THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 57 March 1978 Number 3

A Low-Noise Gallium Arsenide Field Effect Transistor Amplifier for 4 GHz Radio

By R. H. KNERR and C. B. SWAN

(Manuscript received June 14, 1977)

A low-noise amplifier for 4 GHz radio has been designed and is in manufacture. The noise figure is ≤2 dB and the gain is typically 10 dB. Input and output return losses are ≥25 dB. The insertion loss with failure of either the power supply or the low-noise transistor is typically 5 to 8 dB. The amplifier uses a single gallium arsenide field effect transistor in conjunction with a passive failsafe by-pass network utilizing circulators. This approach permits the noise figure and the gain flatness to be optimized for each amplifier without compromising the input and output matches. It is concluded that this single-transistor amplifier design has significant advantages both in performance and in simplicity over the balanced amplifier design.

I. INTRODUCTION

Gallium arsenide Field Effect Transistors (GaAs FETs) are effecting a revolution in both the design philosophy and the performance capability of new microwave systems. In addition, these devices can often provide an economical means for significantly upgrading the performance of existing systems. Such is the case with the 4 GHz radio system, where an RF preamplifier with a maximum noise figure of 2 dB is achieved with GaAs FETs. In this application, each common multichannel amplifier permits the output power of typically five transmitters to be dropped 4 dB, from 5 watts to 2 watts, while still maintaining the system thermal noise objective for 1500 channels. This significantly increases the life of the transmitter amplifier triodes, thus improving the overall system reliability.

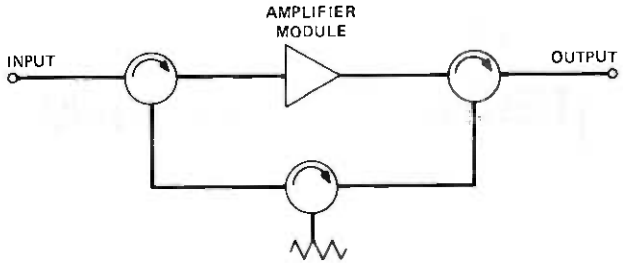


Fig. 1—Single-ended amplifier with provision for unpowered transmission.

II. GENERAL DESIGN CONSIDERATIONS

The use of the GaAs FET amplifier as an RF preamplifier for FM systems requires low intermodulation as well as a low noise figure. In addition, since the amplifier is common to several channels (including the protection channel), reliability is of utmost importance. The two most serious failure mechanisms envisioned are: (*i*) transistor failure and (*ii*) power supply failure. With either type of failure, the GaAs FET amplifier inherently exhibits an unacceptable transmission loss (>20 dB) for radio applications. Use of a balanced amplifier with two transistors coupled with input and output 3 dB hybrid couplers would reduce the gain by only 6 dB for failure of a single transistor. But this redundancy and extra cost gives no relief for loss of the dc supply voltage for the transistors.

Schemes, without active devices, for reducing the loss to <10 dB for either type of failure and which apply to the balanced as well as the single-ended amplifier are shown schematically in Fig. 1 and 2. The signal reflected from the unpowered FETs is fed to the output by interconnecting the normally terminated arms of the coupler (Fig. 2) or isolator (Fig. 1). We have designed, constructed, and evaluated both balanced and single-ended amplifiers.

The requirements for this application are shown in Table I. The choice of the design approach to meet these requirements was based on a de-

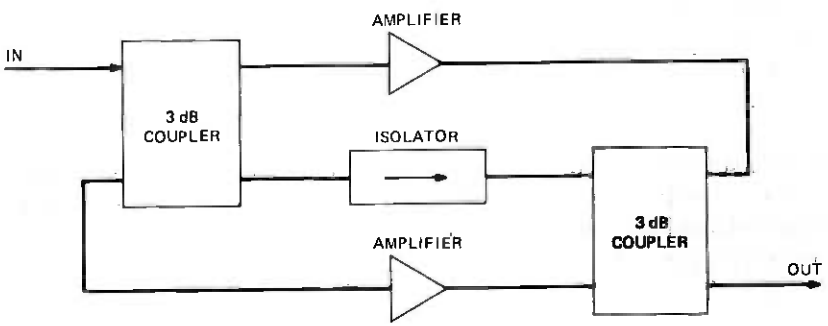


Fig. 2—Balanced amplifier with provision for unpowered transmission.

Table I — Electrical requirements for 3.7 GHz to 4.2 GHz amplifier

	Max.	Min.	Units
Input return loss	—	25	dB
Output return loss	—	25	dB
Noise figure	2.0	—	dB
Gain	11.0	8.0	dB
Gain flatness	±0.5	—	dB
Intermodulation (2A-B intercept)	—	23	dBm
Unpowered insertion loss	10	—	dB

Table II — Single-ended versus balanced amplifier

	Single	Balanced
Gain	Same (8–11 dB)	
Unpowered loss	Same	
Output return loss	≥25 dB	≥20 dB
Input return loss	≥25 dB	≥17 dB
Transistor failure	6–8 dB loss	2–5 dB gain
Intermodulation (2A-B intercept)		3 dB advantage
Noise figure	0.3 dB advantage	
Transistors required	1	2
Couplers required	0	2
Circulators required	3	1

tailed comparison of the capabilities of the two amplifiers. Based on our laboratory experience, Table II compares the performances that we consider practical in manufacture. We realized that meeting the intermodulation and failsafe requirements with a single transistor would allow significant cost savings. The single-ended GaAs FET amplifier reported here not only meets these requirements but also has match and noise figure advantages. This results from the low loss input circulator which allows us to independently optimize the input circuit match and the transistor source impedance for minimum noise.

III. AMPLIFIER MODULE

The GaAs FET is mounted in a microstrip circuit (Fig. 3). This transmission line permits easy mounting of the transistor and MOS dc-blocking capacitors. The amplifier module per se has no adjustments. Tuning screws near the input and output of the module and in the circulator arms are used to adjust the amplifier for optimum noise figure and gain flatness. This feature compensates for variations in transistor parameters as well as for manufacturing tolerances of the piece parts.

3.1 The GaAs FET output circuit design

In a first order approximation the output circuit elements were determined using BAMP.* Supplying the S-parameters and the input re-

* Basic Analysis and Mapping Program.

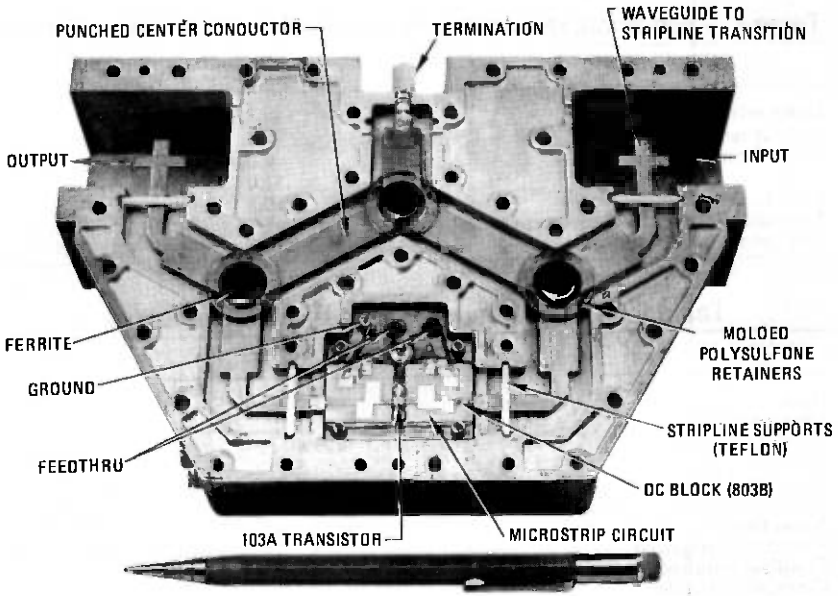


Fig. 3—4 GHz MIC amplifier (open).

flection coefficient (Γ_{MNF}) that results in a minimum noise figure, circles of constant gain are drawn (Fig. 4). If the output circuit reflection coefficient equals Γ_{ML}, optimum gain is obtained. Any deviation from Γ_{ML} results in a loss of gain corresponding to the values indicated on the circles of Fig. 4. Strictly speaking, the set of circles is only valid for one frequency (in our specific case 4 GHz), and a corresponding set would have to be drawn for each frequency under consideration. Since the S-parameter variation over the 12.5 percent frequency band of interest is smooth and relatively small, one set of circles suffices to demonstrate that the output impedance, shown in a dashed line, is reasonably close to match. The actual circuit which produced the impedance was trimmed empirically for bandwidth and flatness of gain.

3.2 The input circuit

The theory of noisy four poles has been treated extensively in the literature.[1–5] It essentially says that the noise figure of the four pole depends solely on the impedance of the input circuit. The noisy four pole is completely characterized by the S-, Y-, or Z-parameters, the source reflection coefficient (Γ_{MNF}) at which the noise figure is minimum (NF_{MIN}), and the equivalent noise resistance (R_n). The measurement of R_n is somewhat cumbersome and is described in Ref. 1. Once the parameters are known, circles of constant noise figure[4,5] can be drawn (Fig.

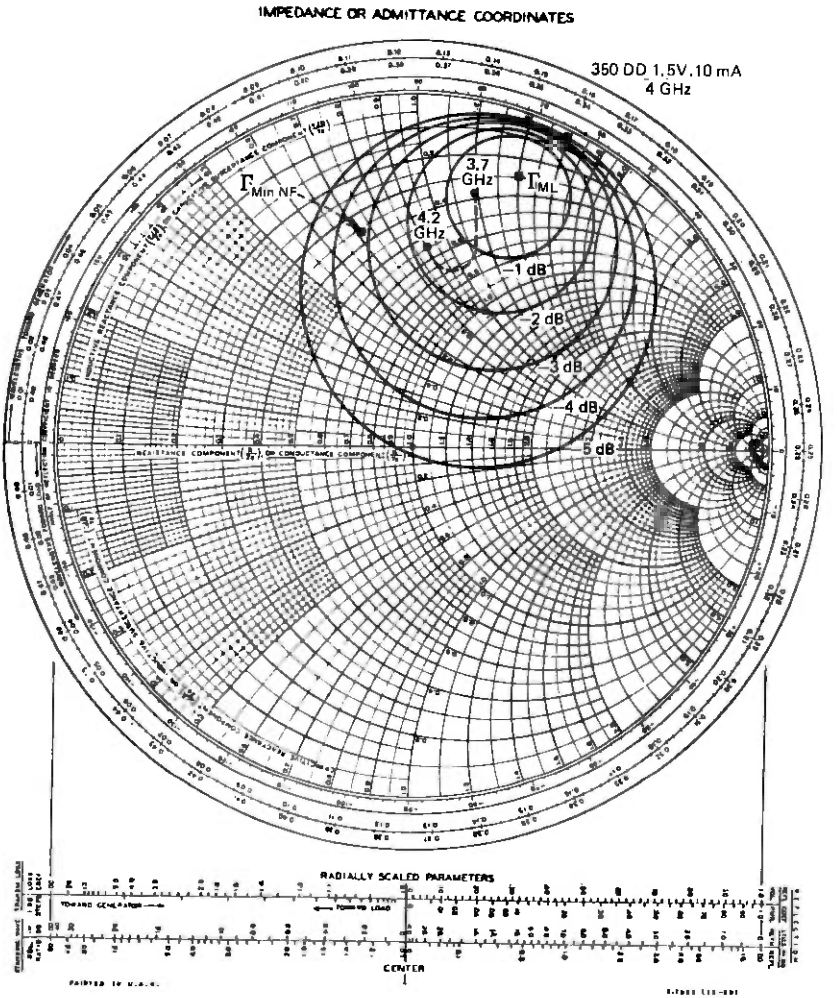


Fig. 4—Circles of constant gain with input tuned for minimum noise figure.

5, solid circles). This set of circles is very insensitive to frequency and independent of the load impedance. The spread of the circles increases with increasing R_n. In our specific case $R_n = 14\ \Omega$. Γ_{ML}, the maximum gain load impedance, has been explained in the output circuit design. The reflection coefficient, Γ_{MS}, in Fig. 5 represents the reflection coefficient of the source that would yield maximum gain, which in our case is about 15.5 dB. It is quite obvious that the points for optimum noise figure and optimum gain are significantly apart. A set of circles similar to the ones in Fig. 4 can be constructed around Γ_{MS}, assuming that the load reflection coefficient is Γ_{ML}. To keep Fig. 5 from becoming over-

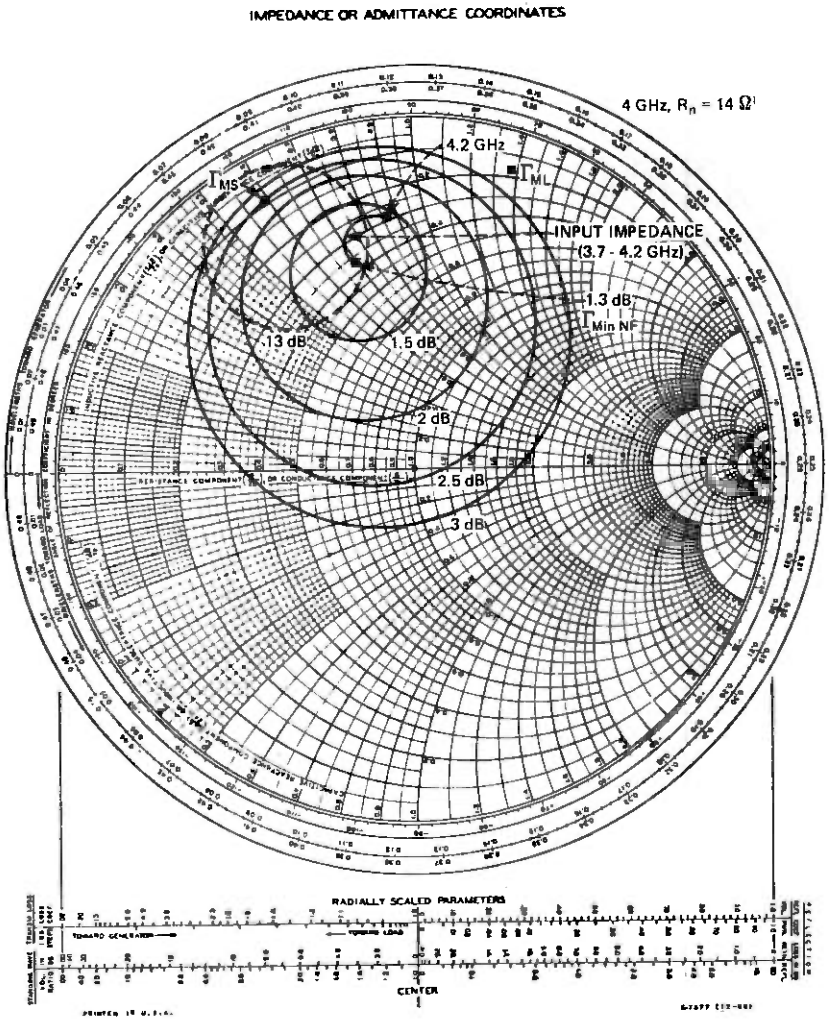


Fig. 5—Circles of constant noise figure.

crowded, only one circle is shown. It is seen that a gain of about 13 dB for optimum noise figure versus 15.5 dB for optimum match can be obtained. This figure, of course, is further reduced by broadbanding and gain flattening, as can be deduced from the source impedance trace in Fig. 5. The performance of the single-ended amplifier module is shown in Fig. 6. The gain of 11.6 dB and corresponding noise figure of 1.5 dB are in good agreement with the values that can be extrapolated from Figs. 4 and 5.

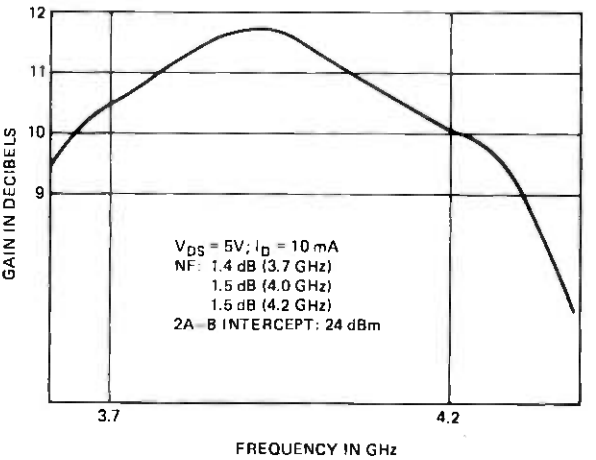


Fig. 6—Performance of amplifier module.

3.3 Final circuit

For the amplifier to be manufacturable, some adjustability is required to compensate for variations in transistor parameters as well as mechanical tolerances on all components. This adjustability is not readily provided in the microstrip circuit, but can be economically introduced in the air dielectric stripline circuit. Pairs of tuning screws are thus located in the air line just in front and just after the microstrip module (Fig. 7). These permit tuning of the amplifier for optimum noise figure and gain flatness.

IV. FAILSAFE BYPASS CIRCUIT

When the GaAs FET is unpowered, both the gate and drain circuits appear approximately as open circuits. The transmission loss typically exceeds 20 dB. If the transistor fails, we expect a short circuit. In either case, the input and output return losses at 4 GHz are typically 2 to 4 decibels.

The provision of three circulators, as shown in Fig. 1, provides an effective passive by-pass circuit. In the normal state, the relatively small reflected input signal is recombined with the amplifier signal at the output of the transistor. This appears as a small ripple on the gain characteristic which can be compensated by output tuning. In the unpowered or failed state, both the gate and drain circuits are "switched" to open or short circuits. The input signal, with relatively small loss, is then directed to the drain circuit of the GaAs FET where it is reflected to the output circulator and directed to the load. The total insertion loss is typically 5 to 8 dB.

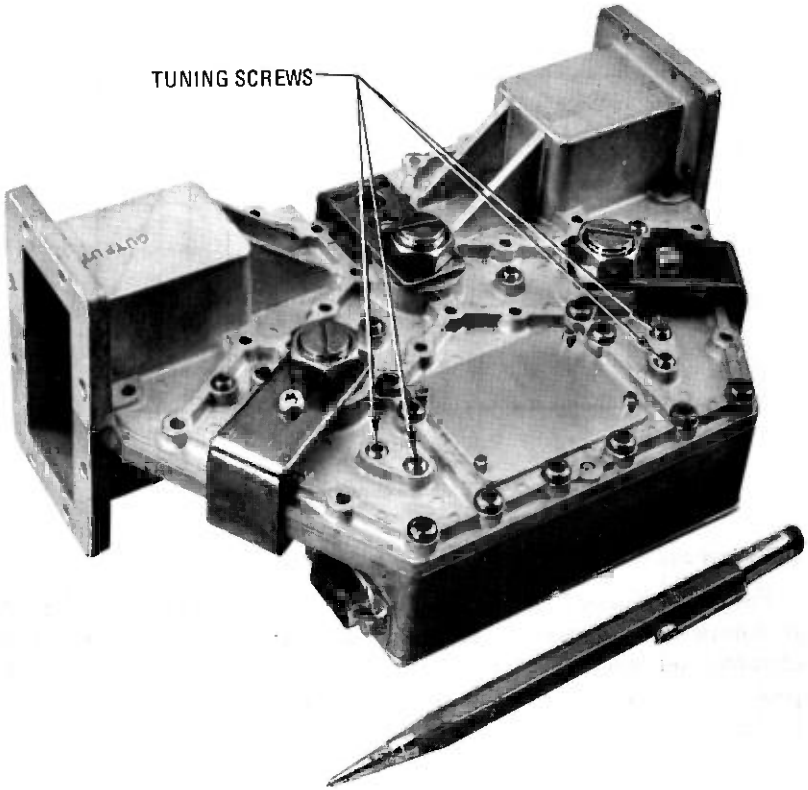


Fig. 7—4 GHz MIC amplifier.

The circulators for the bypass circuit and the waveguide-to-stripline transition[6] were developed in air dielectric stripline (Fig. 8). This simple technology assures minimum circuit losses, low cost parts and assembly, and very high yields. The intermediate circulator is terminated with 50 ohms to provide >25 dB isolation. Since this isolation is only maintained over the 3.7–4.2 GHz band, positive feedback can cause the amplifier to oscillate at lower frequencies. The "low-pass filter" on the output substrate (Fig. 9) eliminated this oscillation which, for our particular by-pass loop, occurred at about 800 MHz.

V. PDWER REGULATOR AND ALARM CIRCUIT

The dc operating point for the GaAs FET is a compromise between minimum noise and acceptable linearity. A regulator automatically sets the gate voltage so that I_D = 15 mA and V_{DS} = 4.8 volts. All GaAs FETs are thus powered identically and require no bias adjustment in manufacture. The amplifier (Fig. 10) operates from a −24 volt supply at 60 milliamperes.

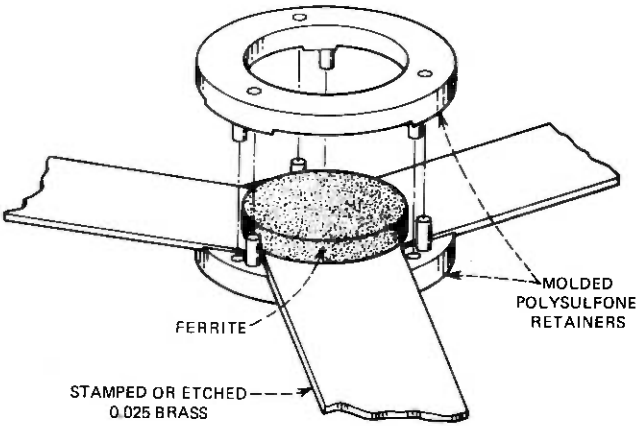


Fig. 8—4 GHz stripline circulator.

In case of transistor or power supply failure ($I_D < 5$ mA or $I_D > 25$ mA), a contact to ground is provided which energizes a remote alarm.

VI. THE LOW-NOISE TRANSISTOR

The GaAs FET was developed at the Murray Hill, New Jersey Laboratory.[7] The gate length and width are 0.8 μm and 2 × 250 μm. The typical noise figure is about 1.2 to 1.4 dB at 4 GHz.

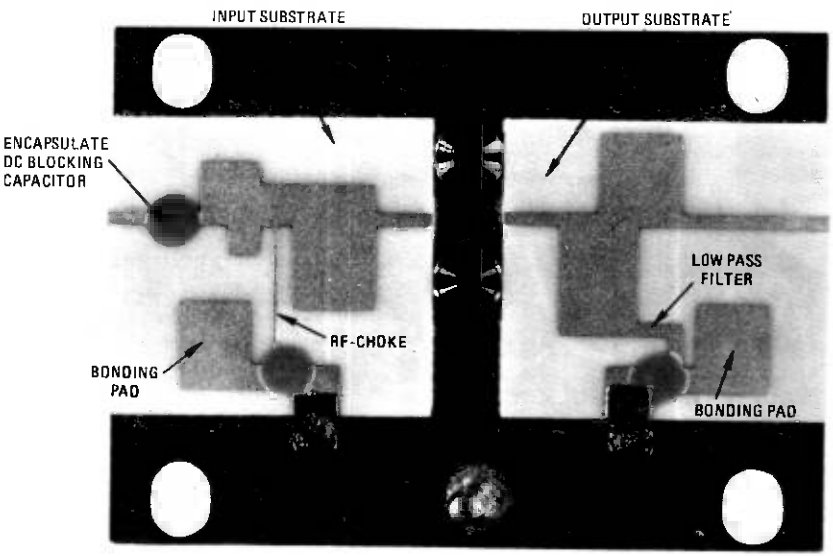
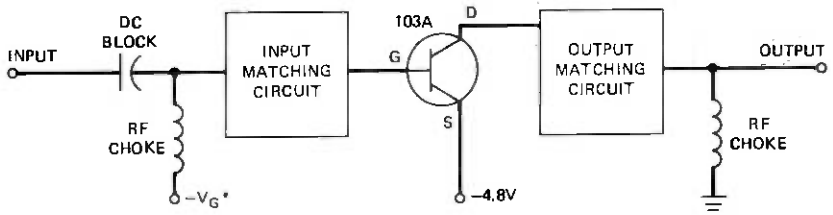


Fig. 9—Amplifier module without transistor.



*V_G AUTOMATICALLY ADJUSTED FOR I_{DS} = 15 mA

Fig. 10—Amplifier module (schematic).

VII. PHYSICAL DESIGN

The completed amplifier is shown in Figs. 3 and 7. The aluminum housing is die cast in two parts. The stripline center conductor is stamped in a single piece from sheet brass. Interlocking molded plastic locating rings are used to locate both the circulator ferrites and the center conductor in the lower housing channel. The printed circuit board with power regulator and alarm circuits (Fig. 11) is mounted on the bottom side of the lower housing.

