negative) going zero crossing of the timing wave. In the phase detector used in the experiments described here, the deviations of the zero crossings from their ideal periodic nature are measured and used as a sequence of numbers or held and smoothed by low pass filter to approximate the phase deviation function $\varphi(t)$. Thus in both these processes, it is the samples of the derived timing wave which are used. When the deviations are not too large, the magnitudes of the zero crossing deviations are equivalent to the magnitudes of samples of the phase deviation wave $\varphi(t)$, taken at the pulse rate. Since the high frequency part of the tuned circuit response which was neglected in (81) lies above one-half the sampling rate, the process described above as equivalent to sampling, may convert some of this high frequency part into very low frequency energy in the final timing noise result. In particular, if the high frequency part of the pulse spectrum has energy at or very near the harmonics of the pulse rate, this will be converted to zero or very low frequency energy in the phase noise spectrum. The analysis by S. O. Rice deals with this situation also.

First the expression (81) for $I_o(t)$ must be enlarged to include a dc term and all values of m from 1 to ∞. This may be rewritten in the following form which corresponds to (83). We have

$$\begin{aligned} I_o(t) = 2 \text{ Re } Y_N C_N \exp\,(j2\pi t/T)\Big\{1 &+ \tfrac{1}{2}\, d_o \exp\,(-j2\pi t/T) \\ &+ \sum_{l=2}^{\infty} d_l \exp\,[j2\pi(l-1)t/T] + \tfrac{1}{2} b_o \exp\,(-j2\pi t/T) \\ &+ \sum_{m=1}^{\infty} b_m \exp\,[j2\pi(m-N)t/NT]\Big\} \end{aligned} \tag{93}$$

where l is an integer,

$$d_l = Y_{Nl}\langle C_{Nl}\rangle_{\text{av}}/Y_N\langle C_N\rangle_{\text{av}}$$

and the coefficient b_m is defined as in (86). This expression is the complete response of the tuned circuit to the random train of pulses.

Again assuming that the term within the square brackets in (93) always remains near 1,

$$\begin{aligned}\varphi(t) \approx{} & \arg Y_N \langle C_N \rangle_{\text{av}} \\ & + \operatorname{Im}\left\{\tfrac{1}{2}d_o \exp(-j2\pi t/T) + \sum_{l=2}^{\infty} d_l \exp[j2\pi(l-1)t/T]\right\} \\ & + \operatorname{Im}\left\{\tfrac{1}{2}b_o \exp(-j2\pi t/T) + \sum_{m=1}^{\infty} b_m \exp[j2\pi(m-N)t/NT]\right\}. \qquad (94)\end{aligned}$$

This is an improved version of (85). The first line on the right is approximately the ensemble average $\langle\varphi(t)\rangle_{\text{av}}$. It consists of a dc term and harmonics of the pulse repetition frequency $1/T$. While it may appear that the assumption about the bracketed part of (94) being small is unjustified because this represents the adding of harmonics to yield the pulse wave form, this is not the case since all the harmonics included are reduced at least by Q. Furthermore, since the sampling operation occurs near the zeros of the response, where the harmonics are zero or very small, an additional reduction of magnitude is involved. The noise portion of the power spectrum of $\varphi(t)$ arises from the second line, which may be written as

$$\begin{aligned}\varphi(t) - \langle\varphi(t)\rangle_{\text{av}} \\ \approx \operatorname{Im}\left[\tfrac{1}{2}b_o \exp(-j2\pi t/T) + \sum_{k=1-N}^{\infty} b_{N+k} \exp(j2\pi kt/TN)\right]. \qquad (95)\end{aligned}$$

The sampling operation mentioned above, which is performed on the phase function $\varphi(t)$ in both the phase detector used in the experiments reported and in regenerative PCM repeaters, generates a new phase function $\theta(t)$. This is

$$\theta(t) \approx \varphi(t)T \sum_{n=-\infty}^{\infty} \delta(t - nT) = \varphi(t) \sum_{n=-\infty}^{\infty} \exp(j2\pi nt/T) \qquad (96)$$

where $\delta(t)$ denotes the unit impulse function. For some frequencies in $\varphi(t)$, the extraneous modulation products introduced by the impulses may be negligible. However, for the higher frequencies and for a single tuned circuit with slowly decreasing $Y(i\omega)$, some of the products may become appreciable and should be taken into account. The sampling times in (96) were arbitrarily set at $t = 0, T, 2T \cdots$ for convenience in the analysis to follow. These can be varied to occur at or near the zeros of the response wave $I_o(t)$ by shifting the time scale of the description of the pulse train using the parameter ν as indicated by (54) and related equations as illustrated in Fig. 19. Forming the ensemble average

$\langle\theta(t)\rangle_{\text{av}}$, subtracting it from $\theta(t)$, and noticing that the sums in the expression (96) for $\theta(t)$ are real, leads to

$$\begin{aligned}\theta(t) - \langle\theta(t)\rangle_{\text{av}} = \operatorname{Im}\Bigg\{&\tfrac{1}{2}b_o \sum_{n=-\infty}^{\infty} \exp(j2\pi nt/T) \\ &+ \sum_{n=-\infty}^{\infty}\sum_{k=1-N}^{\infty} b_{N+k} \exp[j2\pi(k+nN)t/NT]\Bigg\}. \qquad (97)\end{aligned}$$

The component of $\theta(t)$ of frequency l/NT for $1 \leqq l \leqq N-1$, that is, for frequencies which lie between 0 and f_o, is the sum of terms having exponential factors $\exp(\pm j2\pi lt/NT)$. For $k + nN = +l$, the values of k and n are

$$\begin{array}{llll} k = -N+l, & k = l, & k = N+l, & k = 2N+l, \cdots, \\ n = 1, & n = 0, & n = -1, & n = -2, \cdots, \end{array}$$

and for $k + nN = -l$ they are

$$\begin{array}{lll} k = -l, & k = N-l, & k = 2N-l, \cdots, \\ n = 0, & n = -1, & n = -2, \cdots, \end{array}$$

Therefore the component of $\theta(t)$ of frequency l/NT is

$$\begin{aligned}&\operatorname{Im}[(b_l + b_{N+l} + b_{2N+l} + \cdots)\exp(j2\pi lt/NT) \\ &\qquad + (b_{N-l} + b_{2N-l} + b_{3N-l} + \cdots)\exp(-j2\pi l/NT] \\ &\quad = \operatorname{Im}[\{(b_l + b_{N+l} + b_{2N+l} + \cdots) - (b_{N-l} + b_{2N-l} + \cdots)^*\} \\ &\qquad \cdot\exp(j2\pi lt/NT). \qquad (98)\end{aligned}$$

When b_{N+l} and b_{N-l} are the dominant terms in (98), comparison with expression (87) (with $k = l$) shows that the component of $\theta(t)$ of frequency l/NT is nearly equal to the corresponding component in $\varphi(t)$. By means of the procedure used earlier in (91) the average power in the component of $\theta(t)$ of frequency k/NT is obtained where we have returned to k from the l in (98). First, the time average of the power in the component is

$$\begin{aligned}\langle\theta_k^2\rangle_{\text{av }t} &= \frac{1}{2}\left|\sum_{n=0}^{\infty} b_{nN+k} - \sum_{n=1}^{\infty} b_{nN-k}^*\right|^2 \\ &= \frac{1}{2}\left[\sum_{n=0}^{\infty} b_{nN+k} - \sum_{n=1}^{\infty} b_{nN-k}^*\right]\left[\sum_{n=0}^{\infty} b_{nN+k}^* - \sum_{n=1}^{\infty} b_{nN-k}\right]\end{aligned} \qquad (99)$$

where $0 < k < n$. To average (99) over the ensemble we make use of (75) and (76) slightly rewritten by substituting $\exp(-j\pi m/N)$ for $(-1)^{m/N}$ and including it with S_m. Thus

$$\langle (C_m - \langle C_m \rangle_{\text{av}})(C_l - \langle C_l \rangle_{\text{av}}) \rangle$$
$$= \frac{pq}{N} [S_m \exp(-j\pi m/N)][S_l \exp(-j\pi l/N)] + \frac{p^2q^2}{N} \delta_m \, \delta_l \, ,$$
$$m + l = 0, \pm N, \pm 2N, \cdots , \qquad (100)$$

$$\langle (C_m - \langle C_m \rangle_{\text{av}})(C_l - \langle C_l \rangle_{\text{av}})^* \rangle_{\text{av}}$$
$$= \frac{pq}{N} [S_m \exp(-j\pi m/N)][S_l \exp(-j\pi l/N)]^* + \frac{p^2q^2}{N} \delta_m \, \delta_l^*$$
$$m - l = 0, \pm N, \pm 2N, \cdots . \qquad (101)$$

The averages are equal to zero unless m and l satisfy the respective conditions.

A typical term encountered in the averaging of (99) is

$$\langle b_{nN+k} b^*_{n'N+k} \rangle_{\text{av}} = \langle b_m b^*_l \rangle_{\text{av}}$$

where $m = nN + k$, $l = n'N + k$, and $m - l = (n - n')N$. From (101)

$$\langle b_m b^*_l \rangle_{\text{av}} = \frac{pq}{N} [U_m \exp(-j\pi m/N)][U_l \exp(-j\pi l/N)]^* + \frac{p^2q^2}{N} V_m V^*_l \, ,$$
$$m - l = (n - n')N.$$

In this U and V are generalized from the definitions of (91), so that for example

$$U_{nN+k} = \frac{S_{nN+k}}{\langle C_N \rangle_{\text{av}}} \frac{Y_{nN+k}}{Y_N}. \qquad (102)$$

Considering all four forms of $\langle b_m b \rangle_{\text{av}}$ as before, the ensemble average of (99) is found to be

$$\langle \theta_k^2 \rangle_{\text{av}} = \frac{pq}{2N} \left| \sum_{n=0}^{\infty} U_{nN+k} \exp[-j\pi(nN + k)/N] - \sum_{n=1}^{\infty} U^*_{nN-k} \exp[+j\pi(nN - k)/N] \right|^2$$
$$+ \frac{p^2q^2}{2N} \left| \sum_{n=0}^{\infty} V_{nN+k} - \sum_{n=1}^{\infty} V^*_{nN-k} \right|^2 (\text{radians})^2. \qquad (103)$$

The $\exp(-j\pi k/N)$ can be factored from the first absolute value and then $\exp(\pm j\pi nN/N)$ replaced by $(-1)^n$, so that

$$\langle\theta_k^2\rangle_{\text{av}} = \frac{pq}{2N}\left|\sum_{n=0}^{\infty}(-)^n U_{nN+k} - \sum_{n=1}^{\infty}(-)^n U_{nN-k}^*\right|^2 + \frac{p^2q^2}{2N}\left|\sum_{n=0}^{\infty} V_{nN+k} - \sum_{n=1}^{\infty} V_{nN-k}^*\right|^2 \text{(radians)}^2, \tag{104}$$

which is the generalization of (92) when the complete response of the tuned circuit is used. To repeat what was said before, since the Fourier components in the spectrum of θ are f_c/N apart, this expression for $\langle\theta_k^2\rangle_{\text{av}}$ at $f = k/NT$ is equivalent to the "phase power" in a band f_c/N wide in a continuous spectrum.

To obtain the continuous power spectrum $w_\theta(f)$ of the "noise" part of θ, we let $m/NT = f'$ and $N \to \infty$ in the above expressions. First (77) becomes

$$\delta(f') = \frac{\exp(j2\pi\nu f'/f_c)}{T}\int_0^T \exp(-j2\pi f't)[F_4 - F_3 - F_2]\,dt$$

$$S(f') = \frac{\exp(j2\pi\nu f'/f_c)}{T}\int_0^T \exp(-j2\pi f't)\{[2pF_4 + (q-p)F_3 + (q-p)F_2]\cos\pi f'T + j[F_2 - F_3]\sin\pi f'T\}\,dt. \tag{105}$$

Also

$$U_{nN+k} \to U(f') = U(nf_c + f) = \frac{Y[j2\pi(nf_c+f)]}{Y(j2\pi f_c)}\,\frac{S(nf_c+f)}{p[S(f_c) + p\,\delta(f_c)]} \tag{106}$$

where $\langle C_N\rangle_{\text{av}}$ has been replaced by $p(S_N + p\delta_N)$ as derived from (64) and (74).

Since the expressions derived earlier for the average power in a component of $\varphi(t)$ refer to positive frequency only, we shall deal with the one-sided power spectrum $w_\theta(f)$ of $\theta(t)$. As before $f = k/NT$ denotes the frequency associated with the average power expressed in (104). Then the value of the right side of (104) tends to $w(f)\Delta f = w_\theta(f)/NT$ as $N \to \infty$ and consequently

$$w_\theta(f) \approx \frac{Tpq}{2}\left|\sum_{n=0}^{\infty}(-)^n U(nf_c+f) - \sum_{n=1}^{\infty}(-)^n U^*(nf_c-f)\right|^2 + \frac{Tp^2q^2}{2}\left|\sum_{n=0}^{\infty} V(nf_c+f) - \sum_{n=1}^{\infty} V^*(nf_c-f)\right|^2 \text{(radians)}^2 \text{ per Hz}, \tag{107}$$

where $0 < f < f_c$.

2.6.5 *Application of Results in Section 2.6.4 to Particular Pulse Shapes*

Comparing the result (104) with the earlier (92), it is seen that each component of frequency $f = k/NT$ in the spectrum of θ is made up of contributions from the pairs of components spaced f from all the harmonics of the pulse rate f_c. The sampling of $\phi(t)$ has brought in all these additional contributions.

Now we are in a position to find out what the effect is of neglecting the higher frequencies as was done in the previous work and especially in deriving the result (92) to which (104) reduces when only $n = 1$ is considered.

To do this, we will investigate several particular pulse shapes. As will be seen below, certain simplifications arise for $n = 1$ so that in some cases, approximate expressions for the noise may be derived. But, in general, and especially when the sums of (107) are to be calculated, the expressions become so complex that they cannot be dealt with readily in an analytical way. However, numerical calculations of $w_\theta(f)$ in (107) answers most of our questions. Some of the computations for the rectangular pulses were first done by S. O. Rice. The others are extensions of them. The sums were carried to 30 terms for the rectangular pulses and to 15 or 20 in the other cases. Leveling off occurred before these cutoff points were reached.

If each pulse is confined to one time slot, then $I_n(t)$ is determined entirely by a_n. Thus, in the specifications (54), we have

$$F_3(t') = 0; \qquad F_2(t') = F_4(t'). \tag{108}$$

As a consequence, it is seen from (59), (60), and (74) that

$$\delta_m = 0, \qquad \gamma_m = 0 \tag{109}$$

$$S_m = \beta_m \exp(j\pi m/N) \tag{110}$$

with corresponding simplifications in (75), (76), (79), and (92) and (104).

2.6.5.1 *Narrow Rectangular Pulse.* First take the rectangular pulse of duration τ used in Fig. 19. Here

$$\begin{aligned} F_2(t') = F_4(t') &= 1, \qquad (T - \tau)/2 < t' < (T + \tau)/2 \\ &= 0, \qquad \text{elsewhere.} \end{aligned} \tag{111}$$

Calculation of the pulse spectrum function S_m from (110), (58), and (61) yields

$$\frac{S_m}{S_N} = (-)^{n-1} \exp\left[j\pi x(1 - 2\nu)\right] \frac{\exp\left[-j\pi 2\nu(n - 1)\right]}{(n + x)} \frac{\sin \pi(n + x)\tau/T}{\sin \pi\tau/T} \tag{112}$$

using $m = nN + k$ and $x = k/N$.

The expression (3) in Section 2.1 for the ratio of noise power in a band B to carrier power is derived from (112) using (79) and (64), putting f'/f_c for $n + x$, and noting that $\sigma_N = 2pS_N$ for a positive frequency only spectrum as in (79). Thus

$$\begin{aligned}\frac{\sigma_m^2}{\sigma_N^2} &= \frac{q}{pN} \frac{S_m S_m^*}{S_N^2} \\ &= \frac{f_c^2}{N} \frac{\sin^2 \pi f'\tau}{(f')^2 \sin^2 \pi\tau/T}.\end{aligned} \tag{113}$$

These components are f_c/N apart and so the power of each corresponds to that in a band B of this width. The ratio σ_m^2/σ_N^2 corresponds exactly to S^2/A_i^2 in (3).

To apply these results to the single tuned tank circuit centered on the pulse rate, the general form of Y_{m+}/Y_N as given in (15) is necessary. This is

$$Y_{nN+k}/Y_N = \frac{n + x}{n + x + jQ[(n + x)^2 - 1]}. \tag{15}$$

Combining (15) with (112) in (102) or (106) gives

$$U(nf_c + f) = \frac{(-)^{n-1} \exp\left[j\pi x(1 - 2\nu)\right] \exp\left[-j\pi 2\nu(n - 1)\right]}{p[n + x + jQ[(n + x)^2 - 1]]} \frac{\sin \pi(n + x)\tau/T}{\sin \pi\tau/T} \tag{114}$$

with $f' = nf_c + f$ and $m/N = f'/f_c$.

First, notice that the factor $\exp(j2\pi x\nu)$ in (114) disappears in the calculation of (107) since $\exp(j\pi x 2\nu) = \exp{}^*\,(j\pi(-x)2\nu)$ and hence factors out of both sums.

Next consider the unique conditions that arise when $n = 1$. The second exponential factor containing the parameter ν, which determines the sampling time, disappears from (114). And the factor multiplying jQ in the denominator becomes proportional to x. The first means that the sampling time has no effect on the contribution of the $n = 1$ terms to $w_\theta(f)$. The second means that $(U_+ - U_-^*) \rightarrow 0$ as $x \rightarrow 0$ and hence that the contribution of $n = 1$ to $w_\theta(f)$ approaches zero as $f \rightarrow 0$.

As indicated in Fig. 19, the zero crossing of the fundamental component comes at the $t = 0$ sampling time for $\nu = \frac{1}{4}$; hence this value was used in the computations. Timing noise spectra obtained in this way for several durations of rectangular pulse are plotted in Fig. 20. In order that the effect of the high frequencies may be easily seen, the corresponding spectra considering only the $n = 1$ term are shown in Fig. 21.

Several interesting points are evident:

(*i*) Comparing the two sets of spectra, it is seen that the high frequencies have brought in noise at and near zero frequency, except for the $\tau/T = 0.6$ duration.*

(*ii*) There is a great difference in the full spectra for $\tau/T > 0.5$. This was observed experimentally for other values of τ/T than 0.6.

(*iii*) The high frequencies in the timing tank response, through the sampling process, have not only brought in very low frequency phase noise, but reduced the higher frequency noise except for $\tau/T = 0.6$. It is the phase structure of this noise which makes cancellation as well as addition possible.

(*iv*) In the case of $\tau/T = 0.6$, there is only a small difference between the spectrum obtained with the full spectrum and that from considering only $n = 1$.

(*v*) There is very little difference between the spectra for $\tau/T = 0.1$ and $\tau/T = 0.02$. This suggests that the phase noise may not disappear for pulses which approach spikes in shape.

(*vi*) The magnitude of the noise, when compared with that of the usually more practical rounded shapes, is quite small, as seen in Fig. 4. It will be seen below, that the effect of the high frequencies is much less for these other pulse shapes.

For the narrow pulse case, $\tau/T = 0.1$, the timing noise spectrum can be changed greatly by small amounts of tank circuit mistuning, as shown in Fig. 22. This also can be a cancellation or an addition effect.

When the tank circuit is mistuned from the pulse rate, $\pm x$ is replaced by $x_o \pm x$ in (15) for the normalized admittance. Here x_o is the relative mistuning as defined earlier in (18).

2.6.5.2 *1T Raised Cosine Pulse.* A photograph of an oscilloscope display of this pulse shape is given in Fig. 18c. Each pulse is confined to

* The possibility of this property of the noise was pointed out by H. E. Rowe. See Ref. 6.

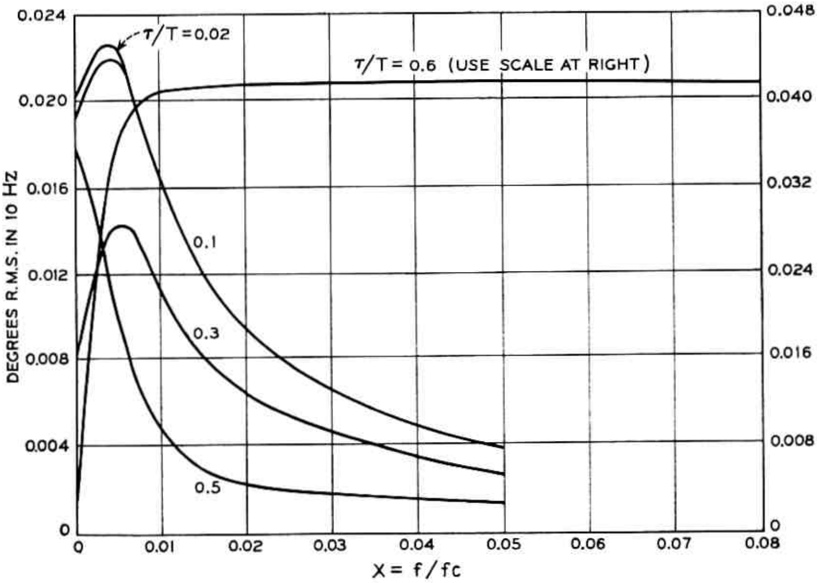


Fig. 20 — Calculated spectra of timing noise for several durations of rectangular pulses.

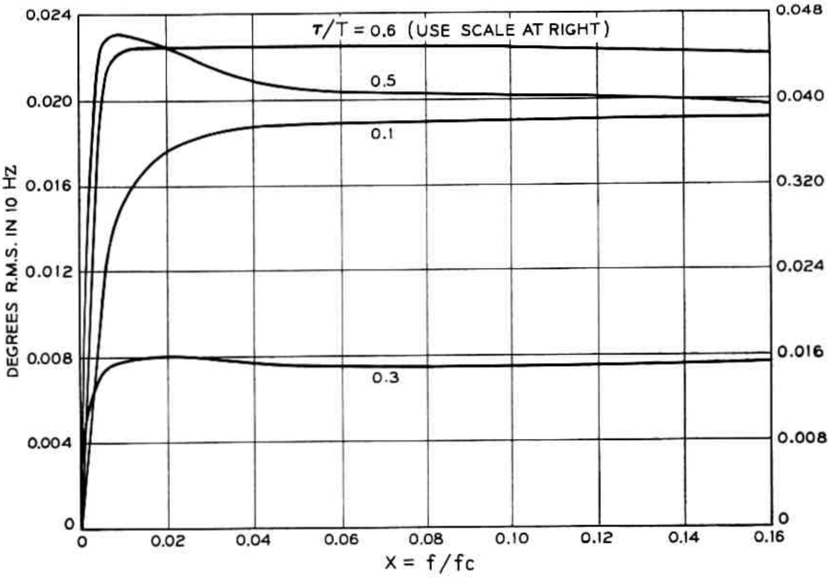


Fig. 21 — Calculated spectra of timing noise for rectangular pulses ($n = 1$ only).

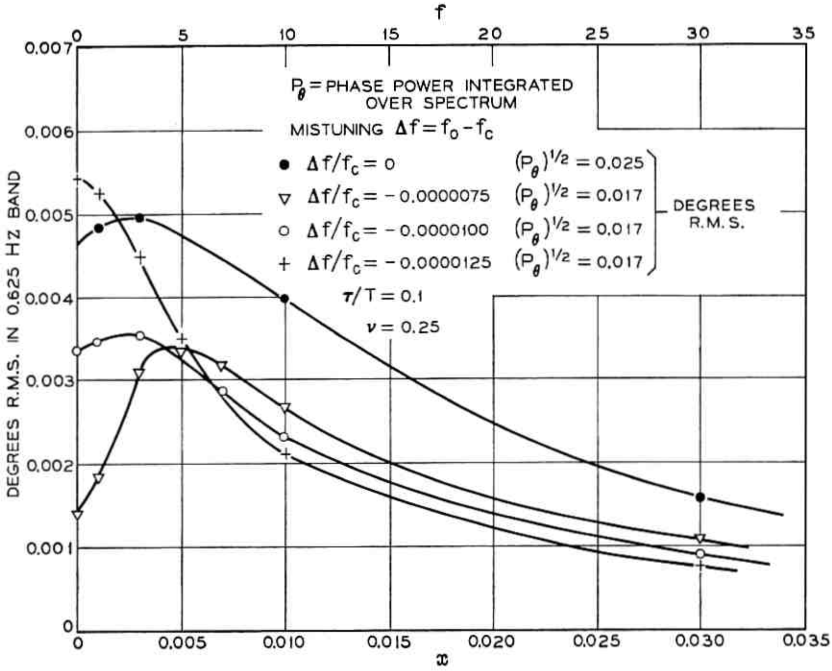


Fig. 22 — Spectrum of phase noise for rectangular pulse.

a single time slot and so the relations (108), (109), (110) hold. Here

$$F_2(t') = F_4(t') = 1 - \cos(2\pi t/T) \tag{115}$$

and

$$\frac{S(nf_c + f)}{S(f_c)} = -\frac{2}{\pi}\,\frac{\exp(j2\pi\nu x)\,\exp[j2\pi\nu(n-1)]\sin\pi(n+x)}{(n+x)[(n+x)^2-1]}. \tag{116}$$

The proper value of ν is 1/4.

The spectrum of timing noise is shown in Fig. 35, where it is seen that there is no noise energy at zero frequency. The pulse spectrum not only has nulls at all the pulse rate harmonics except the fundamental, but has very small energy in all the high frequencies. Thus, in this case, the high frequencies have very little effect on the timing noise. This was verified by both further calculation and measurement.

For x not too large, the noise spectrum is nearly

$$(w_\theta(f))^{\frac{1}{2}} \approx (T/2)^{\frac{1}{2}}\,\frac{\sin\pi x}{\pi}\,\frac{(9+16x^2Q^2)^{\frac{1}{2}}}{1+4x^2Q^2}\ \text{rad/Hz}. \tag{117}$$

2.6.5.3 *1.5T Raised Cosine Pulse.* This pulse shape is drawn in Fig. 23a. A single pulse is described by

$$f(t) = 1 - \cos(4\pi t/3T). \tag{118}$$

Consideration of the specification (54) shows that

$$\left.\begin{aligned} F_2(t') - F_3(t') &= f(t') - f(t' + T) \\ F_4(t') &= f(t') + f(t' + T) \end{aligned}\right\} \quad 0 < t' < T/2$$
$$\left.\begin{aligned} F_2(t') - F_3(t') &= f(t') \\ F_4(t') &= f(t') \end{aligned}\right\} \quad T/2 < t' < T. \tag{119}$$

Thus, even though there is overlapping of adjacent pulses, $\delta_m = 0$ and the pulse spectrum is completely specified by S_m. Consideration of the waveform of Fig. 23(a) for all pulses present shows that $\nu = 1/2$ is the proper value here. S_m/S_N is independent of ν for this value and is

$$\frac{S_m}{S_N} = \exp(j\pi x)(-)^n \frac{5}{8} \frac{[1 - \exp(-j\pi m/N)]}{(m/N)(4/9 - m^2/N^2)} \cdot \left\{\frac{m^2}{N^2}[1 + \exp(-j\pi m/N)] + \frac{4}{9}\exp(j\pi m/N)\right\}. \tag{120}$$

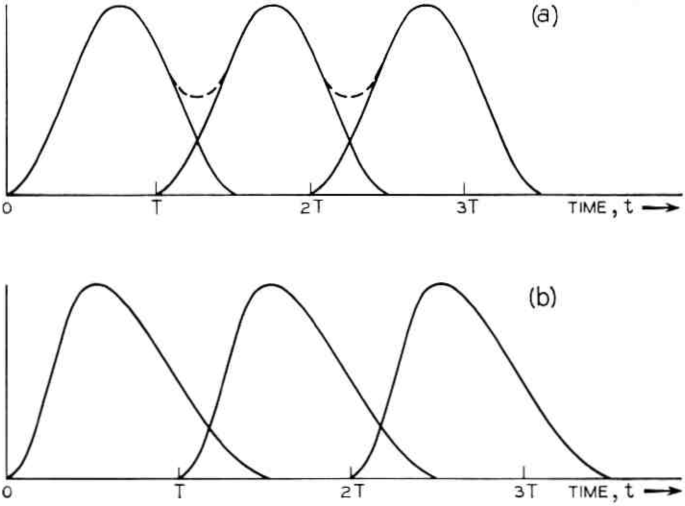


Fig. 23 — Pulse waveforms.

The spectra of timing noise, caused by this pulse shape, calculated and measured, are plotted in Fig. 36. While the magnitude (rms) of this noise is about 4 times that for $1T$ pulses, the energy at zero frequency is 0.00075 degree in a 10 Hz band, which is quite small. Hence overlapping of pulses does not appear to be a source of very low frequency timing noise. As in the previous case, the high frequency part of the response is small and so has only a small effect on the timing noise.

2.6.5.4 *Asymmetrical Overlapping Pulses.* The particular pulse shape chosen here is pictured in Fig. 23b, where it is seen that the rise occurs in one-half pulse period while the decline takes a whole period. The auxiliary function $F_2(t)$ is described by

$$\begin{aligned} F_2(t') &= 1 - \cos 2\pi t'/T \qquad && 0 < t' < T/2 \\ &= 1 - \cos \pi(t' + T/2)/T \qquad && T/2 < t' < 3T/2. \end{aligned} \tag{121}$$

From this and F_3 and F_4, it is found that the δ function is zero as in the previous case. The spectrum function S_m is

$$\begin{aligned} \frac{S_m}{S_N} = \frac{2j}{\pi} & \frac{\exp(j2\pi x\nu)\exp[j2\pi\nu(n-1)]}{(1 - j4/3\pi)(n+x)} \\ & \cdot \left\{ \frac{1 + \exp[-j2\pi(n+x)]}{1 - 4(n+x)^2} - \frac{1 + \exp[j\pi(n+x)]}{1 - (n+x)^2} \right\}. \end{aligned} \tag{122}$$

Estimates indicate that $\nu = 0.35$ will bring the zero crossing of the fundamental term very close to the sampling time of $t = 0$ and so this value was used in the calculations. The spectrum of timing noise caused by this pulse shape is plotted in Fig. 36, where it is seen that the amount at zero frequency is quite small, although somewhat larger than that of the symmetrical overlapping pulses. Hence, asymmetry of pulse shape does not appear to be a significant factor in very low frequency timing noise.

Since no readily available means for generating this pulse shape in the laboratory was found, there is no measured data. The good agreement between calculation and measurement in the other cases gives considerable weight to the calculated curve.

2.6.5.5 *Rectified 2T Raised Cosine Pulses.* Rectified $2T$ raised cosine pulses are pictured in Figs. 17 and 18. Before rectification, a single pulse occupying two time slots is represented by

$$f_2(t) = 1 + \cos(\omega_c/2)t \qquad -T < t < T \tag{123}$$

where $f_c = \omega_c/2\pi$ is the pulse rate. After rectification, the auxiliary functions are shown in Fig. 17c and described by

$$\begin{aligned}
F_2(t') &= 0, & 0 \leqq t' \leqq T/2 \\
&= -\cos(\pi t'/T), & T/2 \leqq t' \leqq T \\
F_3(t') &= \cos(\pi t'/T), & 0 \leqq t' \leqq T/2 \\
&= 0, & T/2 \leqq t' \leqq T \\
F_4(t') &= 1, & 0 \leqq t' \leqq T.
\end{aligned} \tag{124}$$

Because of the nonlinearity, the δ function is not zero. The pulse spectrum is described by

$$\begin{aligned}
S_m &= \frac{1}{\pi}\,\frac{\exp[-j(3\pi/2)(m/N)]\cos(\pi m/N)\sin(\pi m/N)}{(m/N)(1 - 4m^2/N^2)} \\
\delta_m &= \frac{1}{\pi}\,\frac{\exp[-j(3\pi/2)(m/N)][\sin(\pi m/N) - 2m/N]}{(m/N)(1 - 4m^2/N^2)}
\end{aligned} \tag{125}$$

$$S_N = 0, \quad \delta_N = j2/3\pi, \quad \langle C_N\rangle_{\text{av}} = j/6\pi \ \text{ for } \nu = 1/4, \quad p = 1/2.$$

The spectrum of timing noise calculated from (125) is plotted in Fig. 37. The sums in (107) were carried to 15 terms, but the higher order terms added very little. The results are very close to those for $n = 1$ except at zero frequency. The higher frequency terms with aliasing do generate some noise at these. The amount, which is difficult to see in the Fig. 37, is 0.0023 degree rms in a 10 Hz band, about three times that for $1.5T$ pulses and one-half that for the asymmetrical pulses.

2.6.6 *Relation of the General Theory to the Simpler One*

The operation $U_+ - U_{\pm}^*$ which appears in (92) for the calculation of phase deviation is essentially the same as the operation used in Section 2.2 for determining symmetrical and antisymmetrical components or in phase and quadrature components. The expression $U_+ - U_{\pm}^*$ is

$$U_+ - U_{\pm}^* = (S_+/\langle C_N\rangle_{\text{av}})(Y_+/Y_N) - (S_-/\langle C_N\rangle_{\text{av}})^*(Y_-/Y_N)^*. \tag{126}$$

In the simpler derivations of Section 2.2, it was assumed that the pulse train was such that $(S_+/\langle C_N\rangle_{\text{av}}) = (S_-/\langle C_N\rangle_{\text{av}})^*$ and so could be factored out leaving

$$U_+ - U_-^* = (S/\langle C_N\rangle_{\text{av}})[Y_+/Y_N - (Y_-/Y_N)^*]. \tag{127}$$

Or, in other words, the pulse train contributed to the phase deviation in magnitude only. The part within the brackets is the conversion factor caused by mistuning, if any, in the tuned circuit described by Y. In the case of the offset trigger, a separate conversion factor was derived.

The simple theory cannot be applied generally for the wider pulses and particularly in those cases when the δ function enters into the pulse spectrum description. In some cases considered in detail, it was found that the strength of noise components σ_m (and hence the dissymmetry between side frequency pairs about the pulse rate) depends largely on the δ_m part of (79), and very little on the S_m part, while the situation is just the reverse for the phase deviation ϕ_k in (92).

III. SYSTEM USED FOR MEASUREMENTS

3.1 *Principal Apparatus*

In Fig. 24, the connection diagram of Fig. 1 has been revised to show the detection and remodulation process. This also shows why the actual apparatus used (bottom diagram) is really parts of two repeaters.

The complete block diagram of the apparatus used, corresponding to the simplified diagram at the bottom of Fig. 24 is shown in Fig. 25. The functioning of this apparatus will now be described in more detail. The principal sections are:

3.1.1 *Pulse Regenerator*

The pulse regenerator has been especially developed so that it will not add any timing noise of its own. It was worked out mainly by C. R. Crue following plans made by S. L. Freeny. To make the first record (simulating repeater 1), the regenerator input is switched to the source of clipped random noise, thus generating a random train of pulses. Thereafter it is switched to the recorder playback so that the same sequence of pulses, though random, is used for each transmission through the apparatus. A fixed bias may also be connected to the regenerator input so that it sends an all pulses present train to the system for calibration and testing.

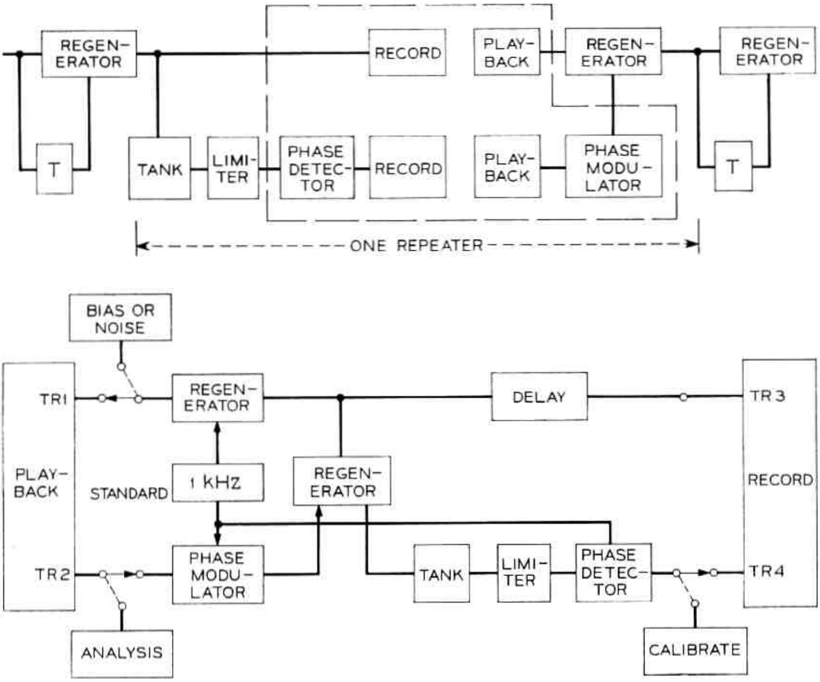


Fig. 24 — Block diagram of simulation of chain of regenerative repeaters. Part within dashed section at top is redrawn at bottom.

3.1.2 *LC Tank Circuit*

The LC tank circuit which derives the timing wave from the incoming pulse train uses an air core coil and positive feedback to achieve a Q of 100. It is necessary to use an air core coil and to keep it in a temperature controlled oven in order to measure phase with the required precision of less than 0.1°.

3.1.3 *Amplitude Limiter*

A very important section and one which is difficult to achieve is the amplitude limiter which removes very nearly all the amplitude variation from the timing wave so that in the detection process, only the phase deviation of the timing wave is measured. For most of the measurements, the limiter has two stages, each consisting of amplifier, cathode follower, and series limiter made up of resistance and a pair

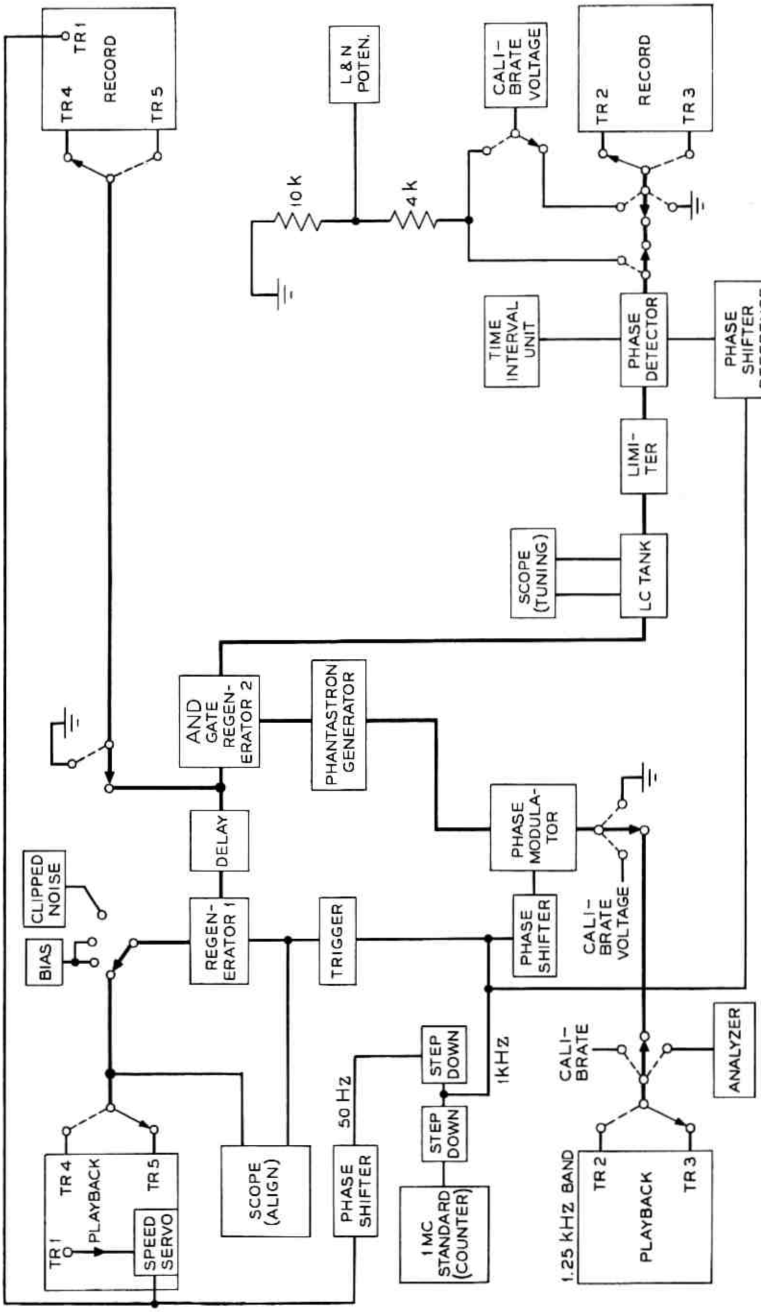


Fig. 25 — Detailed block diagram of system used for simulation of chain of regenerative repeaters. Heavy lines show signal paths.

of diodes. The amplifiers were made to approach linearity very closely and are isolated from the limiters by cathode followers in order to make amplitude to phase conversion negligible.

3.1.4 *Phase Detector Characteristic*

The phase detector characteristic, volts versus phase, extends through zero equally in both directions, and is linear over a wide range. How it works is explained briefly in the simplified diagram of Fig. 26. There are two inputs to the phase detector, the signal wave and the reference standard. At the positive zero crossings of these waves, sharp pulses are generated to operate the flip-flop whose output is the square wave at *E* when *A* and *B* are opposite in phase as shown. The edge of the square wave at *E*, controlled by the refer-

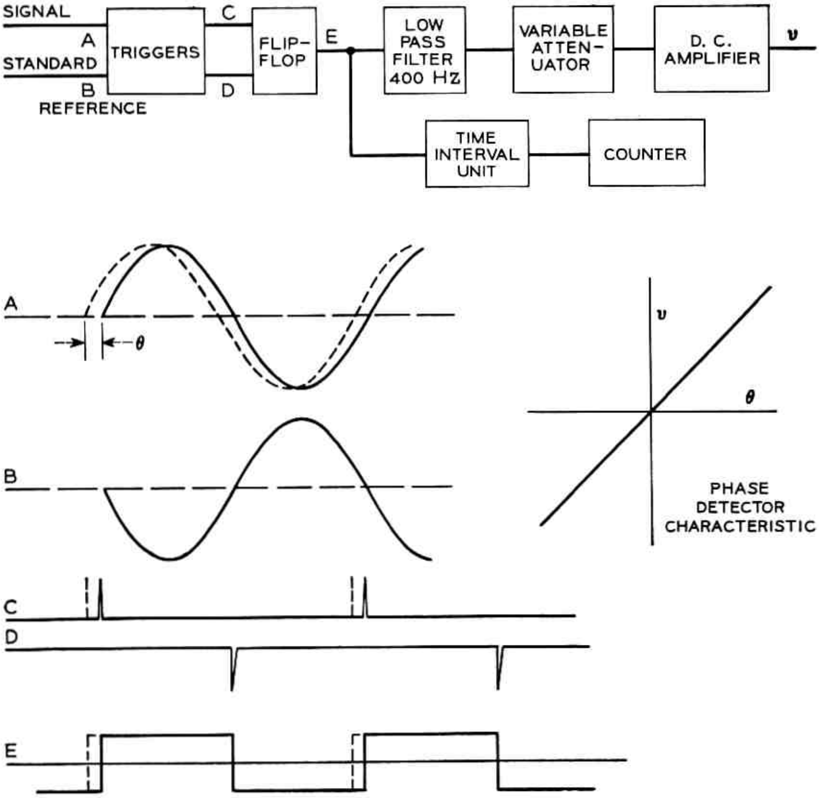


Fig. 26 — Block diagram and operation of phase detector.

ence standard, is fixed; the other edge varies as the phase of the signal wave. The average value of E appearing at the low-pass filter output is then proportional to the time variations of the positive zero crossings of the signal wave, that is, its phase. The low-pass filter and a little shaping in the dc amplifier are such that the bandwidth available for the detector output as indicated in Fig. 26 is about 0.4 the pulse rate. The noise wave thus derived is nearly the best continuous representation of the zero crossing deviations.

Since part of this detector is like a sampler operating at 1 kHz, there will be aliasing in the process if the original time variations contain frequency components higher than 0.5 kHz, as discussed in Section 2.6.4 and 2.6.5.

With the time interval unit and counter connected at E, the time duration from a fixed edge of the square wave to the adjacent variable edge can be measured so that individual zero crossing variations as well as the smoothed wave v may be obtained from the detector; or the time interval unit may be operated to measure from one variable edge to the next. In this way "spacing jitter" may be measured.

This connection is also used in the standardization of the detector characteristic. With an all pulses present train applied to the tuned circuit, a steady sine wave is obtained at the signal point A of the phase detector. The precision phase shifter which supplies the reference at point B is adjusted until the duration of the positive part of E is 500.0 microseconds. Then a bias adjustment is made to bring v to zero volts. After this the phase shifter is moved by various amounts from this reference condition and the amplifier gain adjusted to give the proper detector output voltage. In this calibration a Leeds and Northrup potentiometer is used for voltage measurement.

3.1.5 *Phase Modulator*

The phase modulator generates a sharp pulse at the instant voltage coincidence occurs between the input signal wave and a very "linear sawtooth" wave generated from the 1 kHz standard. This pulse is then used in the regenerator for timing the new pulses. The modulator must generate this timing pulse just as precisely as the phase detector detects the zero crossings of the original timing wave. It is calibrated by applying a known voltage which then causes a phase shift throughout the system. This phase shift is then converted to a voltage by the phase detector and the result compared with the input. The modulator sensitivity is adjusted so that the detector output is the same as the modulator input. Thus all calibration of the detection and modulation process is in terms of two absolute standards, the Leeds and Northrup potentiometer for voltages and the phase shifter for phase angles.

The circuit for the modulator was originally developed and worked out by L. R. Wrathall. This was revised somewhat and built in final form and thoroughly tested by C. R. Crue.

3.1.6 *Tape Recorder*

The tape recorder used is an Ampex FR-100B with servo speed control such that reproduced signal time never varies by more than ±0.25 ms (1 ms = 1 pulse period) from precise time. Because the noise wave is essentially a low frequency signal with most of its energy below 5 Hz, and because frequency modulation (FM) recording is used, these small variations in time will affect only the time at which this old noise is added to the new. Tests have demonstrated that whatever time variation there is has a negligible effect. FM recording at 30 inches per second is used on five tracks, one of which is for speed servo. The machine is operated at 50 Hz derived by step-down chain from the 1 kHz standard which in turn is derived from the 1 MHz crystal oscillator in the counter. Two tracks provide a 5 kHz band for the pulse train. The other two tracks are for the phase noise signal and the bandwidth for these has been reduced to 1.25 kHz to lessen recorder noise. The gain through these channels is made unity. A phase shifter in the 50 Hz supply to the recorder makes it possible to align the recorded pulses with the gating triggers at the regenerator input each time the recorder is started.

3.1.7 *Delay Network*

The delay network in the pulse path makes up for the delays in the noise path caused mainly by the detection and recording process so that the noise record and pulse record correspond at any time.

3.1.8 *Trigger Circuit and Envelope Detector*

A trigger circuit with adjustable triggering level was introduced into the system in place of the limiter for the investigation of amplitude to phase conversion effects. Operation and data taking were simplified by removing the limiter entirely even though in a real repeater some limiting, though imperfect, would be used.

For the measurement of the properties of amplitude variations of the tuned circuit response, an envelope detector was connected at the tuned circuit output.

3.1.9 *Pulse Shaping Networks*

For the study of wide pulses and intersymbol interference, pulse shaping networks were introduced between the pulse regenerator and the tank circuit. These were RC circuits and low-pass filters for producing a sine-squared shape. In some cases a half-wave rectifier was introduced between filter and tank circuit.

3.1.10 *Longword Pulse Pattern*

In order to investigate some aspects of timing noise, an attachment to the signal generator was made so that a longword pulse pattern with a period of 240 bits, and hence having important components within the band of the tank circuit, could be applied to the system. The basic parts of this attachment are (*i*) a code plate with a rectangular 15 by 16 array of holes, placed before (*ii*) a cathode ray tube, the electron beam of which is swept across all the holes successively, and (*iii*) a light sensitive device to convert the light coming through the code plate holes into pulses. The longword signal pattern which is desired is then brought about by blocking out with black tape the proper holes in the code plate. This apparatus was developed by C. R. Crue.

3.2 *Method of Operating the System*

In using the system to obtain a series of noise records corresponding to the timing noise at successive repeaters in a chain, the procedure is as follows. After calibrations have been made with the all pulses present condition, the regenerator is connected to the clipped noise source (which has been adjusted to give the desired pulse density) and recordings are made of the pulse train on track 5, the generated phase noise in the timing wave on track 3, and the 50 Hz servo control wave on track 1, for about 15 minutes. Then the tape is rereeled, calibration checked, and playback started with track 3 going to the phase modulator and track 5 going to the pulse regenerator. This time the pulse train is recorded on track 4 and the timing noise on track 2. We now have two timing noise records corresponding to the timing waves at the first and second repeaters of a chain and these are now analyzed following the procedure outlined below.

Notice that it is the phase noise on the timing wave which is analyzed here rather than the repeater output pulses. These are the same, of course, in a completely retimed repeater. Therefore, it is not necessary that the recorded pulse train have the accumulated timing noise; it is a perfectly timed replica of the original pulse train. The first part of the regenerator removes any time variations acquired in the recording-playback process. In the first recording, a few minutes of all pulses present is recorded before switching to the random train to give the recorder servo time to synchronize and also to allow time for the alignment of recorder to system each time it is started.

To obtain the next pair of records corresponding to transmission through repeaters 3 and 4 of the chain, track 2 is played back to the phase modulator, while the detected accumulated noise is recorded on track 3 thus erasing the record first made there. Then track 3 is played back to the modulator with the new noise record being made on track 2. And so on, as long as desired or until some difficulty in the process arises.

During the playback, the alignment of system and recorder is monitored continuously to make sure that the recorder stays in synchronism. If it does not, the pulse train may be altered.

A great amount of time and effort has gone into the building of the system just described to make it sufficiently stable and accurate for measuring phase deviations to within less than 0.1° out of 10°. It is necessary to measure with this precision in order to be able to describe accurately the change in noise from one repeater to the next because of the small amounts involved. The rms value of noise generated at one repeater with $Q = 100$ and 0.1 percent detuning is a little less than 1°.

Another factor in the reliability of the data is that a long enough signal was used for analyzing so that fluctuations in the plotted parameters of the noise were fairly small. As described below, 32 ten-second averages of the time the noise wave spends below each threshold are used for each point on the cumulative distributions. Since the counter rests for 10 seconds after each adding period, the length of signal involved is 640,000 pulse positions. Each plotted point is well enough established, so that the curve connecting them is smooth without the necessity of further averaging.

The residual noise at the detector output for all pulses present is about 0.006° rms. For a random train of pulses (narrow) there should be no phase noise generated if the tank circuit is centered exactly on the pulse rate. In this situation, the residual noise is about 0.012° rms. See Fig. 28. This is not only a good test of the system as a whole, but is a good dynamic test of the limiter, which is hard to do in any other way.

3.3 *Apparatus for and Process of Analyzing Data*

3.3.1 *Cumulative Distribution-Slicer Circuit*

To obtain the cumulative distribution of the noise, a slicer circuit was developed. This is an adjustable threshold device which generates a standard height pulse whenever the noise wave is below the threshold. The duration of each of these pulses is measured by counting the number of cycles of a 100 kHz wave which the pulse gates to the counter. The accumulated durations of all these pulses which occur in a standard interval of 10 seconds is then totaled by the counter.

At each threshold setting, 32 of these totals are obtained and plotted in control chart fashion as shown in Fig. 27. This helps us to see if the data are statistically acceptable. If appreciable trouble has occurred in the apparatus during the run, it will show up in this picture. The median taken from each chart of data is used to plot one point on the distribution. After the distribution is plotted, the rms value is taken from 1/3 the difference of the values at the 93.3

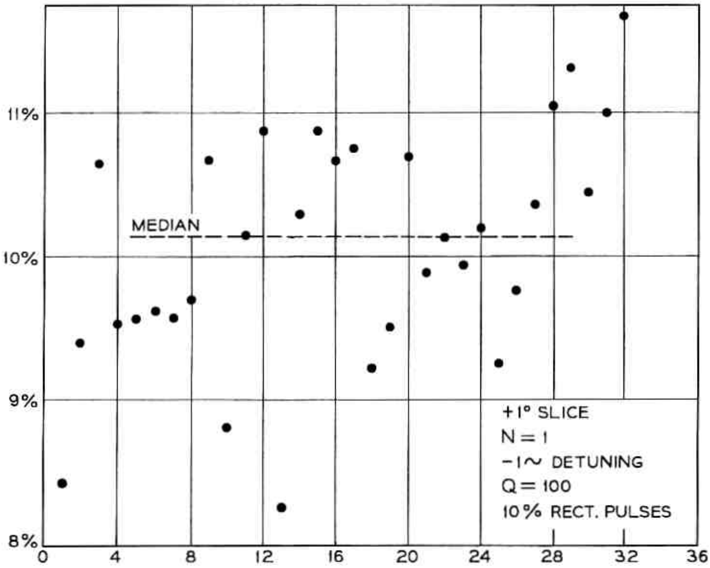


Fig. 27 — Control chart plot of data to obtain one point of cumulative distribution (as in Fig. 31) of timing noise amplitudes.

percent and 6.7 percent points, following a method suggested by E. B. Ferrell for skewed distributions.

3.3.2 *Spectrum Density—Wave Analyzer*

To obtain the spectrum density of the noise wave, a General Radio 1900A wave analyzer with external rms indicator is used. In order to extend the nominal 20 Hz lower end of the analyzer to around 1 Hz, a 16-to-1 speedup of the noise wave is made in a second recording process. Each point on the cumulative distribution requires about 11 minutes of signal, so the original noise record is played back and then rereeled for each point. During each playback the noise record is duplicated at lower speed by bridging the input of another FR-100 recorder operating at 1⅞ inches per second at the main recorder output. The new record then consists of about 8 to 10 serial duplications of the noise which, when played back at 30 inches per second, provides about 5 minutes total of the original noise with all frequency components multiplied by 16. A wave analyzer band of 10 Hz then is equivalent to a 0.625 Hz band for the original noise.

In order to obtain consistent and reliable measurements of noise power in the wave analyzer band, it was found necessary to replace the linear detector provided in the wave analyzer with a square law detector. To do this, the wave analyzer IF output was connected to a Ballantine rms meter and the square law response of this applied to a 4-second time constant RC smoothing circuit and a linear scale meter. The pointer of this meter still fluctuates, thus requiring some kind of average. This average was obtained arithmetically from 20 successive meter readings made at 5-second intervals and, after calibration, was taken as the measure of the mean square power of the noise falling within the wave analyzer band at each frequency setting.

The variance of each of these readings depends partly on the analyzing filter bandwidths and partly on the lengths of signal record available. The latter part could be obtained only by calculation and appears to be the dominating part. Variance estimated this way indicates that any reading taken by the above method is within 10 percent of the true value with 95 percent confidence. The data taken appear to have less variation than this. Some further smoothing of an unknown amount occurs in the plotting of the individual points to give a smooth spectrum curve. Integration of these curves consistently gave results agreeing (within a very few percent) with measurements of total noise power.

IV. RESULTS OF EXPERIMENTS

4.1 *Use of Narrow Rectangular Pulses*

The first measurements of phase noise on the timing wave recovered from a random train of narrow pulses by a tank circuit mistuned from the pulse rate showed that the magnitude of the noise is proportional to the amount of mistuning when this is only a few tenths of a percent, as predicted by W. R. Bennett.[1] Hence there should be no noise for zero mistuning. It was found difficult to define, though, just what zero mistuning is for a real LC tank circuit. In the experiments reported here, the tank circuit consisted of an air core coil and capacitor in shunt at the collector of a transistor, the emitter of which had a resistor equal to the resonant impedance of the LC tank. The transistor was driven at its base and there was positive feedback using another transistor to bring the effective Q of the tank to a value of 100. The reference point from which mistuning was measured, chosen because it could be easily set with sufficient precision, was that of 180° phase between the collector and emitter voltages of the tank transistor as determined by an oscilloscope Lissajou figure.

That the minimum of phase noise does not occur at this point is shown by the curve of Fig. 28. Neither is its value zero, being made up partly of the residual noise of the system as indicated by APP (for all pulses present) and partly of the noise from the train of random narrow rectangular pulses. However, in the investigation of noise caused by pulse shape alone (Section 4.3), it was found that this minimum does not coincide with zero mistuning. Rather, the minimum is the result of cancellation by small amounts of mistuning and trigger offset of part of the noise attributable to pulse shape. While the minimum is about 0.012 degree, the noise from the pulses is about 0.036 degree. This is the true zero mistuning and occurs about where the curve crosses this ordinate. Even though the noise contributed by the narrow rectangular pulses is not zero, it is quite small as may be seen from the spectrum curves of Figs. 4 and 20. Hence using the narrow rectangular pulses in the measuring of noise caused by tank circuit mistuning and by offset trigger gives results which very nearly isolates these as sources for individual scrutiny.

4.2 *Tank Circuit Mistuning*

The curves of Fig. 2 show how the spectrum changes as the noise is examined at successive repeaters, each mistuned by the same amount and direction, in a chain of six repeaters. Distinctive features of these

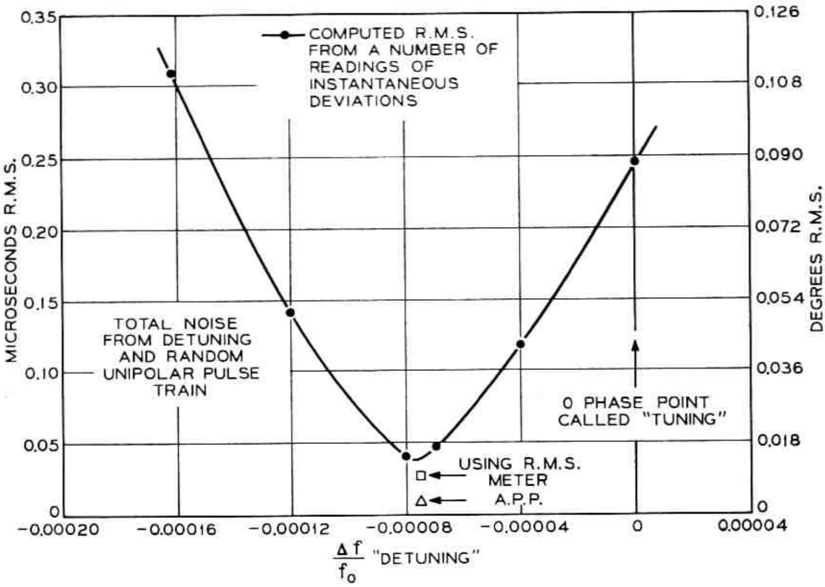


Fig. 28 — Measured timing noise as a function of tank circuit detuning near minimum. This shows residual noise of the system.

spectra are the zeros at zero frequency and the maxima which occurs at $f = f_o/2Q$ for $N = 1$ and closer to zero as N increases. Additional measurements on longer chains up to 20 in length show that this trend continues and agrees with magnitudes predicted by the theory presented in Section II. In Fig. 14, these predictions are plotted and extended to a chain of 100 repeaters. It is evident from these results that while the maximum continues to rise, its magnitude will never be greater than four times what it was at the first repeater. The reason for this is that the noise spectrum generated at each repeater is zero at zero frequency and the succeeding tank circuits continually attenuate the higher frequency components. These effects are seen also in the data on the total phase noise along the chain. For chains up to 20 in length, measured and calculated values are plotted in Fig. 29. Calculated values for longer chains are plotted in Fig. 15.

Figure 30 shows the measured and calculated spectra of phase noise at two adjacent repeaters when the second one is mistuned in the opposite direction but by the same amount as the first one. This has the effect of reversing the dissymmetry of side frequencies about the carrier in the second repeater, and so there is a partial canceling

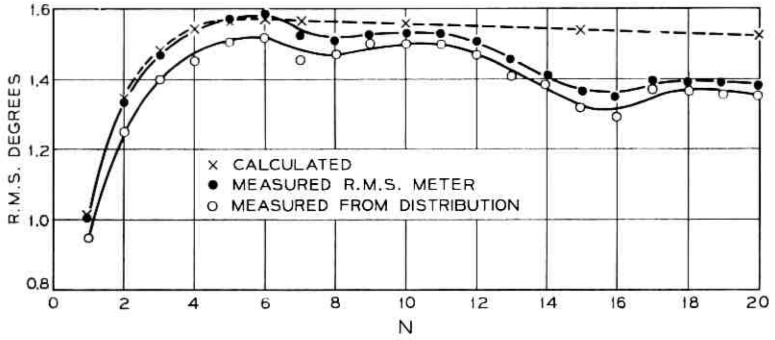


Fig. 29 — Total timing noise caused by 0.1 percent mistuning of timing tanks in chains of N like regenerative repeaters.

of phase noise. Thus the greatest accumulation of timing noise caused by mistuning comes when all repeaters in a chain are mistuned the same way.

In all these cases, we have seen how well the values of timing noise calculated from the theoretical model of Section II agree with those

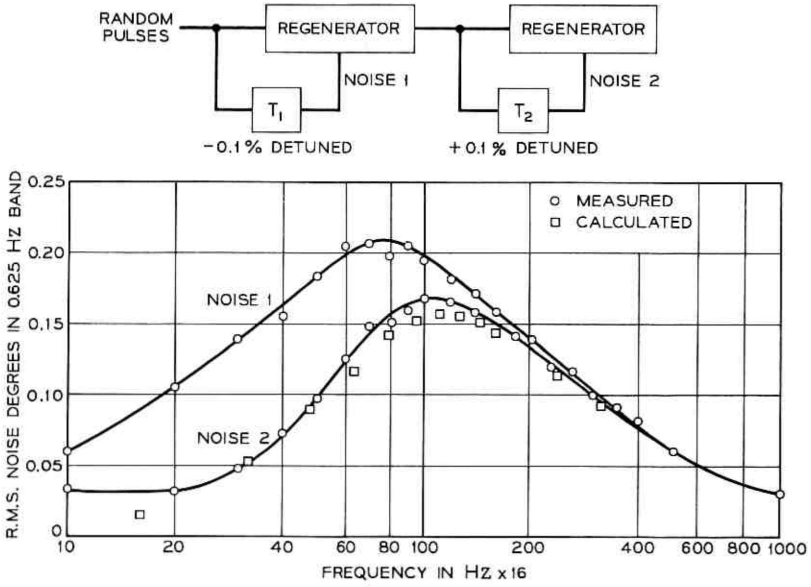


Fig. 30 — Spectra of timing noise at two successive regenerative repeaters with oppositely detuned timing tanks.

obtained by measurement. The theoretical model was developed from the simple picture of the equivalence of phase modulation of the recovered carrier and dissymmetry of the side frequencies about this carrier introduced by the mistuning of the tank circuit into the otherwise symmetrical side frequency and carrier representation of the random pulse train.

The cumulative distribution curve of phase noise generated at a mistuned repeater, obtained from measurements, is shown in Fig. 31 and is seen to have a shape which is approximately log-normal. Calculation of a distribution curve which agrees quite well with these results has been made by M. R. Aaron and J. R. Gray.[9] When the distributions of timing noise at successive repeaters along a chain are examined, it is found that the skewness is gradually reduced.

Another set of data from the measurements of timing noise caused by mistuning is that concerning spacing noise which is displayed in Fig. 32. Spacing noise is defined as the deviations from normal of the spacing between successive positive (or negative) going zero crossings of the timing wave. As suggested by M. R. Aaron and H. E.

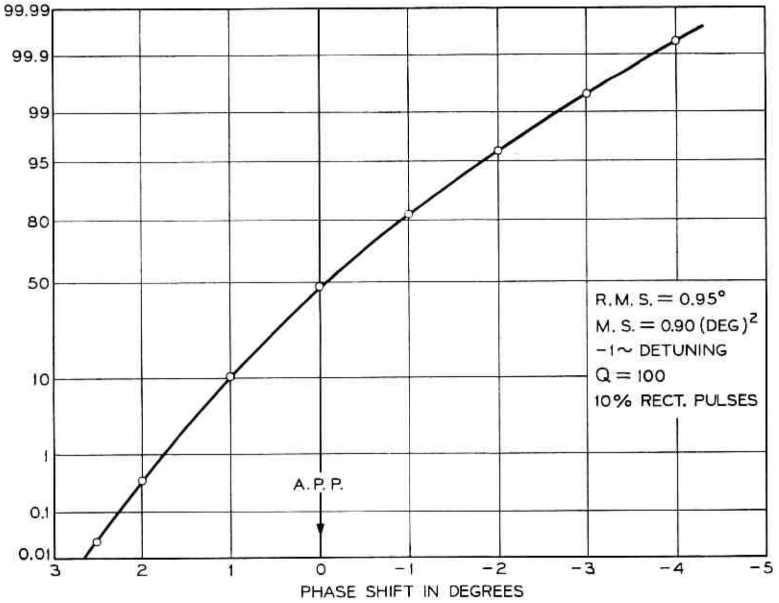


Fig. 31 — Cumulative distribution of timing noise amplitudes obtained from measurements (noise caused by tank detuning).

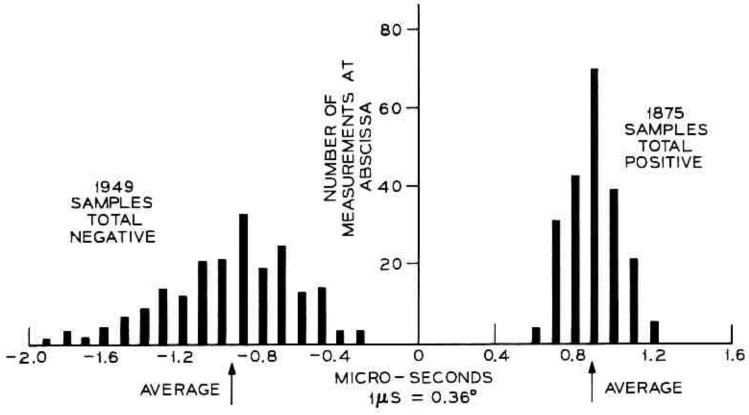


Fig. 32 — Timing wave spacing noise caused by tank mistuning. Measured deviations of zero-crossing differences from 1.0 millisecond.

Rowe, the reason for this lies in the quantized character of changes in the pulse pattern. That is, at any point in time, the next time slot has either a pulse or no pulse. The magnitudes can be found from the results shown in Fig. 39a for one absent pulse in 240 and mistuning. Scaling this to 0.1 percent mistuning (condition for Fig. 32), gives a peak phase change of 0.36° or 1 μs. For a single pulse added, the phase change would have the opposite sign. Since the most likely change in the random pulse train is that of a single pulse added or left out and other changes are less likely, the distribution of Fig. 32 should have peaks near +1 μs and −1 μs as observed.

4.3 *Amplitude-to-Phase Conversion*

The pulse rate fundamental recovered from the signal pulse train by narrowband tank circuit has appreciable phase deviations only if the tank circuit is mistuned from the pulse rate. But this fundamental has noisy amplitude variations even when the tank circuit is perfectly tuned.