VIII. AMPLIFIER PERFORMANCE AND TESTS

8.1 Tests

In order to meet the requirements in Table I, the amplifier was subjected to several tests, most of which used straightforward test procedures. Special test sets were constructed for noise figure and inter-

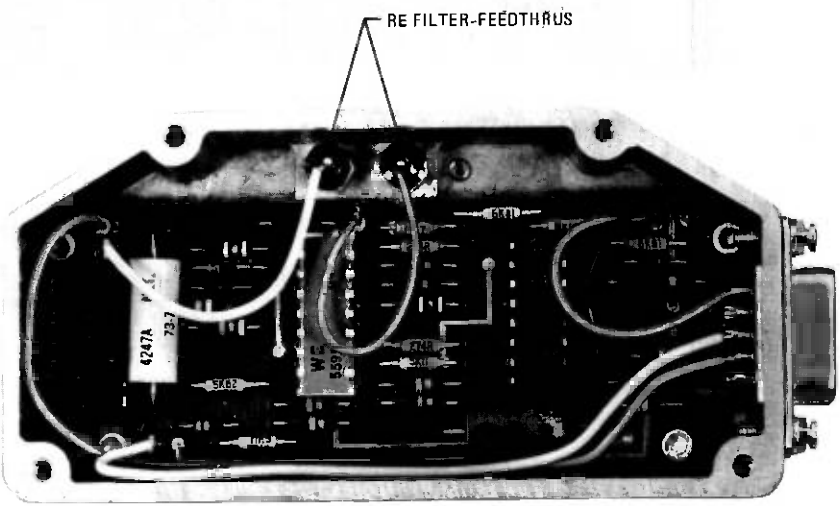


Fig. 11—Power regulator and alarm circuit of 4 GHz MIC amplifier.

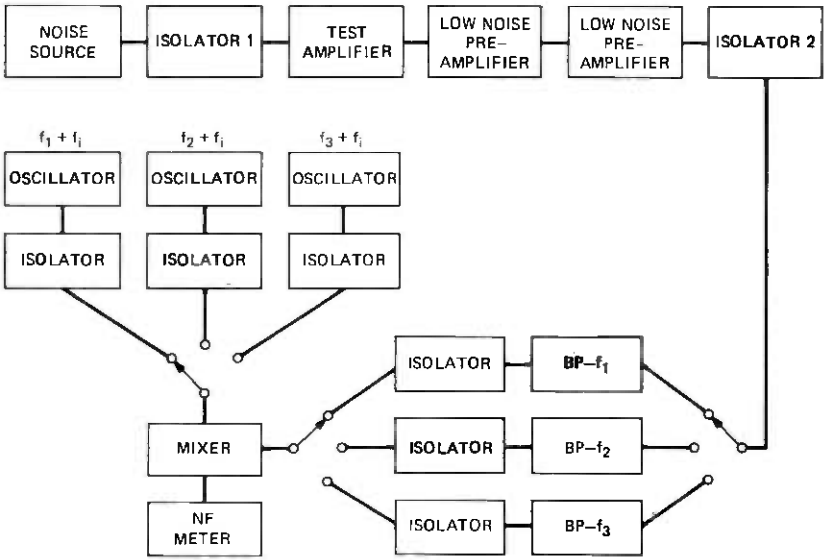


Fig. 12—Complete three-frequency NF test set.

modulation measurements. The noise figure can be accurately and rapidly measured at three frequencies in the test set shown in Fig. 12. Intermodulation (IM) tests were done using a three-tone measurement set. Since not all stations are air-conditioned, a humidity test was started. The amplifier was placed in an 85°C (185°F)—85% humidity environment for two months with DC bias applied. No change in performance was detected. Six field trial models were cycled over the temperature range of 4°C (40°F) to 60°C (140°F) with no significant change in performance.

8.2 Performance

The amplifier is in manufacture at Western Electric Company and meets the requirements summarized in Table I. Typical performance values obtained are:

NF: 1.6–1.8 dB
Input and output return loss: 28 dB
2A-B intercept: 26 dBm
Unpowered transmission loss: 5–8 dB
Gain: 10 dB.

We find that the amplifier tuning arrangement permits the present spread in transistor parameters to be accommodated easily.

IX. CONCLUSION

We have demonstrated a simple 4 GHz microwave amplifier design which achieves a noise figure of 2 dB in manufacture. This has been achieved with a single low noise GaAs field effect transistor in conjunction with a passive failsafe by-pass circuit. It is concluded that the single-ended amplifier with input and output isolator has significant advantages both in performance and in simplicity over the balanced amplifier design for this application. The housing and major piece parts are die replicated so as to fit together with minimal assembly effort. Tuning screws are provided to accommodate variations in transistor characteristics and to allow relaxed piece part tolerances.

X. ACKNOWLEDGMENTS

We gratefully acknowledge the team efforts of Bell Laboratories and Western Electric engineers in the development of this amplifier. We mention especially the efforts of J. J. Kostelnick, G. M. Keltz, G. M. Palmer, and J. L. Brown. The GaAs FET was developed by J. V. DiLorenzo and W. O. Schlosser. L. F. Moose and L. J. Varnerin, Jr., provided essential coordination and technical direction for the program.

REFERENCES

1. H. Rothe and W. Dahlke, "Theorie rauschender Vierpole," Archiv der Elektrischen Übertragung, 9 (March, 1955), pp. 117–121.
2. H. Rothe, "Die Theorie rauschender Vierpole und ihre Anwendung", Nachrichtentechnische Fortschritte, Issue 2 (1955), pp. 24–26.
3. A. G. Th. Becking, H. Groendijk, and K. S. Knol, "The Noise Factor of Four-Terminal Networks," Philips Res. Rep. 10, 1955, pp. 349–357.
4. H. Rothe and W. Dahlke, "Theory of Noisy Four Poles," Proc. IRE, *44*, June 1956, pp. 811–818.
5. H. Fukui, "Available Power Gain, Noise Figure, and Noise Measure of Two-Ports and Their Graphical Representations," IEEE Trans. Circ. Theory, *CT-13*, No. 2 (June 1966), pp. 137–142.
6. R. H. Knerr, "A New Type of Waveguide-to-Stripline Transition," IEEE Trans. Microw. Theory Tech. *MTT-16*, No. 3 (March 1968), pp. 192–194.
7. B. S. Hewitt, H. M. Cox, H. Fukui, J. V. DiLorenzo, W. O. Schlosser, and D. E. Iglesias, "Low Noise GaAs MESFET's—Fabrication and Performance," 6th International Symposium on GaAs and Related Compounds, Edinburgh, September 19–22, 1976.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 3, March 1978
Printed in U.S.A.

Theory of Analytic Modulation Systems

By B. F. LOGAN, JR.

(Manuscript received March 18, 1977)

A general theory of analytic modulation systems is developed where the transmitted signal is of the form $\sigma(t) = Re\{e^{i\omega_c t} f(z(t))\}$. *Here* $f(z)$ *is an analytic function (modulation law), and* $z(t) = x(t) + iy(t)$ *is the analytic baseband signal whose real part* $x(t)$ *is a bounded bandlimited signal of spectral support* $[-\Omega,\Omega]$ *which is assumed to have a bounded Hilbert transform* $y(t)$. *It is shown for a large class of* $\{z(t)\}$ *and modulation laws that* $z(t)$ *may be recovered using a receiver incorporating the inverse function of* f *as a detector with appropriate pre- and post-detection filtering. The theory also shows that in the procedure for factoring certain positive bandlimited signals, an approximate Hilbert transform operator (bandlimited) may be used. A related result is that signals subjected to logarithmic companding (one-sided) and filtering may be recovered by a non-iterative method.*

I. INTRODUCTION

In 1962, Bedrosian proposed a modulation system called single sideband phase (or frequency) modulation.* (See Ref. 5.) The modulated signal is of the form

$$Re\{e^{i(\omega_c t + x(t) + iy(t))}\} = e^{-y(t)} \cos(\omega_c t + x(t))$$

where $x(t)$ is the "baseband" signal, $y(t)$ is the Hilbert transform of $x(t)$, and ω_c is the carrier frequency. The special relation between the amplitude modulation $e^{-y(t)}$ and the phase modulation $x(t)$ results in the modulated signal having no spectrum in the interval $(-\omega_c,\omega_c)$; i.e., the amplitude modulation removes the lower sideband. However, the spectrum is still infinite in extent. We adopt the terminology "single sideband exponential modulation" (SSBEM) for this system.

Bedrosian pointed out that SSBEM was compatible with conventional FM receivers, and suggested that the single-sideband system might offer

* K. H. Powers received a U. S. Patent (No. 3,054,073) on such a system shortly before the appearance of Bedrosian's paper. See Voelcker.[21]

some savings in transmission bandwidth over the conventional system. However, since filtering operations could radically alter the zero crossings of the modulated signal, it was not clear to what degree one could maintain compatibility and at the same time realize some saving in bandwidth.

Others[2,10,17] have compared the spectral distribution of single sideband and conventional frequency modulated signals for the cases of sinusoidal and Gaussian noise modulation. They have shown that the "effective" bandwidths, as measured by central second moments, of the single-sideband signals may be greater or less than that of the conventional FM signal depending on the nature of the modulation. At any rate, it is not clear how one would translate these results into relative bandwidth requirements of the two systems, each employing a conventional FM receiver.

Aside from the compatibility question, Barnard[4] has shown that the transmission bandwidth requirements of single-sideband exponential modulation are, in a strict sense, minimal. He showed that if the modulation $x(t)$ belonged to a certain subclass of bandlimited signals with spectral support $[-\Omega,\Omega]$, that the modulation could be recovered, within an additive constant, from a knowledge of the spectral distribution of the single-sideband signal in the interval $[\omega_c,\omega_c + \Omega + \epsilon]$, provided $\epsilon > 0$. This was proved by demonstrating the convergence of an iterative recovery scheme.

Here we consider a class of single-sideband systems wherein the modulated signal is of the form

$$s(t) = Re\,\{f(z(t))e^{i\omega_c t}\}$$

where $f(z)$, the "modulation law", is an analytic function and $z(t) = x(t) + iy(t)$ is the "analytic signal" of which, say, the real part $x(t)$ is the information to be transmitted. We suppose that $x(t)$ is bounded and bandlimited with spectral support $[-\Omega,\Omega]$ and that $s(t)$ is transmitted over a channel whose transmission function is the Fourier transform of an absolutely integrable function (impulse response) and is equal to unity over $(\omega_c,\omega_c + \alpha)$. It is shown under fairly weak conditions on $f(z)$ and $z(t)$ that $x(t)$ may be recovered (by a relatively simple non-iterative method) from the received signal, provided $\alpha > \Omega$.

The gist of the method can be grasped by considering periodic signals; e.g.,

$$x(t) = \frac{1}{2}\sum_{-n}^{n} X_k e^{ikt} \qquad (X_0 = 0, \text{ say})$$

$$z(t) = \sum_{1}^{n} X_k e^{ikt}.$$

We suppose that $\sup |z(t)| = m$, and $f(z)$, the modulation law, is analytic for $|z| \leq m$, ($f(0) = 0$, say)

$$f(z) = \sum_1^\infty a_k z^k, \quad |z| \leq m.$$

Then setting $w(t) = f\{z(t)\}$ we have

$$w(t) = \sum_1^\infty W_k e^{ikt}$$

where W_k depends only on those X_j and a_j for which $1 \leq j \leq k$. This sort of dependence allows us to determine $x(t)$ from a bandlimited version of $w(t)$, say,

$$w_n(t) = \sum_1^n W_k e^{ikt},$$

by what amounts to reversion of power series, provided $f'(0) \neq 0$. We have

$$z = \phi(w) = \sum_1^\infty b_k w^k, \quad \text{for } |w| \text{ sufficiently small.}$$

$$(b_1 = (a_1)^{-1} = \{f'(0)\}^{-1})$$

Assuming that the series converges when w is replaced by $w_n(t)$ we set

$$z_n(t) = \phi\{w_n(t)\} = \sum_1^\infty b_k \{w_n(t)\}^k$$

and then by formal composition of the power series find that

$$z_n(t) = \sum_1^n X_k e^{ikt} + \sum_{n+1}^\infty c_k e^{ikt}.$$

So the first n Fourier coefficients of $z(t)$ and $z_n(t)$ agree; i.e., under the stated assumptions, $x(t)$ can be recovered by bandlimiting $\phi\{w_n(t)\}$, where ϕ is the inverse function, $z = \phi(w)$, and $w_n(t)$ is the partial sum of the Fourier series of $w(t)$.

The simplicity of this procedure owes to the fact that $z(t)$ has a one-sided spectrum and the *analytic* modulation law $f(z)$ then gives a function $w(t) = f\{z(t)\}$ which also has a one-sided spectrum. Since $z(t)$ contains no negative-frequency components, the usual difference terms do not appear; i.e., the spectrum of $w(t)$ in the frequency interval $[0,\alpha]$ depends only on the spectrum of $z(t)$ in the same interval.

The recovery procedure is not so transparent for more general bandlimited signals $x(t)$. First of all, filtering $w(t)$ with a filter whose transmission function is unity over $[0,\alpha]$ and zero for frequencies greater than

β ($\beta > \alpha$) will give a function $w_{\alpha,\beta}(t)$, analogous to $w_n(t)$, which may differ considerably from $w(t)$. It may be that $\phi\{w_{\alpha,\beta}(t)\}$ does not have a one-sided spectrum; i.e., $\phi\{w_{\alpha,\beta}(\tau)\}$, $\tau = t + iu$, is not analytic in the upper half-plane $u > 0$. Indeed $w_{\alpha,\beta}(t)$, $(-\infty < t < \infty)$, may not even be in the domain of definition of ϕ; i.e., ϕ may have a natural boundary beyond which it cannot be extended. Even if $\phi\{w_{\alpha,\beta}(t)\}$ does have a one-sided spectrum, the function ϕ need not have a power series representation over the range of $w_{\alpha,\beta}(t)$ so that one cannot use convolution arguments to show that the Fourier transforms of $\phi\{w(t)\}$ and $\phi\{w_{\alpha,\beta}(t)\}$ agree over $[0,\alpha]$. This particular problem is met by using generalizations of the Paley-Wiener theorem.

The problem arising when $\phi\{w_{\alpha,\beta}(t)\}$ does not have a one-sided spectrum ($w_{\alpha,\beta}(t)$ not in the range of the inverse function) is met by imposing restrictions on $z(t)$, namely that for sufficiently large u, the range of $z(t + iu)$ is sufficiently small that $w(t + iu)$ will be in the range of the inverse function. This implies that $w_{\alpha,\beta}(t)$ may be filtered (with a Poisson filter) to obtain $w_{\alpha,\beta}(t + ib)$, which for sufficiently large b will be in the range of the inverse function. One can then obtain $z(t + ib)$ and then use inverse Poisson filtering to recover $z(t)$.

Although one could conceivably recover $z(t)$ from $w_{\alpha,\beta}(t)$ by other procedures when $\phi\{w_{\alpha,\beta}(\tau)\}$ is not analytic for $u \geq b$, the method here avoids any decision process and gives a simple receiver model incorporating the inverse function and (possibly) a Poisson filter with its (bandlimited) inverse and appropriate low-pass filter.

Generally speaking, given $f(z)$ and the channel transmission function, one can design a receiver which will work for a certain subclass $\{z(t)\}$ of signals. Or given $f(z)$ and a fixed receiver design, one may ask for the minimum bandwidth channel required for transmitting a given subclass of signals. In this connection, some estimates are given for the bandwidth requirements of "compatible" single-sideband exponential modulation with $\{z(t)\}$ all functions of spectral support $[0,1]$ satisfying $\sup |z(t)| \leq m$. In a recent work, Werner[23] has considered the same problem for $z(t)$ in L_2 and gives upperbounds in terms of the L_2-norm of z.

There are rather dramatic mathematical simplifications in the detection theory when the signals $\{x(t)\}$ are restricted to be of the band-pass type, allowing radical changes in the system design.

The theory also shows that the factorization of certain positive bandlimited signals can be effected with an *approximate* Hilbert transform operator acting on the logarithm of the signal. A related result pertains to the signal recovery problem considered by Landau and Miranker[11,12]; viz., there is one companding function ("log") for which the signal can be recovered by a non-iterative method.

Another interesting consequence of the general theory is the fact that for n arbitrary numbers a_k, $k = 1,2,\ldots,n$, there exists an integer $\nu \geq$

n and corresponding numbers a_k, $k = n+1, \ldots, \nu$, such that the polynomial

$$P_\nu(z) = 1 + \sum_1^\nu a_k z^k$$

is zero-free for $|z| < 1$.

Although the general theory is interesting from a mathematical viewpoint, it would appear that the practical interest in analytic modulation systems, other than the linear system, is limited to SSBEM, i.e., to the case $f(z) = e^z$ (or e^{iz}). In this case one can trade bandwidth for simplicity of detection. However, the trade-off is attractive only for moderate amplitudes of $z(t)$ where SSBEM offers an interesting alternative to other systems employing envelope detection.

It should be noted that the method here is naturally confined to bounded bandlimited signals $z(t)$, since otherwise we would require both $f(z)$ and its inverse $\phi(w)$ to be entire functions, i.e., $w = f(z) = a + bz$. The exception would be the band-pass case where f could be an entire function and ϕ replaced by an equivalent polynomial (see Section 5).

Of course, the theory here has to be extrapolated to practice with the appropriate "epsilons"; i.e., an analytic modulation law $f(z)$ can only be approximated within ϵ_1 over the disk $|z| \leq m$, and an analytic signal $z(t)$ having one-sided spectrum can be realized in practice within ϵ_2, and the impulse response of an ideal filter can be approximated (in L_1) within ϵ_3, etc. Then the continuity of the overall transformation may be used to bound the errors.

In order to deal rigorously with "communication type" signals which do not have ordinary Fourier transforms a considerable amount of preliminary mathematics is required. However, one can follow the theory assuming that the signals are either periodic or have ordinary Fourier transforms, with one cautionary note in mind. Abrupt bandlimiting operations (spectral projections) such as convolution with $\sin t/\pi t$ are not permissible (not defined) for the general signals of interest.

II. PRELIMINARIES

A measurable function $g(t)$ is said to belong to $L_p(-\infty, \infty)$ abbreviated hereafter as L_p, $(1 \leq p < \infty)$ if

$$\int_{-\infty}^{\infty} |g(t)|^p dt < \infty.$$

The L_p-norm of g is defined by

$$\|g\|_p = \left\{ \int_{-\infty}^{\infty} |g(t)|^p dt \right\}^{1/p}$$

and if α is a scalar

$$\|\alpha g\|_p = |\alpha| \cdot \|g\|_p.$$

If $|g(t)|$ is uniformly bounded, with the possible exclusion of a set of measure zero, g is said to belong to L_∞ and the norm of g is

$$\|g\|_\infty = \operatorname*{essup}_t |g(t)|$$

where "essup" over t is the essential supremum of $|g(t)|$, which is the infimum of numbers M such that

$$|g(t)| \leq M \text{ for almost all } t.$$

We will be mainly concerned with continuous bounded functions $g(t)$ in which case

$$\|g\|_\infty = \sup_t |g(t)|.$$

For $1 \leq p \leq \infty$, the L_p norm satisfies the *triangle inequality*

$$\|g_1 + g_2\|_p \leq \|g_1\|_p + \|g_2\|_p \tag{1}$$

which for any sequence of numbers a_k satisfying $\Sigma|a_k| < \infty$ and sequences of functions g_k such that $\|g_k\| \leq M$, leads to

$$\|\sum a_k g_k\|_p \leq \sum |a_k| \, \|g_k\|_p. \tag{2}$$

There are functions which belong to L_p for only one value of p. However, it is easy to see by considering the set where $|g(t)| \leq 1$ and the set where $|g(t)| > 1$, that if g belongs to L_r and L_s where $1 \leq r < s$, then g belongs to L_p for every p satisfying $r \leq p \leq s$. For example, the function $\sin t/t$ belongs to L_p for every $p > 1$.

Associated with the space L_p is the *conjugate* or *complementary* space L_q where

$$\frac{1}{p} + \frac{1}{q} = 1.$$

For functions in complementary spaces we have *Hölder's inequality*

$$\left| \int_{-\infty}^{\infty} g(t)h(t)dt \right| \leq \|g\|_p \|h\|_q, \quad 1 \leq p \leq \infty, \tag{3}$$

which for the case $p = q = 2$ is the familiar *Schwarz's inequality*.

In connection with Hölder's inequality we note that the norm of a function may be equivalently defined as

$$\|g\|_p = \sup_h \left| \int_{-\infty}^{\infty} h(t)g(t)dt \right|, \quad \|h\|_q = 1, \, q = p/(p-1).$$

A convolution kernel K in L_1 carries L_p into L_p. We have

$$K \otimes g(t) = \int_{-\infty}^{\infty} g(s)K(t-s)ds \quad \begin{matrix} g \text{ in } L_p \\ K \text{ in } L_1 \end{matrix}$$

and

$$\|K \otimes g\|_p \leq \|K\|_1 \|g\|_p \tag{4}$$

which amounts to a generalization of (2) to weighted sums of translates of g. The convolution integral is not in general defined for each t unless K belongs to the conjugate space of g. In general the convolution is defined as the limit in L_p of

$$g_m(t) = \int_{-\infty}^{\infty} g(s)K_m(t-s)ds$$

where K_m is a sequence of bounded functions of L_1 (hence K_m belongs to L_q) satisfying

$$\lim_{m\to\infty} \int_{-\infty}^{\infty} |K(t) - K_m(t)|dt = 0.$$

2.1 The Fourier transform on L_p

A function g in L_p has an ordinary Fourier transform, provided $1 \leq p \leq 2$, (a theorem of M. Riesz, cf. Ref. 20) in the sense that

$$\hat{g}_T(\omega) = \int_{-T}^{T} g(t)e^{-i\omega t}dt$$

converges in norm as $T \to \infty$ to a function $\hat{g}(\omega)$ belonging to the complementary space L_q, i.e., there exist a function $\hat{g}$ in L_q such that

$$\lim_{T\to\infty} \|\hat{g} - \hat{g}_T\|_q = 0.$$

However the Fourier transform on L_p, $1 \leq p \leq 2$, does not carry L_p into all of L_q except in the case $p = 2$. In particular, the Fourier transform of a function g of L_1 is a continuous function. Furthermore, (the Riemann-Lebesgue Lemma)

$$\lim_{\omega\to\pm\infty} \int_{-\infty}^{\infty} g(t)e^{-i\omega t}dt = 0$$

for g in L_1. Unfortunately, there is no simple description of functions $\hat{g}(\omega)$ which are the Fourier transforms of functions $g(t)$ of L_1. A useful sufficient cordition is that $\hat{g}(\omega)$ belong to L_2 and have a "derivative in L_2" (meaning only that $\hat{g}(\omega)$ is the integral of a function of L_2 denoted by $d\hat{g}/d\omega$). The sufficiency of this condition may be seen by writing

$$\int_{-\infty}^{\infty} |g(t)|dt = \int_{-\infty}^{\infty} \frac{1}{|t+ia|} \cdot |(t+ia)g(t)|dt \quad (a>0)$$

and then applying Schwarz's inequality and Parseval's theorem. The result is (choosing the best value of a)

$$\left\{\int_{-\infty}^{\infty} |g(t)|\,dt\right\}^2 \leq \|\hat{g}\|_2 \cdot \left\|\frac{d\hat{g}}{d\omega}\right\|_2. \tag{5}$$

In obtaining this result we use the fact that $\hat{g}$ in L_2 *and* $d\hat{g}/d\omega$ in L_2 imply that $\hat{g}(\omega)$ tends to zero at $\pm\infty$. (By Schwarz's inequality, $\hat{g}d\hat{g}/d\omega$ belongs to L_1, so $|\hat{g}(\omega)|^2$ is absolutely continuous and tends to limits at $\pm\infty$. The limits must be zero in order for $\hat{g}$ to belong to L_2.)

2.2 Bounded functions whose Fourier transforms vanish over certain sets

It is not necessary to attempt to define the Fourier transform of a bounded function $g(t)$ in order to give precise meaning to the statement that the Fourier transform of $g(t)$ vanishes over some open set E. This can be done in a way which is consistent with the ordinary Fourier transform, should it exist, of $g(t)$ vanishing over E. Here we restrict E to be the union of a finite number of disjoint open intervals.

Definition: *The Fourier transform of a bounded function g is said to vanish over E if and only if*

(*i*) $$\int_{-\infty}^{\infty} g(t)\bar{h}(t)dt = 0$$

for all h in L_1 *whose Fourier transforms satisfy*

(*ii*) $$\hat{h}(\omega) \equiv \int_{-\infty}^{\infty} h(t)e^{-i\omega t}dt = 0, \quad \omega \notin E.$$

This definition has its logical basis in Parseval's formula for functions of L_2. The bar over h in (i) denotes the complex conjugate of h. It is readily verified that (i) may be replaced by

(*iii*) $$\int_{-\infty}^{\infty} g(t)h(-t)dt = 0$$

which is more directly applicable to convolutions. That is, if the Fourier transform of g vanishes over E we have

$$(g \otimes h)(t) = \int_{-\infty}^{\infty} g(s)h(t-s)ds \equiv 0 \tag{6}$$

for all h in L_1 whose Fourier transforms vanish outside E.

We also note that in case the set E is symmetric with respect to the origin, $\bar{h}(t)$ in (i) may be replaced by $h(t)$. [See Ex. 4 below.]

We say that *the Fourier transforms of two bounded functions g_1 and g_2 agree over E if and only if the Fourier transform of $(g_1 - g_2)$ vanishes over E*. We also say that *g has spectral support E_c* (a closed set), or *the spectrum of g is confined to E_c*, meaning that *the Fourier transform of g vanishes over the complement of E_c*.

The following are some elementary consequences of the definition. The proofs are left as simple exercises. It is understood throughout that g, g_1, and g_2 are bounded functions.

Example 1. *Suppose all the intervals composing E are finite and $K(t)$ is any function of L_1 whose Fourier transform $\hat{K}(\omega)$ satisfies*

$$\hat{K}(\omega) = 1 \quad \text{for} \quad \omega \in E. \tag{7}$$

Let

$$g_2(t) = \int_{-\infty}^{\infty} g_1(s)K(t-s)ds. \tag{8}$$

Then the Fourier transforms of g_1 and g_2 agree over E.

Example 2. *If the Fourier transform of $g(t)$ vanishes over (α,β), then the Fourier transform of $e^{i\lambda t}g(t)$ vanishes over $(\alpha + \lambda,\beta + \lambda)$.*

Example 3. *If the Fourier transform of g_1 vanishes over E_1 and the Fourier transform of g_2 vanishes over E_2, then the Fourier transform of $(g_1 + g_2)$ vanishes over $E_1 \cap E_2$.*

Example 4. *If the Fourier transform of g vanishes over E, then the Fourier transform of the complex conjugate $\bar{g}$ vanishes over $E^{(-)}$, where $E^{(-)}$ denotes the reflection of E with respect to the origin.*

Example 5. *If the Fourier transform of g vanishes over E, then the Fourier transform of $Re\ \{g\}$ (or $I_m\ \{g\}$) vanishes over $E \cap E^{(-)}$.*

Note: $E \cap E^{(-)}$ may be the null set. However, if E is symmetric with respect to the origin, $E = E^{(-)}$. Hence a class of functions whose Fourier transforms vanish over a set E which is symmetric with respect to the origin is essentially a class of real-valued functions, since the real and imaginary parts of the functions separately belong to the class.

Example 6. (Reproducing Kernels) *Suppose the spectrum of g is confined to a set E_c consisting of n finite disjoint closed intervals. Let $K(t)$ be any function of L_1 whose Fourier transform $\hat{K}(\omega)$ satisfies*

$$\hat{K}(\omega) = 1, \quad \omega \in E_c. \tag{9}$$

Then for almost all t we have

$$g(t) = \int_{-\infty}^{\infty} g(s)K(t-s)ds. \tag{10}$$

(Set $g_1 = K \otimes g$ and show that $(g_1 - g)$ is orthogonal to all of L_1.)

Note: The qualification "almost all t" arises because the definition of a function g on a set of measure zero is irrelevant to the condition for its Fourier transform to vanish outside E_c. However, $(K \otimes g)(t)$ is a continuous function of t and so we will adopt the *convention* that *a function such as g in Ex. 6 is continuous.*

We note further that the condition that K in Ex. 6 belong to L_1 can be relaxed in case g belongs to L_p for some p satisfying $1 \leq p < \infty$. In this case we can take $\hat{K}(\omega) = 0$ for $\omega \notin E_c$. It is sufficient to prove this when E_c is a single interval $[-\Omega,\Omega]$ and this has been done (Ref. 14).

2.3 The Paley-Wiener Theorems for L_∞

There is an important connection between functions whose Fourier transforms vanish over a half-line and functions analytic and of exponential type in a half-plane. The following theorems are extensions to L_∞ of the classical "one-sided" and "two-sided" Paley-Wiener Theorems[18] for L_2.

Theorem 1. The Fourier transform of a bounded function g vanishes over $(-\infty,\alpha)$ if and only if $g(t)$ is the boundary value of a function $g(\tau)$, $\tau = t + iu$, analytic in the upper half-plane $u > 0$ and satisfying

$$\sup_t |g(t + iu)| \leq e^{-\alpha u} \sup_t |g(t)| \quad \text{for} \quad u \geq 0. \tag{11}$$

There is the analogous theorem connecting functions $g(t)$ whose Fourier transforms vanish over (β,∞) and functions $g(\tau)$ analytic in the lower half plane. The specialization of Theorem 1 to functions whose Fourier transforms vanish over $(-\infty,\alpha)$ and (β,∞) is the following. (We assume that $-\infty < \alpha < \beta < \infty$ and according to the convention above qualify g to be continuous.)

Theorem 2. The Fourier transform of a continuous bounded function g vanishes outside $[\alpha,\beta]$ if and only if $g(t)$ is the restriction to the real line of an entire function $g(\tau)$, $\tau = t + iu$, satisfying

$$\begin{aligned} \sup_t |g(t + iu)| &\leq e^{-\alpha u} \sup_t |g(t)|, \quad u \geq 0 \\ &\leq e^{-\beta u} \sup_t |g(t)|, \quad u \leq 0. \end{aligned} \tag{12}$$

These theorems are essential to the theory of single-sideband systems for bounded signals which do not have ordinary Fourier transforms.

Actually we do not need a uniform bound on the rate of growth (decay) of $g(t + iu)$ to infer that the Fourier transform of $g(t)$ vanishes over

$(-\infty,\alpha)$. In fact, an asymptotic bound implies a uniform bound.

Theorem 3. If $g(\tau)$ is analytic in the upper half-plane and satisfies

$$\sup_t |g(t+iu)| < \infty \quad \text{for} \quad u \geq 0$$

then the asymptotic estimate

$$\sup_t |g(t+iu)| = O\{e^{-\alpha u}\} \quad \text{as } u \to \infty$$

implies

$$\sup_t |g(t+iu)| \leq e^{-\alpha u} \sup_t |g(t)| \quad \text{for} \quad u \geq 0.$$

Proofs of Theorems 1, 2, and 3 are given in Appendix A.

We note the following corollaries of Theorems 1 and 2, concerning the Fourier transforms of products.

Corollary 1. If the Fourier transform of g_1 vanishes over $(-\infty,\alpha)$ and the Fourier transform of g_2 vanishes over $(-\infty,\beta)$, then the Fourier transform of $g_1 g_2$ vanishes over $(-\infty,\alpha+\beta)$.

Corollary 2. If the Fourier transform of g_1 vanishes outside $[\alpha_1,\beta_1]$ and the Fourier transform of g_2 vanishes outside $[\alpha_2,\beta_2]$, then the Fourier transform of $g_1 g_2$ vanishes outside $[\alpha_1+\alpha_2,\beta_1+\beta_2]$.

2.4 Terminology

Functions whose Fourier transforms vanish outside a finite interval are called *bandlimited* functions. Generally, we think of the interval centered at the origin and refer to bandlimited functions also as *low-pass* functions.

Functions whose Fourier transforms vanish over an interval centered at the origin are called *high-pass* functions, and functions which are both high-pass and low-pass are called *band-pass* functions.

Functions (signals) whose Fourier transforms vanish over a half-line, usually $(-\infty,0)$, are generally called *analytic* signals.

2.5 The Hilbert transform and the analytic signal

We would like to map the space of real-valued bounded signals $x(t)$ of spectral support $[-\Omega,\Omega]$ into the space of complex-valued bounded signals $z(t)$ of spectral support $[0,\Omega]$. We would like the mapping to be linear and also have the property that translates of x map into translates of z, so that no "time stretching" is involved. The usual way of doing this is to take

$$z(t) = x(t) + iy(t) \tag{13}$$

where $y = \tilde{x}$, the Hilbert transform of x.

The Hilbert transform is defined by

$$\tilde{x}(t) = \frac{1}{\pi} \oint_{-\infty}^{\infty} \frac{x(s)}{t-s} ds \tag{14}$$

where the cut in the integral sign indicates a Cauchy principal value at $s = t$. The difficulty we encounter is that an arbitrary bounded band-limited function does not have a Hilbert transform so we have to restrict x somehow.

We may regard the Hilbert transform as the limit as $a \to 0$ of convolution transforms

$$\tilde{x}_a(t) = \int_{-\infty}^{\infty} x(s) K_a(t-s) ds \tag{15}$$

with regular kernels K_a given by

$$K_a(t) = \frac{1}{\pi} \frac{t}{t^2 + a^2}, a > 0. \tag{16}$$

Now $K_a(t)$ belongs to L_p for every $p > 1$ and has a Fourier transform $\hat{K}_a(\omega)$ given by

$$\hat{K}_a(\omega) = -i(\operatorname{sgn} \omega) e^{-a|\omega|}. \tag{17}$$

Thus if $x(t)$ has a Fourier transform $\hat{x}(\omega)$, the Hilbert transform $\tilde{x}(t)$ has a Fourier transform given by

$$\int_{-\infty}^{\infty} \tilde{x}(t) e^{-i\omega t} dt = -i(\operatorname{sgn} \omega) \hat{x}(\omega) \tag{18}$$

and consequently the Fourier transform of $z(t)$ as defined in (13) vanishes over $(-\infty,0)$. Now $z(t)$ is the boundary value of the function $z(\tau)$, $\tau = t + iu$, defined by

$$z(\tau) = \frac{i}{\pi} \int_{-\infty}^{\infty} \frac{x(s)}{\tau - s} ds, \quad u > 0. \tag{19}$$

In case x does not have an ordinary Fourier transform, but has a bounded Hilbert transform $\tilde{x}$, the function $z(\tau)$ is bounded and analytic in the upper half-plane and according to our definition and Theorem 1, the Fourier transform of $z(t)$ vanishes over $(-\infty,0)$. Also if the Fourier transform of $x(t)$ vanishes outside $[-\Omega,\Omega]$ then the Fourier transform of $z(t)$ vanishes outside $[0,\Omega]$.

The subclass of bounded functions which have bounded Hilbert transforms does not have a simple alternate description. In order for $x(t)$ to have a bounded Hilbert transform it is sufficient that $x(t)$ have a bounded derivative and a bounded integral (Ref. 15). If the Fourier transform of x vanishes outside $[-\Omega,\Omega]$ then $x(t)$ has a bounded derivative ("Bernstein's Theorem," cf. Theorem 11.12, Ref. 6) satisfying

$$\sup_t |x'(t)| \leq \Omega \sup_t |x(t)|, \tag{20}$$

and there are various ways to restict $x(t)$ to have a bounded integral. For example, we may assume that $g(t)$ is an arbitrary bandlimited signal and set

$$x(t) = g'(t). \tag{21}$$

(Also any high-pass function has a bounded integral (Ref. 14).) Then we can determine $g(t)$ within an additive constant from the real part of $z(t)$. Assuming that the Fourier transform of g vanishes outside $[-\Omega,\Omega]$ we have the inequality (Theorem 11.4.3, Ref. 6) implied by (21),

$$\sup_t |\tilde{x}(t)| \leq \Omega \sup_t |g(t)|. \tag{22}$$

An interesting subclass of bandlimited functions which have bounded Hilbert transforms are the band-pass functions. For these functions, there are equivalent Hilbert transform kernels which belong to L_1. If the Fourier transform of $x(t)$ vanishes outside the intervals $[-\Omega,-r\Omega]$ and $[r\Omega,\Omega]$ where $0 < r < 1$, then x has a bounded Hilbert transform satisfying (Ref. 7)

$$\sup_t |\tilde{x}(t)| \leq \left\{A + \frac{2}{\pi} \log \frac{1}{r}\right\} \sup_t |x(t)| \tag{23}$$

where $A < 4/\pi$ and $2/\pi$ cannot be replaced by a smaller number (i.e., as r approaches zero).

Other subclasses worthy of mentioning may be generated by convolution transforms on L_∞. Thus if $|g| \leq M$ and k is a kernel in L_1 which has a Hilbert transform also in L_1 then the class of functions of the form

$$x(t) = \int_{-\infty}^{\infty} g(s)k(t-s)ds \tag{24}$$

have bounded Hilbert transforms given by

$$\tilde{x}(t) = \int_{-\infty}^{\infty} g(s)\tilde{k}(t-s)ds. \tag{25}$$

For example, $x(t)$ may be the output of some crude sort of band-pass filter (like an a-c amplifier) as would be the case for

$$\begin{aligned} k(t) &= ae^{-at} - be^{-bt}, \quad t \geq 0 \\ &= 0, \qquad t < 0 \end{aligned} \tag{26}$$

where $a > b > 0$. Then

$$\int_0^{\infty} k(t)e^{-i\omega t}dt = \frac{a}{a+i\omega} - \frac{b}{b+i\omega} = \frac{(a-b)i\omega}{(a+i\omega)(b+i\omega)} \tag{27}$$

and

$$\int_{-\infty}^{\infty} \tilde{k}(t)e^{-i\omega t}dt = \frac{(a-b)|\omega|}{(a+i\omega)(b+i\omega)}. \tag{28}$$

It follows from (28) and (5) that $\tilde{k}$ belongs to L_1.

Another sufficient condition for a bandlimited function $x(t)$ to have a bounded Hilbert transform is the requirement that x belong to L_p where $1 \leq p < \infty$. Then if the Fourier transform of x vanishes outside $[-\Omega,\Omega]$ we have (Ref. 14)

$$x(t) = \int_{-\infty}^{\infty} x(s) \frac{\sin \Omega(t-s)}{\pi(t-s)} ds. \tag{29}$$

Then x has a Hilbert transform given by

$$\tilde{x}(t) = \int_{-\infty}^{\infty} x(s) \frac{1-\cos \Omega(t-s)}{\pi(t-s)} ds \tag{30}$$

and thus by Hölder's inequality

$$|\tilde{x}(t)| \leq \|x\|_p \left\{ \int_{-\infty}^{\infty} \left| \frac{1-\cos \Omega t}{\pi t} \right|^q dt \right\}^{1/q}$$

$$\text{where } \frac{1}{p} + \frac{1}{q} = 1. \tag{31}$$

However, the condition that x belong to L_p $(1 \leq p < \infty)$ is not a satisfactory condition for communication signals.

Hereafter we will suppose that $x(t)$ is so restricted that $z(t) = x(t) + iy(t)$ is a bounded function whose Fourier transform vanishes outside $[0,\Omega]$. We should note that this assumption does not imply that y is the Hilbert transform of x (even within an additive constant). That is, the assumption does not imply that the integral

$$\int_{-T}^{T} \frac{x(t)-x(0)}{t} dt$$

tends to a limit as $T \to \infty$, so in effect we are allowing some functions that are not of the form $z(t) = x(t) + i\tilde{x}(t)$. In case we further restrict $z(t)$ to be a bounded function whose Fourier transform vanishes outside $[r\Omega,\Omega]$ where $0 < r < 1$, then we can assert that $z(t)$ is of the form $x(t) + i\tilde{x}(t)$ where the Fourier transforms of x and $\tilde{x}$ vanish outside $[-\Omega,-r\Omega]$ and $[r\Omega,\Omega]$. Conversely if x is a real-valued (band-pass) function whose Fourier transform vanishes outside these two intervals, then $x(t)$ has a bounded Hilbert transform $\tilde{x}(t)$ and $\{x(t) + i\tilde{x}(t)\}$ is a bounded function whose Fourier transform vanishes outside $[r\Omega,\Omega]$. In other words, we can always regard *any* real-valued band-pass signal $x(t)$ as the real part of an analytic signal $z(t)$. This is a special case of a representation theorem for high-pass signals (Ref. 13).

III. MODULATION AND EQUIVALENT BASEBAND TRANSMISSION

We suppose that $z(t)$ is a bounded bandlimited signal with spectrum confined to $[0,\Omega]$. Now let $f(z)$ be a function which is analytic over a region which includes the disk $|z| \leq m$, where $m = \sup |z(t)|$. We take $f(z)$ as a *modulation law* and generate

$$w(t) = f\{z(t)\} \tag{32}$$

which is the boundary value of a function $w(\tau)$ bounded and analytic in the uhp. Hence the Fourier transform of $w(t)$ vanishes over $(-\infty,0)$. Generally, the modulation process also includes translation of the spectrum of $w(t)$ leading to a transmitter output

$$\sigma(t) = Re\,\{w(t)e^{i\omega_c t}\}. \tag{33}$$

where $\omega_c > 0$ is the carrier frequency. The Fourier transform of $\sigma(t)$ then vanishes over $(-\omega_c,\omega_c)$ and hence $\sigma(t)$ is called a single-sideband signal, although the upper side-band may be infinite in extent.

In conventional single-sideband amplitude modulation (SSBAM) the modulation law is the linear law, $f(z) = z$, in which case the Fourier transform of $\sigma(t)$ vanishes outside the intervals $[\omega_c,\omega_c + \Omega]$ and $[-\omega_c - \Omega, -\omega_c]$, so that the bandwidth required for transmitting $\sigma(t)$ is (counting positive and negative frequencies) $2\Omega + \epsilon$ where ϵ is an arbitrarily small positive number. In other words, $\sigma(t)$ has a reproducing kernel in L_1 of bandwidth slightly larger than 2Ω. (Recall that the Fourier transform of a function of L_1 is continuous.) We will see that the spectral economy of SSBAM carries over to more general modulation laws $f(z)$. So we assume that $\sigma(t)$ is transmitted over a channel (characterized by an L_1 impulse response) which has unity transmission over the frequency bands $[\omega_c,\omega_c + \alpha]$ and $[-\omega_c - \alpha,-\omega_c]$ where $\alpha > \Omega$. The transmission may be zero outside slightly larger intervals.

We denote the received signal by $\sigma_R(t)$ and since its Fourier transform vanishes over $(-\omega_c,\omega_c)$ it has a Hilbert transform $\tilde{\sigma}_R(t)$. We assume that the carrier frequency and phase are known at the receiver so that we can form

$$w_\alpha(t) = e^{-i\omega_c t}\{\sigma_R(t) + i\tilde{\sigma}_R(t)\}. \tag{34}$$

In engineering parlance the real part of $w_\alpha(t)$ is obtained by in-phase synchronous demodulation of $\sigma_R(t)$, while the imaginary part of $w_\alpha(t)$ is obtained by quadrature synchronous demodulation of $\sigma_R(t)$. The Fourier transform of $w_\alpha(t)$ vanishes over $(-\infty,0)$ and we have

$$w_\alpha(t) = \int_{-\infty}^{\infty} w(s)K_\alpha(t-s)ds \tag{35}$$

where $K_\alpha(t)$ is the impulse of an equivalent baseband channel satisfying

for some $\alpha > \Omega$

$$\int_{-\infty}^{\infty} K_\alpha(t)e^{-i\omega t}dt = 1, \quad \text{for } 0 \le \omega \le \alpha \tag{36}$$

$$\int_{-\infty}^{\infty} |K_\alpha(t)|dt < \infty. \tag{37}$$

Hereafter we will be concerned with recovering $z(t)$, and hence $x(t)$, from $w_\alpha(t)$ in the equivalent baseband transmission of $w(t)$ as given by (35).

IV. THE INVERSE FUNCTION AS A DETECTOR

We would like to solve (35) for $z(t)$ where $w(t) = f\{z(t)\}$. This is (superficially) similar to the problem of Landau and Miranker[11,12] where $w(t) = f\{x(t)\}$ and f is a real function of a real variable, $x(t)$ is a real-valued bandlimited function of L_2 whose Fourier transform vanishes outside $[-\Omega,\Omega]$, and $K_\alpha(t) = (\sin \Omega t)/\pi t$. In order for $x(t)$ to be recovered (by an iterative process) they require that f have an inverse over the range of x and that is essentially what we require. The inverse of an analytic f is more complicated, but the fact that the Fourier transforms of $z(t)$ and $f\{z(t)\}$ vanish over $(-\infty,0)$ simplifies the recovery problem.

Let us write

$$w = f(z) \tag{38}$$

and

$$z = \varphi(w) \tag{39}$$

for the inverse and think first of the problem of recovering $z(t)$ from $w(t)$. In case f maps $|z| \le m$ one-one onto some region D^*, there is no problem since φ is single valued over D^*. In general φ is not single valued and we have to know something about $z(\tau)$ in order to decide what element of φ is "the" inverse. For example, suppose $f(z) = 2z + z^2$. Then given

$$w(t) = 2ae^{it} + a^2e^{i2t}$$

we do not know whether $z(t) = ae^{it}$ or $z(t) = -2 - ae^{it}$ without some additional knowledge, such as for example, $\lim_{u\to\infty} z(t + iu) = 0$, or $|z(t)| \le 1$. In this example the inverse function,

$$z = \varphi(w) = -1 + (1 + w)^{1/2}$$

is not a single-valued function of the complex variable w and one generally speaks of two branches of the inverse function. The branches have singularities at $w = -1$, the image of $z = -1$ where $f'(z) = 0$. Clearly in this case, if we require $|z(t)| \le 1$ then we know

$$z(t) = -1 + \sqrt{1 + w(t)}$$

where $\sqrt{1} = 1$. Then $\varphi\{w(t + iu)\}$ is bounded for $u \geq 0$ and analytic for $u > 0$. If we relax the requirement to $|z(t + iu)| \leq 1$ for $u > b$ then $\varphi\{w(t + iu)\}$ will be analytic for $u > b$, but not necessarily for $u > 0$.

In general, *we require that $z(\tau)$ and $f(z)$ are so constrained that $\varphi\{w(\tau)\}$ is analytic for $u \geq u_0$.*

4.1 Received signal in the range of the inverse function

Now we are not given $w(t)$ but instead we have a filtered version $w_\alpha(t)$. Suppose $\varphi\{w(\tau)\}$ is analytic in the upper half-plane $u \geq 0$ and $w_\alpha(t)$ is sufficiently close to $w(t)$ that $\varphi\{w_\alpha(\tau)\}$ is also analytic in the uhp $u \geq 0$. We say then that $w_\alpha(t)$ *is in the range of the inverse function.* A simple sufficient condition for this is that φ be an entire function. Also the channel could have sufficiently large bandwidth for $w_\alpha(t)$ to be close enough to $w(t)$.

We assume then that

$$w_\alpha(t + iu) \in D^*, \quad u \geq 0 \tag{40}$$

where

$$\varphi(w) \text{ is analytic for } w \in D^* \tag{41}$$

$$|\varphi'(w)| \leq M \text{ for } w \in D^* \tag{42}$$

Then we may take the inverse of $w_\alpha(t)$ to obtain

$$z_\alpha(t) = \varphi\{w_\alpha(t)\}. \tag{43}$$

Now we will see that the Fourier transforms of $z_\alpha(t)$ and $z(t)$ agree over $(-\infty,\alpha)$.

First, it follows from (35)–(37) and Ex. 1 of Sec. 2.2 that the Fourier transforms of $w(t)$ and $w_\alpha(t)$ agree over $(-\infty,\alpha)$; i.e., the Fourier transform of $\{w(t) - w_\alpha(t)\}$ vanishes over $(-\infty,\alpha)$. Then from Theorem 1 we have

$$|w(t + iu) - w_\alpha(t + iu)| \leq e^{-\alpha u} \sup_t |w(t) - w_\alpha(t)|. \tag{44}$$

From (42) we have

$$|\varphi(w) - \varphi(w_\alpha)| \leq M|w - w_\alpha| \text{ for } \begin{matrix} w \in D^* \\ w_\alpha \in D^* \end{matrix} \tag{45}$$

Thus $\{z(t) - z_\alpha(t)\}$ is the boundary value of a function bounded and analytic in the uhp satisfying

$$|z(t + iu) - z_\alpha(t + iu)| \leq Me^{-\alpha u} \sup_t |w(t) - w_\alpha(t)|. \tag{46}$$

Hence from Theorem 1,

$$z(t) - z_\alpha(t) = h_\alpha(t) \tag{47}$$

where the Fourier transform of $h_\alpha(t)$ vanishes over $(-\infty,\alpha)$. Since the Fourier transform of $z(t)$ vanishes outside $[0,\Omega]$ and $\alpha > \Omega$, (47) implies that we can bandlimit $z_\alpha(t)$ with an appropriate low-pass filter to obtain $z(t)$. Thus if $K_{\Omega,\alpha}(t)$ is any kernel of L_1 satisfying

$$\int_{-\infty}^{\infty} K_{\Omega,\alpha}(t)e^{-i\omega t}dt = 1, \quad 0 \le \omega \le \Omega$$

$$0, \qquad \omega \ge \alpha \tag{48}$$

we have from Ex. 6, Sec. 2.2, with the convention that $z(t)$ is continuous,

$$z(t) = \int_{-\infty}^{\infty} z(s)K_{\Omega,\alpha}(t-s)ds \tag{49}$$

and since the Fourier transform of h_α vanishes over $(-\infty,\alpha)$ we have $(K_{\Omega,\alpha} \otimes h_\alpha)(t) \equiv 0$; i.e.,

$$z(t) = \int_{-\infty}^{\infty} z_\alpha(s)K_{\Omega,\alpha}(t-s)ds. \tag{50}$$

4.2 Pre-detection filtering

In case the received signal $w_\alpha(t)$ is not in the range of the inverse function, we may under suitable conditions recover $z(t)$ by appropriate filtering before (and after) detection. Here then we replace (40) with the condition

$$w(t+iu) \in D^* \quad \text{for } u \ge u_0 \qquad (\ge 0). \tag{51}$$

It follows from (44) and (42) that for sufficiently large b we have

$$w_\alpha(t+iu) \in D_1^* \quad \text{for } u \ge b \tag{52}$$

where D_1^* is slightly larger than D^* and

$$\varphi(w) \text{ is analytic for } w \in D_1^* \tag{53}$$

$$|\varphi'(w)| < M_1 \text{ for } w \in D_1^*. \tag{54}$$

Then we have $w_\alpha(t+ib)$ in the range of the inverse function.