By connecting an envelope detector at the tank circuit output, recordings were made of the response amplitude variations. The spectrum is shown in Fig. 33 where it is seen to have a nonzero value as frequency approaches zero as predicted by the calculated values superimposed and as also shown by the calculated curve of Fig. 8. That this must be so may be seen by considering that the noise side frequencies continue to exist at about the same magnitude as they get closer to the carrier (pulse rate). That is, the amplitude modulation is not reduced, only the dissymmetry disappears. The measured cumulative distribution of the amplitude variations is very nearly normal.

One way in which these amplitude variations may be converted into phase variations is through an imperfection in the trigger circuit which, from the timing wave, generates sharp pulses for retiming the signal pulses. Such will be the case if the triggering level of this circuit is, for some reason, offset from the zeros of the timing wave.

This kind of timing noise was generated in the system for simulating a chain of regenerative repeaters by replacing the amplitude limiter which follows the tank circuit in Fig. 9 with a trigger circuit having an adjustable threshold. In a real repeater, some amplitude limiting, though imperfect, would be used between tank and trigger, but here it is more convenient to leave out all limiting.

Spectra of this amplitude to phase timing noise, when there is no mistuning, at successive-like repeaters in a chain of six are plotted in Fig. 3 along with calculated points. The spectra have the same shape as that of the amplitude variations since the two phenomena are directly related. The rms magnitudes at very low frequencies

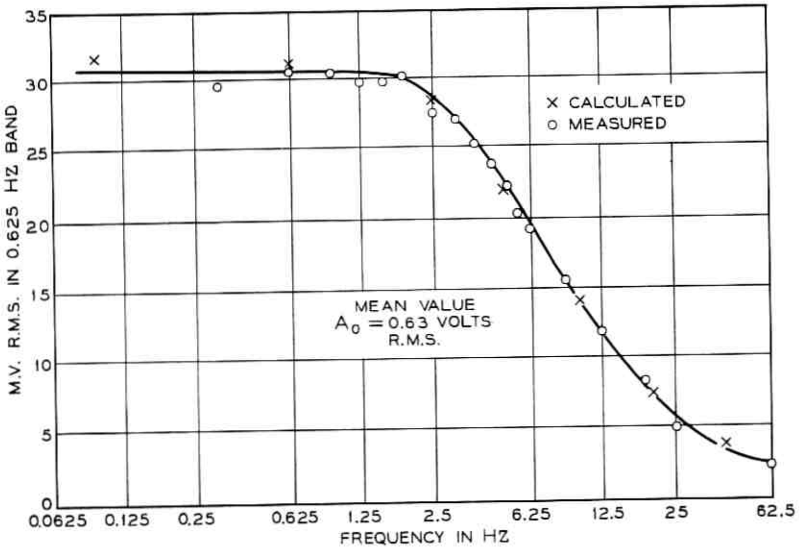


Fig. 33 — Spectrum of amplitude variation of tuned circuit response to random unipolar rectangular pulses.

grow directly with repeater number because of the nonzero magnitude of noise and nearly zero tank circuit attenuation at very low frequencies. Total timing noise in the same situation is shown in Fig. 34. The cumulative distribution of the noise is somewhat curved, about like the amplitude to phase conversion characteristic of the trigger circuit.

Timing noise was measured also when mistuning and amplitude to phase conversion were both present in the same repeater. This is shown in Fig. 10, along with calculated values.

We see in all these results very good agreement between the measured values and those calculated, as outlined in Section II, by the method developed first in the investigation of noise caused by mistuning.

4.4 *Phase Noise Attributable to Pulse Form*

4.4.1 *Rectangular Pulses*

After it was found that the low sharp minimum of timing noise shown in the curve of Fig. 28 was attained by small deviations from zero mistuning and zero trigger offset, fairly good agreement between measured and calculated noise spectra for rectangular pulses was obtained. Most of the curves of Fig. 4 are from measurements while those of Figs. 20 and 21 are from calculations.

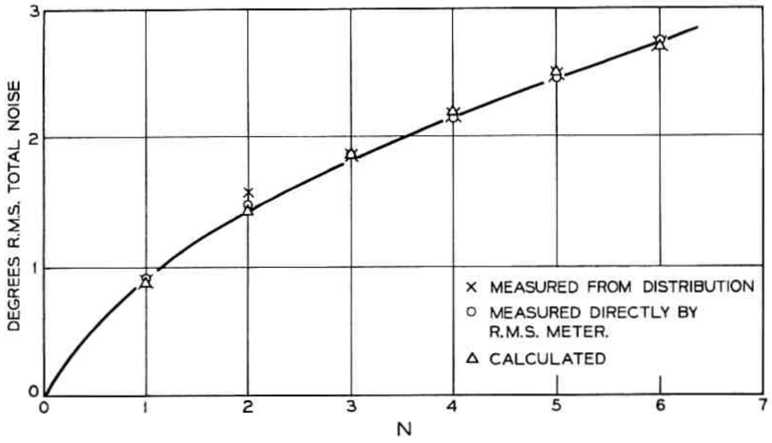


Fig. 34 — Total timing noise from amplitude variation of timing wave and offset trigger as a function of the number (N) of repeaters in a chain.

It is only the narrow rectangular pulses which have the magnitude and shape of noise spectrum which can be changed to produce the result of Fig. 28 by small adjustments of the two indicated factors. The calculated curves of Fig. 22 show how this can be. For rectangular pulses with $\tau/T = 0.6$, the total noise changes only about 10 percent as the tuning varies over the range of Fig. 22.

An attempt was made to measure the effect of high frequencies in the tank circuit response by cutting out these components; it was not entirely satisfactory because of the difficulty of obtaining a suitable filter. The results were sufficient though to verify the general features of the differences between the curves of Fig. 20 and Fig. 21.

4.4.2 *Raised Cosine Pulses*

Trains of pulses approximating the raised-cosine shape were generated by applying the 10 percent duty factor rectangular pulses to a 4-section low-pass filter built from a design by W. E. Thomson.[12] Filters were built to generate pulses $1T$, $1.5T$, and $2T$ wide at their bases. Oscilloscope presentations of the $1T$ pulses are shown in Fig. 18c and of the $2T$ pulses in Fig. 18a.

The measured spectrum of phase noise caused by the $1T$ pulses is shown in Fig. 35 along with the calculated values and it is seen that agreement is fairly good. This spectrum has appreciable energy at considerably higher frequencies than does that caused by mistuning or trigger offset and narrow rectangular pulses. The reason for this is that the dissymmetry of side frequencies extends to much higher frequencies. The same is true of course for the rectangular pulses wider than $\tau/T = 0.5$.

Measured and calculated spectra for the $1.5T$ pulses are plotted in Fig. 36 where it may be seen that the total amount of noise is about four times as great as that for the $1T$ pulses.

Since there seemed to be no suitable way to generate the asymmetrical pulses, no experimental data is available for this case.

Some of the data obtained with rectified $2T$ pulses are presented in Fig. 37. It is difficult to duplicate experimentally the idealized waveform of Fig. 17b assumed in the calculations of this case. For example the wiggles at the top of the real waveform Fig. 18a come from small departures from ideal of the pulse shapes, and the rectifier does not produce cusps at its cutoff point but rounded transitions as in Fig. 18b. The data plotted in Fig. 37 agrees fairly well with calculation, and it is believed to be reliable. But other data has been

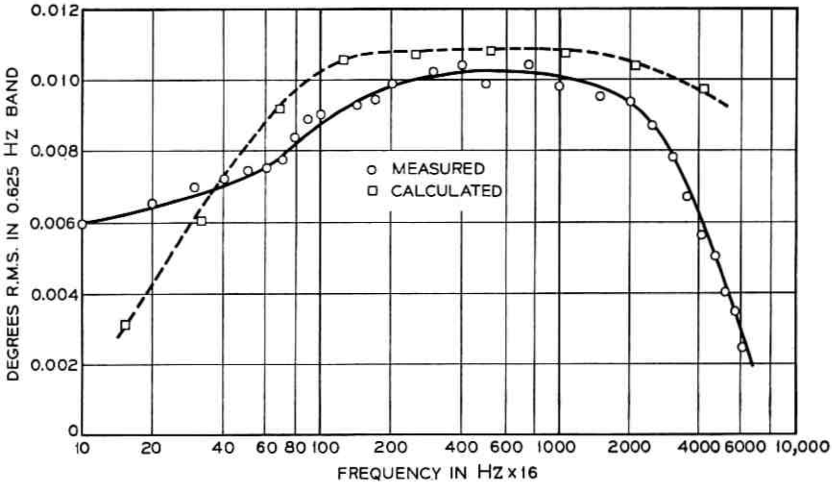


Fig. 35 — Spectrum of timing noise from raised cosine random pulses one time slot wide at base. No tank mistuning.

obtained which gives spectra both smaller and larger than that shown.

When the driving pulses are ac coupled to the shaping filter, there is a large increase in very low-frequency noise components as shown. This illustrates the statement made before that if there is low-frequency distortion in the transmission of pulses, then nonlinearity

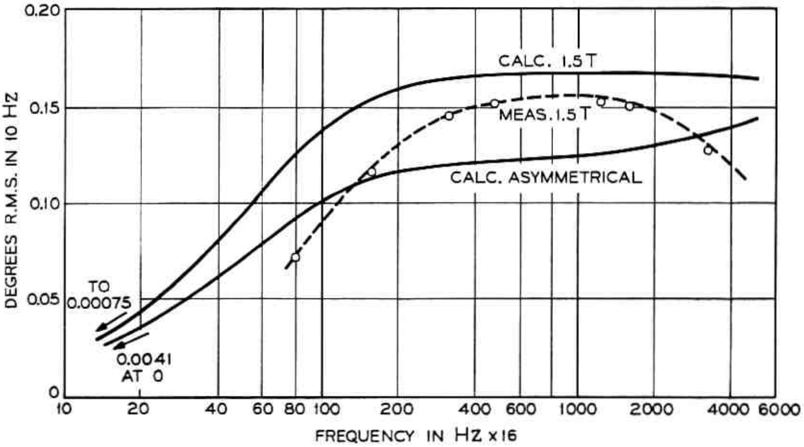


Fig. 36 — Timing noise spectra for 1.5 T and asymmetrical raised cosine pulses.

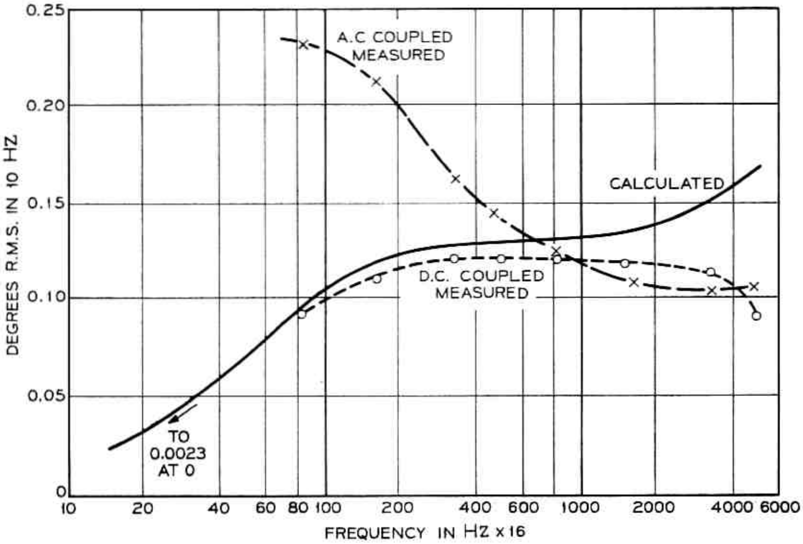


Fig. 37 — Spectrum of timing noise from rectified raised cosine puses two time slots wide. No tank mistuning.

which follows this can convert this distortion into very low-frequency phase noise. With ac coupling, the pulse waveform of Fig. 18(a) is changed to that of Fig. 38.

4.5 *Long Word Periodic Pulse Pattern*

In addition to the previous results obtained with a random pulse train, a few measurements were made using a periodic pulse pattern of period 240 time slots.

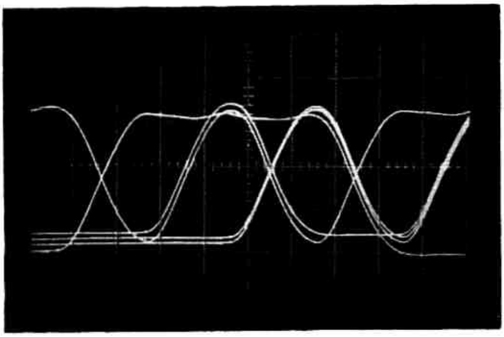


Fig. 38 — Photograph of random pulse oscilloscope traces for raised cosine pulses two time slots wide at base. AC coupling.

For a particular pattern chosen at random, the measured phase modulation on the recovered pulse rate agrees very closely with that calculated from the theory developed in Section II. Hence there is nothing in periodic pulse patterns as such to cause behavior different from that predicted by the theory.

Another particular pattern, with one pulse per period missing was used to measure a sort of phase impulse response. Under this condition phase deviation was observed and photographed under two conditions. Figure 39a shows the phase detector response for tank mistuning and Fig. 39b that for an offset trigger circuit following a perfectly tuned tank. In Fig. 39a, the mistuning is −0.2 percent and the peak phase deviation 0.73 degree. In Fig. 39b, the trigger offset is 12.6 degrees and the peak phase deviation is 0.43 degree. The wiggles on both waveforms are a residual noise in the system and have nothing to do with the phenomena being discussed.

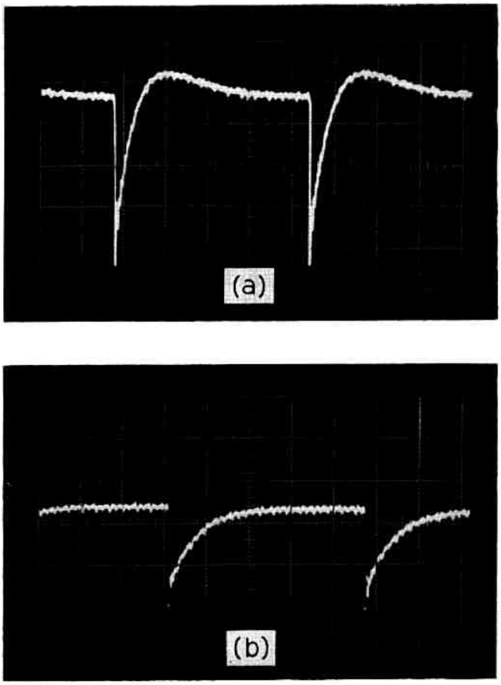


Fig. 39 — Photographs of oscilloscope traces—Phase deviation of recovered fundamental pulse rate for pattern of 239 pulses, 1 space, with (a) tank circuit mistuning, (b) amplitude to phase conversion by means of trigger offset.

Both of these resemble the response of an RC circuit to an impulse. A simple picture of this situation then is that deletion of one pulse from the all pulses present condition is converted, by one or the other of the two imperfections considered, into an equivalent impulse of phase change which through the tank phase transfer characteristic produces one of the responses of Fig. 39.

But there is a small, though significant, difference between the two responses. The simpler of the two, Fig. 39b, with its sharp rise and exponential decline is very close to an RC impulse response. Its measured time constant is about 30 ms which is quite close to the value of 32 ms which describes the phase transfer characteristic derived from sine-wave measurements. The Laplace transform of this pulse is $F(p) = 1/(p + \alpha)$. Its amplitude spectrum, with its finite value at zero frequency, is like the measured spectra of phase noise caused by amplitude to phase conversion.

The pulse in Fig. 39a differs from the other, principally in its crossing of the baseline and undershooting as it declines to zero. The observed pulse can be approximated very closely by the modified exponential $f_1(t) = e^{-\alpha t}(1 - \alpha t)$ which has the Laplace transform

$$F(p) = \frac{p}{(p + \alpha)^2}.$$

The corresponding amplitude spectrum is

$$| F(j\omega) | = \frac{| \omega |}{\omega^2 + \alpha^2}.$$

It is seen that $f_1(t)$ goes through zero at $t = t_0 = 1/\alpha$ and has a minimum value of $-e^{-2} = -0.135$ at $t - = 2t_0$. Also, it is seen that $| F(j\omega) |$ has its maximum at $\omega = \alpha$. From the photograph, we estimate that $t_0 = 33$ ms and that the minimum of 0.135 occurs at $2t_0$. Further, from the earlier work we find that both measurement and theory show a maximum in the spectrum of phase noise caused by tank circuit mistuning at $\omega = \omega_0/2Q$. Equating this to α gives $1/\alpha = 32$ ms which is very good agreement with the estimation of t_0.

Thus the "impulse response" view of phase deviation in the fundamental recovered by means of a narrowband tank circuit is consistent with the spectrum modification view worked out earlier.

V. CONCLUSIONS

Two important sources of timing noise in a self-timed regenerative PCM repeater, namely tank circuit mistuning and amplitude to phase conversion by means of an offset trigger circuit, have been identified and studied both experimentally and theoretically. Another source, pulse shape and duration has been studied less than the other two. Another subject investigated is the way these noises accumulate along a chain of repeaters and how they combine.

As a part of this investigation, a simple theoretical picture of the process has been developed whereby many important properties of timing noise and its accumulation along a chain of repeaters can be calculated with results which agree very well with corresponding measurements.

The process of calculation begins with the spectrum of the pulse train, then proceeds to the modifications of it which arise as the pulse train is transmitted through the timing tank and retiming process of each repeater. The properties of the timing noise which appear on the retiming wave depend on the particular character of these modifications. The further modifications of the spectrum of the pulse train, acquired at each repeater, are used to find the way in which the timing noise accumulates along the repeater chain.

These investigations have shown that the characters of the spectra of these timing noises are important in determining how they accumulate along a chain of repeaters. It has been found that the spectrum of the noise caused by tank circuit mistuning has a zero at zero frequency. It has been demonstrated that because of its property, the noise increases only for the first few repeaters of the chain, soon reaching a limit. Further, it has been found that the spectrum of timing noise which depends on amplitude variations of the timing wave (as in a trigger circuit where the firing point has been offset from a zero crossing) has a nonzero value at zero frequency. It has been shown that because of this property, this timing noise increases without limit along the repeater chain. The total amount of the noise varies inversely with the Q of the tank circuit.

Thus, whether or not timing noise increases without limit along a repeater chain depends, not on whether the same noise is generated and added on at each repeater, but on whether or not the spectrum of the noise added at each repeater has a nonzero value at zero frequency.

Study of the effect of pulse shape and duration have shown that while the total noise from this source is greater than it is for the others, the amount at very low frequencies is quite small, though not zero in a number of cases. These latter components are the result of aliasing of the high-frequency parts of the tank circuit response.

If there is nonlinearity in the fundamental recovery path, low-frequency pulse distortion during transmission can be converted to very low frequency timing noise.

VI. ACKNOWLEDGMENTS

S. L. Freeny has had an active part in this project during much of the time it was carried on. Also, consultation with him has been frequent and valuable. The work of C. R. Crue in developing and building and testing quite a bit of the apparatus used was very helpful. Many helpful conversations with S. O. Rice, and the analysis carried out by him for the case of wide overlapping pulses, are gratefully acknowledged. Numerous consultations with B. G. King and J. A. Young, who supervised the work during much of the time, were very valuable in bringing the project to a successful conclusion.

REFERENCES

1. Bennett, W. R., "Statistics of Regenerative Digital Transmission," B.S.T.J., *37,* No. 6 (November 1958), pp. 1501–1542.
2. Wrathall, L. R., "Transistoried Binary Pulse Regenerator," B.S.T.J., *35,* No. 5 (September 1956), pp. 1059–1084.
3. Sunde, E. D., "Self-Timing Regenerative Repeaters," B.S.T.J., *36,* No. 4 (July 1957), pp. 891–938.
4. DeLange, O. E., "The Timing of High-Speed Regenerative Repeaters," B.S.T.J., *37,* No. 6 (November 1958), pp. 1455–1486.
5. DeLange, O. E. and Pusetlnyk, M., "Experiments on the Timing of Regenerative Repeaters," B.S.T.J., *37,* No. 6 (November 1958), pp. 1487–1500.
6. Rowe, H. E., "Timing in a Long Chain of Regenerative Binary Repeaters," B.S.T.J., *37,* No. 6 (November 1958), pp. 1543–1598.
7. Aaron, M. R., "PCM Transmission in the Exchange Plant," B.S.T.J., *41,* No. 1 (January 1962), pp. 99–141.
8. Byrne, C. J., Karafin, B. J., and Robinson, D. B., "Systematic Jitter in a Chain of Digital Regenerators," B.S.T.J., *42,* No. 6 (November 1963), pp. 2679–2714.
9. Aaron, M. R. and Gray, J. R., "Probability Distribution for the Phase Jitter in Self-Timed Reconstructive Repeaters for PCM," B.S.T.J., *41,* No. 2 (March 1962), pp. 503–558.
10. Kinariwala, B. K., "Timing Errors in a Chain of Regenerative Repeaters, I and II," B.S.T.J., *41,* No. 9 (November 1962), pp. 1769–1797.
11. Kinariwala, B. K., "Timing Errors in a Chain of Regenerative Repeaters, III," B.S.T.J., *43,* No. 4 (July 1964), pp. 1481–1504.
12. Thomson, W. E., "The Synthesis of a Network to Have a Sine-Squared Impulse Response," Inst. of Elec. Eng. Proc. (London), *99,* part III (November 1952), pp. 373–376.

Synchronizing Digital Networks

By J. R. PIERCE

(Manuscript received September 11, 1968)

It appears that stable synchronization of large digital transmission networks should be easy, granted accurate clocks, buffers which accept pulses at the incoming rate and deliver them at the local clock rate, and adequate delay for making frames coincide. An electric network analog of a simple linear system in which the clock frequency depends on the fullness of buffers and the departure of frequency setting from midsetting makes it obvious that the system is stable. System frequency should be made to depend strongly on accurate or master clocks; criteria are given for choosing parameters to achieve this. Strategies are given for periodic infrequent adjustments to compensate for changes in transmission time, and for adding new clocks to the network. The practical realization of a synchronized network calls for more information concerning variations of transmission time and for adequate components, particularly, buffers and adjustable delay devices.

The synchronization of digital networks has been studied theoretically and experimentally.[1–9] This paper does not purport to review excellent previous work, some of which has been highly theoretically oriented, through some results of earlier work are referred to. Rather, it discusses some problems of synchronization and illustrates them by means of a simple analysis of a simple example. We reach the following conclusions: that with good clocks, buffers and adequate delay both to compensate for changes in transmission time and to make frames coincide, there should be no trouble in stably synchronizing a nationwide network. This is in accord with earlier analysis and experiment.[6]

I. SYNCHRONIZING FRAMES OR BITS

In some papers, synchronization has been discussed in terms of synchronizing frames, that is, successive groups of bits identified by some framing signal present in each group.[1, 6] In this paper, synchronization will be discussed in terms of synchronization of bit streams.

The choice of the bit stream as the signal to be synchronized is partly arbitrary. However, there may be advantages in performing, at a terminal, as many operations as possible on an accurately timed binary bit stream.

For some purposes, and especially for time division switching, it seems essential to synchronize frames. This can be done by passing the bit stream through a suitable adjustable delay device. The delay measured in pulses, which is needed to make frames coincide, is not small. The frame time for a digital system designed for speech transmission is commonly 1/8000 second. If a transmission system runs at the rate of 5×10^8 pulses per second, the frame time includes about 70,000 pulses. This may be an akwardly large number of pulses to store.

II. CHANGES IN TRANSMISSION TIME AND CLOCK RATE

A scheme of synchronization must take into account both errors in clock frequency and changes in transmission time. These pose rather different problems. Both for this reason, and because it makes the presentation simpler, changes in transmission time are disregarded in the earlier portions of this paper, and treated separately later on.

III. THE SYSTEM CONSIDERED

Various approaches to synchronization are possible. One solution would be to transmit synchronizing signals from a central master source. Few people seem to like this method because of problems of reliability.

The system considered here is composed of a lot of centers (the dots of Fig. 1) with highly stable clocks, interconnected by two-way digital circuits (the lines of Fig. 1). The common frequency of operation will be determined by the characteristics of the clocks and of the transmission circuits.

We assume that each clock is equipped with a frequency control, so that its frequency can be adjusted to be above or below its central, "correct" value. We also assume that each receiver is equipped with a buffer which will accept a bit stream from a terminal at some received rate and emit bits at the rate or frequency determined by the local clock.

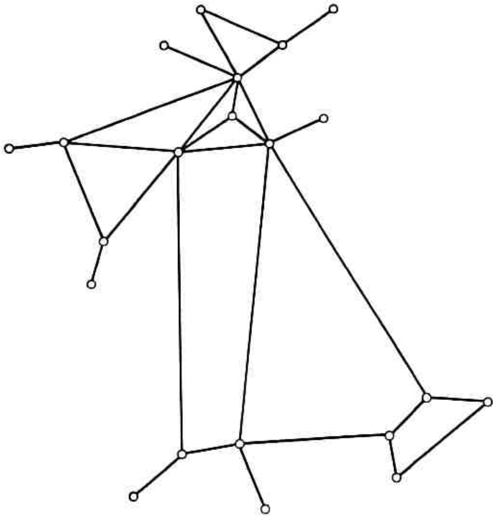


Fig. 1 — A network of centers to be synchronized.

We might assume a system in which each center knew the state of every other center's buffers and clock settings. In Section XI we discuss how such knowledge may be used in dealing with changes in transmission time. Initially, we will consider the case in which each center knows only the state of its buffers and the setting of its clock frequency. Thus, any adjustments of the clock frequency will be based on the frequency of each clock with respect to its center value, and on the state of the buffers, each of which reflects both the clock frequency at the other end of a transmission circuit relative to the local clock frequency, and any changes in the transmission time.

The case considered is illustrated schematically in Fig. 2. The elements concerned with automatic adjustment of the clock frequency are enclosed by a dashed line; they consist of the clock, buffers, and a network whose inputs are buffer readings and (optionally) the reading of the clock setting and whose output is a signal which adjusts the frequency setting of the clock. Other elements shown in Fig. 2 are adjustable delay for framing the bit stream from the buffer output, a decoder for going from received pulses to bit stream, and an adjustable delay before or after the decoder which can be used to compensate for changes in transmission time.

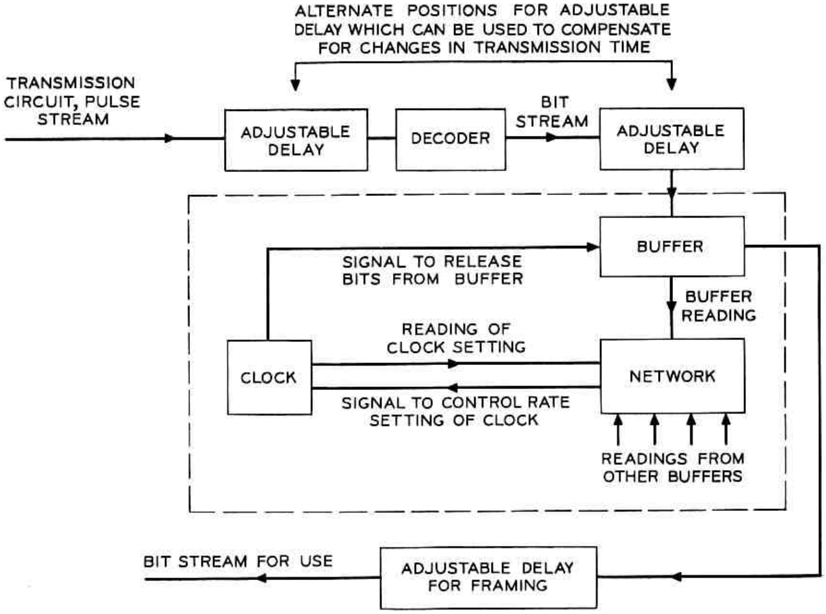

ALTERNATE POSITIONS FOR ADJUSTABLE DELAY WHICH CAN BE USED TO COMPENSATE FOR CHANGES IN TRANSMISSION TIME
TRANSMISSION CIRCUIT, PULSE STREAM
ADJUSTABLE DELAY
DECODER
BIT STREAM
ADJUSTABLE DELAY
BUFFER
SIGNAL TO RELEASE BITS FROM BUFFER
BUFFER READING
READING OF CLOCK SETTING
CLOCK
NETWORK
SIGNAL TO CONTROL RATE SETTING OF CLOCK
READINGS FROM OTHER BUFFERS
BIT STREAM FOR USE
ADJUSTABLE DELAY FOR FRAMING

Fig. 2 — Block diagram showing components used in synchronization.

IV. STABILITY OF CLOCKS

It is clear that both the stability of the clocks and the pulse rate are overwhelmingly important. If the clocks were perfectly stable (if they could be exactly synchronized at the factory and if they maintained their frequency exactly after being shipped to the centers), then the buffers would merely have to take care of fluctuations in transmission time. As the transmission time will not change without bound, finite, realizable buffers would insure the satisfactory operation of the system.

The clocks cannot be synchronized perfectly, but will differ in rate by some fraction, d, which may be 10^{-8} for a good crystal oscillator or perhaps as good as 10^{-12} for an atomic clock. Comparatively inexpensive ($1,800) commercial frequency standards are now available which have a short-term d of 10^{-11} and a long-term value of 2×10^{-11} (see Ref. 10).

Consider two centers interchanging pulses at a rate r per second. If no effort is made to synchronize the clocks, the number N of pulses the buffer must absorb per day would be roughly $N = 86{,}400\ rd$.

Let us consider some cases:

r	d	N
5×10^7	10^{-7}	4.32×10^5
5×10^7	2×10^{-11}	864
5×10^7	10^{-12}	43.2

This is instructive. If the clocks were stable enough, the buffers could be dumped and the delays readjusted at strategic times, with resulting errors through loss of message bits. This might be feasible with the Hewlett-Packard clock.[10] If very stable clocks are used, any adjustments can be very slow or very infrequent.

V. TRANSMISSION TIME

The time delay in transmission is a complication in any analysis of synchronization. This makes the idea of infrequent periodic adjustments or very slow continuous adjustments attractive. If the time intervals between adjustments or the time constants involved in the adjustment process are long compared with the transmission time, then a sort of kinetic model, in which transmission time can be disregarded, will apply.

Transcontinental transmission time via coaxial cable is of the order of 0.02 second, and via microwave radio somewhat less. For a clock stability $d = 10^{-8}$ and a pulse rate 5×10^8, around five pulses would accumulate in the buffers each second, and an adjustment once a second seems reasonable. For a $d = 10^{-10}$ and the same pulse rate, 5 pulses would accumulate in 100 seconds, a time long compared with the 0.3 second transmission time for a synchronous satellite.

Thus, it appears from the outset that very slow adjustments of clock frequency are permissible. We will see later that very slow adjustments are desirable as well. There is every reason to believe that we can disregard transmission time in considering the stability and performance of the synchronizing scheme which we discuss in Section VIII.*

VI. THE BUFFERS

We have noticed that at each center we may make use of the departure from the center frequency setting of the clock frequency

* A criterion for stability which is independent of transmission time has been available for some years.[3] In work to be published, I. W. Sandberg gives a less stringent criterion which is dependent on transmission time, and which is in accord with the qualitative statements made here.

adjustment and of the loading of the buffers, which we call b_1, b_2, and so on. These loadings may, with respect to their design or "normal" values, be positive or negative numbers.

Buffers will have finite capacity and hence will overflow when clock frequencies at two interconnected centers are persistently different. Here it is assumed that when a buffer overflows it remains completely full or empty until the sign of the difference in clock frequencies changes, at which time it again accepts and release pulses. It is also assumed that during the overflow condition the buffer reading b remains at some extreme value $\pm b_m$, and becomes smaller in magnitude once the sign of the difference in clock frequencies changes.

The bounds imposed on the b's by buffer overflow are valuable. The clock at the other end of a circuit may be in really bad trouble. In the extreme, no pulses may be coming in. Thus, we should not allow, or else we should disregard, buffer readings beyond some limiting extremes. As the buffers are finite, drastic malfunction will cause them to overflow and so limit the range of the buffer reading b.

VII. CLOCKS OF DIFFERING ACCURACY

In adjusting clock frequency, account should be taken of the accuracy of the clock. In a large system, it may be desirable to use very accurate master clocks at some important nodes and less accurate subsidiary clocks at other nodes. It is essential that some provision be made so that the final frequency of operation depends most strongly on the accurate master clocks and less on the less accurate subsidiary clocks. Thus, in adjusting the clocks, either the magnitude of the adjustments should be suitably smaller for the more accurate clocks than for the less accurate clocks, or else the adjustments of the more accurate clocks should be made less frequently or both.

VIII. MATHEMATICAL DESCRIPTION OF THE SYNCHRONIZING SCHEME

Let us now consider one particular node of the network of Fig. 1, at which the clock frequency is f_1. This node is shown as 1 in Fig. 3. Connected to it are nodes 2, 3, 4, . . . , n, where the frequencies are f_2, f_3, f_4, f_n. If the 1, n buffer is set to b_{1n0} at $t = 0$, the various buffer readings at 1 are

$$b_{1n} = b_{1n0} + \int_0^t (f_n - f_1)\, dt \tag{1}$$

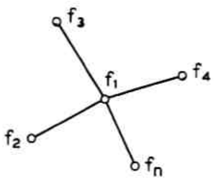


Fig. 3 — A node of the network of Fig. 1.

We will assume that b_{1n} can be positive or negative. If we let $b_{1n} = 0$ when the buffer is half full, then the buffer reading can have any value in some range from $-b_m$ to $+b_m$ where $2b_m$ is the size of the buffer.

Let the "center" clock frequency at 1 for "center setting" of the frequency control be f_{10}. The center settings are intended to adjust the clock to the desired system frequency. The frequency f_{n0} of a given clock at center setting differs from the intended frequency because of the error of the clock.

The buffer reading must be an integer. If the range of variation of this integer is great enough, it should be possible to treat the buffer reading as a continuous variable in a differential equation; this is what we will do.

A suitable linear strategy for adjusting the frequency f_1 was found to be

$$A \sum_n \int_0^t (f_n - f_1)\, dt - B_1(f_1 - f_{10}) - C \frac{df_1}{dt} + A \sum_n b_{1no} = 0. \quad (2)$$

We can regard the equation as applying either to control of the clock frequency or the rate of change of clock frequency. If we make $C = 0$, then the equation prescribes departure from center clock setting, $(f_1 - f_{10})$, in terms of buffer content. If $C \neq 0$, then the equation prescribes rate of change of clock setting, df_1/dt, in terms of buffer content and departure from center clock setting.

In either case B_1 should be larger for more accurate clocks and smaller for less accurate clocks. This will give the departure from center clock setting greater weight for more accurate clocks and less weight for less accurate clocks.

IX. AN ELECTRIC ANALOG

Let us now consider an electric analog, the circuit shown in Fig. 4. Node 1 is connected to ground through a capacitance C. Current

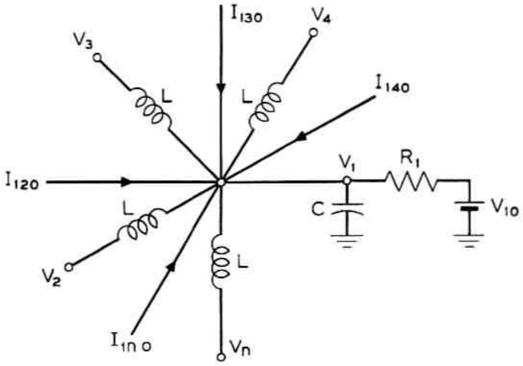


Fig. 4 — Electric analog of a node synchronization system considered.

flows to C through a resistance R_1 to which a bias voltage V_{10} is applied. Current also flows to node 1 because this node is connected to other nodes 2, 3, 4, . . . , n, at which the voltages are V_2, V_3, V_4, . . . , V_n, by inductances L. The differential equation for the voltage V_1 is

$$(1/L) \sum_n \int_0^t (V_n - V_1)\, dt - (1/R_1)(V_1 - V_{10}) - C\frac{dV_1}{dt} + \sum_n I_{1n0} = 0. \qquad (3)$$

Here I_{1n0} is a current flowing into node 1, not from but associated with node n, at time $t = 0$. We see that equation (3) is identical with equation (2) if we let

$$\left.\begin{aligned} V_n &= f_n \\ V_{10} &= f_{10} \\ (1/L) &= A \\ (1/R_1) &= B_1 \\ I_{1n0} &= Ab_{1n0} \\ C &= C \end{aligned}\right\} \qquad (4)$$

The behavior of the frequency of a network of oscillators adjusted according to the strategy of equation (2) will be the same as the behavior of the voltage in an L, R, C network of the form shown in Fig. 5. We should notice that it is perfectly permissible to make

$C = 0$ in (2) or (3). A network analog for this case has been given by Brilliant.[5]

In a network such as that of Fig. 5, no matter how extensive or how interconnected, all the voltages settle down to some final value because of the damping (energy loss) of the resistors R_n. What is the final voltage? It must be such that the total of the currents flowing to all nodes is zero. This means that

$$0 = \sum_m \sum_n I_{nm0} + \sum_n (V_{n0} - V)/R_n \, . \tag{5}$$

The double summation is in each case over all nodes. Not all nodes are connected to one another, and I_{nm0} will be zero for $n = m$ and for all nodes n not connected to m.

The analogous expression for final frequency is

$$0 = A \sum_m \sum_n b_{nm0} + \sum_n B_n(f_{n0} - f). \tag{6}$$

Again, b_{nm0} will be zero for $n = m$ and for all nodes n not connected to m.

We will see that the final frequency depends on the center frequencies of the oscillators, on the A and B coefficients, and on the

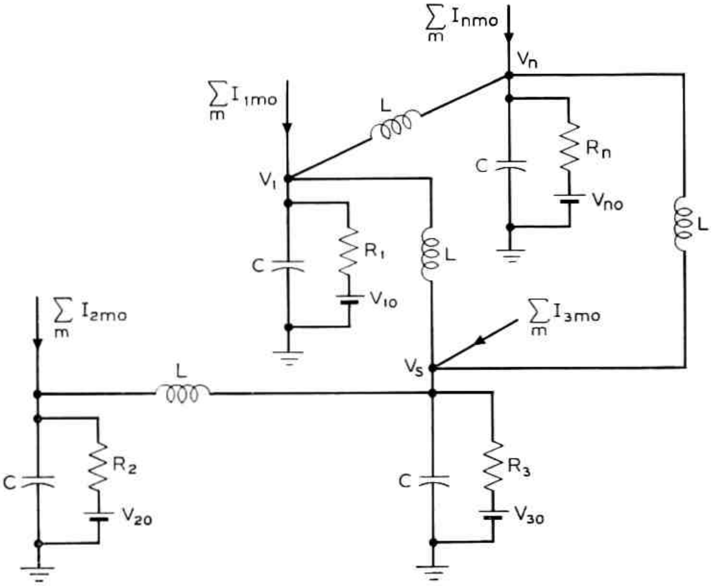


Fig. 5 — Interconnected nodes of electric analog.

initial buffer settings. If the initial buffer settings b_{nm0} are all zero, the final frequency depends only on the f_{n0}'s and the B_n's. If the buffers overflow, this in effect resets them.

We can see from Fig. 5 that the system will be stable in the case of certain nonlinearities.

For example, all the R_n's can be nonlinear as long as no resistance is ever negative at any current. Thus, if in the equations of the oscillator system the second term on the right of (2) is replaced by a nonlinear function of $(f_1 - f_{10})$, the system will be stable as long as the term decreases monotonically with increasing f_1. In the case $C = 0$, this corresponds to a nonlinear control of clock setting as a function of buffer content.

It is easy to show that in a special case buffer overflow cannot result in instability. This is the case in which the buffer readings at two ends of a transmission circuit are complementary, that is, their sum is zero. Disregarding changes in transmission time, if the buffers are both set to zero or to complementary values at the same time, or if both buffers overflow and then recover, the readings will be complementary.

Appendix A shows that for this case, in the electric analog of Figs. 4 and 5 buffer overflow results in the dissipation of energy, and this convinces the writer that in this case overflowing buffers cannot make the network unstable. The writer is mortified that he is unable to demonstrate this for the case of non-complementary readings, but he suspects that buffer overflow will not result in instability in this case, either.

Buffer overflow would affect the final frequency of operation. Further, all the buffers at a given node can conceivably overflow permanently (if the clock center rate shifts drastically, for example). In such a case, the node will operate out of synchronism with the rest of the network. This will cause buffer overflows at nodes connected to the asynchronous node. Such overflows can affect the frequency of operation of the rest of the network, but need not prevent its synchronous operation.

If widespread buffer overflow is avoided, adjustment according to equation (2) will result in stable operation.

X. PARAMETERS AND TIME CONSTANTS

Let us consider equation (2) with $C = 0$. By using (1), this can be written

$$f_1 - f_{10} = (1/T_1) \sum_n b_{1n} \tag{7}$$

$$T_1 = B_1/A = L/R_1 . \tag{8}$$

We see that T_1 is a time constant. Equation (7) prescribes the departure of the clock frequency from center setting in terms of the sum of buffer readings b_{1n} for the various transmission circuits terminating at node 1. How shall we choose the parameter T_1?

The smallest amount by which the sum of the buffer readings can change is unity. The frequencies f_1 and f_{10} differ by a very small amount. Thus, the fractional change in frequency caused by unit change in the sum of the buffer readings can be written $1/T_1f_1$. If we are to take full advantage of the stability of the clock at the node, we should make this smallest change small compared with the clock stability expressed as a fraction, which we call d_1. Hence, we should choose

$$1/T_1f_1 < d_1 , \qquad T_1 > 1/d_1f_1 . \tag{9}$$

If this is not so, changes in buffer readings will cause sudden changes in frequency larger than the changes which would occur if the clock were not adjusted.

The buffer readings must, however, be able to change the frequency of the clock by several times its fractional stability d_1 if we are to be sure to bring all the clocks to the common intended frequency. Strictly, it might be possible to accomplish this if $(1/T_1f_1)\ b_m > d_1/n$, where b_m is the maximum buffer reading and n is the number of buffers. It would seem wise to choose b_m large enough so that this criterion is considerably exceeded. We might reasonably ask that

$$b_m/T_1f_1 > d_1 , \qquad b_m > T_1f_1d_1 . \tag{10}$$

As an example, let us consider a case in which

$$T_1 = 10/d_1f_1 \tag{11}$$

$$b_m = 10T_1f_1d_1 = 100. \tag{12}$$

Assume that $f_1 = 5 \times 10^8$. For various values of d_1, the computed values of T_1 are:

Fractional oscillator stability, d_1	*Stability (pulses per day)* 86,400 d_1	*Time constant,* T_1 *(seconds)*
10^{-8}	4.32×10^5	2
2×10^{-11}	864	1000 (17 minutes)
10^{-12}	43.2	20,000 (5.5 hours)

The time constants T_1 are large, implying that the time for the system to come to equilibrium is large. The time constants may not seem excessive, however, when we consider oscillator stability measured in pulses per day. It would be rash to try to adjust an exceedingly stable oscillator in too short a time. Further, time jitter in the received pulses (see Appendix D) and perhaps other quickly changing phenomena could cause undesirable changes in operating frequency if T_1 were made smaller.

XI. COPING WITH CHANGES IN TRANSMISSION TIME

We have not yet considered the effect of changes in transmission time. Changes in buffer reading have been ascribed to differences in frequency. But changes in transmission time can also cause changes in buffer reading.

Consider two interconnected nodes. If the clock at one speeds up, the buffer at the node will tend to empty and the buffer at the node to which the fast clock is connected will tend to fill. Thus, a change in clock rate will change the buffer readings at interconnected nodes in opposite sense.

Consider an increase in transmission time between nodes, for example, an increase in transmission time in both directions. Because of the increase in transmission time, more pulses will be stored in the lines and the buffer readings will decrease at both ends. Thus, changes in transmission time will change the buffer readings at interconnected nodes in the same senses.

If we can compare the buffer readings at two interconnected nodes, we can distinguish changes caused by changes in transmission time from changes caused by changes in clock frequencies. Once we identify a change in transmission time, we can correct for it by means of an adjustable delay between the transmission system and the buffer input. Such corrections have been provided in some synchronization schemes.[6] It is not clear, however, that an automatic system acting in the same way among all nodes is best in coping with changes in transmission time.

Another course would be to use no adjustable delay, and merely provide buffers large enough to accommodate changes in transmission time. Then, changes in transmission time would cause buffers to fill or empty. This would have some effect on system frequency. If the system included highly stable clocks, such changes would necessarily be very small.

What should be done about changes in transmission time depends on the magnitude of such changes and on how rapidly they occur. Unfortunately, adequate information is not available.

In cable and waveguide systems, transmission time can vary with temperature and with the gas pressure within the waveguide or cable. What is known is discussed in Appendix B. For a circuit 3,000 miles long, we might expect variations of more than a thousand pulses over the year. However, because cable is buried and waveguide would be, we would expect the transmission time to vary little during the day, or over a period of several days. Experience with the L4 system tends to confirm this.

If the short-term stability of transmission time of cable and waveguide is as good as would seem, it would be satisfactory to make compensating adjustments in delay at intervals of days or weeks; this would argue for a scheme of adjustment separate from that used for clock synchronization.

Variations in transmission time for microwave radio systems (see Appendix C) may be comparable to those for cable or waveguide systems, but changes may be more rapid. Diurnal changes might be taken care of by the buffers, and slower changes by daily delay adjustments.

A node may often be connected to the rest of the network by cable and/or waveguide circuits as well as by microwave circuits. In this case, it seems attractive to the writer to use the cable or waveguide circuits for adjusting clock rate. Then the buffers and delay adjustments associated with the microwave circuits could be used solely to compensate for changes in microwave transmission time.

Thus, a somewhat mixed strategy of adjustment may be called for. The following seems reasonable to the writer:

(*i*) If possible, avoid the use of microwave circuits for synchronization. Use parallel cable or waveguide circuits for synchronization, and use buffers or adjustments of delay to absorb changes in microwave transmission time.

(*ii*) The system will probably contain several master clocks which are more accurate than other subsidiary clocks. The parameters T_1 will be chosen [according to (9) or (11)] so that frequency is chiefly dependent on the center frequencies of the master clocks. Hence, the correct function of the buffer readings at the subsidiary clocks is to adjust these clocks to the (correct) operating frequency. At each node with a subsidiary clock, periodic adjustments of delay should be made, such as to make all buffer readings zero. At the same time a delay adjustment should be made to make any buffer at a master clock on a line from the subsidiary clock zero. Simultaneously, the subsidiary clock should be readjusted so that its center frequency f_{n0} is equal to the current operating frequency f_n. This adjusts for changes in transmission time by making buffer readings at the ends of links to and from subsidiary clocks complementary (and equal to zero). It may be desirable to make these adjustments at a common time at all nodes concerned. Notice that in making the adjustments at a node, no knowledge of buffer readings or clock settings at other nodes is needed. Such adjustments might be made once a day or once a week.

(*iii*) It is desirable that the network link all master clocks together by cable or waveguide, directly or indirectly, but with no intervening buffering and retiming by less accurate clocks. When this is so, it would seem desirable to make periodic adjustments of the delays in each transmission circuit connecting two master clocks, such as to render the buffer readings complementary at the two ends of the link. This compensates for changes in transmission time. It is undesirable to adjust the frequency at center setting, since we rely on the frequencies of the master clocks at center settings to determine the operating frequency. This adjustment of delays need be made only infrequently (once a day or once a week). To make it, we must know at each master clock the buffer readings at the other ends of the circuits connecting it to other master clocks.

(*iv*) In adding a subsidiary clock to a network, it should be adjusted so that its frequency at center setting is equal to the current operating frequency and the buffers at both ends of all circuits connecting it to the network should be set to zero. If a master clock is added, the adjustment of frequency at center setting should be omitted.

XII. CONCLUDING REMARKS

The behavior of a simple scheme of synchronization was investigated and found to be stable in the absence of buffer overload if clock adjustments are sufficiently slow compared with transmission time.

The criteria given for choice of parameters result in time constants very long in comparison with transmission time.

A strategy of infrequent periodic adjustment (once a day or once a week) has been suggested. This can compensate for changes in transmission time and correct the frequencies at center setting of subsidiary clocks. A strategy has been given for adding new clocks to the system. If these strategies are followed, and if parameters are chosen as prescribed, it seems likely that the buffers will not overload except in case of clock failure. Clock failure will cause loss of synchronization at a node.

It has been shown that in a special case, buffer overload from other causes will not result in instability; it seems plausible that this is so in general.

It appears that there are no inherent obstacles to the synchronization of large digital networks. In the practical realization of such networks it would be desirable to have more information concerning the variation of transmission time with time, and on the availability of suitable components, including:

(*i*) Adequate buffers which will accept pulses at one rate, emit them at another, and behave under overload in the fashion described earlier.

(*ii*) Adequate delay of the order of 10^5 pulses, to bring frames into coincidence.

XIII. ACKNOWLEDGMENTS

The writer is indebted to I. W. Sandberg, J. C. Candy, and J. R. Gray for valuable comments and suggestions, to S. L. Freeny and S. E. Miller for information in Appendix B, and to D. C. Hogg for most of the material in Appendix C.

APPENDIX A

Buffer Overflow

The purpose of this section is to consider the consequences of buffer overflow by studying the behavior of the electrical analog.

Consider equation (1):

$$b_{1n} = b_{1n0} + \int_0^t (f_n - f_1)\,dt. \tag{1}$$

Because of the finite content of the buffer, if for $b_{1no} = 0$ the buffer is set half full at $t = 0$, we must have

$$-b_m < b_{1n} < b_m\,, \qquad | b_{1n} | < b_m \tag{13}$$

One satisfactory way to treat buffer overflow is to let b_{1n0} of (1) change during time intervals when the integral would otherwise cause an increasing violation of (13). For example, suppose that the integral is increasing with time. When b_{1n} reaches the value b_m, b_{1n0} starts to decrease and decreases so as to make $b_{1n} = b_m$ for as long as the integral continues to increase. As soon as the integral starts to decrease, b_{1no} remains constant (until another buffer overflow) at whatever value it had when the integral started to decrease, and b_{1n} starts to decrease. A little thought shows that this results in just the behavior that an overflowing buffer of the type described in the paper would exhibit.

In equations (3) and (4) and in Fig. 4 we see that the exact analog of b_{1n} is LI_{1n} given by

$$LI_{in} = L\left[I_{1n0} + (1/L)\int_0^t (V_n - V_1)\,dt\right]. \tag{14}$$

In exploring the effect of buffer overload it is sufficient to consider a circuit element consisting of an inductor and two bias currents. For simplicity we will assume that these bias currents are equal and opposite, with magnitudes I_o, as shown in Fig. 6. The input and output currents are thus equal and of magnitude I. The input and output voltages are V_2 and V_1. The current I_L through the inductance L is

$$I_L = I + I_0 . \tag{15}$$

Assume the same buffer overload current at each end, of magnitude I_m. Thus, the magnitude of the current I cannot become greater than I_m.

Suppose that $I = 0$ at $t = 0$ and the voltages are such as to increase the magnitude of I. How much energy have we put into the circuit by the time $I = I_m$? This energy E_m is

$$E_m = \int_{I=0}^{I_m} (V_2 - V_1)I\,dt. \tag{16}$$

Now

$$V_2 - V_1 = L\frac{dI_L}{dt} = L\frac{d}{dt}(I + I_o) = L\frac{dI}{dt}. \tag{17}$$

Hence

$$E_m = L\int_0^{I_m} I\,dI = (1/2)LI_m^2 . \tag{18}$$

Notice that E_m does not depend on I_o and that it is recoverable, that is, we get it all back if we change the current from I_m to 0 in such a manner that the magnitude of I is always less than I_m.

The voltage difference $V_2 - V_1$ can cause the current I_L to change even after I has reached its limiting magnitude I_m. I_o must then change to keep the magnitude of I from exceeding I_m. In this regime of buffer overload,

$$I = I_m \tag{19}$$

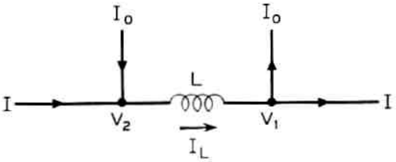


Fig. 6 — Electric analog of buffer, used in studying buffer overflow.

$$I_L = I_o + I_m \ . \tag{20}$$

What about the energy E that is supplied to the circuit during this period? This energy is

$$\begin{aligned} E &= \int_{t_1}^{t_2} (V_2 - V_1) I_m \, dt \\ &= \int_{t_1}^{t_2} L \frac{dI_L}{dt} I_m \, dt \\ &= LI_m(I_{L2} - I_{L1}) \\ &= LI_m(I_{o2} - I_{o1}). \end{aligned} \tag{21}$$

Overload operation persists only as long as I_o is increasing. If we come to a point where I_o would decrease, we hold I_o constant. The buffer is no longer overloaded and we return to the regime of a fixed bias current.

Thus,

$$I_{02} > I_{01} \tag{22}$$

$$E > 0. \tag{23}$$

In fixed bias operation, the recoverable energy is merely E_m as given by (18). The positive energy E has been dissipated. Hence, the effect of buffer overload is to dissipate energy, and buffer overload cannot result in instability.

APPENDIX B

Cable and Waveguide

The transmission time through cable or waveguide can change for several reasons. The gas pressure in a cable is controlled, and changes in gas pressure cause changes in transmission time, as could changes in gas temperature. While large changes in pressure could produce large effects (see Appendix C), it seems likely that another effect will dominate.

This is the linear expansion of the cable or waveguide because of changes in temperature. Structurally, waveguide would probably be largely steel; cable might be considered as copper. The thermal coefficient of expansion of steel is about 12×10^{-6}; for copper it is about 18×10^{-6}. For a change in temperature of 40°F or 22°C, the fractional change FC in length would be approximately

Material	*Fractional change in length, FC, for 22°C change in temperature*
Steel	260×10^{-6}
Copper	400×10^{-6}

During an experiment on the L-4 field installation in Dayton, Ohio, T. J. Pedersen observed phase shift versus time of a 12 MHz sine wave sent through an 84-mile loop of the L-4 system. During a 24-hour period he measured a peak-to-peak change of about 1.5°. The velocity of propagation is about 175,000 miles per second, so the total phase shift in the 84-mile loop was about 2×10^{6} degrees. This is a fractional change of

$$0.75 \times 10^{-6}.$$

During the time that this change took place, the temperature of the cable was changing about 0.3°F per day. If this temperature change was indeed the source of the phase shift, the fractional change in transmission time for a 40°F change in temperature would be

$$\text{FC} = (0.75 \times 10^{-6})(40/0.3) = 100 \times 10^{-6}.$$

This value may not be accurate, but an estimate based on the thermal expansion of a metal may not be accurate either.

Change in temperature increases the resistivity of copper, and this causes a change in the reactive as well as the resistive component of skin impedance. Further, the capacitance may change with temperature. Both calculations and measured values of cable parameters as a function of temperature indicate that such effects will change transmission time much less than linear expansion.

Change in diameter of a waveguide causes a change in group velocity. The change in transmission time with temperature which this would cause is considerably smaller than a change proportional to linear expansion.

It may be conjectured whether cable or waveguide is or need to be free to expand in length as temperature changes. We have no waveguide systems in operation at present, and variations of transmission time for coaxial cable systems have not been adequately measured. We can only conclude:

(*i*) Change in transmission time will be very slow, so that infrequent adjustments would be satisfactory.

(*ii*) Total fractional changes in transmission time may be from 100–400 parts per million.

For a path length of 3,000 miles, a velocity of 175,000 miles per second and a pulse rate of 5×10^8 pulses per second, the number of pulses stored in the line would be 10^7. For the fractional changes in transmission time quoted above, the changes in number of pulses stored in a 3,000 mile cable would be

Fractional change in transmission time	*Change in number of pulses stored*
0.75×10^{-6} (observed change in one day)	7.5
100×10^{-6} (estimated effect of 40°*F* change in temperature)	1000
220×10^{-6} (from expansion of steel caused by 40°F change in temperature)	2200
400×10^{-6} (from expansion of copper caused by 40°F change in temperature)	4000

APPENDIX C

Microwave Radio

The velocity of radio waves traveling through the atmosphere depends on temperature, pressure, humidity and probably on rain.

The effect of the first three factors is given in terms of N units. In terms of the index of refraction n (ratio of velocity in vacuo to velocity in the medium)

$$N = (n - 1)10^6.$$

A simple expression gives N quite accurately[11]

$$N = \frac{77.6}{T}\left(p + 4{,}810\,\frac{e}{T}\right)$$

p = total pressure in millibars
e = partial pressure of water vapor in millibars
T = absolute temperature °K = °C + 273.

Here we are concerned with rough estimates of changes in N over short and long periods.

The diurnal change in temperature may account for the most rapid fluctuations in N. For a constant pressure and disregarding water vapor, N is inversely proportional to absolute temperature. For a 20°C (36°F) change in temperature around 20°C (68°F), at a pressure of one bar the change in N would be about 18.

Data taken over a six-year period from forty-five U. S.[12] weather stations show that the monthly mean value of N at the earth's surface (N_s) varies from 230 to 400 over the country and through the year. In the U. S. the largest local variation in the monthly means is on the southeast and Gulf Coasts and amounts to a charge of N of 50 units.

During a given month, the variation in N_s (for one and ninety-nine percent probability) can be as much as 100 N units.

Concerning rain, it has been calculated that 150mm per hour rain over a 1 km path will introduce a phase shift of some 500 degrees at 30 GHz.[13] This is equivalent to a change in N of 14 units.

It is not easy to arrive at a reasonable estimate of short-term and long-term changes in the average value of N over a long transmission system. On the basis of the foregoing data, it appears to the writer that changes during the day would probably not exceed 20 units, while changes during the month might be as much as 100 units, and changes during the year might be several hundred units.

If we assume a 3000 mile path, the nominal transmission time is about 0.016 second, and at a pulse rate of 5×10^8 pps the number of pulses in the transmission system will be 8×10^6. The change in this number of pulses will be the number times N times 10^{-6}. For some values of change in N, the change in number of pulses will be

Change in N	*Change in number of pulses*
20	160
100	800
200	1600

APPENDIX D

Jitter Due to Regenerative Repeaters

In an experimental digital repeater line[1] the systemic jitter introduced by a single repeater had an rms value θ_1 of about 3.3°. The rms jitter after N repeaters, θ_N, is given by[14]

$$\theta_N = \theta_1 \sqrt{\frac{P(N)}{P(1)}} \tag{24}$$

$$P(N) = \frac{N}{2} - \frac{(2N - 1)!}{4^N[N - 1!]^2} \tag{25}$$

The function $P(N)$ has been tabulated:[15] $P(1) = 0.250$ and for $N > 100$, $P(N) = N/2$. Thus, for a large $N > 100$ number of repeaters,

$$\theta_n = \theta_1 \sqrt{2N} \tag{26}$$

Some computed values of θ_n are:

Number of repeaters	*θ_n, rms phase jitter, (degrees)*
1	3.3
100	47
300	81
1000	148
3000	256

REFERENCES

1. Karnaugh, M., "A Model for the Organic Synchronization of Communications Systems," B.S.T.J., *45*, No. 10 (December 1966), pp. 1705–1735.
2. Bosworth, R. H., Kammerer, F. W., Rowlinson, D. E., and Scattaglia, J. V., "Design of a Simulator for Investigating Organic Synchronization Systems," B.S.T.J., *47*, No. 2 (February 1968), pp. 209–226.
3. Gersho, A., and Karafin, B. J., "Mutual Synchronization of Geographically Separated Oscillators," B.S.T.J., *45*, No. 10 (December 1966), pp. 1689–1704.
4. Brilliant, M. B., "The Determination of Frequency in Systems of Mutually Synchronized Oscillators," B.S.T.J., *45*, No. 10 (December 1966), pp. 1737–1748.
5. Brilliant, M. B., "Dynamic Response of Systems of Mutually Synchronized Oscillators," B.S.T.J., *46*, No. 2 (February 1967), pp. 319–356.
6. Candy, J. C., and Karnaugh, M., "Organic Synchronization: Design of the Controls and Some Simulation Results," B.S.T.J., *47*, No. 2 (February 1968), pp. 227–259.
7. Inose, H., Fujisaki, H., and Saito, T., "Theory of Mutually Synchronised Systems," Electronics Letters, *2*, No. 3 (March 1966), pp. 96–97.
8. Inose, H., Fujisaki, H., and Saito, T., "System Design of a Mutually Synchronised System," Electronics Letters, *3*, No. 1 (January 1967), pp. 15–16.
9. Inose, H., Fujisaki, H., and Saito, T., "Phase Relation between Offices in a Mutually Synchronized System," Electronics Letters, *3*, No. 6 (June 1967), pp. 243–244.

10. Throne, Darwin H., "A Rubidium-Vapor Frequency Standard for Systems Requiring Superior Frequency Stability," Hewlett-Packard J., *19,* No. 10 (June 1968), pp. 8–14.
11. Smith, E. K., Jr. and S. Weintraub, "The Constants in the Equation for Atmosphere Refractive Index at Radio Frequencies," Proc. IRE, *41,* No. 8 (August 1953), pp. 1035–1037.
12. Bean, B. A., and Dutton, E. J., "Radio Meteorology," NBS Monograph, *92* (March 1966), pp. 63, 100, and 416.
13. Hogg, D. C., private communication.
14. Dorros, I., Sipress, J. M., and Waldhauer, F. D., "An Experimental 224 Mb/s Digital Repeater Line," B.S.T.J., *45,* No. 7 (September 1966), pp. 993–1043.
15. Byrne, C. J., Karafin, B. J., and Robinson, D. B., Jr., "Systematic Jitters in a Chain of Digital Repeaters," B.S.T.J., *42,* No. 6 (November 1963), pp. 2679–2714.

Extension of Bode's Constant Resistance Lattice Synthesis of Transfer Impedance Function*

By S. Y. LEE

(Manuscript received August 30, 1968)

Bode developed some explicit formulas in terms of the poles and zeros of the transfer impedance function for each element of the first and second degree constant resistance lattice structures. To extend work in this area, we derive explicit formulas for two of Bode's structures using coupled coils; we give two new structures which avoid coupled coils. Illustrative examples show the usage of these formulas. Finally, we include a general procedure for synthesizing any physically realizable, rational transfer impedance function by a constant resistance lattice network. A flow chart aids in detailing this procedure.

With the addition of these results, a general method for synthesizing any physically realizable, rational transfer impedance function with explicit formulas is complete. The explicit formulas method developed in this paper gives more rapid results and introduces fewer round-off errors than the step-by-step procedures used in the past.

I. INTRODUCTION

An important characteristic of constant resistance lattice networks is the absence of reflection effects when such two-port lattice networks are connected in tandem. The synthesis of a given transfer impedance function is simplified by representing the function by a partial product expansion. Thus the transfer impedance may be represented by a tandem connection of a number of constant resistance structures (one for each partial product). This process will result in

[*] Some results presented in this paper are based upon the author's thesis, "Explicit Formulas for Constant Resistance Lattice Synthesis of Transfer Impedance," presented to the Moore School of Electrical Engineering, University of Pennsylvania in December 1965 in partial fulfillment of the requirements for the degree of Master of Science in engineering.

realizable, simpler, transfer impedance functions provided that the constant multiplier of the given transfer impedance function is made large enough to permit each of the constituent networks to have nonnegative loss on the real axis.[1] In general, then, there will be additional fixed loss for the overall two-port network.

For physical realizability, it is required that both members of any conjugate complex pair of zeros or poles be retained within a given partial product. Hence, each of the elementary constituent networks must be represented by a biquadratic factor. When there are single zero and single pole pairs on the σ axis, the partial product factor for each pair is reduced to the bilinear form. It is recognized that the expansion can be performed with the zeros and the poles collected in a variety of ways and assigned to the individual networks. The elementary lattice networks for these bilinear and biquadratic factors are first and second degree constant resistance lattice structures, respectively. Therefore, one can realize a complicated rational transfer impedance function, to within a constant loss, by a combination of elementary structures in tandem.

Bode[1] developed the basic first and second degree structures, which are given in Fig. 2. They cover all the possible pole-zero combinations. Furthermore, he derived the explicit formula for each element of structures I to VI in terms of the poles and the zeros of the transfer impedance function. The object of this paper is to extend work in this area. Explicit formulas are obtained for structures VII and VIII and for two additional structures (Fig. 7). The structures of Fig. 7 avoid coupled coils. Illustrative examples are given to show the usage of these formulas. Finally, a general procedure for synthesizing any physically realizable, rational transfer impedance function by a constant resistance lattice network is included. This procedure is detailed with the aid of a flow chart. The appendix gives a method of obtaining the physical realizability conditions for one of the struc-

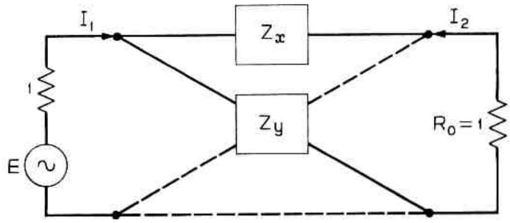


Fig. 1 — General constant resistance lattice network.

tures as an illustration. Reference 2 supplies the derivations of the physical realizability conditions of other structures.