Now the Poisson kernel with parameter u

$$P_u(t) = \frac{1}{\pi}\frac{u}{t^2+u^2}, \quad u > 0 \tag{55}$$

reproduces functions bounded and analytic in the uhp from their

boundary values. (See the proof of Theorem 1 in Appendix A.) We have

$$w_\alpha(t + ib) = \int_{-\infty}^{\infty} w_\alpha(s)P_b(t - s)ds. \tag{56}$$

That is, we may determine $w_\alpha(\tau)$ along a line $u = b$ parallel to the real axis by convolving $w_\alpha(t)$ with the Poisson kernel (parameter $u = b$). This operation we term *Poisson filtering*. Since, by assumption, $w_\alpha(t + ib)$ is in the range of the inverse we may take the inverse of $w_\alpha(t + ib)$ to obtain

$$z_\alpha(t + ib) = \varphi\{w_\alpha(t + ib)\} \tag{57}$$

which is analytic in the uhp and then as argued before

$$|z(t + iu + ib) - z_\alpha(t + iu + ib)| \leq M_2 e^{-\alpha u}. \tag{58}$$

So the Fourier transforms of $z(t + ib)$ and $z_\alpha(t + ib)$ agree over $(-\infty,\alpha)$.

Thus if the conditions (51), (41), and (42) are met we may by suitable pre-detection filtering (Poisson filtering) obtain a function $z_\alpha(t + ib)$ which corresponds to replacing $z(t)$ at the transmitter by $z(t + ib)$; i.e., from the reproducing property of the Poisson kernel

$$w(t + iu) = \int_{-\infty}^{\infty} w(s)P_u(t - s)ds$$

$$= f\{z(t + iu)\} = \int_{-\infty}^{\infty} f\{z(s)\}P_u(t - s)ds. \tag{59}$$

Then interchanging the order of convolutions with $P_b(t)$ and $K_\alpha(t)$ in (35) we have

$$w_\alpha(t + ib) = \int_{-\infty}^{\infty} f\{z(s + ib)\}K_\alpha(t - s)ds. \tag{60}$$

This relation has been noted by Foschini.[8]

The Poisson kernel is a contraction operator; i.e., it averages the values of a function so that the range of the resultant is no larger than the range of the function. Also as a filter it has the frequency response

$$\int_{-\infty}^{\infty} P_u(t)e^{-i\omega t}dt = e^{-u|\omega|}, \quad (u > 0). \tag{61}$$

We have

$$z(t + ib) = \int_{-\infty}^{\infty} z(s)P_b(t - s)ds \tag{62}$$

and therefore for large b we would expect the range of $z(t + ib)$ to be

appreciably less than the range of $z(t)$ for a wide class of $z(t)$ since for very large b it is only the low-frequency content of $z(t)$ that contributes appreciably to $z(t + ib)$. It is possible, though, for the low-frequency content to be such that, for every $u \geq 0$

$$z(t + iu) \sim \cos \sqrt{t} + i \sin \sqrt{t} \quad \text{as } t \to \infty.$$

If we require $z(t + iu)$ to tend uniformly to a limit z_0 as $u \to \infty$; i.e.,

$$|z(t + iu) - z_0| \leq \epsilon(u), \quad -\infty < t < \infty, \tag{63}$$

where $\epsilon(u) \to 0$ as $u \to \infty$ and in addition

$$f'(z_0) \neq 0, \tag{64}$$

then $\varphi(w)$ will be analytic in the neighborhood of $w_0 = f(z_0)$ and so (51) will be satisfied for sufficiently large u_0 with (41) and (42) holding.

The condition (63) is not a severe constraint. In fact, all the simple sufficient conditions, given in (21)–(29), for $x(t)$ to have a Hilbert transform imply (63) with $z_0 = 0$. However, $x(t)$ may have a Hilbert transform without (63) holding.

In connection with pre-detection filtering, we note that equivalent Poisson filtering can be effected at the carrier frequency (or an intermediate frequency) in the receiver before the synchronous demodulation of the received signal $\sigma_R(t)$ indicated in (34). That is, if the signal $\sigma_R(t)$ is passed through a filter whose frequency response $F_b(\omega)$ satisfies

$$F_b(\omega) = e^{-b(\omega - \omega_c)}, \quad \omega \geq \omega_c \tag{65}$$

$$F_b(-\omega) = \overline{F}_b(\omega)$$

a signal $\sigma_R(t;b)$ is obtained which we may identify as

$$\sigma_R(t;b) = Re\,\{e^{i\omega_c t} w_\alpha(t + ib)\}. \tag{66}$$

Then synchronous in-phase and quadrature detection of $\sigma_R(t;b)$ yields the real and imaginary parts of $w_\alpha(t + ib)$. Thus the Poisson filtering may be accomplished with a single equivalent frequency-translated filter, whereas the direct Poisson filtering of the complex signal $w_\alpha(t)$ requires two Poisson filters acting separately on the real and imaginary parts.

4.3 Post-detection filtering

When Poisson filtering is required to bring the received analytic signal within the range of the inverse function the low pass filtering after detection must be modified. The output of the detector is $z_\alpha(t + ib)$ and we need $z(t)$. Now the Fourier transforms of $z(t + ib)$ and $z_\alpha(t + ib)$ agree over $(-\infty,\alpha)$ and

$$z(t + ib) = \int_{-\infty}^{\infty} z(s) P_b(t - s) ds. \tag{67}$$

The Poisson filtering operation has an inverse for bandlimited functions; i.e.,

$$z(t) = \int_{-\infty}^{\infty} z(s + ib)Q_b(t - s)ds \tag{68}$$

where $Q_b(t)$ is any function of L_1 satisfying

$$\int_{-\infty}^{\infty} Q_b(t)e^{-i\omega t}dt = e^{b\omega}, \quad 0 \leq \omega \leq \Omega. \tag{69}$$

Since the Fourier transforms of $z(t + ib)$ and $z_\alpha(t + ib)$ agree over $(-\infty,\alpha)$, the Fourier transform of

$$z(t) - \int_{-\infty}^{\infty} z_\alpha(s + ib)Q_b(t - s)ds \tag{70}$$

vanishes over $(-\infty,\alpha)$, and since the Fourier transform of $z(t)$ vanishes outside $[0,\Omega]$, we have [cf. (47)–(50)]

$$z(t) = \int_{-\infty}^{\infty} z_\alpha(s + ib)k(t - s)ds \tag{71}$$

where $k(t) \equiv k(t;b,\Omega,\alpha)$ is any kernel in L_1 satisfying

$$\int_{-\infty}^{\infty} k(t)e^{-i\omega t}dt = e^{\omega b}, \quad 0 \leq \omega \leq \Omega \tag{72}$$

$$= 0, \qquad \omega \geq \alpha.$$

That is, the post-detection filtering must invert the pre-detection filtering over the band $[0,\Omega]$ and remove frequencies greater than α. (Of course, in a practical system we are interested in recovering only $x(t)$ so that only one post-detection filter is required, acting on the real part of $z_\alpha(t + ib)$.)

V. SPECIALIZATION TO BAND-PASS SIGNALS

In case the base-band signal $x(t)$ is of the band-pass type, i.e., a signal whose Fourier transform vanishes outside $[r\Omega,\Omega]$ and $[-\Omega,-r\Omega]$ where $0 < r < 1$, the detection theory may be modified so that no Poisson filtering is required. In this case the inverse function may be replaced by an entire function, in particular, a polynomial. All we require for the recovery of band-pass signals is that $f[z(\tau)]$ be analytic in the uhp and

$$f'(0) \neq 0. \tag{73}$$

Then

$$w = f(z) = \sum_{k=0}^{\infty} a_k z^k \quad \text{for } |z| \text{ sufficiently small} \tag{74}$$

$$z = \varphi(w) = \sum_{k=1}^{\infty} b_k (w - a_0)^k \quad \text{for } |w - a_0| \text{ sufficiently small.} \tag{75}$$

Here $b_1 = (a_1)^{-1} = [f'(0)]^{-1}$.

Now for band-pass signals $x(t)$, the Fourier transform of the analytic signal $z(t)$ vanishes over $(-\infty, r\Omega)$ where $\Omega > 0$ and $0 < r < 1$. Thus by the Paley-Wiener Theorem for L_∞,

$$|z(t + iu)| \leq e^{-r\Omega u} \sup_t |z(t)|, \quad u \geq 0. \tag{76}$$

Then we have

$$w(t + iu) - a_0 = \sum_{k=1}^{\infty} a_k \{z(t + iu)\}^k \quad \text{for sufficiently large } u. \tag{77}$$

It follows that the Fourier transform of $\{w(t) - a_0\}$ also vanishes over $(-\infty, r\Omega)$. Hence the Fourier transform of $\{w_\alpha(t) - a_0\}$ vanishes over $(-\infty, r\Omega)$.

Now let $\varphi^*(w)$ be any entire function of the form

$$\varphi^*(w) = \sum_{k=1}^{\infty} c_k (w - a_0)^k \tag{78}$$

where

$$c_k = b_k, \quad \text{for } k = 1,2, \ldots, n \tag{79}$$

and n is an integer such that

$$nr \geq 1. \tag{80}$$

Defining

$$z_\alpha^*(t) = \varphi^*\{w_\alpha(t)\} \tag{81}$$

we have $z_\alpha^*(\tau)$ analytic in the uhp and for sufficiently large u

$$\begin{aligned} z(t + iu) - z_\alpha^*(t + iu) &= \varphi\{w(t + iu)\} - \varphi^*\{w_\alpha(t + iu)\} \\ &= \sum_{k=1}^{n} b_k [\{w(t + iu) - a_0\}^k - \{w_\alpha(t + iu) - a_0\}^k] \\ &\quad + \sum_{k=n+1}^{\infty} b_k \{w(t + iu) - a_0\}^k \\ &\quad - \sum_{k=n+1}^{\infty} c_k \{w_\alpha(t + iu) - a_0\}^k. \end{aligned} \tag{82}$$

Since the Fourier transforms of $\{w(t) - a_0\}$ and $\{w_\alpha(t) - a_0\}$ vanish over $(-\infty, r\Omega)$, the last two sums in (82) are of the order of $\exp\{-(n + 1)r\Omega u\}$.

Also, for $k \geq 1$

$$\begin{aligned} &|\{w(t+iu)-a_0\}^k - \{w_\alpha(t+iu)-a_0\}^k| \\ &\quad = O\{|w(t+iu)-w_\alpha(t+iu)|\} \\ &\quad = O(e^{-\alpha u}). \end{aligned} \tag{83}$$

So it follows from the Paley-Wiener Theorem (with Theorem 3) that the Fourier transforms of $z(t)$ and $z_\alpha^*(t)$ agree over $(-\infty,B)$ where

$$B = \min\{\alpha,(n+1)r\Omega\} > \Omega. \tag{84}$$

Therefore

$$z(t) = \int_{-\infty}^{\infty} z_\alpha^*(s) K_{\Omega,B}(t-s)\,ds \tag{85}$$

where $K_{\Omega,B}$ is any function of L_1 satisfying

$$\begin{aligned} \int_{-\infty}^{\infty} K_{\Omega,B}(t)e^{-i\omega t}dt &= 1, \quad 0 \leq \omega \leq \Omega \\ &= 0, \qquad \omega \geq B. \end{aligned} \tag{86}$$

Thus for band-pass signals we have the option of replacing the inverse function $\varphi(w)$ by an equivalent entire function $\varphi^*(w)$ so that it does not matter whether or not the received analytic signal $w_\alpha(t)$ is in the range of the inverse function $\varphi(w)$. In particular, $\varphi^*(w)$ may be a polynomial of degree n where n is roughly the ratio of the upper and lower cut-off frequencies of the base-band signal.

VI. DETECTION OF EXPONENTIAL MODULATION

The exponential modulation law $f(z) = e^z$ offers the unique advantage of eliminating the need for preliminary in-phase and quadrature detection of the received single-sideband signal $\sigma_R(t)$. In this case we have

$$z = \varphi(w) = \log w \tag{87}$$

or using Log to denote the real part of the logarithm,

$$x(t) = \text{Log}\,|w(t)| \tag{88}$$

$$y(t) = \arg\{w(t)\}. \tag{89}$$

We may regard either $x(t)$ or $y(t)$ as the signal to be recovered. The transmitted signal is

$$\sigma(t) = Re\,\exp\{i\omega_c t + z(t)\} \tag{90}$$

and

$$\sigma(t) + i\tilde{\sigma}(t) = \exp\{i\omega_c t + z(t)\} \tag{91}$$

where $\tilde{\sigma}(t)$ denotes the Hilbert transform of $\sigma(t)$. The envelope of $\sigma(t)$ is

$$E\{\sigma(t)\} = |\sigma(t) + i\tilde{\sigma}(t)| = e^{x(t)} \tag{92}$$

The instantaneous phase of $\sigma(t)$ is

$$A\{\sigma(t)\} = \arg\{\sigma(t) + i\tilde{\sigma}(t)\} = \omega_c t + y(t) \tag{93}$$

For perfect transmission; i.e., $\sigma_R(t) = \sigma(t)$, we have

$$x(t) = \text{Log } E\{\sigma(t)\} \tag{94}$$

and using an ideal discriminator (FM detector) we obtain

$$y'(t) = \frac{d}{dt} A\{\sigma(t)\} - \omega_c. \tag{95}$$

Replacing $\sigma(t)$ by $\sigma_R(t)$ we have

$$x_\alpha(t) = \text{Log } E\{\sigma_R(t)\} = \text{Log } |w_\alpha(t)| \tag{96}$$

$$y'_\alpha(t) = \frac{d}{dt} A\{\sigma_R(t)\} - \omega_c = \frac{d}{dt} \text{Arg } \{w_\alpha(t)\}. \tag{97}$$

Now if $w_\alpha(\tau)$ is zero-free in the uhp, then

$$z_\alpha(\tau) = \log w_\alpha(\tau) \tag{98}$$

is analytic in the uhp and by the previous theory the Fourier transforms of $z(t)$ and $z_\alpha(t)$ agree over $(-\infty,\alpha)$. In this case the Fourier transforms of $x_\alpha(t)$ and $x(t)$ agree over $(-\alpha,\alpha)$. Also the Fourier transforms of $y'_\alpha(t)$ and $y'(t)$ agree over $(-\alpha,\alpha)$. So if the received analytic signal $w_\alpha(t)$ is zero-free in the uhp, $x(t)$ may be recovered by taking the Log of the envelope of the received signal and then filtering with an ideal low-pass filter having unity transmission in the band $[-\Omega,\Omega]$ and zero transmission outside the band $[-\alpha,\alpha]$. Similarly $y'(t)$ may be recovered by filtering the output of an ideal discriminator acting on the received signal.

Later, in examining the bandwidth requirements of single-sideband exponential modulation we give sufficient conditions for $w_\alpha(t)$ to be zero-free in the uhp so that the simple detectors described above may be used.

The simple detectors can always be used with appropriate pre-detection and post-detection filters, since for sufficiently large b, $w_\alpha(\tau + ib)$ will be zero-free in the uhp. We can give an estimate for b under the condition

$$\sup_t |z(t)| \leq m. \tag{99}$$

We have

$$e^{-m} \leq |w(t)| \leq e^m \tag{100}$$

and consequently

$$e^{-m} \leq |w(t + iu)| \leq e^{m}, \quad u \geq 0. \tag{101}$$

Also, $w(\tau)$ is zero-free in the uhp and hence [cf. (219)] if

$$|w(t + ib) - w_\alpha(t + ib)| < e^{-m} \tag{102}$$

then $w_\alpha(\tau + ib)$ will be zero-free in the uhp. From (44) we have

$$|w(t + ib) - w_\alpha(t + ib)| \leq e^{-\alpha b} \sup_t |w(t) - w_\alpha(t)| \tag{103}$$

and since

$$w_\alpha(t) = \int_{-\infty}^{\infty} w(s)K_\alpha(t - s)ds \tag{104}$$

we have

$$\sup_t |w_\alpha(t)| \leq \sup_t |w(t)| \cdot \|K_\alpha\|_1 \leq e^{m}\|K_\alpha\|_1 \tag{105}$$

and hence

$$\sup |w(t) - w_\alpha(t)| \leq \{1 + \|K_\alpha\|_1\}e^{m}. \tag{106}$$

Thus if

$$e^{-\alpha b}\{1 + \|K_\alpha\|_1\} < e^{-2m} \tag{107}$$

then (102) will be satisfied. That is, $w_\alpha(\tau + ib)$ will be zero-free in the uhp for

$$b > \frac{2m + \log\{1 + \|K_\alpha\|_1\}}{\alpha} \tag{108}$$

Thus if (99) is satisfied we may use a frequency-translated Poisson filter with parameter b satisfying (108) to obtain $\sigma_R(t;b)$ [cf. (66)] and then operate on the envelope and phase of $\sigma_R(t;b)$ as before. Then the appropriate post-detection filtering may be employed to recover $x(t)$ and $y'(t)$.

VII. NOTE ON THE FACTORIZATION OF CERTAIN POSITIVE FUNCTIONS

The detection theory for SSBEM has important application to the problem of factoring certain positive functions of exponential type, i.e., certain positive bandlimited functions.

Voelcker[22] has proposed a scheme for demodulating conventional single-sideband signals via envelope detection. Conventional SSBAM is characterized by linear modulation; i.e., $f(z) = z$. There is no bandwidth expansion so we assume that the Fourier transform of $z(t)$ vanishes

outside $[0,\Omega]$ and the channel is such that the received signal is simply the transmitted signal; i.e.,

$$\sigma_R(t) = \sigma(t) = Re\, e^{i\omega_c t} z(t). \tag{109}$$

The envelope of the received signal is $|z(t)|$. Voelcker's scheme requires first that $z(\tau)$ be zero-free in the uhp in order that $z(t)$ may be recovered from $|z(t)|$. This is insured by requiring $Re\, z(t) = x(t) > 0$. We have

$$z(t + iu) = \int_{-\infty}^{\infty} z(s) P_u(t - s) ds \tag{110}$$

and since the Poisson kernel is positive,

$$Re\, z(t + iu) = \int_{-\infty}^{\infty} x(s) P_u(t - s) ds > 0. \tag{111}$$

Therefore $z(\tau)$ is zero-free in the uhp. The function $\log z(\tau)$ is analytic in the uhp and with some additional conditions on $z(\tau)$, e.g.,

$$\lim_{u \to \infty} z(t + iu) = 1, \tag{112}$$

the imaginary part of $\log z(t)$ can be determined from $\text{Log}\, |z(t)|$ and hence $z(t)$ can be recovered from $|z(t)|$. In particular, if

$$x(t) = 1 + g(t) \tag{113}$$

where

$$g(t) > -1 \text{ and } g(t) \text{ belongs to } L_p\ (1 \leq p < \infty), \tag{114}$$

then $\text{Log}\, |z(t)|$ will belong to L_p and will therefore have a Hilbert transform. A more attractive condition for recovering $x(t)$ is the condition

$$g(t) > -1 \text{ and } g(t) \text{ of band-pass type.} \tag{115}$$

Then if (115) is satisfied, the Fourier transform of $\{z(t) - 1\}$ vanishes outside $[r\Omega,\Omega]$, where $0 < r < 1$, and hence

$$w(\tau) = \log z(\tau) \tag{116}$$

is analytic in the uhp and satisfies

$$\begin{aligned} w(t + iu) &= \log [1 + \{z(t + iu) - 1\}] \\ &= O\{|z(t + iu) - 1|\} = O(e^{-r\Omega u}), \qquad u \to \infty. \end{aligned} \tag{117}$$

Therefore, if (115) is satisfied, the Fourier transform of $w(t)$ vanishes over $(-\infty,r\Omega)$ and hence the Fourier transforms of $\text{Log}\, |z(t)|$ and $\arg \{z(t)\}$ vanish over $(-r\Omega,r\Omega)$. That is, *if (115) is satisfied, then the log of the envelope of $z(t)$ and the phase (arg) of $z(t)$ are high-pass functions.*

If either (114) or (115) are satisifed we have

$$\arg\{z(t)\} \equiv \varphi(t) = \oint_{-\infty}^{\infty} \frac{\text{Log}\,|z(s)|}{\pi(t-s)}\,ds \tag{118}$$

and then

$$z(t) = |z(t)|e^{i\varphi(t)}. \tag{119}$$

The practical problem encountered here is in approximating the Hilbert transform in (118). The function Log $|z(t)|$ is not band-limited† and the implementation of (118) requires a filter whose frequency characteristic is, formally,

$$H(\omega) = \int_{-\infty}^{\infty} \frac{e^{-i\omega t}}{\pi t}\,dt = -i\int_{-\infty}^{\infty} \frac{\sin \omega t}{\pi t}\,dt = -i\,\text{sgn}\,\omega. \tag{120}$$

$$(-\infty < \omega < \infty)$$

These stringent filter requirements can be avoided, for we can, by proper application of the previous theory, ignore the frequency content of Log $|z(t)|$ outside the band $(-\alpha,\alpha)$ where $\alpha > \Omega$. Actually if (114) is satisfied we may take $\alpha = \Omega$.

The Hilbert transform problem is simplified if we begin with a filtered version of Log $|z(t)|$; viz.,

$$\lambda_\alpha(t) = \int_{-\infty}^{\infty} \text{Log}\,|z(s)|h_\alpha(t-s)ds \tag{121}$$

where $h_\alpha(t)$ is an even real-valued function whose Fourier transform satisfies

$$\hat{h}_\alpha(\omega) = \int_{-\infty}^{\infty} h_\alpha(t)e^{-i\omega t}dt = 1 \quad \text{for } -\alpha \le \omega \le \alpha. \tag{122}$$

We suppose further that $h_\alpha(t)$ is sufficiently smooth to have a Hilbert transform $\tilde{h}_\alpha(t)$. Then the Hilbert transform of $\lambda_\alpha(t)$ is given by

$$\tilde{\lambda}_\alpha(t) \equiv \varphi_\alpha(t) = \int_{-\infty}^{\infty} \text{Log}\,|z(s)|\tilde{h}_\alpha(t-s)ds \tag{123}$$

$$= \int_{-\infty}^{\infty} \varphi(s)h_\alpha(t-s)ds.$$

Then defining

$$w_\alpha(t) = \lambda_\alpha(t) + i\varphi_\alpha(t) \tag{124}$$

we have

$$w_\alpha(t) = \int_{-\infty}^{\infty} \log\{z(s)\}h_\alpha(t-s)ds. \tag{125}$$

† Unless $z(t)$ = constant. See Theorem 6 in Section IX for a stronger statement.

Then according to the previous theory

$$z(t) = \int_{-\infty}^{\infty} k(t-s) \exp\{w_\alpha(s)\} ds \tag{126}$$

where $k(t)$ is any function of L_1 satisfying

$$\int_{-\infty}^{\infty} k(t) e^{-i\omega t} dt = 1, \quad 0 \le \omega \le \Omega \tag{127}$$

$$= 0, \qquad \omega \ge \alpha.$$

In case (115) is satisfied the function $h_\alpha(t)$ need satisfy (122) only over the intervals $(r\Omega,\alpha)$ and $(-\alpha,-r\Omega)$, since the Fourier transform of Log $|z(t)|$ vanishes over $(-r\Omega,r\Omega)$. That is, $h_\alpha(t)$ can then be chosen so that the equivalent Hilbert transform kernel $\tilde{h}_\alpha(t)$ has a Fourier transform that is more easily approximated (within a linear phase factor $e^{-i\omega T}$) by practical filters.

VIII. NOTE ON LOGARITHMIC COMPANDING

Suppose $g(t)$ is a function belonging to L_p for some p satisfying $1 \le p < \infty$ and suppose the Fourier transform of $g(t)$ vanishes outside $[-\Omega,\Omega]$. Companding functions f are sometimes used to compress the range of $g(t)$ for transmission; i.e. $f\{g(t)\}$ is transmitted rather than $g(t)$. Landau and Miranker[11,12] showed that $g(t)$ (in L_2) can be recovered from the bandlimited version of $f\{g(t)\}$ with suitable conditions on f. The recovery is accomplished by an iterative scheme. Here we use the detection theory to give an explicit solution to the problem of Landau and Miranker for the case

$$f(x) = \frac{1}{2} \text{Log } (1+x), \quad x > -1. \tag{128}$$

Accordingly, we further require $g(t)$ to satisfy

$$g(t) > -1. \tag{129}$$

The function $f(x)$ given by (128) is not an odd function, as one might desire for companding purposes, but is interesting because the recovery problem is simple.