II. DEVELOPMENT OF SECOND DEGREE CONSTANT RESISTANCE LATTICE STRUCTURE INTO GENERAL FORMULAS

The transfer impedance function exp θ of the constant resistance lattice given in Fig. 1 of second degree can be written as the biquadratic factor

$$\exp\,\theta = \frac{E_1}{2I_2} = K\frac{(s - a_1)(s - a_2)}{(s - b_1)(s - b_2)} \tag{1}$$

where exp θ is related to the series branch impedance z_x by the expression

$$z_x = \frac{[\exp\,\theta] - 1}{[\exp\,\theta] + 1} \quad \text{for} \quad z_x z_y = R_0^2 = 1 \tag{2}$$

or

$$\exp\,\theta = \frac{1 + z_x}{1 - z_x}. \tag{3}$$

Hence z_x is a biquadratic of the form

$$z_x = \frac{A_5 s^2 + A_3 s + A_1}{A_6 s^2 + A_4 s + A_2} \tag{4}$$

The solution for A_j's can be expressed in terms of the zeros and the poles of the transfer impedance function exp θ by setting (1) and (3) equal

$$K\frac{s^2 - (a_1 + a_2)s + a_1 a_2}{s^2 - (b_1 + b_2)s + b_1 b_2} = \frac{(A_6 + A_5)s^2 + (A_4 + A_3)s + (A_2 + A_1)}{(A_6 - A_5)s^2 + (A_4 - A_3)s + (A_2 - A_1)}. \tag{5}$$

For convenience let

$$\alpha_1 = a_1 + a_2\,, \qquad \alpha_2 = a_1 a_2 \tag{6}$$

and

$$\beta_1 = b_1 + b_2\,, \qquad \beta_2 = b_1 b_2\,. \tag{7}$$

Then equating coefficients in (5) and solving for A_j's yields

$$A_1 = \tfrac{1}{2}[\alpha_2(A_6 + A_5) - \beta_2(A_6 - A_5)] \tag{8}$$

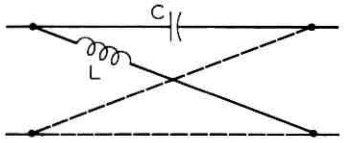
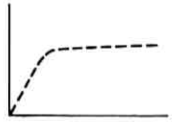
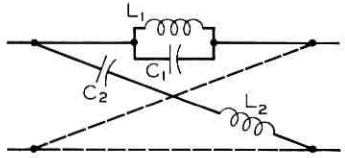
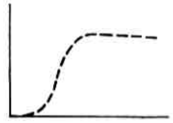
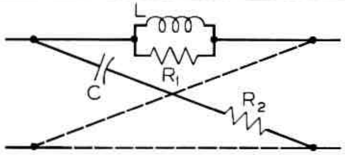
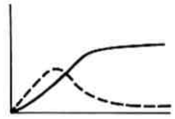
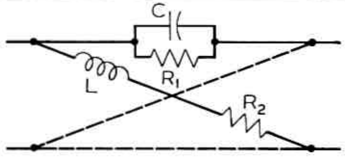
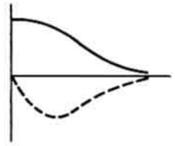
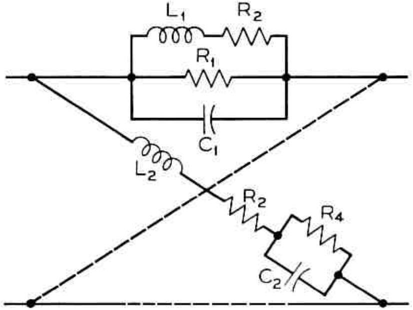
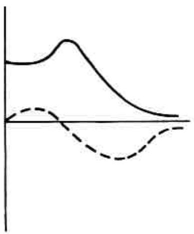
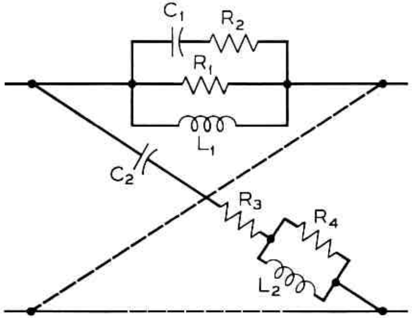
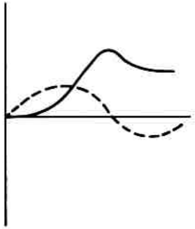
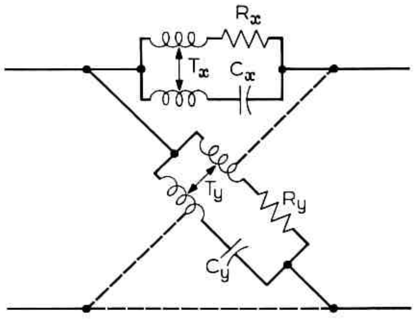
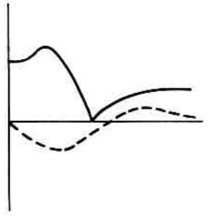
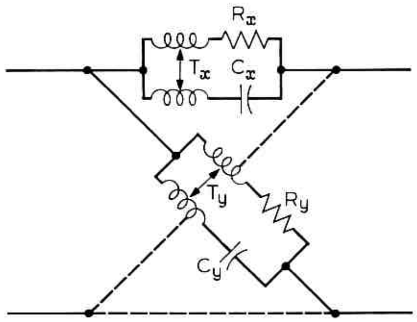
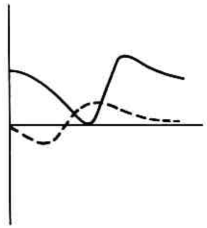
	STRUCTURE	GENERAL FORM OF EXP θ	REQUIREMENTS FOR PHYSICAL REALIZABILITY	TYPICAL ATTENUATION PHASE CHARACTERISTICS
I		$\frac{s-a}{s-b}$	$a = -b$	
II		$\frac{(s-a_1-ja_2)(s-a_1+ja_2)}{(s-b_1-jb_2)(s-b_1+jb_2)}$	$a_1 = -b_1$ $a_2 = -b_2$	
III		$\frac{K(s-a)}{(s-b)}$	$\lvert a\rvert \le \lvert b\rvert$	
IV		$\frac{K(s-a)}{(s-b)}$	$\lvert a\rvert \ge \lvert b\rvert$	

Fig. 2 — Constant resistance lattice equalizers.

	STRUCTURE	GENERAL FORM OF EXP θ	REQUIREMENTS FOR PHYSICAL REALIZABILITY	TYPICAL ATTENUATION PHASE CHARACTERISTICS
V		$\dfrac{K(s-a_1)(s-a_2)}{(s-b_1)(s-b_2)}$	$\lvert a_1 a_2\rvert \geq \lvert b_1 b_2\rvert$ $b_1^2+b_2^2 \leq a_1^2+a_2^2$ THE b's ARE COMPLEX THE a's MAY BE REAL OR COMPLEX	
VI		$\dfrac{K(s-a_1)(s-a_2)}{(s-b_1)(s-b_2)}$	$\lvert a_1 a_2\rvert \leq \lvert b_1 b_2\rvert$ $\dfrac{1}{b_1^2}+\dfrac{1}{b_2^2} \leq \dfrac{1}{a_1^2}+\dfrac{1}{a_2^2}$ THE b's ARE COMPLEX THE a's MAY BE REAL OR COMPLEX	

	STRUCTURE	GENERAL FORM OF EXP θ	REQUIREMENTS FOR PHYSICAL REALIZABILITY	TYPICAL ATTENUATION PHASE CHARACTERISTICS
VII		$\frac{K(s-a_1)(s-a_2)}{(s-b_1)(s-b_2)}$	$\lvert a_1 a_2 \rvert \geq \lvert b_1 b_2 \rvert$ $b_1^2 + b_2^2 \geq a_1^2 + a_2^2$ THE a's ARE COMPLEX THE b's MAY BE REAL OR COMPLEX	
VIII		$\frac{K(s-a_1)(s-a_2)}{(s-b_1)(s-b_2)}$	$\lvert a_1 a_2 \rvert \leq \lvert b_1 b_2 \rvert$ $\frac{1}{b_1^2} + \frac{1}{b_2^2} \geq \frac{1}{a_1^2} + \frac{1}{a_2^2}$ THE a's ARE COMPLEX THE b's ARE REAL OR COMPLEX	

$$A_2 = \tfrac{1}{2}[\alpha_2(A_6 + A_5) + \beta_2(A_6 - A_5)] \tag{9}$$

$$A_3 = \tfrac{1}{2}[\alpha_1(A_6 + A_5) - \beta_1(A_6 - A_5)] \tag{10}$$

$$A_4 = \tfrac{1}{2}[\alpha_1(A_6 + A_5) + \beta_1(A_6 - A_5)] \tag{11}$$

$$K = \frac{A_6 + A_5}{A_6 - A_5}. \tag{12}$$

Expressing exp θ on the real frequency axis we have

$$\exp(\alpha + j\beta) = \frac{K[(a_1a_2 - \omega^2) - (a_1 + a_2)j\omega]}{[(b_1b_2 - \omega^2) - (b_1 + b_2)j\omega]} \tag{13}$$

where $\theta = \alpha + j\beta$ may be called the transfer loss and phase. From this the expression for the transfer loss is obtained as

$$\exp(2\alpha) = \frac{K^2[(-\omega^2 + a_1a_2)^2 + (a_1 + a_2)^2\omega^2]}{(-\omega^2 + b_1b_2)^2 + (b_1 + b_2)^2\omega^2} \tag{14}$$

by letting

$$k = K^2 \qquad \text{and} \qquad x = \omega^2 \tag{15}$$

and from (6) and (7), equation (14) becomes

$$\exp(2\alpha) = \frac{k(\alpha_2 - x)^2 + \alpha_1^2 x}{(\beta_2 - x)^2 + \beta_1^2 x}. \tag{16}$$

It can be shown that in general the attenuation characteristic of a lattice for which exp θ is a biquadratic function exhibits a minimum at a real frequency.[1] One can shift this minimum loss to have zero loss at that particular frequency; thus the transfer impedance obtained will be within a constant loss. By doing this the attenuation characteristics of all elementary structures will have zero transfer loss at one frequency ω_0. Corresponding to ω_0, exp $(2\alpha) = 1$. Thus k can be determined in terms of the zeros and the poles of the transfer impedance function by (16). If A_6 is equated to unity, A_5 is obtained by (12). With the relationships (8) to (11) one can determine A_j's in terms of the zeros and the poles of the transfer impedance function. Hence from (4) and (2) z_x and z_y can be obtained respectively.

III. EXPLICIT FORMULAS FOR STRUCTURE VII

The physical realizability conditions and the typical attenuation characteristic of this structure are given in Fig. 2. Zero attenuation occurs at a frequency ω_0; therefore we must choose k such that the attenuation becomes zero at the same frequency.

For zero attenuation, exp (2α) must be equal to 1; thus from (16) after rearranging and letting $\omega_0^2 = x$, we obtain a quadratic in x_0

$$(k - 1)x_0^2 + (k\alpha_1^2 - 2\alpha_2 k + 2\beta_2 - \beta_1^2)x_0 + (k\alpha_2^2 - \beta_2^2) = 0 \tag{17}$$

which can be written compactly as

$$ax_0^2 + bx_0 + c = 0 \tag{18}$$

where

$$a = k - 1 \tag{19}$$

$$b = (\alpha_1^2 - 2\alpha_2)k + 2\beta_2 - \beta_1^2 \tag{20}$$

$$c = k\alpha_2^2 - \beta_2^2 . \tag{21}$$

In order that the frequency be real and the attenuation equal to zero, the solution of (18) must have a double root at x_0. This condition holds only when the discriminant $b^2 - 4ac = 0$. First we find from (18)

$$x_0 = \frac{-b}{2a}. \tag{22}$$

Secondly we obtain the following quadratic in k

$$Ak^2 + 2Bk + C = 0 \tag{23}$$

where

$$A = \alpha_1^2(\alpha_1^2 - 4\alpha_2) \tag{24}$$

$$B = (\alpha_1^2 - 2\alpha_2)(2\beta_2 - \beta_1^2) + 2(\beta_2^2 + \alpha_2^2) \tag{25}$$

$$C = \beta_1^2(\beta_1^2 - 4\beta_2). \tag{26}$$

It can be shown that the larger root of (23) must be used to insure that z_x is a positive real function.* Denote this larger real root by k_m. Thus, from (15) and (23)

$$k_m = K_m^2 = \left\{\frac{-2B \pm [(2B)^2 - 4AC]^{\frac{1}{2}}}{2A}\right\}_{\max} \tag{27}$$

Hence K_m can be obtained quite easily and it is in terms of the zeros and the poles of exp θ.

* Notice that k_m as well as the discriminant of (23) must be positive and real because of the physical realizability conditions. By considering all the possible sign combinations for A, B and C, one of the positive roots of K_m is greater than, or equal to, unity.

By substituting K_m into (12) and letting $A_6 = 1$ we obtain

$$A_5 = \frac{K_m - 1}{K_m + 1}. \tag{28}$$

Notice that from (12) and (28), the root K_m must be greater or equal to unity in order for z_x to be positive real functions.

Using (8), (9), (10) and (11) we can determine A_1, A_2, A_3 and A_4. Then multiplying each coefficient by $K_m + 1$ we get

$$A_1 = \alpha_2 K_m - \beta_2 \tag{29}$$

$$A_2 = \alpha_2 K_m + \beta_2 \tag{30}$$

$$A_3 = \beta_1 - K_m \alpha_1 \tag{31}$$

$$A_4 = -\beta_1 - K_m \alpha_1 \tag{32}$$

$$A_5 = K_m - 1 \tag{33}$$

$$A_6 = K_m + 1. \tag{34}$$

Thus the coefficients of z_x and K_m are expressed implicitly in terms of the zeros and the poles of the transfer impedance function exp θ. Furthermore from (33) and (34), we must have the positive root $K_m > 1$ for A_j's to be positive.

Before considering the realization of the biquadratic z_x in (4) with its coefficients given from (29) to (34), we will show that z_x is a minimum resistance function.

Rewriting (3) as

$$\exp \theta = \frac{1 + z_x}{1 - z_x} = \frac{(1 + R_x) + jX_x}{(1 - R_x) - jX_x} \tag{35}$$

where

$$z_x = R_x + jX_x , \tag{36}$$

the corresponding magnitude is

$$\exp (2\alpha) = \frac{(1 + R_x)^2 + X_x^2}{(1 - R_x)^2 + X_x^2}. \tag{37}$$

Since we require that

$$\exp (2\alpha) = 1 \text{ at } \omega_0 \tag{38}$$

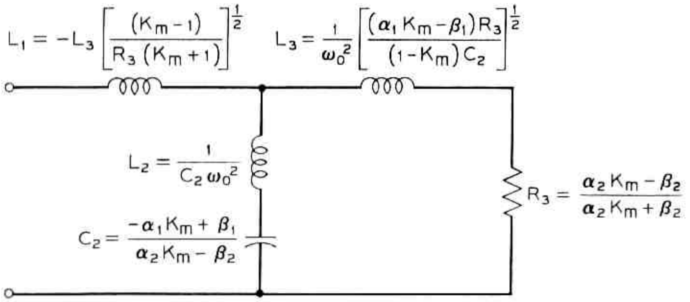
then R_x must equal zero and hence z_x must be a minimum resistance function.

Now we can determine the element values for structure VII by using the results given in Chapter 4 of Boghosian and Bedrosian, in that the element values of a Brune network were expressed explicitly in terms of the coefficients of a minimum resistance biquadratic impedance function.[3] Since we have shown that the biquadratic function in (4) is also minimum resistance, it is a simple matter to express the element values in terms of the zeros and the poles of exp θ. The case when the coefficients of z_x satisfy the inequality

$$[A_5A_2]^{\frac{1}{2}} - [A_1A_6]^{\frac{1}{2}} < 0 \tag{39}$$

is suitable for structure VII. Thus the Brune network for the series arm z_x of structure VII is shown in Fig. 3, where the equivalent-T is used instead of the transformer. Since $z_xz_y = 1$, we have for the cross arm of these lattices

$$z_y = \frac{A_6s^2 + A_4s + A_2}{A_5s^2 + A_3s + A_1}. \tag{40}$$


$L_1 = -L_3\left[\frac{(K_m-1)}{R_3(K_m+1)}\right]^{\frac{1}{2}}$
$L_3 = \frac{1}{\omega_0^2}\left[\frac{(\alpha_1K_m-\beta_1)R_3}{(1-K_m)C_2}\right]^{\frac{1}{2}}$
$L_2 = \frac{1}{C_2\omega_0^2}$
$C_2 = \frac{-\alpha_1K_m+\beta_1}{\alpha_2K_m-\beta_2}$
$R_3 = \frac{\alpha_2K_m-\beta_2}{\alpha_2K_m+\beta_2}$

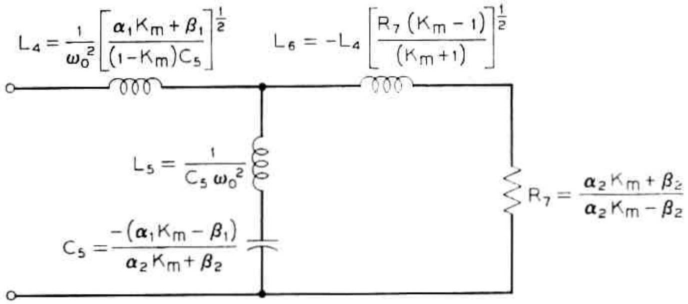
$$\omega_0^2 = \left[\frac{(\alpha_2K_m)^2-\beta_2^2}{K_m^2-1}\right]^{\frac{1}{2}} = \left[\frac{(a_1a_2K_m)^2-(b_1b_2)^2}{K_m^2-1}\right]^{\frac{1}{2}}$$

$$L_3 = \frac{1}{\omega_0^2}\left[\frac{[(a_1+a_2)K_m-(b_1+b_2)]R_3}{(1-K_m)C_2}\right]^{\frac{1}{2}}$$

$$C_2 = -\frac{(a_1+a_2)K_m+(b_1+b_2)}{a_1a_2K_m-b_1b_2}$$

$$R_3 = \frac{a_1a_2K_m-b_1b_2}{a_1a_2K_m+b_1b_2}$$

Fig. 3 — Series arm z_x for structure VII of Fig. 2.


$L_4 = \frac{1}{\omega_0^2}\left[\frac{\alpha_1 K_m + \beta_1}{(1-K_m)C_5}\right]^{\frac{1}{2}}$
$L_6 = -L_4\left[\frac{R_7(K_m-1)}{(K_m+1)}\right]^{\frac{1}{2}}$
$L_5 = \frac{1}{C_5\,\omega_0^2}$
$R_7 = \frac{\alpha_2 K_m + \beta_2}{\alpha_2 K_m - \beta_2}$
$C_5 = \frac{-(\alpha_1 K_m - \beta_1)}{\alpha_2 K_m + \beta_2}$

$$\omega_0^2 = \left[\frac{(a_1 a_2 K_m)^2 - (b_1 b_2)^2}{K_m^2 - 1}\right]^{\frac{1}{2}}$$

$$L_4 = \frac{1}{\omega_0^2}\left[\frac{(a_1 + a_2) K_m + (b_1 + b_2)}{(1 - K_m)\, C_2}\right]^{\frac{1}{2}}$$

$$C_5 = -\frac{(a_1 + a_2) K_m - (b_1 + b_2)}{a_1 a_2 K_m + b_1 b_2}$$

$$R_7 = \frac{a_1 a_2 K_m + b_1 b_2}{a_1 a_2 K_m - b_1 b_2}$$

Fig. 4 — Cross arm z_y for structure VII of Fig. 2.

Then z_y is given by the network in Fig. 4 where the values of K_m, the zeros and the poles are the same as those given in Fig. 3.

IV. EXPLICIT FORMULAS FOR STRUCTURE VIII

The physical realizability conditions and the typical attenuation characteristic of this structure are given in Fig. 2. It can be shown that general formulas in terms of the zeros and the poles of exp θ for K_m and the A_j's are the same as those for structure VII, with the exception that for structure VIII the coefficients of z_x satisfy the following inequality

$$[A_5 A_2]^{\frac{1}{2}} - [A_1 A_6]^{\frac{1}{2}} > 0 \tag{41}$$

yielding a positive sign for the reactance $j\omega_0 L_1$ and a negative sign for the reactance $j\omega_0 L_3$ of Fig. 3. Similarly, the cross arm z_y of structure VIII yielding a negative sign for reactance $j\omega_0 L_4$ and a positive sign for reactance $j\omega_0 L_6$ of Fig. 4. Thus the Brune network for the series arm z_x and the cross arm z_y of structure VIII is shown in Figs. 5 and 6 respectively.

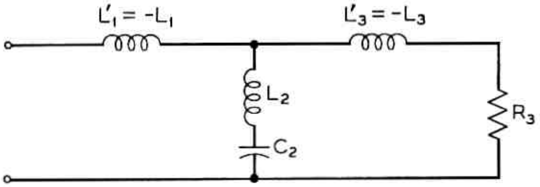


Fig. 5 — Series arm z_x for structure VIII of Fig. 2. (L's, C_2, and R_3 are same as for Fig. 3.)

V. EQUIVALENT NETWORKS TO STRUCTURES VII AND VIII

To avoid the need for coupled coils in the lattices developed previously, we can introduce the Bott-Duffin impedance arms in structures IX and X to obtain equivalent networks to structures VII and VIII respectively in Fig. 2.[4] The series and cross arm of structure IX have the same configuration as the cross and series arm respectively of structure X. Hence only the series arm of each lattice is shown in Fig. 7. These new lattice structures necessarily have the same physical realizability requirements and exhibit the same typical characteristics as sketched in Fig. 2 for structures VII and VIII. The element values for the general case of these lattice networks without mutual inductance are given in Table I.

VI. EXAMPLE

We now illustrate by an example the methods we have developed to obtain realization of constant resistance lattice networks. Let us find such a realization given the transfer impedance function

$$\exp\,\theta = \frac{K(s^4 + 2.268s^3 + 6.517s^2 + 3.302s + 4.905)}{s^4 + 2s^3 + 4.778s^2 + 5.556s + 5.556} \tag{42}$$

In order to represent the given function as tandem lattices, we ob-

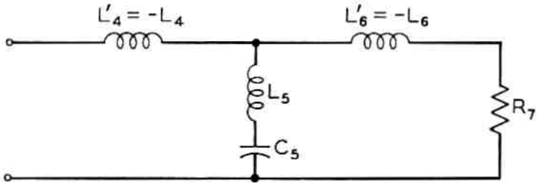


Fig. 6 — Cross arm z_y for structure VIII of Fig. 2. (L's, C_5 and R_7 are same as for Fig. 4.)

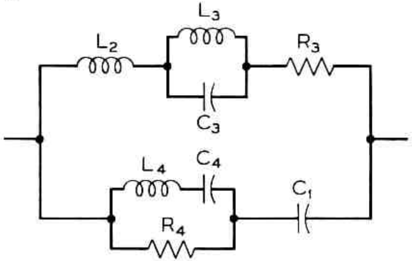
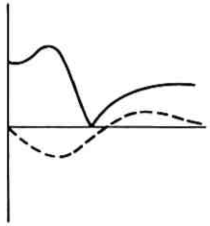
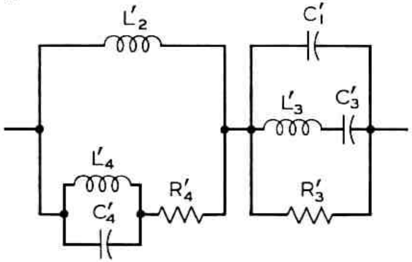
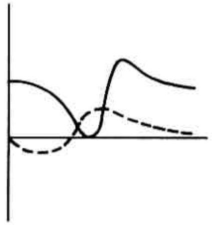
	STRUCTURE	GENERAL FORM OF EXP θ	REQUIREMENTS FOR PHYSICAL REALIZABILITY	TYPICAL ATTENUATION PHASE CHARACTERISTICS
IX	Z_x (SERIES ARM) 	$\frac{K(s-a_1)(s-a_2)}{(s-b_1)(s-b_2)}$	$\lvert a_1 a_2\rvert \geq \lvert b_1 b_2\rvert$ $b_1^2 + b_2^2 \geq a_1^2 + a_2^2$ THE a's ARE COMPLEX THE b's MAY BE REAL OR COMPLEX	
X	Z_x (SERIES ARM)  Z_y HAVE THE SAME CONFIGURATION AS Z_x IN (IX)	$\frac{K(s-a_1)(s-a_2)}{(s-b_1)(s-b_2)}$	$\lvert a_1 a_2\rvert \leq \lvert b_1 b_2\rvert$ $\frac{1}{b_1^2} + \frac{1}{b_2^2} \geq \frac{1}{a_1^2} + \frac{1}{a_2^2}$ THE a's ARE COMPLEX THE b's ARE REAL OR COMPLEX	

Fig. 7 — Constant resistance lattice equalizers.

TABLE I(a)—ELEMENT VALUES FOR STRUCTURE IX

Z_x	Z_y
$L_2 = \dfrac{R_3R_4}{C_1}$	$L_2' = \dfrac{1}{C_1}$
$L_3 = \dfrac{M}{[(\alpha_2K_m + \beta_2)^3(K_m + 1)]^{\frac{1}{2}}}$	$C_2' = L_2$
$\quad = \dfrac{M}{[(a_1a_2K_m + b_1b_2)^3(K_m + 1)]^{\frac{1}{2}}}$	
$C_3 = \dfrac{R_3R_4}{L_4}$	$L_3' = \dfrac{1}{C_3}$
$R_3 = \dfrac{\alpha_2K_m - \beta_2}{\alpha_2K_m + \beta_2} = \dfrac{a_1a_2K_m - b_1b_2}{a_1a_2K_m + b_1b_2}$	$C_3' = L_3$
$L_4 = \dfrac{[(K_m - 1)^3(\alpha_2K_m - \beta_2)]^{\frac{1}{2}}}{M}$	$R_3' = \dfrac{1}{R_3}$
$\quad = \dfrac{[(K_m - 1)^3(a_1a_2K_m - b_1b_2)]^{\frac{1}{2}}}{M}$	
$C_4 = \dfrac{R_3R_4}{L_3}$	$L_4' = \dfrac{1}{C_4}$
$R_4 = \dfrac{K_m - 1}{K_m + 1}$	$C_4' = L_4$
$C_1 = \left[\dfrac{(\alpha_2K_m - \beta_2)(\beta_1 - K_m\alpha_1)}{(-\beta_1 - K_m\alpha_1)(K_m + 1)}\right]^{\frac{1}{2}}$	$R_4' = \dfrac{1}{R_4}$
$\quad = \left\{\dfrac{(a_1a_2K_m - b_1b_2)[(b_1 + b_2) - K_m(a_1 + a_2)]}{[-(b_1 + b_2) - K_m(a_1 + a_2)](K_m + 1)}\right\}^{\frac{1}{2}}$	

where

$$
\begin{aligned}
M &= (\beta_1 - K_m\alpha_1)[(\alpha_2K_m + \beta_2)(K_m + 1)]^{\frac{1}{2}} \\
&\quad - (\beta_1 + K_m\alpha_1)[(\alpha_2K_m - \beta_2)(K_m - 1)]^{\frac{1}{2}} \\
&= [(b_1 + b_2) - K_m(a_1 + a_2)][(a_1a_2K_m + b_1b_2)(K_m + 1)]^{\frac{1}{2}} \\
&\quad - [(b_1 + b_2) + K_m(a_1 + a_2)][(a_1a_2K_m - b_1b_2)(K_m - 1)]^{\frac{1}{2}}
\end{aligned}
$$

TABLE I(b)—ELEMENT VALUES FOR STRUCTURE X

Z_x	Z_y
$L_2' = \left\{\dfrac{(K_m - 1)(\beta_1 - K_m\alpha_1)}{(\alpha_2 K_m + \beta_2)(-\beta_1 - K_m\alpha_1)}\right\}^{\frac{1}{2}}$	$L_2 = C_1'$
$= \left\{\dfrac{(K_m - 1)[(b_1 + b_2) - K_m(a_1 + a_2)]}{(a_1a_2K_m + b_1b_2)[-(b_1 + b_2) - K_m(a_1 + a_2)]}\right\}^{\frac{1}{2}}$	
$C_1' = \dfrac{L_2'}{R_3R_4}$	$L_3 = C_3'$
$L_3' = R_3'R_4'C_4'$	$C_3 = \dfrac{1}{L_3'}$
$C_3' = \dfrac{M}{[(\alpha_2K_m - \beta_2)^3(K_m - 1)]^{\frac{1}{2}}}$	$R_3 = \dfrac{1}{R_3'}$
$= \dfrac{M}{[(a_1a_2K_m - b_1b_2)^3(K_m - 1)]^{\frac{1}{2}}}$	
$R_3' = \dfrac{\alpha_2K_m - \beta_2}{\alpha_2K_m + \beta_2} = \dfrac{a_1a_2K_m - b_1b_2}{a_1a_2K_m + b_1b_2}$	$L_4 = C_4'$
$L_4' = R_3'R_4'C_3'$	$C_4 = \dfrac{1}{L_4'}$
$C_4' = \dfrac{[(K_m + 1)^3(\alpha_2K_m + \beta_2)]^{\frac{1}{2}}}{M}$	$R_4 = \dfrac{1}{R_4'}$
$= \dfrac{[(K_m + 1)^3(a_1a_2K_m + b_1b_2)]^{\frac{1}{2}}}{M}$	
$R_4' = \dfrac{K_m - 1}{K_m + 1}$	$C_1 = \dfrac{1}{L_2'}$

where

$$\begin{aligned} M &= (\beta_1 - K_m\alpha_1)[(\alpha_2K_m + \beta_1)(K_m + 1)]^{\frac{1}{2}} \\ &\quad - (\beta_1 + K_m\alpha_1)[(\alpha_2K_m - \beta_2)(K_m - 1)]^{\frac{1}{2}} \\ &= [(b_1 + b_2) - K_m(a_1 + a_2)][(a_1a_2K_m + b_1b_2)(K_m + 1)]^{\frac{1}{2}} \\ &\quad - [(b_1 + b_2) + K_m(a_1 + a_2)][(a_1a_2K_m - b_1b_2)(K_m - 1)]^{\frac{1}{2}} \end{aligned}$$

tain a partial product expansion of exp θ wherein the factors are bilinear or biquadratic forms. In the present example, we find

$$\exp\,\theta = K\left(\frac{s^2 + 2s + 5}{s^2 + 2s + 2}\right)\left(\frac{s^2 + 0.268s + 0.981}{s^2 + 2.778}\right) \tag{43}$$

or

$$\exp\,\theta = \left[K_1\left(\frac{s^2 + 2s + 5}{s^2 + 2s + 2}\right)\right]\left[K_2\left(\frac{s^2 + 0.268s + 0.981}{s^2 + 2.778}\right)\right] \tag{44}$$

where K or K_1K_2 are constant multipliers to allow for any corresponding net increase in loss required by the overall network. Each biquadratic factor must be physically realizable if it is to be synthesized using one of the basic structures. For the first factor we get the following zeros for the polynomials

$$\begin{aligned} a_1 &= -1 + j2, & a_2 &= -1 - j2, \\ b_1 &= -1 + j, & b_2 &= -1 - j. \end{aligned} \tag{45}$$

From these a's and b's we determine

$$\begin{aligned} \alpha_1 &= -2, & \alpha_2 &= 5, \\ \beta_1 &= -2, & \beta_2 &= 2, \\ a_1^2 + a_2^2 &= -6, & b_1^2 + b_2^2 &= 0. \end{aligned} \tag{46}$$

Substituting into the physical realizability conditions of structures VII and IX, we find that these conditions are satisfied. The second factor in (44) has the following values

$$\begin{aligned} \alpha_1 &= -0.268, & \alpha_2 &= 0.981, \\ \beta_1 &= 0, & \beta_2 &= 2.778, \\ \frac{1}{a_1^2} + \frac{1}{a_2^2} &= -1.964, & \frac{1}{b_1^2} + \frac{1}{b_2^2} &= -0.72. \end{aligned} \tag{47}$$

With these values the second factor satisfies the requirements for structures VIII and X. Hence, the given transfer function (42) can be represented by two second degree lattices in tandem with or without mutual coupled coils.

Using (27), K_1 and K_2 of (44) become

$$K_1 = K_{m1} = 1.2895, \tag{48}$$

$$K_2 = K_{m2} = 7.035, \tag{49}$$

and the constant multiplier K is

$$K = K_1 K_2 = 9.0716. \tag{50}$$

Thus by substituting (46) and (48) into the explicit formulas for structures VII and IX, and substituting (47) and (49) into the explicit formulas for structures VIII and X, the element values for each corresponding structure can be obtained. These element values are summarized in Tables II, III, IV and V. It should be apparent that another realization may be obtained for (42) by interchanging the numerators of the two biquadratic factors given in (43). Then the counterparts would have to be re-examined to see which basic structure would be realizable.

VII. GENERAL SYNTHESIS PROCEDURE

The flow chart shown in Fig. 8 is a guide for the general synthesis procedure of any physical realizable, rational transfer impedance function exp θ. This flow chart can be summarized as follows:

(*i*) Factor the given transfer impedance function into first and second degree functions with both members of any conjugate complex pair of zeros and poles retained in each given partial product.

(*ii*) Synthesize all first degree functions by structure III or IV according to their physical realizability conditions.

(*iii*) Examine the poles and zeros of these second degree functions to see whether they are real or complex; then use the appropriate group of structures indicated. If the poles and zeros are real, factor the second degree function into first degree functions.

(*iv*) Examine the physical realizability conditions further to determine to which sub-group of structures the function belongs.

TABLE II—ELEMENT VALUES OF STRUCTURES VII FOR (43)

z_x (series arm)	z_y (cross arm)
$L_1 = -0.0658$	$L_4 = 2.0178$
$L_2 = 0.1290$	$L_5 = 1.9379$
$C_2 = 1.0296$	$C_5 = 0.0685$
$L_3 = 0.1344$	$L_6 = -0.9876$
$R_3 = 0.5265$	$R_7 = 1.8994$

TABLE III—ELEMENT VALUES OF STRUCTURE IX FOR (43) (EQUIVALENT TO STRUCTURE VII)

z_x (series arm)		z_y (cross arm)	
$C_1 = 0.4956$	$L_3 = 0.2084$	$C_1' = 0.1342$	$C_3' = 0.2084$
$L_2 = 0.1342$	$R_4 = 0.1264$	$L_2' = 2.0178$	$R_4' = 7.9113$
$R_3 = 0.5265$	$L_4 = 0.0424$	$R_3' = 1.8994$	$L_4' = 3.1319$
$C_3 = 1.5695$	$C_4 = 0.3193$	$L_3' = 0.6371$	$C_4' = 0.0424$

(*v*) Connect the synthesized elementary structures in tandem.
(*vi*) Raise the impedance level to the desired R_0 .

The realizability conditions of structures VII and IX are the same and also those for structure VIII and X. If one wishes to avoid having coupled coils, he should use structures IX and X. Since structures IX and X are generally more complex, one may elect to use the structures VII and VIII to save on the number of elements in the final network.

VIII. CONCLUSION

In the field of classical network theory, Bode developed explicit formulas in terms of the poles and zeros of the transfer impedance function for synthesizing constant resistance lattice structures of the types I, II, III, IV, V, and VI. This paper has shown the detailed development and derivations of explicit formulas in terms of the poles and zeros of the transfer impedance function for synthesizing types VII and VIII, with coupled coils by Brune Method, and for types IX and X which are new types of structures that may be used to avoid having coupled coils by Bott-Duffin Procedure. With the addition of these results, a general method for synthesizing any physically realizable, rational transfer impedance function with explicit formulas is complete. The explicit formulas method developed in this

TABLE IV—ELEMENT VALUES OF STRUCTURE VIII FOR (43)

z_x (series arm)	z_y (cross arm)
$L_1' = 0.7896$	$L_4' = -1.3958$
$L_2 = 2.4106$	$L_5 = 5.6767$
$C_2 = 0.4572$	$C_5 = 0.1948$
$L_3' = -0.5949$	$L_6' = 1.8564$
$R_3 = 0.4266$	$R_7 = 2.3475$

TABLE V—ELEMENT VALUES OF STRUCTURE X FOR (43) (EQUIVALENT TO STRUCTURE VIII)

z_x (series arm)		z_y (cross arm)	
$C_1' = 2.4677$	$C_3' = 1.2658$	$C_1 = 1.2665$	$C_3 = 1.1481$
$L_2' = 0.7896$	$R_4' = 0.7511$	$L_2 = 2.4677$	$L_4 = 2.7222$
$R_3' = 0.4260$	$C_4' = 2.7222$	$R_3 = 2.3470$	$C_4 = 2.4691$
$L_3' = 0.8710$	$L_4' = 0.4050$	$L_3 = 1.2658$	$R_4 = 1.3314$

paper gives more rapid results and introduces fewer round-off errors than the step-by-step procedures used in the past.

IX. ACKNOWLEDGMENT

The author would like to thank Drs. W. H. Boghosian and S. D. Bedrosian of The Moore School of Electrical Engineering for their many helpful suggestions and criticism, and Messrs. H. A. Helm, W. D. Wynn and L. Schuchman of Bellcomm, Inc., for their fruitful comments.

APPENDIX

Derivation of Physical Realizability Conditions

We now develop the physical realizability conditions for second degree lattices in terms of the poles and the zeros of exp θ. We shall find that the requirement for non-negative loss at real frequencies for such two-ports leads to both a product and a summation condition on the poles and zeros of the transfer impedance function. This analysis is carried out in terms of an example utilizing structures VII and IX. These structures exhibit zero loss at a finite frequency ω_0 (see Figs. 2 and 7). To evaluate the constant multiplier for these structures we set exp $(2\alpha) = 1$ at ω_0, and let $x_0 = \omega_0^2$. Then from (16), k becomes

$$k = \frac{(\beta_2 - x_0)^2 + \beta_1^2 x_0}{(\alpha_2 - x_0)^2 + \alpha_1^2 x_0} \tag{51}$$

where α's and β's are given by (6) and (7) respectively.

Substituting (51) back into (16) and applying the non-negative loss condition, that is, exp $(2\alpha) \geqq 1$ for all frequencies we have

$$\left[\frac{(\beta_2 - x_0)^2 + \beta_1^2 x_0}{(\alpha_2 - x_0)^2 + \alpha_1^2 x_0}\right]\left[\frac{(\alpha_2 - x)^2 + \alpha_1^2 x}{(\beta_2 - x)^2 + \beta_1^2 x}\right] \geqq 1. \tag{52}$$

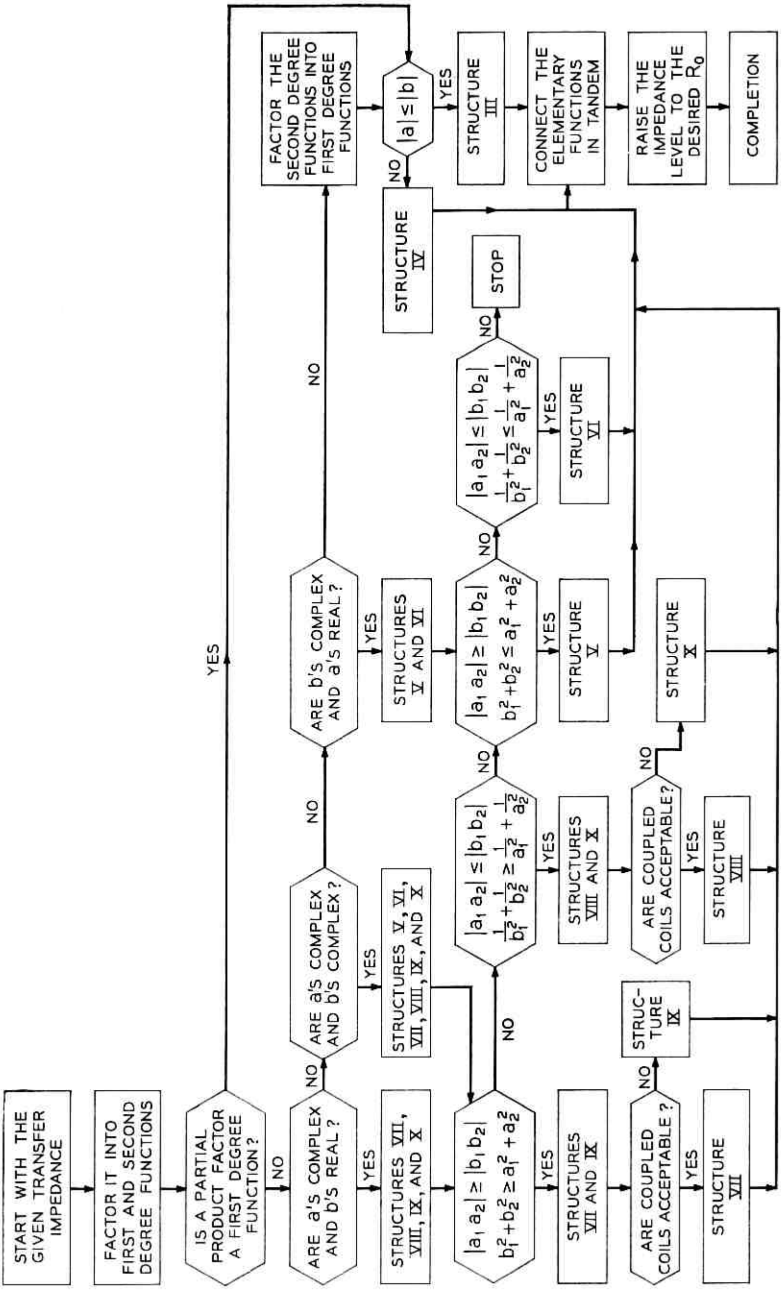

START WITH THE GIVEN TRANSFER IMPEDANCE
FACTOR IT INTO FIRST AND SECOND DEGREE FUNCTIONS
IS A PARTIAL PRODUCT FACTOR A FIRST DEGREE FUNCTION?
YES
NO
FACTOR THE SECOND DEGREE FUNCTIONS INTO FIRST DEGREE FUNCTIONS
$|a| \leq |b|$
YES
NO
STRUCTURE III
STRUCTURE IV
CONNECT THE ELEMENTARY FUNCTIONS IN TANDEM
RAISE THE IMPEDANCE LEVEL TO THE DESIRED R_0
COMPLETION
ARE a'S COMPLEX AND b'S REAL?
YES
NO
STRUCTURES VII, VIII, IX, AND X
ARE a'S COMPLEX AND b'S COMPLEX?
YES
NO
STRUCTURES V, VI, VII, VIII, IX, AND X
ARE b'S COMPLEX AND a'S REAL?
YES
NO
STRUCTURES V AND VI
$|a_1 a_2| \geq |b_1 b_2|$ $b_1^2+b_2^2 \geq a_1^2+a_2^2$
YES
NO
STRUCTURES VII AND IX
ARE COUPLED COILS ACCEPTABLE?
YES
NO
STRUCTURE VII
STRUC-TURE IX
$|a_1 a_2| \leq |b_1 b_2|$ $\frac{1}{b_1^2}+\frac{1}{b_2^2} \geq \frac{1}{a_1^2}+\frac{1}{a_2^2}$
YES
NO
STRUCTURES VIII AND X
ARE COUPLED COILS ACCEPTABLE?
YES
NO
STRUCTURE VIII
STRUCTURE X
$|a_1 a_2| \geq |b_1 b_2|$ $b_1^2+b_2^2 \leq a_1^2+a_2^2$
YES
NO
STRUCTURE V
$|a_1 a_2| \leq |b_1 b_2|$ $\frac{1}{b_1^2}+\frac{1}{b_2^2} \leq \frac{1}{a_1^2}+\frac{1}{a_2^2}$
YES
NO
STRUCTURE VI
STOP

Fig. 8 — Flow chart for synthesizing transfer impedance functions with constant resistance lattice networks.

We shall obtain one of the realizability conditions by letting the frequency approach infinity. The result is the expression

$$(\beta_2 - x_0)^2 + \beta_1^2 x_0 \geqq (\alpha_2 - x_0)^2 + \alpha_1^2 x_0 \ . \tag{53}$$

Expanding this expression and substituting for α's and β's from (6) and (7), we obtain

$$(b_1 b_2)^2 + (b_1^2 + b_2^2)x_0 \geqq (a_1 a_2)^2 + (a_1^2 + a_2^2)x_0 \ . \tag{54}$$

For

$$| \, a_1 a_2 \, | \geqq | \, b_1 b_2 \, | \tag{55}$$

then

$$(a_1 a_2)^2 \geqq (b_1 b_2)^2. \tag{56}$$

Thus (56) can be rewritten as

$$(a_1 a_2)^2 = (b_1 b_2)^2 + \epsilon \tag{57}$$

where ϵ is a positive quantity. Substituting (57) into (54) we get

$$(b_1 b_2)^2 + (b_1^2 + b_2^2)x_0 \geqq (b_1 b_2)^2 + \epsilon + (a_1^2 + a_2^2)x_0 \ . \tag{58}$$

Simplifying and dividing both sides by x_0 yields

$$b_1^2 + b_2^2 \geqq a_1^2 + a_2^2 + \frac{\epsilon}{x_0}. \tag{59}$$

Since ϵ and x_0 are positive quantities their ratio may be deleted without altering the inequality of (59). Thus we have shown that the realizability requirements for second degree structures VII and IX are given by the pair of expressions

$$b_1^2 + b_2^2 \geqq a_1^2 + a_2^2 \tag{60}$$

and

$$| \, a_1 a_2 \, | \geqq | \, b_1 b_2 \, | .$$

REFERENCES

1. Bode, H. W., *Network Analysis and Feedback Amplifier Design,* Princeton, New Jersey: D. Van Nostrand Company, Inc., 1959, Chapter 12, pp. 250–258.
2. Lee, S. Y., "Explicit Formulas for Constant Resistance Lattice Synthesis of Transfer Impedance," Master's Thesis, The Moore School of Electrical Engineering, University of Pennsylvania (December 1965).
3. Boghosian, W. H. and Bedrosian, S. D., unpublished work.
4. Foster, R. M., "Passive Network Synthesis," Proceedings of the Polytechnic Institute of Brooklyn, Symposium on Modern Network Synthesis, 1955, pp. 3–9.

Recirculating Ultrasonic Stores: An Economical Approach to Sequential Storage with Bit Rates Beyond 100 MHz

By E. K. SITTIG and F. M. SMITS

(Manuscript received September 30, 1968)

State of the art integrated circuits and ultrasonic delay lines can be combined to form batch-fabricated digital storage modules having random access to sequentially stored blocks of information. Greatest economy is indicated if such stores are designed for as high a bit rate as is technologically feasible, at present limited by the speed of available integrated circuitry. A store of optimum design will have a block size of approximately 1000 bits which, for a bit rate of 100 MHz, gives a maximum latency time of approximately 10 microseconds. Such designs are realizable with zero temperature coefficient material. The stores can be used as main memories for small computers or as fast transfer stores shuttling information between a slow external bulk memory and a very fast random access memory in large computers.

A variety of accessing modes permit these stores to operate over a large range of access rates without requiring large buffer stores.

I. INTRODUCTION

In a computer of conventional organization, a central processor communicates with a large array of randomly accessible storage locations, each of which contains one word of a given number of bits. The assembly of these locations, the "random access memory," typically consists of one discrete element for each bit stored, which occupies a fixed location in space. This approach is comparatively costly. At present, the cycle time for such a memory of megabit size is of the order of one microsecond.

Since cost and size normally prohibit providing storage for more than a few million bits in this form, additional bulk memory is provided in which bits are stored in homogeneous media at lower cost and higher density. Since the bit locations in this bulk memory are basically defined by sequential scanning from a given addressable starting location, the information has to be stored or read out sequentially as a block. Therefore, once the desired block is addressed, a certain latency time passes until the information is available. This typically ranges from 10 to 100 milliseconds in mechanically scanned systems such as drums or disks and is even longer if heads have to be repositioned. If the information is stored on magnetic tape, this latency time may be several minutes.

The present trend is to have shorter processor cycle times and multiple access facilities in evolving computers; increasing emphasis is put on the ability to transfer blocks of information quickly between a bulk store and the random-access memory with which the processor interacts. In order to avoid a bottleneck in the throughput of information, transfer stores of lower capacity but shorter latency time are provided; these transfer stores can be loaded from a slower store without intervention of the central processor but can also transfer data on demand with minimum waiting time. Drum stores and random access memory blocks are often used for this application.

It is the purpose of this paper to point out that ultrasonic delay line stores have been developed to the point where bit rates of 100 MHz and higher have become feasible; therefore, stores with operational properties similar to those of a drum can be built which have maximum latency times of about 10 microseconds. Organized in parallel tracks, these stores can transfer many giga bit per second. In contrast to magnetic storage, these stores share with semiconductor stores the disadvantage of volatility with respect to power failure; but they have the advantage of high storage density and absence of moving parts.

These devices appear to be a strong contender for buffer stores of relatively short latency time. Since they are sequential stores with the information stored in a homogeneous medium, one may expect that the storage cost could be considerably lower than would be the case for a random access memory. This paper will show this to be the case. As a matter of fact, the higher the frequency of operation the more economical a delay line store becomes since its components become more and more compact. At 100 MHz bit rate, for example, stores with packing densities exceeding 6000 bits per cubic centimeter are readily possible.

In optimizing a store, the interrelation between delay line and auxiliary circuitry needs to be considered. Details of delay line design analysis have been discussed elsewhere.[1] The present paper uses the delay line design analysis in an optimization with the auxiliary circuitry to find a combination which is optimized from functional and economical considerations.

The most repetitive elements in a delay line store are individual recirculating delay line loops. Accordingly, the optimization of individual loops will be considered first. This will be followed by giving detailed design considerations for the delay lines meeting such requirements. Finally, operational characteristics of a delay line store will be covered.

II. GENERAL TRADE-OFF CONSIDERATIONS FOR A SINGLE DELAY LINE LOOP

A basic recirculating delay line loop typically has the configuration shown in Fig. 1. Binary coded data in the form of pulses appearing at terminal DI are inserted into the delay line through gates B and A when a "write" command pulse appears at terminal W. As the pulses appear at the other end of the delay line, they are amplified to make up for the insertion loss of the line and are detected in the amplifier-detector AD. In gate E the detected signals are retimed with respect to an external clock frequency inserted at terminal Cl. They can be monitored at the data output terminal DO_1 or gated

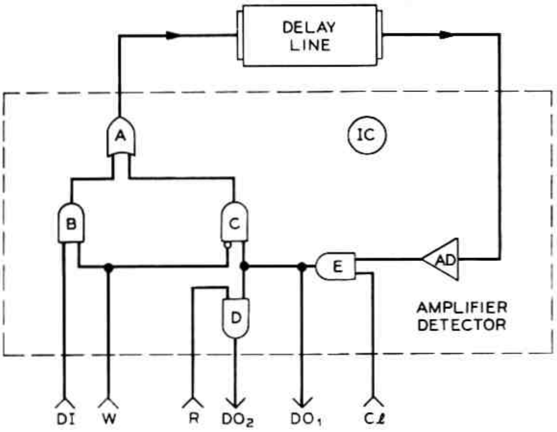


Fig. 1 — Basic recirculating storage loop. The terminals are as designated: DI, data insertion; W, write command; R, read command; DO_1, DO_2, data output; Cl, external clock frequency.

through D to the output terminal DO_2 by a "read" command applied at R. Gate C provides reconnection to the input of the delay line unless disabled by the write command at W while reading in new data. Storage loops of similar configuration have in the past been used with bit rates of a few MHz and storage capacities of up to about 20,000 bits in electronic desk calculators and the like.

A large store will have to contain a multiplicity of identical storage loops, and a limited number of control circuits, registers, and clock frequency supplies which are common to all loops. To minimize the over-all cost of the system, it is most important to minimize the cost and complexity of the components constituting the individual loops. In principle a given storage capacity may be obtained by using many relatively simple delay lines with a corresponding number of regeneration circuits or by using fewer but more complex delay lines usually requiring more complex circuitry. As long as the circuits have to be built from discrete components, it is more economical to use long and fairly complex delay lines so that maximum use could be made of the expensive circuitry.

As an example of this approach, a store has been built with a capacity of 1.3×10^6 bits, using 48 delay lines, storing 28,000 bits each at a bit rate of 40 MHz, which gives a resulting latency time of approximately 707 microseconds.[2] The materials available for such a large storage capacity per delay line exhibit a sizeable absorption loss and a temperature coefficient of delay around 80 ppm per centigrade degree. With such a large temperature coefficient some form of temperature stabilization is necessary. The high insertion loss of the delay lines—typically 50 dB pulse-to-pulse—requires amplifiers with carefully controlled linear gain at the output of each delay line to bring the signal back up to logic level. To retime unavoidable drift in temperature between the individual delay lines, the regeneration circuitry, in addition, has to provide the largest retiming margin possible.

These stringent requirements might be relaxed considerably by an alternative approach using delay lines which store only about 1000 bits each, so that the individual lines can be of a simple rectangular block configuration; the rectangular block configuration, in contrast to the polygons required in the above-mentioned example, can readily be batch fabricated in large numbers. Thus a cost saving appears possible in spite of the 28-fold increase in the number of delay lines over the example mentioned before. Also, the shorter delay line length permits the use of a delay medium, with higher absorption but with a lower temperature coefficient, so that temperature stabilization equipment becomes unnecessary. Finally, with shorter individual delay lines a 28-fold reduction in the latency time is achieved which, if combined with lower cost for the devices, could make it more attractive for computer applications.

This approach will only be economical if the concomitant increase in the number of regeneration circuits can be obtained at minimal cost. This should be feasible if each circuit could be built as an individual integrated circuit of reasonable size. In order to make this practical, certain requirements are posed on the delay line performance. Since with an integrated circuit level detector pulses of 30 millivolts amplitude or more can easily be detected, and since an integrated circuit driver can deliver readily pulses of the order of 1 volt amplitude, the delay line pulse-to-pulse insertion loss should not exceed 30 dB. If kept to such levels, closely gain-controlled linear amplifiers can be avoided. Also, the transducers of the delay line should have an impedance falling into the range of 10 to 100 ohms so as to permit coupling the delay line to integrated circuitry without the use of transformers or tuning inductors.

III. DESIGN CONSIDERATIONS FOR THE ULTRASONIC DELAY LINES

The simple delay line to be considered has the configuration shown in Fig. 2; two equal piezoelectric transducers of thickness ℓ_c and active diameter $2r$ are affixed to a delay medium in the shape of a bar of length x and square cross-section having a width D. The delay medium is characterized by its sound velocity c_d, density ρ_d and amplitude absorption index μ (absorption per wavelength). The transducer material is characterized by its sound velocity c_c, density ρ_c, permittivity ϵ, and electromechanical coupling factor k. At the frequency f_o, for which the thickness of the transducer equals half a sound wave length, that is, for

$$f_o = c_c/2\ell_c \tag{1}$$

the electrical impedance Z_i appearing at its electrical terminals is

$$Z_i = \frac{1}{\omega_o C_o}(-j + 4k^2/\pi z_t), \tag{2}$$

where $C_o = \pi r^2 \epsilon/\ell_c$ and $z = \rho_d c_d/\rho_c c_c$ is the acoustic impedance ratio of the delay medium with respect to the transducers. Without tuning networks, maximum power is transferred between a source of impedance R_s and the delay line, and likewise between the line and a

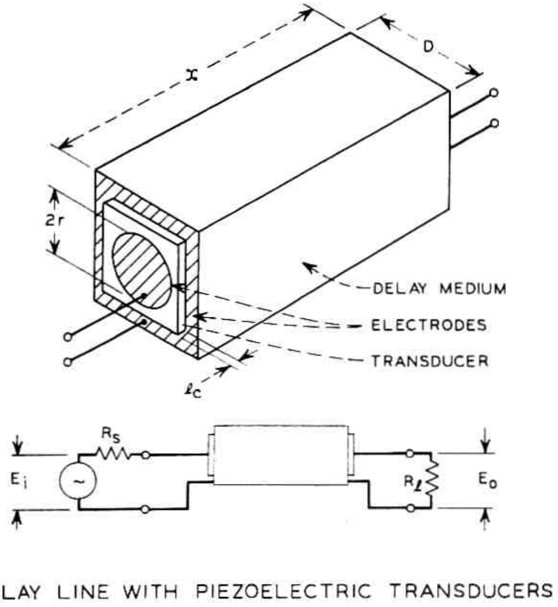


Fig. 2 — The basic delay line configuration. Two transducers of thickness l_c and diameter $2r$ are attached to a delay medium of length x and lateral dimension D. This delay line is connected between a source of resistance R_s and load R_l without tuning networks.

load resistance R_l, if

$$R_s = R_l \approx 1/\omega_o C_o \,; \qquad \omega_o = 2\pi f_o \,. \tag{3}$$

provided that $4k^2/\pi z \ll 1$ which is fulfilled in all the cases of interest here.

As shown in detail in Ref. 1, with these electrical terminations one obtains a reasonably linear phase response and a pass band centered near f_o which rolls off approximately like $\sin^4 (\pi f/f_o)$ on either side of f_o if z is selected to fulfill the condition

$$z = 1 - k^2. \tag{4}$$

With this pass band, a unipolar input pulse of rectangular envelope and a nominal width

$$T = 1/2f_o \tag{5}$$

gives rise to an output pulse having the shape shown in Fig. 3. The center lobe of this pulse is about 6.5 dB below the amplitude of the input pulse. This reduction in pulse amplitude is due to side lobes being generated by the band limiting characteristic of the delay line.

Pulses can be inserted at a maximum bit rate equaling f_o with adequate margin for binary detection. Thus one bit can be stored for each sound wave length λ_d in the delay medium so that the storage capacity N is given by

$$N = x/\lambda_d = c_d x/f_o \, . \tag{6}$$

The transducer insertion loss would be minimized if k is chosen as large as possible. This is evident from (2) in that more of the input voltage is dropped across the resistive part of the input impedance. However, $k = 0.6$ is about the maximum available in transducer materials with reasonable technological properties so that a transducer insertion loss minimum of a few dB seems unavoidable. An attempt to trade off bandwidth for a reduction of loss would, as a rule, impair the detection margin of the output signal, and would probably increase the pulse amplitude insertion loss due to the pass-band characteristic above the value of 6.5 dB mentioned before.

Within these limitations one may now choose a transducer-delay medium combination by criteria such as a small temperature coefficient and adequate sound absorption. As shown in Ref. 1 one can combine the expressions for the length of the delay line x, the thickness of the

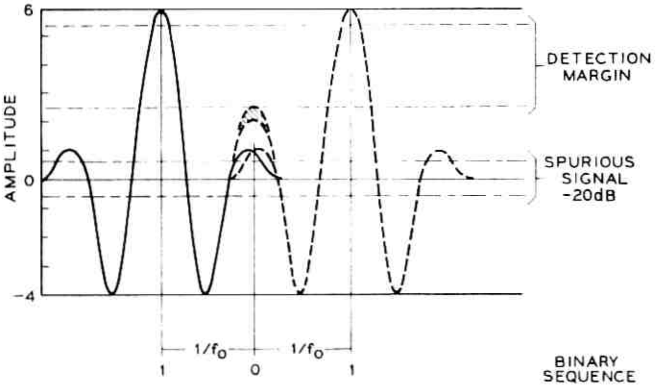


Fig. 3 — Output signal from a delay line when rectangular pulses of duration $1/2f_o$ are applied to its input. The worst case for detection is obtained when such pulses are entered at the rate $f_o/2$ representing a binary 1–0–1 sequence as shown; since the outermost sidelobes of the nearest neighbors will fill in the "zero" slot between "one" pulses, this creates intersymbol interference. A spurious signal at a —20 dB level further reduces the detection margin to the one shown.

transducer ℓ_c, its capacitance C_o, and the length of its Fresnel zone x_o which determines its directivity, namely

$$x = N\lambda; \qquad \ell_c = c_c/2f_o\,; \qquad C_o = \epsilon 4\pi r^2/\ell_c\,; \qquad x_o = r^2/\lambda, \tag{7}$$

to obtain a compatibility condition

$$N = m\cdot\omega_o C_o\cdot x/x_o \tag{8}$$

with a materials constant

$$m = c_c/4\pi^2\epsilon c_d^2\,. \tag{9}$$

This condition implies that the beam spreading loss L_b (which is approximately x/x_o in dB if $x/x_o \leqq 10$) and the impedance levels of the delay lines are interrelated for a given material combination and storage capacity.

The total delay path loss L_d is composed in part of the absorption in the delay medium L_a and the beam spreading loss L_b .

$$L_d = L_a + L_b \approx x/x_o + \mu N \tag{10}$$

with μ in dB per bit. In order to keep the total loss below 30 dB, the above loss should be restricted to about 20 dB, since in addition there are a few dB transducer loss and the 6.5 dB loss in pulse amplitude due to the band-pass characteristic of the delay line.

The absorption loss L_a increases with increasing frequency and thus introduces by itself additional distortions. Since, the beam spreading loss L_b decreases with increasing frequency, it is possible at least to first order, to compensate these two losses by choosing them approximately equal at the center of the pass band leading to the condition

$$x/x_o = \mu\cdot N. \tag{11}$$

By restricting the total loss to 20 dB, one is limited to $x/x_o \leqq 10$. To reduce the spurious signals, (primarily the triple travel signal due to multiple reflection between the transducers) to at least 20 dB below the main response, a minimum propagation loss L_d of 10 dB is necessary. This requirement limits the design range to

$$5 \leqq x/x_o \leqq 10. \tag{12}$$

The lateral dimension D of the delay medium is determined by the requirement that the directional response of the transducers suppresses glancing reflections from the side walls by at least 20 dB. This is assured if[1]

$$D/r = x/x_o \,. \tag{13}$$

The above relations contain all information necessary for a complete design of an optimized delay line.

It is interesting to note that (7), (9), and (11) do not contain the frequency explicitly. It enters only in implicit form through the absorption index μ which for most suitable materials increases less than linearly with frequency.

The other factor to consider is the potential cost of the delay line when compared to other methods of information storage. The technology required in fabricating delay lines is very similar to the semiconductor device technology and, as in that case, the variable giving a measure of the cost is the area that has to be precision finished, plated, and so on. For delay lines this is the area of the two end faces, each of which is given by

$$D^2 = N\lambda_d^2 x/x_o \,, \tag{14}$$

a relation obtained by combining (6), (7) and (13).

Thus, the area per bit to be finished, $2D^2/N$, is seen to decrease as $1/f_o^2$. Moreover, (6) and (14) combined state that the delay line volume xD^2 decreases like $1/f_o^3$. Thus at high frequencies the materials cost can be expected to be negligible compared to the finishing cost. These relations imply that, for economical reasons, the delay line should be operated at as high a frequency as possible, in spite of the reduction in storage capacity with increasing frequency which is imposed by the loss limit.

With present-day technology, delay lines have been made with storage capacities of approximately 1000 bits and pulse-to-pulse insertion loss in the vicinity of 30 dB with bit rates beyond 100 MHz.[3] There is no reason that this frequency could not be further increased. However, at present integrated circuitry with toggle rates much above 100 MHz has barely become available commercially so that at present 100 MHz is the highest frequency that can be considered from a practical point of view.

Once the frequency f_o, the material constant m, and the impedance $1/\omega_o C_o$ have been chosen, (9) indicates that the storage capacity N can only be varied in proportion to x/x_o. This in combination with (14) implies that the finished area per bit $2D^2/N$ increases with the storage capacity N. The permissible range of N is limited by the considerations leading to (12). As mentioned before, in the optimum case the beam spreading loss should equal the bulk loss, a condition which usually can

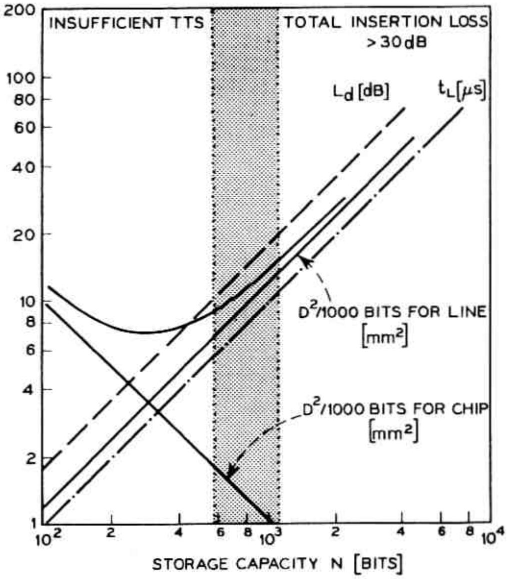


Fig. 4 — For a delay line consisting of sodium-potassium niobate transducers attached to a Bausch & Lomb T-40 glass delay medium and $f_o = 100$ MHz: the loss in the delay medium L_d, the latency time t_L, the combined endface area of the delay line, silicon chip area both per 1000 bits, and their sum. Outside the shaded region either the loss is too high or the triple travel suppression too low.

be met only approximately since the conditions imposed by (4) and (9) on the materials data will also have to be considered.

A useful compromise on all counts consists of a delay line using as a delay medium, a glass with nearly zero temperature coefficient of delay, such as Bausch & Lomb's T40 glass, and using ceramic sodium potassium niobate transducers. The materials constants for this system are as follows:[1,4] T40 glass: $c_d = 2.58$ millimeter per microsecond, $z = 0.51$, $\mu = 9 \times 10^{-3} \times (f/f_0)^{0.3}$ dB at $f_o = 100$ MHz; sodium potassium niobate ceramic: $c_c = 3.68$ millimeter per microsecond, $\epsilon \approx 500\epsilon_o$, $k = 0.6$. This combination fulfills (4) closely enough to be usable at bit rates up to f_o with adequate detection margins. Figure 4 shows the delay medium loss L_d of (10), the maximum latency time t_L, and the processed area per 1000 bit, $2D^2$ of (14) as a function of the storage capacity N for $f_o = 100$ MHz. Condition (12) in this case limits the value of N between 560 and 1100 bits as is also indicated in Fig. 4.

For $N = 1000$ the delay medium length is $x = 25.8$ millimeters and the lateral dimension $D = 2.45$ millimeters, so that the material cost can be considered insignificant compared with the processing cost of the end faces. These, however, tend to be proportional to the end face area $2D^2$, but will certainly be less than the cost of an equal area of integrated circuit chips, in view of the less complex procedures involved in fabricating the delay lines. The regeneration circuit comprises about 30 transistors and should, at the present state of the art, require about 1 square millimeters of silicon.[5] One circuit of this area is required for each delay line loop. The processed area per bit decreases, therefore, inversely with the number of bits stored in a single delay line as indicated in Fig. 4. As a result of this the sum of the processed area of the chip and the delay line itself has a minimum near $N = 300$ bits. If the processing costs per unit area were equal for the delay line and the chip, this minimum would correspond to the cost minimum. If, as appears likely, the cost per unit area is lower for the delay line, this minimum shifts to a higher N, close to the values of N between 560 and 1100 bit, permitted by the spurious signal supression and insertion loss limit. With a transducer impedance of 28.5 ohms the bulk loss in the delay medium at 100 MHz can be made equal to the beam spreading loss. This impedance will pose no difficulty with standard integrated circuitry.

It appears, therefore, that some presently available materials have close to optimum properties for the design of delay line storage loops operating at a bit rate of 100 MHz and storing 1024 bits with a resulting latency time of 10.24 microseconds. If such delay lines were produced by batch techniques, their cost should be comparable to those of a few square millimeters of silicon integrated circuits. The storage density in these devices is approximately 6000 bits per cubic centimeter of volume. Individual storage loops with delay lines storing 1024 bits have been built and operated at bit rates of 100 MHz.[5]

IV. ORGANIZATION OF A DELAY LINE STORE

Delay lines of the design described above are highly compact and could be built in modules of, for example, 18 delay lines built in a single glass plate of approximate dimensions 2 inches by 1 inch by 0.1 inch. The transducers would be mounted on the 2 inch by 0.1 inch faces while the 2 inch by 1 inch faces would be available as the substrate for interconnections, and the integrated circuit chips performing the recirculation, clock, and control functions. If one uses two of the 18 tracks for parity check and timing functions, then each module would store 16,384 bits. Such a module might constitute a repetitive element of any larger store so that the latency time for any randomly accessible block of information need not exceed 10.24 microseconds at a bit rate of $f_o = 100$ MHz, regardless of the store size.

It thus seems sufficient to discuss the organization of an individual module. The simplest organization would consist of providing random access by a 4-bit track selector to each track which stores a block of data words in a word and bit sequential organization. The start of each track would be delineated either by a 10-bit cyclic counter, counting off the clock frequency, or by reading signals from one or two additional delay lines serving as timing tracks. By adding an address register and a circuit, comparing its content with the counter, individual words in each track can be addressed individually. This type of organization, shown schematically in Fig. 5 using as elements the storage loops of Fig. 1, is natural for bit-serial processing with its economy of equipment so that it may well find application as main memory in small computers, where a cycle time of 10 microseconds is quite adequate.