The fact that $g(t)$ is a bandlimited function belonging to L_p for some p satisfying $1 \le p < \infty$ implies [cf. (29)]

$$g(t) = \int_{-\infty}^{\infty} g(s) \frac{\sin \Omega(t-s)}{\pi(t-s)} ds \tag{130}$$

from which one can conclude with the aid of Hölder's inequality that

$$\lim_{t \to \pm\infty} g(t) = 0. \tag{131}$$

It follows from (129) and (131) and (128) that

$$f\{g(t)\} \text{ belongs to } L_p. \tag{132}$$

Since $g(t)$ is bounded and belongs to L_p it follows that $g(t)$ belongs to $L_{p'}$ for every p' satisfying $p \leq p' \leq \infty$. Hence $f\{g(t)\}$ also belongs to $L_{p'}$ for such p'.

Now we suppose we are given

$$\lambda_\Omega(t) = \int_{-\infty}^{\infty} f\{g(s)\} \frac{\sin \Omega(t-s)}{\pi(t-s)} ds \tag{133}$$

where the integral is absolutely convergent by Hölder's inequality. In fact (Ref. 16) $\lambda_\Omega(t)$ belongs to L_p, and therefore has a Hilbert transform. Furthermore, since $\lambda_\Omega(t)$ is bandlimited, its Hilbert transform is given by

$$\begin{aligned} \bar{\lambda}_\Omega(t) \equiv \varphi_\Omega(t) &= \int_{-\infty}^{\infty} \lambda_\Omega(s) \frac{1-\cos \Omega(t-s)}{\pi(t-s)} ds \\ &= \int_{-\infty}^{\infty} f\{g(s)\} \frac{1-\cos \Omega(t-s)}{\pi(t-s)} ds. \end{aligned} \tag{134}$$

Defining

$$w_\Omega(t) = \lambda_\Omega(t) + i\varphi_\Omega(t) \tag{135}$$

we have

$$w_\Omega(t) = \int_{-\infty}^{\infty} f\{g(s)\} K_\Omega(t-s) ds \tag{136}$$

where

$$K_\Omega(t) = \frac{e^{i\Omega t} - 1}{i\pi t}. \tag{137}$$

So $w_\Omega(\tau)$ is an entire function which is bounded in the uhp.

Now $\{1 + g(t)\}$ is a positive bandlimited function which can be represented as (Theorem 7.5.1 with Theorem 6.4.5, Ref. 6)

$$1 + g(t) = \gamma(t)\overline{\gamma}(t) \tag{138}$$

where the Fourier transform of $\gamma(t)$ vanishes outside $[-\Omega/2, \Omega/2]$ and $\gamma(\tau)$ is zero-free in the uhp. Then $z(t)$ defined by

$$z(t) = \gamma(t) e^{i\Omega t/2} \tag{139}$$

is a function whose Fourier transform vanishes outside $[0, \Omega]$ and $z(\tau)$ is zero-free in the uhp. Thus we have

$$1 + g(t) = |z(t)|^2. \tag{140}$$

We may assume that [cf. (131)]

$$\lim_{t\to\pm\infty} z(t) = 1. \tag{141}$$

Then $z(t)$ is given by

$$z(t) = \exp w(t) \tag{142}$$

where

$$w(\tau) = \frac{i}{2}\int_{-\infty}^{\infty} \operatorname{Log}\{1 + g(s)\}\frac{ds}{\pi(\tau - s)} \tag{143}$$

and

$$\lim_{u\to 0+} w(t + iu) = \lambda(t) + i\varphi(t) \tag{144}$$

$$\lambda(t) = \frac{1}{2}\operatorname{Log}\{1 + g(t)\} \tag{145}$$

$$\varphi(t) = \tilde{\lambda}(t), \text{ the Hilbert transform of } \lambda(t). \tag{146}$$

We see from (136) and (143) that

$$w_\Omega(t) = \int_{-\infty}^{\infty} w(s)K_\Omega(t - s)ds \tag{147}$$

$$= \int_{-\infty}^{\infty} \log\{z(s)\}K_\Omega(t - s)ds.$$

Then the Fourier transform of $z_\Omega(t)$ defined by

$$z_\Omega(t) = \exp\{w_\Omega(t)\} \tag{148}$$

agrees over $(-\infty,\Omega)$ with the Fourier transform of $z(t)$. Since $\{z(t) - 1\}$ belongs to L_p and its Fourier transform vanishes outside $[0,\Omega]$ we have

$$z(t) - 1 = \int_{-\infty}^{\infty} \{z_\Omega(s) - 1\}\frac{\sin \Omega(t - s)}{\pi(t - s)}ds. \tag{149}$$

Writing $z(t) = x(t) + iy(t)$ we have

$$x(t) = 1 + \int_{-\infty}^{\infty} \{e^{\lambda_\Omega(s)}\cos \varphi_\Omega(s) - 1\}\frac{\sin \Omega(t - s)}{\pi(t - s)}ds \tag{150}$$

$$y(t) = \int_{-\infty}^{\infty} e^{\lambda_\Omega(s)}\sin \varphi_\Omega(s)\frac{\sin \Omega(t - s)}{\pi(t - s)}ds \tag{151}$$

Then we have

$$g(t) = x^2(t) + y^2(t) - 1. \tag{152}$$

Thus the recovery problem is solved by means of the Hilbert transform in (134) and the formulas (150)–(152).

Expressing the solution in terms of the bandlimiting operator $\mathcal{B}_\Omega$, defined for h in L_p, $1 \le p < \infty$, by†

$$\mathcal{B}_\Omega h(t) = \int_{-\infty} h(s) \frac{\sin \Omega(t-s)}{\pi(t-s)}\, ds, \tag{153}$$

the resulting class of functions denoted by $B_p(\Omega)$, we have

$$g(t) = |\mathcal{B}_\Omega \exp\{\lambda_\Omega(t) + i\tilde{\lambda}_\Omega(t)\}|^2 - 1 \tag{154}$$

where $\lambda_\Omega(t)$ is the given function

$$\lambda_\Omega(t) = \frac{1}{2}\mathcal{B}_\Omega \operatorname{Log}\{1 + g(t)\} \tag{155}$$

$$g \text{ in } B_p(\Omega), \quad g > -1,$$

and $\tilde{\lambda}_\Omega$ is the Hilbert transform of λ_Ω.

The solution (154) is deceptive in that it suggests that $\lambda_\Omega(t)$ may be any function of $B_p(\Omega)$, $1 \le p < \infty$, since $g(t)$ given by (154) is a function of $B_p(\Omega)$ satisfying $g > -1$. However the solution was obtained on the premise that $\lambda_\Omega(t)$ is a given function of the form (155). All functions in $B_p(\Omega)$ do not have the representation (155). The crucial point is that the function

$$z(t) = \mathcal{B}_\Omega \exp\{\lambda_\Omega(t) + i\tilde{\lambda}_\Omega(t)\}, \quad \{z(t) - 1\} \text{ in } B_p(\Omega), \tag{156}$$

whose Fourier transform vanishes outside $[0,\Omega]$ should extend as a function zero-free in the upper half-plane. Then, and only then, according to the general theory, will we have

$$\mathcal{B}_\Omega \log\{z(t)\} = \lambda_\Omega(t) + i\tilde{\lambda}_\Omega(t) \tag{157}$$

and hence

$$\mathcal{B}_\Omega \operatorname{Log}|z(t)| = \frac{1}{2}\mathcal{B}_\Omega \operatorname{Log}\{1 + g(t)\} = \lambda_\Omega(t), \quad g \text{ in } B_p(\Omega), \tag{158}$$

$$g > -1.$$

On the other hand, if (158) is known to hold, implying (157), then $z(t)$ must necessarily extend as a function zero-free in the upper half-plane.

We state this important result as

Theorem 4. Given a function $\lambda_\Omega(t)$ in $B_p(\Omega)$, for some p satisfying $1 \le$

† The operator $\mathcal{B}_\Omega$ can be extended to certain other classes of functions. For example, $\mathcal{B}_\Omega$ is an identity for the constant function, which fact is used in (154).

$p < \infty$, the equation (155) has a solution $g(t)$ in the same class $B_p(\Omega)$ satisfying

$$g(t) > -1$$

if and only if the function

$$z(t) = \mathcal{B}_\Omega \exp\{\lambda_\Omega(t) + i\tilde{\lambda}_\Omega(t)\}$$

where $\tilde{\lambda}_\Omega(t)$ is the Hilbert transform of $\lambda_\Omega(t)$, extends as a function zero-free in the upper half-plane. Then the solution of (155) is given by (154).

IX. BANDWIDTH REQUIREMENTS FOR EXPONENTIAL MODULATION

We have seen that for a wide class of analytic signals $z(t)$ and modulation laws $f(z)$ the bandwidth requirement for transmitting $f\{z(t)\}$ and recovering $z(t)$ is $\Omega + \epsilon$ (for any $\epsilon > 0$) where Ω is the bandwidth of $z(t)$, *provided* we allow the use of Poisson filtering at the receiver. In case the inverse function $z = \varphi(w)$ is an entire function there is no need for Poisson filtering. If we look at the overall system design, as contrasted to a detection problem, it is reasonable to ask for the bandwidth requirements for a given $f(z)$ and a *fixed* receiver, namely *the inverse function $\varphi(w)$ followed by a low-pass filter*, such that we recover all $z(t)$ whose Fourier transforms vanish outside $[0,\Omega]$ and satisfy some sort of norm constraint, say $|z(t)| \leq m$. The problem then is to specify a channel of finite bandwidth, i.e., a function $K_{\alpha,\beta}(t)$ in L_1, with $\Omega < \alpha < \beta$, satisfying

$$\hat{K}_{\alpha,\beta}(\omega) = \int_{-\infty}^{\infty} K_{\alpha,\beta}(t)e^{-i\omega t}dt = 1, \quad 0 \leq \omega \leq \alpha \tag{159}$$

$$= 0, \qquad \omega > \beta$$

such that the received (bandlimited) analytic signal $w_{\alpha,\beta}(t)$, given by

$$w_{\alpha,\beta}(t) = \int_{-\infty}^{\infty} f\{z(s)\}K_{\alpha,\beta}(t - s)ds, \tag{160}$$

satisfies

$$\varphi\{w_{\alpha,\beta}(\tau)\} \text{ analytic in the uhp} \tag{161}$$

for all $z(t)$ whose Fourier transforms vanish outside $[0,\Omega]$ and which satisfy

$$|z(t)| \leq m, \quad -\infty < t < \infty. \tag{162}$$

We assume of course that $f(z)$ is analytic for $|z| \leq m$. We would like to make the channel bandwidth β as small as possible consistent with (161) and (162). Clearly, we may take $\Omega = 1$ with no loss in generality. We de-

fine the minimum bandwidth $\beta_0(m)$ as

$$\beta_0(m) = \inf \beta(m) \tag{163}$$

where the infimum is over all functions $K_{\alpha,\beta}(t)$ subject to (159), (161), and (162) with $\Omega = 1$. In taking the infimum we may allow $\alpha = 1$.

The determination of $\beta_0(m)$ is in general a very difficult problem. We give some estimates here for $\beta_0(m)$ for the case $f(z) = e^z$, $\varphi(w) = \log w$. In this case, (161) is satisfied if and only if

$$w_{\alpha,\beta}(\tau) \text{ is zero-free in the uhp.} \tag{164}$$

We have

$$w(t) = \exp\{z(t)\} \tag{165}$$

and with (162)

$$e^{-m} \leq |w(t)| \leq e^m. \tag{166}$$

If $w_{\alpha,\beta}(t)$ is sufficiently close to $w(t)$, (164) will be satisfied; i.e., a sufficient condition for (164) is

$$|w(t) - w_{\alpha,\beta}(t)| < e^{-m}. \tag{167}$$

It is intuitively obvious, with the freedom we have in defining $\hat{K}_{\alpha,\beta}$, that for sufficiently large β we can find a function $K_{\alpha,\beta}(t)$ such that $w_{\alpha,\beta}(t)$ given by (160), with $f(z) = e^z$, will satisfy (167). It is important to note in this connection that, although the definition of $\hat{K}_{\alpha,\beta}(\omega)$ for $\omega < 0$ does not affect $w_{\alpha,\beta}(t)$, we are free to define $\hat{K}_{\alpha,\beta}(\omega)$ for $\omega < 0$ (as well as for $\alpha < \omega < \beta$) in the most favorable way to obtain the estimate (167).

First we obtain lower bounds for $\beta_0(m)$.

9.1 Lower bounds for $\beta_0(m)$

We can obtain a lower bound for $\beta_0(m)$ by taking

$$z(t) = me^{it}. \tag{168}$$

We have

$$w(t) = \exp\{me^{it}\} = \sum_{k=0}^{\infty} \frac{m^k}{k!} e^{ikt}. \tag{169}$$

Now assume the channel cutoff frequency β satisfies

$$n < \beta \leq n + 1 \tag{170}$$

where $n \geq 1$ is an integer. Then

$$w_{\alpha,\beta}(t) = 1 + me^{it} + \sum_{k=2}^{n} a_k e^{ikt} \tag{171}$$

where the a_k depend on m and the definition of $\hat{K}_{\alpha,\beta}(\omega)$ for $1 < \omega \leq n$. We have

$$w_{\alpha,\beta}(t) = \prod_{k=1}^{n} \left(1 - \frac{\zeta}{\zeta_k}\right) \tag{172}$$

where $\zeta = e^{it}$ and

$$\sum_{k=1}^{n} \frac{1}{\zeta_k} = -m. \tag{173}$$

We require $w_{\alpha,\beta}(t)$ to be zero-free in the uhp; i.e.,

$$|\zeta_k| \geq 1. \tag{174}$$

Thus

$$m = \left|\sum_{k=1}^{n} \frac{1}{\zeta_k}\right| \leq \sum_{k=1}^{n} \frac{1}{|\zeta_k|} \leq n. \tag{175}$$

Therefore, we must have $\beta > n \geq m$ in order for $w_{\alpha,\beta}(t)$ to be zero-free in the uhp. Then

$$\beta_0(m) > [m]^+ \tag{176}$$

where $[m]^+$ is the smallest integer which is not less than m.

9.2 Lower bound for small m

We know that $\beta_0(m) > 1$ for any $m > 0$ but (176) does not say how much $\beta_0(m)$ must exceed 1 for $0 < m < 1$. For sufficiently small ϵ and correspondingly small m we can show that $\beta_0(m) > 1 + \epsilon$.

For small m, we have

$$w(t) = 1 + z(t) + \frac{z^2(t)}{2} + O(m^3). \tag{177}$$

Now by Corollary 2 of Theorem 2 the Fourier transform of $z^2(t)$ vanishes outside [0,2] and we would like to find a $z(t)$ such that a channel filter with a sharp cut-off, i.e., $\beta = 1 + \epsilon$, acting on a small $z^2(t)$ gives a large (negative) output at $t = 0$. For sufficiently small ϵ and fixed m we can accomplish this by taking

$$z(t) = -\frac{me^{it/2}}{\sqrt{2}}\{1 + iS_n(\epsilon t)\} \tag{178}$$

where $S_n(t)$ is a sine polynomial,

$$S_n(t) = \sum_{k=1}^{n} a_k \sin kt \tag{179}$$

with real coefficients a_k and

$$\max |S_n(t)| = 1. \tag{180}$$

Also we require

$$n\epsilon = \frac{1}{2} \tag{181}$$

so that the Fourier transform of $z(t)$ given by (178) vanishes outside [0,1]. Also we have

$$\max |z(t)| = m \tag{182}$$

and

$$w(t) = 1 - \frac{me^{it/2}}{\sqrt{2}}\{1 + iS_n(\epsilon t)\} + \frac{m^2}{4}e^{it}\{1 - S_n^2(\epsilon t) + 2iS_n(\epsilon t)\} + R_3(t), \tag{183}$$

where

$$R_3(t) = \sum_{k=3}^{\infty} \frac{z^k(t)}{k!} \tag{184}$$

$$|R_3(t)| \leq \frac{m^3}{3!}\frac{1}{\left(1 - \frac{m}{4}\right)}, \qquad (m < 4). \tag{185}$$

We have

$$w_{\alpha,\beta}(t) = \int_{-\infty}^{\infty} w(\xi)K_{\alpha,\beta}(t - \xi)d\xi = 1 - \frac{me^{it/2}}{\sqrt{2}}\{1 + S_n(\epsilon t)\} + \frac{m^2}{4}\int_{-\infty}^{\infty} e^{i\xi}\{1 - S_n^2(\epsilon\xi) + 2iS_n(\epsilon\xi)\}K_{\alpha,\beta}(t - \xi)d\xi + \int_{-\infty}^{\infty} R_3(\xi)K_{\alpha,\beta}(t - \xi)d\xi \tag{186}$$

and $w_{\alpha,\beta}(t)$ is a polynomial of degree $2n$ in exp $(i\epsilon t)$ where $\epsilon = 1/2n$. We have

$$2ie^{it}S_n(\epsilon t) = -e^{it}\sum_{k=1}^{n} a_k e^{-ik\epsilon t} + e^{it}\sum_{k=1}^{n} a_k e^{ik\epsilon t} \tag{187}$$

$$e^{it}S_n^2(\epsilon t) = e^{it}\sum_{k=-2n}^{2n} b_k e^{ik\epsilon t}. \tag{188}$$

So

$$w_{\alpha,\beta}(t) = 1 - \frac{me^{it/2}}{\sqrt{2}}\{1 + iS_n(\epsilon t)\}$$
$$+ \frac{m^2}{4}\left\{1 - e^{it}\sum_{k=-2n}^{0} b_k e^{ik\epsilon t} - e^{it}\sum_{k=1}^{n} a_k e^{-ik\epsilon t}\right\} \tag{189}$$
$$+ r_3(t)$$

where

$$r_3(t) = \int_{-\infty}^{\infty} R_3(\xi)K_{\alpha,\beta}(t - \xi)d\xi. \tag{190}$$

Since $R_3(t)$ is a periodic function of the form $\Sigma_0^\infty c_k e^{-ik\epsilon t}$ we may take for $K_{\alpha,\beta}(t)$ any function of L_1 whose Fourier transform satisfies

$$\begin{aligned} \hat{K}_{\alpha,\beta}(\omega) &= 1, \qquad 0 \le \omega \le 1 \\ &= 0, \qquad \omega \ge \beta = 1 + \epsilon \end{aligned} \tag{191}$$

It is shown in Appendix B that there exists a function $K_{\alpha,\beta}(t)$ whose Fourier transform satisfies (191) with

$$\int_{-\infty}^{\infty} |K_{\alpha,\beta}(t)|dt < 1 + \frac{1}{\pi}\log\left(1 + \frac{4}{3\epsilon}\right). \tag{192}$$

Thus

$$|r_3(t)| \le \max |R_3(t)| \int_{-\infty}^{\infty} |K_{\alpha,\beta}(t)|dt \tag{193}$$
$$< \frac{m^3}{3!\left(1 - \frac{m}{4}\right)}\left\{1 + \frac{1}{\pi}\log\left(1 + \frac{4}{3\epsilon}\right)\right\}.$$

We have from (189)

$$w_{\alpha,\beta}(0) = 1 - \frac{m}{\sqrt{2}} + \frac{m^2}{4}\left\{1 - \sum_{k=-2n}^{0} b_k - \sum_{k=1}^{n} a_k\right\} \tag{194}$$
$$+ r_3(0).$$

Now $S_n^2(t)$ is an even function and $S_n(0) = 0$. Thus

$$\sum_{k=-2n}^{2n} b_k = 0, \qquad b_{-k} = b_k \tag{195}$$

$$\sum_{k=-2n}^{-1} b_k = \sum_{k=1}^{2n} b_k = \frac{-b_0}{2} \tag{196}$$

$$\sum_{k=-2n}^{0} b_k = \frac{b_0}{2} \tag{197}$$

and

$$b_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_n^2(t)dt = \frac{1}{2}\sum_{k=1}^{n} a_k^2 < 1. \tag{198}$$

Now we may take $S_n(t)$ to be an approximation to $\operatorname{sgn}|\sin t|$. In particular we may take $S_n(t)$ to be the function found by Szegö[19] which maximizes $\Sigma_1^n a_k$ subject to (180). He gives the inequality

$$\sum_{1}^{n} a_k \le M_n = \frac{2}{n+1}\sum_{\substack{1\le k\le n\\ k \text{ odd}}} \cot\left(\frac{k\pi}{2(n+1)}\right) \sim \frac{2}{\pi}\log n, \quad n \to \infty. \tag{199}$$

Here we should not identify a_k with the terms in the second sum. For equality in (199) we must have

$$S_n\left(\frac{k\pi}{n+1}\right) = 1, \quad 1 \le k \le n, \quad k \text{ odd}. \tag{200}$$

Equality in (199) is attained for

$$S_n(t) = \sum_{\substack{1\le k\le n\\ k \text{ odd}}} \left\{I_n\left(t - \frac{k\pi}{n+1}\right) - I_n\left(t + \frac{k\pi}{n+1}\right)\right\} \tag{201}$$

where

$$I_n(t) = \left\{\frac{\sin\dfrac{n+1}{2}t}{(n+1)\sin\dfrac{t}{2}}\right\}^2. \tag{202}$$

It is easy to show that

$$\sum_{\substack{k=-n\\ (k \text{ odd})}}^{n} I_n\left(t - \frac{k\pi}{n+1}\right) \equiv 1 \quad \text{for } n \text{ odd} \tag{203}$$

$$\sum_{\substack{k=-(n-1)\\ (k \text{ odd})}}^{n-1} I_n\left(t - \frac{k\pi}{n+1}\right) + \frac{1}{2}\{I_n(t-\pi) + I_n(t+\pi)\} \tag{204}$$

$$\equiv 1 \qquad \text{for } n \text{ even}$$

It follows from (203), (204), and (201) that $S_n(t)$ given by (201) satisfies

$$-1 \le S_n(t) \le 1. \tag{205}$$

From (201) and (202) we find that

$$a_k = \frac{4}{n+1}\left(1 - \frac{k}{n+1}\right)\frac{\sin^2 \frac{k\pi}{2}}{\sin \frac{k\pi}{n+1}}, \qquad \begin{matrix} n \text{ odd} \\ k = 1,2,\ldots,n \end{matrix} \tag{206}$$

$$a_k = \frac{4}{n+1}\left(1 - \frac{k}{n+1}\right)\frac{\left(\sin \frac{nk\pi}{2(n+1)}\right)^2}{\sin \frac{k\pi}{n+1}}, \qquad \begin{matrix} n \text{ even} \\ k = 1,2,\ldots,n \end{matrix} \tag{207}$$

It is shown in Appendix C that M_n given by (199) satisfies

$$M_n > \frac{2}{\pi}\log n + \frac{2}{\pi}\left(\log \frac{4}{\pi} + \gamma\right) \tag{208}$$

where

$$\gamma = 0.5772\ldots \qquad \text{(Euler's constant)}. \tag{209}$$

So with $S_n(t)$ given by (201) and $z(t)$ given by (178) we have from (193), (194), (197), and (208)

$$w_{\alpha,\beta}(0) < 1 - \frac{m}{\sqrt{2}} + \frac{m^2}{4}\left\{1 - \frac{b_0}{2} - \frac{2}{\pi}\log n - \frac{2}{\pi}\left(\log \frac{4}{\pi} + \gamma\right)\right\} + \frac{m^3}{3!\left(1 - \frac{m}{4}\right)}\left\{1 + \frac{1}{\pi}\log\left(1 + \frac{8n}{3}\right)\right\} \tag{210}$$

where $1 < \alpha < \beta$,

$$\beta = 1 + \frac{1}{\epsilon} = 1 + \frac{1}{2n}$$

$$b_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_n^2(t)dt \sim 1 \text{ as } n \to \infty.$$

Now if we set

$$\frac{m^2}{2\pi}\log n = 1 \tag{211}$$

it is clear that

$$w_{\alpha,\beta}(0) < 0 \text{ for sufficiently large } n. \tag{212}$$

Since

$$w_{\alpha,\beta}(t) = \sum_{k=0}^{2n} c_k e^{ikt/2n} \tag{213}$$

where the c_k are real and $c_0 > 0$, we have $w_{\alpha,\beta}(\mathrm{iu})$ real and

$$\lim_{u \to \infty} w_{\alpha,\beta}(\mathrm{iu}) = c_0 > 0. \tag{214}$$

So (212) and (214) imply that $w_{\alpha,\beta}(t)$ has at least one zero on the positive imaginary axis; i.e. for sufficiently large n, and m given by (211), we must have $\beta > 1 + 1/2n$ (to pick up at least another harmonic) in order for $w_{\alpha,\beta}(t)$ to be zero free in the uhp. Thus

$$\beta_0\left(\sqrt{\frac{2\pi}{\log n}}\right) > 1 + \frac{1}{2n} \text{ for sufficiently large } n. \tag{215}$$

In connection with obtaining lower bounds for β_0 the idea comes to mind that we might be able to find a $z(t)$ satisfying $|z(t)| \leq m$ (for sufficiently large m) such that the Fourier transform of $w(t)$ would vanish over a large interval $(1,\beta)$. Then if $w_{\alpha,\beta}(t)$ were not zero free in the uhp we would have $\beta_0 > \beta(m)$. The idea is to obtain a $w_{\alpha,\beta}(t)$ which would be independent of the choice of $K_{\alpha,\beta}(t)$. However, we cannot make the Fourier transform of $w(t)$ vanish over large intervals unless $z(t) \equiv$ constant.