Faster data transfer at the cost of more equipment is obtained in the word serial, bit parallel organization shown in Fig. 6. There the bits of any one word are contained in parallel tracks, so that a complete word is accessed by comparing its address with the counter reading. Data input and output have to be provided by parallel registers; transfer of words to the outside world can occur at the bit rate f_b. The store, whatever its size, can be written or read completely

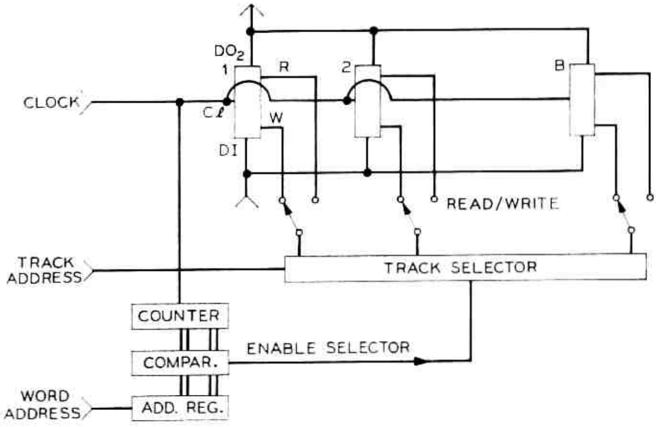


Fig. 5 — Bit-serial, word-serial organization of B storage loops as in Fig. 1.

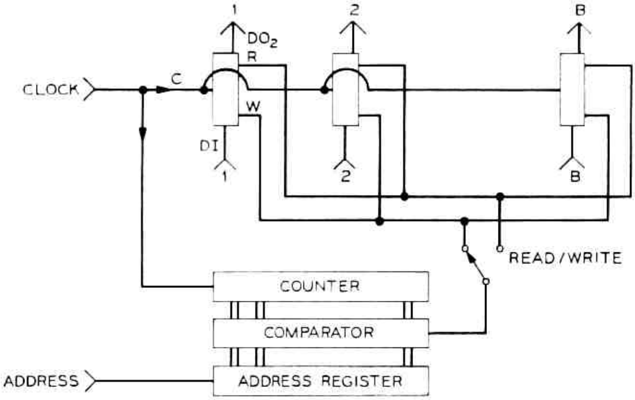


Fig. 6 — Bit-parallel, word-serial organization of B storage loops.

within 10.24 microseconds. In this form the store would cooperate best with a random access memory of about 10 nanoseconds cycle time without excessive buffer pile-up. As will be discussed in the Section V, slower, nonsequential operating modes can provide a match with memories of any cycle time between 10 nanoseconds and 10 microseconds.

Finally, like any sequential memory the delay line store can be organized associatively as shown in Fig. 7. All words are compared with a word preselected in the content register during the 10.24 microseconds it takes for all words to pass by the test location. If the comparison extends only over a part of the bit forming a word, these bits can serve as a pointer address for sorting sequences, and so on. In this version the store acts as an associative memory with a cycle time of 10.24 microseconds.

V. TIMING PROBLEMS

In the operation of a delay line store, the timing requires careful attention since in contrast to a digital shift register the delay line has a characteristic recirculation time. This time is dominated by the delay time t_d of the individual delay lines while the circuitry will contribute only a minor additional delay. A delay line designed for a characteristic frequency f_o can be operated at a bit frequency f_b up to the frequency f_o. If the frequency is chosen below f_o, it is important that the pulse width is nevertheless maintained at $1/2f_o$. Deviation to longer or shorter values causes the insertion loss to increase, since less

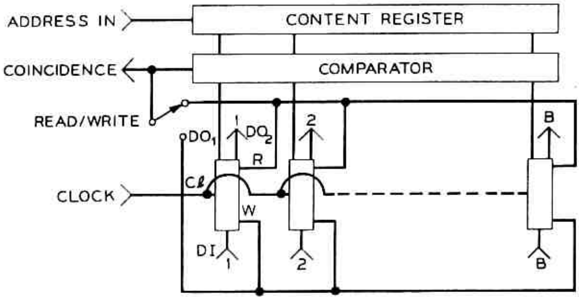


Fig. 7 — Bit-parallel, word-serial organization of B storage loops with associative addressing. A coincidence signal is produced when the content register agrees with a stored word, as the latter passes through the comparator.

of the spectral energy of the input pulse falls into the pass band of the delay line. Storage capacity N and bit rate f_b are interrelated by

$$N = f_b \cdot t_d , \tag{15}$$

so that within certain limits storage capacity can be traded off in order to synchronize the store with an external clock frequency determining the bit rate. However, usually it will be more advantageous to operate the store asynchronously with respect to the outside world. One then avoids the problems in distributing frequencies of 100 MHz over an extended system with a predetermined phase. Moreover, one can then slave the store clock to a delay line used as a timing track.[2]

As in any asynchronous organization, buffer registers must be provided which temporarily store information entered at one transfer rate and withdrawn at a lower one. Information accumulates in the buffer at a rate equal to the difference between the two rates until the complete block is transferred. This determines the buffer capacity required. It is therefore of advantage to know how a delay line store can be accessed at rates nearly equal to those of the equipment it serves, so that the buffer capacity can be kept low.

If bits are loaded every t_ℓ second into a delay line loop of capacity N, bit $(N + 1)$ will coincide in time with the first bit loaded if

$$Nt_\ell = pt_d \tag{16}$$

where $p \geqq 1$ may be any positive integer. However, if p has common factors with N, bits earlier than the $(N + 1)$-st one will already coincide in time, so that such values of p should be avoided. For $N = 1024$, p may thus assume all odd numbers.

Combining (16) and (15) one obtains

$$f_\ell = 1/t_\ell = f_b/p \tag{17}$$

as the permissible loading rates. These generally cause the bits to be stored internally in a scrambled sequence, with the exception of the choices

$$p = qN + 1 \qquad q = 1, 2, 3, \cdots, \tag{18}$$

which cause storage in the original sequence, and

$$p = qN - 1 \qquad q = 1, 2, 3, \cdots, \tag{19}$$

which cause storage in time-inverted sequence. Equation (18) is the basis of the well known delay line time compression (DELTIC) signal processing systems described in the literature in which digital data are written into storage at a slow rate f_ℓ but are read at the fast bit rate $f_b = (qN + 1)f_\ell$.[6] Also, all pairs of loading rates $f_{\ell 1}$ and $f_{\ell 2}$ given by

$$f_{\ell 1}/f_{\ell 2} = p/(qN + p) \tag{20}$$

cause storage in the same, although internally scrambled, sequence so that the bit addresses can be set by sequential counting rather than address comparison.

Series-parallel conversion can be utilized to keep the difference of transfer rates between stores at a minimum. For instance, certain core memories provide access to, say, 72 bits every cycle with a cycle time of 1 microsecond. Parallel to serial conversion would make use of a bit rate of 72 MHz quite feasible, which would increase even further if various checking bits were added to each 72-bit word.

VI. CONCLUSIONS

Sequential storage in recirculating loops, consisting of ultrasonic delay lines in combination with integrated regeneration circuitry of medium scale complexity, has reached a state of the art where bit rates can be obtained which are, at present, unattainable with large scale integrated circuit registers. At bit rates around 100 MHz and beyond such storage appears economically competitive with LSI implementations operating at much lower speeds. Such bit rates have already been demonstrated.

With modules storing blocks of 1000 bits at 100 MHz bit rate, larger stores can be built with latency times of about 10 microseconds, for possible use as main memory in small computers or as fast transfer stores shuttling information between a slow external bulk memory and a fast random access memory in large computers.

In spite of the inherently fixed recirculation time of such a store, it should be adaptable to various processor and bulk store speeds without having to provide more than a few words of buffer storage by a judicious combination of measures such as clock frequency variation, DELTIC modes, and series-parallel conversion. This opens the prospect of using only one or a few basic types of storage modules with a standardized delay time so that maximum advantage could be taken of the savings inherent in mass fabrication.

BIBLIOGRAPHY

1. Sittig, E. K., "High Speed Ultrasonic Delay Line Design: A Restatement of Some Basic Considerations," Proc. IEEE, *56,* No. 7 (July 1968), pp. 1194–1202.
2. Fuss, P. S., "Delay Line Memory and Time Compression System," IEEE Int. Conv. Record, part 2 (March 1967), pp. 94–98.
3. Sittig, E. K. and Cook, H. D., "A Method for Preparing and Bonding Ultrasonic Transducers Used in High Frequency Digital Delay Lines," Proc. IEEE, *56,* No. 8 (August 1968), pp. 1375–1376.
4. Chapin, D. E., "Frequency and Temperature Dependence of Shear Wave Attenuation in Bausch & Lomb T40 Glass," IEEE Trans. Sonics & Ultrasonics, *SU-15,* No. 3 (July 1968), pp. 178–181.
5. Heiter, G. L. and Young, E. H., "A 100 MHz, 1024 Bit Recirculating Ultrasonic Delay Line Store," Digest Solid-State Circuits Conf., Paper FAM 10.3, Philadelphia, Pennsylvania, February 16, 1968.
6. Allen, W. B. and Westerfield, E. C., "Digital Compressed-Time Correlators and Matched Filters for Active Sonar," J. Acoust. Soc. Amer., *36,* No. 1 (January 1964), pp. 121–139.

Queues Served in Cyclic Order

By R. B. COOPER and G. MURRAY*

(Manuscript received September 30, 1968)

We study two models of a system of queues served in cyclic order by a single server. In each model, the i*th queue is characterized by general service time distribution function* $\mathrm{H}_i(\cdot)$ *and Poisson input with parameter* λ_i .

In the exhaustive service model, the server continues to serve a particular queue until the server becomes idle and there are no units waiting in that queue; at this time the server advances to and immediately starts service on the next nonempty queue in the cyclic order.

The gating model differs from the exhaustive service model in that when the server advances to a nonempty queue, a gate closes behind the waiting units. Only those units waiting in front of the gate are served during this cycle, with the service of subsequent arrivals deferred to the next cycle.

We find expressions for the mean number of units in a queue at the instant it starts service, the mean cycle time, and the Laplace-Stieltjes transform of the cycle time distribution function.

I. INTRODUCTION

We consider a system of queues served in cyclic order by a single server. The ith queue is characterized by general service time distribution function $H_i(\cdot)$ and Poisson input with parameter λ_i.

We study two variations of this model. In the first, called the exhaustive service model, the process begins with the arrival of a unit at some queue, say A, when the system is otherwise empty. The server begins on this unit immediately, and continues to serve queue A until for the first time the server becomes idle and there are no units waiting in queue A. The server then looks at the next queue in the cyclic order, queue $A + 1$, and serves those units, if any, that have accumulated during the serving period of queue A. The server continues to serve queue $A + 1$ until for the first time the server

[*] The RAND Corporation.

becomes idle and there are no units waiting in queue $A + 1$. The process continues in this manner, with the queues being served in cyclic order, until for the first time the system becomes completely empty. The process is then re-initiated by the arrival of the next unit. No time is required to switch from one queue to the next.

The second variation, called the gating model, differs from the first in the following way: When the server moves to a queue with at least one waiting unit, the server accepts only those units that were waiting when the server arrived, deferring service of all subsequent arriving units until the next cycle. That is, in the gating model, at the instant the server advances to a nonempty queue a gate closes behind the waiting units, and only those units waiting in front of the gate are served during that cycle.

The exhaustive service model is analyzed in detail. We obtain the generating function for the joint probability distribution of the number of units in each queue at an instant at which the server finishes serving any one of the queues. We then obtain expressions for the mean number of units in a queue at the instant it starts service, the mean cycle time, and, in a form suitable for numerical computation, the Laplace-Stieltjes transform of the cycle time distribution function.

Finally, we note that the equations describing the gating model differ only trivially from those of the exhaustive service model, and that the same method of solution applies to each.

Systems in which a single server is shared among several queues are common. For example, in the No. 1 Electronic Switching System the central control spends much of its time polling various hoppers and performing work requests that it finds in these hoppers. Similarly, in a time-shared computer system the users have access through teletypewriters to a central computer which is shared among them. The cyclic queueing models studied here are of a type which may be useful in the analyses of these and similar problems.

The exhaustive service model for the special case of two queues has been studied by L. Takács,[1] B. Avi-Itzhak, W. L. Maxwell, and L. W. Miller,[2,3] and M. F. Neuts and M. Yadin.[4] Avi-Itzhak *et al* used an argument based on the properties of mean values, to obtain an expression for the mean waiting time suffered by a unit in either queue. Takács, by a more direct argument, utilizing the Markov chain imbedded at the epochs of service completion, obtained the corresponding Laplace-Stieltjes transform and formulas for the waiting time moments assuming service in order of arrival. Neuts and Yadin obtained waiting time results for the transient case.

The more general two-queue model in which the time required to switch from one queue to the next has some arbitrary distribution function is also studied in Ref. 3, and has been investigated in addition by M. Eisenberg[5] and J. S. Sykes.[6] M. A. Leibowitz[7,8] has studied a multiqueue model similar to the gating model studied here. A nonprobabilistic approach to cyclic queueing problems has been used by J. B. Kruskal.[9]

In the present paper we use the imbedded Markov chain approach but, as with Neuts and Yadin, our chain is imbedded at the instants at which the server completes serving a queue, rather than at the set of all instants of service completion used by Takács. Whereas Takács and Neuts and Yadin obtained waiting time results, our analysis yields cycle time results. The mathematical analyses characterizing the three approaches share some common ground, although the differences, especially those arising from our consideration of an arbitrary number of queues, are significant.

Also, in a recent nontechnical article on queues by Leibowitz,[10] the present problem is offered as a prime example of an important, difficult, unsolved queueing problem.

II. PRELIMINARIES

In the analysis of the exhaustive service model, we take the number of queues to be $N + 1 \geqq 2$. Units arrive at the ith queue according to the Poisson process with rate λ_i; that is, the probability $Q_i(k;t)$ that k units arrive at the ith queue in an interval of length t is

$$Q_i(k;t) = \frac{(\lambda_i t)^k}{k!} \exp(-\lambda_i t) \qquad (k = 0, 1, 2, \cdots; i = 0, 1, \cdots, N).$$

The length of time required to serve a unit from queue i has distribution function $H_i(\cdot)$ with mean h_i $(i = 0, 1, \cdots, N)$.

In the analysis of the exhaustive service model, we shall use the concept of busy period, discussed at length by Takács[11]. For the ordinary single-server queue, the busy period is defined as the length of time from the instant a unit enters a previously empty system until the next instant at which the system is completely empty. Both the distribution function of the busy period and its Laplace–Stieltjes transform are known explicitly for the $M/G/1$ queue. In particular, the $M/G/1$ queue with arrival rate λ and mean service time h has a busy period with mean $b = h/(1 - \lambda h)$ if $\lambda h < 1$ and $b = \infty$ if $\lambda h \geqq 1$.

Consider now the $M/G/1$ queue with j waiting units; define the

j-busy period as the length of time from the instant at which service starts on the first of the j units until the next instant at which the system is completely empty. (When $j = 1$, the j-busy period and the busy period are identical.) Each of the j units, which together generate a j-busy period, individually generates a 1-busy period. Thus (as Takács shows) the distribution function of the j-busy period is the j-fold convolution with itself of the distribution function of the 1-busy period.

Denote by $B_i(\cdot)$ the distribution function of a 1-busy period for queue i, by $\beta_i(\cdot)$ its Laplace-Stieltjes transform, and by $b_i = h_i/(1 - \lambda_i h_i)$ its mean. Let $B_i^{*j}(\cdot)$ be the j-fold convolution of $B_i(\cdot)$ with itself, $B^{*1}_i(\cdot) = B_i(\cdot)$. Then a j-busy period for the ith queue has distribution function $B_i^{*j}(\cdot)$ and Laplace-Stieltjes transform $(\beta_i(\cdot))^j$.

III. FORMULATION OF IMBEDDED MARKOV CHAIN STATE EQUATIONS FOR THE EXHAUSTIVE SERVICE MODEL

There are $N + 1 \geqq 2$ queues. Suppose that the system is idle and a unit arrives at some queue at epoch τ_0. The server immediately commences service at that queue, and continues to serve units at that queue until the first instant τ_1 at which that queue becomes empty. If the system is not empty at τ_1, the server advances to the next queue in the cyclic order. The server immediately commences work at this queue until the instant τ_2 at which this queue becomes empty (where $\tau_2 = \tau_1$ if the server finds the queue empty), and continues on in this manner until for the first time, τ_n say, the server finishes serving a queue and there are no units waiting anywhere in the system. The process terminates at τ_n and is reinitiated by the next arrival.

Thus the process generates a set of points $\tau_0, \tau_1, \cdots, \tau_n$, where τ_0 is the arrival instant of a unit at some queue in the previously empty system, and τ_n is the first instant at which the system becomes completely empty again. The next arrival, at epoch $\tau_{0'}$ say, reinitiates the process, and a new set of points, $\tau_{0'}, \tau_{1'}, \cdots, \tau_{n'}$ is generated. We call the points $\tau_1, \cdots, \tau_n$ (and $\tau_{1'}, \cdots, \tau_{n'}$) switch points.

Note that τ_0 is not a switch point, whereas τ_n is a switch point. Successive switch points may occur simultaneously in time, but are nevertheless considered distinct. Thus, with each switch point is associated a queue, namely, that queue at which the server has just completed its visit.

When the server finishes serving a queue and finds the system completely empty, a switch point associated with that queue is recorded. The next switch point is recorded when the server leaves the queue at which the process is reinitiated, and is associated with that queue.

Let $(i; n_1, \cdots, n_N)$ denote the state of the system at an arbitrary switch point, where i is the index of the associated queue, and n_k is the number of units waiting in queue $i + k (k = 1, \cdots, N)$. [For simplicity, no special notation will be used to denote arithmetic mod $(N + 1)$.] Let the state $(i; n_1, \cdots, n_N)$ have probability $P_i(n_1, \cdots, n_N)$; that is, $P_i(n_1, \cdots, n_N)$ is the joint probability that at a switch point, the server has just completed a visit to queue i $(i = 0, 1, \cdots, N)$ and n_1 units are waiting in queue $i + 1$, n_2 units in queue $i + 2$, $\cdots$, and n_N units in queue $i + N$.

The state $(i; n_1, \cdots, n_N)$ can occur through the following exhaustive and mutually exclusive contingencies:

(*i*) The server leaves queue $i - 1$ and finds $j \geqq 1$ units waiting for service in queue i, where it thus spends a length of time equal to a j-busy period.

(*ii*) The server leaves queue $i - 1$ and finds $j = 0$ units waiting for service in queue i, but at least one unit waiting for service somewhere else in the system, so that the server then "passes through" queue i in zero time. [That is, the state $(i; n_1, \cdots n_{N-1}, 0)$ necessarily follows the state $(i - 1; 0, n_1, \cdots, n_{N-1})$ where at least one of the $n_k \neq 0 (k = 1, \cdots, N-1)$.]

(*iii*) The server leaves some queue and finds no units waiting anywhere in the system. With probability λ_i/λ $(\lambda = \lambda_0 + \cdots + \lambda_N)$ the next arrival (which reinitiates the process) occurs at queue i, where the server then spends a 1-busy period.

These considerations lead directly to the imbedded (at the switch points) Markov chain probability state equations:

$$
\begin{aligned}
P_i(n_1, \cdots, n_N)
&= \sum_{j=1}^{\infty} \sum_{k_1=0}^{n_1} \cdots \sum_{k_{N-1}=0}^{n_{N-1}} P_{i-1}(j, k_1, \cdots, k_{N-1}) \\
&\quad \cdot \int_0^{\infty} \prod_{m=1}^{N-1} Q_{i+m}(n_m - k_m; t) Q_{i+N}(n_N; t) \, dB_i^{*j}(t) \\
&\quad + P_{i-1}(0, n_1, \cdots, n_{N-1}) \left(1 - \delta\left(\sum_{m=1}^{N-1} n_m\right)\right) \delta(n_N) \\
&\quad + \frac{\lambda_i}{\lambda} \sum_{k=0}^{N} P_k(0, \cdots, 0) \int_0^{\infty} \prod_{m=1}^{N} Q_{i+m}(n_m; t) \, dB_i(t) \\
&\quad \cdot \left[\delta(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x \neq 0 \end{cases}; \qquad i = 0, 1, \cdots, N \right] \cdot
\end{aligned}
\tag{1}
$$

[Throughout the analysis, arithmetic mod $(N + 1)$ in subscripts will not be specially denoted.] Assuming it exists, the distribution $\{P_i(n_1, \cdots, n_N)\}$ is uniquely determined by (1) and the normalization equation

$$\sum_{i=0}^{N} \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} P_i(n_1, \cdots, n_N) = 1. \tag{2}$$

(Intuitively, one would expect a unique stationary distribution to exist when $\sum_{i=0}^{N} \lambda_i h_i < 1$.)

IV. FUNCTIONAL EQUATIONS FOR GENERATING FUNCTIONS

We define the probability generating functions $g_i(x_1, \cdots, x_N)$:

$$g_i(x_1, \cdots, x_N) = \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} P_i(n_1, \cdots, n_N) x_1^{n_1} \cdots x_N^{n_N} \qquad (i = 0, 1, \cdots, N). \tag{3}$$

Substitution of (1) into (3) yields, after some rearrangement,

$$\begin{aligned} g_i(x_1, \cdots, x_N) &= \sum_{j=1}^{\infty} \sum_{k_1=0}^{\infty} \cdots \sum_{k_{N-1}=0}^{\infty} P_{i-1}(j, k_1, \cdots, k_{N-1}) x_1^{k_1} \cdots x_{N-1}^{k_{N-1}} \\ &\quad \cdot \int_0^{\infty} \exp\left(-t \sum_{m=1}^{N} \lambda_{i+m}(1 - x_m)\right) dB_i^{*i}(t) \\ &\quad + \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} \left(1 - \delta\left(\sum_{m=1}^{N-1} n_m\right)\right) \\ &\quad \cdot \delta(n_N) P_{i-1}(0, n_1, \cdots, n_{N-1}) x_1^{n_1} \cdots x_N^{n_N} \\ &\quad + \frac{\lambda_i}{\lambda} \sum_{k=0}^{N} P_k(0, \cdots, 0) \int_0^{\infty} \exp\left(-t \sum_{m=1}^{N} \lambda_{i+m}(1 - x_m)\right) dB_i(t) \\ &\qquad (i = 0, 1, \cdots, N). \end{aligned} \tag{4}$$

The integrals on the right side of (4) are recognized as the Laplace–Stieltjes transform $(\beta_i(\cdot))^j$ of the j-busy period distribution function with argument $\sum_{m=1}^{N} \lambda_{i+m}(1 - x_m)$. Hence (4) yields the set of simultaneous functional equations

$$\begin{aligned} g_i(x_1, \cdots, x_N) = g_{i-1}\left(\beta_i\left(\sum_{m=1}^{N} \lambda_{i+m}(1 - x_m)\right), x_1, \cdots, x_{N-1}\right) \\ + \frac{\lambda_i}{\lambda} \beta_i\left(\sum_{m=1}^{N} \lambda_{i+m}(1 - x_m)\right) \sum_{k=0}^{N} P_k(0, \cdots, 0) \\ - P_{i-1}(0, \cdots, 0) \qquad (i = 0, 1, \cdots, N). \end{aligned} \tag{5}$$

V. SOLUTION OF THE FUNCTIONAL EQUATIONS

For notational convenience, we define the nesting operator Ξ for any sequence of functions $\{f_k(\cdot)\}$ for which it is meaningful:

$$\mathop{\Xi}_{k=0}^{n} f_k(x) = f_n(\cdots(f_2(f_1(f_0(x))))\cdots).$$

We shall denote by $\mathbf{x}$ the vector with components $x_1, \cdots, x_N$, and by $\mathbf{0}$ the vector with all components zero. Both vectors and vector-valued functions will be denoted by boldface type, and square brackets will be used to enclose vector arguments of vector-valued functions. Finally, we will denote by $\phi(\mathbf{v})$ the first component of a vector $\mathbf{v}$.

Define the vector functions

$$\mathbf{Z}_i[x_1, \cdots, x_N] = \left[\beta_i\left(\sum_{m=1}^{N} \lambda_{i+m}(1 - x_m)\right), x_1, \cdots, x_{N-1}\right] \qquad (i = 0, 1, \cdots, N) \tag{6}$$

so that (5) can be rewritten

$$g_i(\mathbf{x}) = g_{i-1}(\mathbf{Z}_i[\mathbf{x}]) + \frac{\lambda_i}{\lambda}\phi(\mathbf{Z}_i[\mathbf{x}]) \sum_{k=0}^{N} P_k(\mathbf{0}) - P_{i-1}(\mathbf{0}) \qquad (i = 0, 1, \cdots, N). \tag{7}$$

Iterating $\nu - 1$ times on i in (7) we obtain

$$g_i(\mathbf{x}) = g_{i-\nu}\left(\mathop{\Xi}_{k=0}^{\nu-1} \mathbf{Z}_{i-k}[\mathbf{x}]\right) + \frac{1}{\lambda} \sum_{k=0}^{N} P_k(\mathbf{0}) \sum_{m=0}^{\nu-1} \lambda_{i-m}\phi\left(\mathop{\Xi}_{k=0}^{m} \mathbf{Z}_{i-k}[\mathbf{x}]\right) - \sum_{m=1}^{\nu} P_{i-m}(\mathbf{0}) \qquad (i = 0, 1, \cdots, N; \nu = 1, 2, \cdots). \tag{8}$$

In particular, when $\nu = N + 1$ (8) can be written

$$g_i\left(\mathop{\Xi}_{k=0}^{N} \mathbf{Z}_{i-k}[\mathbf{x}]\right) - g_i(\mathbf{x}) = P(\mathbf{0})\left(1 - \frac{1}{\lambda} \sum_{m=0}^{N} \lambda_{i-m}\phi\left(\mathop{\Xi}_{k=0}^{m} \mathbf{Z}_{i-k}[\mathbf{x}]\right)\right) \qquad (i = 0, 1, \cdots, N) \tag{9}$$

where we have set

$$P(\mathbf{0}) = \sum_{k=0}^{N} P_k(\mathbf{0}). \tag{10}$$

We shall now solve (9) by extending a method devised by M. F. Neuts[12] for the solution of a related equation in one variable.

Define the iteration procedure

$$\mathbf{V}_i^{(j)}[\mathbf{x}] = \mathop{\Xi}_{k=0}^{N} \mathbf{Z}_{i-k}[\mathbf{V}_i^{(j-1)}[\mathbf{x}]]$$

$$(i = 0, 1, \cdots, N; j = 1, 2, \cdots; \mathbf{V}_i^{(0)}[\mathbf{x}] = \mathbf{x}). \qquad (11)$$

Using (11) in (9) gives

$$g_i(\mathbf{V}_i^{(j)}[\mathbf{x}]) - g_i(\mathbf{V}_i^{(j-1)}[\mathbf{x}]) = P(\mathbf{0})\left(1 - \frac{1}{\lambda} \sum_{m=0}^{N} \lambda_{i-m}\phi\left(\mathop{\Xi}_{k=0}^{m} \mathbf{Z}_{i-k}[\mathbf{V}_i^{(j-1)}[\mathbf{x}]]\right)\right)$$

$$(i = 0, 1, \cdots, N; j = 1, 2, \cdots). \qquad (12)$$

Adding equations (12) for $j = 1, 2, \ldots, n$ yields

$$g_i(\mathbf{V}_i^{(n)}[\mathbf{x}]) - g_i(\mathbf{x}) = P(\mathbf{0}) \sum_{j=0}^{n-1} \left(1 - \frac{1}{\lambda} \sum_{m=0}^{N} \lambda_{i-m}\phi\left(\mathop{\Xi}_{k=0}^{m} \mathbf{Z}_{i-k}[\mathbf{V}_i^{(j)}[\mathbf{x}]]\right)\right)$$

$$(i = 0, 1, \cdots, N; n = 1, 2, \cdots). \qquad (13)$$

Now let $n \to \infty$ in (13). We will show in the next section that

$$\lim_{n \to \infty} \mathbf{V}_i^{(n)}[\mathbf{x}] = \mathbf{1} \qquad (x_1 \leqq 1, \cdots, x_N \leqq 1; i = 0, 1, \cdots, N) \qquad (14)$$

where $\mathbf{1} = [1, 1, \cdots, 1]$, so that (13) becomes

$$g_i(\mathbf{1}) - g_i(\mathbf{x}) = P(\mathbf{0}) \sum_{j=0}^{\infty} \left(1 - \frac{1}{\lambda} \sum_{m=0}^{N} \lambda_{i-m}\phi\left(\mathop{\Xi}_{k=0}^{m} \mathbf{Z}_{i-k}[\mathbf{V}_i^{(j)}[\mathbf{x}]]\right)\right)$$

$$(i = 0, 1, \cdots, N). \qquad (15)$$

Notice that

$$\sum_{i=0}^{N} g_i(\mathbf{1}) = 1 \qquad (16)$$

and

$$\sum_{i=0}^{N} g_i(\mathbf{0}) = P(\mathbf{0}) \qquad (17)$$

so that upon setting $\mathbf{x} = \mathbf{0}$ in (15) and adding for $i = 0, 1, \cdots, N$ we obtain

$$P(\mathbf{0}) = \left(1 + \sum_{i=0}^{N} A_i(\mathbf{0})\right)^{-1} \qquad (18)$$

where

$$A_i(\mathbf{x}) = \sum_{j=0}^{\infty} \left(1 - \frac{1}{\lambda} \sum_{m=0}^{N} \lambda_{i-m} \phi\left(\mathop{\Xi}_{k=0}^{m} \mathbf{Z}_{i-k}[\mathbf{V}_i^{(j)}[\mathbf{x}]]\right)\right)$$

$$(i = 0, 1, \cdots, N). \qquad (19)$$

(We remark that $P(0) \neq 1 - \sum_{i=0}^{N} \lambda_i h_i$ because the set of switch points is not an arbitrary subset of the set of all points at which units leave the server.)

It remains to calculate $g_i(\mathbf{1})$. Physically, $g_i(\mathbf{1})$ is the probability that at the instant the server leaves some queue, that queue is queue i. This event occurs if

(*i*) the last time the server left a queue the system was empty, and the next arrival occurred at queue i, or

(*ii*) the last time the server left a queue the system was not empty, and the queue was queue $i - 1$.

Event (*i*) has probability $(\lambda_i/\lambda)P(\mathbf{0})$; event (*ii*) has probability $g_{i-1}(\mathbf{1}) - g_{i-1}(\mathbf{0})$. Hence

$$g_i(\mathbf{1}) = \frac{\lambda_i}{\lambda} P(\mathbf{0}) + (g_{i-1}(\mathbf{1}) - g_{i-1}(\mathbf{0})) \qquad (i = 0, 1, \cdots, N). \qquad (20)$$

[Equation (20) can also be obtained directly from (7) with $\mathbf{x} = \mathbf{1}$.] But the difference $(g_{i-1}(\mathbf{1}) - g_{i-1}(\mathbf{0}))$ can be evaluated from (15) with $\mathbf{x} = \mathbf{0}$. Hence

$$g_i(\mathbf{1}) = \frac{\lambda_i}{\lambda} P(\mathbf{0}) + P(\mathbf{0}) A_{i-1}(\mathbf{0}) \qquad (i = 0, 1, \cdots, N) \qquad (21)$$

so that (15) can be rewritten

$$g_i(\mathbf{x}) = \frac{\dfrac{\lambda_i}{\lambda} + A_{i-1}(\mathbf{0}) - A_i(\mathbf{x})}{1 + \displaystyle\sum_{j=0}^{N} A_j(\mathbf{0})} \qquad (i = 0, 1, \cdots, N). \qquad (22)$$

The quantities on the right side of (22) are completely specified; the set of simultaneous functional equations (5) has been solved in the sense that $g_i(\mathbf{x})$ may be calculated for any $\mathbf{x} \leqq \mathbf{1}$.

VI. PROOF OF CONVERGENCE

We wish to prove statement (14):

$$\lim_{n\to\infty} \mathbf{V}_i^{(n)}[\mathbf{x}] = \mathbf{1} \qquad (x_1 \leqq 1, \cdots, x_N \leqq 1; i = 0, 1, \cdots, N).$$

Note first that $\mathbf{V}_i^{(n)}[\mathbf{x}]$ is a vector whose $(N + 1 - m)$th element $(m = 1, 2, \cdots, N)$ is

$$\phi\left(\overset{m}{\underset{k=0}{\Xi}} \mathbf{Z}_{i-k}[\mathbf{V}_i^{(n-1)}[\mathbf{x}]]\right).$$

Therefore, we need show only that

$$\lim_{n\to\infty} \phi\left(\overset{m}{\underset{k=0}{\Xi}} \mathbf{Z}_{i-k}[\mathbf{V}_i^{(n)}[\mathbf{x}]]\right) = 1$$

$$(i = 0, 1, \cdots, N;\ m = 1, 2, \cdots, N;\ x_1 \leqq 1, \cdots, x_N \leqq 1). \qquad (23)$$

From the definition (11) it is clear that the sequence $\{\mathbf{V}_i^{(n)}[\mathbf{x}]\}$ is bounded as $n \to \infty$ for $\mathbf{x} \leqq \mathbf{1}$, and therefore the sequence $\{g_i(\mathbf{V}_i^{(n)}[\mathbf{x}])\}$ is bounded as $n \to \infty$ for $\mathbf{x} \leqq \mathbf{1}$. Also,

$$0 \leqq \phi\left(\overset{m}{\underset{k=0}{\Xi}} \mathbf{Z}_{i-k}[\mathbf{V}_i^{(n)}[\mathbf{x}]]\right) \leqq 1 \qquad (n > 1, \mathbf{x} \leqq \mathbf{1}). \qquad (24)$$

We now turn our attention to equation (13). From (24) we see that the right side of (13) increases monotonically with n for $\mathbf{x} \leqq \mathbf{1}$, and therefore the sequence $\{g_i(\mathbf{V}_i^{(n)}[\mathbf{x}])\}$ increases monotonically with n for $\mathbf{x} \leqq \mathbf{1}$. Thus the sequence $\{g_i(\mathbf{V}_i^{(n)}[\mathbf{x}])\}$ is monotonically increasing and bounded for $\mathbf{x} \leqq \mathbf{1}$, and therefore has a limit. Hence the left side of (13) has a limit, which implies that the series of nonnegative terms on the right side of (13) converges. This in turn implies that

$$\lim_{n\to\infty} \frac{1}{\lambda} \sum_{m=0}^{N} \lambda_{i-m} \phi\left(\overset{m}{\underset{k=0}{\Xi}} \mathbf{Z}_{i-k}[\mathbf{V}_i^{(n)}[\mathbf{x}]]\right) = 1 \qquad (\mathbf{x} \leqq \mathbf{1}). \qquad (25)$$

Statements (24) and (25) together imply (23), completing the proof.

VII. MEAN NUMBERS OF WAITING UNITS

Denote by $\bar{n}_i(k)$ the mean number of units waiting in queue $i + k$ when the server leaves queue i $(i = 0, 1, \cdots, N;\ k = 0, 1, \cdots, N;\ \bar{n}_i(0) = 0)$. For convenience, let $g_i(\mathbf{1})\bar{n}_i(k) = \bar{m}_i(k)$ and $\bar{m}_i(1) = \bar{m}_i$. Then $\sum_{i=0}^{N} \bar{m}_i$ is the mean number of waiting units found by the server in the next queue in cyclic order at a switch point, and $\sum_{i=0}^{N} \bar{m}_i(k - i)$ is the mean number of waiting units in queue k at a switch point. We shall evaluate $\bar{m}_i(k)$ $(i = 0, 1, \cdots, N;\ k = 1, 2, \cdots, N)$.

We first note that $\bar{m}_i(k)$ is given by

$$\bar{m}_i(k) = \frac{\partial}{\partial x_k} g_i(x_1, \cdots, x_N) \bigg|_{x_1 = \cdots = x_N = 1}$$

$$(i = 0, 1, \cdots, N;\ k = 1, 2, \cdots, N) \qquad (26)$$

and the mean 1-busy period $b_i = h_i/(1 - \lambda_i h_i)$ generated by a unit in queue i is given by

$$b_i = -\frac{d}{ds}\beta_i(s)\bigg|_{s=0} \qquad (i = 0, 1, \cdots, N). \tag{27}$$

Differentiating through (5) we obtain

$$\begin{aligned}\frac{\partial}{\partial x_k} g_i(x_1, \cdots, x_N) &\\ = \frac{\partial}{\partial x_k}\beta_i\Big(\sum_{m=1}^{N}\lambda_{i+m}(1 - x_m)\Big)&\frac{\partial}{\partial\beta_i} g_{i-1}(\beta_i, x_1, \cdots, x_{N-1})\\ &+ (1 - \delta(N-k))\frac{\partial}{\partial x_k} g_{i-1}(\beta_i, x_1, \cdots, x_{N-1})\\ &+ \frac{\lambda_i}{\lambda}P(\mathbf{0})\frac{\partial}{\partial x_k}\beta_i\Big(\sum_{m=1}^{N}\lambda_{i+m}(1 - x_m)\Big)\\ &(i = 0, 1, \cdots, N; k = 1, 2, \cdots, N)\end{aligned} \tag{28}$$

which upon setting $x_1 = \ldots = x_N = 1$ gives the two-dimensional set of linear equations

$$\bar{m}_i(k) = \lambda_{i+k}b_i\bar{m}_{i-1} + \frac{\lambda_i}{\lambda}P(\mathbf{0})\lambda_{i+k}b_i + (1 - \delta(N-k))\bar{m}_{i-1}(k+1)$$

$$(i = 0, 1, \cdots, N; k = 1, 2, \cdots, N). \tag{29}$$

For each i, (29) can be solved successively starting with $k = N$ and working backward:

$$\bar{m}_i(N - j) = \lambda_{i+N-j}\sum_{\nu=1}^{j+1} b_{i+1-\nu}\bar{m}_{i-\nu} + \lambda^{-1}P(\mathbf{0})\lambda_{i+N-j}\sum_{\nu=1}^{j+1}\lambda_{i+1-\nu}b_{i+1-\nu}$$

$$(i = 0, 1, \cdots, N; j = 0, 1, \cdots, N - 1). \tag{30}$$

In particular, when $j = N - 1$ equation (30) can be written

$$\bar{m}_i = \lambda_{i+1}\sum_{\nu=i+1}^{i+N} b_{\nu+1}\bar{m}_\nu + \lambda^{-1}P(\mathbf{0})\lambda_{i+1}\sum_{\nu=i+1}^{i+N}\lambda_{\nu+1}b_{\nu+1}$$

$$(i = 0, 1, \cdots, N). \tag{31}$$

When $\lambda_{i+1}b_{i+1}\bar{m}_i$ is added to both sides of the ith equation of the set (31) we have after rearrangement

$$\bar{m}_i(1 + \lambda_{i+1}b_{i+1})\lambda_{i+1}^{-1} - \lambda^{-1}P(\mathbf{0})\sum_{\nu=i+1}^{i+N}\lambda_{\nu+1}b_{\nu+1} = \sum_{\nu=i}^{i+N} b_{\nu+1}\bar{m}_\nu$$

$$(i = 0, 1, \cdots, N). \tag{32}$$

The sum on the right side of (32) is a constant independent of the value of the index i. Hence

$$\bar{m}_i(1 + \lambda_{i+1}b_{i+1})\lambda_{i+1}^{-1} - \lambda^{-1}P(\mathbf{0}) \sum_{\nu=i+1}^{i+N} \lambda_{\nu+1}b_{\nu+1}$$
$$= \bar{m}_{i+j}(1 + \lambda_{i+j+1}b_{i+j+1})\lambda_{i+j+1}^{-1} - \lambda^{-1}P(\mathbf{0}) \sum_{\nu=i+j+1}^{i+j+N} \lambda_{\nu+1}b_{\nu+1}$$
$$(i = 0, 1, \cdots, N;\ j = 0, 1, \cdots, N). \qquad (33)$$

Combining (33) and (31) yields

$$\bar{m}_i = \frac{\lambda_{i+1}}{\lambda} P(\mathbf{0}) \frac{\rho - \rho_{i+1}}{1 - \rho} \qquad (i = 0, 1, \cdots, N) \qquad (34)$$

where we define $\rho_i = \lambda_i h_i$ and $\rho = \sum_{i=0}^{N} \rho_i$. Note that for (34) to be meaningful we must have $\rho < 1$. The $\{\bar{m}_i(k)\}$ can now be calculated from equations (34) and (30).

VIII. LAPLACE-STIELTJES TRANSFORM OF CYCLE TIME DISTRIBUTION FUNCTION

Consider the set of switch points associated with queue i, and append to this set every switch point associated with queue $i - 1$ at which the server finds the system completely empty. Call the elements of this augmented set the record points associated with queue i.

We define the partial cycle time for queue i as the elapsed time between a switch point associated with queue $i - 1$ and the temporally preceding record point associated with queue i. Denote by $\hat{G}_i(\cdot)$ the distribution function of the partial cycle time for the ith queue, and by $\hat{\gamma}_i(\cdot)$ its Laplace–Stieltjes transform.

Since queue i is necessarily empty at an associated record point, all of the units waiting for service in queue i at a switch point of queue $i - 1$ must have arrived during the preceding partial cycle time. Let $P_{i-1}(j)$ be the conditional probability that $j \geqq 0$ units will be waiting for service in queue i, given that a switch point associated with queue $i - 1$ has just occurred. Then the distribution function $\hat{G}_i(\cdot)$ of the partial cycle time for the ith queue and the distribution $\{P_{i-1}(j)\}$ of the number of units that arrive (according to the Poisson process with rate λ_i) during the partial cycle time are related as follows:

$$P_{i-1}(j) = \int_0^\infty \frac{(\lambda_i t)^j}{j!} \exp(-\lambda_i t)\, d\hat{G}_i(t)$$
$$(i = 0, 1, \cdots, N;\ j = 0, 1, \cdots). \qquad (35)$$

Notice also that the distribution $\{P_{i-1}(j)\}$ has probability generating function

$$\sum_{j=0}^{\infty} P_{i-1}(j)x^j = \frac{g_{i-1}(x, 1, \cdots, 1)}{g_{i-1}(\mathbf{1})} \qquad (i = 0, 1, \cdots, N). \qquad (36)$$

Substitution of (35) into (36) yields, for the Laplace–Stieltjes transform $\hat{\gamma}_i(\cdot)$ of the partial cycle time distribution function for queue i,

$$\hat{\gamma}_i(s) = \frac{g_{i-1}\left(\dfrac{\lambda_i - s}{\lambda_i}, 1, \cdots, 1\right)}{g_{i-1}(\mathbf{1})} \qquad (i = 0, 1, \cdots, N). \qquad (37)$$

We define the (full) cycle time for queue i as the partial cycle time plus the time required to serve those units, if any, waiting in queue i when the server finishes queue $i - 1$. (Notice that in order to be counted as a cycle for queue i, a time interval must contain a partial cycle ending at a switch point at queue $i - 1$.) Denote by $G_i(\cdot)$ the distribution function of the cycle time for the ith queue, and by $\gamma_i(\cdot)$ its Laplace–Stieltjes transform.

The cycle time distribution function $G_i(\cdot)$ is related to the partial cycle time distribution function $\hat{G}_i(\cdot)$ as follows:

$$G_i(t) = \int_0^t \sum_{j=0}^{\infty} \frac{(\lambda_i\xi)^j}{j!} \exp(-\lambda_i\xi) B_i^{*j}(t - \xi)\, d\hat{G}_i(\xi)$$
$$(B_i^{*0}(\cdot) = 1;\ i = 0, 1, \cdots, N). \qquad (38)$$

Taking Laplace-Stieltjes transforms throughout (38) we obtain

$$\gamma_i(s) = \hat{\gamma}_i(\lambda_i + s - \lambda_i\beta_i(s)) \qquad (i = 0, 1, \cdots, N). \qquad (39)$$

Hence we have for the Laplace-Stieltjes transform $\gamma_i(s)$ of the cycle time distribution function for the ith queue

$$\gamma_i(s) = \frac{g_{i-1}\left(\dfrac{\lambda_i\beta_i(s) - s}{\lambda_i}, 1, \cdots, 1\right)}{g_{i-1}(\mathbf{1})} \qquad (i = 0, 1, \cdots, N). \qquad (40)$$

By differentiating through (40) we obtain for the mean cycle time $\bar{t}_i$ the intuitively obvious result

$$\bar{t}_i = (b_i + \lambda_i^{-1})\bar{n}_{i-1} \qquad (i = 0, 1, \cdots, N). \qquad (41)$$

IX. THE GATING MODEL

Consider now a system of $N \geqq 1$ cyclic queues described by the gating model of Section I. Define $P_i(n_1, \cdots, n_N)$ as the joint probability that at the instant the server leaves a queue, that queue is queue i ($i = 0, 1, \cdots, N - 1$) *and* n_1 units are waiting in queue $i + 1$, n_2 units in queue $i + 2, \cdots$, and n_N units are waiting in queue i (that is, n_N units arrived at queue i after the closing of the gate). Denote by $H_i^{*j}(\cdot)$ the j-fold convolution with itself of the service time distribution function $H_i(\cdot)$. Then

$$\begin{aligned} P_i(n_1, \cdots, n_N) &= \sum_{j=1}^{\infty} \sum_{k_1=0}^{n_1} \cdots \sum_{k_{N-1}=0}^{n_{N-1}} P_{i-1}(j, k_1, \cdots, k_{N-1}) \int_0^\infty \prod_{m=1}^{N-1} Q_{i+m}(n_m - k_m; t) \\ &\quad \cdot Q_{i+N}(n_N; t)\, dH_i^{*j}(t) + P_{i-1}(0, n_1, \cdots, n_{N-1}) \left(1 - \delta\left(\sum_{m=1}^{N-1} n_m\right)\right) \\ &\quad \cdot \delta(n_N) + \frac{\lambda_i}{\lambda} \sum_{k=0}^{N-1} P_k(0, \cdots, 0) \int_0^\infty \prod_{m=1}^{N} Q_{i+m}(n_m; t)\, dH_i(t) \\ &\qquad (i = 0, 1, \cdots, N - 1) \qquad (42) \end{aligned}$$

where $Q_t = Q_{i+N}$.

Equation (42), for the N-queue gating model, is only trivially different from (1), which describes the $(N + 1)$ − queue exhaustive service model. The analogue of (5) is

$$\begin{aligned} g_i(x_1, \cdots, x_N) &= g_{i-1}\left(\eta_i\left(\sum_{m=1}^{N} \lambda_{i+m}(1 - x_m)\right), x_1, \cdots, x_{N-1}\right) \\ &\quad + \frac{\lambda_i}{\lambda} \eta_i\left(\sum_{m=1}^{N} \lambda_{i+m}(1 - x_m)\right) \sum_{k=0}^{N-1} P_k(0, \cdots, 0) \\ &\quad - P_{i-1}(0, \cdots, 0) \qquad (i = 0, 1, \cdots, N - 1) \qquad (43) \end{aligned}$$

where $\eta_i(\cdot)$ is the Laplace-Stieltjes transform of the distribution function $H_i(\cdot)$, and $g_i(x_1, \ldots, x_N)$ is now the generating function for the gating model state probabilities. The solution of (43) follows that given for (5), and a complete analysis may now be carried out in a manner similar to that employed for the exhaustive service model. (We remark in passing that the equations originally considered by Neuts are those of the gating model with $N = 1$.)

X. SUMMARY

Two models of a system of queues served in cyclic order by a single server have been presented. One of these, the exhaustive service model, has been analyzed in detail. This model is described by the imbedded Markov chain probability state equations (1), from which a set of functional equations (5) for the probability generating functions are derived. The functional equations are solved with the help of a generalization of an iteration procedure used by Neuts. The equations (5) are then used to obtain explicit expressions for various mean values, such as the mean number of units found waiting by the server in the ith queue, given by equation (34), and the mean cycle time, given by equation (41). The Laplace-Stieltjes transform of the cycle time distribution function is given, in a form suitable for numerical computation (and hence numerical inversion), by equation (40).

It is then shown that the gating model is described by state equations only trivially different from those of the exhaustive service model. It is now easy to adapt the methods and results of the detailed analysis of the exhaustive service model to a similar analysis of the gating model.

It is noteworthy that all results are expressed directly in terms of the single state probability $P(\mathbf{0})$ and the relevant generating functions, so that there is no need to evaluate the individual state probabilities. The calculations are thus reduced to the iteration algorithm, which may be suited to digital computer solution.

REFERENCES

1. Takács, L., "Two Queues Attended by a Single Server," Operations Research, *16,* No. 3 (May–June 1968), pp. 639–650.
2. Avi-Itzhak, B., Maxwell, W. L., and Miller, L. W., "Queuing With Alternating Priorities," Operations Research, *13,* No. 2 (March–April 1965), pp. 306–318.
3. Conway, R. W., Maxwell, W. L., and Miller, L. W., *Theory of Scheduling,* New York: Addison-Wesley, 1967.
4. Neuts, M. F. and Yadin, M., "The Transient Behavior of the Queue with Alternating Priorities, with special reference to the Waitingtimes," Mimeo Series No. 136, Dept. of Statistics, Purdue University (January 1968).
5. Eisenberg, M., "Multi-Queues With Changeover Times," MIT Doctoral Dissertation, (September 1967).
6. Sykes, J. S., unpublished work.
7. Leibowitz, M. A., "An Approximate Method for Treating a Class of Multiqueue Problems," IBM J. Research Development, *5,* No. 3 (July 1961), pp. 204–209.
8. Saaty, T. L., *Elements of Queueing Theory,* New York: McGraw-Hill, 1961, pp. 298–301.
9. Kruskal, J. B., unpublished work.
10. Leibowitz, M. A., "Queues," Scientific American, *219,* No. 2 (August 1968), pp. 96–103.
11. Takács, L., "Introduction to the Theory of Queues," New York: Oxford University Press, 1962, pp. 32, 57–65.
12. Neuts, M. F., "The Queues With Poisson Input and General Service Times, Treated as a Branching Process," Purdue University, (September 1966), unpublished paper distributed by the Clearinghouse for Federal Scientific and Technical Information, Dept. of Commerce, AD640483.

A Hybrid Coding Scheme for Discrete Memoryless Channels*

By D. D. FALCONER

(Manuscript received June 6, 1968)

We consider a coding-decoding scheme which can permit reliable data communication at rates up to the capacity of a discrete memoryless channel, and which offers a reasonable trade off between performance and complexity. The new scheme embodies algebraic and sequential coding-decoding stages. Data is initially coded by an algebraic (Reed–Solomon) encoder into blocks of N *symbols, each symbol represented by* n *binary digits. The* N n*-bit symbols in a block are transmitted separately and independently through* N *parallel subsystems, each consisting of a sequential coder, an independent discrete memoryless channel, and a sequential decoder in tandem. Those coded* n*-bit symbols which would require the most sequential decoding computations are treated as erasures and decoded by a Reed–Solomon decoder. We show that the hybrid technique reduces the variability of the amount of sequential decoding computation. We also derive asymptotic results for the probabilities of error and buffer overflow as functions of the system complexity.*

I. INTRODUCTION

It is well known that the use of block coding and maximum-likelihood decoding permits transmission of information at rates up to the capacity of a discrete memoryless channel with an error probability which decreases exponentially with the code block length.[1–4] A discrete memoryless channel (DMC) may be an adequate model for some types of real one-way digital communication channels consisting of a transmission medium, transmitting and receiving equipment and modulation-demodulation scheme. An arbitrary DMC is assumed to have

* This research is partly based on a Ph.D. thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, 1966, carried out at the Research Laboratory of Electronics, Massachusetts Institute of Technology, supported by the National Aeronautics and Space Administration (Grant NsG-334) and a Hughes Industrial Fellowship.

a P-symbol input alphabet and a Q-symbol output alphabet. During each channel use, an input symbol is selected and transmitted and an output symbol is received. Successive input-to-output transitions are random and statistically independent; the probability that the output is symbol j ($j = 1, 2, \ldots, Q$), given that the input is symbol i ($i = 1, 2, \ldots, P$), is q_{ij}. (Table I contains a list of the symbols used throughout this paper)

Maximum-likelihood decoding, which is known to be optimum, involves the cross-correlation of a received block code word with all possible transmitted code words. The number of code words, and hence the required number of decoding operations, grows exponentially with the block length; this exponential growth in decoding complexity makes maximum-likelihood decoding impractical, even for moderate block lengths. There has thus been considerable incentive to find suitable classes of codes having nonoptimum decoding schemes, for which the complexity (reflecting the number of components and the number of decoding operations per unit of transmitted information) does not increase exponentially with the block length.

A number of coding-decoding schemes have previously been proposed. Among the most widely known are:

(*i*) Algebraic coding and decoding schemes.[5, 6]
(*ii*) Elias' iterated coding and decoding.[7]
(*iii*) Massey's threshold decoding of convolutional codes.[8]
(*iv*) Gallager's low density parity check codes.[9]
(*v*) Sequential coding and decoding.[10–12]

For some performance-versus-complexity criteria, one or more of these schemes may be well suited. However, lower bounds on the performance and complexity of these schemes show that none can yield an exponentially low error probability for a rate arbitrarily close to channel capacity without incurring exponentially growing complexity; Ziv, Pinsker, and Forney have proposed some more general coding-decoding schemes for use with discrete memoryless channels.[13–16] The common feature of these schemes and of the earlier scheme of Elias is that they incorporate two or more separate stages of coding and decoding as Fig. 1 illustrates.[7] The "inner stage" is an arbitrary block coding-decoding scheme, generally using maximum-likelihood decoding, which has just enough complexity to guarantee a fairly low probability of decoding error. Then the chain consisting of the inner coder, DMC, and inner decoder constitutes another dis-

Table I—List of Symbols

Symbol	Definition
P	Size of channel input alphabet
Q	Size of channel output alphabet
q_{ij}	Transition probability that output is j if input is i
N	Block length of RS code
K	Number of information symbols per RS code word
R	Dimensionless rate of RS code. $R = K/N$
d	Minimum distance of RS code
S	Number of erasures to be corrected per parallel block
T	Maximum number of correctable errors per parallel block
v	Number of channel symbols per tree branch
r	Rate of sequential code in bits per channel use
R_{comp}	Computational cutoff rate
τ	Time interval for transmission of a single channel symbol
n	Number of tree branches per serial block
m	Number of redundant (known) branches per serial block
R'	Overall information rate in bits per channel use
δ	Defined by: $S = N\delta - 1$
$p_u(e)$	Probability of decoding error for one serial block
$p_u'(e)$	Upper bound on $P_u(e)$
A_e, A_c	Constants, for a given sequential code
$E_u(r)$	Sequential decoding error exponent
$p(e)$	Probability of error for a super block
$T_e\,(x, y)$	$= -x\ell ny - (1-x)\,\ell n(1-y)$
$H(x)$	$= -x\ell nx - (1-x)\,\ell n(1-x)$
S_e	Overall block length
c_j	Number of sequential decoding computations to decode the jth super block
p_x	Upper bound on the probability that c_j exceeds x
α	Pareto exponent
α'	$= \max(\alpha, 1)$
C	Number of computation units to decode a given super block
A_1	$= n^{(\alpha'/\alpha)}A_c \exp\,[H(\delta)/\alpha\delta]$
A_2	$= N\delta\alpha/(N\delta\alpha - 1)\;A_c \exp\,[H(\delta)/\alpha\delta]$
B	Size of buffer allotted to each sequential decoder
$p_L(B)$	Probability that buffer overflows before first L super blocks are decoded
q_i	Queue size after ith super block is decoded
X_i	Number of new super blocks joining queue during the decoding of the ith super block
μ	Maximum number of computations each sequential decoder can do per received branch
C_l	Number of computation units to decode the lth super block
D	$= 1 + e^{\delta}$
μ_0	$= \mu n/A_1$
$S^{\dagger}$	Total decoder buffer storage

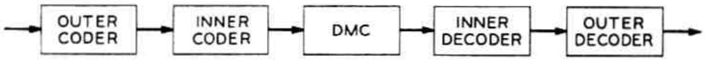


Fig. 1 — Two-stage coding-decoding scheme.

crete channel with a low probability of error or erasure. Scrambling and descrambling may be necessary to make this new channel memoryless. The "outer" stage or stages embody available coding and decoding techniques with long block length, which drive the probability of decoding error down to a negligibly small value with a relatively small degree of complexity. The overall block length is the product of the block lengths of the individual coding stages, and the overall information rate is the product of the individual rates.

The overall block lengths for these schemes are much larger than those known to be necessary to achieve a given error probability with a given information rate. However, this penalty, which is reflected in increased coder complexity, may be compensated for by the more favorable tradeoff between performance and decoder complexity.

These multistage schemes allow transmission at any information rate up to channel capacity with error probabilities which decrease exponentially with overall block length (or its square root in Ziv's scheme); the total decoder complexity may be large but it increases only algebraically with the overall block length. Notice that if the inner stage uses maximum likelihood decoding in order to achieve a low error probability for a rate close to channel capacity, its complexity increases exponentially with its block length. Thus the complexity of the inner stage may well dominate the total complexity, for rates close to capacity.

We propose yet another two-stage coding-decoding scheme, which we call a *hybrid scheme* and which is described in detail in Section II. The inner stage involves sequential coding-decoding, which is known to be capable of yielding exponentially small error probability for any rate less than the channel capacity. The decoding effort required of the inner stage is actually alleviated by the use of the outer stage, which involves algebraic coding-decoding. Section III contains derivations of upper bounds on error probability, distribution of decoding computation, average decoding computation and probability of buffer overflow for the hybrid scheme. These bounds display the asymptotic performance capabilities of the scheme. The bounds are not sufficiently tight to be useful in obtaining detailed performance parameters for actual systems, but must be supplemented by simulations. Section IV contains some simple calculations, based on a previous simulation, for the performance of a hybrid scheme. Before describing the new scheme, we briefly review some salient features of algebraic coding and of sequential coding.

1.1 *Algebraic Coding and Decoding*

Any algebraic code has an underlying algebraic structure, upon which the coding and decoding algorithms are based.[5] For a code with block length N, each code word consists of N symbols picked from a finite field. Thus the symbol alphabet size must be a prime or power of a prime. The channel is assumed to either change a symbol to a different symbol in the field with some probability p (thus making an error) or change it to a symbol not in the field with some probability q (thus making an erasure), or pass the symbol on unchanged with probability $1-p-q$.

Algebraic codes may be put in systematic form; K of the N symbols in a code word are information symbols and the remaining $N-K$ are check symbols. The ratio K/N is the *dimensionless* rate of the code. The required coder complexity is generally proportional to N.

An important property of an algebraic code is its *minimum distance, d,* which is the minimum number of symbols in which any two code words differ. Practical decoding algorithms are available for certain classes of algebraic codes with specified minimum distance properties. These decoding algorithms generally involve a finite number of algebraic (finite field) operations, and guarantee the correction of up to T errors and S erasures for any T and S such that

$$2T + S \leqq d - 1. \tag{1}$$

The best known algebraic block codes are the BCH codes, for which both the number of decoding operations per block and the number of components vary with N approximately as $N \log N$ and with T approximately as $T \log N$, as shown by Berlekamp.[6] A special case of BCH codes, involving roughly the same order of decoder complexity, is the class of Reed-Solomon (RS) Codes.[17, 18] A RS code can be defined with any rate R and block length N, provided that the size of the symbol alphabet exceeds N. It can be shown that a RS code's minimum distance is the largest possible, given R and N, that is

$$d = d_{\max} = (1 - R)N + 1. \tag{2}$$

Reed-Solomon codes are useful where the size of the code's symbol alphabet can be large.

1.2 *Sequential Coding and Decoding*

Sequential coding and decoding is applicable in principle to any DMC. Sequential coding is also known as *tree* coding.[10–12] Included in the class of tree codes are the easily implemented convolutional codes.[10]

A sequential coder accepts a sequence of consecutive binary information digits and, for each, generates v channel input symbols. Coding is sequential; each channel input symbol depends only on previous binary input digits.

Implicit in the structure of a sequential coder is a tree, as typified in Fig. 2 for $v = 3$. Each branch is labeled with v channel input symbols. A sequence of binary inputs to the coder is conceptually a sequence of directions which sequentially steer the coder along a path (called the *correct path*) starting at the origin of the tree. Successive branches along the correct path are transmitted over the DMC as v-tuples of channel symbols. The rate of the tree code in bits per channel use is $r = 1/v$. If a rate $r = u/v$ is required, bits entering the coder would be grouped into u-tuples, and there would be 2^u branches stemming from each node of the tree.

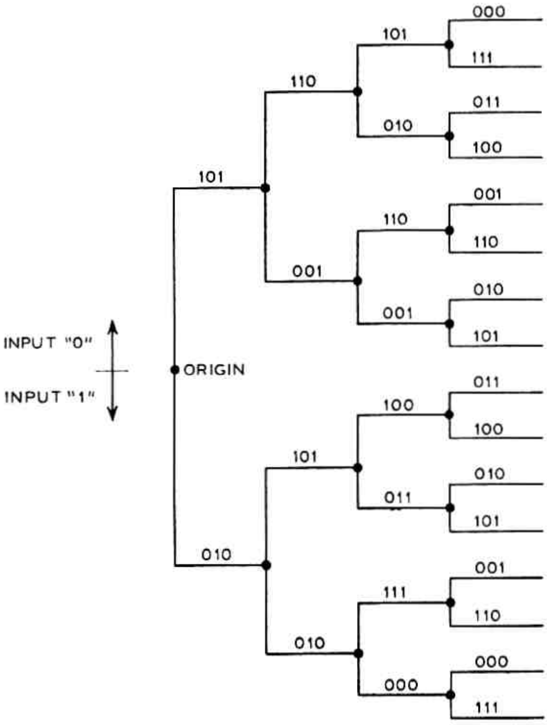


Fig. 2 — Tree structure of a sequential code.

Sequential decoding is a form of *probabilistic decoding,* which is applicable to tree codes. It is termed "probalistic" because the general decoding procedure applies to any randomly selected tree code and because the decoder is guided to a final decision by probabilistic considerations rather than by a fixed sequence of algebraic operations. A sequential decoder implicitly contains a copy of the tree, and must hypothesize a path through the tree, starting at the origin, which with high probability is the correct path.

The Fano sequential decoding algorithm is a specific sequential tree search procedure which is efficient, practical to implement, and is amenable to analysis.[11, 12] The decoder examines received branches successively, makes tentative hypotheses for the corresponding branches of the correct path, and advances along them through the tree, if their likelihood, measured by an appropriate "path metric," appears high enough. If the current hypothesized path appears not sufficiently likely, the decoder retreats one branch and starts searching for a more likely path. Thus there is backward and forward searching through the tree, with a trend toward the right, as the decoder continually extends and revises its estimate of the correct path. If the rate r is less than the capacity of the DMC, the Fano algorithm sequential decoder can be shown to eventually trace out the correct path with high probability.

The number of branch examinations, or computations done by the decoding algorithm to advance one branch deeper into the tree is a random variable. Analysis and simulation have shown that its mean is bounded, independent of the coder complexity, only if the code rate r is less than a "computational cutoff rate," R_{comp}, which is characteristic of the channel and is always less than the channel capacity.

Since the rate of transmission and the decoder's operating speed are fixed, a buffer must be provided at the decoder to store arriving branches which accumulate during periods of intensive tree searching. The buffer is necessarily of finite size, and hence may overflow if a span of received branches requires an unusually large amount of computations. Buffer overflow is catastrophic, since it is accompanied by loss of data and subsequent disruption of the decoding process. It is generally the most prevalent mode of failure in systems which use sequential decoding.

Restarting the decoding process after an overflow occurs is generally possible only if the sequence of transmitted channel symbols is divided into blocks which are coded and decoded independently. That is, at regular intervals, the coder starts afresh at the tree origin and erases its memory of previous information bits. Then if an overflow occurs, decoding can resume at the beginning of the next block.

It will be shown that the hybrid coding-decoding scheme described in the Section II reduces the severe variability in decoding effort that is characteristic of sequential decoding, and furthermore, that for any rate up to channel capacity, the probability of decoding failure (error or overflow) asymptotically decreases nearly exponentially with the total system's complexity.

II. DESCRIPTION OF CODER AND DECODER

2.1 *The Coder*

Figure 3 shows the structure of the hybrid coder. We assume that N parallel independent DMC's are available, each of which is used for transmission once every τ seconds. These N parallel channels could be created by time-multiplexing a single DMC which is used once every τ/N seconds. The input to each DMC is from a separate sequential coder. The code rate is $r = 1/v$ bits per channel use. Every $v\tau$ seconds each sequential coder accepts a binary input digit and generates v successive channel input symbols which, in accordance with the tree structure of the code, depend on present and past coder inputs. However, each coder's memory of past input bits is erased

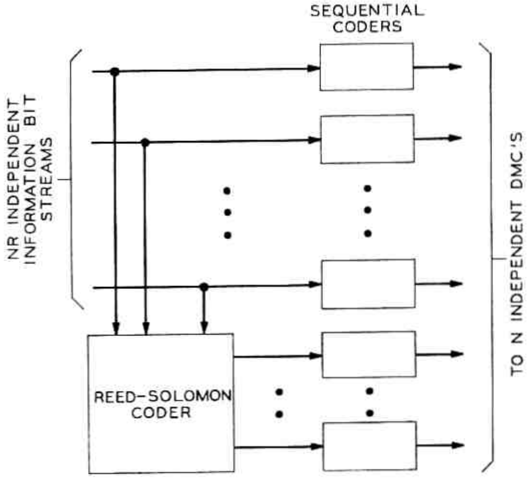


Fig. 3—Hybrid coder structure.

at $nv\tau$-second intervals. Thus, successive blocks of n inputs are coded independently into blocks of nv channel input symbols; such independently coded blocks are called *serial blocks,* and the corresponding blocks of n coder input digits are called *n-symbols.*

If a coder input digit is to be decodable with a low error probability, it must affect a certain minimum number of subsequent channel input symbols. However since the coder's memory of previous inputs is erased at the beginning of each serial block, the final coder input digits in any n-symbol can affect relatively few channel input symbols. The error probability is kept low by making the last m $(m < n)$ digits of each n-symbol a fixed sequence known to the decoder.[12] Then each *a priori* unknown coder input digit can affect at least mv channel input symbols. The last m coder input digits are redundant; the net information rate of each sequential coder is then $(1-m/n/)v$ bits per channel use. In general, n is chosen to be much greater than m, so that the decrease in net rate resulting from the periodic "resynchronization" is acceptably small.

The N serial blocks simultaneously coded and transmitted in parallel over the N DMC's comprise a *super block.* The corresponding set of N n-symbols which enter the coders in parallel is called a *parallel block.* NR of the n-symbols in a parallel block are independent subblocks each consisting of n–m information bits followed by m known bits. The remaining $N(1-R)$ n-symbols in a parallel block are parity check symbols generated from the information n-symbols by an algebraic block coder operating on a field of 2^n elements (that is, the coder operates on n-symbols rather than individual bits). Each n-symbol is made to enter its respective sequential coder serially, as a sequence of binary digits at $v\tau$-second intervals.

A parallel block is thus a member of a block code with block length N and a 2^n-symbol alphabet. The code's dimensionless rate is R, and the number of words in the code is 2^{nNR}.

The overall information rate of the system is

$$R' = R(1 - m/n)/v \text{ bits per channel use.} \tag{3}$$

Since each DMC is used once every τ seconds, the overall information rate is NR'/τ bits per second. A source producing information at this rate would determine which of the 2^{nNR} block code words would be generated in each $nv\tau$-second interval.