Theorem 5. Suppose $z(t)$ is a bounded continuous function whose Fourier transform vanishes outside $[0,\Omega]$, and suppose that the Fourier transform of $w(t)$, where

$$w(t) = \exp\{z(t)\},$$

vanishes over (a,b) where

$$a \geq 0 \quad \text{and} \quad b - a > \Omega.$$

Then

$$z(t) \equiv \text{constant}.$$

A similar result holds for the logarithmic function.

Theorem 6. Suppose $z(t)$ is a bounded continuous function whose Fourier transform vanishes outside $[0,\Omega]$ and suppose its analytic continuation $z(\tau)$ satisfies

$$|z(t + iu)| \geq \epsilon > 0 \quad \text{for } u \geq 0$$
$$-\infty < t < \infty.$$

Suppose further that the Fourier transform of $w(t)$, where

$$w(t) = \log\{z(t)\}$$

vanishes over (a,b) where

$$a \geq 0, \quad b - a > \Omega.$$

Then

$$z(t) \equiv \text{constant}.$$

As an application of the last theorem, we may take $\Omega = n$ and

$$z(t) = \prod_{k=1}^{n} (1 - \lambda_k e^{ikt})$$

Then, assuming $|\lambda_k| < 1$, we have

$$\log z(t) = -\sum_{m=1}^{\infty} \frac{\mu_m e^{imt}}{m}$$

where

$$\mu_m = \sum_{k=1}^{n} (\lambda_k)^m$$

Then the following is true.

Corollary. If $\{\lambda_k\}$, $k = 1,2, \ldots, n$, *is any set of n complex numbers and*

$$\sum_{k=1}^{n} (\lambda_k)^m = 0 \text{ for } m = p, p+1, p+2, \ldots p+n-1$$

where p is a positive integer, then $\lambda_k = 0$, $k = 1,2, \ldots, n$.

Proofs of Theorems 5 and 6 are given in Appendices D and E.

9.3 Upper bound for $\beta_0(m)$

We have

$$w(t) = e^{z(t)} \quad \text{and} \quad |z(t)| \leq m \tag{216}$$

so $\{w(t)\}^{-1}$ is bounded and analytic in the uhp. Thus the quotient

$$\frac{w_{\alpha,\beta}(t)}{w(t)} = 1 + \frac{w_{\alpha,\beta}(t) - w(t)}{w(t)} \tag{217}$$

is bounded and analytic in the uhp, and is reproduced by the Poisson kernel from its values on the real line. Then if

$$\left| \frac{w(t) - w_{\alpha,\beta}(t)}{w(t)} \right| < 1 \text{ for } -\infty < t < \infty \tag{218}$$

the function $w_{\alpha,\beta}(t)$ is necessarily zero-free in the uhp. Thus a sufficient condition for $w_{\alpha,\beta}(t)$ to be zero-free in the uhp is

$$|w(t) - w_{\alpha,\beta}(t)| < e^{-m}. \tag{219}$$

For lack of something better we will use this condition to obtain an upper bound for $\beta_0(m)$. To meet this condition for large m will require con-

siderably more bandwidth than the lower bound for $\beta_0(m)$. In fact for the case $z(t) = me^{it}$ we have

$$w(t) = \sum_{k=0}^{\infty} \frac{m^k}{k!} e^{ikt} \tag{220}$$

and

$$w_{\alpha,\beta}(t) = 1 + me^{it} + \sum_{2}^{[\beta]} a_k e^{ikt} \tag{221}$$

where the a_k depend on the definition of $\hat{K}_{\alpha,\beta}(\omega)$ for $\omega > 1$. Since

$$\frac{m^k}{k!} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \{w(t) - w_{\alpha,\beta}(t)\} e^{-ikt} dt, \quad k \geq \beta \tag{222}$$

we have

$$\frac{m^k}{k!} \leq \max |w(t) - w_{\alpha,\beta}(t)| \quad \text{where } k \geq \beta. \tag{223}$$

Hence in order to satisfy (219) for the case $z(t) = me^{it}$ we must have

$$\max_{k \geq \beta} \left\{ \frac{m^k}{k!} \right\} < e^{-m}. \tag{224}$$

For large m we must have β large since

$$k! < \sqrt{2\pi k}\, k^k e^{-k} e^{1/12k}. \tag{225}$$

We find for large m that

$$\beta > \rho m + o(m) \tag{226}$$

where ρ is the root of

$$\rho = e^{1+1/\rho} \doteq 3.591121477.$$

Thus for large m the upper bound we obtain for $\beta_0(m)$ from the condition (219) must be something like 3.6 times as large as the lower bound ($\sim m$) we obtained previously. We can in fact obtain an upper bound for $\beta_0(m)$ that is close to ρm for large m and, as it turns out, is close to the lower bound for $\beta_0(m)$ as $m \to 0$.

To do this we suppose that

$$\alpha = n < \beta \tag{227}$$

where n is a positive integer. We take

$$\begin{aligned} \hat{K}_{\alpha,\beta}(\omega) &= 1, \quad 0 \leq \omega \leq n \\ &= \frac{\beta - \omega}{\beta - n}, \quad n < \omega \leq \beta \\ &= 0, \quad \omega > \beta. \end{aligned} \tag{228}$$

$\hat{K}_{\alpha,\beta}(\omega)$ is defined for $\omega < 0$ in such a way (see Appendix B) that

$$\int_{-\infty}^{\infty} |K_{\alpha,\beta}(t)|dt < \frac{1}{\pi}\log\left\{1 + \frac{4n}{3(\beta - n)}\right\} + 1. \tag{229}$$

Now we write

$$w(t) = e^{z(t)} = P_n\{z(t)\} + \sum_{k=n+1}^{\infty} \frac{\{z(t)\}^k}{k!} \tag{230}$$

where

$$P_n\{z(t)\} = \sum_{k=0}^{n} \frac{\{z(t)\}^k}{k!}. \tag{231}$$

Since by Corollary 2 the Fourier transform of $P_n\{z(t)\}$ vanishes outside $[0,n]$, we have

$$\begin{aligned} w_{\alpha,\beta}(t) &= \int_{-\infty}^{\infty} w(s)K_{\alpha,\beta}(t - s)ds \\ &= P_n\{z(t)\} + R_{n+1}(t) \end{aligned} \tag{232}$$

where

$$R_{n+1}(t) = \int_{-\infty}^{\infty} \left\{\sum_{k=n+1}^{\infty} \frac{z^k(s)}{k!}\right\} K_{\alpha,\beta}(t - s)ds \tag{233}$$

$$|R_{n+1}(t)| < \left\{\sum_{k=n+1}^{\infty} \frac{m^k}{k!}\right\} \int_{-\infty}^{\infty} |K'_{\alpha,\beta}(t)|dt. \tag{234}$$

Thus

$$|w(t) - w_{\alpha,\beta}(t)| < \left\{\sum_{k=n+1}^{\infty} \frac{m^k}{k!}\right\} \left\{1 + \int_{-\infty}^{\infty} |K_{\alpha,\beta}(t)|dt\right\}. \tag{235}$$

Therefore if

$$1 + \int_{-\infty}^{\infty} |K_{\alpha,\beta}(t)|dt < \frac{e^{-m}}{\sum_{k=n+1}^{\infty} \frac{m^k}{k!}} \equiv Q_n(m) \tag{236}$$

$w_{\alpha,\beta}(t)$ will be zero-free in the uhp. Then from (229) and (236) we can get an upper bound on β for each choice of n. Since

$$\int_{-\infty}^{\infty} K_{\alpha,\beta}(t)e^{-i\omega t}dt = 1 \quad \text{for } 0 \le \omega \le \alpha \tag{237}$$

it follows that

$$\int_{-\infty}^{\infty} |K_{\alpha,\beta}(t)|dt > 1. \tag{238}$$

So the inequality (236) can hold only if m and n are such that the right-hand member of (236) is greater than 2. Then setting

$$A_n(m) = \max\{0,[Q_n(m) - 2]\} \tag{239}$$

we see from (229) and (236) that

$$\frac{1}{\pi}\log\left\{1 + \frac{4n}{3[\beta_n(m) - n]}\right\} = A_n(m) \tag{240}$$

defines for each positive integer n an upper bound for $\beta_0(m)$, viz.,

$$\beta_n(m) = n + \frac{\frac{4}{3}n}{e^{\pi A_n(m)} - 1}. \tag{241}$$

We take $\beta_n(m) = \infty$ for $A_n(m) = 0$. For fixed m, we have $A_n(m) > 0$ for sufficiently large n. Then

$$\beta_0(m) \leq B(m) \equiv \min_n \beta_n(m) \quad n = 1,2,3,\ldots \tag{242}$$

Since $\beta_n(m) > n$, it is clear that the minimum in (242) will be $\beta_1(m)$ for sufficiently small m. We have

$$A_1(m) = 2m^{-2} + O(m^{-1}), \quad m \to 0. \tag{243}$$

So the upper bound $\beta_1(m)$ for small m compares favorably with the lower bound (215). At least we have the dominant exponential behavior pinned down as $m \to 0$; i.e.,

$$\lim_{m\to 0} m^2 \log\{\beta_0(m) - 1\} = -2\pi. \tag{244}$$

The function $A_n(m)$ defined in (239) behaves like $(n + 1)!/m^{n+1}$ as $m \to 0$ and decreases to zero at $m = m_n$ where

$$m_n \approx \frac{(n+1)}{\rho} + \frac{1}{2(1+\rho)}\log(n+1) + \frac{1}{1+\rho}\left\{\frac{1}{2}\log 2\pi + \log\frac{\rho - 1}{2\rho}\right\} \tag{245}$$

and $\rho \doteq 3.591121477$ is defined in (226),

$$\frac{1}{\rho} \doteq .278464543$$

$$\frac{1}{2(1+\rho)} \doteq .1089058525$$

$$\frac{1}{1+\rho}\left\{\frac{1}{2}\log 2\pi + \log\frac{\rho - 1}{2\rho}\right\} \doteq -.0219080253.$$

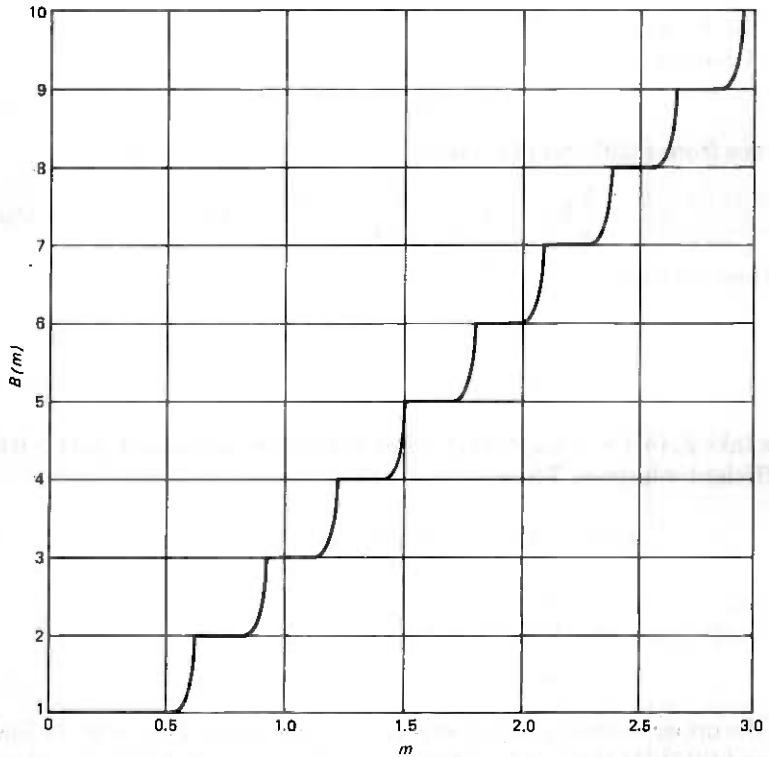


Fig. 1—Upper bound for transmission bandwidth required in ESSB for simple detection of signals $z(t)$ whose Fourier transforms vanish outside $[0,1]$ and satisfy $|z(t)| \leq m$.

The behavior of $A_n(m)$ is such that $\beta_n(m)$ defined in (241) for $0 \leq m \leq m_n$ is very close to n over most of this range and increases suddenly as $m \to m_n$. Consequently, the upper bound $B(m)$ defined in (242) is roughly a staircase function as shown in Figure 1. For $0 \leq m \leq .62$ we have $B(m) = \beta_1(m)$. In Figure 2 a graph of $\log_{10}\{B(m) - 1\}$ is plotted for $.48 \leq m \leq .62$. It is seen that only .1% increase in transmission bandwidth is required for $m \leq .48$, and 10% increase suffices for $m \leq .57$. We know that $\beta_0(m) > 2$ for $m > 1$, so without Poisson filtering SSBEM is interesting perhaps only for $|z| < .6$.

X. A POLYNOMIAL PROBLEM

To each polynomial

$$\bar{P}_n(\zeta) = 1 + \sum_{k=1}^{n} a_k \zeta^k \tag{246}$$

we can assign a positive integer ν which is the smallest integer $\geq n$ such

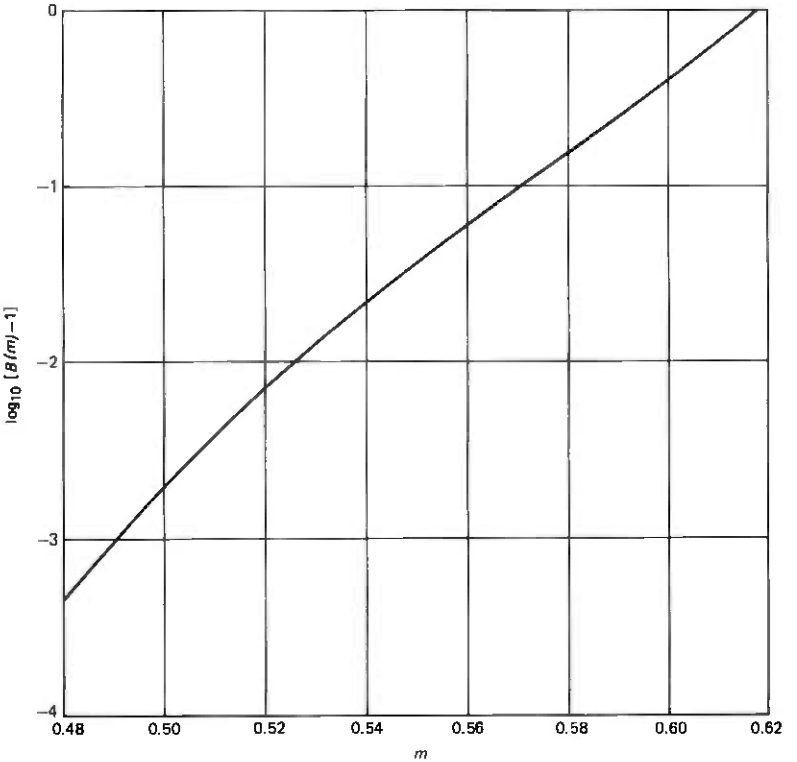


Fig. 2—Logarithmic expansion of Fig. 1 for small m.

that for some choice of a_k for $k = n + 1, n + 2, \ldots, \nu$, the polynomial

$$P_{n,\nu}(\zeta) = 1 + \sum_{k=1}^{n} a_k \zeta^k + \sum_{n+1}^{\nu} a_k \zeta^k \tag{247}$$

is zero free for $|\zeta| < 1$. The integer ν is some complicated function of the coefficients a_k, $k = 1,2, \ldots n$. The fact that ν is finite is a rather remarkable fact that follows from the previous theory. To see this we set

$$\zeta = re^{it}, \quad r > 0. \tag{248}$$

For sufficiently small r we have $P_n(re^{it})$ zero-free in the uhp. Then taking

$$Q(re^{it}) = \log P_n(re^{it}) \tag{249}$$

$$= \sum_{1}^{\infty} b_k r^k e^{ikt}$$

and

$$Q_n(re^{it}) = \sum_1^n b_k r^k e^{ikt} \tag{250}$$

we have

$$\exp\{Q_n(e^{it})\} = 1 + \sum_1^\infty c_k e^{ikt} \tag{251}$$

where

$$c_k = a_k, \quad k = 1,2,\ldots,n. \tag{252}$$

Now we identify $Q_n(e^{it/n})$ with $z(t)$ in the previous section where we obtained upper bounds on $\beta_0(m)$. We know we can bandlimit $\exp\{Q_n(e^{it})\}$ to obtain a function of the form

$$1 + \sum_1^n c_k e^{ikt} + \sum_{n+1}^N d_k e^{ikt}$$

which for sufficiently large N is zero free in the uhp. In particular, we have shown that this is possible for

$$N \leq nB(m) \tag{253}$$

where $B(m)$ is defined in (242) and

$$m = \max_t |Q_n(e^{it})|. \tag{254}$$

Thus

$$\nu \leq N \leq nB(m). \tag{255}$$

Given the a_k, or equivalently the b_k, for $k = 1,2,\ldots,n$, we are interested in obtaining a lower bound for ν.

Writing

$$P_{n,\nu}(\zeta) = \prod_{k=1}^\nu (1 - \lambda_k \zeta) \tag{256}$$

$$\text{where } |\lambda_k| \leq 1$$

we have

$$\log P_{n,\nu}(\zeta) = \sum_{k=1}^\infty b_k \zeta^k, \quad |\zeta| < 1 \tag{257}$$

where

$$b_k = -\frac{1}{k} \sum_{j=1}^\nu (\lambda_j)^k. \tag{258}$$

Since $|\lambda_k| \leq 1$ we have $|b_k| \leq \nu/k$ and hence

$$\nu \geq \max_k |kb_k|, \quad k = 1,2,\ldots,n. \tag{259}$$

Consider, for example, the case

$$P_n(\zeta) = 1 + m\zeta^n, \quad m \geq 0 \tag{260}$$

$$P_{n,\nu}(\zeta) = 1 + m\zeta^n + a_{n+1}\zeta^{n+1} + \ldots a_\nu \zeta^\nu. \tag{261}$$

We have

$$\log P_{n,\nu}(\zeta) = m\zeta^n + b_{n+1}\zeta^{n+1} + \ldots \tag{262}$$

So for the case (260) we have

$$\nu \geq nm. \tag{263}$$

This inequality is clearly best possible in case m is a positive integer. Now suppose $m = 1 + \epsilon$, where $0 < \epsilon < 1/n$. Since ν is an integer we conclude from (263) that $\nu \geq n + 1$. However, we can show (see Appendix F) by another method that $m > 1$ in (260) implies $\nu \geq 2n$. Then it is apparent from the example

$$P_{n,2n}(\zeta) = 1 + m\zeta^n + a_{2n}\zeta^{2n}$$

(with appropriate choice of a_{2n}) that $\nu = 2n$ for $1 < m \leq 2$. It is conjectured that this large jump in ν at $m = 1$ also occurs at all integer values of m, but we have not been able to show, for example, that $m > 2$ implies $\nu \geq 3n$.

In order to improve the lower bounds obtained for $\beta_0(m)$ we are interested in maximizing the ratio ν/n subject to the constraint

$$\max_t \left| b_o + \sum_{rf1}^{n} b_k e^{ikt} \right| \leq m. \tag{264}$$

For any choice of b_k, $k = 1,2,\ldots n$, we are free to choose b_o so as to minimize the maximum modulus of the sum. That is, in the bandwidth problem of the previous section we take $\Omega = n$ and

$$z(t) = \sum_{k=0}^{n} b_k e^{ikt} \tag{265}$$

with the constraint (257). Then assuming $n \leq \alpha \leq n + 1$, and $\nu < \beta \leq \nu + 1$, we have

$$w_{\alpha,\beta}(t) = e^{b_o}\left\{1 + \sum_{k=1}^{n} a_k e^{ikt} + \sum_{k=n+1}^{\nu} a_k e^{ikt}\right\} \tag{266}$$

$$= e^{b_o} \prod_{k=1}^{\nu} (1 - \lambda_k e^{it}), \quad |\lambda_k| \leq 1.$$

Now we want lower bounds on ν implied by (258), subject to (264), giving us $\beta_0(m) > \nu/n$. For a given n (large) we would like to choose the b_k so as to maximize ν. For a particular choice of b_k we can in principle determine the minimum value of ν required to satisfy (258). In order to do this we may assign values to the $(n - \nu)$ coefficients a_k, $k = n + 1, \ldots, \nu$, in (264) and see if it is possible to make $|\lambda_k| \leq 1$ for $k = 1,2, \ldots, \nu$. Recall that the first n coefficients are determined by

$$\exp\left\{\sum_{k=1}^{n} b_k e^{ikt}\right\} = 1 + a_1 e^{it} + \ldots + a_n e^{int} + \ldots \tag{267}$$

Perhaps a computer study could shed some light on this very difficult and challenging problem.

APPENDIX A

Proofs of Theorems 1, 2, and 3

In view of Ex. 2 in Sec. 2.2 it is sufficient to prove Theorem 1 for the interval $(-\infty,0)$. So we assume first that $g(t)$ is a bounded function whose Fourier transform vanishes over $(-\infty,0)$ and we wish to show that $g(t)$ is the boundary value of a function bounded and analytic in the upper half-plane. For this purpose we define

$$g_u(t) = \int_{-\infty}^{\infty} g(s) P_u(t - s) ds, \quad u > 0 \tag{268}$$

where

$$P_u(t) = \frac{u}{\pi} \frac{1}{t^2 + u^2}. \tag{269}$$

We have

$$|g_u(t)| \leq \int_{-\infty}^{\infty} |g(s)||P_u(t - s)| ds.$$

Hence

$$\sup_t |g_u(t)| \leq \sup_t |g(t)|. \tag{270}$$

Also

$$\lim_{u \to 0} g_u(t) = g(t) \quad \text{for almost all } t. \tag{271}$$

Now all we have to show is that $g_u(t)$ is an analytic function of $\tau = t + iu$. Since the Fourier transform of $g(t)$ vanishes over $(-\infty,0)$ we may replace $P_u(t)$ in (268) by any function of L_1 whose Fourier transform agrees (for each $u > 0$) over $(0,\infty)$ with that of $P_u(t)$. In particular, we

may replace $P_u(t)$ by the analytic kernel,

$$K(t+iu) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{i\omega(t+iu)}d\omega, \quad u>0 \tag{272}$$

$$\text{where } F(\omega) = 1 \quad \text{for} \quad \omega \geq 0$$

$$= \omega + 1 \quad \text{for} \quad -1 \leq \omega < 0$$

$$= 0 \quad \text{for} \quad \omega < -1.$$

The definition of $F(\omega)$ for $\omega < 0$ is important only to the extent that the integral in (272) converges for $u > 0$ and implies

$$\int_{-\infty}^{\infty} |K(t+iu)|dt < \infty \quad \text{for} \quad u > 0. \tag{273}$$

It is sufficient for (273) that $e^{-\omega u}F(\omega)$ belong to L_2 and have a derivative in L_2 [see (5)].

Thus we have

$$g_u(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} g(s)K(t+iu-s)ds = G(t+iu) \tag{274}$$

where $G(\tau)$ is analytic in the uhp and from (270) and (271)

$$|G(\tau)| \leq \sup_t |g(t)| \tag{275}$$

$$\lim_{u\to 0} G(t+iu) = g(t) \quad \text{for almost all } t. \tag{276}$$

This proves the first half of the theorem and we may as well write

$$G(t+iu) = g(t+iu) = g_u(t). \tag{277}$$

Now for the second half of the theorem we wish to establish that if $g(\tau)$ is bounded and analytic in the uhp, then

$$\int_{-\infty}^{\infty} g(t)h(-t)dt = 0 \tag{278}$$

for all functions h of L_1 whose Fourier transforms vanish over $(0,\infty)$, or equivalently

$$\int_{-\infty}^{\infty} g(t)h(t)dt = 0 \tag{279}$$

for all functions h of L_1 whose Fourier transforms vanish over $(-\infty,0)$. To do this we need some lemmas concerning analytic functions belonging to L_1 on lines parallel to the real axis.

Lemma 1. If $h(t)$ belongs to L_1 and its Fourier transform vanishes over

$(-\infty,0)$ then $h(t)$ has an analytic continuation $h(t+iu)$ in the upper half-plane $u>0$ satisfying

$$\int_{-\infty}^{\infty} |h(t+iu)|dt \le \int_{-\infty}^{\infty} |h(t)|dt. \tag{280}$$

Proof. We have

$$h(t) = \int_0^{\infty} \hat{h}(\omega)e^{i\omega t}d\omega \tag{281}$$

with the Fourier integral providing the analytic continuation

$$h(t+iu) = \int_0^{\infty} \hat{h}(\omega)e^{i\omega(t+iu)}d\omega, \quad u>0. \tag{282}$$

Since $\hat{h}(\omega)=0$ for $\omega \le 0$ we may write

$$h(t+iu) = \int_{-\infty}^{\infty} \hat{h}(\omega)e^{-u|\omega|}e^{i\omega t}d\omega, \quad u>0 \tag{283}$$

and hence conclude that

$$h(t+iu) = \int_{-\infty}^{\infty} h(s)P_u(t-s)ds \tag{284}$$

where $P_u(t)$ is the Poisson kernel defined in (269). Then (280) follows from (284), since the L_1 norm of the Poisson kernel is 1 for every $u>0$.

Lemma 2. Suppose $h(\tau)$, $\tau = t+iu$, is analytic in the strip $0<u<b$ and satisfies

$$\int_{-\infty}^{\infty} |h(t+iu)|dt < \infty \quad \text{for} \quad 0 \le u < b. \tag{285}$$

Then

$$h(t+iu) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{h}(\omega)e^{i\omega(t+iu)}d\omega, \quad 0 \le u < b \tag{286}$$

where

$$\hat{h}(\omega) = \int_{-\infty}^{\infty} h(t)e^{-i\omega t}dt. \tag{287}$$

Proof: Defining

$$\hat{h}(\omega,u) = \int_{-\infty}^{\infty} h(t+iu)e^{-i\omega t}dt, \quad 0 \le u < b \tag{288}$$

we wish to show that

$$\hat{h}(\omega;u) = e^{-\omega u}\hat{h}(\omega). \tag{289}$$

We have

$$h(t+iu) = \lim_{\Omega\to\infty} \frac{1}{2\pi} \int_{-\Omega}^{\Omega} \left\{1 - \frac{|\omega|}{\Omega}\right\} \hat{h}(\omega;u) e^{i\omega t} d\omega. \tag{290}$$

Also, from the analyticity of $h(\tau)$, we have

$$\frac{\partial}{\partial t} h(t+iu) = \frac{1}{i} \frac{\partial}{\partial u} h(t+iu), \quad 0 < u < b. \tag{291}$$

We would like to establish (289) from (291) by differentiating inside the integral of (290) but at this point we do not know enough about $\hat{h}(\omega;u)$ to justify the differentiation. Therefore, we will define

$$g(t+iu) = \int_{-\infty}^{\infty} k(s) h(t+iu-s) ds \tag{292}$$

where $k(t)$ is a function of L_1 whose Fourier transform $\hat{k}(\omega)$ does not vanish for any argument and such that $\omega\hat{k}(\omega)$ belongs to L_1. We would also like $k'(t)$ to belong to L_1. We may take

$$k(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \tag{293}$$

Then $g(\tau)$ is analytic in the strip and

$$\int_{-\infty}^{\infty} |g(t+iu)| dt \leq \int_{-\infty}^{\infty} |h(t+iu)| dt. \tag{294}$$

We have

$$g(t+iu) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{k}(\omega) \hat{h}(\omega;u) e^{i\omega t} d\omega, \quad 0 \leq u < b, \tag{295}$$

and since $\omega\hat{k}(\omega)$ is in L_1,

$$\frac{\partial}{\partial t} g(t+iu) = \frac{1}{2\pi} \int_{-\infty}^{\infty} i\omega \hat{k}(\omega) \hat{h}(\omega;u) e^{i\omega t} d\omega, \quad 0 \leq u < b. \tag{296}$$

Now

$$g'_u(t) \equiv \frac{\partial}{\partial t} g(t+iu) \equiv \frac{\partial}{i\partial u} g(t+iu) \tag{297}$$

belongs to L_1 for $0 < u < b$ since

$$g'_u(t) = \int_{-\infty}^{\infty} h(s+iu) k'(t-s) ds \tag{298}$$

and k' belongs to L_1. Hence the function of t

$$\frac{\partial}{\partial u} g(t+iu)$$

has a Fourier transform for $0 < u < b$. Thus from (295)

$$\frac{\partial}{\partial u} g(t + iu) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{k}(\omega) \frac{\partial}{\partial u} \hat{h}(\omega;u) e^{i\omega t} d\omega, \quad 0 < u < b. \quad (299)$$

Since $\hat{k}(\omega) \neq 0$ for $-\infty < \omega < \infty$, we conclude from (296), (297), and (299) that

$$\frac{1}{i} \frac{\partial}{\partial u} \hat{h}(\omega;u) = i\omega \hat{h}(\omega;u). \quad (300)$$

Then (289) follows from (300).

A corollary of Lemma 2 is the following

Corollary. If $g(\tau)$, $\tau = t + iu$, is analytic in the strip $a < u < b$ and satisfies

$$\int_{-\infty}^{\infty} |g(t + iu)| dt < \infty \quad \text{for} \quad a < u < b, \quad (301)$$

then

$$\int_{-\infty}^{\infty} g(t + iu) dt = \text{constant}, \quad \text{for} \quad a < u < b. \quad (302)$$

The corollary follows by applying Lemma 2 to the function $g(t + ia + i\epsilon)$ for arbitrarily small positive ϵ.

Lemma 3. If $g(\tau)$, $\tau = t + iu$, is analytic in the upper half plane $u > 0$ and satisfies

$$\int_{-\infty}^{\infty} |g(t + iu)| dt < \infty \quad \text{for} \quad u \geq 0$$

then the asymptotic estimate

$$\int_{-\infty}^{\infty} |g(t + iu)| dt = O\{e^{-au}\} \quad \text{as} \quad u \to \infty \quad (303)$$

implies

$$\int_{-\infty}^{\infty} |g(t + iu)| dt \leq e^{-au} \int_{-\infty}^{\infty} |g(t)| dt \quad \text{for} \quad u \geq 0 \quad (304)$$

and

$$g(t + iu) = \int_{-\infty}^{\infty} \hat{g}(\omega) e^{i\omega(t+iu)} d\omega \quad \text{for} \quad u > 0 \quad (305)$$

where

$$\hat{g}(\omega) = \int_{-\infty}^{\infty} e^{-i\omega t} g(t) dt = 0 \quad \text{for} \quad \omega \leq a. \quad (306)$$

Proof: From Lemma 2 we have the representation (305) so that

$$\hat{g}(\omega)e^{-\omega u} = \int_{-\infty}^{\infty} g(t + iu)e^{-i\omega t}dt. \tag{307}$$

Thus

$$|\hat{g}(\omega)e^{-\omega u}| \leq \int_{-\infty}^{\infty} |g(t + iu)|dt. \tag{308}$$

Then (303) and (308) imply

$$\hat{g}(\omega) = 0 \quad \text{for} \quad \omega < a \tag{309}$$

and since $\hat{g}(\omega)$ is continuous, (309) implies

$$\hat{g}(\omega) = 0 \quad \text{for} \quad \omega \leq a. \tag{310}$$

Thus we may write

$$g(t + iu) = \frac{e^{-au}}{2\pi} \int_{-\infty}^{\infty} e^{-u|\omega - a|}\hat{g}(\omega)e^{i\omega t}dt. \tag{311}$$

Then since

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-u|\omega - a|}e^{i\omega t}dt = e^{iat}P_u(t) \tag{312}$$

where $P_u(t)$ is the Poisson kernel of (269), we have

$$g(t + iu) = e^{-au} \int_{-\infty}^{\infty} g(s)e^{ia(t-s)}\, P_u(t - s)ds \tag{313}$$

and (304) follows from (313), and Lemma 3 is proved.

Now we are prepared to prove the second half of the Paley-Wiener Theorem for L_∞. We have $g(\tau)$ analytic for $u > 0$ and

$$\sup_t |g(t + iu)| \leq M \quad \text{for} \quad u \geq 0. \tag{314}$$

Now suppose $h(t)$ is any function of L_1 whose Fourier transform vanishes over $(-\infty,0)$. We have from Lemma 1 that $h(t)$ is the boundary value of a function $h(\tau)$ analytic in the upper half-plane $u > 0$, satisfying

$$\int_{-\infty}^{\infty} |h(t + iu)dt \leq \int_{-\infty}^{\infty} |h(t)|dt, \quad u \geq 0. \tag{315}$$

Now we consider the function

$$f(\tau) = g(\tau)h(\tau) \tag{316}$$

which is analytic in the uhp and satisfies

$$\int_{-\infty}^{\infty} |f(t + iu)|dt \leq M \int_{-\infty}^{\infty} |h(t)|dt. \tag{317}$$

Then from Lemma 3

$$\int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt = 0 \quad \text{for} \quad \omega \leq 0. \tag{318}$$

In particular

$$\int_{-\infty}^{\infty} f(t)dt = \int_{-\infty}^{\infty} g(t)h(t)dt = 0 \tag{319}$$

which completes the proof of the first (one-sided) Paley-Wiener Theorem.

Theorem 2. It is sufficient to prove the two-sided Paley-Wiener Theorem for functions $g(t)$ whose Fourier transforms vanish outside $[-a,a]$. We show that $g(t)$ is the boundary value of an entire function of exponential type by defining

$$G_1(t+iu) = \int_{-\infty}^{\infty} g(s)e^{au}\frac{ue^{-ia(t-s)}}{\pi\{(t-s)^2+u^2\}}ds, \quad u>0 \tag{320}$$

$$G_2(t+iu) = \int_{-\infty}^{\infty} g(s)e^{-au}\frac{|u|e^{ia(t-s)}}{\pi\{(t-s)^2+u^2\}}ds, \quad u<0. \tag{321}$$

We have

$$|G_1(t+iu)| \leq e^{au}\sup_t |g(t)|, \quad u>0 \tag{322}$$

$$|G_2(t+iu)| \leq e^{-au}\sup_t |g(t)|, \quad u<0 \tag{323}$$

and since $g(t)$ is continuous

$$\lim_{u\to 0} G_1(t+iu) = g(t) \quad \text{for all } t \tag{324}$$

$$\lim_{u\to 0} G_2(t+iu) = g(t) \quad \text{for all } t. \tag{325}$$

Now we define

$$G_3(\tau) = \int_{-\infty}^{\infty} K(\tau-s)g(s)ds \tag{326}$$

where

$$K(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{K}(\omega)e^{i\omega\tau}d\omega$$

and

$$\hat{K}(\omega) = 1, \quad -a \leq \omega < a$$

$$= 2\left(1 - \frac{\omega}{2a}\right), \quad a \leq \omega \leq 2a$$

$$= 0, \quad \omega \geq 2a$$

$$\hat{K}(-\omega) = \hat{K}(\omega). \tag{327}$$

Then $K(\tau)$ is an entire function belonging to L_1 along each line parallel to the real axis. So $G_3(\tau)$ is an entire function bounded on each line parallel to the real axis. Since the Fourier transform of $g(t)$ vanishes outside $[-a,a]$ we may replace the convolution kernels in (320) and (321) by $K(t + iu)$ since in each case their Fourier transforms agree over $[-a,a]$.

Thus

$$G_1(t + iu) = G_3(t + iu), \quad u > 0 \tag{328}$$

$$G_2(t + iu) = G_3(t + iu), \quad u < 0. \tag{329}$$

Hence $g(t)$ is the restriction to the real line of an entire function $G_3(t + iu) \equiv g(t + iu)$ satisfying

$$|g(t + iu)| \leq e^{a|u|} \sup_t |g(t)|. \tag{330}$$

Now for the second half of the theorem we suppose that $g(\tau)$ is an entire function satisfying

$$\sup_t |g(t + iu)| \leq e^{a|u|} \sup_t |g(t)| \tag{331}$$

and wish to conclude that

$$\int_{-\infty}^{\infty} g(t)h(t)dt = 0 \tag{332}$$

for all functions h in L_1 whose Fourier transforms vanish over $(-a,a)$ (and hence over $[-a,a]$). From the one-sided Paley-Wiener Theorem we have

$$\int_{-\infty}^{\infty} g(t)h_1(t)dt = 0 \tag{333}$$

for all functions h_1 in L_1 whose Fourier transforms are supported on $(-\infty,-a)$ and

$$\int_{-\infty}^{\infty} g(t)h_2(t)dt = 0 \tag{334}$$

for all function h_2 in L_1 whose Fourier transforms are supported on (a,∞). The difficulty we encounter in establishing (332) is that an arbitrary function h in L_1 whose Fourier transform vanishes over $[-a,a]$ cannot be decomposed as

$$h(t) = h_-(t) + h_+(t) \tag{335}$$

where h_- and h_+ belong to L_1, and the Fourier transform of h_- is supported on $(-\infty,-a)$, and the Fourier transform of h_+ is supported on (a,∞). In order to deduce (332) from (333) and (334) we have to approximate the test function h in (332) with bandlimited functions h_b. We may take

$$h_b(t) = \int_{-\infty}^{\infty} bK(bs)h(t-s)ds, \quad b > 0 \tag{336}$$

where

$$K(t) = \frac{2}{\pi}\frac{\sin^2\frac{t}{2}}{t^2}. \tag{337}$$

Since

$$\int_{-\infty}^{\infty} K(t)dt = \int_{-\infty}^{\infty} |K(t)|dt = b\int_{\infty}^{\infty} |K(bt)|dt = 1,$$

we have

$$\int_{-\infty}^{\infty} |h_b(t)|dt \le \int_{-\infty}^{\infty} |h(t)|dt. \tag{338}$$

Also we may write

$$h(t) - h_b(t) = \int_{-\infty}^{\infty} bK(bs)\{h(t) - h(t-s)\}ds \tag{339}$$

which gives

$$\int_{-\infty}^{\infty} |h(t) - h_b(t)|dt \le \int_{-\infty}^{\infty} bK(bs)\mu_1(s;h)ds \tag{340}$$

where

$$\mu_1(s;h) = \int_{-\infty}^{\infty} |h(t) - h(t-s)|ds. \tag{341}$$

The function $\mu_1(s)$ is called the modulus of continuity of h. It is an even, *continuous*, bounded function of s (see Ref. 3) and

$$\begin{cases} \mu_1(0;h) = 0 \\ \mu_1(s;h) \le 2\displaystyle\int_{-\infty}^{\infty} |h(t)|dt = 2\|h\|_1. \end{cases} \tag{342}$$

Then writing

$$\int_{-\infty}^{\infty} bK(bs)\mu_1(s;h)ds = \int_{-\infty}^{\infty} K(t)\mu_1\left(\frac{t}{b};h\right)dt \tag{343}$$

$$\leq \int_{-\sqrt{b}}^{\sqrt{b}} K(t)\mu_1\left(\frac{t}{b};h\right)dt + 2\|h\|_1 \int_{|t|>\sqrt{b}} K(t)dt$$

it is clear that given $\epsilon > 0$ we can choose b so large ($a < b < \infty$) that

$$\int_{-\infty}^{\infty} |h(t) - h_b(t)|dt < \epsilon. \tag{344}$$

Now the Fourier transform of h_b is supported on the intervals $(-b,-a)$ and (a,b), so h_b does have the decomposition

$$h_b(t) = h_-(t) + h_+(t) \tag{345}$$

desired in (335). This follows from the existence† of a function $K_{a,b}(t)$ in L_1 whose Fourier transform satisfies

$$\hat{K}_{a,b}(\omega) = 1, \quad a \leq \omega \leq b$$
$$= 0, \quad \omega \leq -a \tag{346}$$

so that

$$h_+(t) = \int_{-\infty}^{\infty} h_b(s)K_{a,b}(t-s)ds \tag{347}$$

and

$$\|h_+\|_1 \leq \|h_b\|_1 \cdot \|K_{a,b}\|_1. \tag{348}$$

We have

$$h_-(t) = h_b(t) - h_+(t). \tag{349}$$

So h_- also belongs to L_1.

Returning to (332) we have

$$\int_{-\infty}^{\infty} g(t)h(t)dt = \int_{-\infty}^{\infty} g(t)h_b(t)dt + \int_{-\infty}^{\infty} g(t)\{h(t) - h_b(t)\}dt. \tag{350}$$

Since

$$\int_{-\infty}^{\infty} g(t)h_b(t)dt = \int_{-\infty}^{\infty} g(t)h_-(t)dt + \int_{-\infty}^{\infty} g(t)h_+(t)dt \tag{351}$$
$$= 0 + 0$$

† See Appendix B for a good choice of $K_{a,b}(t)$.

we have

$$\left| \int_{-\infty}^{\infty} g(t)h(t)dt \right| \leq \sup_t |g(t)| \int_{-\infty}^{\infty} |h(t) - h_b(t)|dt$$

$$\leq \epsilon \sup_t |g(t)|. \tag{352}$$

Since we may choose b sufficiently large to make ϵ arbitrarily small we conclude that

$$\int_{-\infty}^{\infty} g(t)h(t)dt = 0 \tag{353}$$

for all h in L_1 whose Fourier transforms vanish over $(-a,a)$. This completes the proof of the two-sided Paley-Wiener Theorem.

Theorem 3. Here we wish to show that if $g(\tau)$ is analytic in the uhp and bounded on the real line as well as every line parallel to the real axis in the uhp, then the asymptotic estimate

$$\sup_t |g(t + iu)| = O\{e^{-au}\} \quad \text{as } u \to \infty \tag{354}$$

implies

$$\sup_t |g(t + iu)| \leq e^{-au} \sup_t |g(t)|, \quad u \geq 0. \tag{355}$$

Now if x is any real number and y is any positive number, the function

$$h(\tau) = \frac{y}{\pi} \frac{\{g(\tau)e^{-ia\tau} - g(x + iy)e^{-ia(x+iy)}\}}{(\tau - x)^2 + y^2} \tag{356}$$

where we think of x and y fixed, is analytic in the upper half-plane $u > 0$ and satisfies

$$\int_{-\infty}^{\infty} |h(t + iu)|dt < \infty, \quad u \geq 0 \tag{357}$$

and

$$\int_{-\infty}^{\infty} |h(t + iu)|dt = O(1) \quad \text{as } u \to \infty. \tag{358}$$

It follows from Lemma 3 that

$$\int_{-\infty}^{\infty} h(t)dt = 0. \tag{359}$$

Hence

$$\frac{y}{\pi}\int_{-\infty}^{\infty}\frac{g(t)e^{-iat}dt}{(t-x)^2+y^2}=g(x+iy)e^{-ia(x+iy)}. \tag{360}$$

Thus

$$|g(x+iy)|\leq e^{-ay}\sup_t|g(t)|. \tag{361}$$

This inequality holds for any $-\infty<x<\infty$ and any $y>0$. Then (355) follows from (361), if we note that (355) holds trivially for $u=0$.

APPENDIX B

Reproducing Kernels of Small L_1 Norm

We would like to find a kernel $K_{\alpha,\beta}(t)$ of minimum L_1 norm whose Fourier transform satisfies

$$\begin{aligned}\hat{K}_{\alpha,\beta}(\omega)&=1, \quad 0\leq\omega\leq\alpha\\ &=0, \qquad \omega>\beta\end{aligned} \tag{362}$$

where $0<\alpha<\beta$.

Replacing $K_{\alpha,\beta}(t)$ by $1/\alpha K_{\alpha,\beta}(t/\alpha)$ we see that the minimum norm is a function of β/α, or if we like, a function of $\alpha/(\beta-\alpha)$. It is sufficient to consider functions $K_\lambda(t)$ whose Fourier transforms satisfy

$$\begin{aligned}\hat{K}_\lambda(\omega)&=0, \quad \omega<0\\ &=1, \quad 1\leq\omega\leq 1+\lambda\end{aligned} \tag{363}$$

where we make the identification

$$\lambda=\frac{\alpha}{\beta-\alpha}. \tag{364}$$

We will not treat the minimization problem here. Instead we give a construction for a particular function $\hat{K}_\lambda(\omega)$ which can be shown† to be the solution for the case $\lambda=n$, $n=1,2,3,\ldots$. The construction provides an interpolation between the minimal norm values in case $n<\lambda<n+1$.

We write

$$\lambda=n+\theta \tag{365}$$

where $n=[\lambda]$ is the largest integer contained in λ and $0\leq\theta<1$. Then we set

$$K_\lambda(t)=2\pi\{F_\lambda(t)\}^2 \tag{366}$$

† The details will be given in a future paper.

where the Fourier transform of $F_\lambda(t)$ vanishes for negative argument, and otherwise is defined by

$$\begin{aligned}\hat{F}_\lambda(\omega) &= a_k, \quad k \le \omega < k+1 \le n+1 \\ &= a_{n+1}, \quad n+1 \le \omega < n+1+\theta \\ &= 0, \quad \omega \ge n+1+\theta = \lambda+1\end{aligned} \tag{367}$$

where the a_k are defined by

$$\frac{1}{\sqrt{1-z}} = \sum_0^\infty a_k z^k \tag{368}$$

i.e.,

$$\begin{aligned}a_k = \frac{(1/2)_k}{k!} &= \frac{\left(\frac{1}{2}\right)\left(\frac{1}{2}+1\right)\left(\frac{1}{2}+2\right)\cdots\left(\frac{1}{2}+k-1\right)}{k!} \\ &= \frac{\Gamma\left(\frac{1}{2}+k\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma(1+k)}\end{aligned} \tag{369}$$

We then have

$$\hat{K}_\lambda(\omega) = \int_0^\omega \hat{F}_\lambda(x)\hat{F}_\lambda(\omega - x)dx \tag{370}$$

which is a piecewise linear function satisfying

$$\hat{K}_\lambda(m) = \sum_{k=0}^{m-1} a_k a_{m-1-k} = 1 \tag{371}$$

for $m = 1,2,\ldots,n+1$. Thus

$$\begin{aligned}\hat{K}_\lambda(\omega) &= 1 \quad \text{for} \quad 1 \le \omega \le n+1 \\ &= \omega \quad \text{for} \quad 0 \le \omega \le 1.\end{aligned} \tag{372}$$

For $n+1 \le \omega \le n+1+\theta$ the convolution in (370) is independent of the definition of $\hat{F}_\lambda(x)$ for $x > n+1+\theta$; i.e.

$$\hat{K}_\lambda(\omega) = \hat{K}_{n+1}(\omega) \quad \text{for} \quad 0 \le \omega \le n+1+\theta. \tag{373}$$

So

$$\hat{K}_\lambda(\omega) = 1, \quad n+1 \le \omega \le n+1+\theta. \tag{374}$$

Thus

$$\begin{aligned}\hat{K}_\lambda(\omega) &= 1 \quad \text{for} \quad 1 \le \omega \le n+1+\theta \\ &= \omega \quad \text{for} \quad 0 \le \omega \le 1.\end{aligned} \tag{375}$$

We have

$$\mu(\lambda) \equiv \int_{-\infty}^{\infty} |K_\lambda(t)|dt = 2\pi \int_{-\infty}^{\infty} |F_\lambda(t)|^2 dt$$

$$= \int_{-\infty}^{\infty} |\hat{F}_\lambda(\omega)|^2 d\omega$$

$$= \sum_0^n a_k^2 + \theta a_{n+1}^2 \tag{376}$$

which is a piecewise linear function of λ, i.e.,

$$\mu(\lambda) = (n + 1 - \lambda)\mu(n) + (\lambda - n)\mu(n + 1), \quad n \leq \lambda \leq n + 1. \tag{377}$$

We have

$$\mu(0) = 1, \quad \mu(1) = \frac{5}{4}, \quad \mu(2) = \frac{89}{64}. \tag{378}$$

The remainder of this appendix is devoted to estimating $\mu(n)$ for large n. We need an upper bound.