For moderate-to-large parallel and serial block lengths (greater than, say 50) the most eligible available block code would be a Reed-

Solomon code, since the required alphabet size is generally large, and RS codes have the largest possible minimum distance for given rate and block length. The alphabet size must be a power of two and must exceed $(N + 1)$. This imposes a constraint on n,

$$n \geqq \log_2 (N + 1). \tag{4}$$

Typically, m might be between 10 and 100, n might be 10 or 20 times m, and N might be between 10 and 1000. Forney[16] has pointed out that if $n = n'I$ (n' and I integers) and $2^{n'} \geqq N$ then a RS code of block length N on a field of 2^n elements can be implemented more simply as I repetitions of a RS code of block length N on the subfield of $2^{n'}$ elements. Use of this smaller field for algebraic operations makes for simpler implementation of the RS coder and decoder. Figure 4 shows the structure of a super block.

The Reed–Solomon coder may be implemented with a number of components proportional to N. Each of the N sequential coders may be realized as a convolutional coder, constructed from at most n shift register stages. Thus, the overall coder complexity is proportional to nN.

2.2 *The Decoder*

Not surprisingly, a decoder appropriate to the two-stage coding scheme just described consists of sequential and algebraic stages, as illustrated in Fig. 5. The first stage consists of N parallel sequential decoders which simultaneously and independently utilize the Fano sequential decoding algorithm to decode serial blocks emerging in

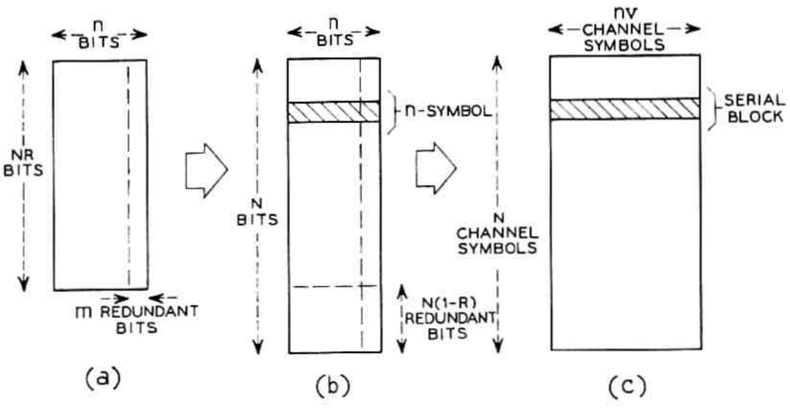


Fig. 4 — Code structure: (a) block of bits entering RS codes, (b) parallel block (output of RS coder), (c) super block (output of sequential coders).

parallel from the N DMC's. This stage might be implemented by a time-sharing technique, in which a single logic unit is allocated to one decoder after another in turn. The second stage is an algebraic decoder for the RS code.

During the decoding of a super block, all N sequential decoders attempt to decode their respective serial blocks into the original input n-symbols. In general, some serial blocks require more computations, and therefore more computing time, than others. After all but some fixed number S ($S < N$) of the N serial blocks have been sequentially decoded, the S sequential decoders still at work are halted, and then all sequential decoders are free to start work on the n-symbols of the following super block.

Meanwhile the present super block is passed on to the RS decoder in the form of a parallel block consisting of $N - S$ sequentially decoded n-symbols and S undecoded n-symbols which are treated as erasures. If the RS code's minimum distance is d, and no more than T of the sequentially decoded n-symbols contain errors, where

$$2T + S = d - 1, \tag{5}$$

then the RS decoder is guaranteed to decode the parallel block correctly, using a fixed number of decoding computations that varies roughly as $N \log N$ and as $T \log N$.[6, 16] In this way, those S serial blocks which normally would be sequentially decoded last are essentially all corrected by the algebraic decoder as soon as the first $(N - S)$ serial blocks have been sequentially decoded. Thus the algebraic decoder's assistance should tend to curtail the very long decoding times which occasional serial blocks may require and should thereby reduce the chances for overflow of the sequential decoders' buffers.

From relation (2), governing the minimum distance of an RS code,

$$2T + S = (1 - R)N; \tag{6}$$

the numbers of correctable errors and erasures are proportional to N, for fixed rate R.

A hybrid scheme closely related to the one described here was described and analyzed in Ref. 19. In that scheme the sequence of channel input symbols is not divided into independently coded serial blocks. Instead, once the sequential decoding algorithm advances a certain fixed number of branches beyond a given n-symbol, that n-symbol is considered irrevocably decoded, and thus is presented to the block decoder as a nonerased symbol in a parallel block. As in the

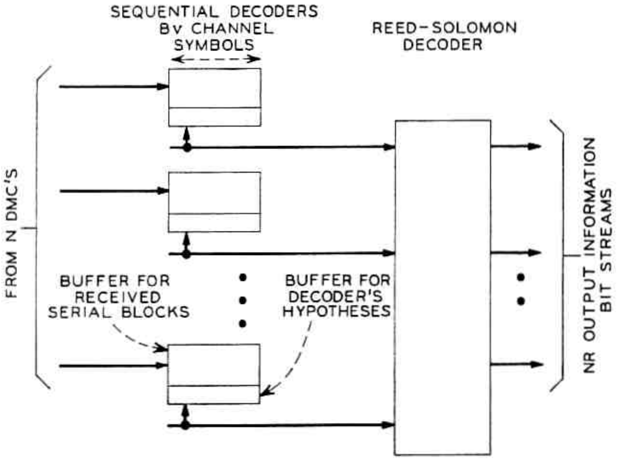

SEQUENTIAL DECODERS
Bv CHANNEL SYMBOLS
REED-SOLOMON DECODER
FROM N DMC'S
BUFFER FOR RECEIVED SERIAL BLOCKS
BUFFER FOR DECODER'S HYPOTHESES
NR OUTPUT INFORMATION BIT STREAMS

Fig. 5 — Hybrid decoder structure.

scheme described here, n-symbols which would require excessive numbers of sequential decoding computations may be decoded by the Reed–Solomon decoder. The asymptotic bounds on computation statistics are essentially similar for both hybrid schemes. The scheme described here appears somewhat more practical to implement. Reference 19 also describes a simulation of the earlier scheme in which there are ten parallel sequential coding-decoding systems, and the block code word rate is either 8/10 or 9/10. The outer stage was intended to correct erasures only. The tail of the observed distribution of sequential decoding computation behaved as predicted by the upper bound of Section 3.2; the frequency of very large peaks of computation was considerably reduced.

III. BOUNDS ON PERFORMANCE AND COMPLEXITY

In deriving bounds on the probability of error, distribution of computation, average computation, and probability of buffer overflow, we assume arbitrarily that the RS decoder corrects $T = N\delta/2 - 1$ errors and $S = N\delta - 1$ erasures per parallel block, where $0 < \delta < \frac{1}{2}$. Half the RS code's minimum distance is then used to correct erasures and half to correct errors. The value of δ is then fixed by (6);

$$\delta = \frac{1-R}{2} + \frac{3}{2N} \geqq \frac{1-R}{2}, \tag{7}$$

and δ is essentially independent of the block length N for large values of N.

Arbitrarily set $m/n = \delta$. Then the overall rate is

$$R' = rR(1 - m/n) = r(1 - \delta)\left(1 - 2\delta + \frac{3}{N}\right) > r(1 - \delta)(1 - 2\delta). \tag{8}$$

It will turn out that the performance of the hybrid scheme depends on the distribution of computation and on the error probability for the Fano sequential decoding algorithm. Previously known upper bounds on these statistics are summarized in Appendix A. The bounds are on averages over ensembles of tree codes. Following an argument of Shannon, one can show that most tree codes picked at random satisfy all the bounds at least to within a small constant factor.[1] For example, suppose the ensemble averages of error probability and mean computation per decoded bit are upper bounded respectively by X and Y. Then at least 9/10 of all possible tree codes have error probabilities less than $10X$, at least 9/10 have mean computations less than $10Y$, and therefore at least 8/10 satisfy both of these bounds.

The upper bounds on the error probability[20] and on the distribution of computation[23] for rates r exceeding R_{comp} are known to apply also to the ensemble of convolutional codes, for which the coder's complexity is proportional to n. This extension to convolutional codes has not been analytically established for the distribution of computation for rates below R_{comp};[21, 22] however, it seems a reasonable conjecture that the degradation in performance due to the implementation of a tree code by a convolutional code is small for all rates.

3.1 *Error Probability*

From a result of Yudkin, it is inferred in Appendix A that the probability $p_u(e)$ that a sequential decoder decodes a serial block incorrectly is bounded by a negative exponential function of m, the number of redundant coder input bits in each n-symbol.[20] With $m = n\delta$,

$$p_u(e) < p_u'(e) = nA_e \exp\left[-n\delta v E_u(r)\right] \tag{9}$$

where A_e is a constant and $E_u(r)$ is a function of the tree code rate r and of the transition probabilities of the DMC. The exponent $E_u(r)$ is positive for any rate less than the capacity of the DMC. It is sketched for a typical DMC in Fig. 6. The probability of error $p(e)$ for the hybrid decoder is the probability that $N\delta/2$ or more undetected

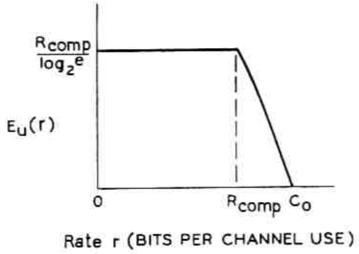


Fig. 6 — Sequential code error exponent $E_u(r)$ for a typical DMC.

serial block errors occur within a parallel block. Thus

$$p(e) = \sum_{\ell=N\delta/2}^{N} \begin{bmatrix} N \\ \ell \end{bmatrix} p_u(e)^{\ell}[1 - p_u(e)]^{N-\ell}. \tag{10}$$

The asymptotically tight Chernoff bound for the distribution of sums of binomially distributed random variables may be applied to the right-hand side of (10).[10]

$$p(e) \leqq \exp\left(-N\{T_e[\delta/2, p_u(e)] - H(\delta/2)\}\right) \qquad 0 \leqq p_u(e) < \delta/2 \tag{11}$$

where

$$T_e(x, y) = -x \,\ell n\, y - (1 - x)\,\ell n\,(1 - y)$$
$$H(x) = -x\,\ell n\, x - (1 - x)\,\ell n\,(1 - x).$$

It can readily be shown that for $y < x < 1/2$,

$$T_e(x, y) - H(x) > 0. \tag{12}$$

Thus the bound decreases exponentially with N. Notice that

$$\frac{\partial}{\partial p_u(e)} T_e[\delta/2, p_u(e)] < 0 \qquad p_u(e) < \delta/2. \tag{13}$$

Thus, the exponent in (11) is monotone decreasing in $p_u(e)$, provided that $p_u(e) < \delta/2$; therefore $p(e)$ can be further upper bounded by substituting $p'_u(e)$ for $p_u(e)$ in (11)

$$p(e) < \exp\left(-N\{T_e[\delta/2, p'_u(e)] - H(\delta/2)\}\right) \qquad p'_u(e) < \delta/2. \tag{14}$$

The exponent in (14) will be positive if $p'_u(e) < \delta/2 < \frac{1}{2}$. By virtue of (9), this will be true if

$$n > \frac{1}{\delta v E_u(r)}\,\ell n\,(2nA_e/\delta). \tag{15}$$

Thus $p(e)$ decreases exponentially with N if (15) is satisfied. But

$$\delta = \frac{1-R}{2} + \frac{3}{2N}; \tag{7}$$

$$\delta > 0 \quad \text{if} \quad R < 1$$

and

$$E_u(r) > 0 \qquad \text{if } r < \text{channel capacity.}$$

Thus, values of r, δ, and n can be found for which the constraint (15) is satisfied, while the overall rate, given by (8), is arbitrarily close to the channel capacity; that is, δ arbitrarily close to zero and r arbitrarily close to capacity.

The overall block length is $S_e = nN$. The serial block length n is constrained by (15) and by the constraint on the alphabet size of an RS code:

$$n \geqq \log_2 (N+1). \tag{4}$$

Thus for fixed overall rate R', and very large values of N, n behaves essentially as $\log_2 N$, or at most as $\log_2 S_e$. This implies that for a fixed rate less than the channel capacity, the probability of error is bounded by a quantity that asymptotically decreases almost exponentially (approximately as $S_e/\log_2 S_e$) with overall block length S_e. Notice also that the quantity S_e is proportional to the complexity of the hybrid coder, if the tree codes are convolutional codes. As mentioned earlier, it seems a reasonable assumption that the bounds on error probability and distribution of computation apply to convolutional codes of any rate.

The choice of $T = N\delta/2 - 1$ was arbitrary but convenient. For practical systems where N is less than, say 50, it would undoubtedly be more efficient to make m large enough that $p_u'(e)$ is negligible and to use the RS decoder to correct only erasures, that is, set $T = 0$ and $S = N(1 - R)$.

3.2 *Distribution of Computation*

A sequential decoding computation is done every time a tree branch is examined and compared to a received branch. Let c be the total number of computations to decode a given serial block, that are done by a sequential decoder operating alone, without aid or relief from an algebraic decoder. Appendix A uses the results of References 19, 21, 22 and 23 to show that the probability distribution function of c is bounded by a function which asymptotically is a pareto distribution. That is,

$$\text{pr}\ (c \geqq X) < [n^{\alpha'/\alpha} A_c/X]^{\alpha} \tag{16}$$

where $\alpha' = \max\ \{1, \alpha\}$, A_c is a constant, and α is the pareto exponent, a function of tree code rate r and of the channel statistics. The pareto exponent is positive for all rates less than channel capacity, and is greater than unity for all rates less than R_{comp}, which is less than channel capacity. The pareto exponent is sketched as a function of r for a typical DMC in Fig. 7. Note that the average of c is finite if and only if α is greater than one. It is clear that the smaller α is, the slower is the asymptotic decrease in pr ($c \geqq x$), and hence the greater is the variability of the random variable c. The bound on pr ($c \geqq X$) will be used to upper bound the distribution of the number of computations done by the hybrid decoder in decoding a super block.

For analytical convenience it will be assumed that sequential decoding of any serial block within a super block does not start until:

(*i*) The preceding super block has been decoded.

(*ii*) The entire serial block has been received and stored in the sequential decoder's buffer.

These conditions ensure that successive super blocks are decoded independently, and that during the decoding of any super block there is no idle time spent by the sequential decoders waiting for new branches to arrive. These assumptions can only delay the operation of the sequential decoders in our model, and hence lead to a conservative estimate of the buffer overflow probability.

Decoding of a super block is essentially completed when all but S of its N serial blocks have been sequentially decoded. The number of decoding operations then done by the RS decoder is bounded by a fixed quantity, and will be neglected. Accordingly we define C, the

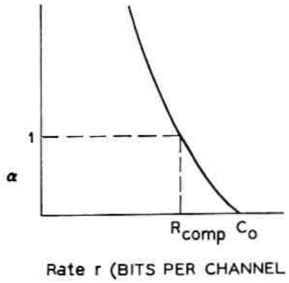


Fig. 7 — Pareto exponent α for a typical DMC.

number of *computation units* to decode the super block to be the $(S+1)$th largest of $\{c_1, c_2, \ldots, c_n\}$, where c_j is the number of computations that the jth sequential decoder, acting alone, would require to decode the jth serial block. Then, no more than C computations are done by any one sequential decoder during the decoding of the super block. One computation unit represents one or more (up to N) sequential decoding computations done simultaneously by the corresponding number of sequential decoders.

The number of computation units C exceeds X if $(S+1)$ or more of $\{c_1, c_2, \ldots, c_N\}$ exceed X. From (16),

$$\text{pr}\,(c_i \geqq X) \leqq p_x = [n^{\alpha'\alpha} A_e/X]^\alpha. \tag{17}$$

Then analogous to (14) we have, for $S + 1 = N\,\delta$,

$$\text{pr}\,(C \geqq X) \leqq \exp\,\{-N[T_e(\delta, p_x) - H(\delta)]\} \quad p_x < \delta. \tag{18}$$

A cruder but simpler bound is obtained by bounding $T_c(\delta, p_x)$ by $-\delta\,ln\,p_x$. Thus for $p_x < \delta$

$$\begin{aligned}\text{pr}\,(C > X) &\leqq \exp\,[NH(\delta)]p_x^{N\delta}\\ &= [A_1/X]^{N\delta\alpha}\end{aligned} \tag{19}$$

where

$$A_1 = \exp\,[H(\delta)/\alpha\delta]n^{\alpha'/\alpha}A_c\,.$$

From the definitions of A_1, and $H(\delta)$, and expression (17) for p_x, it is easy to show that the condition $p_x < \delta$ is certainly true if $X > A_1$. Also, pr $(C \geqq X)$, being a probability, is certainly bounded by unity. Thus

$$\text{pr}\,(C \geqq X) \leqq \begin{cases}[A_1/X]^{N\delta\alpha} & X > A_1\\ \quad 1 & X \leqq A_1\end{cases}. \tag{20}$$

Notice that the right-hand side of (20) asymptotically has the form of a pareto probability distribution, but that the effective pareto exponent is $N\delta$ times the pareto exponent for pure sequential decoding. Now,

$$\delta = \frac{1-R}{2} + \frac{3}{2N}\,; \tag{7}$$

$$\delta > 0 \quad \text{if} \quad 0 < R < 1$$

and

$$\alpha > 0 \qquad \text{if } r < \text{channel capacity.}$$

As the overall rate approaches channel capacity, α and δ both approach zero and the "break point" A_1 grows very large [A_1 also increases as the $1/\alpha$th power of $\log_2 (N + 1)$ for very large values of N]. However, for arbitrarily small but fixed values of α and δ, the RS code's block length N may be chosen sufficiently large that the effective pareto exponent $N\delta\alpha$ can be arbitrarily large and hence pr $(C \geqq X)$ arbitrarily small, for any X greater than A_1.

For a fixed value of N, the upper bound (20) is interesting only for $X \gg A_1$ or for values of α and δ large enough that $N\delta\alpha \gg 1$. For values of X for which (20) is not tight, the probability pr $(C \geqq X)$ is upper bounded by the probability that the largest of $\{c_1, c_2, \cdots, c_N\}$ exceeds X; that is, it is bounded by N pr $(C \geqq X)$ where pr $(C \geqq X)$ is bounded in (16).

3.3 *Average Computation*

Presumably, the average number of computation units done per super block is bounded if $N\delta\alpha > 1$, even if $0 < \alpha \leqq 1$. This is true, as will now be shown. The average of C is written

$$\begin{aligned}\langle C\rangle_{\text{av}} &= \sum_{X=1}^{\infty} X \text{ pr } (C = X)\\ &= \sum_{X=1}^{\infty} X[\text{pr } (C \geqq X) - \text{pr } (C \geqq X + 1)] \qquad (21)\\ &= \sum_{X=1}^{\infty} \text{pr } (C \geqq X).\end{aligned}$$

Then by (20)

$$\langle C\rangle_{\text{av}} \leqq A_1 + \sum_{X=A_1+1}^{\infty} (A_1/X)^{N\delta\alpha}.$$

The sum can be bounded by an integral from A_1 to infinity, since the integrand is positive and monotone decreasing.

$$\begin{aligned}\langle C\rangle_{\text{av}} &\leqq A_1 + \int_{A_1}^{\infty} (A_1/X)^{N\delta\alpha}\, dX\\ &= \frac{N\delta\alpha A_1}{N\delta\alpha - 1} < \infty \quad \text{if} \quad N\delta\alpha > 1 \qquad (22)\\ &= \frac{N\delta\alpha}{N\delta\alpha - 1}\, [\exp\, [H(\delta)/\alpha\delta] n^{\alpha'/\alpha} A_\epsilon].\end{aligned}$$

Thus the average number of computation units per super block is bounded if the effective pareto exponent $N\delta\alpha$ exceeds unity for any overall rate that is arbitrarily close to capacity, if N is chosen sufficiently large.

The bound on $\langle C\rangle_{\text{av}}$ varies with n as $n^{\alpha'/\alpha}$. Note that the number of computation units C is a bound on the number of computations done by each of the N sequential decoders, and that the number of information bits decoded by each sequential decoder per super block is no more than n. Thus the average number of sequential decoding computations per information bit is bounded by

$$\langle C\rangle_{\text{av}}/n \leqq A_2 n^{\alpha'/\alpha-1} \qquad N\delta\alpha > 1 \tag{23}$$

where

$$A_2 = \frac{N\delta\alpha}{N\delta\alpha - 1}[\exp[H(\delta)/\alpha\delta]A_c].$$

Since the block code is Reed–Solomon, n is constrained by $n \geqq \log_2(N+1)$. The overall block length (reflecting the complexity of the hybrid coder) is nN. Thus the minimum possible value of n behaves as the logarithm of the overall block length, and the average computation per bit increases as the $(\alpha'/\alpha - 1)$th power of the logarithm of overall block length. Furthermore, if $r < R_{\text{comp}}$ then $\alpha' = \alpha > 1$, and the average computation per bit is independent of the overall block length.

For rates above R_{comp}, the exponent α'/α increases rapidly with rate, approaching infinity at channel capacity. Thus the bound on the average computation, although finite, increases very rapidly with rate above R_{comp}. The average computation observed in the simulation reported in Reference 19 did indeed increase very rapidly with rate above R_{comp}.

3.4 *Probability of Buffer Overflow*

A new super block arrives to be decoded once every nv_T seconds. Each of the N sequential decoders is provided with a buffer which is assumed to store the latest Bv received output symbols from its respective DMC. Since we have assumed that all symbols comprising a super block must have been received before any decoding of the super block can start, the total storage must be large enough to contain one or more super blocks, that is, B must exceed n. Whole or partial super blocks stored but not yet decoded form a queue.

If the queue exceeds B/n super blocks (Bv channel output symbols per DMC) buffer overflow occurs. We wish to upper bound $P_L(B)$, the probability that the buffer overflows before the first L consecutive super blocks are decoded, given that the decoder starts with initially empty buffers.

Let q_i be the number of undecoded super blocks in the queue just after i consecutive super blocks have been decoded. Let X_i be the number of new super blocks which arrive to join the queue during the decoding of the ith super block. Because of our convention that decoding of any super block does not begin until the entire block has joined the queue, the number X_i does not include the ith super block itself or later super blocks. The random variables X_i and q_i are not necessarily integers, since a fraction $1/n$ of a super block arrives to be decoded every υ_T seconds.

When decoding of the first super block starts, the queue consists of only the first super block. Just after the first super block is decoded, the queue is thus diminished by one but has been increased by X_1. Thus

$$q_1 = 1 - 1 + X_1 = X_1 . \tag{24}$$

Just after the second super block is decoded,

$$q_2 = \begin{cases} q_1 - 1 + X_2 & \text{if} \quad q_1 \geqq 1 \\ \quad X_2 & \text{if} \quad q_1 < 1. \end{cases} \tag{25}$$

This is upper bounded by $q_1 + X_2$ for any $q_1 \geqq 0$. Therefore

$$q_2 \leqq X_1 + X_2 . \tag{26}$$

Similarly,

$$\begin{aligned} q_3 &= \begin{cases} q_2 - 1 + X_3 & \text{if} \quad q_2 \geqq 1 \\ \quad X_3 & \text{if} \quad q_2 < 1 \end{cases} \\ &\leqq X_1 + X_2 + X_3 \quad \text{for any} \quad q_2 \geqq 0. \end{aligned} \tag{27}$$

By induction then,

$$q_i \leqq \sum_{\ell=1}^{i} X_\ell . \tag{28}$$

This upper bound increases monotonically with i. It is clearly a crude approximation for large i. However it will turn out to yield a theoretically interesting upper bound on $p_L(B)$, at least for values of L which are small relative to B.

$$\begin{aligned} p_L(B) &= \text{pr}\,[(q_1 + 1 \geqq B/n) \quad \text{or} \quad (q_2 + 1 \geqq B/n) \quad \text{or} \quad \cdots \\ &\qquad (q_L + 1 \geqq B/n)] \\ &= \text{pr}\,[\max\,\{q_1\,,\,q_2\,,\,\cdots\,,\,q_L\} \geqq (B - n)/n] \\ &\leqq \text{pr}\left[\sum_{\ell=1}^{L} X_\ell \geqq (B - n)/n\right]. \end{aligned} \tag{29}$$

This inequality follows from (28) and the fact that all $X_\ell \geqq 0$.

Suppose each sequential decoder is capable of doing up to μ computations in each v_T-second interval, during which time a new branch arrives in each buffer. The parameter μ must be several times greater than the average number of computation units that the hybrid decoder does per information bit, if the decoder is to keep up with the incoming data. The hybrid decoder is "busy" (doing exactly μ computation units every v_T-second interval) until it is about to start decoding a super block which has not yet completely entered the buffer. From that instant it is idle until the entire super block has entered the buffer, at which time it becomes busy again. Thus, a busy interval can only be initiated just after the arrival of some super block, and can end only upon completion of the decoding of some subsequent super block. Suppose that during a particular busy interval, the νth through $(\nu + \eta)$th super blocks are decoded (ν, η integers; $L \geqq \nu \geqq 1$, $\eta \geqq 0$). Let C_ℓ be the number of computation units to decode the lth super block. Thus $\sum_{\ell=\nu}^{\nu+\eta} C_\ell$ is the total number of computation units done during the busy period. The first new super block to arrive during the busy interval arrives after η computation units have been done; thereafter, super blocks arrive every μn computation units. Thus $(1/\mu n) \sum_{\ell=\nu}^{\nu+\eta} C_\ell$ super blocks arrive during the entire busy interval. Successive busy intervals do not overlap, and therefore until the Lth super block is decoded,

$$\sum_{\ell=1}^{L} X_\ell \leqq (1/\mu n) \sum_{\ell=1}^{L} C_\ell\,. \tag{30}$$

Thus, from (29),

$$p_L(B) \leqq \text{pr}\left[\sum_{\ell=1}^{L} C_\ell \geqq \mu(B - n)\right]. \tag{31}$$

Since coding and decoding is independent from one super block to the next, the random variables $\{C_\ell\,,\,\ell = 1, 2, \cdots, L\}$ are statistically independent, and have a common cumulative probability distribution function which is bounded by the asymptotically-pareto distribution function (20).

The probability that overflow occurs before the first block is decoded is

$$p_1(B) \leqq \text{pr}\,[C_1 \geqq (B - n)] < \left[\frac{A_1}{\mu(B - n)}\right]^{N\delta\alpha}. \tag{32}$$

In appendix B an upper bound is obtained for the probability distribution of a sum of L statistically independent pareto-distributed random variables.* If the distribution of each random variable is upper bounded by pr $(C_i \geqq X) \leqq (A/X)^s$, $s > 1$ then it is shown that

$$\text{pr}\left[\sum_{i=1}^{L} C_i \geqq y\right] < DL(Ae/y)^s, \tag{33}$$

where $D = 1 + e^6$. This bound is valid for values of L which are small relative to y; specifically, for

$$(LA/y)\ell n(y^s/A^sL)\ell n(y/A) < e^{-1}. \tag{33a}$$

Applying inequality (33) to (31), we obtain the following bound for the probability of buffer overflow before L super blocks are decoded:

$$p_L(B) < DL\left[\frac{A_1e}{\mu(B - n)}\right]^{N\delta\alpha}, \qquad N\delta\alpha > 1 \tag{34}$$

provided that

$$\frac{LA_1}{\mu(B - n)}\,\ell n\left\{\frac{[\mu(B - n)]^{N\delta\alpha}}{A_1^{N\delta\alpha}L}\right\}\ell n\left\{\frac{\mu(B - n)}{A_1}\right\} < e^{-1}. \tag{34a}$$

Condition (34a) will be satisfied for values of L which are small relative to the product of decoder speed and available buffer size $\mu(B–n)$. Inequality (34) then indicates that $p_L(B)$ tends to increase linearly toward one with L and to decrease asymptotically as the negative $(N\delta\alpha)$th power of $\mu(B–n)$.

The techniques used to bound $p_L(B)$ were too crude to yield a useful result for small values of $\mu(B–n)$ or relatively large values of L; if condition (34a) is not satisfied, $p_L(B)$ can only be estimated by heuristic reasoning. The waiting line of undecoded super blocks can increase during the decoding of a given super block only if C, the number of computation units to decode the super block exceeds the number of computation units the decoder can do in $n\nu_T$ seconds, that is,

* Jelinek has given a more easily derived upper bound, which in its dependence on L, is at least as tight as our bound for $1 \leqq s \leqq 2$.[30]

if $C > n\mu$. The probability that the queue increases is bounded by (20) with $X = n\mu$.

$$\text{pr}\ (C > n\mu) < (A_1/n\mu)^{N\delta\alpha}.$$

If the decoder's speed factor μ is made large enough so that $n\mu > \langle C \rangle_{\text{av}}$ where $\langle C \rangle_{\text{av}}$ is bounded by (22), then the probability that the queue increases during the decoding of any super block approaches zero as N approaches infinity. Then the queue would be expected to remain close to zero most of the time, and consequently the probability that the buffer overflows during the decoding of a given super block would be approximated by $p_1(B)$, the probability that the first super block causes buffer overflow. For this reason, we use $p_1(B)$, bounded by (32) as a measure of buffer overflow probability.

It was shown in Section (3.3) that if $N\delta\alpha > 1$, the mean computation per super block is bounded by

$$\langle C \rangle_{\text{av}} \leqq \frac{N\,\delta\alpha}{N\,\delta\alpha - 1} A_1\ . \tag{22}$$

A hybrid decoder which can perform at least $\langle C \rangle_{\text{av}}$ computation units in a $n\upsilon_T$-second interval can, on the average, keep up with the incoming stream of super blocks arriving at $n\upsilon_T$-second intervals. A necessary condition for $n\mu > \langle C \rangle_{\text{av}}$ is

$$\mu = \mu_o \frac{A_1}{n} = \mu_o n^{\alpha'/\alpha - 1} A_c \exp\,[H(\delta)/\alpha\delta], \tag{35}$$

where μ_o is any number greater than $N\delta\alpha/(N\delta\alpha - 1)$.

Under condition (35) $p_1(B)$, given by (32), is bounded by

$$p_1(B) < \left[\frac{n}{\mu_o(B - n)}\right]^{N\delta\alpha}. \tag{36}$$

A fairly realistic measure of the cost of the hybrid scheme is the total amount of buffer storage utilized. If each of the N individual sequential decoders has a buffer capable of storing Bv channel output symbols, the total number of symbols which can be stored is

$$S_t = BNv. \tag{37}$$

Suppose we set

$$B = n(1 + e/\mu_0). \tag{38}$$

Then

$$\begin{aligned} p_1(B) &< \exp\,[-N\,\delta\alpha] \\ &= \exp\,[-S_t\ \delta\alpha/Bv] \\ &= \exp\left[\frac{-S_t\ \delta\alpha}{vn(1 + e/\mu_o)}\right]. \end{aligned} \tag{39}$$

For very large values of N (and therefore of S_t), the necessary value of serial block length n increases no faster than $\log_2 S_t$ to fulfill the constraints (4), (37) and (38). Consequently, the buffer overflow probability $p_1(B)$ is bounded by a quantity that asymptotically decreases almost exponentially with the total decoder storage S_t (that is, as $S_t/\log_2 S_t$). Furthermore, the exponent in (39) is positive provided that $\delta > 0$ and $\alpha > 0$. These conditions may be met for any overall rate R' which is less than channel capacity if the tree code rate r is less than channel capacity, and δ is small enough so that condition (8) is fulfilled. The derivation of this result suggested that best use would be made of a large but fixed amount of buffer storage if the number of parallel sequential coding-decoding systems is as large as possible, while the amount of storage allocated to each is a relatively small fixed multiple of the serial block length n.

IV. A NUMERICAL EXAMPLE

The upper bounds of the previous section are generally useful only if one is interested in asymptotic performance. Calculation of performance parameters for an implementable system should be based on the results of simulations. In this section we illustrate the estimation of performance parameters, based on a simulation of a sequential decoder.

Reference 25 describes the computer simulation of a Fano algorithm sequential decoder which decodes convolutionally coded binary antipodal signals received from a quantized phase-coherent white gaussian noise channel. For a convolutional code rate $r = 1/7$ bits per channel use, a signal-to-noise ratio of -6.5 dB, and an 8-level channel output quantization scheme, the pareto exponent α was very close to unity, that is, R_{comp} was close to 1/7. Other parameters are:

(*i*) serial block length $n = 360$ branches

(*ii*) number of redundant branches per serial block $m = 24$

(*iii*) convolutional code constraint length $= 24$ branches.

The net information rate of this system was then

$$\frac{1}{7}\,\frac{360 - 24}{360} = 0.133 \quad \text{bits per channel use.}$$

Assume the following RS code parameters

(*i*) Block length $N = 31 = 2^5 - 1$.
(*ii*) Alphabet size $= 32 = 2^5$, so that each super block is a sequence of 72 RS code words.
(*iii*) Rate $R = 26/31$, so that 5 serial blocks out of 31 are check symbols.
(*iv*) The RS decoder is designed to correct no errors and up to 5 erasures per RS code word.

The RS decoder would be easy to implement. A 155-bit register is required to store a RS code word consisting of 31 32-ary symbols. In addition, circuitry must be provided to solve 5 parity check equations to find the values of up to 5 erased 32-ary symbols. Forney has described efficient techniques for finding values of erasures.[16] The number of RS decoding operations is on the order of the square of the number of erasures which can be corrected.

Reference 25 shows empirical probability distribution functions for the total number of computations per serial block as observed in the simulation. For example, for the -6.5 dB channel, the probability that c, the number of computations per serial block exceeds 36,000 is approximately 10^{-2}. Thus the probability pr ($C \geqq 36{,}000$) that the number of computation units to decode a super block exceeds 36,000 equals the probability that 6 or more of the 31 serial blocks require more than 36,000 computations. This probability is obtained from tables (S. Weintraub, *Tables of the Cumulative Binomial Probability Distribution for Small Values of p*, London: Collier–Macmillan, 1963).

$$\text{pr}\,(C \geqq 36{,}000) = \sum_{i=6}^{36} \begin{bmatrix} 36 \\ i \end{bmatrix} p^i(1-p)^{36-i} = 6 \times 10^{-7} \quad (p = 10^{-2}).$$

Now assume that each sequential decoder is fast enough to do $\mu = 50$ computations between received branches. Then, up to $360\mu = 18{,}000$ computations can be done by each decoder in the time taken for one new serial block to enter the buffer of each; hence if each sequential decoder has a buffer with a storage capacity of three serial blocks, the buffer storage will overflow (starting from the initially empty state and assuming that decoding of a block starts after it is within the buffer) if the first super block requires more than $2 \times 18{,}000 = 36{,}000$ computation units. Then, assuming overflows are rare enough to be nearly statistically independent, the buffer overflow probability per super block would be about 6×10^{-7}. Each decoder's buffer stores

$3 \times 360 = 1080$ received branches, and the total number of branches stored is thus $1080\ N = 33{,}480$. The total number of bits (one per branch) per super block is $360 \times 31 = 11{,}160$.

Taking $n\delta = m = 24$, $v = 7$, $E_u(r) \log_2 e \approx R_{\text{comp}} \approx 1/7$, and assuming that $A_e \approx 1$ and that the upper bound (9) holds for convolutional codes, we have a rough upper bound for $p_e(e)$, the probability of undetected error per serial block.

$$p_u(e) \lesssim 360 \times 2^{-24} = 2.23 \times 10^{-5}.$$

(In the simulation, none of 1331 decoded blocks contained undetected errors.)

The probability of an undetected error for a super block is the probability that one or more of the 31 serial blocks has undetected errors; this probability is upper-bounded by $31 \times 2.23 \times 10^{-5} = 6.9 \times 10^{-4}$. This probability may be considered too high. It may be decreased about 3 orders of magnitude by increasing the value of m from 24 to 34. The resulting increase in the serial block length from 360 to 370 should cause negligible effect on the distribution of computation per serial block.

The net information rate of this system is $rR(n-m)/n = 0.109$ bits per channel use. It can be shown that the required signal-to-noise ratio per information bit is about 4.7 dB above Shannon's theoretical minimum for the infinite bandwidth white gaussian noise channel.

By such simple calculations based on extensive simulations, one can optimize the parameters of a hybrid scheme to meet given cost and performance criteria.

V. CONCLUSIONS

In the hybrid decoding scheme the number of decoding computation units per super block is a random variable, reflecting the probabilistic character of the sequential decoders' operations. However the pareto exponent is proportional to N; the frequency of large peaks of computational effort is reduced by algebraic decoding of the occasional serial blocks which otherwise would require excessive sequential decoding computation.

It was shown that for any overall information rate that is strictly less than the channel capacity, a finite minimum value of parallel block length N can be specified such that the average number of sequential decoding computations per bit is bounded by a quantity varying as

$n^{\alpha'/\alpha-1}$, where α is the original pareto exponent for the sequential decoding components and $\alpha' = \max\{\alpha, 1\}$. The number of algebraic decoding computations per bit is a fixed number which is almost independent of parallel or serial block length.[6]

It was also shown that for a proper choice of parameters, the error probability decreases nearly exponentially with the overall block length, and (heuristically) that the probability of buffer overflow asymptotically decreases almost exponentially with the total amount of storage at the decoder. These results can hold for any overall information rate which is strictly less than the channel capacity.

A rigorous upper bound was also obtained on $p_L(B)$, the probability that the buffer overflows before L super blocks are decoded. The bound is valid for $\mu(B - n) \gg L$, and behaves as $L[A_1e/\mu B]^{N\delta\alpha}$ for $B \gg n$ and fixed effective pareto exponent $N\delta\alpha$.

The hybrid scheme shares the multistage feature with the schemes of Ziv, Pinsker, and Forney.[13–16] In Ziv's scheme, there is an intermediate stage in which errors made by the inner block coding stage are detected and treated as erasures. After a scrambling-descrambling procedure these erasures are corrected by an outer block coding-decoding stage. Forney's scheme has two stages; a large alphabet RS code outer stage corrects errors and/or erasures made by an arbitrary inner block coding-decoding stage. Pinsker's scheme utilizes sequential coding-decoding for the outer stage. The principle is that if the inner stage has a sufficiently low error probability, the rate R_{comp} seen by the outer stage is little different from channel capacity. (This is, in a sense, the inverse of our hybrid scheme.)

In the hybrid scheme described in this paper, the inner and outer stages embody sequential (probabilistic) coding-decoding and algebraic coding-decoding respectively. Sequential coding and decoding is practical to implement and is efficient for any given DMC, which might be created from a physical communication channel by efficient modulation, demodulation, and quantization.[26, 27] The number of computation units per super block is a random variable, reflecting the probabilistic nature of sequential decoding and of short-term channel behavior. However, the variability of the sequential decoding computational load is eased substantially by the outer (algebraic) stage. Thus, in contrast with previous multistage schemes, the outer decoding stage assists the inner decoding stage, as well as correcting its errors.

Modifications and generalizations of the hybrid scheme are possible. A related scheme, in which channel symbols are not organized into independently coded blocks, was studied in Reference 19. Another modified hybrid scheme, falling into the general class of concatenated schemes considered by Forney, is implemented by imposing an upper limit X_o on the number of computations any sequential decoder can do on a serial block.[16] Assuming the speed factor μ is large enough that X_o computation units may be done in the time taken to receive one super block, no queue of undecoded super blocks can build up, and the buffer overflow problem is eliminated. Instead, any super blocks requiring more than X_o computation units are passed on to the user as erasures. The probability of erasure is then bounded by the right-hand side of (19) with $X = X_o > A_1$, that is, it decreases exponentially with parallel block length N.

The multistage approach embodied in the hybrid scheme would also appear to be useful for real channels with memory, where errors or severe channel disturbances occur in bursts, usually separated by fairly long intervals with only scattered random errors. If the N serial blocks comprising a super block are transmitted consecutively, a burst occurring during the transmission of one or more consecutive serial blocks would likely render them nearly undecodable by sequential decoding. Then if the burst did not extend over more than S serial blocks, an outer Reed-Solomon or other burst-correcting stage could correct the resulting erasures. The application of hybrid or other multistage coding schemes to real channels with memory is an interesting area for future investigation.

Any "hybrid" or "concatenated" coding-decoding scheme, incorporating a number of separate parallel coders and decoders would likely be orders of magnitude more complex than present day coding-decoding schemes for discrete memoryless channels. However the additional complexity may be a tolerable price to pay for the benefits of increased reliability and more efficient utilization of the communication channel. It is also well to remember that highly complex digital systems are becoming increasingly feasible as a result of rapid progress in integrated circuit technology.

VI. ACKNOWLEDGMENT

The author is grateful to his Massachusetts Institute of Technology thesis supervisor, Professor Irwin M. Jacobs, and his thesis readers, Professor Robert G. Gallager and Professor Robert S. Kennedy, for their guidance and encouragement during the thesis research.

APPENDIX A

Bounds on Performance for Sequential Decoding

Various upper bounds on the probability of error and the distribution of computation for the Fano sequential decoding algorithm have been given in References 20, 21, 22, and 23. All these bounds were obtained by random coding arguments, that is, by averaging over an ensemble of tree codes with a given probability distribution. The results apply to an arbitrary DMC with a P-symbol input alphabet and Q-symbol output alphabet, and a transition probability matrix $\{q_{ij}\}$. We shall summarize some of these previous bounds and then shall relate them to the performance of the hybrid scheme.

Using Gallager's notation, we define the function

$$E_o(\rho) = -\ell n \sum_{j=1}^{Q} \left[\sum_{i=1}^{P} p_i q_{ij}^{1/1+\rho} \right]^{1+\rho}; \qquad 0 \leqq \rho < \infty$$

where

$\{p_i\}$ $i = 1, 2, \ldots, P$ is the probability distribution on the channel input symbols which maximizes $E_o(\rho)$.[3]

It can be shown that $E_o(\rho)$ is a nondecreasing function of ρ, that

$$E_o(0) = 0$$

and that

$$\lim_{\rho \to 0} \frac{1}{\rho} E_o(\rho) = C_o$$

where C_o is the capacity of the DMC in bits per channel use.

Any transmitted serial block is a sequence of nv channel input symbols which label the corresponding correct path through the code tree. A path which diverges from the correct path is termed an *incorrect path.* A sequential decoder makes an undetected error at some node lying on the correct path, if the pattern of channel symbol transitions causes the decoder to reach the end of the serial block while on some incorrect path stemming from that node. One or more branches following the node will then have been decoded incorrectly. Of the n coder input bits which generate a serial block, m (the final m) are known to the sequential decoder. Hence a necessary condition for an undetected error to occur in decoding any serial block is that an incorrect path exists whose corresponding sequence of coder input digits matches that of the correct path in m or more places, and which the decoding algorithm can follow past those m places. The probability $p_h(e)$ that this necessary condition is fulfilled for say the hth node lying on the correct path has been upper-bounded by Yudkin, by random coding arguments.[20]

$$p_h(e) < A_e \exp\left[-mvE_u(r)\right] \tag{40}$$

where A_e is a constant and $E_u(r) > 0$ for $0 \leqq r < C_o$. The exponent $E_u(r)$ is sketched for a typical DMC in Fig. 6. It is considerably greater than the unexpurgated error exponent for block codes with the same rate.[3] In fact, $E_u(r) = E_o(1)$ for rates below $r = E_o\,(1) \log_2 e$ bits per channel use. This result for convolutional codes was also shown by Viterbi.[29] The probability of error $p_u(e)$ for a serial block is upper bounded by the probability that the necessary condition for undetected error occurs for one or more of the n nodes on the correct path. By the union bound,

$$p_u(e) \leqq \sum_{h=1}^{n} p_h(e) \leqq nA_e \exp\left[-mvE_u(r)\right]. \tag{41}$$

Inequality (9) follows from this result with $n\delta$ substituted for m.

Consider tree codes of rate r bits per channel input, where the tree extends infinitely to the right. The *incorrect subset* of the hth node lying on the correct path is defined to consist of that node plus the infinite set of nodes lying on incorrect paths which stem from the hth node. Let γ_h be the total number of computations (examinations of branches) ever done on nodes within this incorrect ubset. Then γ_h iss a random variable over the ensemble of tree codes and channel transition sequences. The sth ($s > 0$) moment of γ_h is bounded by a fixed quantity A_c^s for rates r such that

$$r < (E_0(s)/s) \log_2 e. \tag{42}$$

The quantity A_c^s is a function of s, r, $\{p_i\}$ and $\{q_{ij}\}$. This was established for integral values of s by Savage, for all $s \geqq 1$ by Yudkin, and for $0 < s \leqq 1$ by Falconer.[19,21-23] In particular, note that the mean of γ_h is only bounded for $r < E_0(1) \log_2 e$. The quantity $E_0(1)$ for a DMC is also denoted by R_{comp}, that rate below which the mean computation is finite. This bound on $\langle\gamma_h^s\rangle_{\text{av}}$ leads to an upper bound on the probability distribution pr $(\gamma_h \geqq x)$ by use of the Chebyshev inequality.[21]

$$\begin{aligned} \text{pr}\,(\gamma_h \geqq x) &\leqq \langle\gamma_h^s\rangle_{\text{av}} x^{-s} \qquad s > 0 \\ &\leqq A_c^s x^{-s} \qquad r < (E_o(s)/s) \log_2 e. \end{aligned} \tag{43}$$

The *pareto exponent* α is defined parametrically by

$$r = (E_o(\alpha)/\alpha) \log_2 e. \tag{44}$$

Then for any $\epsilon > 0$

$$\text{pr}\ (\gamma_h \geqq x) \leqq (A_c/x)^{(\alpha - \epsilon)}. \tag{45}$$

The right-hand side of (45) is proportional to a pareto probability distribution. The positive quantity ϵ may be made arbitrarily small by setting A_c large enough. Henceforth, we shall ignore ϵ as trivial since it would not affect our asymptotic results. Thus, we write

$$\langle \gamma_h^\alpha \rangle_{\text{av}} < A_c^\alpha \tag{46}$$

and

$$\text{pr}\ (\gamma_h \geqq x) < (A_c/x)^\alpha \tag{47}$$

where

$$E_0(\alpha)/\alpha = r/\log_2 e,$$

where the rate r is in bits per second. The pareto exponent for a typical DMC is shown in Fig. 7. The exponent on the right-hand side of (47) agrees asymptotically with that of a lower bound on the distribution of sequential decoding computation derived by Jacobs and Berlekamp.[24]

Let us now relate this upper bound on the distribution of computation for the Fano sequential decoding algorithm to the sequential decoding of serial blocks in the hybrid system. Only the portion of the code tree to a depth n branches from the origin is used to code and decode a serial block. Furthermore the last m information digits are known to the sequential decoder. Truncating a tree at a depth of n branches and making known the final m information letters can only reduce the number of branches a sequential decoder must examine before completing all computations in the first n incorrect subsets of an infinitely deep tree. Furthermore, it can be shown that for the Fano sequential decoding algorithm, allowing the decoder to search branches beyond depth n cannot reduce the number of computations ultimately done within a depth of n branches. Therefore, if c is the total number of computations to decode a serial block,

$$c \leqq \sum_{h=1}^{n} \gamma_h \tag{48}$$

and

$$\langle c^\alpha \rangle_{\text{av}} \leqq \left\langle \left[\sum_{h=1}^{n} \gamma_h \right]^\alpha \right\rangle_{\text{av}} \qquad (\alpha > 0). \tag{49}$$

The right-hand side of (49) may be bounded with well-known inequalities.[28]

$$\left\langle \left[\sum_{h=1}^{n} \gamma_h \right]^\alpha \right\rangle_{\text{av}} < \begin{cases} \displaystyle\sum_{h=1}^{n} \langle \gamma_h^\alpha \rangle_{\text{av}} & 0 < \alpha \leqq 1. \quad (50) \\ \left\{ \displaystyle\sum_{h=1}^{n} [\langle \gamma_h^\alpha \rangle_{\text{av}}]^{1/\alpha} \right\}^\alpha & \alpha \geqq 1. \quad (51) \end{cases}$$

Since

$$\langle \gamma_h^\alpha \rangle_{\text{av}} < A_c^\alpha ; \qquad 0 < \alpha < \infty, \quad \text{for all} \quad h,$$

we have

$$\langle c^\alpha \rangle_{\text{av}} \leqq \begin{cases} nA_c^\alpha & 0 < \alpha \leqq 1 \quad (52) \\ n^\alpha A_c^\alpha & \alpha \geqq 1 \quad (53) \end{cases}$$

or

$$\langle c^\alpha \rangle_{\text{av}} < n^{\alpha'} A_c^\alpha \tag{54}$$

where

$$\alpha' = \max (1, \alpha).$$

Then pr $(c \geqq x)$ is bounded using Chebyshev's inequality

$$\text{pr } (c \geqq x) \leqq \langle c^\alpha \rangle_{\text{av}} x^{-\alpha} \leqq n^{\alpha'} (A_c/x)^\alpha \tag{55}$$

where α is given parametrically by $r = [E_0(\alpha)/\alpha] \log_2 e$.

APPENDIX B

Probability Distribution of a Sum of Independent Pareto-Distributed Random Variables

It is required to upper-bound

$$\text{pr} \left[\sum_{i=1}^{L} C_i \geqq y \right],$$

where the $\{C_i\}$ are a set of independent positive integer-valued random variables whose distribution is asymptotically bounded by a pareto distribution function

$$\operatorname{pr}[C_i \geqq x] \leqq \begin{cases} (A/x)^s, & x \geqq A \\ 1, & 0 < x < A \end{cases} \tag{56}$$

where s is greater than one. The following assumption will be found necessary

$$(A/y) \ln (y^s/A^sL) \ln (y/A) < \frac{1}{eL}. \tag{57}$$

This assumption is tantamount to requiring that y be large relative to L.

We shall split the required probability into two parts, one of which is bounded by a union argument, and the other by use of a Chernoff technique (Reference 12, p. 97). That is, we write

$$\operatorname{pr}\left[\sum_{i=1}^{L} C_i \geqq y\right] = p_1 + p_2 \tag{58}$$

where

$$p_1 = \operatorname{pr}\left[\sum_{i=1}^{L} C_i \geqq y; \quad \text{one or more of} \quad \{C_i\} \geqq y\right]$$

$$p_2 = \operatorname{pr}\left[\sum_{i=1}^{L} C_i \geqq y; \quad \text{all} \quad \{C_i\} < y\right].$$

But

$$\begin{aligned} p_1 &< \operatorname{pr}[\text{one or more of } \{C_i\} \geqq y] \\ &\leqq \sum_{i=1}^{L} \operatorname{pr}(C_i \geqq y) \end{aligned} \tag{59}$$

by the union probability bound. So, substituting (56) into (59), we get

$$p_1 < L(A/y)^s \quad y > A > 1 \quad s > 1. \tag{60}$$

The probability p_2 may be bounded using the Chernoff technique, since each random variable C_i, being upper-bounded by y, has a finite moment generating function. To bound p_2 we first define

$$f_{z_i} = \operatorname{pr}(C_i = z_i) \qquad z_i = 1, 2, \cdots, y; \quad i = 1, 2, \cdots, L \tag{61}$$

and

$$\Phi(x) = \begin{cases} 1 & x \geqq 0. \\ 0 & x < 0 \end{cases} \tag{62}$$

Then by definition,

$$p_2 = \sum_{z_1=1}^{y-1} f_{z_1} \sum_{z_2=1}^{y-1} f_{z_2} \cdots \sum_{z_L=1}^{y-1} f_{z_L} \Phi\left[\sum_{i=1}^{L} z_i - y\right]. \tag{63}$$

We upper-bound the step function $\Phi(x)$ by the exponential function $\exp(\lambda x)$, where λ is an arbitrary positive quantity. We shall later choose a convenient value for λ. The right-hand side of (63) can now be bounded by a product of sums.

$$\begin{aligned} p_2 &\leqq \sum_{z_1=1}^{y-1} f_{z_1} \sum_{z_2=1}^{y-1} f_{z_2} \cdots \sum_{z_L=1}^{y-1} f_{z_L} \exp\left[\lambda\left(\sum_{i=1}^{L} z_i - y\right)\right] \\ &= \exp(-\lambda y) \prod_{i=1}^{L} \left[\sum_{z_i=1}^{y-1} f_{z_i} \exp(\lambda z_i)\right] \qquad (64) \\ &= \exp(-\lambda y) \left[\sum_{z=1}^{y-1} f_z \exp(\lambda z)\right]^L, \qquad (65) \end{aligned}$$

since the random variables $\{z_i\}$ are identically-distributed.

Now let

$$\psi = \sum_{z=1}^{y-1} f_z \exp(\lambda z). \tag{66}$$

This may be expressed in terms of the distribution function pr $(C \geqq z)$.

$$f_z = \text{pr}\ (C = z) = \text{pr}\ (C \geqq z) - \text{pr}\ (C \geqq z + 1) \tag{67}$$

So,

$$\begin{aligned} \psi &= \sum_{z=1}^{y-1} \exp(\lambda z)[\text{pr}\ (C \geqq z) - \text{pr}\ (C \geqq z + 1)] \\ &= 1 + \sum_{z=1}^{y-1} [\exp(\lambda z) - \exp(\lambda(z - 1))]\ \text{pr}\ (C \geqq z) \\ &\quad - \exp(\lambda(y - 1))\ \text{pr}\ (C \geqq y), \qquad (68) \end{aligned}$$

since

$$\text{pr}\ (C \geqq 1) = 1.$$

Taking out the common factor $[1 - \exp(-\lambda)]$ and upper-bounding it by λ, we get

$$\psi < 1 + [1 - \exp(-\lambda)] \sum_{z=1}^{y-1} \exp(\lambda z) \operatorname{pr}(C \geqq z)$$

$$\leqq 1 + \lambda \sum_{z=1}^{y-1} \exp(\lambda z) \operatorname{pr}(C \geqq z). \tag{69}$$

The function ψ is further bounded by employing the upper bound (56) for pr $(C \geqq z)$.

$$\psi < 1 + \lambda \sum_{z=1}^{A} \exp(\lambda z) + \lambda \sum_{z=A+1}^{y-1} \exp(\lambda z)(A/z)^s. \tag{70}$$

We now express the exponential functions as convergent power series and interchange the order of summation to yield

$$\psi < 1 + \sum_{z=1}^{A} \lambda \exp(\lambda z) + \sum_{h=1}^{\infty} \frac{\lambda^h A^s h}{h!} \sum_{z=A+1}^{y-1} z^{h-1-s}. \tag{71}$$

The sum over z may be upper bounded by an integral, which can be evaluated and bounded by simple expressions

$$\sum_{z=A+1}^{y-1} z^{h-1-s} \leqq \int_A^y z^{h-1-s}\, dz \leqq \begin{cases} \dfrac{A^{h-s}}{s-h} & 1 \leqq h \leqq s-1 \\ A^{h-s} \ell n\,(y/A) & s-1 < h \leqq s \\ y & s < h \leqq s+1 \\ \dfrac{y^{h-s}}{h-s} & h > s+1 \end{cases}. \tag{72}$$

These bounds will be used to bound the right-hand side of (71). The first sum in (71) is bounded by the number of terms times the largest (last) term.

$$\sum_{z=1}^{A} \lambda \exp(\lambda z) < A \exp(\lambda A). \tag{73}$$

Therefore, defining h_o to be that integer for which $h_0 + 1 > s \geqq h_o$, we have

$$\psi < 1 + \lambda A \exp(\lambda A) + \sum_{h=1}^{h_o-1} \frac{\lambda^h A^h}{(h-1)!\,(s-h)} + \frac{\lambda^{h_o} A^{h_o}}{(h_o-1)!} \ell n\,(y/A)$$

$$+ \frac{\lambda^{h_o+1} A^s y}{h_o!} + (A/y)^s \sum_{h=h_o+2}^{\infty} \frac{(\lambda y)^h h}{h!\,(h-s)}. \tag{74}$$

In the final sum in (74), $h \geqq h_o + 2 > s + 1$, and hence the sum may be upper bounded by bounding $h/(h - s)$ by $h_o + 2$ and then extending the summation down to $h = 0$. Thus,

$$(A/y)^s \sum_{h=h_o+2}^{\infty} \frac{(\lambda y)^h h}{h!\,(h - s)} \leqq (h_o + 2)(A/y)^s \sum_{h=0}^{\infty} \frac{(\lambda y)^h}{h!}$$

$$= (h_o + 2)(A/y)^s \exp(\lambda y). \qquad (75)$$

Furthermore in the first sum in (74), $h \leqq h_0 - 1 \leqq s - 1$, and hence the sum may be bounded by bounding $1/(s - h)$ by 1 and then extending the summation to infinity. Thus

$$\sum_{h=1}^{h_o-1} \frac{(\lambda A)^h}{(h - 1)!\,(s - h)} < \lambda A \sum_{h=0}^{\infty} \frac{(\lambda A)^h}{h!} = \lambda A \exp(\lambda A). \qquad (76)$$

Since $s \geqq h_0$,

$$\psi < 1 + 2\lambda A \exp(\lambda A) + (s + 2)(A/y)^s \exp(\lambda y)$$

$$+ \frac{(\lambda A)^s}{(h_o - 1)!} \ell n\,(y/A) + \frac{(\lambda A)^s}{h_o!} \lambda y. \qquad (77)$$

We shall now choose a particular value for λ:

$$\lambda = \lambda_o = \frac{1}{y} \ell n \left(\frac{y^s}{LA^s}\right). \qquad (78)$$

We also assume that L is small enough relative to y so that

$$\lambda_o A[\ell n\,(y/A)] < \frac{1}{eL}. \qquad (79)$$

This assumption is equivalent to (57). This condition also ensures that $\lambda_o A < 1/eL < 1$. The terms of (77) may now be bounded separately to yield a convenient upper bound on ψ. Thus,

$$2\lambda_0 A \exp(\lambda_0 A) < 2/L. \qquad (80)$$

From (78),

$$(s + 2)(A/y)^s \exp(\lambda_o y) = (s + 2)/L. \qquad (81)$$

Finally, using (78) and (79) it is easy to show that

$$\frac{(\lambda_o A)^s}{(h_o - 1)!} \ell n\,(y/A) + \frac{(\lambda_o A)^s}{h_o!} \lambda_o y < 2/L. \qquad (82)$$

The function ψ is now upper bounded for $\lambda = \lambda_o$ by using (80), (81), and (82) in (77),

$$\psi < 1 + 4/L + (s + 2)L = 1 + (6 + s)/L. \qquad (83)$$

Then,

$$\psi^L = \left[\sum_{s=1}^{\nu-1} f_s \exp(\lambda z) \right]^L < [1 + (6 + s)/L]^L. \tag{84}$$

Now for any L, $a \geqq 0$,

$$\begin{aligned}[1 + a/L]^L &= 1 + a + \frac{L(L-1)}{2!}(a/L)^2 \\ &\quad + \frac{L(L-1)(L-2)}{3!}(a/L)^3 \cdots + (a/L)^L. \\ &< 1 + a + a^2/2! + \cdots = \exp(a). \end{aligned} \tag{85}$$

Hence,

$$\psi^L < \exp(6 + s). \tag{86}$$

Substituting (86) in (65), we obtain

$$p_2 < L(Ae/y)^s \exp(6). \tag{87}$$

Finally, after substitution of (87) in (58) and (60),

$$\text{pr}\left[\sum_{i=1}^{L} C_i \geqq y \right] < DL(Ae/y)^s, \tag{88}$$

where

$$D = 1 + e^6.$$

This bound is valid under the condition (57),

$$(A/y)\, \ell n\, (y^s/A^sL)\, \ell n\, (y/A) < 1/eL. \tag{57}$$

REFERENCES

1. Shannon, C., "A Mathematical Theory of Communication," B.S.T.J., *27*, Nos. 3 and 4 (July and October 1948), pp. 279–423, 623–656.
2. Fano, R. M., *Transmission of Information,* New York: M.I.T. Press and J. Wiley, 1961, Chapter 9.
3. Gallager, R. G., "A Simple Derivation of the Coding Theorem and Some Applications," IEEE Trans. Inform. Theory, *IT-11,* No. 1 (January 1965), pp. 3–18.
4. Shannon, C. E., Gallager, R. G., and Berlekamp, E. R., "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels I," Inform. and Control, *10,* No. 1 (January 1967), pp. 65–103.
5. Peterson, W. W., *Error-Correcting Codes,* New York: M.I.T. Press and J. Wiley, 1961.
6. Berlekamp, E. R., "Nonbinary BCH Decoding," University of North Carolina, Institute of Statistics Mimeo Series No. 502 (December 1966).
7. Elias, P., "Error-Free Coding," IRE Trans. Inform. Theory, *PGIT-4,* No. 4 (September 1954), pp. 29–37.

8. Massey, J. L., *Threshold Decoding,* Cambridge, Massachusetts: M.I.T. Press, 1963.
9. Gallager, R. G., *Low-Density Parity-Check Codes,* Cambridge, Massachusetts: M.I.T. Press, 1963.
10. Wozencraft, J. M. and Reiffen, B., *Sequential Decoding,* New York: M.I.T. Press and J. Wiley, 1961.
11. Fano, R. M., "A Heuristic Discussion of Probabilistic Decoding," IEEE Trans. Inform. Theory, *IT-9,* No. 2 (April 1963), pp. 64–74.
12. Wozencraft, J. M. and Jacobs, I. M., *Principles of Communication Engineering,* New York: J. Wiley, 1965, Chapter 6.
13. Ziv, J., "Asymptotic Performance and Complexity of a Coding Scheme for Memoryless Channels," IEEE Trans. Inform. Theory, *IT-13,* No. 3 (July 1967), pp. 356–359.
14. Ziv, J., "Further Results on the Asymptotic Complexity of an Iterative Coding Scheme," IEEE Trans. Inform. Theory, *IT-12,* No. 2 (April 1966), pp. 168–171.
15. Pinsker, M. S., "Complexity of the Decoding Process," Problems of Information Transmission, *1* (1965), pp. 113–116.
16. Forney, G. D., "Concatenated Codes," M.I.T. Research Laboratory of Electronics, Technical Report 440 (December 1965).
17. Reed, I. S. and Solomon, G., "Polynomial Codes Over Certain Finite Fields," J. Soc. Ind. Appl. Math., *8* (June 1960), pp. 300–304.
18. Gorenstein, D. and Zierler, N., "A Class of Cyclic Linear Error-Correcting Codes in p^m Symbols," J. Soc. Ind. Appl. Math., *9* (June 1961), pp. 207–214.
19. Falconer, D. D., "A Hybrid Sequential and Algebraic Decoding Scheme," Ph.D. dissertation, Dept. of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts (February 1967).
20. Yudkin, H. L., "Channel State Testing in Information Decoding," Sc.D. dissertation, Dept. of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts (September 1964).
21. Savage, J. E., "Sequential Decoding—The Computation Problem," B.S.T.J., *45,* No. 1 (January 1966), pp. 149–175.
22. Yudkin, H. L., unpublished correspondence, 1965.
23. Falconer, D. D., "An Upper Bound on the Distribution of Computation for Sequential Decoding with Rate Above R_{comp}," Massachusetts Institute of Technology Research Laboratory of Electronics, Quart. Progress Rep. 81 (March 1966), pp. 174–179.
24. Jacobs, I. M. and Berlekamp, E. R., "A Lower Bound to the Distribution of Computation for Sequential Decoding," IEEE Trans. Inform. Theory, IT-13, No. 2 (April 1967), pp. 167–174.
25. Falconer, D. D. and Niessen, C. W., "Simulation of Sequential Decoding for a Telemetry Channel" Massachusetts Institute of Technology Research Laboratory of Electronics, Quart. Progress Rep. 80 (January 1966), pp. 183–193.
26. Jordan, K. L., "The Performance of Sequential Decoding in Conjunction with Efficient Modulation," IEEE Trans. Comm. Technology, *COM-14,* No. 3 (June 1966), pp. 283–297.
27. Jacobs, I. M., "Sequential Decoding for Efficient Communication from Deep Space," IEEE Trans. Comm. Technology, *COM-15,* No. 4 (August 1967), pp. 492–501.
28. Hardy, G. H., Littlewood, J. E. and Polya, G., *Inequalities,* London: Cambridge University Press, 1959.
29. Viterbi, A. J., "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," IEEE Trans. Inform. Theory, *IT-13,* No. 2 (April 1967), pp. 260–269.
30. Jelinek, F., *Probabilistic Information Theory,* New York: McGraw-Hill, 1968, Chapter 10.

Convolutional Reed-Solomon Codes

By P. M. EBERT and S. Y. TONG

(Manuscript received July 28, 1968)

We derive a new family of convolutional character-error-correcting codes which are a convolutional form of the Reed-Solomon block codes and as such have nonbinary symbols. We also derive a bound on the error correcting capabilities of these codes in which the error-correcting capability per constraint length grows approximately with the square root of the constraint length.

When these codes are used on a binary channel they are effective for both random and burst error correction because a single character spans several channel digits.

These codes have greater error-correcting capabilities than the Robinson-Bernstein self-orthogonal codes but are harder to decode. The single-character-error-correcting codes, when interleaved, are shown to be more powerful than the equivalent Hagelbarger code and appear to be simpler to implement. They are also slightly better than the interleaved version of Berlekamp's code.

We discuss encoding and decoding algorithms and illustrate a simple decoding algorithm for some of the codes. These codes are closely related to the Bose–Chaudhuri–Hocquenghem block codes and share with them the decoding simplification for character erasures in place of errors. Any Bose–Chaudhuri–Hocquenghem decoding algorithm can be used to decode these codes.

I. INTRODUCTION

This paper is concerned with a family of character error correcting convolutional codes which are derived from Reed-Solomon block codes.[1] The derivation of the error correcting capabilities is easy because of the algebraic nature of the code; the convolutional nature of the code allows the use of a simple encoder even at high rates. Also, sequential decoding techniques might be applicable.

Throughout the paper we use elements of a finite field as symbols instead of just 1 and 0. The elements can be represented by k-tuples of ones and zeros if the field has 2^k elements; these k tuples are called symbols or characters. Thus with this code we are able to correct character errors which may be bursts of binary errors. All that one need know about finite fields is that each nonzero element has an inverse and that there exists at least one element which when taken to successive powers will generate the entire field with the exception of the zero element. This is called a primitive element.

The capability of the codes is given by the number of errors that can be corrected within a fixed number of characters (the constraint length). Suppose a code can correct three errors within a constraint length of 12. Then the code can correct any pattern of errors as long as no sequence of 12 characters has more than three errors in it. In the context of error correcting ability one can define a minimum distance of the code. For linear codes the minimum distance (d), is equal to the minimum weight code word segment, one constraint length long, which has a nonzero first character. With this definition the code can correct up to $(d\text{-}1)/2$ errors occurring within one constraint length.

The rate of a code, R, is the fraction of characters used as information symbols.

A convolutional code has its check symbols formed as a convolution or weighted sum of a fixed number of the past information symbols. The weighted sum is formed in the algebra of the finite field.

The codes described in Section II are capable of correcting t character errors within a constraint length of $(2t^2 - t + 1)/(1 - R)$ channel characters. The channel characters must be elements of a finite field of size at least $[R(2t - 1)\ (t - 1) + 1]/(1 - R)$. If one uses as channel symbols elements of a much larger field it is possible to construct a code which will correct t errors with a constraint length of only $2t/(1 - R)$. Those codes can be encoded by a standard convolutional encoder or by a number of accumulators which can also perform multiplication. Decoding can be accomplished by a modified Bose-Chaudhuri-Hocquenghem decoder.

II. CONSTRUCTION OF THE CODES

For a code of rate $R = (b - 1)/b$ one can use every bth symbol as a check symbol. To define the code completely one need only give the weights used in the convolution of past channel symbols. For convenience we put these weights in an N by $b = 1/(1-R)$ matrix, $\mathbf{B}$. $B_{i,j} =$

W_{i-1+bj} (The constraint length is Nb.) If x_i is a check symbol we write

$$x_i = -\sum_{k=1}^{bN-1} x_{i-k}W_k\,, \qquad i = b, 2b, \cdots$$

or

$$\sum_{k=0}^{bN-1} x_{i-k}W_k = 0, \qquad W_0 = 1. \tag{1}$$

Following Wyner and Ash we take the N by b matrix called **B** and form it into an infinite **A** matrix by shifting **B** to the right b places and down one place:[2]

$$\mathbf{A} = \begin{bmatrix} \begin{bmatrix} \\ B \\ \\ \end{bmatrix} & \begin{matrix} 00 \\ \begin{bmatrix} \\ B \\ \\ \end{bmatrix} \end{matrix} & \begin{matrix} 00 \\ 00 \\ \begin{bmatrix} \\ B \\ \\ \end{bmatrix} \end{matrix} & \begin{matrix} \\ \\ \cdot \\ \cdot \\ \cdot \end{matrix} \\ \begin{matrix} 00 \\ 00 \end{matrix} & 00 & & \\ & \cdots & & \end{bmatrix}$$

Then by (1) any code sequence written as a vector **x** will satisfy

$$\mathbf{Ax} = \mathbf{0}. \tag{2}$$

We define the code by defining the elements of the matrix **B**. **B** is used instead of the weights W_i because the notation is clearer. Denote the elements of **B** by B_{ij}. Then let $B_{ij} = \beta_j\gamma_i^{i-1}$ where the b γ_i's are α^0, $\alpha^1, \cdots, \alpha^{b-2}$, 0. The symbol α is a primitive element of the field, and 0^0 is taken as 1. The γ_i's are called locators. Let: $\beta_j = \alpha^{v^{(j)}\ (b-1)}$, where

$$v^{(j)} = \sum_{i=0}^{j-2} i = \frac{(j-1)(j-2)}{2}.$$

Thus for any b, N, and finite field a code is defined.

As an example we take the special case which was presented by Wyner, where $N = 2$ and the B_{ij} are given by[3]

$$B_{ij} = \begin{cases} \alpha^{(i-1)(j-1)} & i \neq b \\ 0^{(j-1)} & i = b \end{cases}$$

$$\mathbf{B} = \begin{bmatrix} 1 & 1 & \cdot & \cdots & 1 & 1 \\ 1 & \alpha & \alpha^2 & \cdots & \alpha^{b-2} & 0 \end{bmatrix}. \tag{3}$$

This code has $d = 3$ and thus can correct single errors, or double erasures in a constraint length of $2b$. It has the additional advantages that it can be easily implemented, it is optimal,* and it has no error propagation. We describe, in Sections III through V, the implementation and some properties of the single-character-correcting codes.

III. DECODING TECHNIQUE FOR SINGLE CHARACTER CORRECTING CODES

Single errors are particularly easy to correct because the first syndrome (difference between the calculated check character and the received check character) is equal to the error, and the second syndrome is equal to the error multiplied by the error location. Since we take

$$\begin{aligned} \gamma_i &= \alpha^{i-1}, & i &\neq b \\ \gamma_i &= 0, & i &= b \end{aligned} \tag{4}$$

the error location i can be found by dividing the second syndrome S_1 by α and comparing the result to the first syndrome, S_0. This is repeated until they agree. The division can be implemented by a shift register whose feedback path corresponds to the coefficients of $g(x)$ the generator polynominal of α, the primitive element.[4]

IV. HARDWARE IMPLEMENTATION FOR SINGLE CHARACTER CORRECTING CODE

The implementation is described through the example: symbols in $GF(2^k)$ $k = 2$, $b = 4$. The elements of $GF(4)$ are represented as binary 2-tuples. Since $x^2 + x + 1$ is a primitive polynomial in $GF(2)$, division by α can be instrumented by a two-state shift register with appropriate taps. The decoder takes form of Fig. 1. The entire received vector is shifted into the data buffer and then the syndromes S_0 and S_1 are computed and stored in two registers R_0 and R_1, respectively. S_0 identifies the error pattern. If S_0 is nonzero, it is assumed that a single character error of the pattern shown occurred. For each character shifted out of data buffer the R_1 register is shifted once with the feedback path connected (which corresponds to a division by α). The error pattern S_0 in R_0 matches that contained in register R_1 when the erroneous character just emerges out of the data buffer. This character need only have S_0 subtracted from it to complete correction. It is possible to do all of this with simple logic circuitry and a storage of $2bk$ bits.

*No other code, of the same constraint length, which corrects single errors, can have a higher rate (that is, fewer check symbols).

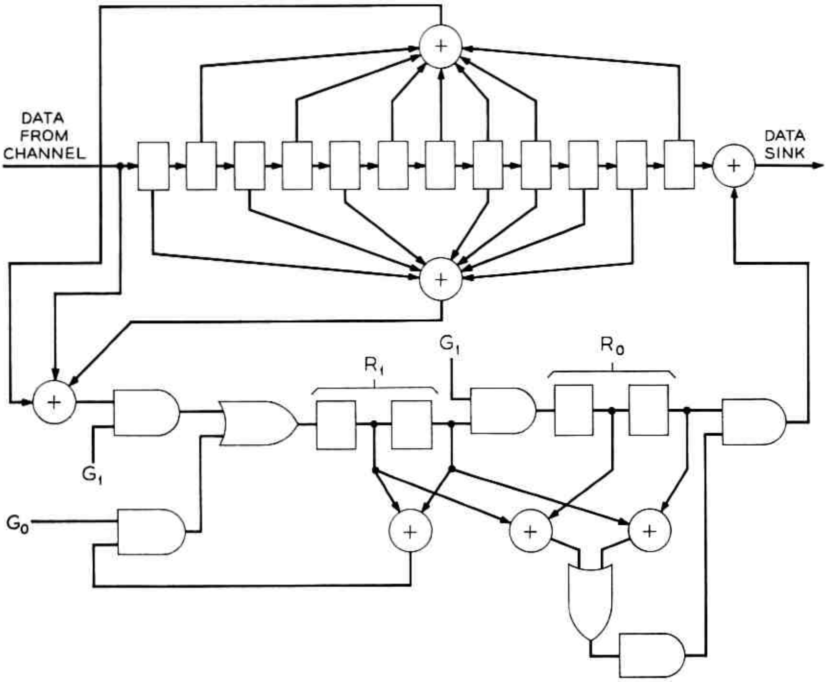


Fig. 1 — Decoder for single error correcting code. G_0 and G_1 are timing signals: G_0 is high during the time information characters appear; G_1 is high when parity character appears.

V. PROPERTIES OF THE SINGLE CHARACTER CORRECTING CODE

Observe that for every b character one must decide, based on the current syndromes of $2k$ bits, the error pattern and the location of the error. Since there are b possible error locations one needs at least $[\log_2 b] = k$ bits for identification if $2^{k-1} < b$, and as there are 2^k error patterns in a character (including the no-error pattern) that calls for k bits of information; thus the lower bound on the number of syndrome bits is $2k$ which shows the code is optimal* when $b > 2^{k-1}$.

Notice that since the syndrome is reset to zero after correction (corresponding to the removal of the effect of the error on syndrome) no error propagation is possible.

Interleaving can be applied to this class of codes to achieve a class of near-optimal burst-error correcting codes. If the interleaving degree is

* It is optimal in the number of check symbols.

m then two adjacent characters in the original code are separated by $m - 1$ characters or $L = k(m - 1)$ binary digits. A burst of length $L + 1$ bits will never affect two adjacent characters of the original codes; hence the interleaved code is capable of correcting an $L + 1$ bit burst. Since the original code has a storage requirement of $2bk$ bits, the interleaved code requires, at most, $2kbm$ bits. Observe that to correct a particular character one need not store the last character of the other $m - 1$ interleaved codes; therefore the storage requirement is $2\ bkm - (m - 1)k = 2\ bk(1 + L/k) - L$.

The required guard space is simply one less than the constraint length.* This can be seen by observing that in correcting a burst the syndrome must be set to zero when the last character in the burst is corrected. It follows that the guard space must always be shorter than the constraint length. Table I compares some members of this class of codes with some Hagelbarger burst-error-correcting recurrent codes as well as the Berlekamp burst-error-correcting recurrent codes (assuming Massey's decoding algorithm).[5–7] Table I shows this class of codes requires less guard space and hence is more powerful.

Table I—Comparison of Burst Correcting Codes

Rate	Burst	Decoder cost*			Guard space		
		A	B	C	A	B	C
$R = 1/2$	20	44	—	61	61	—	60
	19	—	60	58	—	61	57
$R = 2/3$	21	132	—	112	170	—	111
	22	—	120	—	—	122	—
$R = 3/4$	20	183	—	—	223	—	—
	21	—	168	156	—	171	155
$R = 4/5$	20	326	—	—	384	—	—
	21	—	225	—	—	229	—
	22	—	—	219	—	—	218

* More precisely, the number of shift registers.
Note: A – Hagelbarger codes.
B – Berlekamp's type B2 burst-error correcting codes modified by interleaving.
C – Single-character correcting codes modified by interleaving.

Although the number of shift register stages is not an accurate measure of decoder cost, it is seen that, except for the rate of ½, Hagelbarger's codes generally cost more, especially at higer rates. Although

* In our case the constraint length is equal to the storage requirement.

the interleaved single character correcting codes are slightly better than Berlekamp's code for correcting type B1 bursts, they are both the same (and optimal) for type B2 burst correction.[2]

VI. GENERALIZATION

In order to demonstrate the minimum distance of codes with a longer constraint length we rely on the following lemma.

Lemma: *If the code with* $N = N_1 \geqq 1$ *has a minimum distance of at least* d_1 , *then the code with* $N_2 = N_1 + d_1 - 1$ *has a minimum distance of at least* $d_1 + 1$.

The proof of this lemma depends on showing that no code word with d_1 or less nonzero elements can satisfy $\mathbf{Ax} = \mathbf{0}$, which is equivalent to showing that any d_1 columns of $\mathbf{A}$ form a matrix of rank at least d_1 and thus equation (2) can be satisfied only by $\mathbf{x} = \mathbf{0}$ among all $\mathbf{x}$ with weight d_1 or less.

Proof: By the assumption that the code with $N = N_1$ has minimum distance d_1 , there must be at least d_1 nonzero elements in the first N_1b symbols of $\mathbf{x}$. Assume that there are just d_1 and no more, as well as no additional nonzero symbols in the rest of the constraint length N_2 . We write $\mathbf{x}'$ as a d_1 dimensional vector consisting of only the nonzero elements of $\mathbf{x}$. Accordingly $\mathbf{A}'\mathbf{x}' = \mathbf{0}$, where $\mathbf{A}'$ is a N_2 by d_1 matrix with columns corresponding to the elements of $\mathbf{x}'$. We must choose d_1 rows of $\mathbf{A}'$ which have a nonzero determinant. Assume for the moment that none of error locations are zero. We then choose the bottom d_1 of $\mathbf{A}'$. We have a d_1 by d_1 array of nonzero elements. Each of the matrix element is guaranteed to be nonzero by the fact that all the nonzero elements of $\mathbf{x}$ lie within the first N_1b elements and consequently $\mathbf{A}'$ can only have no zeros in the d_1th or lower rows.

We now take the d_1 by d_1 array and divide each column by the first element. The first row is now all 1's and the kth row is

$$\begin{aligned}\frac{B_i(j + k - 1)}{B_{ij}} &= \frac{\beta_{j+k-1}\gamma_i^{j+k-2}}{\beta_j\gamma_i^{j-1}} \\ &= (\alpha^{j(b-1)}\gamma_i)^{k-1}\alpha^{[(k-1)(k-4)(b-1)/2]}. \qquad (5)\end{aligned}$$

We next divide each row by $\alpha^{[(k-1)(k-4)(b-1)/2]}$ thus obtaining a Vander-Monde matrix. The determinant of the matrix cannot be zero since the quantity

$$\alpha^{j(b-1)}\gamma_i \qquad (6)$$

is unique for each column. If q of the locators are zero, we choose the bottom $d_1 - q$ rows and those q rows which correspond to the 1's in the q zero locator columns. This determinant is nonzero by the same logic. Consequently, the equation $\mathbf{A}'\mathbf{x}' = \mathbf{0}$, can only be satisfied by $\mathbf{x}' = \mathbf{0}$, which contradicts our assumption. Therefore, the assumption is untrue and the minimum weight code word which satisfies (2) must have weight $d_1 + 1$ or larger, therefore providing the lemma.

We now prove the general result by induction. When $N = 1$ the matrix **1** obviously has rank 1 and thus the minimum distance is 2. To obtain a minimum distance of 3 we need to add $2 - 1 = 1$ to b, thus $N = 2$. Each time we increase the minimum distance by 1 we increase N by $d - 1$, thus

$$N = 1 + \sum_{i=1}^{d-2} i = \frac{(d-2)(d-1)}{2} + 1$$

$2N - 4 = d(d - 3)$ for any value of $d \geqq 2$. Since this is a lower bound on d we can be assured that for any $d \geqq 2$ we can build a code with $2N - 4 \leqq d(d - 3)$.

The field must be of sufficiently large size so that all the "locators" of (6) be different. This can be done if the field has at least $(b - 1)$ $[(d - 2)\ (d - 3) + 2]/2$ nonzero elements. Thus for example we could build a code which corrects two errors ($d = 5$), has a rate of ¾ ($b = 4$) with a 16 element alphabet. The constraint length would be $Nb = 7 \times 4 = 28$. This could be implemented by using 4-tuples as symbols with an overall constraint of 112 binary digits.

VII. ALTERNATIVE CODE

It is also possible to obtain a minimum distance of $N + 1$ if one uses only a certain set of symbols as locators. In the previous section one needed a field with at least $(b - 1)$ $[(d - 2)\ (d - 3) + 2]/2$ nonzero symbols. In this section we show that N can be made equal to $(d - 1)$ if one is willing to use symbols from a much larger field. For this purpose we define B by:

$$\mathbf{B} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \gamma_1^1 & \gamma_2^2 & \cdots & \gamma_b^b \\ \beta\gamma_1^2 & \beta\gamma_2^2 & \cdots & \beta\gamma_b^2 \\ \cdots & \cdots & \cdots & \cdots \\ \beta^{v(N)}\alpha_1^{N-1} & \cdots & \cdots & \beta^{v(N)}\alpha_b^{N-1} \end{bmatrix}$$

$$v(i) = \sum_{i=1}^{j-2} i = \frac{(j-2)(j-1)}{2}$$

where γ_i is the locator and β will be defined later. As in Section VI we wish to show that

$$\mathbf{Ax} = \mathbf{0} \tag{7}$$

only if $\mathbf{x}$ has d or more nonzero elements, given that there is at least one nonzero element among the first b. Suppose there exists a code word $\mathbf{x}$ with weight $d - 1$ or less such that (7) is met. Then take the N by $d - 1$ matrix consisting of the $d - 1$ columns of $\mathbf{A}$ corresponding to the nonzero components of $\mathbf{x}$, and call it $\mathbf{A}'$. The determinant of $\mathbf{A}'$ is a polynomial in β, where the lowest power of β is no lower than that found along the main diagonal. This comes about because the right-hand columns are shifted down and thus contribute least to the power of β in the bottom rows. The highest power of β is shown in the appendix to differ from the lowest power by no more than:

$$\left[\frac{N}{3}\right]\left[\frac{N+1}{3}\right]\left\{2\left[\frac{N+2}{3}\right] - 1\right\}^{*}.$$

If we call the difference between the highest and lowest power of β, r, then the determinant can only be zero when β is a root of an rth order or smaller polynomial with coefficients from the field containing the γ_i's. Consequently we can choose β from an $r + 1$ order extension field such that it is a root of an irreducible $r + 1$ order polynomial. Then it cannot be the root of a polynomial of degree r or less, and the determinant cannot be zero. Consequently $\mathbf{x} = \mathbf{0}$ for (7) to be met.

We have now defined α as a member of the field with q^{r+1} elements where $GF(q)$ is the locator field. In other words our symbols are taken from a q^{r+1} element field and the locators α_i are confined to the q element subfield. For example, suppose we require a double-error-correcting code ($N = 4$) with a rate ¾. We use the four element field as locators, and since $r + 1 = 4$ the symbols must come from a $4^4 = 256$ element field. This results in an overall constraint length of $8 \times 4 \times 4 = 128$ bits.

VIII. IMPLEMENTATION

The encoding can be accomplished by the standard convolutional encoder. The decoding presents the same problems as the decoding of

* [·] is the symbol for the integer part of ·.

a nonbinary Bose-Chaudhuri-Hocquenghem code. One takes increasing estimates of the number of errors within a constraint length and solves for the error locations and values. If the estimate is incorrect there is no consistent solution for the error values. The problem of error propagation can be easily handled by introducing a second set of syndrome calculators which are not changed when errors are corrected. A guard space equal to the constraint length, with no errors, will then produce zeros in all these syndromes, indicating that there have been no errors in the last constraint length.

In both coding and decoding, symbols must be added and multiplied. This is easy to implement when the symbols are represented as binary k tuples. Addition is just bit by bit modulo two addition; multiplication by α (the primitive element) is accomplished by a linear shift register.[4]

IX. CONCLUSIONS

We have presented a class of algebraic convolutional codes. These codes can be compared to other convolutional codes, such as Robinson and Bernstein's self-orthogonal codes.[8] The Robinson–Bernstein codes are binary error correcting codes which have a constraint length bounded by $N \geqq [(2t^2 - t)(b - 1) + 1]b$. Our codes have a constraint length bounded by $N \leqq (2t^2 - t + 1)b$, which is less than the above for $b \geqq 2$. Ignoring the fact that our code is a character error correcting code, our code is more powerful than the Robinson–Bernstein code.

If one takes account of the fact that characters are several bits long, one can still make a comparison by taking the constraint length of our code as

$$N \leqq (2t^2 - t + 1)b(\{ n_2[(b - 1)(2t^2 - 3t + 1) + b]\} + 1),$$

$[\cdot]$ integer of.

One now finds that our code is more powerful than the Robinson–Bernstein code for sufficiently large b. More specifically, if $t^2 \leqq 2^b/16b$ our code is more powerful. The Robinson-Bernstein codes have the advantage that they can be decoded by majority logic devices while we require much more complex devices.

The codes described in this section may be useful when one desires a high rate code. Here the amount of storage needed at the encoder can be made equal to the redundancy per constraint length. The decoder must store the entire constraint length in order to make corrections but the part of storage needed to compute the error locations and values is only equal to the redundancy. The disadvantage is that the decoder must perform finite field multiplication, division, and addition.

The code is also capable of correcting erasures; decoding is much simpler in this case. Therefore it may be advantageous to let the k tuples (which represent symbols) be code words in a binary subcode. The subcode then could be used to detect binary errors (producing an erasure), and the convolutional code used to correct these erasures. This is a convolutional version of Forney's concatenated block codes.[9]

The decoding algorithm is also simple for any code of distance four or less. This includes the single-error correcting code, the double-erasure correcting code, and the single-error-plus-single-erasure code. The decoding complexity is comparable to that of Hamming codes.

APPENDIX

Bound on the Powers of β

We wish to bound the difference between the maximum and minimum power of β in the determinant of $\mathbf{A}'$. The matrix will be made up of columns taken from $\mathbf{B}$ or columns from $\mathbf{B}$ shifted downward. This shifting downward of columns is the only parameter which effects our bound. Therefore we take Z_i as the amount that the ith column is shifted; the ith column is headed by Z_i zeros. By ordering we make Z_i a nondecreasing function of i. We can assume that $Z_i < i$; otherwise this would put a zero on the main diagonal and produce a zero determinant. One could then take the largest nonsingular upper left minor $\mathbf{A}''$ and write $\mathbf{A}''\mathbf{X}'' = \mathbf{0}$. The only difference here is that $\mathbf{A}''$ has smaller size; but the bound will still be valid.

The determinant of the matrix is given by

$$\sum_{\sigma} \prod_{i=1}^{N} b_{i,\sigma(i)} \tag{8}$$

where σ is a permutation of the numbers 1 to N. The term $b_{i,\sigma(i)}$ is zero if $\sigma(i) \leqq Z_i$; otherwise it contains β to the $\{[\sigma(i) - Z_i - 2] \cdot[\sigma(i) - Z_i - 1]\}/2$ power. For any given σ the product of (8) contains β to the

$$\tfrac{1}{2} \sum_{i=1}^{N} [\sigma(i) - Z_i - 2][\sigma(i) - Z_i - 1] \tag{9}$$

power, provided that for all i $\sigma(i) > Z_i$. Equation (9) can be rewritten

$$\tfrac{1}{2} \sum_{i=1}^{N} \sigma^2(i) - 3\sigma(i) - 2(i)Z_i + (Z_i + 2)(Z_i + 1). \qquad (10)$$

The only term which depends on σ is

$$-\sum_{i=1}^{N} \sigma(i)Z_i \, . \qquad (11)$$

The minimum power of β in (8) is given by (10) when we minimize (11). Since Z_i is nondecreasing the term which minimizes (11) is clearly $\sigma(i) = i$. Furthermore, the only other terms of (8) which contribute to this lowest power of β come from σ's which permute the i's over regions where Z_i is constant.

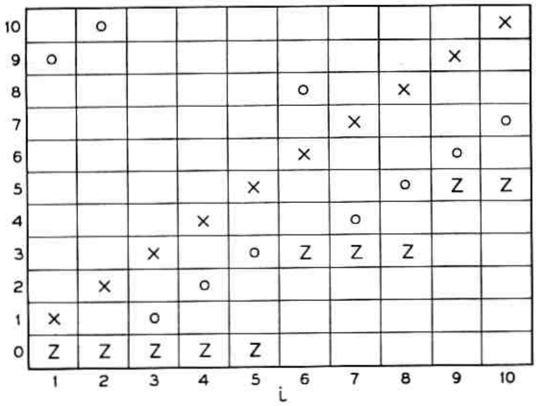
The difference between the maximum power of β and the minimum power is

$$\begin{aligned} D &= \sum_{i=1}^{N} iZ_i - \min_{\sigma,\sigma(i)>Z_i} \sum_{i=1}^{N} \sigma(i)Z_i \\ &= \max_{\sigma,\sigma(i)>Z_i} \sum_{i=1}^{N} [i - \sigma(i)]Z_i = \sum_{i=1}^{N} (i - \sigma_{\min}(i)]Z_i \, . \qquad (12) \end{aligned}$$

The minimization in (12) is accomplished by assigning the smallest i to the largest Z_i, subject to the constraint that $\sigma(i) > Z_i$. This can be done by starting with Z_N and assigning the smallest possible integer to $\sigma(N)$, namely $Z_N + 1$. The term Z_{N-1} then receives the next smallest integer which is still free, and so on. This is the minimum as any other acceptable σ can be converted into this $\sigma_{\min}$ by a sequence of pairwise permutations which do not increase the sum. In Fig. 2 we show an example of the assignment on a square array where the 0's indicate the σ which minimizes the sum and the x's that which maximize it $[\sigma(i) = i]$. The 0's in Fig. 2 are not exactly as described above but are permuted somewhat above equal value Z_i's for reasons which will be clarified in this appendix.

We now show that we can increase various Z_i's to Z_i''s so that $\Delta_i' = Z_{i+1}' - Z_i' \leqq 1$ (in Fig. 2 we increase Z_4, Z_5 and Z_8) and D will not decrease. When we change the Z_i, $\sigma_{\min}(i)$ does not change. When $\Delta_i \geqq 1$, our construction of $\sigma_{\min}(i)$ gives $\sigma_{\min}(i) = Z_{i+1} < i + 1$; thus $\sigma_{\min}(i) \leqq i$, and $i - \sigma_{\min}(i) \geqq 0$. Once we have the new Z_i' with $Z_{i+1}' - Z_i' \leqq 1$ we write $D' \geqq D$ in terms of a sequence

$$\Delta i' = Z_{i+1}' - Z_i' = \begin{cases} 0 \\ \blacksquare \\ 1 \end{cases}.$$



	1	2	3	4	5	6	7	8	9	10
10		o								×
9	o								×	
8						o		×		
7							×			o
6						×			o	
5					×			o	Z	Z
4				×			o			
3			×		o	Z	Z	Z		
2		×		o						
1	×		o							
0	Z	Z	Z	Z	Z					

Fig. 2 — Bounds on the minimum and maximum powers of β in the determinant of **A**.

Since

$$Z_i' = \sum_{j=1}^{i-1} \Delta_j' ,$$

then

$$\begin{aligned} D' &= \sum_{i=1}^{N} [i - \sigma(i)] Z_i' = \sum_{i=1}^{N} [i - (i)] \sum_{j=1}^{i-1} \Delta_i' \\ &= \sum_{j=1}^{N-1} \Delta_j' \sum_{i=j+1}^{N} [i - \sigma(i)] \\ &= \sum_{j=1}^{N-1} \Delta_j' \left[\frac{(N-j)(N+j+1)}{2} - \sum_{i=j+1}^{N} \sigma(i) \right]. \end{aligned} \tag{13}$$

We observe that our choice of $\sigma(i)$ gives

$$\sum_{i=j+1}^{N} \sigma(i) = \sum_{i=Z_j+2}^{N-j+Z_j+1} i = [N - j] \left[\frac{N-j+3}{2} + Z_j \right], \qquad \text{when} \quad \Delta_j' = 1,$$

and (13) becomes

$$\sum_{j=1}^{N-1} \Delta_j'(N-j)(j-1-Z_j).$$

For each value of $Z_j(0, 1, 2, \cdots)$ there can be only one nonzero Δ_j' ; if one takes K to be the number of nonzero Δ_j', then

$$D' = \sum_{k=0}^{K-1} (N - f_k)(f_k - 1 - k),$$

where f_k is that value of j for which $Z_j = k$ and $\Delta_j' = 1$. We now show that

$$\sum_{k=0}^{K-1} (N - f_k)(f_k - 1 - k) \leqq \sum_{i=f_0}^{N} (N - i)(f_0 - 1).$$

The terms on the right are all nonnegative and for every k on the left there is an i on the right with $i = f_k$. For that term $f_k \leqq f_0 + k$, as f_k cannot increase faster than k.

$$f_k - 1 - k \leqq f_0 - 1,$$

$$(N - f_k)(f_k - 1 - k) \leqq (N - f_k)(f_0 - 1).$$

$$D' \leqq \sum_{i=f_0}^{N} (N - i)(f_0 - 1) = (f_0 - 1)\,\frac{(N - f_0)(N - f_0 + 1)}{2}. \tag{14}$$

The integer value of f_0 which maximizes this is

$$f_0 = \left[\frac{N}{3}\right] + 1, \qquad [\cdot] = \text{integer part of}\cdot$$

and

$$D \leqq \left[\frac{N}{3}\right]\left[\frac{N+1}{3}\left[\left\{2\left[\frac{N+2}{3}\right] - 1\right\}\right.\right.. \tag{15}$$

This bound on D can actually be achieved for any N.

REFERENCES

1. Reed, I. S., and Solomon, G., "Polynomial Codes over Certain Finite Fields," J. Soc. Ind. Appl. Math., *8*, No. 2 (June 1960), pp. 300–304.
2. Wyner, A., and Ash, R., "Analysis of Recurrent Codes," IEEE Trans. on Inform. Theory, *9*, No. 3 (July 1963), pp. 143–160.
3. Wyner, A., "Some Results on Burst-Correcting Recurrent Codes," IEEE Conv. Rec., Part 4 (1963), 139–152.
4. Peterson, W., *Error Correcting Codes,* Cambridge: M.I.T. Press, 1961.
5. Hagelbarger, D. W., "Recurrent Codes; Easily Mechanized Burst-Correcting Codes," B.S.T.J., *38*, No. 4 (July 1959), pp. 969–984.
6. Berlekamp, E. R., "Note on Recurrent Codes," IEEE Trans. on Inform. Theory, *10*, No. 3 (July 1964), p. 257.
7. Massey, J. L., "Implementation of Burst-Correcting Convolutional Codes," IEEE Trans. on Inform. Theory, *11*, No. 3 (July 1965), pp. 416–422.
8. Robinson, J. P. and Bernstein, A. J., "A Class of Binary Recurrent Codes with Limited Error Propagation," IEEE Trans. on Inform. Theory, *13*, No. 1 (January 1967), pp. 106–113.
9. Forney, G. D., *Concatenated Codes,* Cambridge: M.I.T. Press, 1966.

Error Rate Considerations for Coherent Phase-Shift Keyed Systems with Co-Channel Interference

By V. K. PRABHU

(Manuscript received July 23, 1968)

In this paper we present a theoretical analysis of the performance of an m-phase coherent phase-shift keyed system in the presence of random gaussian noise and interference. An explicit expression is given for the probability of error of the phase angle of the received signal; we show that this probability of error can be expressed as a converging power series. We show that the coefficients of this series are expressible in terms of well-known and well-tabulated functions, and we give methods of evaluating the character error rates of the systems. We also show that this error rate is minimum when all the interference power is concentrated in a single interferer, and that it attains its maximum $[\mathrm{P_m}]_{\max}$ *when the total interference power is equally distributed amongst the K interferers. The limiting case when* K *goes to infinity is considered. The cases of* $\mathrm{K} = 1$, *and* $\mathrm{m} = 2, 4, 8,$ *and* 16 *are treated in some detail, and the results are given graphically. The usefulness of the results presented in this paper is that the designer can have at his disposal very simple expressions with which to evaluate the performance of any given Coherent Phase-Shift Keyed system when the received signal is corrupted by both interference and random gaussian noise.*

I. INTRODUCTION

The performance of coherent phase-shift keyed (CPSK) systems has been investigated by many authors;[1-5] in the transmission of information the CPSK system has been shown to be one of the most efficient techniques for trading bandwidth for signal-to-noise ratio. However, the type of noise considered by these authors is almost always limited to be random gaussian noise although most authors admit that interference other than normal noise must be considered in the design of any modulation scheme for digital transmission.

Consider the following situation. In the frequency bands above 10 GHz where the signal attenuation resulting from rain storms could be very severe, close spacings of the repeaters are almost always mandatory for reliable communication from point to point and for all periods of time.[6] In such cases the problem of interference may be much more important than the problem of noise in the optimum detection of the desired signal; hence it is very desirable to evaluate the performance of a CPSK system with co-channel and adjacent channel interference so that, for the selection of an optimum transmission scheme, comparative advantages of CPSK over other broadband modulation techniques (like FM) in combating interference can be determined.

We consider in this paper the performance of a CPSK system when the received signal is corrupted by both interference and random gaussian noise.* We first discuss binary (2-phase) and quaternary (4-phase) CPSK systems and show that exact expressions can be obtained for their probability of error P_m. These expressions are in the form of infinite power series which are shown to converge for all values of signal-to-noise ratio and for all signal-to-interference ratios above a certain level determined by the system. For $m = 2$ and 4, these error rates are calculated and the results are given in graphical form.

For $m = 3$ and for $m > 4$ we show that exact expressions for P_m are very complicated functions of signal-to-noise ratio, and signal-to-interference ratios; in this paper we only indicate how these expressions can be obtained. However, we do obtain expressions for upper and lower bounds to P_m and show that the difference between these two bounds is a monotonically decreasing function of signal-to-noise ratio, signal-to-interference ratios, and the number m of phases used in the system. For $m \geqq 4$, signal-to-noise ratio $\rho^2 \geqq 5$ dB,† and for signal-to-interference ratio $1/L^2 \geqq 20$ dB, we show that this difference is less than 5 percent, and that the upper bound can be used as a good approximation to P_m. For $m = 8$ and 16, we calculate these upper bounds and we present the results graphically.

For a given amount of interference power, we show that the character error rate is minimum when all the power is concentrated in a single interferer. If the total number of interferers is K we also show that the error rate P_m reaches its maximum $[P_m]_{\max}$ when the interference power is equally distributed among all the interferers. It

* The word "noise" indicates random gaussian noise corrupting the desired received signal.

† We use the notation $b = a$ dB if $10 \log_{10} b = a$.

follows that $[P_m]_{\max}$ is a monotonically increasing function of K and attains its maximum when K goes to infinity. We show that the case of K going to infinity can be treated in a simple manner.

For the computation of error rates P_m (or upper bounds to P_m, $m > 2$) it is necessary to calculate the central moments μ_{2n}'s of a certain random variable η defined in terms of the K interfering carriers. For large values of K the conventional method of evaluating μ_{2n}'s can be rather tedious; we give some simple methods of evaluating these moments.

In conclusion, this paper determines the performance of m-phase CPSK systems for the important case of signals corrupted by random gaussian noise and interference. The cases of $m = 2$, 4, 8, and 16 are treated in some detail.

II. PHASE ANGLE DISTRIBUTION IN CPSK SYSTEMS

Let us consider an m-phase CPSK system. We assume that there is a steady received signal* which is corrupted by random gaussian noise and interference. The gaussian noise is assumed to have zero mean and variance σ^2. The signals under consideration consist of phase-modulation pulses of specified width transmitted at a known repetition rate; we assume that there are K interferers, each interferer having the same form as the signal.

If we assume that each signal transmitted has a duration T, the received signal waveform in the absence of noise during the Nth interval can be represented as

$$s_N(t) = (2S)^{\frac{1}{2}} \cos (\omega_0 t + \theta), NT \leqq t \leqq (N + 1)T, \qquad (1)$$

where S is the received signal power, ω_0 is the angular frequency of the signal, and θ will have some value in the discrete set $2\pi k/m$, $0 \leqq k \leqq m - 1$, corresponding to the Nth message. All m messages are assumed to be equally likely. In the absence of noise and interference, the set of m possible received signals is described by a set of m equally-spaced vectors in the complex plane as shown in Fig. 1. The noise and interference corrupting the signal distort the signal both in amplitude and in phase; a zero-phase signal (corresponding to $k = 0$), as disturbed by noise and interference, is also shown in Fig. 1.

If we now assume that powei in the jth interferer is I_j, the jth inter-

* In this paper we do not consider the effects of fading on the error rates of CPSK systems. The effects of fading can usually be accounted for by a further integration of error rates obtained in this paper.[7]

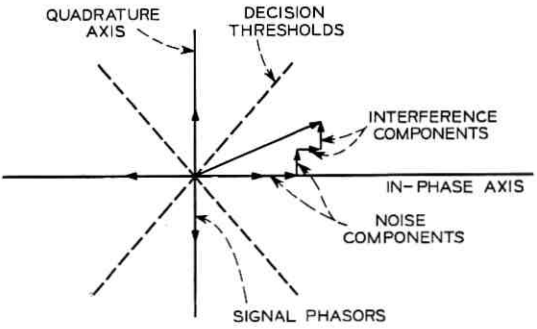


Fig. 1 — Phasor representation of CPSK signals for $m = 4$.

ferer as received during the Nth interval can be represented as*

$$i_{jN}(t) = (2I_j)^{\frac{1}{2}} \cos \{\omega_j t + \theta_j + \mu_j\}, NT \leqq t \leqq (N+1)T \qquad (2)$$

where ω_j is the angular frequency of the jth interferer, θ_j is some value in the discrete set $(2\pi/m)\, k$, $0 \leqq k \leqq m-1$, and the probability density $\pi_{\mu_j}(\mu_j)$ of μ_j is given by

$$\pi_{\mu_j}(\mu_j) = \begin{cases} \dfrac{1}{2\pi}, & 0 \leqq \mu_j < 2\pi \\ 0, & \text{otherwise.} \end{cases} \qquad (3)$$

Since the K interferers are assumed to originate from K different sources, it is reasonable to assume that all μ_j's are statistically independent of each other and are also independent of gaussian noise $n(t)$.

The total received signal during the Nth interval can then be written as

$$r_N(t) = (2S)^{\frac{1}{2}} \cos(\omega_0 t + \theta) + \sum_{j=1}^{K} (2I_j)^{\frac{1}{2}} \cos(\omega_j t + \theta_j + \mu_j) + n(t),$$

$$NT \leqq t \leqq (N+1)T \qquad (4)$$

where $n(t)$ has zero mean and variance σ^2.

Assuming that the receiver used in the system detects only the phase angle Φ of $r_N(t)$ and does not respond to its amplitude variations,† we can write[8]

* We assume that all i_{jN}'s, $1 \leqq j \leqq K$, are in the passband of the CPSK receiver used in the system.

† This can be achieved in practice by using an ideal limiter at the front end of the receiver. If $A(t)e^{j\varphi(t)}$ is the input to an ideal limiter, its output is given by $A_0 e^{j\varphi(t)}$ where A_0 is a constant.

$$\Phi = \tan^{-1} \frac{\hat{r}_N(t)}{r_N(t)} - \omega_0 t \tag{5}$$

where $\hat{r}_N(t)$ is the Hilbert transform of $r_N(t)$ and is given by

$$\hat{r}_N(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{r_N(\tau)}{t - \tau} \, d\tau. \tag{6}$$

Let us write

$$n(t) = I_c \cos (\omega_0 t + \theta) - I_s \sin (\omega_0 t + \theta). \tag{7}$$

We can show[9] that I_c and I_s are two independent gaussian random variables each distributed with mean zero and variance σ^2.* From (4)–(7), we can now show that

$$\Phi = \theta + \tan^{-1} \cdot \frac{I_s + \sum_{j=1}^{K} (2I_j)^{\frac{1}{2}} \sin [(\omega_j - \omega_0)t + \theta_j - \theta + \mu_j]}{(2S)^{\frac{1}{2}} + I_c + \sum_{j=1}^{K} (2I_j)^{\frac{1}{2}} \cos [(\omega_j - \omega_0)t + \theta_j - \theta + \mu_j]} . \tag{8}$$

Let us now write

$$\rho = \frac{S^{\frac{1}{2}}}{\sigma}, \tag{9}$$

$$\frac{I_s}{(2S)^{\frac{1}{2}}} = v, \tag{10}$$

$$\frac{I_c}{(2S)^{\frac{1}{2}}} = u, \tag{11}$$

$$\delta = \sum_{j=1}^{K} R_j \sin \lambda_j , \tag{12}$$

$$\eta = \sum_{j=1}^{K} R_j \cos \lambda_j , \tag{13}$$

where

$$R_j = \left(\frac{I_j}{S}\right)^{\frac{1}{2}}, \tag{14}$$

and

$$\lambda_j = (\omega_j - \omega_0)t + \theta_j - \theta + \mu_j . \tag{15}$$

* It is assumed that the spectrum of gaussian noise is symmetrical around the frequency $\omega = \omega_0$.

Let us also denote the set $\{\lambda_1, \lambda_2, \cdots, \lambda_i, \cdots, \lambda_K\}$ of random variables λ_i's by $\underset{\sim}{\lambda}$.

We can now write eq. (8) as

$$\Phi = \theta + \tan^{-1} \frac{v + \delta}{1 + u + \eta} \tag{16}$$

where δ and η are functions of $\underset{\sim}{\lambda}$.

If K is a finite number, we can show[10] that the probability density $p_\eta(\eta)$ can be represented as*

$$p_\eta(\eta) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\eta t} \prod_{j=1}^{K} J_0(tR_j)\, dt, \tag{17}$$

where $J_0(x)$ is the Bessel function of the first kind and of order zero. For $K = 1$, we can show that[10]

$$p_\eta(\eta) = \begin{cases} \dfrac{1}{\pi} \dfrac{1}{(R_1^2 - \eta^2)^{\frac{1}{2}}}, & |\eta| \leqq R_1 \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

For $K = 2$, $p_\eta(\eta)$ can be expressed in terms of elliptic functions, and for $K > 2$, no closed form expressions can be obtained for $p_\eta(\eta)$. In Ref. 10 $p_\eta(\eta)$ has been expressed as a converging sum and has been evaluated for $K = 10$. It is easy to show that

$$p_\eta(\eta) = 0, \quad \text{for} \quad |\eta| > \sum_{j=1}^{K} R_j, \tag{19}$$

and

$$\int_{-\infty}^{\infty} p_\eta(\eta) e^{-i\eta t}\, d\eta = \prod_{j=1}^{K} J_0(tR_j). \tag{20}$$

III. CPSK RECEIVER

An ideal CPSK receiver is shown in Fig. 2. The ideal limiter removes all the amplitude variations of the received signal before it reaches the ideal phase detector of the system. We shall assume maximum likelihood detection for our analysis of the receiver. Let us assume that the receiver shown in Fig. 2 has zero-width decision thresholds as shown in Fig. 1.

* We can also write similar expressions for $p_\delta(\delta)$.

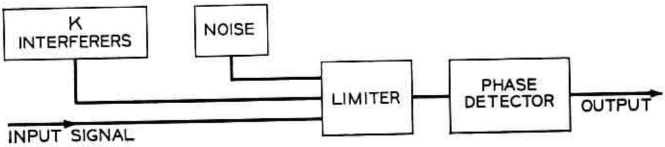


Fig. 2 — CPSK receiver.

3.1 *Error Rates for Binary CPSK Systems*

For a binary CPSK system the set of two possible received signals in the absence of noise and interference is shown in Fig. 3. The noise and interference corrupting the desired signal distort the signal both in amplitude and in phase; a zero-phase signal (corresponding to $k = 0$) as disturbed by noise and interference is also shown in Fig. 3.

When the message $k = 0$ is sent, and when the phase angle Φ of the received signal lies in the second and third quadrants of the complex plane shown in Fig. 3, an error is made in detecting the received signal. For a given ρ^2, and for an arbitrary set of λ_i's let us assume that the origin of the gaussian noise vector is at point G in Fig. 3. When the terminus or tip of the gaussian noise vector lies in the left half of the complex plane (the shaded portion of Fig. 3) an error is made by the receiver. Since I_c and I_s are two independent gaussian random variables and since they are distributed independently of λ_i's, the probability $P_2(\lambda)$ that the terminus of the gaussian noise vector lies in the

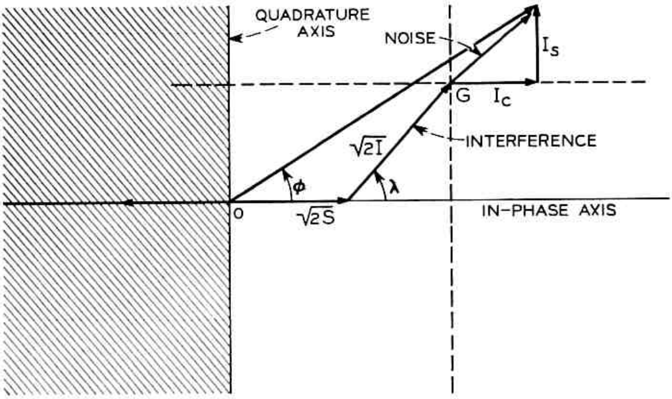


Fig. 3 — Phasor representation of CPSK signals for $m = 2$. I_c and I_s are the in-phase and quadrature components of gaussian noise corrupting the desired received signal.

left half of the complex plane is given by*

$$
\begin{aligned}
P_2(\underline{\lambda}) &= \Pr\,[-\infty < I_s < \infty] \\
&\quad \cdot \Pr\left[-\infty < I_c < -\left((2S)^{\frac{1}{2}} + \sum_{j=1}^{K} (2I_j)^{\frac{1}{2}} \cos \lambda_j\right)\right] \\
&= \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \int_{-\infty}^{-\left\{(2S)^{\frac{1}{2}} + \sum_{j=1}^{K} (2I_j)^{\frac{1}{2}} \cos \lambda_j\right\}} \exp\,(-t^2/2\sigma^2)\, dt. \qquad (21)
\end{aligned}
$$

We can show from Equation (21) that

$$P_2(\underline{\lambda}) = \tfrac{1}{2}\,\text{erfc}\,[\rho + \rho\eta], \qquad (22)$$

where

$$\text{erf}\,(x) = \frac{2}{\pi^{\frac{1}{2}}} \int_0^x \exp\,(-u^2)\, du \qquad (23)$$

and

$$\text{erfc}\,(x) = 1 - \text{erf}\,(x). \qquad (24)$$

The character error rate P_2 for a binary CPSK system is, therefore, given by

$$P_2 = E[P_2(\underline{\lambda})], \qquad (25)$$

where $E[P_2(\underline{\lambda})]$ represents the mathematical expectation of the random function $P_2(\underline{\lambda})$.

From Equations (22) and (25) we have

$$P_2 = \tfrac{1}{2} E[\text{erfc}\,\{\rho + \rho\eta\}]. \qquad (26)$$

We now note that we can write[11, 12]

$$\text{erfc}\,[x + z] = \text{erfc}\,[x] + \frac{2}{\pi^{\frac{1}{2}}} \exp\,(-x^2) \sum_{\ell=1}^{\infty} (-1)^{\ell} H_{\ell-1}(x) \frac{z^{\ell}}{\ell!}\,, \qquad (27)$$

where $H_n(x)$ represents the Hermite polynomial of order n. The series converges for all values of $x + z$ such that

$$x + z \geqq 0. \qquad (28)$$

From Equations (26) and (27) we have

$$P_2 = \tfrac{1}{2}\,\text{erfc}\,(\rho) + \frac{1}{\pi^{\frac{1}{2}}} \exp\,(-\rho^2) \sum_{\ell=1}^{\infty} (-1)^{\ell} H_{\ell-1}(\rho) \frac{\rho^{\ell}}{\ell!} E(\eta^{\ell}). \qquad (29)$$

* The notation $\Pr[a < x < b]$ denotes the probability that the random variable x satisfies the inequality $a < x < b$. It may also be noted that $P_2(\underline{\lambda})$ is a conditional probability conditioned on $\underline{\lambda}$.

Let us denote by μ_n the nth central moment of η.[8] It can then be shown that[10]

$$\mu_{2\ell+1} = 0, \qquad \ell = 0, 1, 2, \cdots . \tag{30}$$

We, therefore, have

$$P_2 = \tfrac{1}{2} \operatorname{erfc}(\rho) + \frac{1}{\pi^{\frac{1}{2}}} \exp(-\rho^2) \sum_{\ell=1}^{\infty} H_{2\ell-1}(\rho) \frac{\rho^{2\ell}}{(2\ell)!} \mu_{2\ell} . \tag{31}$$

The series given in Equation (31) converges for all values of ρ and R_j's such that

$$\rho + \rho\eta \geqq 0 \qquad \text{for all } \lambda. \tag{32}$$

From Equations (13) and (32) we can show that the series converges when

$$\Omega \leqq 1, \tag{33}$$

where

$$\Omega = \sum_{j=1}^{K} R_j . \tag{34}$$

Equation (34) states that the sum of the normalized amplitudes of all the interfering carriers may not exceed the normalized amplitude of the desired signal. This is not a very stringent requirement and it is almost always satisfied when low error rates are desired.

Since we also know that*

$$\left\{\frac{1}{K} \sum_{j=1}^{K} R_j\right\} \geqq \prod_{j=1}^{K} R_j^{1/K}, \tag{35}$$

when Equation (33) is satisfied, we have

$$\prod_{j=1}^{K} \left(\frac{I_j}{S}\right) \leqq \left(\frac{1}{K}\right)^{2K}. \tag{36}$$

The expression S/I_j denotes the signal-to-interference ratio of the jth interfering carrier.

When there is only one interfering carrier we can show that,

$$\mu_{2\ell} = R^{2\ell} \frac{(2\ell)!}{2^{2\ell}\{\ell!\}^2}, \tag{37}$$

* Equation (35) states that the arithmetic mean of a set of real variables is always greater than or equal to its geometric mean.

and equation (31) can be written as

$$P_2 = \tfrac{1}{2}\,\mathrm{erfc}\,(\rho) + \frac{1}{\pi^{\frac{1}{2}}}\exp(-\rho^2)\sum_{\ell=1}^{\infty} H_{2\ell-1}(\rho)\,\frac{\left[\dfrac{\rho R}{2}\right]^{2\ell}}{[\ell!]^2}. \tag{38}$$

The series in equation (38) converges for all signal-to-interference ratios such that

$$I/S \leqq 1. \tag{39}$$

The values of P_2 have been calculated from equation (38) and the results are given in graphical form in Fig. 4.*

Notice that we need to calculate only the even order moments μ_{2n}'s of the random variable η in determining P_2 from equation (31). Some methods of calculating these moments are given in Appendix A.

3.2 *Error Rates for Quaternary CPSK Systems*

Let us now consider a 4-phase CPSK system. For this system the set of four possible signal phasors and the four optimum decision thresholds are shown in Fig. 5. A signal phasor (corresponding to $k = 1$) as disturbed by noise and interference is also shown in Fig. 5.

For a given set of λ_j's let us assume that the gaussian noise is represented by a vector from the point G. If the message $k = 1$ is transmitted, an error is made if the received phase angle lies in areas marked 1, 2, and 3. The phase angle of the received signal will lie in areas marked 1, 2, and 3 if the terminus of the gaussian noise vector lies in this area of the plane.[14]

We notice that

$$GA = (2S)^{\frac{1}{2}}\sin\frac{\pi}{4} + \sum_{j=1}^{K}(2I_j)^{\frac{1}{2}}\sin\left(\frac{\pi}{4} + \lambda_j\right), \tag{40}$$

and

$$GB = (2S)^{\frac{1}{2}}\sin\frac{\pi}{4} + \sum_{j=1}^{K}(2I_j)^{\frac{1}{2}}\cos\left(\frac{\pi}{4} + \lambda_j\right). \tag{41}$$

Let us denote by $\Pi_{k1,k2}, \cdots, {}_{kn}(\lambda)$ the probability that the terminus of the gaussian noise vector lies in area

$$\bigcup_{i=1}^{n} k_i.\dagger$$

* The results obtained in Fig. 4 indicate that the error rates obtained in Refs. 13, 14, and 15 agree well with those obtained in this paper.

† The notation $\cup_{i=1}^{n} k_i$ denotes the union of all elements of the set $\{k_1, k_2, \cdots, k_n\}$.

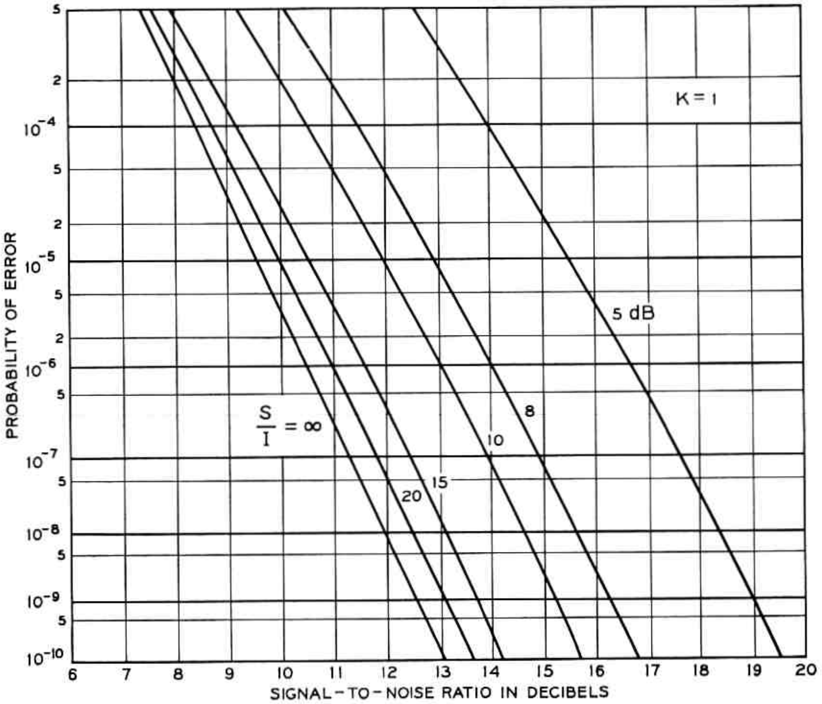


Fig. 4 — Error rates for a 2-phase CPSK system with one interferer.

We can show from Fig. 5 that

$$\Pi_{1,2}(\lambda) = \tfrac{1}{2} \operatorname{erfc}\left[\rho \sin\frac{\pi}{4} + \rho \sum_{j=1}^{K} R_j \cos\left(\frac{\pi}{4} + \lambda_j\right)\right], \tag{42}$$

$$\Pi_{2,3}(\lambda) = \tfrac{1}{2} \operatorname{erfc}\left[\rho \sin\frac{\pi}{4} + \rho \sum_{j=1}^{K} R_j \sin\left(\frac{\pi}{4} + \lambda_j\right)\right], \tag{43}$$

and

$$\begin{aligned}\Pi_2(\lambda) = \tfrac{1}{4} \operatorname{erfc}&\left[\rho \sin\frac{\pi}{4} + \rho \sum_{j=1}^{K} R_j \cos\left(\frac{\pi}{4} + \lambda_j\right)\right] \\ &\cdot \operatorname{erfc}\left[\rho \sin\frac{\pi}{4} + \rho \sum_{j=1}^{K} R_j \sin\left(\frac{\pi}{4} + \lambda_j\right)\right].\end{aligned} \tag{44}$$

The probability $P_4(\lambda)$ of an error due to noise is, therefore, given by

$$P_4(\lambda) = \Pi_{1,2}(\lambda) + \Pi_{2,3}(\lambda) - \Pi_2(\lambda). \tag{45}$$

The probability of an error due to noise and interference is therefore

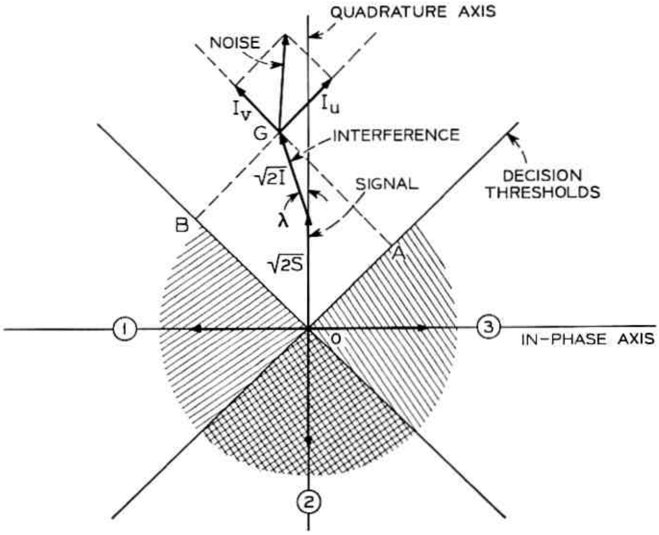


Fig. 5 — Phasor representation of CPSK signals for $m = 4$. I_u and I_v are two orthogonal components of gaussian noise.

given by

$$P_4 = E[P_4(\lambda)]. \tag{46}$$

From equations (27), and (42) through (46) we can show that

$$\begin{aligned} P_4 = {} & \operatorname{erfc}\left[\rho \sin \frac{\pi}{4}\right] - \tfrac{1}{4} \operatorname{erfc}^2 \left[\rho \sin \frac{\pi}{4}\right] \\ & + \frac{1}{(\pi)^{\frac{1}{2}}} \exp\left(-\rho^2 \sin^2 \frac{\pi}{4}\right)\left\{2 - \operatorname{erfc}\left[\rho \sin \frac{\pi}{4}\right]\right\} \\ & \cdot \sum_{\ell=1}^{\infty} H_{2\ell-1}\left(\rho \sin \frac{\pi}{4}\right) \frac{\rho^{2\ell}}{(2\ell)!} \mu_{2\ell} - \frac{1}{\pi} \exp\left(-2\rho^2 \sin^2 \frac{\pi}{4}\right) \\ & \cdot \sum_{\ell=1}^{\infty} \sum_{j=1}^{\infty} \frac{H_{2\ell-1}\left(\rho \sin \frac{\pi}{4}\right) H_{2j-1}\left(\rho \sin \frac{\pi}{4}\right)}{(2\ell)!\,(2j)!} \rho^{2(\ell+j)} \mu^*_{2\ell,2j} \end{aligned} \tag{47}$$

where $\mu^*_{2\ell,2j}$'s are given by

$$\begin{aligned} \mu^*_{2\ell,2j} = {} & \frac{1}{(2\pi)^K} \int_0^{2\pi} d\theta_1 \int_0^{2\pi} d\theta_2 \cdots \int_0^{2\pi} d\theta_K \left\{\sum_{j=1}^{K} R_j \cos \theta_j\right\}^{2\ell} \\ & \cdot \left\{\sum_{\ell=1}^{K} R_\ell \sin \theta_\ell\right)^{2j} \cdot \end{aligned} \tag{48}$$

For a given set of R_i's, $\mu^*_{2\ell,2j}$'s may be evaluated from equation (48). For $K = 1$, we can show that

$$\mu^*_{2\ell,2s} = R^{2(\ell+s)} \frac{(2\ell)!\,(2s)!}{2^{2(\ell+s)}\ell!\,s!\,(\ell+s)!}. \tag{49}$$

For $K = 1$, we have calculated P_4 from equation (47) and the results are presented in Fig. 6.

We can again show that the series given in equation (47) converges for all values of ρ and R_j's such that

$$\Omega \leqq \sin\frac{\pi}{4} = \frac{1}{\sqrt{2}}. \tag{50}$$

For $K = 1$, equation (50) becomes

$$S/I \geqq 2. \tag{51}$$

Equation (50) is usually satisfied by systems encountered in practice.

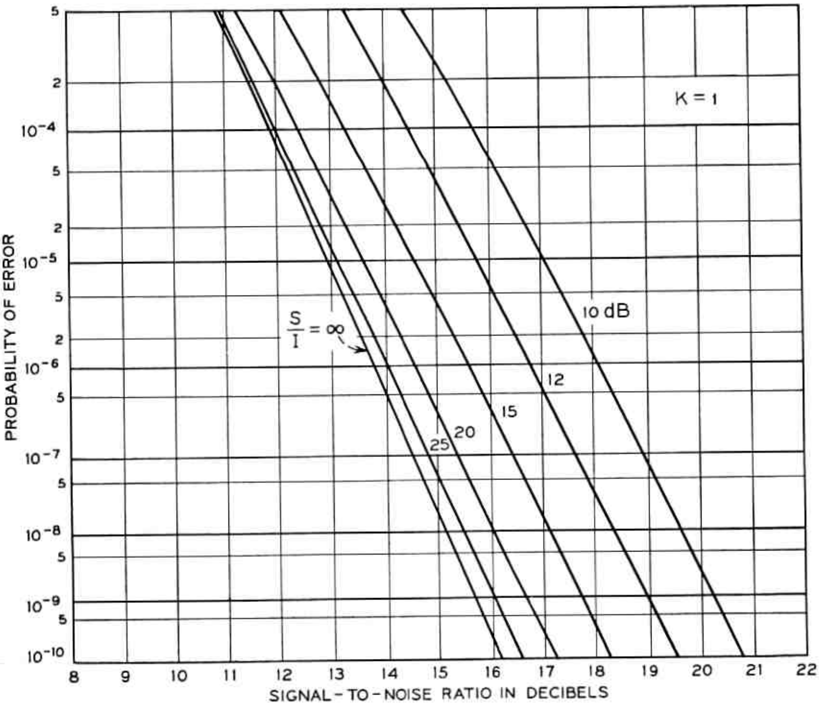


Fig. 6 — Error rates for a 4-phase CPSK system with one interferer.

3.3 *Error Rates for Multilevel CPSK Systems*

In this section we shall investigate the performance of a multilevel ($m \geqq 3$) CPSK Systems and indicate a method in which an exact expression can be obtained for the probability of error of the system. This exact expression is a very complicated function of signal-to-noise ratio and R_i's; we do not obtain this expression in this paper. However, we obtain upper and lower bounds to P_m and show that the difference between these two bounds is a monotonically decreasing function of ρ, m, and signal-to-interference ratios. For $K = 1$, $m \geqq 4$, $\rho^2 \geqq 5$ dB, and $S/I \geqq 20$ dB, we show that this difference is less than 5 percent of the lower bound, and hence the upper bound is a good approximation to P_m when low error rates are desired.

A signal phasor corresponding to $k = 0$ as disturbed by noise and interference is shown in Fig. 7. For a given set of λ_i's let us again assume that random gaussian noise is represented by a vector from the point G shown in Fig. 7. If the message $k = 0$ is transmitted, an error is made if the terminus of the noise vector lies in areas marked 1, 2, and 3.

We can show that *

$$\Pi_{1,2}(\lambda) = \tfrac{1}{2} \operatorname{erfc} \left[\rho \sin \frac{\pi}{m} + \rho \sum_{j=1}^{K} R_j \sin \left(\frac{\pi}{m} - \lambda_j \right) \right] \tag{52}$$

and

$$\Pi_{2,3}(\lambda) = \tfrac{1}{2} \operatorname{erfc} \left[\rho \sin \frac{\pi}{m} + \rho \sum_{j=1}^{K} R_j \sin \left(\frac{\pi}{m} + \lambda_j \right) \right]. \tag{53}$$

The probability of error due to noise is, therefore, given by

$$P_m(\lambda) = \Pi_{1,2}(\lambda) + \Pi_{2,3}(\lambda) - \Pi_2(\lambda). \tag{54}$$

By looking at Fig. 7 we can see that no simple expression can be obtained for $\Pi_2(\lambda)$ (except when $m = 4$). $\Pi_2(\lambda)$ denotes the probability that the terminus of the gaussian noise vector lies in area 2; we shall now obtain upper and lower bounds to $\Pi_2(\lambda)$. Assume that

$$\rho \sin \frac{\pi}{m} - \rho \sum_{j=1}^{K} R_j \geqq 0 \tag{55}$$

* Note that

$$GA = (2S)^{\frac{1}{2}} \sin \frac{\pi}{m} + \sum_{j=1}^{K} (2I_j)^{\frac{1}{2}} \sin \left(\frac{\pi}{m} - \lambda_j \right)$$

and

$$GB = (2S)^{\frac{1}{2}} \sin \frac{\pi}{m} + \sum_{j=1}^{K} (2I_j)^{\frac{1}{2}} \sin \left(\frac{\pi}{m} + \lambda_j \right).$$

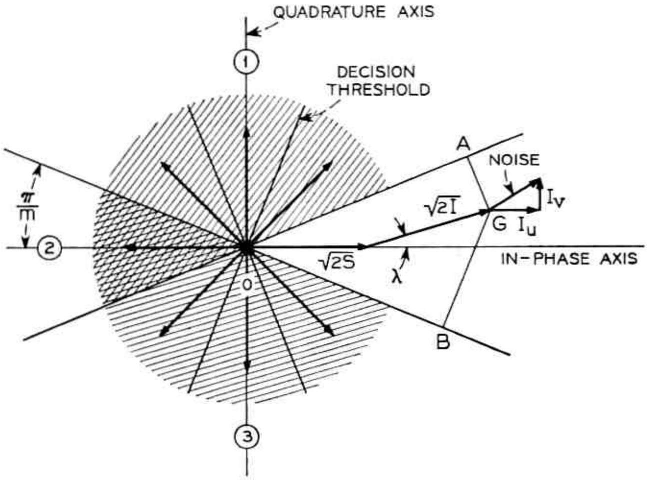


Fig. 7 — Phasor representation of CPSK signals for $m = 8$. I_u and I_v are the in-phase and quadrature components of gaussian noise.

so that $\Pi_{1,2}(\lambda)$ and $\Pi_{2,3}(\lambda)$ are nonnegative for all values of λ. If equation (55) is satisfied, it is easy to see (see Fig. 7) that

$$\Pi_2(\lambda) \geqq 0 \qquad \text{for all } \lambda, \tag{56}$$

and $\Pi_2(\lambda)$ reaches its maximum when*

$$\eta = -\sum_{j=1}^{K} R_j = -\Omega. \tag{57}$$

For this value of η it can be shown (see Fig. 8) that

$$\Pi_2(\lambda) = \frac{1}{\pi\sigma^2}\int_{-\infty}^{-y_0} \exp(-y^2/2\sigma^2)\,dy \int_0^{-(y+y_0)\tan\pi/m} \exp(-x^2/2\sigma^2)\,dx$$

or

$$\Pi_2(\lambda) = \frac{1}{\pi\sigma^2}\int_{y_0}^{\infty} \exp(-y^2/2\sigma^2)\,dy \int_0^{(y-y_0)\tan\pi/m} \exp(-x^2/2\sigma^2)\,dx \tag{58}$$

where

$$y_0 = (2S)^{\frac{1}{2}}[1 - \Omega]. \tag{59}$$

Since we always have

$$0 \leqq \exp(-x^2/2\sigma^2) \leqq 1 \text{ for all real } x, \tag{60}$$

* We can show that $\eta = -\Omega$ when all λ_j's are odd multiples of π.

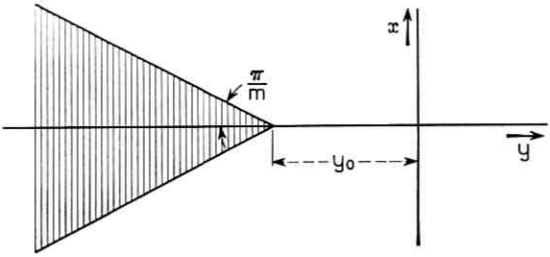


Fig. 8 — Computation of lower bound to P_m.

we have

$$\Pi_2(\lambda) \leqq \frac{\tan \pi/m}{\pi\sigma^2} \int_{y_0}^{\infty} (y - y_0) \exp(-y^2/2\sigma^2)\, dy. \tag{61}$$

Equation (61) can be simplified to*

$$\Pi_2(\lambda) \leqq Q_{m0} = \frac{\tan \pi/m}{\pi} \exp[-\rho^2(1 - \Omega)^2] \cdot [1 - (\pi)^{\frac{1}{2}}\rho(1 - \Omega) \exp[\rho^2(1 - \Omega)^2] \operatorname{erfc}\{\rho(1 - \Omega)\}]. \tag{62}$$

From equations (54), (56), and (62) we have

$$\Pi_{1,2}(\lambda) + \Pi_{1,3}(\lambda) - Q_{m0} \leqq P_m(\lambda) \leqq \Pi_{1,2}(\lambda) + \Pi_{1,3}(\lambda). \tag{63}$$

Since

$$P_m = E[P_m(\lambda)], \tag{64}$$

we can show from equations (52), (53), (55), and (63) that

$$Q_m - Q_{m0} \leqq P_m \leqq Q_m \tag{65}$$

where

$$Q_m = \operatorname{erfc}\left(\rho \sin \frac{\pi}{m}\right) + \frac{2}{(\pi)^{\frac{1}{2}}} \exp\left(-\rho^2 \sin^2 \frac{\pi}{m}\right) \sum_{\ell=1}^{\infty} \frac{H_{2\ell-1}\left(\rho \sin \frac{\pi}{m}\right)}{(2\ell)!} \rho^{2\ell} \mu_{2\ell}. \tag{66}$$

* For large values of ρ and small values of Ω, Q_{m0} is approximately equal to

$$\frac{\tan \pi/m}{2\pi} \frac{\exp[-\rho^2(1 - \Omega)^2]}{\rho^2(1 - \Omega)^2}.$$

The series given in equation (66) converges if equation (55) is satisfied, or if

$$\Omega \leqq \sin \frac{\pi}{m}. \tag{67}$$

When low error rates are desired, equation (67) must be satisfied.

Equation (65) gives an upper and a lower bound to P_m; as can be seen from equation (62) the difference Q_{m0} between these two bounds is a monotonically decreasing function of ρ, m, and signal-to-interference ratios. From equation (65) we have

$$-\frac{Q_{m0}}{Q_m - Q_{m0}} \leqq \frac{P_m - Q_m}{P_m} \leqq 0. \tag{68}$$

For $K = 1$, $R_1 = \frac{1}{10}$, and for $m = 4$, 8, and 16, we have plotted in Fig. 9 $Q_{m0}/(Q_m - Q_{m0})$ as a function of ρ^2. From Fig. 9 we see that $Q_{m0}/(Q_m - Q_{m0})$ is less than 5 percent for $\rho^2 \geqq 5$ dB and for $m \geqq 4$. We can, therefore, use Q_m as a good approximation to P_m for high values of signal-to-noise ratio ($\rho^2 \geqq 5$ dB) and for high values of signal-to-interference ratio ($1/R_1 \geqq 10$ dB).

In these cases we then have

$$P_m \approx \operatorname{erfc}\left(\rho \sin \frac{\pi}{m}\right) + \frac{2}{(\pi)^{\frac{1}{2}}} \exp\left(-\rho^2 \sin^2 \frac{\pi}{m}\right) \sum_{k=1}^{\infty} \frac{H_{2k-1}\left(\rho \sin \frac{\pi}{m}\right)}{(2k)!} \rho^{2k} \mu_{2k} . \tag{69}$$

For $K = 1$, and for $m = 8$ and 16, the values of P_m obtained from equation (69) are given in Figs. 10 and 11. The error made in this approximation can be estimated from equation (68).

IV. ERROR RATE AS A FUNCTION OF NUMBER OF INTERFERERS

Let us now investigate how P_m varies as a function of K for a total given interference power. Let us assume that the total interference power is some number SL^2 where

$$\sum_{j=1}^{K} I_j = SL^2. \tag{70}$$

This power SL^2 can be distributed among the K interferers in a variety of ways; every one of these distributions will in general lead to a different character error rate of the system. Let us find out those

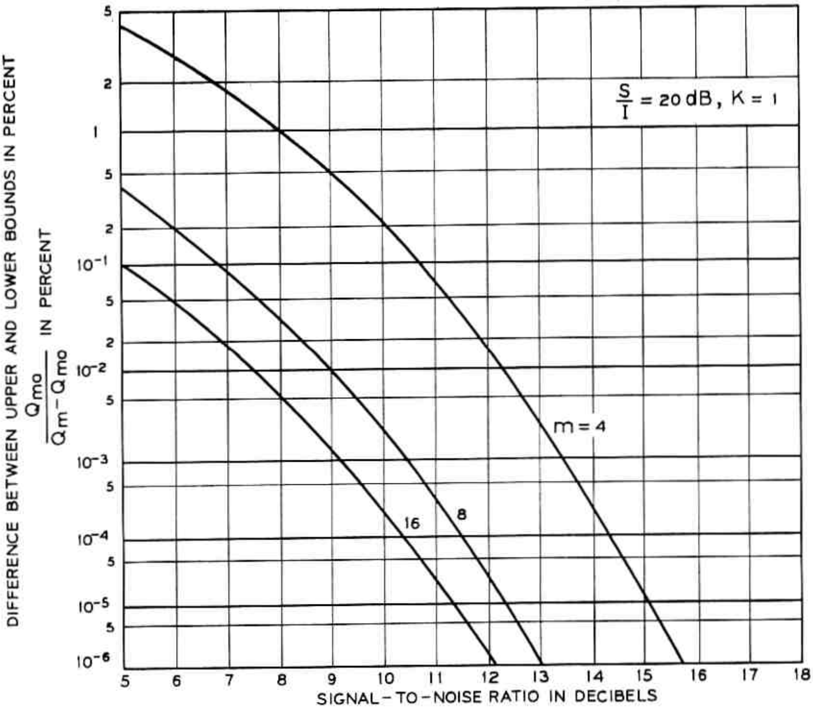


Fig. 9 — $Q_{m0}/(Q_m - Q_{m0})$ as a function of ρ.

distributions of power (if they exist) which make this character error rate a maximum or a minimum.

4.1 *Error Rates for K Interferers*

Let us first consider the case when $\rho \gg 1$ and $\Omega \ll 1$. In this case the series corresponding to P_m (or Q_m) converges very rapidly; let us say that the first N terms of the series are sufficient to evaluate P_m to the desired degree of accuracy.

For all ℓ and z, we have

$$H_{2\ell-1}(z) = 2zH_{2\ell-2}(z) - 2\,(2\ell - 2)H_{2\ell-3}(z). \tag{71}$$

From equation (71) we can show that

$$H_{2j-1}\left(\rho \sin \frac{\pi}{m}\right) \geqq 0, \qquad 1 \leqq j \leqq N, \tag{72}$$

if

$$\rho \geqq \frac{2N - \frac{3}{2}}{\sin \frac{\pi}{m}}, \qquad N > 1. \tag{73}$$

If Equations (72) and (73) are satisfied notice from Equations (31) and (69) that P_m's are monotonically increasing functions of $\mu_{2\ell}$'s, $\ell \geqq 1$. For a given μ_2, it can be shown from Equation (13) that $\mu_{2\ell}$'s $\ell \geqq 2$, reach their minimum when Ω is minimum and they reach their maximum when Ω is maximum.

We can then say that P_m's (or Q_m's) attain their minimum when Ω is minimum and that they are at their maximum when Ω is maximum.

From Figs. 3, 5, and 7 this seems to be true for all values of ρ and Ω which satisfy Equation (55).

Let us now find out when Ω is minimum for a given value of signal-

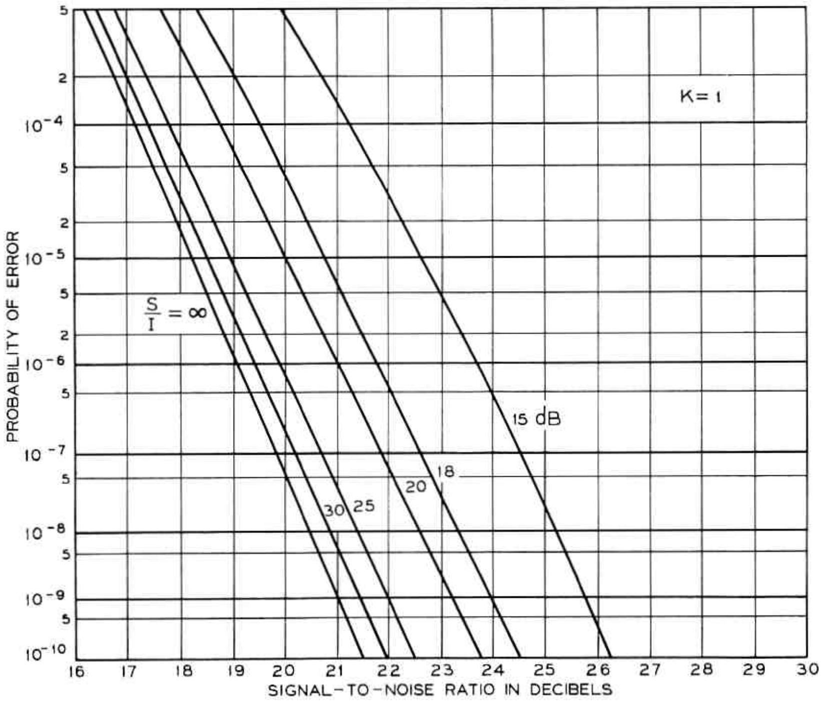


Fig. 10 — Error rates for an 8-phase CPSK system with one interferer.

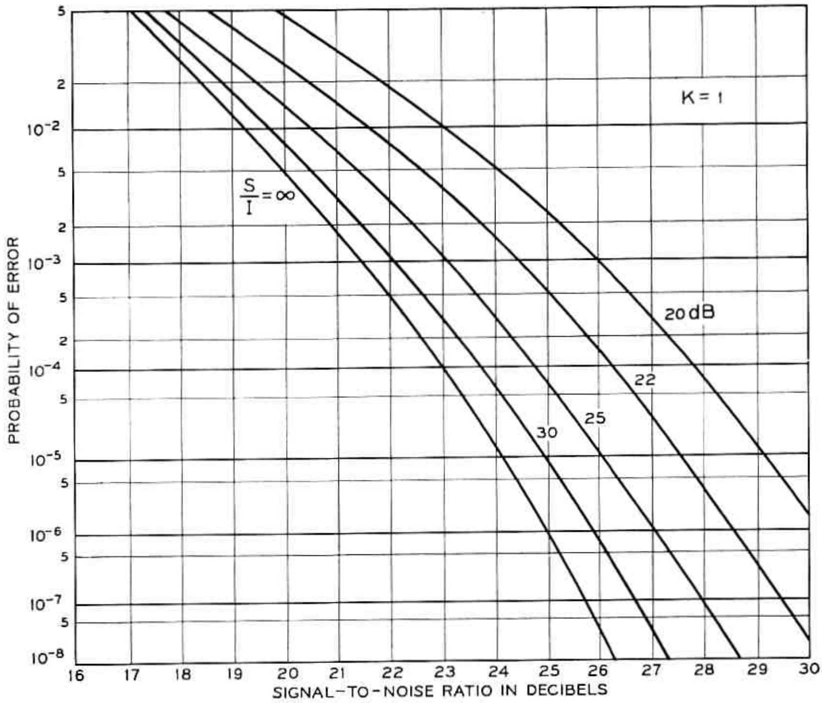


Fig. 11 — Error rates for a 16-phase CPSK system with one interferer.

to-interference ratio. The signal-to-interference ratio $1/L^2$ is given by

$$L^2 = \sum_{j=1}^{K} \frac{I_j}{S}. \tag{74}$$

Clearly Ω is minimum when

$$I_j = SL^2, \qquad 1 \leqq j \leqq K, \tag{75}$$

and

$$I_\ell = 0, \qquad 1 \leqq \ell \leqq K, \ell \neq j. \tag{76}$$

We can then say that the character error rate P_m is minimum when the total interference power is concentrated in a single interferer.

Now from equations (14), (34) and (74), Ω is a maximum when

$$\frac{\partial}{\partial I_j}\left[\sum_{j=1}^{K}\left(\frac{I_j}{S}\right)^{\frac{1}{2}} - \epsilon \sum_{j=1}^{K}\left(\frac{I_j}{S}\right)\right] = 0, \qquad 1 \leqq j \leqq K. \tag{77}$$

ϵ is a constant and is the Lagrange multiplier used in finding the extremum of Ω.

Solving equation (77) we observe that Ω is a maximum or that P_m is a maximum when

$$I_j = \frac{SL^2}{K}, \qquad 1 \leqq j \leqq K \tag{78}$$

or that the total interference power is equally distributed among the K interferers.

Let us now assume that K is a variable number. It is clear from equation (78) that $[P_m]_{\max}$ is a monotonically increasing function of K.

4.2 *Error Rates for a Large Number of Interferers**

Let us now consider the limiting case when K goes to infinity and

$$\sum_{j=1}^{K} I_j = SL^2. \tag{79}$$

We can show[10] that the probability distribution function of†

$$y(t) = \sum_{j=1}^{K} (2I_j)^{\frac{1}{2}} \cos \{\omega_j t + \theta_j + \mu_j\} \tag{80}$$

as K goes to infinity approaches that of gaussian noise with mean zero and variance SL^2 under certain conditions.

In this case we have from equation (4)

$$r_N(t) = (2S)^{\frac{1}{2}} \cos (\omega_0 t + \theta) + y(t) + n(t). \tag{81}$$

Since $y(t)$ and $n(t)$ are independent gaussian random variables their sum

$$b(t) = y(t) + n(t) \tag{82}$$

is also a random gaussian variable with mean zero and variance $SL^2 + \sigma^2$.

From equations (81) and (82) we can write

$$r_N(t) = (2S)^{\frac{1}{2}} \cos (\omega_0 t + \theta) + b(t) \tag{83}$$

where $b(t)$ is a gaussian random variable.

* The results of this section are applicable for any signal-to-noise ratio and any signal-to-interference ratio.

† Ruthroff has shown that for $K \geqq 50$ the distribution of $y(t)$ can be considered to be gaussian in practice for the computation of distortion in PM systems.[16]

The case where $r_N(t)$ can be described by equation (83) has been considered in detail in Ref. 17;* we can easily determine the deterioration in performance produced by interference from the results presented in that paper. For example, suppose that $m = 4$, $S/\sigma^2 = 16$ dB, and $L^2 = -16$ dB. Clearly

$$\frac{\sigma^2}{S} + L^2 = -13 \text{ dB} \tag{84}$$

and P_4 from Ref. 17 is given by

$$P_4 = 7.9 \times 10^{-6}. \tag{85}$$

For the calculation of the effect of interference in CPSK systems, we note that we have not shown the validity of the gaussian approximation of $y(t)$ for $K \gg 1$. However, this assumption seems to be justified for large signal-to-noise ratios and small interference-to-signal ratios.[15]

In conclusion, this section gives methods of evaluating character error rates of CPSK systems for all values of m and for all values of K. It shows that the error rate P_m is minimum when all the interference power is concentrated in a single interferer and that it attains its maximum value $[P_m]_{\max}$ when the interference power is equally distributed amongst all the interferers. We further show that $[P_m]_{\max}$ is a monotonically increasing function of the number K of interferers. We also show that the case, K going to infinity, can be treated and that the deterioration in performance produced by interference can be determined.

V. CONCLUSIONS

A method to evaluate the character error rates of CPSK systems has been presented in this paper. The received signal is assumed to be corrupted by both interference and random gaussian noise. When the number of interferers is very large it can be shown that the interference and random gaussian noise can be combined together to give rise to an equivalent noise source having gaussian properties. The variance of this random variable is the sum of variance of random gaussian noise and total interference power. In this case the analysis of the CPSK system can be done by methods presented in Ref. 17.

When K is a finite number and when $m = 2$ or 4, exact expressions

* The results presented in this paper for $S/I = \infty$ are also sufficient to determine P_m for a large number of interferers.

are given for the probability of error P_m. When $m \geqq 3$, upper and lower bounds to P_m are derived. We show that the difference between these two bounds is a monotonically decreasing function of signal-to-noise ratio ρ^2, signal-to-interference ratio $1/L^2$, and the number m of phases used in the system. For $K = 1$, $m \geqq 4$, $\rho^2 \geqq 5$ dB, and $1/R_1 \geqq 10$ dB we show that this difference is less than 5 per cent, and that the upper bound can be used as a good approximation to P_m.

We then show that for any m-phase CPSK system the character error rates can be expressed in terms of the central moments of a certain random variable η and that they can be calculated to any desired degree of accuracy by using a set of tables or by using a digital computer.

For a total given interference power we show that the character error rate P_m attains its minimum when all the power is concentrated in a single interferer, and that it reaches its maximum $[P_m]_{\max}$ when the power is equally distributed among all the K interferers. It is also shown that $[P_m]_{\max}$ is a monotonically increasing function of K.

The cases of $K = 1$, $m = 2, 4, 8$, and 16, have been treated in some detail and the results are given in graphical form. The required signal-to-noise ratio for any value of signal-to-interference ratio can be determined from these figures.

The usefulness of the presented results is that they provide the designer with some relatively simple expressions with which to evaluate the performance of any given CPSK system with interference and random gaussian noise. The only quantities he must have at his disposal are the central moments of a certain random variable η defined in terms of the K interfering carriers.

VI. ACKNOWLEDGMENT

I would like to express my very sincere thanks to Clyde L. Ruthroff for his help and advice in the preparation of this paper, as well as, for suggesting the approach taken in this paper.

APPENDIX

Evaluation of Central Moments of η

In the computation of character error rates for CPSK systems it is necessary to calculate the even order moments of the random variable η; we shall give in this section two alternate methods to evaluate these moments.

By definition μ_{2n} is given by

$$\mu_{2n} = \frac{1}{(2\pi)^K} \int_0^{2\pi} d\theta_1 \int_0^{2\pi} d\theta_2 \cdots \int_0^{2\pi} d\theta_K \left[\sum_{j=1}^{K} R_j \cos \theta_j \right]^{2n}. \tag{86}$$

By the multinomial theorem

$$\left[\sum_{j=1}^{K} R_j \cos \theta_j \right]^{2n} = \sum \frac{(2n)!}{\prod_{j=1}^{K} n_j!} \prod_{j=1}^{K} (R_j)^{n_j} \cos^{n_j} \theta_j \tag{87}$$

where n_j's are positive integers such that

$$\sum_{j=1}^{K} n_j = 2n. \tag{88}$$

Since θ_j's are statistically independent of each other, and since $\mu_{2\ell+1} = 0$ for all ℓ, we have from Equations (86) and (87)*

$$\mu_{2n} = \sum \frac{(2n)!}{\prod_{\ell=1}^{K} n_\ell!} \prod_{\ell=1}^{K} (R_\ell)^{n_\ell} \frac{(n_\ell)!}{2^{n_\ell}\left[\left(\frac{n_\ell}{2}\right)!\right]^2}, \tag{89}$$

where n_ℓ's are a set of even positive integers satisfying Equation (88).

Even though equation (89) gives an exact expression to evaluate μ_{2n}'s, it can be rather tedious to evaluate μ_{2n}'s from equation (89) for large values of n and K. We shall therefore give an alternate method to evaluate the central moments of the random variable η.

It can be shown that the probability density function $p_\eta(\eta)$ of the random variable η can be expressed as[10]

$$p_\eta(\eta) = \frac{1}{2\Omega}\left[1 + 2 \sum_{s=1}^{\infty} \cos \frac{s\pi\eta}{\Omega} \prod_{j=1}^{K} J_0\left(\frac{s\pi R_j}{\Omega}\right)\right]. \tag{90}$$

The $2n$th moment of η can be represented as

$$\mu_{2n} = \int_{-\Omega}^{\Omega} z^{2n} p_\eta(z)\, dz. \tag{91}$$

From equations (90) and (91) we can show that

$$\mu_{2n} = \Omega^{2n}\left(\frac{1}{2n+1} + 2 \sum_{\ell=1}^{\infty} (-1)^{\ell+1} \cdot \left\{\left[\prod_{j=1}^{K} J_0\left(\frac{\ell\pi R_j}{\Omega}\right)\right] \sum_{k=1}^{n} (-1)^k \frac{(2n)!}{[2n-2k+1]!\,(\ell\pi)^{2k}}\right\}\right). \tag{92}$$

* For $K = 1$, equation (89) reduces to equation (37).

It can be seen that the infinite series appearing in equation (92) converges rapidly for all values of R_j's; we need take only a finite number of terms from equation (92) to estimate μ_{2n}'s. It is, therefore, easier to evaluate μ_{2n}'s from equation (92) than from equation (89) when there are a large number of interferers, and we have to take a large number of terms in estimating P_m.

REFERENCES

1. Lawton, J. G., "Comparison of Binary Data Transmission Systems," Proc. Nat'l. Convention on Military Electronics, Washington, D. C., August 1958, pp. 54–61.
2. Helstrom, C. W., "The Resolution of Signals in White, Gaussian Noise," Proc. IEEE, *43,* No. 9 (September 1955), pp. 1111–1118.
3. Cahn, C. R., "Performance of Digital Phase-Modulation Systems," IEEE Trans. on Communication Systems, *CS-7,* No. 1 (May 1959), pp. 3–6.
4. Cahn, C. R., "Comparison of Coherent and Phase-Comparison Detection of a Four-Phase Signal," Proc. IEEE, *47,* No. 9 (September 1959), pp. 1662.
5. Arthurs, E. and Dym, H., "On the Optimum Detection of Digital Signals in the Presence of White Gaussian Noise—A Geometric Interpretation and a Study of Three Basic Data Transmission Systems," IEEE Trans. on Communication Systems, *10,* No. 4 (December 1962), pp. 336–372.
6. Tillotson, L. C. and Ruthroff, C. L., "The Next Generation of Short Haul Radio Systems," unpublished work.
7. Montgomery, G. F., "A Comparison of Amplitude and Angle Modulation for Narrow-Band Communication of Binary-Coded Messages in Fluctuation Noise," Proc. IEEE, *42,* No. 2 (February 1954), pp. 447–454.
8. Rowe, H. E., *Signals and Noise in Communication Systems,* Princeton, N. J.: D. Van Nostrand Co., Inc., 1965, pp. 13–16.
9. Rice, S. O., "Mathematical Analysis of Random Noise," B.S.T.J., *23,* No. 3 (July 1944), pp. 282–332.
10. Bennett, W. R., "Distribution of the Sum of Randomly Phased Components," Quart. Appl. Math., *5,* No. 1 (April 1948), pp. 385–393.
11. Morse, P. M. and Feshbach, H., *Methods of Theoretical Physics,* New York: McGraw-Hill Book Co., Inc., 1953, pp. 786–787.
12. Magnus, W. and Oberhettinger, F., *Formulas and Theorems for the Functions of Mathematical Physics,* New York: Chelsea Publishing Co., 1943, pp. 80–82.
13. Rosenbaum, A. S., "Error Performance of Coherently Detected PSK Signals in the Presence of Gaussian Noise and Co-channel Interference," unpublished work.
14. Pagones, M. J., "Error Probability Upper Bound of a Coherently Detected PSK Signal Corrupted by Interference and Gaussian Noise," unpublished work.
15. Koerner, M. A., "Effect of Interference on a Binary Communication Channel Using Known Signals," Technical Rep. 32-1281, Jet Propulsion Lab., California Inst. Technology, Pasadena, Calif., December 1968.
16. Ruthroff, C. L., "Computation of FM Distortion in Linear Networks for Bandlimited Periodic Signals," B.S.T.J., *47,* No. 6 (July–August 1968), pp. 1043–1063.
17. Prabhu, V. K., "Error Rate Considerations for Digital Phase-Modulation Systems," to be published in IEEE Trans. on Communication Technology, *17,* No. 1 (February 1969).

Spectral Density Bounds of a PM Wave

By V. K. PRABHU and H. E. ROWE

(Manuscript received November 5, 1968)

In this paper we derive upper and lower bounds of the spectrum of a sinusoidal carrier phase modulated by gaussian noise having a rectangular power spectrum. It has been found in practice that such a random process adequately simulates for some purposes, a frequency division multiplex signal, a composite speech signal, and so on. We show that these upper and lower bounds of the spectrum are very close to each other if the root mean square phase deviation of the carrier is even moderately high. Also, a simple method called the saddle-point method can be used at all frequencies f *to estimate the spectrum with less than ten percent error. We also show that the results obtained from the quasistatic approximation, often used in such cases, are too small for large* f, *and that this low-frequency approximation cannot be used in cases where the behavior on the tails is of importance.*

I. INTRODUCTION

It has been found in practice that a bandlimited random gaussian noise having a rectangular power spectrum adequately simulates for some purposes a wideband composite speech signal, a frequency division multiplex baseband signal consisting of a group of single sideband carrier telephone channels, and so on.[1] In the design of communication systems, the spectral characteristics of a sinusoidal carrier phase modulated by such a baseband signal are of great interest; various methods have been used in recent years to estimate this spectrum for large and small values of mean square phase deviation of the wave, both close to and far from the carrier frequency (that is, in the principal part of the spectrum and far down on the tails of the spectrum respectively).[1–8]

It has been shown that the spectrum may be expanded as an infinite series of weighted convolution terms.[2,5,7,8] This series may be used to estimate the principal part of the spectrum (close to the carrier) for small or moderate index (that is, small or moderate values of rms phase deviation). However, for large index, or far down on the tails for small index, too many terms would have to be included if this series is to be used directly.

The simplest analysis—often called the quasistatic approximation—yields a gaussian spectrum for large-index angle-modulated waves[2, 4–8] in most cases.* This approximation fails far out on the tails of the spectrum; a careful investigation has been given in only a few cases.[7] We obtain below upper and lower bounds for the spectrum of an angle-modulated wave with white, band-limited phase modulation; far out on the tails the spectrum far exceeds that predicted by the quasistatic approximation.

This problem is of interest in considering interference between two (or more) phase modulation (PM) systems in neighboring locations. Consider the following situation. In the frequency bands above 10 GHz, where the signal attenuation due to rain storms could be very severe, close spacings of the repeaters are almost mandatory for reliable communication from point to point.[9] In such cases the problem of interference between neighboring systems may be much more important than the problem of noise; the system may thus be interference limited. In order to combat this interference it is very likely that broadband modulation techniques like PM [or frequency modulation (FM)] or pulse code modulation (PCM) will have to be used. In order to investigate the effect of this interference between two co-channel PM (or FM) waves it is necessary to evaluate the spectrum of a PM wave, so that the parameters (such as rms phase deviations) of the two PM systems can be properly chosen to keep the interchannel interference below a certain desired level.

We first obtain an expression for the covariance function of the PM wave. From this covariance function we then derive an expression for the spectrum of the PM wave and show that it can be expressed as an infinite series. This series has been evaluated for certain values of rms phase deviation N.[6]

We then show that the autocorrelation function of the PM wave is analytic at all points in the finite part of the complex plane determined by the argument of the autocorrelation function. In determining the Fourier transform of the autocorrelation function we change the path of integration† so that the contour is very close to the path

* For exceptional cases see Ref. 7, Ch. 4, pp. 131–135.

† The method used in this paper to evaluate the spectrum is a close relative of the method of steepest descent (or the saddle-point method) used in evaluating certain kinds of integrals.[1, 7, 10, 11, 12]

of steepest descent of the integrand. We then divide this contour into four (or five) disjoint sections and show that the major contribution to the integral comes from one of these sections.

We next derive upper and lower bounds to the spectrum $S_V(f)$ of the PM wave and show that these bounds are very close for all f and for all values of rms phase deviation $N \geqq 5$. For $N \geqq (10)^{\frac{1}{2}}$ we show that the spectrum can be evaluated by this saddle-point method in a very simple manner with a very small fractional error (less than 10 percent), and we give this method.

We finally compare the quasistatic approximation to the saddle-point approximation. For large values of frequency f, we show that the quasistatic approximation gives too small a value for the spectrum, and that it cannot be used in cases where the spectral behavior on the tails is of importance. However, as we show, the saddle-point method can be used in all cases in which N is moderately high.

In conclusion, this paper gives a simple method of evaluating the spectrum of a sinusoidal carrier phase modulated by gaussian noise having a rectangular power spectrum and having a moderately high modulation index.

II. SPECTRAL ANALYSIS OF PM WAVES WITH RANDOM PHASE MODULATION

A sinusoidal wave of constant-amplitude phase modulated by a signal $n(t)$ may be written as

$$W(t) = A \cos [\omega_0 t + n(t) + \theta], \tag{1}$$

where A is the amplitude of the wave, $f_0 = \omega_0/2\pi$ is the carrier frequency of the wave, and θ is a random variable with probability density*

$$\pi_\theta(\theta) = \begin{cases} \dfrac{1}{2\pi}, & 0 \leqq \theta < 2\pi \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Assume that $n(t)$ is a stationary bandlimited white gaussian random process with mean zero and variance N^2.† Its spectral density

* If $n(t)$ is a stationary random process the introduction of random variable θ in equation (1) makes $W(t)$ a random process which is at least wide-sense stationary so that its spectrum can be calculated from the Wiener-Khintchine theorem.[2, 7] If we do not have θ in equation (1), $W(t)$ is no longer (even wide-sense) stationary, and the spectrum of $W(t)$ is usually calculated from the time autocorrelation function of $W(t)$.[7] The results obtained in the two cases are identical.

† The parameter N represents the rms phase deviation (or modulation index) of the PM wave given in equation (1).

$S_n(f)$ can be represented (see Fig. 1) as

$$S_n(f) = \begin{cases} N^2/2W, & |f| < W, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Such a random process $n(t)$, is often used to simulate a multiplex signal, a composite speech signal, and so on.[1, 5]

We can show from equation (3) that the covariance function $R_n(\tau)$ of $n(t)$ is given by

$$R_n(\tau) = N^2 \frac{\sin 2\pi W\tau}{2\pi W\tau} ; \tag{4}$$

this function $R_n(\tau)$ is shown in Fig. 2. Since $n(t)$ is a stationary gaussian random process it can be shown that $W(t)$ is at least wide-sense stationary and that its covariance function $R_W(\tau)$ can be represented as[2, 7]

$$R_W(\tau) = \frac{A^2}{2} \exp\,[-R_N(0)] \exp\,[R_N(\tau)] \cos \omega_0\tau. \tag{5}$$

From the Wiener-Khintchine theorem, and from equation (5), the spectrum $S_W(f)$ of $W(t)$ can be written as

$$S_W(f) = \int_{-\infty}^{\infty} R_W(\tau) \exp\,[-j2\pi f\tau]\, d\tau, \tag{6}$$

or

$$S_W(f) = \frac{A^2}{4} [S_V(f - f_0) + S_V(f + f_0)], \tag{7}$$

where

$$S_V(f) = \int_{-\infty}^{\infty} \exp\,[-R_N(0)] \exp\,[R_N(\tau)] \exp\,[-j2\pi f\tau]\, d\tau. \tag{8}$$

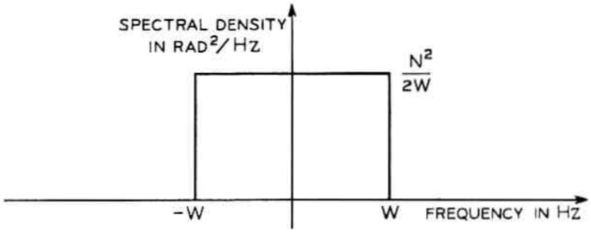


Fig. 1 — Spectral density of phase modulation.

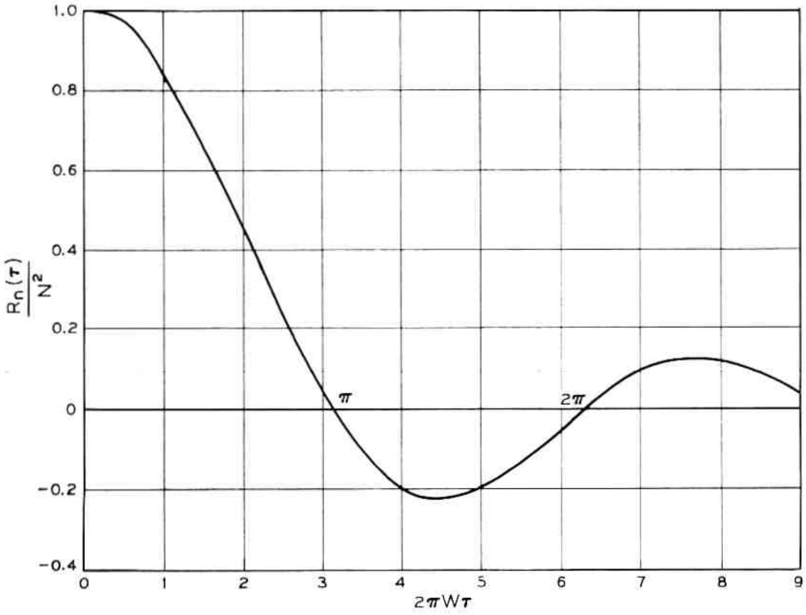


Fig. 2 — Covariance function of $n(t)$. Since $R_n(\tau)$ is an even function of τ we only show $R_n(\tau)$ for $\tau \geqq 0$.

From equations (4) and (8) we have

$$S_v(f) = \frac{1}{2\pi W} \int_{-\infty}^{\infty} \exp\left\{-N^2\left[1 - \frac{\sin p}{p}\right]\right\} \exp\left[-j\lambda p\right] dp, \qquad (9)$$

where

$$\lambda = \frac{f}{W}. \qquad (10)$$

III. SERIES EXPANSION OF SPECTRUM FOR GAUSSIAN MODULATION

The integral in Equation (9) can be evaluated by expanding

$$\exp\left\{-N^2\left[1 - \frac{\sin p}{p}\right]\right\}$$

into a Taylor series; integrating term by term we can write*

$$\exp\left\{-N^2\left[1 - \frac{\sin p}{p}\right]\right\} = \exp\left[-N^2\right] \sum_{\ell=0}^{\infty} \frac{N^{2\ell}}{\ell!} \left(\frac{\sin p}{p}\right)^{\ell}. \qquad (11)$$

* We note that $\sum_{n=0}^{\infty} x^n/n!$ converges uniformly to $\exp[x]$ for all finite values of x.

From equations (9) and (11) we have*

$$S_V(f) = \exp[-N^2] \cdot \left\{\delta(f) + \frac{1}{2\pi W}\sum_{\ell=1}^{\infty}\frac{N^{2\ell}}{\ell!}\int_{-\infty}^{\infty}\left(\frac{\sin p}{p}\right)^{\ell}\exp[-j\lambda p]\right\}dp, \tag{12}$$

where $\delta(f)$ is the Dirac delta (unit impulse) function.

We now note that

$$\int_{-\infty}^{\infty}\frac{\sin p}{p}\exp[-j\lambda p]\,dp = F_1(\lambda) = \begin{cases}\pi, & |\lambda| < 1,\\ 0, & \text{otherwise},\end{cases} \tag{13}$$

or

$$F_1(\lambda) = \pi[u_{-1}(\lambda + 1) - u_{-1}(\lambda - 1)], \tag{14}$$

where $u_{-1}(x)$ is the unit step function defined by

$$u_{-1}(x) = \begin{cases}1, & x > 0,\\ 0, & x < 0,\end{cases} \tag{15}$$

and that[1, 13]

$$\int_{-\infty}^{\infty}\left(\frac{\sin p}{p}\right)^{\ell}\exp[-j\lambda p]\,dp = F_\ell(\lambda)$$
$$= \begin{cases}\dfrac{\ell\pi}{2^{\ell-1}}\displaystyle\sum_{k=0}^{M}(-1)^k\frac{(|\lambda| + \ell - 2k)^{\ell-1}}{k!\cdot(\ell-k)!}, & 0 \leqq |\lambda| < \ell,\ \ell \geqq 2,\\ 0, & \text{otherwise},\end{cases} \tag{16}$$

where

$$M = INT\left[\frac{\ell + |\lambda|}{2}\right], \tag{17}$$

and $INT[x]$ represents the largest integer not greater than x.

It can be shown that $F_\ell(\lambda)$, $\ell \geqq 2$ is a continuous function of λ and that $F_1(\lambda)$ is discontinuous at $\lambda = 1$. For large ℓ, we can show from the central-limit theorem† that[2]

$$F_\ell(\lambda) \sim \left(\frac{6\pi}{\ell}\right)^{\frac{1}{2}}\exp\left[\frac{-3\lambda^2}{2\ell}\right]. \tag{18}$$

* The term containing $\delta(f)$ in equation (12) represents the dc component of $S_V(f)$.

† See pp. 362–366 of Ref. 2. It can be shown that $\left(\frac{\sin p}{p}\right)^{\ell}$ can be interpreted as the characteristic function of the sum Ω of ℓ independent random variables with identical uniform probability distributions.[5] The function $F_\ell(\lambda)/2\pi$ therefore represents the probability density function of Ω. Alternatively $F_\ell(\lambda)$ is the $(\ell - 1)$-fold convolution of the flat spectrum with itself.

From equations (12), (13), and (16) we can write

$$S_V(f) = \exp\,[-N^2]\left[\delta(f) + \frac{1}{2\pi W}\sum_{\ell=1}^{\infty}\frac{N^{2\ell}}{\ell!}F_\ell(f/W)\right]\cdot \qquad (19)$$

For $N^2 = 6$, we have calculated the spectrum from equation (19) and the results are shown in Fig. 3. Notice that the spectral density is discontinuous at $f/W = 1$.

For $N^2 \ll 1$ (low-index modulation), we have from (19)

$$S_V(f) \approx S_{V0}(f) = \exp\,[-N^2]\left[\delta(f) + \frac{N^2}{2\pi W}F_1(f/W)\right], \qquad (20)$$

and the error in this approximation may be investigated from equations (9) and (20).* The approximation given in equation (20) represents the low-index approximation for the spectrum; this result has been obtained by many authors.[3-6]

The series given in equation (19) may be used to estimate the principal part of the spectrum (close to the carrier) for small or moderate index, since only a small number of terms need to be included in the partial sum. However, for large N^2, or far down on the tails of the spectrum for small N^2, too many terms would have to be included to estimate the spectral density. In fact for $N^2 \gg 1$, or for $f/W \gg 1$, the degree of complexity required in estimating $S_V(f)$ from equation (19) becomes inordinately high.

When $N^2 \gg 1$, and for low frequencies, several authors have given[1-7] the quasistatic approximation†

$$S_V(f) \approx \exp\,(-N^2)\,\delta(f) + \frac{1}{NW}\left(\frac{3}{2\pi}\right)^{\frac{1}{2}}\exp\left[-\frac{3}{2N^2}\left(\frac{f}{W}\right)^2\right] \qquad (21)$$

for the spectrum. The question arises whether equation (21) can be used for large f. Since $R_W(\tau)$ is infinitely differentiable there is no simple way (known to the authors) of investigating, for large f, the error is this approximation.[7]

In the problem of interference between two neighboring channels it is necessary to evaluate $S_V(f)$ for large f so that the effect of this interference can be determined. As we shall show later on in this paper equation (21) gives too small values to $S_V(f)$ for large f; it is therefore essential to have a simple and elegant method (different from the series method) to evaluate $S_V(f)$ for large f and for large N^2.

* At times the low-index approximation for the spectrum is written as $\exp\,[-N^2]\delta(f) + (N^2/2\pi W)F_1(f/W)$. For $N^2 \ll 1$, $\exp\,[-N^2] \approx 1$.

† Note that mean square frequency deviation is $N^2W^2/3$.

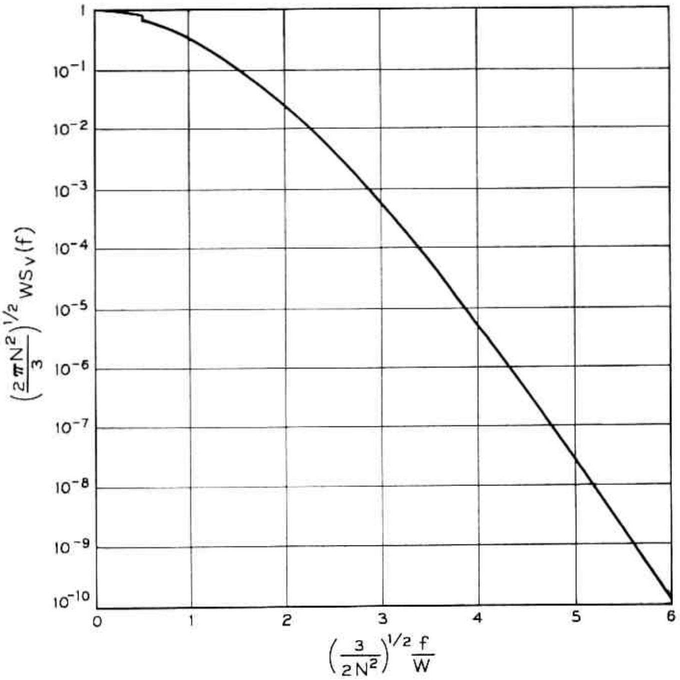


Fig. 3 — Spectral density of a PM wave for $N^2 = 6$. The discrete part of $S_V(f)$ for $f = 0$ is not shown in this figure. Note the discontinuity in the spectrum at $f/W = 1$.

Readers who might be interested in the final results without wishing to work through the detailed analysis, might skip Section IV of this paper.

IV. SPECTRUM EVALUATION BY CONTOUR INTEGRATION

Let us now consider the integral given in equation (9). Since $\delta(f)$ and $F_1(f/W)$ are discontinuous functions of f, let us define an integral*

$$S(f) = \int_{-\infty}^{\infty} \exp\left[-N^2\right]\left\{\exp\left[N^2 \frac{\sin p}{p}\right] - \left(1 + N^2 \frac{\sin p}{p}\right)\right\} \cos \lambda p \, dp$$

* In Ref. 14 this integral has been studied by Lewin for $\lambda = 0$ and $\lambda = 1$. It also occurs in several limiting cases in Ref. 1. It is sometimes referred to by the name Lewin's integral.[1]

or

$$S(f) = \int_{-\infty}^{\infty} \exp\,[-N^2]\left\{\exp\left[N^2 \frac{\sin p}{p}\right] - \left(1 + N^2 \frac{\sin p}{p}\right)\right\} \cdot \exp\,(-j\lambda p)\; dp. \tag{22}$$

Since all $F_\ell(f)$, $\ell \geqq 2$ are continuous, it can be shown from equations (19), and (22) that $S(f)$ is a continuous function of f.

From equations (9), (19), and (22) we can then write*

$$S_V(f) = \exp\,(-N^2)\left\{\delta(f) + \frac{N^2}{2W}\,[u_{-1}(f + W) - u_{-1}(f - W)]\right\} + \frac{1}{2\pi W}\,\mathrm{Re}\;S(f). \tag{23}$$

Notice from equation (22) that the integration is carried out along the real axis, and that for large $|\lambda|$ (or $|f|/W$), the final factor of the integrand $\exp\,(-j\lambda p)$ is a very rapidly oscillating function of p. From Refs. 7, 10–13 notice that in such circumstances the method of steepest descent (or saddle-point method), or one of its close relatives, is often useful to get an approximate expression for the integral; we shall now apply such a method to evaluate $S_V(f)$.

In applying this method to the evaluation of an integral with a real variable of integration, we must first be able to regard the integral as a contour integral along the real axis of the complex plane, with an analytic integrand. We note that the integrand in equation (22) is an analytic function of p, and that it has no singularities in the finite part of the complex plane (defined by p). From Cauchy's theorem it therefore follows that the contour of integration can be arbitrarily deformed as long as one end is at $p = -\infty + j0$ and the other at $p = \infty + j0$.[11]

In making use of the method of steepest descent the contour must be deformed so that the phase of the integrand remains constant (or almost so), while the magnitude of the integrand is small except in one or more localized regions, where it varies rapidly. This is usually accomplished by deforming the contour so that it goes through one or more saddle points. In other cases there may be some additional constraints on the contour;[7] then only a portion of the path of steep-

* Re z and Im z denote respectively the real and imaginary parts of complex number z.

est descent through a saddle point may be used in finding the integral, and the deformed contour may not actually reach the saddle point. In this case the original integral is usually reduced to a virtually real integral whose integrand behaves sufficiently simply on the modified path of integration so that an approximate evaluation of the integral with rigorous (upper and lower) bounds on the error may be obtained.

Departures from the strict method of steepest descent occur in this paper in that approximate paths of steepest descent are chosen. Although not quite optimum, they are analytically tractable and serve to give useful bounds on the integral under consideration.

Consider equation (22). Since the integrand in equation (22) is an analytic function of p, let us assume that $p = x + jy$ is a complex variable, and let us deform the contour so as to obtain the path of steepest descent. Since the integrand behaves properly on the contour for large $|p|$, it is clear that the contour of integration can be deformed in quite a general way in the complex p-plane without modifying the value of the integral.

From equation (22) it can be shown that the major portion of the integrand

$$R(p) \equiv \exp\,[-N^2]\,\exp\left[N^2\,\frac{\sin p}{p}\right]\exp\,[j\lambda p] \tag{24}$$

has a saddle point on the imaginary axis, with the path of steepest descent parallel to the real axis at this point. The location $p_s = jy_s$ of this saddle point is given by

$$\frac{\cosh y_s}{y_s} - \frac{\sinh y_s}{y_s^2} = \frac{\lambda}{N^2} = \frac{f}{N^2 W}\,; \tag{25}$$

for a given f/N^2W, equation (25) can be solved numerically to give the required y_s. We plot y_s as a function of f/N^2W in Fig. 4, and ln $R(jy_s)$ in Fig. 5.

Let us now deform the contour so that it passes through the point $p_s = jy_s$ and is parallel to the real axis at this point. From equation (22) we then have

$$S(f) = \exp\left[-2N^2\left(\cosh^2\frac{y_s}{2} - \frac{\sinh y_s}{y_s}\right)\right]\,\mathrm{Re}\,I \tag{26}$$

where

$$
\begin{aligned}
I = {} & \exp\left[-N^2 \frac{\sinh y_s}{y_s}\right] \\
& \cdot \int_{-\infty}^{\infty} \left\{ \exp\left[N^2 \frac{\sin (x + jy_s)}{x + jy_s}\right] - \left[1 + N^2 \frac{\sin (x + jy_s)}{x + jy_s}\right]\right\} \\
& \cdot \exp\,[j\lambda x]\, dx.
\end{aligned} \tag{27}
$$

Rewriting equation (27)

$$
I = \int_{-\infty}^{\infty} G(x, y_s)\, dx, \tag{28}
$$

where

$$
\begin{aligned}
G(x, y_s) \equiv {} & \exp\,[-Q_R(x, y_s)] \exp\,[jQ_I(x, y_s)] \\
& - \exp\left[-N^2 \frac{\sinh y_s}{y_s}\right]\left\{1 + N^2 \frac{\sin (x + jy_s)}{x + jy_s}\right\} \exp\,(j\lambda x),
\end{aligned} \tag{29}
$$

where $Q_R(x, y_s)$ and $Q_I(x, y_s)$ are real and

$$
Q_R(x, y_s) = N^2 \left\{ \frac{\sinh y_s}{y_s} - \operatorname{Re}\left[\frac{\sin (x + jy_s)}{x + jy_s}\right]\right\} \tag{30}
$$

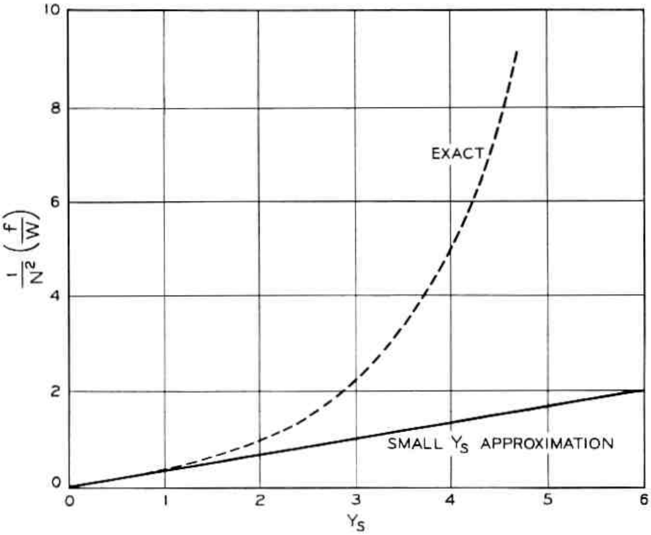


Fig. 4 — Location of saddle-point y_s.

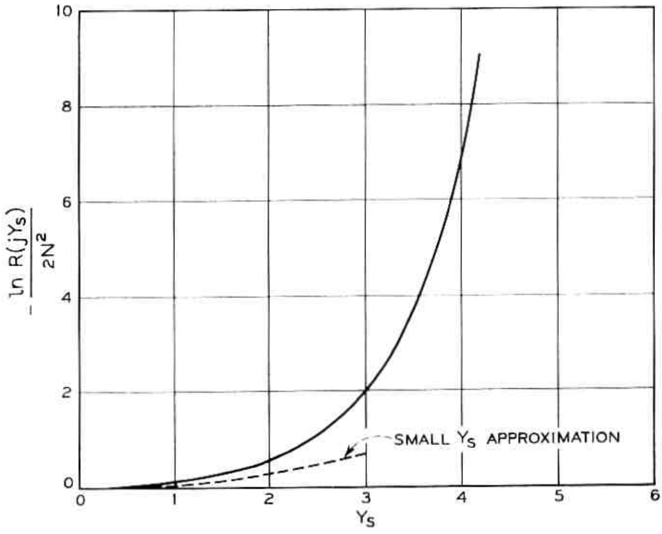


Fig. 5 — Value of $-[\ln R(jy_s)]/(2N^2)$.

or

$$Q_R(x, y_s) = N^2 \frac{\sinh y_s}{y_s} x^2 \frac{\dfrac{1 - \dfrac{y_s}{\tanh y_s} \dfrac{\sin x}{x}}{y_s^2} + \dfrac{1 - \cos x}{x^2}}{1 + (x/y_s)^2} \tag{31}$$

and

$$Q_I(x, y_s) = N^2 \operatorname{Im}\left[\frac{\sin (x + jy_s)}{x + jy_s}\right] + \lambda x, \tag{32}$$

or

$$Q_I(x, y_s)$$

$$= N^2 \left\{\frac{\sinh y_s \cos x - y_s \cosh y_s \dfrac{\sin x}{x}}{x^2 + y_s^2} + \left[\frac{\cosh y_s}{y_s} - \frac{\sinh y_s}{y_s^2}\right]\right\} x. \tag{33}$$

The functions $Q_R(x, y_s)$ and $Q_I(x, y_s)$ have been plotted in Fig. 6 for a set of values of y_s .

Since we are primarily interested in the high-index case let us assume that $N^2 \gg 1$. From equations (27)–(33) we now observe that the principal contribution to the integral I comes from small x. For small x,

$$G(x, y_s) \approx \exp\left[-Q_R(x, y_s)\right]. \tag{34}$$

It can be shown (see Figs. 6, 7, and Appendix A) that $Q_R(x, y_s)$ is a monotonically increasing function of x for $0 \leqq x \leqq \pi$ and that it oscillates for values of $x > \pi$. For large x,

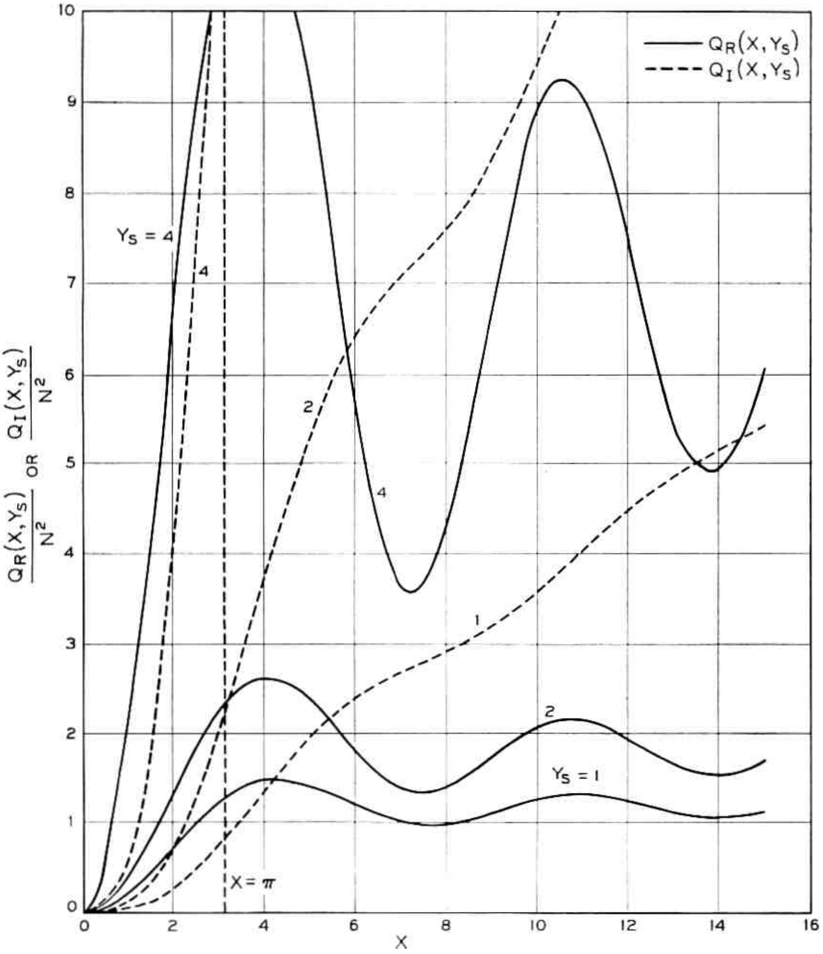


Fig. 6 — Functions $Q_R(x, y_s)$ and $Q_I(x, y_s)$.

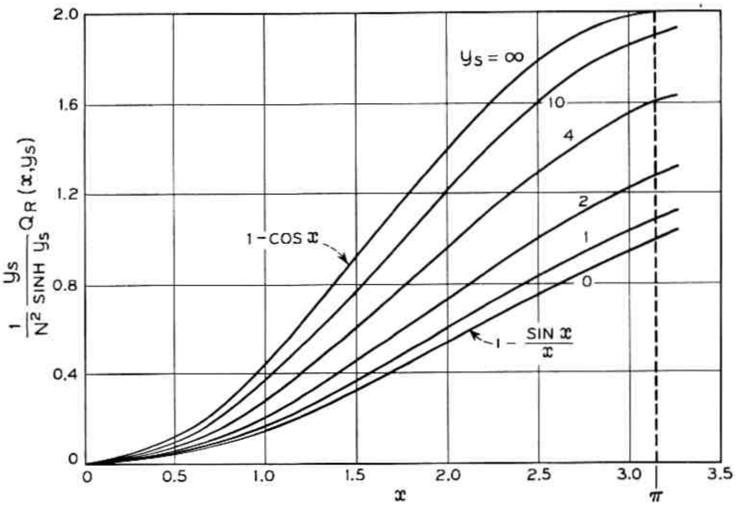


Fig. 7 — Function $Q_R(x, y_s)$. From this figure, it can be seen that $Q_R(x, y_s)$ is a monotonically increasing function of x for $0 \leqq x \leqq \pi$.

$$G(x, y_s) \approx \exp\left[-N^2 \frac{\sinh y_s}{y_s}\right] \cdot \left\{ \exp\left[N^2 \cosh y_s \frac{\sin x}{x}\right] \exp\left[jN^2 \sinh y_s \frac{\cos x}{x}\right] - \left[1 + N^2\left(\cosh y_s \frac{\sin x}{x} + j \sinh y_s \frac{\cos x}{x}\right)\right]\right\} e^{j\lambda x}, \tag{35}$$

and we note that the first and second terms both have small and almost equal magnitude, and almost opposite phase angle, so that they almost cancel. As $|x| \to \infty$ the cancellation becomes exact. For these reasons it is convenient to divide the range of integration in equation (27) into at least four regions:

$$\begin{aligned} 0 < |x| < x_1, &\quad \text{small } |x|, \\ x_1 < |x| < \pi, &\quad \text{intermediate } |x|, \\ \pi < |x| < x_2, &\quad \text{intermediate } |x|, \\ x_2 < |x| < \infty, &\quad \text{large } |x|. \end{aligned} \tag{36}$$

From equations (25), (30), and (32) we can show that*

* These expressions for Q_R and Q_I may be obtained from the Taylor series expansion of the function $[\sin(x + jy_s)]/(x + jy_s)$.

$$Q_R(x, y_s) = N^2 \sum_{\ell=1}^{\infty} (-1)^{\ell-1} A_{2\ell} \frac{x^{2\ell}}{(2\ell)!}, \tag{37}$$

and

$$Q_I(x, y_s) = N^2 \sum_{\ell=1}^{\infty} (-1)^{\ell+1} A_{2\ell+1} \frac{x^{2\ell+1}}{(2\ell+1)!}, \tag{38}$$

where

$$A_0 = \frac{\sinh y_s}{y_s}, \tag{39}$$

$$A_1 = \frac{\cosh y_s}{y_s} - \frac{\sinh y_s}{y_s^2} = \frac{\lambda}{N^2}, \tag{40}$$

$$A_{2\ell} = \frac{\sinh y_s}{y_s} - 2\ell \frac{A_{2\ell-1}}{y_s}, \qquad \ell = 1, 2, 3, \cdots, \tag{41}$$

and

$$A_{2\ell+1} = \frac{\cosh y_s}{y_s} - (2\ell + 1) \frac{A_{2\ell}}{y_s}, \qquad \ell = 1, 2, 3, \cdots. \tag{42}$$

It can also be shown that

$$A_{2k-1} = \sum_{\ell=0}^{\infty} \frac{1}{2\ell + 2k + 1} \frac{y_s^{2\ell+1}}{(2\ell + 1)!}, \qquad k = 1, 2, 3, \cdots, \tag{43}$$

and

$$A_{2k} = \sum_{\ell=0}^{\infty} \frac{1}{2\ell + 2k + 1} \frac{y_s^{2\ell}}{(2\ell)!}, \qquad k = 1, 2, 3, \cdots. \tag{44}$$

Since the spectrum is an even function of f we can assume without loss of generality that

$$y_s \geqq 0. \tag{45}$$

For $y_s \geqq 0$, from equations (43) and (44) all A_ℓ's are monotonically increasing functions of y_s, and we can further show that

$$0 < A_{2(k+1)} < A_{2k}, \qquad k = 0, 1, 2, \cdots, \tag{46}$$

and

$$0 \leqq A_{2\ell+1} \leqq A_{2\ell-1}, \qquad \ell = 1, 2, 3, \cdots. \tag{47}$$

For large y_s (for large f/W), it can also be proved that

$$A_{2\ell} \approx A_{2\ell-1} \approx \frac{\exp(y_s)}{2y_s}. \tag{48}$$

Since it appears that the main contribution to the integral I comes from the region of small $|x|$, assume that $|x_1|$ is small and that

$$I = I_1 + I_R , \tag{49}$$

where

$$I_1 = \int_{-x_1}^{x_1} G(x, y_s)\, dx, \tag{50}$$

and

$$I_R = \int_{-\infty}^{-x_1} G(x, y_s)\, dx + \int_{x_1}^{\infty} G(x, y_s)\, dx. \tag{51}$$

For small $|x|$, we have from equations (37) and (38)

$$Q_R(x) \approx N^2 \tfrac{1}{2} A_2 x^2, \tag{52}$$

and

$$Q_I(x) \approx N^2 \tfrac{1}{6} A_3 x^3. \tag{53}$$

Let us choose* x_1 so that $\exp(-\frac{1}{2}N^2 A_2 x^2)$ falls to $\exp(-5) \approx 0.0067$ for $x = x_1$,† or that

$$x_1 = \left(\frac{10}{N^2 A_2}\right)^{\frac{1}{2}}. \tag{54}$$

Since it can be shown from equations (39)–(44) that the minimum value of A_2 is $\frac{1}{3}$ (at $y_s = 0$),

$$x_1 \leqq \left(\frac{30}{N^2}\right)^{\frac{1}{2}}. \tag{55}$$

Assume that

$$x_1 \leqq \pi, \tag{56}$$

so that $\exp\{-Q_R(x, y_s)\}$ is a monotonically decreasing function of x for $0 \leqq x \leqq x_1$. Equation (56) will be satisfied for all y_s if

$$N^2 \geqq \frac{30}{\pi^2} \approx 3.039. \tag{57}$$

Since we are primarily interested in the high-index case, this is not a significant restriction.

* See equation (34).

† While there is some degree of arbitrariness in such choices, the bounds obtained are not too sensitive to small variations in the value of x_1, x_2, and so on.

Noticing that

$$\sin x \leqq x, \qquad 0 \leqq x < \infty, \tag{58}$$

$$\left| \frac{\sin(x + jy_s)}{x + jy_s} \right| = \frac{(\sin^2 x + \sinh^2 y_s)^{\frac{1}{2}}}{(x^2 + y_s^2)^{\frac{1}{2}}} \leqq \left(\frac{x^2 + \sinh^2 y_s}{x^2 + y_s^2} \right)^{\frac{1}{2}}. \tag{59}$$

Since $(\sinh y_s)/(y_s) \geqq 1$, it can be proved from equation (59) that

$$\left| \frac{\sin(x + jy_s)}{x + jy_s} \right| \leqq \frac{\sinh y_s}{y_s}. \tag{60}$$

We can show from equations (29), (50), and (60) that

$$| I_1 | < 2 \int_0^{x_1} \{\exp[-Q_R(x, y_s)] + H(x, y_s)\}\, dx, \tag{61}$$

where

$$H(x, y_s) = \exp\left[-N^2 \frac{\sinh y_s}{y_s} \right]\left[1 + N^2 \frac{\sinh y_s}{y_s} \right]. \tag{62}$$

From equation (37)

$$Q_R(x, y_s) = N^2 \left\{ \frac{1}{2} A_2 x^2 - \frac{1}{24} A_4 x^4 + \frac{1}{6!} A_6 x^6 \left[1 - \frac{6!}{8!} \frac{A_8}{A_6} x^2 \right] \right.$$
$$\left. + \frac{1}{10!} A_{10} x^{10} \left[1 - \frac{10!}{12!} \frac{A_{12}}{A_{10}} x^2 \right] + \cdots \right\}. \tag{63}$$

From equations (46) and (63) it can be shown that for all y_s

$$Q_R(x, y_s) \geqq N^2 \left[\frac{1}{2} A_2 x^2 - \frac{1}{24} A_4 x^4 \right], \qquad 0 \leqq x \leqq x_1 < (56)^{\frac{1}{2}}. \tag{64}$$

We then have

$$| I_1 | \leqq 2 \int_0^{x_1} \exp\left\{ -N^2 \left[\frac{1}{2} A_2 x^2 - \frac{1}{24} A_4 x^4 \right] \right\} + 2 \int_0^{x_1} H(x, y_s)\, dx. \tag{65}$$

One can show that

$$e^t \leqq 1 + \frac{e^R - 1}{R}\, t, \qquad 0 \leqq t \leqq R. \tag{66}$$

Since we have

$$0 \leqq \frac{N^2}{24} A_4 x^4 \leqq \frac{25}{6} \frac{A_4}{A_2} \frac{1}{N^2 A_2}, \qquad 0 \leqq x \leqq x_1, \tag{67}$$

$$\begin{aligned} |I_1| &\leqq 2\int_0^{x_1} \exp\left[-N^2\tfrac{1}{2}A_2x^2\right]\left\{1 + \frac{\exp\left[\dfrac{25}{6}\dfrac{A_4}{A_2}\dfrac{1}{N^2A_2}\right] - 1}{x_1^4}\,x^4\right\} dx \\ &\quad + 2\int_0^{x_1} H(x, y_s)\,dx \qquad < 2\int_0^{\infty} \exp\left[-N^2\tfrac{1}{2}A_2x^2\right] \\ &\quad + 2\,\frac{\exp\left[\dfrac{25}{6}\dfrac{A_4}{A_2}\dfrac{1}{N^2A_2}\right] - 1}{x_1^4}\int_0^{\infty} x^4 \exp\left[-N^2\tfrac{1}{2}A_2x^2\right] dx \\ &\quad + 2\int_0^{x_1} H(x, y_s)\,dx. \end{aligned} \tag{68}$$

From equation (68) it can be shown that

$$|I_1| < \left(\frac{2\pi}{N^2A_2}\right)^{\frac{1}{2}}[1 + E_1], \tag{69}$$

where

$$\begin{aligned} E_1 = \frac{3}{100}&\left\{\exp\left[\frac{25}{6}\frac{A_4}{A_2}\frac{1}{N^2A_2}\right] - 1\right\} \\ &+ 2\left(\frac{5}{\pi}\right)^{\frac{1}{2}} \exp\left[-N^2\frac{\sinh y_s}{y_s}\right]\left[1 + N^2\frac{\sinh y_s}{y_s}\right]. \end{aligned} \tag{70}$$

Since we know that

$$-|p| \leqq \operatorname{Re} p \leqq |p|, \qquad p \text{ any arbitrary complex number}, \tag{71}$$

equation (69) gives an upper bound to Re I_1. Let us now find a lower bound to Re I_1.

From equations (29) and (50) we have

$$\begin{aligned} \operatorname{Re} I_1 &= 2\int_0^{x_1} \exp\left[-Q_R(x, y_s)\right] \cos\left[Q_I(x, y_s)\right] dx \\ &\quad - 2\operatorname{Re}\int_0^{x_1} \exp\left[-N^2\frac{\sinh y_s}{y_s} + j\lambda x\right]\left[1 + N^2\frac{\sin(x + jy_s)}{x + jy_s}\right] dx. \end{aligned} \tag{72}$$

As shown earlier in this paper

$$\begin{aligned} 2\operatorname{Re}\int_0^{x_1} &\exp\left[-N^2\frac{\sinh y_s}{y_s} + j\lambda x\right]\left[1 + N^2\frac{\sin(x + jy_s)}{x + jy_s}\right] dx \\ &\leqq 2x_1 \exp\left[-N^2\frac{\sinh y_s}{y_s}\right]\left\{1 + N^2\frac{\sinh y_s}{y_s}\right\}. \end{aligned} \tag{73}$$

One can also show that for z real

$$\cos z \geqq 1 - \frac{z^2}{2}, \qquad -\infty < z < \infty. \tag{74}$$

Using equation (74) we can write

$$\begin{aligned} 2 \int_0^{x_1} \exp\,[-Q_R(x, y_s)] \cos\,[Q_I(x, y_s)]\,dx \\ \geqq 2 \int_0^{x_1} \exp\,[-Q_R(x, y_s)] \left[1 - \frac{Q_I^2(x, y_s)}{2}\right] dx. \end{aligned} \tag{75}$$

Now from equations (37), and (38) we have

$$\begin{aligned} Q_R(x, y_s) = N^2 \bigg[\frac{1}{2} A_2 x^2 - \frac{1}{4!} A_4 x^4 \left(1 - \frac{4!}{6!} \frac{A_6}{A_4} x^2\right) \\ - \frac{1}{8!} A_8 x^8 \left(1 - \frac{8!}{10!} \frac{A_{10}}{A_8} x^2\right) - \cdots \bigg], \end{aligned} \tag{76}$$

and

$$\begin{aligned} Q_I(x, y_s) = N^2 \bigg[\frac{1}{6} A_3 x^3 - \frac{1}{5!} A_5 x^5 \left\{1 - \frac{5!}{7!} \frac{A_7}{A_5} x^2\right\} \\ - \frac{1}{9!} A_9 x^9 \left\{1 - \frac{9!}{11!} \frac{A_{11}}{A_9} x^2\right\} - \cdots \bigg]. \end{aligned} \tag{77}$$

From equations (76), and (77) one can show that

$$Q_R(x, y_s) \leqq \tfrac{1}{2} N^2 A_2 x^2, \qquad 0 \leqq x \leqq x_1 < (30)^{\frac{1}{2}}, \tag{78}$$

and

$$Q_I(x, y_s) \leqq \tfrac{1}{6} N^2 A_3 x^3, \qquad 0 \leqq x \leqq x_1 < (42)^{\frac{1}{2}}. \tag{79}$$

Equations (75), (78), and (79) yield

$$\begin{aligned} 2 \int_0^{x_1} \exp\,[-Q_R(x, y_s)] \cos\,[Q_I(x, y_s)]\,dx \\ > 2 \int_0^{x_1} \exp\,[-\tfrac{1}{2} N^2 A_2 x^2] \left[1 - \frac{N^4}{72} A_3^2 x^6\right] dx \\ = 2 \int_0^{\infty} \exp\,[-\tfrac{1}{2} N^2 A_2 x^2]\,dx - 2 \int_{x_1}^{\infty} \exp\,[-\tfrac{1}{2} N^2 A_2 x^2]\,dx \\ - \frac{N^4}{36} A_3^2 \int_0^{x_1} x^6 \exp\,[-\tfrac{1}{2} N^2 A_2 x^2]\,dx. \end{aligned} \tag{80}$$

Notice that

$$\int_0^{x_1} x^6 \exp\,[-\tfrac{1}{2}N^2A_2x^2]\,dx < \int_0^\infty x^6 \exp\,[-\tfrac{1}{2}N^2A_2x^2]\,dx$$

$$= \frac{1}{2}\left(\frac{2\pi}{N^2A_2}\right)^{\frac{1}{2}} \frac{15}{(N^2A_2)^3}. \qquad (81)$$

We can also show that

$$\int_{a^2}^\infty \exp\,(-p^2t^2)\,dt < \frac{\exp\,(-p^2a^4)}{2a^2p^2}, \qquad a^2p^2 \neq 0. \qquad (82)$$

We can, therefore, write

$$2\int_{x_1}^\infty \exp\,[-\tfrac{1}{2}N^2A_2x^2]\,dx < \left(\frac{2}{5N^2A_2}\right)^{\frac{1}{2}} e^{-5}. \qquad (83)$$

From equations (72)–(73), (75), (80)–(83) it can now be shown that

$$\text{Re } I_1 > \left(\frac{2\pi}{N^2A_2}\right)^{\frac{1}{2}} [1 - E_1'], \qquad (84)$$

where

$$E_1' = \frac{1}{(5\pi)^{\frac{1}{2}}} e^{-5} + \frac{5}{24}\left(\frac{A_3}{A_2}\right)^2 \frac{1}{N^2A_2} + 2\left(\frac{5}{\pi}\right)^{\frac{1}{2}}$$

$$\cdot \exp\left[-N^2 \frac{\sinh y_s}{y_s}\right]\left[1 + N^2 \frac{\sinh y_s}{y_s}\right]. \qquad (85)$$

We shall now obtain upper and lower bounds to Re I_R in equation (51). According to equation (36) let

$$I_R = I_2 + I_T \qquad (86)$$

where

$$I_2 = \int_{-\pi}^{-x_1} G(x, y_s)\,dx + \int_{x_1}^{\pi} G(x, y_s)\,dx, \qquad (87)$$

and

$$I_T = \int_{-\infty}^{-\pi} G(x, y_s)\,dx + \int_{\pi}^{\infty} G(x, y_s)\,dx. \qquad (88)$$

From equation (87) we have

$$I_2 \mid \leqq 2\int_{x_1}^{\pi} \mid G(x, y_s) \mid dx. \qquad (89)$$

Now it can be shown from equations (29) and (60) that

$$| G(x, y_s) | \leqq \exp [-Q_R(x, y_s)] + \exp \left[-N^2 \frac{\sinh y_s}{y_s} \right] \left[1 + N^2 \frac{\sinh y_s}{y_s} \right]. \tag{90}$$

From equations (89) and (90) we can write

$$| I_2 | \leqq 2 \int_{x_1}^{\pi} \exp [-Q_R(x, y_s)] \, dx + 2(\pi - x_1) \exp \left[-N^2 \frac{\sinh y_s}{y_s} \right] \left[1 + N^2 \frac{\sinh y_s}{y_s} \right]. \tag{91}$$

From equation (64)

$$Q_R(x, y_s) \geqq N^2 \left[\frac{1}{2} A_2 x^2 - \frac{1}{24} A_4 x^4 \right] \equiv N^2 v, \qquad x_1 \leqq x \leqq \pi \tag{92}$$

where

$$v = \frac{1}{2} A_2 x^2 - \frac{1}{24} A_4 x^4. \tag{93}$$

It can be shown (see Fig. 8) that v and dv/dx are positive for $x_1 \leqq x \leqq x_m = (6)^{\frac{1}{2}}(A_2/A_4)^{\frac{1}{2}} \geqq (6)^{\frac{1}{2}}$ and that dv/dx is a monotonically increasing function of x for $x_1 \leqq x \leqq x_n$ where

$$x_n = \sqrt{2} \left(\frac{A_2}{A_4} \right)^{\frac{1}{2}} \geqq \sqrt{2}. \tag{94}$$

We also know that $Q_R(x, y_s)$ is a monotonically increasing function of x for $x_1 \leqq x \leqq \pi$. Let us now assume that $x_1 < \sqrt{2}$.

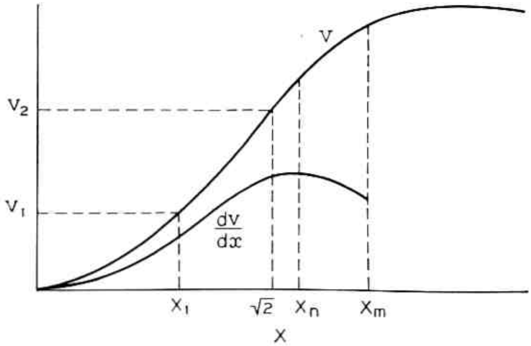


Fig. 8 — Functions v and $\frac{dv}{dx}$ appearing in equation (95).

Equations (92)–(94) therefore yield

$$2 \int_{x_1}^{\pi} \exp [-Q_R(x, y_s)] \, dx$$

$$< 2 \int_{v_1}^{v_2} \exp [-N^2 v] \frac{dx}{dv} \, dv + 2(\pi - \sqrt{2}) \exp [-Q_R(\sqrt{2}, y_s)], \tag{95}$$

where

$$v_1 = \frac{5}{N^2} \left[1 - \frac{5}{6} \frac{A_4}{A_2} \frac{1}{N^2 A_2} \right], \tag{96}$$

$$v_2 = A_2 - \frac{A_4}{6}, \tag{97}$$

and

$$0 \leqq v_1 \leqq v_2 . \tag{98}$$

Since we know that

$$0 \leqq \frac{dx}{dv} = \frac{1}{\dfrac{dv}{dx}} \leqq \frac{1}{\left[\dfrac{dv}{dx} \right]_{x=x_1}} = \frac{N^2 x_1}{10 \left[1 - \dfrac{5}{3} \dfrac{A_4}{A_2} \dfrac{1}{N^2 A_2} \right]}, \tag{99}$$

we can write

$$2 \int_{v_1}^{v_2} \exp [-N^2 v] \frac{dx}{dv} \, dv < \frac{N^2 x_1}{5 \left[1 - \dfrac{5}{3} \dfrac{A_4}{A_2} \dfrac{1}{N^2 A_2} \right]} \int_{v_1}^{v_2} \exp (-N^2 v) \, dv$$

$$< \frac{N^2 x_1}{5 \left[1 - \dfrac{5}{3} \dfrac{A_4}{A_2} \dfrac{1}{N^2 A_2} \right]} \int_{v_1}^{\infty} \exp (-N^2 v) \, dv. \tag{100}$$

Since it can be shown that

$$\int_{a^2}^{\infty} \exp [-p^2 t] \, dt = \frac{\exp [-p^2 a^2]}{p^2}, \qquad p^2 \neq 0, \tag{101}$$

from equations (91), (95) and (100), we have

$$| I_2 | < \left(\frac{2\pi}{N^2 A_2} \right)^{\frac{1}{2}} E_2 , \tag{102}$$

where

$$E_2 = \frac{1}{(5\pi)^{\frac{1}{2}}} \frac{1}{1 - \frac{5}{3}\frac{A_4}{A_2}\frac{1}{N^2 A_2}} \exp\left[-5\left(1 - \frac{5}{6}\frac{A_4}{A_2}\frac{1}{N^2 A_2}\right)\right]$$

$$+ 2(\pi - \sqrt{2})\left(\frac{N^2 A_2}{2\pi}\right)^{\frac{1}{2}} \exp\left[-Q_R(\sqrt{2}, y_s)\right]$$

$$+ 2(\pi - x_1)\left(\frac{N^2 A_2}{2\pi}\right)^{\frac{1}{2}} \exp\left[-N^2 \frac{\sinh y_s}{y_s}\right]\left[1 + N^2 \frac{\sinh y_s}{y_s}\right],$$

$$x_1 < \sqrt{2}. \qquad (103)$$

Similarly, if $x_1 > \sqrt{2}$, one can show that

$$E_2 = 2(\pi - x_1)\left(\frac{N^2 A_2}{2\pi}\right)^{\frac{1}{2}}\left\{\exp\left[-Q_R(x_1 , y_s)\right] + \exp\left[-N^2 \frac{\sinh y_s}{y_s}\right]\left[1 + N^2 \frac{\sinh y_s}{y_s}\right]\right\}. \qquad (104)$$

Let us now consider the range of integration $\pi < x < \infty$. For $x \gg y_s$,

$$\exp\left[-Q_R(x, y_s) + jQ_I(x, y_s)\right] \approx \exp\left[-\left(N^2 \frac{\sinh y_s}{y_s} - N^2 \cosh y_s \frac{\sin x}{x}\right) + j\left(\lambda x + N^2 \sinh y_s \frac{\cos x}{x}\right)\right], \qquad (105)$$

and

$$\exp\left[-N^2 \frac{\sinh y_s}{y_s} + j\lambda x\right] \cdot \left[1 + N^2 \frac{\sin (x + jy_s)}{x + jy_s}\right] \approx \exp\left[-N^2 \frac{\sinh y_s}{y_s} + j\lambda x\right] \cdot \left\{1 + N^2\left[\frac{\sin x}{x} \cosh y_s + j \frac{\cos x}{x} \sinh y_s\right]\right\}. \qquad (106)$$

Let us choose the point $x_2 + jy_s$ along the path of integration so that the amplitudes of the two terms in equations (105) and (106) differ by less than 10.5 percent $[(N^2 \cosh y_s)/x_2 \leqq 0.1]$and their relative angle departs from 180° by less than 0.1 radian $[(N^2 \sinh y_s)/x_2 \leqq 0.1]$.

Such a point $x_2 + jy_s$ is given by

$$x_2 = 10N^2 \cosh y_s \, . \tag{107}$$

We assume that $x_2 > \pi \geqq x_1$, or that

$$N^2 > \frac{\pi}{10} \approx 0.31416. \tag{108}$$

Since from equation (57) $N^2 \geqq 30/\pi^2$, this inequality is always satisfied.

We shall now write

$$I_T = I_3 + I_4 \, , \tag{109}$$

where

$$I_3 = \int_{-x_2}^{-\pi} G(x, y_s) \, dx + \int_{\pi}^{x_2} G(x, y_s) \, dx, \tag{110}$$

and

$$I_4 = \int_{-\infty}^{-x_2} G(x, y_s) \, dx + \int_{x_2}^{\infty} G(x, y_s) \, dx. \tag{111}$$

Noticing that

$$\frac{|\sin (x + jy_s)|}{|x + jy_s|} \leqq \frac{\cosh y_s}{(x^2 + y_s^2)^{\frac{1}{2}}} \leqq \frac{\cosh y_s}{x} \, , \tag{112}$$

we can show that

$$| I_3 | \leqq 2 \int_{\pi}^{x_2} \exp \left[-Q_R(x, y_s) \right] dx$$
$$+ \, 2(x_2 - \pi) \exp \left[-N^2 \frac{\sinh y_s}{y_s} \right] \left[1 + N^2 \cosh y_s \frac{\ln \left(\frac{x_2}{\pi} \right)}{x_2 - \pi} \right], \tag{113}$$

and

$$\int_{\pi}^{x_2} \exp \left[-Q_R(x, y_s) \right] dx \leqq \sum_{\ell=1}^{K} \int_{(2\ell-1)\pi}^{2\ell\pi} \exp \left[-Q_R(x, y_s) \right] dx$$
$$+ \sum_{\ell=1}^{K} \int_{2\ell\pi}^{(2\ell+1)\pi} \exp \left[-Q_R(x, y_s) \right] dx, \tag{114}$$

where K is an integer such that

$$(2K + 1)\pi > x_2 \geqq (2K - 1)\pi. \tag{115}$$

One can show that equation (115) is satisfied if

$$K = INT\left[\frac{x_2}{2\pi} + \frac{1}{2}\right]$$

or

$$K = INT\left[\frac{5N^2 \cosh y_s}{\pi} + \frac{1}{2}\right]. \tag{116}$$

We now have

$$0 < \frac{(n\pi)^2}{(n\pi)^2 + y_s^2} \leqq \frac{x^2}{x^2 + y_s^2} \leqq \frac{(n+1)^2\pi^2}{(n+1)^2\pi^2 + y_s^2},$$
$$n\pi \leqq x \leqq (n+1)\pi, \qquad n = 1, 2, 3, \cdots, \tag{117}$$

and

$$1 - \frac{y_s}{\tanh y_s}\frac{\sin x}{x} + y_s^2\frac{1 - \cos x}{x^2} \geqq 1,$$
$$(2\ell - 1)\pi \leqq x \leqq 2\ell\pi, \qquad \ell = 1, 2, 3, \cdots. \tag{118}$$

From equations (31), (117), and (118) we can now prove that

$$\begin{aligned}\sum_{\ell=1}^{K} \int_{(2\ell-1)\pi}^{2\ell\pi} & \exp\left[-Q_R(x, y_s)\right] dx \\ & \leqq \sum_{\ell=1}^{K} \pi \exp\left[-N^2 \frac{\sinh y_s}{y_s} \frac{1}{1 + \dfrac{y_s^2}{\pi^2(2\ell - 1)^2}}\right] \\ & = \pi \exp\left[-N^2 \frac{\sinh y_s}{y_s}\right] \sum_{\ell=1}^{K} \exp\left[\frac{N^2 y_s \sinh y_s}{\pi^2} \frac{1}{(2\ell - 1)^2 + \dfrac{y_s^2}{\pi^2}}\right].\end{aligned} \tag{119}$$

Further it can be shown that*

$$\begin{aligned}\sum_{\ell=1}^{K} \exp\left[\frac{N^2 y_s \sinh y_s}{\pi^2} \frac{1}{(2\ell - 1)^2 + y_s^2/\pi^2}\right] & \\ < \exp\left[\frac{N^2 y_s \sinh y_s}{\pi^2} \frac{1}{1 + y_s^2/\pi^2}\right] & \\ + (K - 1) \exp\left[\frac{N^2 y_s \sinh y_s}{\pi^2} \frac{1}{9 + y_s^2/\pi^2}\right]. &\end{aligned} \tag{120}$$

* The upper bound derived in equation (120) can be improved in various ways. Since this makes only a minor contribution to the total integral we shall be satisfied with this simple bound.

From equations (119) and (120) we have

$$\sum_{\ell=1}^{K} \int_{(2\ell-1)\pi}^{2\ell\pi} \exp\left[-Q_R(x, y_s)\right] dx$$
$$< \pi \exp\left[-N^2 \frac{\sinh y_s}{y_s}\right]\left\{\exp\left[\frac{N^2 y_s \sinh y_s}{\pi^2} \frac{1}{1 + (y_s/\pi)^2}\right]\right.$$
$$\left. + (K - 1) \exp\left[\frac{N^2 y_s \sinh y_s}{\pi^2} \frac{1}{9 + (y_s/\pi)^2}\right]\right\}. \tag{121}$$

For $2\ell\pi + \pi/2 \leqq x \leqq (2\ell + 1)\pi$, $\ell \geqq 1$, we can show that

$$1 - \frac{y_s}{\tanh y_s} \frac{\sin x}{x} + y_s^2 \frac{1 - \cos x}{x^2}$$
$$\geqq 1 - \frac{y_s}{\tanh y_s} \frac{2}{(4\ell + 1)\pi} \sin x + \frac{y_s^2}{(2\ell + 1)^2\pi^2} (1 - \cos x)$$
$$\equiv J_\ell(x). \tag{122}$$

In this range of x

$$\cos x \leqq 0, \tag{123}$$

$$\sin x \geqq 0, \tag{124}$$

and

$$\frac{\partial J_\ell}{\partial x} = -\frac{y_s}{\tanh y_s} \frac{2}{(4\ell + 1)\pi} \cos x + \frac{y_s^2}{(2\ell + 1)^2\pi^2} \sin x \geqq 0. \tag{125}$$

We, therefore, have

$$1 - \frac{y_s}{\tanh y_s} \frac{\sin x}{x} + y_s^2 \frac{1 - \cos x}{x^2}$$
$$\geqq J_\ell(x) \geqq J_\ell\left(2\ell\pi + \frac{\pi}{2}\right) \equiv V_\ell(y_s), \tag{126}$$

where

$$V_\ell(y_s) = 1 - \frac{y_s}{\tanh y_s} \frac{2}{(4\ell + 1)\pi} + \frac{y_s^2}{(2\ell + 1)^2\pi^2}. \tag{127}$$

It can be shown that

$$V_\ell(0) = 1 - \frac{2}{(4\ell + 1)\pi}, \tag{128}$$

that $V_\ell(y_s)$ reaches its minimum at

$$y_s \approx \left[1 + \frac{4\ell^2}{4\ell + 1}\right]\pi, \tag{129}$$

and

$$[V_\ell(y_s)]_{\min} \approx 1 - \frac{1}{4}\left[1 + \frac{1}{4\ell + 1}\right]^2 > 0, \qquad \ell \geqq 1. \tag{130}$$

Now for $2\ell\pi \leqq x \leqq 2\ell\pi + \pi/2$, it can also be shown that

$$1 - \frac{y_s}{\tanh y_s}\frac{\sin x}{x} + y_s^2\frac{1 - \cos x}{x^2} \geqq 1 - \frac{y_s}{\tanh y_s}\frac{\sin x}{2\ell\pi}$$
$$+ \frac{4y_s^2}{(4\ell + 1)^2\pi^2}(1 - \cos x) \equiv L_\ell(x). \tag{131}$$

One can prove that $L_\ell(x)$ reaches its minimum at

$$x = 2\ell\pi + \tan^{-1}\left[\frac{(4\ell + 1)^2}{8\ell}\frac{\pi}{y_s \tanh y_s}\right], \tag{132}$$

and

$$[L_\ell(x)]_{\min} \equiv U_\ell(y_s) = 1 + \frac{4y_s^2}{(4\ell + 1)^2\pi^2}$$
$$- \left[\frac{16y_s^4}{(4\ell + 1)^4\pi^4} + \frac{y_s^2}{4\ell^2\pi^2 \tanh^2 y_s}\right]^{\frac{1}{2}} \tag{133}$$

Next we can show that

$$U_\ell(0) = 1 - \frac{1}{2\ell\pi} < 1 - \frac{2}{(4\ell + 1)\pi} = V_\ell(0), \tag{134}$$

that $U_\ell(y_s)$ is a monotonically decreasing function of y_s (see Fig. 9), and

$$\lim_{y_s \to \infty} U_\ell(y_s) = 1 - \frac{(4\ell + 1)^2}{32\ell^2} > 0, \qquad \ell \geqq 1. \tag{135}$$

It can also be proved by numerical methods (see Fig. 9) that

$$U_\ell(y_s) \geqq U_1(y_s),\ y_s \geqq 0, \qquad \ell \geqq 1, \tag{136}$$

and

$$V_\ell(y_s) > U_1(y_s),\ y_s \geqq 0, \qquad \ell \geqq 1. \tag{137}$$

We therefore conclude that

$$1 - \frac{y_s}{\tanh y_s}\frac{\sin x}{x} + y_s^2 \frac{1 - \cos x}{x^2} \geqq U_1(y_s) = 1 + \frac{4y_s^2}{25\pi^2} - \left[\frac{16y_s^4}{625\pi^4} + \frac{y_s^2}{4\pi^2 \tanh^2 y_s}\right]^{\frac{1}{2}}, \qquad \ell \geqq 1, \qquad y_s \geqq 0,$$

$$2\ell\pi \leqq x \leqq (2\ell + 1)\pi. \tag{138}$$

Equations (31), (117), and (138) show that

$$Q_R(x, y_s) \geqq N^2 \frac{\sinh y_s}{y_s} U_1(y_s) \frac{(2\ell)^2\pi^2}{(2\ell)^2\pi^2 + y_s^2},$$

$$2\ell\pi \leqq x \leqq (2\ell + 1)\pi. \tag{139}$$

From equation (139) we can now write

$$\sum_{\ell=1}^{K} \int_{2\ell\pi}^{(2\ell+1)\pi} \exp\left[-Q_R(x, y_s)\right] dx \leqq \sum_{\ell=1}^{K} \pi \exp\left[-N^2 \frac{\sinh y_s}{y_s} U_1(y_s) \frac{(2\ell)^2\pi^2}{(2\ell)^2\pi^2 + y_s^2}\right]$$
$$= \pi \exp\left[-N^2 \frac{\sinh y_s}{y_s} U_1(y_s)\right] \cdot \sum_{\ell=1}^{K} \exp\left[\frac{N^2 y_s \sinh y_s}{\pi^2} U_1(y_s) \frac{1}{(2\ell)^2 + (y_s/\pi)^2}\right]. \tag{140}$$

It can be shown that

$$\sum_{\ell=1}^{K} \exp\left[\frac{N^2 y_s \sinh y_s}{\pi^2} U_1(y_s) \frac{1}{(2\ell)^2 + (y_s/\pi)^2}\right] < \exp\left[\frac{N^2 y_s \sinh y_s}{\pi^2} U_1(y_s) \frac{1}{4 + (y_s/\pi)^2}\right] + (K - 1) \exp\left[\frac{N^2 y_s \sinh y_s}{\pi^2} U_1(y_s) \frac{1}{16 + (y_s/\pi)^2}\right]. \tag{141}$$

Equations (140) and (141) yield

$$\sum_{\ell=1}^{K} \int_{2\ell\pi}^{(2\ell+1)\pi} \exp\left[-Q_R(x, y_s)\right] dx < \pi \exp\left[-N^2 \frac{\sinh y_s}{y_s} U_1(y_s)\right] \cdot \left\{\exp\left[N^2 \frac{y_s \sinh y_s}{\pi^2} U_1(y_s) \frac{1}{4 + (y_s/\pi)^2}\right] + (K - 1) \exp\left[N^2 \frac{y_s \sinh y_s}{\pi^2} U_1(y_s) \frac{1}{16 + (y_s/\pi)^2}\right]\right\}. \tag{142}$$

From equations (113), (114), (121), and (142) we can write

$$| I_3 | < \left(\frac{2\pi}{N^2 A_2}\right)^{\frac{1}{2}} E_3 \,, \tag{143}$$

where

$$\begin{aligned} E_3 = \left(\frac{N^2 A_2}{2\pi}\right)^{\frac{1}{2}} \Bigg\{ & 2\pi \exp\left[-N^2 \frac{\sinh y_s}{y_s}\right] \\ & \cdot \left[\exp\left(\frac{N^2 y_s \sinh y_s}{\pi^2} \frac{1}{1 + (y_s/\pi)^2}\right)\right. \\ & \left. + (K - 1) \exp\left(\frac{N^2 y_s \sinh y_s}{\pi^2} \frac{1}{9 + (y_s/\pi)^2}\right)\right] \\ & + 2\pi \exp\left[-N^2 \frac{\sinh y_s}{y_s} U_1(y_s)\right] \\ & \cdot \left\{\exp\left[N^2 \frac{y_s \sinh y_s}{\pi^2} U_1(y_s) \frac{1}{4 + (y_s/\pi)^2}\right]\right. \\ & \left. + (K - 1) \exp\left[N^2 \frac{y_s \sinh y_s}{\pi^2} U_1(y_s) \frac{1}{16 + (y_s/\pi)^2}\right]\right\} \\ & + 2(x_2 - \pi) \exp\left[-N^2 \frac{\sinh y_s}{y_s}\right]\left[1 + N^2 \cosh y_s \frac{\ln\left(\frac{x_2}{\pi}\right)}{x_2 - \pi}\right]\Bigg\} \cdot \end{aligned} \tag{144}$$

Finally, from equation (111) we have

$$| I_4 | \leqq 2 \int_{x_2}^{\infty} | G(x, y_s) | \, dx. \tag{145}$$

Now from equations (27)–(28)

$$\begin{aligned} | G(x, y_s) | = {} & \exp\left[-N^2 \frac{\sinh y_s}{y_s}\right] \\ & \cdot \left| \exp\left[N^2 \frac{\sin (x + jy_s)}{x + jy_s}\right] - \left[1 + N^2 \frac{\sin (x + jy_s)}{x + jy_s}\right] \right| \cdot \end{aligned} \tag{146}$$

If z is a complex variable, it can be shown that

$$| \exp (z) - 1 - z | \leqq \frac{| z |^2}{2} \exp | z |. \tag{147}$$

From equations (112), (145)–(147), we can write

$$\begin{aligned}| I_4 | &\leqq \exp\left[-N^2\frac{\sinh y_s}{y_s}\right]N^4 \cosh^2 y_s \\ &\quad \cdot\int_{x_2}^{\infty}\frac{1}{x^2+y_s^2}\exp\left[\frac{N^2\cosh y_s}{(x^2+y_s^2)^{\frac{1}{2}}}\right]dx \\ &= \frac{N^4\cosh^2 y_s}{y_s}\exp\left[-N^2\frac{\sinh y_s}{y_s}\right] \\ &\quad \cdot\int_0^{N^2\cosh y_s/(x_2{}^2+y_s{}^2)^{\frac{1}{2}}}\frac{\exp(t)}{\left(\dfrac{N^4\cosh^2 y_s}{y_s^2}-t^2\right)^{\frac{1}{2}}}\,dt \\ &= \frac{N^4\cosh^2 y_s}{y_s}\exp\left[-N^2\frac{\sinh y_s}{y_s}\right] \\ &\quad \cdot\int_0^{\sin^{-1} y_s/(x_2{}^2+y_s{}^2)^{\frac{1}{2}}}\exp\left[\frac{N^2\cosh y_s}{y_s}\sin\theta\right]d\theta. \end{aligned} \tag{148}$$

Since

$$0 \leqq \sin\theta \leqq \theta, \qquad 0 \leqq \theta \leqq \sin^{-1}\frac{y_s}{(x_2^2+y_2^2)^{\frac{1}{2}}} < \frac{\pi}{2}, \tag{149}$$

we can show from equation (148) that

$$\begin{aligned}| I_4 | &< \frac{N^4\cosh^2 y_s}{y_s}\exp\left[-N^2\frac{\sinh y_s}{y_s}\right] \\ &\quad \cdot\int_0^{\sin^{-1} y_s/(x_2{}^2+y_s{}^2)^{\frac{1}{2}}}\exp\left[\frac{N^2\cosh y_s}{y_s}\,\theta\right]d\theta \\ &= N^2\cosh y_s\exp\left[-N^2\frac{\sinh y_s}{y_s}\right] \\ &\quad \cdot\left\{\exp\left[N^2\frac{\cosh y_s}{y_s}\sin^{-1}\frac{y_s}{(x_2^2+y_s^2)^{\frac{1}{2}}}\right]-1\right\}. \end{aligned} \tag{150}$$

Now we have

$$0 \leqq \sin^{-1}\sigma \leqq \frac{\pi}{2}\sigma, \qquad 0 \leqq \sigma \leqq 1. \tag{151}$$

From equations (149)–(151), we can write

$$\begin{aligned}| I_4 | &< N^2\cosh y_s\exp\left[-N^2\frac{\sinh y_s}{y_s}\right] \\ &\cdot\left\{\exp\left[\frac{\pi}{2}N^2\frac{\cosh y_s}{(100N^4\cosh^2 y_s+y_s^2)^{\frac{1}{2}}}\right]-1\right\} < N^2\cosh y_s\end{aligned}$$

$$\cdot \exp\left[-N^2 \frac{\sinh y_s}{y_s}\right]\left[\exp\left(\frac{\pi}{20}\right) - 1\right] = \left(\frac{2\pi}{N^2 A_2}\right)^{\frac{1}{2}} E_4 , \tag{152}$$

where

$$E_4 = \left(\frac{N^2 A_2}{2\pi}\right)^{\frac{1}{2}} N^2 \cosh y_s \exp\left[-N^2 \frac{\sinh y_s}{y_s}\right]\left[\exp\left(\frac{\pi}{20}\right) - 1\right] \cdot \tag{153}$$

From equations (27), (49), (69), (71), (84), (86), (102), (109), (143), and (152) we can write the following bounds for Re I:

$$\left(\frac{2\pi}{N^2 A_2}\right)^{\frac{1}{2}} \{1 - E_1' - E_2 - E_3 - E_4\}$$

$$< \operatorname{Re} I < \left(\frac{2\pi}{N^2 A_2}\right)^{\frac{1}{2}} [1 + E_1 + E_2 + E_3 + E_4]. \tag{154}$$

It has been shown that

$$A_2 = \frac{\sinh y_s}{y_s} - \frac{2}{y_s} \frac{f}{N^2 W} \cdot \tag{155}$$

V. UPPER AND LOWER BOUNDS TO $S_V(f)$

We have shown in the previous section that

$$S_V(f) = \exp(-N^2)\left\{\delta(f) + \frac{N^2}{2W}[u_{-1}(f + W) - u_{-1}(f - W)]\right\}$$

$$+ \frac{1}{2\pi W} \exp\left\{-2N^2\left[\cosh^2 \frac{y_s}{2} - \frac{\sinh y_s}{y_s}\right]\right\}\mu \tag{156}$$

where

$$\left(\frac{2\pi}{N^2 A_2}\right)^{\frac{1}{2}} \{1 - E_1' - E_2 - E_3 - E_4\}$$

$$< \mu < \left(\frac{2\pi}{N^2 A_2}\right)^{\frac{1}{2}} [1 + E_1 + E_2 + E_3 + E_4], \tag{157}$$

$$\frac{\cosh y_s}{y_s} - \frac{\sinh y_s}{y_s^2} = \frac{f}{N^2 W} , \tag{25}$$

$$A_2 = \frac{\sinh y_s}{y_s} - \frac{2}{y_s} \frac{f}{N^2 W} \cdot \tag{155}$$

Parameter	*Equation*
E_1	70
E_1'	85
E_2	103 or 104
E_3	144
E_4	153

For $N^2 = 10$ and 25 we plot, in Figs. 10–16, E_1, E_1', E_2, E_3, and E_4. Notice that E_1, E_1', E_2, E_3, and E_4 appearing in these bounds are all very small compared to unity so long as the modulation index is moderately high, and that E_1, E_3, and E_4 are monotonically decreasing functions of f and N^2. Also notice that E_1' and E_2 may first increase (see Figs. 11, 12) with y_s (or f), reach their maxima and then decrease with y_s.* It can be shown that these maxima are all very small compared to unity for all N^2 which are even moderately high.

For all f, we can then write

$$\left(\frac{2\pi}{N^2 A_2}\right)^{\frac{1}{2}}(1 - C) < \mu < \left(\frac{2\pi}{N^2 A_2}\right)^{\frac{1}{2}}(1 + D), \tag{158}$$

where

$$C = E_1' + E_2 + E_3 + E_4 , \tag{159}$$

and

$$D = E_1 + E_2 + E_3 + E_4 . \tag{160}$$

From Figs. 10–16 and expressions for C and D, we can show that C and D are both small ($<2\%$) compared to unity for $N^2 > 25$ and for all f. Hence we deduce that

$$\mu \approx \left(\frac{2\pi}{N^2 A_2}\right)^{\frac{1}{2}}, \tag{161}$$

and that the fractional error in this approximation is very much less than unity ($<2\%$).

For $N^2 = 10$ and 25 the spectral density $S_V(f)$ and the fractional error C and D obtained from equations (158)–(161) are plotted in Figs. 17–20. From these figures notice that C and D are less than 10 percent for $N^2 > 10$,† and that

$$C < 2\%, \quad \text{for} \quad N^2 \geqq 25, \tag{162}$$

$$D < 2\%, \quad \text{for} \quad N^2 \geqq 25, \tag{163}$$

proving the assertion made earlier in this paper.

For $N^2 = 6$ the spectral density obtained from equations (158), and (161) is given in Fig. 21; the percentage error between this spectral

* One of the terms in E_1' is independent of f and N^2.

† By modifying the contour of integration we also have been able to show that C and D are less than 8% for $N^2 \geqq 10$. Since this modified contour leads to unnecessary complications, we have not given that modified contour analysis in this paper.

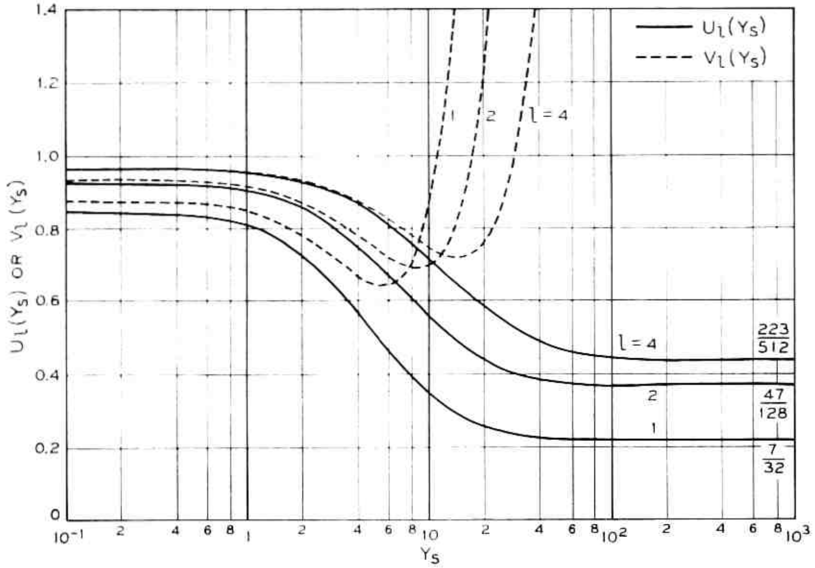


Fig. 9 — Functions $U_\ell(y_s)$ and $V_\ell(y_s)$. It can be observed that $V_\ell(y_s) > U_1(y_s) > 0$, $\ell \geqq 1$.

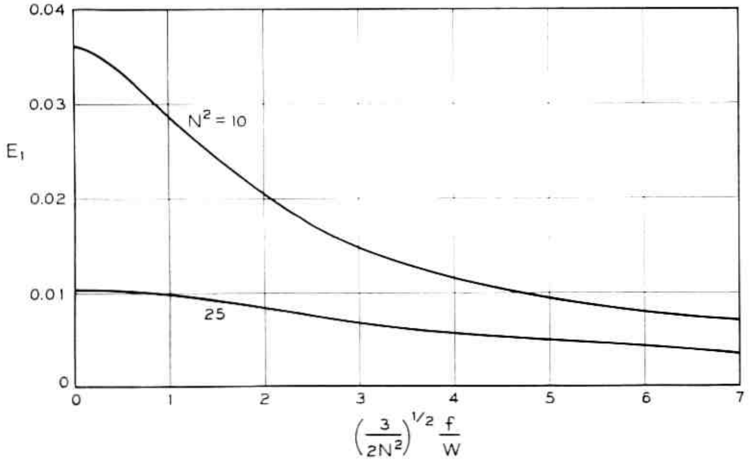


Fig. 10 — Parameter E_1 as a function of $\left(\frac{3}{2N^2}\right)^{\frac{1}{2}} \frac{f}{W}$.

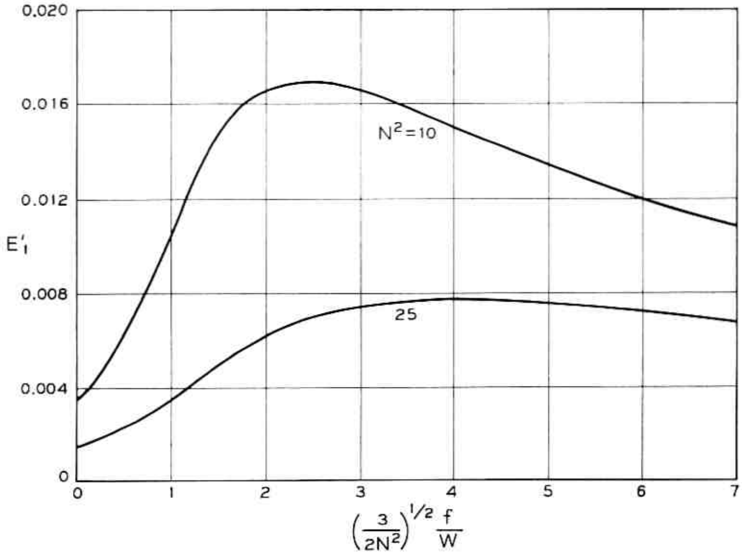


Fig. 11 — Parameter E_1' as a function of $\left(\frac{3}{2N^2}\right)^{\frac{1}{2}}\frac{f}{W}$.

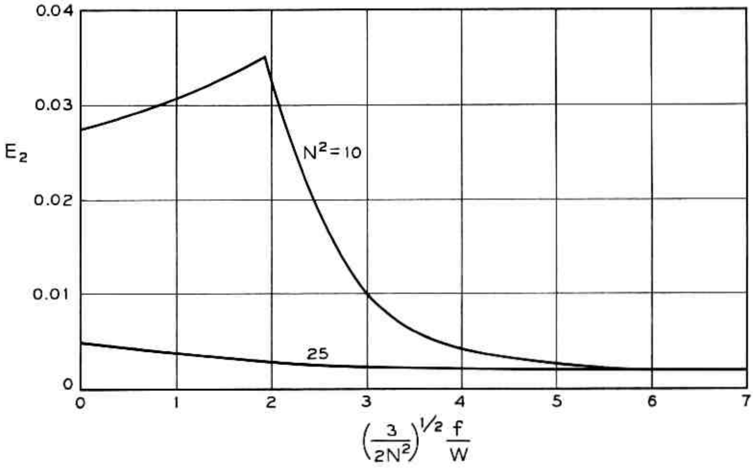


Fig. 12 — Parameter E_2 as a function of $\left(\frac{3}{2N^2}\right)^{\frac{1}{2}}\frac{f}{W}$.

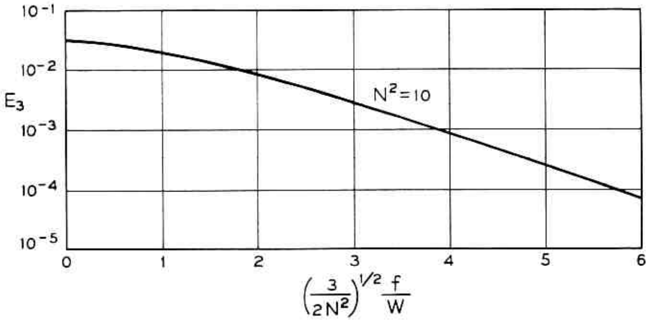


Fig. 13 — Parameter E_3 as a function of $\left[\frac{3}{2N^2}\right]^{\frac{1}{2}}\frac{f}{W}$, with $N^2 = 10$.

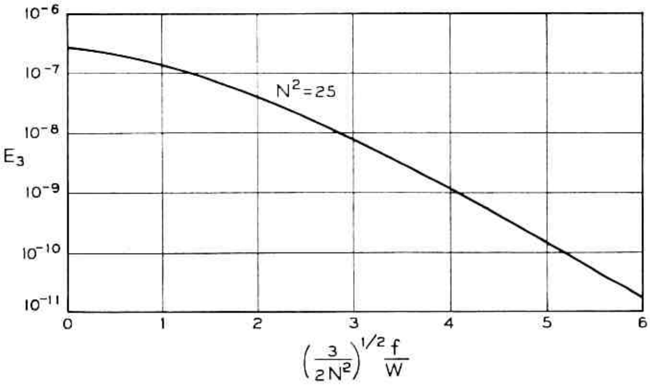


Fig. 14 — Parameter E_3 as a function of $\left[\frac{3}{2N^2}\right]^{\frac{1}{2}}\frac{f}{W}$, with $N^2 = 25$.

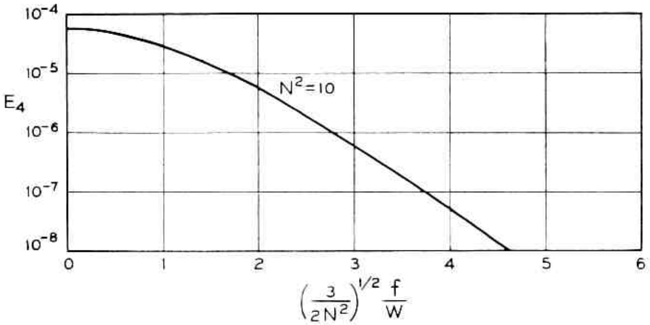


Fig. 15 — Parameter E_4 as a function of $\left[\frac{3}{2N^2}\right]^{\frac{1}{2}}\frac{f}{W}$, with $N^2 = 10$.

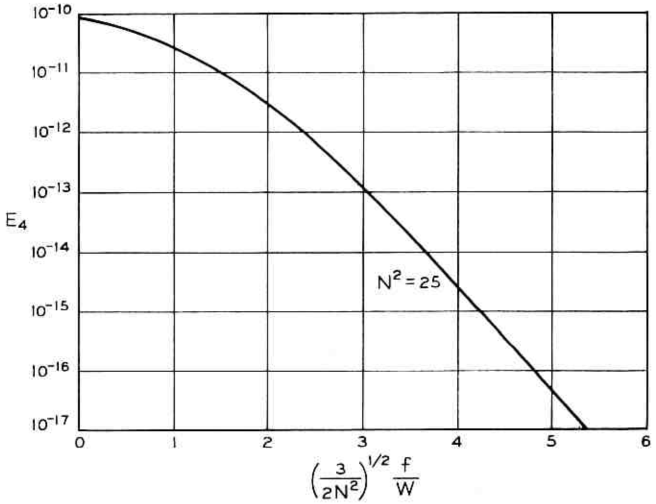


Fig. 16—Parameter E_4 as a function of $\left[\frac{3}{2N^2}\right]^{\frac{1}{2}}\frac{f}{W}$, with $N^2 = 25$.

density and that obtained from equation (19) has been plotted in Fig. 22 (for a set of values of f/W). The scatter diagram in Fig. 22 indicates that the spectral densities obtained from the two methods agree very closely and that the saddle-point approximation error is not related in a simple way to the truncation error (it does not seem possible to draw a smooth curve through the points shown in Fig. 22).

For all practical purposes, including interference calculations, estimation of the spectrum to such an accuracy is almost always sufficient. It can therefore be said that the saddle-point approximation given by equations (25), (155), (158), and (161) is a good approximation to $S_V(f)$ as long as the modulation index is even moderately high ($N^2 > 10$). The spectrum can be estimated by this method for all values of f even when it is millions of decibels smaller than the continuous part of the spectrum at $f = 0$.

Now compare the spectrum obtained from the quasistatic approximation* to that obtained from saddle-point approximation. For this purpose, the spectra obtained from equation (21) for $N^2 = 10$ and 25 are plotted in Figs. 17, and 19. We see that the spectra obtained from the quasistatic approximation agree very closely with those obtained from the saddle-point approximation for low frequencies, but that the quasi-

* See equation (21).

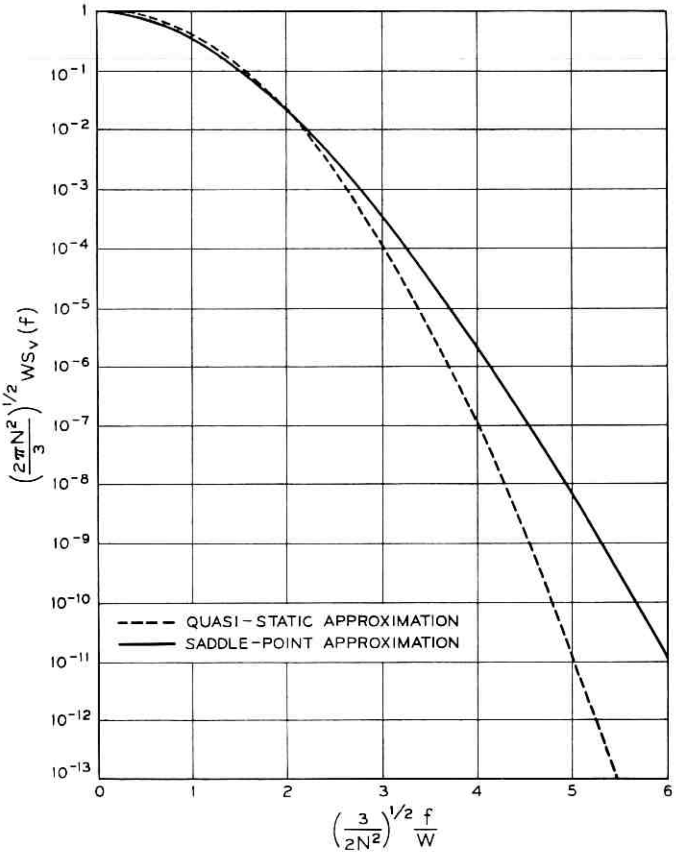


Fig. 17 — Spectral density of an angle-modulated wave, with gaussian phase modulation with a rectangular spectrum. $N = (10)^{1/2} \approx 3.162$ radians, rms phase deviation.

static approximation to $S_V(f)$ is too small for large f.* In fact for $N^2 = 10$ the quasistatic approximation is 30 dB too small for $f/W \approx 13.5$. We have therefore shown that the quasistatic approximation to the spectrum cannot be used in any interference calculations or in any calculations where the behavior of the spectrum on the tails is of importance.† The saddle-point approximation can be used at all frequencies as long as N^2 is moderately high.

* For small f (or small y_s) it can easily be shown that the saddle-point approximation reduces to the quasistatic approximation.

† The higher the rms phase deviation, the further out will the low-frequency (quasistatic) approximation be valid,

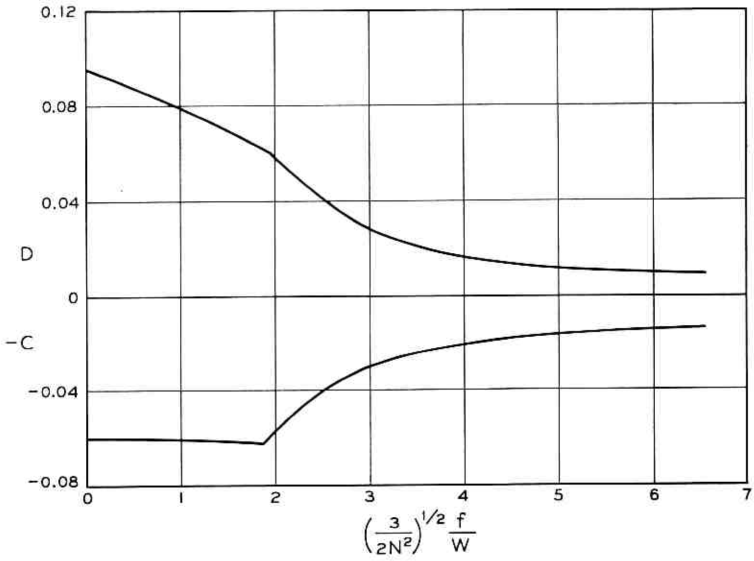


Fig. 18 — Bounds on fractional error in saddle-point approximation to the spectrum. $N^2 = 10$.

VI. RESULTS AND CONCLUSIONS

A simple method (called the saddle-point method) has been presented in this paper to estimate the spectrum of a sinusoidal carrier phase modulated by gaussian noise having a rectangular power spectrum.

This method gives upper and lower bounds to the spectrum and shows that these bounds are very close for all f and for all moderately high phase deviations. We also show that the fractional error in the saddle-point approximation is less than 2 percent for $N^2 \geqq 25$ and for all f.

The calculation of the spectrum by the saddle-point method is rather simple. For a given value of f, N^2, and W, we calculate y_s from equation (25) and A_2 from equation (155). The spectrum $S_V(f)$ is then calculated from equations (156) and (161).

We have also shown in this paper that the quasistatic approximation to $S_V(f)$ is only good at low frequencies, and that for large f the results obtained from that approximation are too small.

APPENDIX

It can be shown (see Ref. 7, p. 114) that

$$S_V(f) = \frac{1}{2\pi W} \int_{-\infty}^{\infty} \exp\left[-\frac{N^2}{2W} \int_{-W}^{W} (1 - \cos 2\pi\mu\tau)\, d\mu \right] \cdot \exp\left[-j2\pi f\tau\right] d\tau, \qquad (164)$$

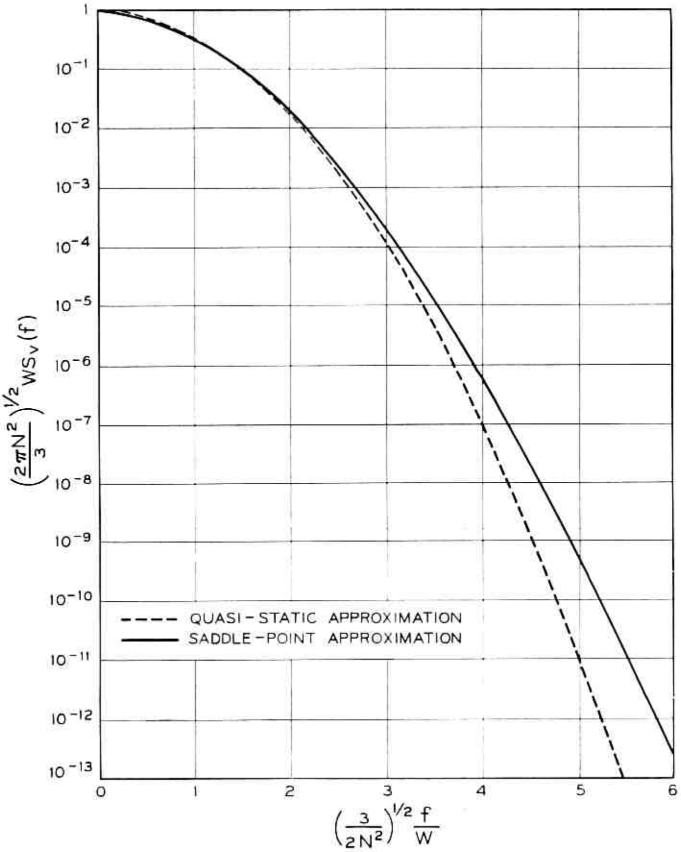


Fig. 19 — Spectral density of an angle-modulated wave, with gaussian phase modulation with a rectangular spectrum. $N = 5$ radians, rms phase deviation.

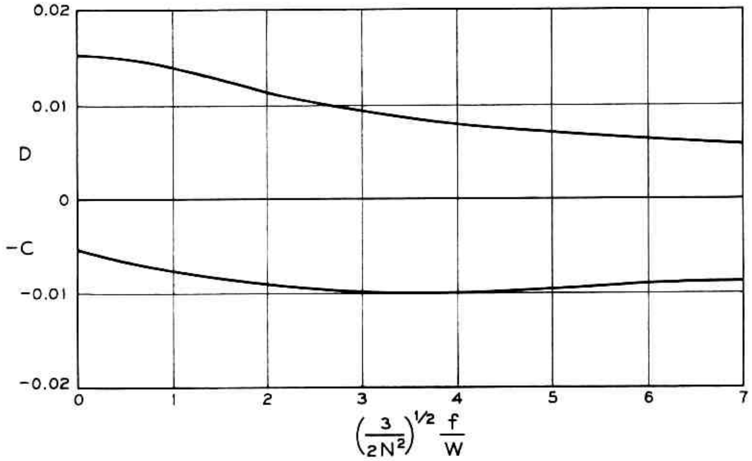


Fig. 20 — Bounds on fractional error in saddle-point approximation to the spectrum. $N^2 = 25$.

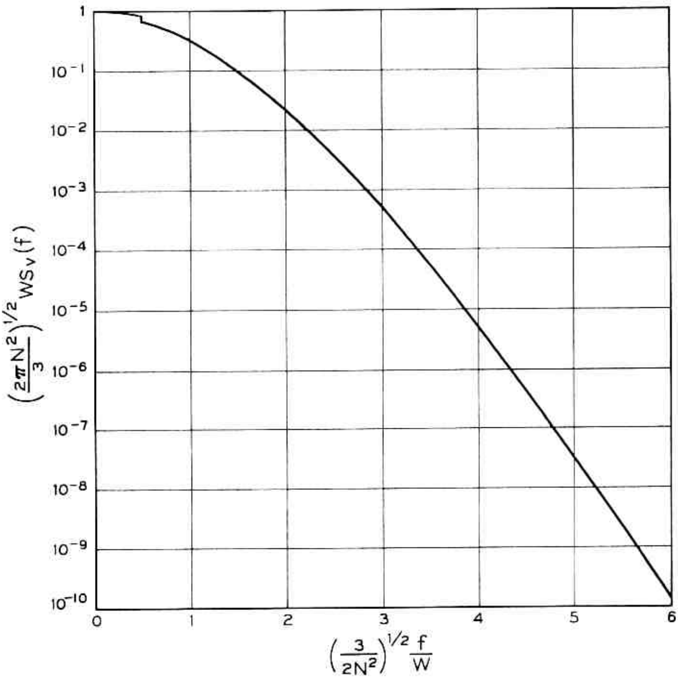


Fig. 21 — Spectral density of an angle-modulated wave, with gaussian phase modulation with a rectangular spectrum. $N = (6)^{1/2} \approx 2.449$ radians, rms phase deviation.

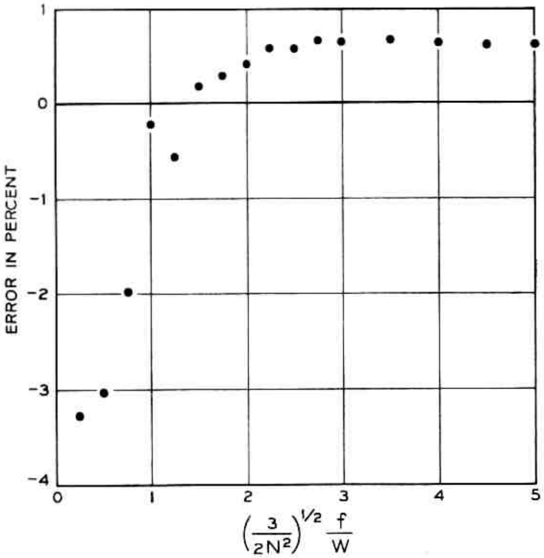


Fig. 22 — Percentage error between the spectral densities obtained from saddle-point approximation and that obtained from equation (19). It does not seem possible to draw a smooth curve through the points shown in this figure. We have, therefore, shown the error as a scatter diagram.

or

$$S(f) = \int_{-\infty}^{\infty} \left\{ \exp\left[-\frac{N^2}{2W} \int_{-W}^{W} \left(1 - \cos\frac{\mu}{W} p\right) d\mu \right] - \exp\left[-N^2\right] \left[1 + N^2 \frac{\sin p}{p} \right] \right\} e^{i\lambda p}\, dp. \tag{165}$$

From equation (165), it can be shown that

$$I = \exp\left[-N^2 \frac{\sinh y_s}{y_s} \right] \cdot \int_{-\infty}^{\infty} \left\{ \exp\left[\frac{N^2}{2W} \int_{-W}^{W} \cos\frac{\mu}{W}(x + jy_s)\, d\mu \right] - \left[1 + N^2 \frac{\sin(x + jy_s)}{x + jy_s} \right] \right\} e^{i\lambda x}\, dx, \tag{166}$$

and

$$Q_R(x, y_s) = N^2 \frac{\sinh y_s}{y_s} - \operatorname{Re} \frac{N^2}{2W} \int_{-W}^{W} \cos\frac{\mu}{W}(x + jy_s)\, d\mu, \tag{167}$$

or

$$Q_R(x, y_s) = N^2 \frac{\sinh y_s}{y_s} - \frac{N^2}{W} \int_0^W \cos \frac{\mu}{W} x \cosh \frac{\mu}{W} y_s \, d\mu. \tag{168}$$

Equation (168) yields

$$\frac{\partial Q_R(x, y_s)}{\partial x} = \frac{N^2}{W} \int_0^W \frac{\mu}{W} \sin \frac{\mu}{W} x \cosh \frac{\mu}{W} y_s \, d\mu. \tag{169}$$

For $0 \leqq \mu x/W \leqq \pi$,

$$\sin \frac{\mu}{W} x \geqq 0. \tag{170}$$

For $y_s \geqq 0$, and $0 \leqq \mu \leqq W$,

$$0 \leqq x \leqq \pi, \tag{171}$$

$$\frac{\mu}{W} \sin \frac{\mu}{W} x \cosh \frac{\mu}{W} y_s \geqq 0, \tag{172}$$

and from (169),

$$\frac{\partial Q_R(x, y_s)}{\partial x} \geqq 0, \qquad 0 \leqq x \leqq \pi. \tag{173}$$

From equation (173) we then conclude that $Q_R(x, y_s)$ is a monotonically increasing function of x for $0 \leqq x \leqq \pi$.

REFERENCES

1. Bennett, W. R., Curtis, H. E., and Rice, S. O., "Interchannel Interference in FM and PM Systems Under Noise Loading Conditions," B.S.T.J., *34,* No. 3 (May 1955), pp. 601–636.
2. Middleton, D., *An Introduction to Statistical Communication Theory,* New York: McGraw-Hill, Inc., 1960, pp. 599–635.
3. Blachman, N., "Limiting Frequency-Modulation Spectra," Information and Control, *1,* No. 3 (September 1957), pp. 26–37.
4. Stewart, J. L., "The Power Spectrum of a Carrier Frequency Modulated by Gaussian Noise," Proc. IEEE, *42,* No. 9 (October 1954), pp. 1539–1542.
5. Abramson, N., "Bandwidth and Spectra of Phase-and-Frequency-Modulated Waves," IEEE Trans. on Communication Systems, *CS-11,* No. 4 (December 1963), pp. 407–414.
6. Ferris, C. C., "Spectral Characteristics of FDM-FM Signals," IEEE Trans. on Communication Technology, *CT-16,* No. 2 (April 1968), pp. 233–238.
7. Rowe, H. E., *Signals and Noise in Communication Systems,"* Princeton, N. J.: D. Van Nostrand Co., Inc., 1965, pp. 98–203.
8. Gilbert, E. N., "Power Spectra of Some Random Frequency-Modulated and Phase-Modulated Waves," unpublished technical memorandum, August 1953.
9. Tillotson, L. C., and Ruthroff, C. L., "The Next Generation of Short Haul Radio Systems," unpublished technical memorandum, April 1965.

10. Erdelyi, A., *Asymptotic Expansions,* New York: Dover Publications, Inc., 1956, pp. 26–57.
11. Morse, P. M., and Feshbach, H., *Methods of Theoretical Physics,* New York: McGraw-Hill, 1953, pp. 437–443.
12. Jeffreys, H., *Asymptotic Approximations,* London, England: Oxford University Press, 1962, pp. 10–50.
13. Erdelyi, A., and others, *Tables of Integral Transforms,* New York: McGraw-Hill, 1954, pp. 18–20.
14. Lewin, L., "Interference in Multi-Channel Circuits," Wireless Engineer, *27* (December 1950), pp. 294–304.

Contributors to This Issue

Robert E. Bogner, B.E., 1956, University of Adelaide (Australia); M.E., 1959, University of Adelaide; Postmaster-General's Department (Australia) Research Laboratories, 1957–61; Lecturer and Senior Lecturer, University of Queensland (Australia), 1961–67; Lecturer in Electrical Engineering, Imperial College of Science and Technology (London), 1967—. Mr. Bogner has been involved in a variety of studies in speech processing, simulation of communication systems including radio wave scattering phenomena, and modulation techniques. The work on digital filters was a byproduct of his researches at Bell Telephone Laboratories where he was a summer employee in 1968, working on speech communication. Member, IEE.

Edwin L. Chinnock, Stevens Institute of Technology; Bell Telephone Laboratories 1939—. Mr. Chinnock has worked on microwave components, microwave radio relay, and helix waveguide fabrication. He is presently working on optical waveguide components.

Robert B. Cooper, B.S., 1961, Stevens Institute of Technology; M.S., 1962, and Ph.D., 1968, University of Pennsylvania; Bell Telephone Laboratories, 1961—. Mr. Cooper has worked on a variety of problems concerned with applications of probability theory to the analysis of telephone systems. He teaches the GSP course, Probability Applied to Traffic Engineering, and is currently writing a set of notes for this course. Member, Tau Beta Pi.

Herbert E. Earl, Jr., Bell Telephone Laboratories 1940—. He was engaged in work on pyrolitic film resistors, ferrites, and ferrimagnetic resonances. More recently he has worked on optical transmission techniques.

P. M. Ebert, B.S., 1958, University of Wisconsin; S.M., 1962, Sc.D., 1965, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1965—. Mr. Ebert has worked on problems in communications and information theory. Member, IEEE.

David D. Falconer, B.A.Sc., 1962, University of Toronto; S.M., 1963, and Ph.D., 1967, Massachusetts Institute of Technology; postdoctoral research fellowship, Royal Institute of Technology, Stockholm, Sweden, 1966–67; Bell Telephone Laboratories, 1967—. Mr. Falconer is concerned with the application of communication theory and error control techniques to data communications. Member, IEEE, Sigma Xi, Tau Beta Pi.

Detlef Gloge, Dipl. Ing., 1961, D.E.E., 1964, Braunschweig Technische Hochschule (Germany); research staff, Braunschweig Technische Hochschule, 1961–1965; Bell Telephone Laboratories, 1965—. In Braunschweig, Mr. Gloge was engaged in research on lasers and optical components. At Bell Telephone Laboratories, he has concentrated on the study of optical transmission techniques. Member, VDE, IEEE.

Shui Yee Lee, B.S.E.E., 1964, University of Maryland; M.S.E.E., 1965, and Ph.D. (E.E.), 1967, University of Pennsylvania; teaching fellow, the Moore School of Electrical Engineering, University of Pennsylvania, 1965–1967; member of research staff of Bockus Research Institute, Graduate Hospital of University of Pennsylvania, 1966–1967; Bellcomm, Inc., 1967—. At the University of Pennsylvania, Mr. Lee was engaged in research on synthesis techniques and methods for determining transfer functions of physical systems. At Bellcomm, he has concentrated on the study of digital-optical and electro-optical information processing. He is also interested in communication systems optimization. Member, Sigma Xi.

Jack M. Manley, B.S. (Electrical Engineering), 1930, University of Missouri; Bell Telephone Laboratories, 1930—. He was first concerned with theoretical and experimental studies of nonlinear electric circuits. He later worked with new multiplex methods for communication systems, including early research work on PCM. Afterward, he was engaged in transmission line research, and at present he is working on noise problems in digital transmission systems. Fellow, IEEE; member, Sigma Xi, Tau Beta Pi and Eta Kappa Nu.

F. W. Mounts, E.E., 1953, and M.S., 1956, University of Cincinnati; Bell Telephone Laboratories, 1956—. Mr. Mounts has been primarily concerned with research in efficient methods of encoding pictorial information for digital television systems. Member, IEEE, Eta Kappa Nu.

Grace Murray, B.A., 1962, Duke University; M.S., 1966, Stevens Institute of Technology; Bell Telephone Laboratories, 1962–68; The RAND Corporation, 1968—. Miss Murray has worked extensively on traffic studies of complex telephone systems, using both stochastic simulation and mathematical techniques. Also, she has taught the GSP course, Advanced Programming. She is working on a study of the deployment and dispatching operations of the New York City Fire Department. Member, Phi Beta Kappa.

Donald E. Pearson, B.Sc. (Eng.), 1957, University of Cape Town; Ph.D., 1965, Imperial College, University of London; Bell Telephone Laboratories, 1965—. Mr. Pearson has been involved with picture coding, especially subjective studies of the effect of various bandwidth compression techniques on picture quality. He presently is engaged in research into the laws of color mixture in complex scenes such as television pictures and the choice of primary colors for optimum rendition of skin tones. Member, IEE, Optical Society of America.

John R. Pierce, B.S., 1933, M.S., 1934, and Ph.D. (E.E.) 1936, California Institute of Technology. He has published 12 technical books, hundreds of papers and articles, a number of science fiction stories (some under the name J. J. Coupling), and a few poems. Some of his computer music appears on a Decca record, Music from Mathematics. His awards include: Eta Kappa Nu, 1942; Morris Liebmann Memorial Prize, 1947; Stuart Ballantine Medal, 1960; Air Force Association H. H. Arnold Trophy, 1962; the Arnold Air Society General Hoyt S. Vandenberg Trophy, 1963; the Edison Medal, 1963; the Valdemar Poulsen Medal, 1963; the National Medal of Science, 1963; the H. T. Cedergren Medal, 1964; Caltech Alumni Distinguished Service Award, 1966; and six honorary degrees.

Dr. Pierce is Executive Director, Research, Communications Sciences Division of Bell Laboratories, with responsibilities in radio, electronics, acoustics and vision, mathematics, economic analysis, and psychology. Member, National Academy of Sciences, National Academy of Engineering, Air Force Association; Fellow, American Academy of Arts and Sciences, IEEE, American Physical Society, Acoustical Society of America. He is a Kentucky Colonel.

Vasant K. Prabhu, B.E. (Dist.), 1962, Indian Institute of Science, Bangalore, India; S.M., 1963, Sc.D., 1966, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1966—. Mr. Prabhu is a member of Radio Research Laboratory, and his areas of interest include systems theory, solid-state microwave devices, noise theory, and optical communication systems. Member, IEEE, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, AAAS.

HARRISON E. ROWE, B.S., 1948, M.S., 1950, Sc.D., 1952, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1952—. His fields of interest have included parametric amplifier theory, noise and communication theory, propagation in random media, and related problems in waveguide, radio, and optical systems. Member, IEEE, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

ERHARD K. SITTIG, Dipl. Imper. College, E.E. 1954, London, U.K.; Dipl., Phys., 1955, Univ. Tübingen, Germany; Dr. rer. nat., 1959, Techn. Hochschule, Stuttgart, Germany; Bell Telephone Laboratories, 1963—. Since 1963, Mr. Sittig has been working on ultrasonic devices, notably diffraction delay lines. He now supervises a group active in ultrasonic device technology, photodetectors, and access circuitry development for an exploratory optical memory. Member, German Phys. Soc., IEEE, Acoust. Soc. of America.

FRIEDOLF M. SMITS, Dipl. Phys., 1950, Dr. rer. nat., 1950, University of Freiburg, Germany; research associate, Physikalisches Institut, University of Freiburg, 1950–54; Bell Telephone Laboratories, 1954–62; Sandia Corporation, 1962–65; Bell Telephone Laboratories, 1965—. Mr. Smits' early work at Bell Telephone Laboratories included studies of solid-state diffusion in germanium and silicon, exploratory semiconductor device development, and radiation damage studies for the *Telstar*® communications satellite. At Sandia Corporation he was responsible for work on radiation effects, particularly electron and neutron damage to semiconductors and semiconductor devices. His more recent responsibilities at Bell Telephone Laboratories were in the field of ultrasonics and acousto-optics. He is presently Director of the Semiconductor Device Laboratory at Murray Hill. Senior Member, IEEE; Member, American Physical Society, German Physical Society.

S. Y. TONG, B.S., 1955, Taiwan University; M.S., 1961, University of Vermont; Ph.D., 1966, Princeton University; Bell Telephone Laboratories, 1964—. Mr. Tong has worked on problems in coding theory. Member, IEEE, AAAS, Sigma Xi.

B.S.T.J. BRIEFS

Correction Concerning Reflecting Objects in Coherent Illumination

By L. H. ENLOE

(Manuscript received January 31, 1969.)

In a recent paper, the author analyzed and discussed the noise-like structure in images of diffuse objects in coherent illumination.* While the author's interest and discussion concerned objects viewed in *reflection,* a model illustrating a diffuse object was unfortunately shown in Fig. 1 as a granular transparency viewed in *transmission.* It turns out that the analysis presented is not sufficiently general to cover transparencies viewed in transmission. The object *must* be viewed in reflection because:

(*i*) The direct beam, that is, the unscattered component, is not included in the fundamental equation (1).

(*ii*) The relative phase angles θ_i of the individual scatterers in equation (1) are not unqualifiedly random in the forward scattering direction.

I would like to thank D. Berkley for bringing this to my attention.

* Enloe, L. H., "Noise-Like Structure in the Image of Diffusely Reflecting Objects in Coherent Illumination," B.S.T.J., *46*, No. 7 (September 1967), pp. 1479–1491.