For convenience we set

$$\gamma(x) = \frac{\Gamma\left(\frac{1}{2} + x\right)}{\Gamma(1 + x)} \tag{379}$$

and then

$$\mu(n) = \frac{1}{\pi} \sum_{k=0}^{n} \gamma^2(k). \tag{380}$$

First we estimate $\gamma(x)$. We have the representation for the Beta function

$$\frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)} = \int_0^1 t^{y-1}(1 - t)^{x-1} dt \ (Re\, x > 0, Re\, y > 0). \tag{381}$$

Then

$$\gamma(x) = \frac{1}{\sqrt{\pi}} \int_0^1 \frac{(1 - t)^x}{\sqrt{t(1 - t)}} dt. \tag{382}$$

Setting $1 - t = e^{-s}$ we obtain

$$\gamma(x) = \frac{1}{\sqrt{\pi}} \int_0^\infty \frac{e^{-s(x+1/2)}}{\sqrt{1 - e^s}} ds$$

$$= \frac{1}{\sqrt{\pi}} \int_0^\infty \frac{e^{-s(x+1/4)}}{\sqrt{2 \sinh s/2}} ds. \tag{383}$$

Since

$$1 - e^{-s} < s$$

and $2 \sinh s/2 > s$ we obtain from (383)

$$\frac{1}{\sqrt{x + \frac{1}{2}}} < \gamma(x) < \frac{1}{\sqrt{x + \frac{1}{4}}}. \tag{384}$$

We can obtain a simple upper bound for $\mu(n)$ from (384). We have

$$\begin{aligned} \mu(n) &= 1 + \frac{1}{\pi} \sum_{k=1}^{n} \gamma^2(k) \\ &< 1 + \frac{1}{\pi} \sum_{k=1}^{n} \frac{1}{k + \frac{1}{4}}. \end{aligned} \tag{385}$$

Since t^{-1} is convex, we have

$$\int_T^{T+1} \frac{dt}{t} > \frac{1}{T + \frac{1}{2}}, \quad T > 0 \tag{386}$$

and thus

$$\mu(n) < 1 + \frac{1}{\pi} \int_{3/4}^{n+3/4} \frac{dt}{t} = 1 + \frac{1}{\pi} \log \left(1 + \frac{4n}{3}\right), \quad (n \geq 1). \tag{387}$$

Since $\log (1 + 4\lambda/3)$ is a concave function of λ and since $\mu(\lambda)$ is piecewise linear between integers, we conclude from (387) that

$$\mu(\lambda) < 1 + \frac{1}{\pi} \log \left(1 + \frac{4\lambda}{3}\right), \quad \lambda > 0. \tag{388}$$

A sharper estimate of $\mu(\lambda)$ for large λ is obtained as follows. We are interested in the constant term in the asymptotic expansion.

From (383) we have

$$\gamma\left(x - \frac{1}{2}\right) = \frac{1}{\sqrt{\pi}} \int_0^{\infty} \frac{e^{-sx} ds}{\sqrt{1 - e^{-s}}}. \tag{389}$$

Then from the convolution theorem for Laplace transforms,

$$\gamma^2\left(x - \frac{1}{2}\right) = \int_0^{\infty} e^{-sx} \varphi(s)\, ds \tag{390}$$

where

$$\varphi(s) = \frac{1}{\pi} \int_0^s \frac{1}{\sqrt{1 - e^{-t}}} \frac{dt}{\sqrt{1 - e^{-(s-t)}}}. \tag{391}$$

Setting $e^{-t} = u$ in (391) we obtain

$$\varphi(s) = \frac{1}{\pi}\int_v^1 \frac{1}{\sqrt{1-u}\sqrt{u-v}}\frac{du}{\sqrt{u}} \tag{392}$$

where

$$v = e^{-s}.$$

Then with the substitution

$$1 - u = (1 - v)t$$

(392) becomes

$$\varphi(s) = \frac{1}{\pi}\int_0^1 \frac{1}{\sqrt{t}\sqrt{1-t}}\frac{dt}{\sqrt{1-(1-v)t}}. \tag{393}$$

Identifying $\varphi(s)$ with the hypergeometric function which has the representation,

$${}_2F_1(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}dt$$

$$(Re\, c > Re\, b > 0) \tag{394}$$

we see that

$$\varphi(s) = {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; 1 - e^{-s}\right). \tag{395}$$

Thus

$$\gamma^2\left(x - \frac{1}{2}\right) \equiv \frac{\Gamma^2(x)}{\Gamma^2\left(x + \frac{1}{2}\right)} = \int_0^\infty e^{-sx}{}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; 1 - e^{-s}\right) ds \tag{396}$$

and

$$\begin{aligned}\mu(n) = \frac{1}{\pi}\sum_{k=0}^{n}\gamma^2(k) &= \frac{1}{\pi}\int_0^\infty \frac{1-e^{-s(n+1)}}{1-e^{-s}} e^{-s/2}\varphi(s)ds \\ &= \frac{1}{\pi}\int_0^\infty \{e^{-s/2} - e^{-s(n+3/2)}\}\left\{\frac{\varphi(s)}{1-e^{-s}} - \frac{1}{s}\right\} ds \\ &\quad + \frac{1}{\pi}\int_0^\infty \frac{e^{-s/2} - e^{-s(n+3/2)}}{s} ds \\ &= \frac{1}{\pi}\log(2n+3) + \frac{1}{\pi}\int_0^\infty e^{-s/2}\left\{\frac{\varphi(s)}{1-e^{-s}} - \frac{1}{s}\right\} ds \\ &\quad - \frac{1}{\pi}\int_0^\infty e^{-s(n+3/2)}\left\{\frac{\varphi(s)}{1-e^{-s}} - \frac{1}{s}\right\} ds. \end{aligned} \tag{397}$$

From (395) and the series

$$ {}_2F_1(a,b;c;z) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k} \frac{z^k}{k!} \tag{398} $$

it is clear that

$$ \varphi(s) \geq \varphi(0) = 1, \quad s > 0 \tag{399} $$

and hence that

$$ \frac{\varphi(s)}{1 - e^{-s}} - \frac{1}{s} \geq 0. \tag{400} $$

Thus from (397) we have

$$ \mu(n) < \frac{1}{\pi} \log (2n + 3) + \frac{1}{\pi} \int_0^{\infty} e^{-s/2} \left\{ \frac{\varphi(s)}{1 - e^{-s}} - \frac{1}{s} \right\} ds \tag{401} $$

and in fact

$$ \lim_{n \to \infty} \left\{ \mu(n) - \frac{1}{\pi} \log (2n + 3) \right\} = \frac{1}{\pi} \int_0^{\infty} e^{-s/2} \left\{ \frac{\varphi(s)}{1 - e^{-s}} - \frac{1}{s} \right\} ds. \tag{402} $$

Now we can evaluate the integral in (402) by an indirect route.

We have

$$ \mu(n) - \frac{1}{\pi} \sum_0^n \frac{1}{k+1} = \sum_0^n \left\{ \frac{(1/2)_k (1/2)_k}{k!k!} - \frac{1}{\pi(k+1)} \right\} $$

$$ = \frac{1}{\pi} \sum_0^n \left\{ \gamma^2(k) - \frac{1}{k+1} \right\}. \tag{403} $$

From the estimate (384) we see that the sum on the right converges as $n \to \infty$; i.e.

$$ \lim_{n \to \infty} \left\{ \mu(n) - \frac{1}{\pi} \sum_0^n \frac{1}{k+1} \right\} = \frac{1}{\pi} \sum_0^{\infty} \left\{ \gamma^2(k) - \frac{1}{k+1} \right\}. \tag{404} $$

Now we can write

$$ A \equiv \frac{1}{\pi} \sum_0^{\infty} \left\{ \gamma^2(k) - \frac{1}{k+1} \right\} = \lim_{x \to 1} \left\{ \sum_{k=0}^{\infty} \frac{(1/2)_k (1/2)_k}{k!k!} x^k - \frac{x^k}{\pi(k+1)} \right\} $$

$$ = \lim_{x \to 1} \left\{ {}_2F_1 \left(\frac{1}{2}; \frac{1}{2}; 1; x \right) - \frac{1}{\pi x} \log \frac{1}{1-x} \right\} \tag{405} $$

Then using (394) we have

$$\pi A = \lim_{x\to 1}\int_0^1 \left\{\frac{1}{\sqrt{t(1-t)(1-xt)}} - \frac{1}{1-xt}\right\} dt$$

$$= \lim_{x\to 1}\int_0^1 \frac{dt}{1-xt}\left\{\frac{\sqrt{1-xt}}{\sqrt{t(1-t)}} - 1\right\}$$

$$= \lim_{x\to 1}\int_0^1 \frac{dt}{1-xt}\left\{\frac{1}{\sqrt{t}} - 1\right\} + \lim_{x\to 1}\int_0^1 \frac{dt}{(1-xt)\sqrt{t}}\left\{\frac{\sqrt{1-xt}}{\sqrt{1-t}} - 1\right\}$$

$$= \int_0^1 \frac{dt}{(1+\sqrt{t})\sqrt{t}} + \lim_{x\to 1}\int_0^1 \frac{dt}{(1-xt)\sqrt{t}}\left\{\frac{\sqrt{1-xt}}{\sqrt{1-t}} - 1\right\}$$

$$= 2\log 2 + 2\log 2. \quad (406)$$

Some care is required in evaluating the last limit. An alternative way of obtaining the result is worth noting, as it makes use of an interesting series obtained from (396). A change of variables gives

$$\frac{\Gamma^2(x)}{\Gamma^2\left(x+\frac{1}{2}\right)} = \int_0^1 (1-t)^{x-1}{}_2F_1\left(\frac{1}{2},\frac{1}{2};1;t\right) dt$$

$$= \int_0^1 (1-t)^{x-1}\left\{\sum_0^\infty \frac{(1/2)_k(1/2)_k}{k!k!}\,t^k\right\} dt$$

$$= \frac{1}{x} + \frac{1}{x(x+1)}\left\{\frac{1}{2}\right\}^2 + \frac{2!}{x(x+1)(x+2)}\left\{\frac{\frac{1}{2}\cdot\frac{3}{2}}{2}\right\}^2 + \ldots \quad (407)$$

Then setting

$$F(x) = \frac{\Gamma^2\left(x+\frac{1}{2}\right)}{\Gamma^2(x+1)}$$

$$\left\{\frac{(1/2)_k}{k!}\right\}^2 = \left\{\frac{\Gamma\left(\frac{1}{2}+k\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma(k+1)}\right\}^2 = \frac{1}{\pi}F(k)$$

we have

$$F\left(x-\frac{1}{2}\right) = \frac{1}{\pi}\sum_{k=0}^{\infty}\frac{k!}{(x)_{k+1}}F(k) \quad (408)$$

which is an interesting formula. In particular $\langle F(0),F(1),F(2),\ldots\rangle$ is

an eigenvector of a certain infinite matrix. Also an interesting series for π is obtained by setting $x = n + 1/2$ (large n) in (408). Returning to (407) and recalling $a_k = (1/2)_k/k!$ we have

$$\frac{x\Gamma^2(x)}{\Gamma^2\left(x+\frac{1}{2}\right)} = \frac{\Gamma^2(x+1)}{x\Gamma^2\left(x+\frac{1}{2}\right)} = 1 + \frac{a_1^2}{x+1} + \frac{2!}{(x+1)(x+2)}a_2^2 + \ldots \tag{409}$$

Then using the series

$$\frac{1}{x} = \int_0^1 (1-t)^{x-1}dt = \int_0^1 (1-t)^x(1+t+t^2+\ldots)dt$$

$$= \frac{1}{x+1} + \frac{1}{(x+1)(x+2)} + \frac{2!}{(x+1)(x+2)(x+3)} + \ldots \quad (x>0)$$

we may write

$$\frac{1}{x}\left\{\frac{\Gamma^2(x+1)}{\Gamma^2\left(x+\frac{1}{2}\right)} - \frac{\Gamma^2(1)}{\Gamma^2\left(\frac{1}{2}\right)}\right\}$$

$$= 1 + \frac{1}{x+1}\left\{a_1^2 - \frac{1}{\pi}\right\} + \frac{2!}{(x+1)(x+2)}\left\{a_2^2 - \frac{1}{2\pi}\right\}$$

$$+ \ldots \frac{k!}{(x+1)_k}\left\{a_k^2 - \frac{1}{k\pi}\right\} + \ldots \tag{410}$$

Since

$$\frac{1}{\pi\left(k+\frac{1}{2}\right)} < a_k^2 < \frac{1}{\pi\left(k+\frac{1}{4}\right)}$$

we may let $x \to 0$ with the result

$$1 + \sum_{k=1}^{\infty}\left\{a_k^2 - \frac{1}{k\pi}\right\} = \frac{d}{dx}\left.\frac{\Gamma^2(x+1)}{\Gamma^2\left(x+\frac{1}{2}\right)}\right|_{x=0} = \frac{4}{\pi}\log 2 \tag{411}$$

and since $a_0^2 = 1$, this sum is the same as the sum on the right in (404). Hence the limit in (404) is

$$\lim_{n\to\infty}\left\{\mu(n) - \frac{1}{\pi}\sum_0^n \frac{1}{k+1}\right\} = \frac{4}{\pi}\log 2 \tag{412}$$

and since

$$\lim_{n\to\infty}\left\{\sum_{1}^{n}\frac{1}{k}-\log n\right\}=C=0.577215\ldots\text{ (Euler's constant)}\quad(413)$$

we have from (412), (413), and (402),

$$\lim_{n\to\infty}\left\{\mu(n)-\frac{1}{\pi}\log(2n+3)\right\}=\frac{1}{\pi}\int_0^\infty e^{-s/2}\left\{\frac{\varphi(s)}{1-e^{-s}}-\frac{1}{s}\right\}ds$$

$$=\frac{3}{\pi}\log 2+\frac{C}{\pi}.\quad(414)$$

Then from (401) and (414) we have

$$\mu(n)<\frac{1}{\pi}\log(2n+3)+\frac{3}{\pi}\log 2+\frac{C}{\pi}\quad(415)$$

and by the same argument used in establishing (388),

$$\mu(\lambda)<\frac{1}{\pi}\log(2\lambda+3)+\frac{3}{\pi}\log 2+\frac{C}{\pi},\quad\lambda>0.\quad(416)$$

We find from (397) and (414) that

$$\pi\mu(n)\sim\log(2n+3)+3\log 2+C$$

$$-\frac{3}{2(2n+3)}-\frac{43}{48(2n+3)^2}-\frac{7}{16(2n+3)^3}$$

$$+O(n^{-4}).\quad(417)$$

For comparing the estimates (389) and (415) we have

$$1+\frac{1}{\pi}\log\frac{4}{3}\doteq 1.0915720476\quad(418)$$

$$\frac{1}{\pi}\{4\log 2+C\}\doteq 1.066275853\quad(419)$$

and for use in (415) and (417)

$$\frac{1}{\pi}\{3\log 2+C\}\doteq\frac{2.656657207}{\pi}$$

$$\doteq .8456402533\quad(420)$$

In the following tabulations the estimates (388) and (415) and the asymptotic formula (417) are compared with the true value of $\mu(n)$.

n	$\mu(n)$	Asymptotic formula	Error
1	1.25	1.249927097	7.29 (−5)
2	1.390625	1.390607982	1.70 (−5)
3	1.48828125	1.488275479	5.77 (−6)
4	1.563049316	1.563046870	2.45 (−6)
5	1.623611450	1.623610248	1.20 (−6)
6	1.674500465	1.674499809	6.56 (−7)
7	1.718379259	1.718378872	3.87 (−7)
8	1.756944605	1.756944363	2.42 (−7)
9	1.791343941	1.791343782	1.59 (−7)
10	1.822389342	1.822389234	1.08 (−7)

n	Upper bound (388)	Upper bound (415)
1	1.269703286	1.357940252
2	1.413574619	1.465042692
3	1.512299999	1.545038559
4	1.587544884	1.608914025
5	1.648359655	1.662088992
6	1.699398305	1.707639405
7	1.743372924	1.747480071
8	1.782003284	1.782884290
9	1.816448738	1.814741844
10	1.847528028	1.843699061

APPENDIX C

Estimates for M_n

$$M_n = \frac{2}{n+1} \sum_{\substack{1 \le k \le n \\ k \text{ odd}}} \cot \frac{k\pi}{2(n+1)} \tag{421}$$

In order to express the sum as an integral we note first that

$$\begin{aligned} \pi \cot \pi x &= \frac{d}{dx} \log \sin \pi x \\ &= \frac{d}{dx} \log \frac{\pi}{\Gamma(x)\Gamma(1-x)} \\ &= \psi(1-x) - \psi(x) \end{aligned} \tag{422}$$

where

$$\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)} = \int_0^1 \left\{ -\frac{1}{\log t} - \frac{t^{x-1}}{1-t} \right\} dt, \quad x > 0. \tag{423}$$

So

$$\pi \cot \pi x = \int_0^1 \{t^{x-1} - t^{-x}\} \frac{dt}{1-t} \tag{424}$$

$$= \int_0^\infty \{e^{-u(x-1)} - e^{ux}\} \frac{e^{-u}}{1-e^{-u}} du, \quad 0 < x < 1.$$

Then for $0 < \theta < 1/\nu$, where ν is an odd integer,

$$\pi\theta \sum_{\substack{k=1 \\ k \text{ odd}}}^{\nu} \cot k\pi\theta = \theta \int_0^\infty \left\{ e^u \frac{(e^{-u\theta} - e^{-(\nu+2)u\theta})}{1-e^{-2u\theta}} - \frac{(e^{u\theta} - e^{(\nu+2)u\theta})}{1-e^{2u\theta}} \right\} \frac{e^{-u}}{1-e^{-u}} du \tag{425}$$

or

$$\pi\theta \sum_{\substack{k=1 \\ k \text{ odd}}}^{\nu} \cot k\pi\theta$$

$$= \int_0^\infty \left\{ e^{at} \frac{(e^{-t} - e^{-(\nu+2)t})}{1-e^{-2t}} + \frac{(e^{-t} - e^{\nu t})}{1-e^{-2t}} \right\} \cdot \frac{e^{-at}}{1-e^{-at}} dt$$

$$= \int_0^\infty \frac{e^{-t} - e^{-(\nu+2)t}}{1-e^{-2t}} dt - \int_0^\infty \frac{\{1-e^{-(\nu+1)t}\}^2}{1-e^{-2t}} \frac{e^{(\nu-a)t}}{1-e^{-at}} dt \tag{426}$$

where $a = 1/\theta$. For n odd we take $\nu = n$, $a = 2(n+1)$. Then

$$M_n = \frac{4}{\pi} \int_0^\infty \frac{e^{-t} - e^{-(n+2)t}}{1-e^{-2t}} dt - \frac{4}{\pi} \int_0^\infty \frac{\{1-e^{-(n+1)t}\}^2}{1-e^{-2t}} \frac{e^{-(n+2)t}}{1-e^{-2(n+1)t}} dt$$

$$= \frac{4}{\pi} \int_0^\infty \frac{e^{-t} - e^{-(n+2)t}}{1-e^{-2t}} dt - \frac{2}{\pi(n+1)} \int_0^\infty \frac{1-e^{-t}}{\sinh \dfrac{t}{n+1}} \frac{dt}{1+e^t}$$

$$n \text{ odd} \tag{427}$$

The first integral is just the sum of the reciprocals of the odd integers from 1 through n. We have

$$\int_0^\infty \frac{e^{-t} - e^{-(n+2)t}}{t} dt = \log(n+2). \tag{428}$$

So

$$2 \int_0^\infty \frac{e^{-t} - e^{-(n+2)t}}{1-e^{-2t}} dt = \log(n+2) + \int_0^\infty e^{-t} \left\{ \frac{2}{1-e^{-2t}} - \frac{1}{t} \right\} dt - \int_0^\infty e^{-(n+2)t} \left\{ \frac{2}{1-e^{-2t}} - \frac{1}{t} \right\} dt. \tag{429}$$

We have

$$\int_0^\infty e^{-t}\left\{\frac{2}{1-e^{-2t}}-\frac{1}{t}\right\}dt=\int_0^1\left\{\frac{2}{1-u^2}+\frac{1}{\log u}\right\}du$$

$$=\int_0^1\left\{\frac{1}{1-u}+\frac{1}{1+u}+\frac{1}{\log u}\right\}du$$

$$=\int_0^1\frac{du}{1+u}+\int_0^1\left\{\frac{1}{1-u}+\frac{1}{\log u}\right\}du$$

$$=\log 2-\psi(1)$$

$$=\log 2+\gamma \tag{430}$$

where

$$\gamma=.577215\ldots\quad\text{(Euler's constant)}. \tag{431}$$

Thus

$$M_n=\frac{2}{\pi}\log(n+2)+\frac{2}{\pi}(\log 2+\gamma)$$

$$-\frac{2}{\pi}\int_0^\infty e^{-(n+2)t}\left\{\frac{2}{1-e^{-2t}}-\frac{1}{t}\right\}dt$$

$$-\frac{2}{\pi(n+1)}\int_0^\infty\frac{1-e^{-t}}{\sinh\dfrac{t}{n+1}}\frac{dt}{1+e^t}$$

(n odd) (432)

We may write

$$\int_0^\infty e^{-(n+2)t}\left\{\frac{2}{1-e^{-2t}}-\frac{1}{t}\right\}dt=\int_0^\infty e^{-(n+1)t}\left\{\frac{1}{\sinh t}-\frac{e^{-t}}{t}\right\}dt$$

$$=\int_0^\infty e^{-(n+1)t}\left\{\frac{1}{\sinh t}-\frac{1}{t}\right\}dt+\int_0^\infty e^{-(n+1)t}(1-e^{-t})\frac{dt}{t}$$

$$=\frac{1}{n+1}\int_0^\infty e^{-t}\left\{\frac{1}{\sinh\dfrac{t}{n+1}}-\frac{n+1}{t}\right\}dt+\log\frac{n+2}{n+1}. \tag{433}$$

Also

$$\frac{1}{n+1}\int_0^\infty\frac{1-e^{-t}}{1+e^t}\frac{dt}{\sinh\dfrac{t}{n+1}}$$

$$= \frac{1}{n+1} \int_0^\infty \frac{1-e^{-t}}{1+e^t} \left\{ \frac{1}{\sinh \dfrac{t}{n+1}} - \frac{n+1}{t} \right\} dt$$

$$+ \int_0^\infty \frac{1-e^{-t}}{1+e^t} \frac{dt}{t} \tag{434}$$

and (Ref. 9, p. 327)

$$\int_0^\infty \frac{1-e^{-t}}{1+e^t} \frac{dt}{t} = \log \frac{\pi}{2}. \tag{435}$$

Thus we find from (432)–(435),

$$M_n = \frac{2}{\pi} \log (n+1) + \frac{2}{\pi} \left(\log \frac{4}{\pi} + \gamma \right)$$

$$+ \frac{4}{\pi(n+1)} \int_0^\infty \left\{ \frac{n+1}{t} - \frac{1}{\sinh \dfrac{t}{n+1}} \right\} \frac{dt}{1+e^t}$$

n odd. (436)

Clearly the integral in (436) is positive. In order to obtain the asymptotic series for M_n, n odd, we use the generating function for the Bernoulli polynomials (Ref. 1, formula 23.1.1)

$$\frac{te^{xt}}{e^t - 1} = \sum_{k=0}^{\infty} B_k(x) \frac{t^k}{k!} \qquad |t| < 2\pi \tag{437}$$

to obtain

$$\frac{t}{2 \sinh \dfrac{t}{2}} = \sum_{k=0}^{\infty} B_k \left(\frac{1}{2} \right) \frac{t^k}{k!} \qquad |t| < 2\pi \tag{438}$$

where (Ref. 1, formula 23.1.21)

$$B_k \left(\frac{1}{2} \right) = -(1 - 2^{1-k}) B_k, \quad B_k \equiv B_k(0). \tag{439}$$

Thus

$$\frac{t}{\sinh t} = -\sum_{k=0}^{\infty} (2^k - 2) B_k \frac{t^k}{k!} \qquad |t| < \pi. \tag{440}$$

We have (Ref. 1, formula 23.1.19)

$$B_{2k+1} = 0, \quad k = 1,2,\ldots \tag{441}$$

and (Ref. 1, Table 23.2)

$$B_0 = 1, \quad B_1 = -\frac{1}{2}, \quad B_2 = \frac{1}{6}, \quad B_4 = -\frac{1}{30}$$

$$B_6 = \frac{1}{42}, \quad B_8 = -\frac{1}{30}, \quad B_{10} = \frac{5}{66}, \quad B_{12} = -\frac{691}{2730},$$

$$B_{14} = \frac{7}{6}. \tag{442}$$

So

$$\frac{1}{t} - \frac{1}{\sinh t} = \sum_{k=1}^{\infty} (2^{2k} - 2) \frac{B_{2k} t^{2k-1}}{(2k)!} \tag{443}$$

$$= \frac{t}{6} - \frac{7}{360} t^3 + \frac{31}{15120} t^5 + \ldots \quad |t| < \pi.$$

From Ref. 9, p. 325, we have

$$\int_0^{\infty} \frac{x^{2k-1}}{e^x + 1} dx = (2^{2k} - 2)|B_{2k}| \frac{\pi^{2k}}{4k}, \quad k = 1,2,\ldots \tag{444}$$

Also, (Ref. 1, formula 23.1.18)

$$B_{2n} = B_{2n}(0) = (-1)^{n+1} \frac{2(2n)!}{(2\pi)^{2n}} \sum_{k=1}^{\infty} \frac{1}{k^{2n}}, \quad n = 1,2,\ldots \tag{445}$$

so

$$B_{2k} = (-1)^{k+1} |B_{2k}|. \tag{446}$$

Thus we obtain the asymptotic series

$$\frac{1}{n+1} \int_0^{\infty} \left\{ \frac{n+1}{t} - \frac{1}{\sinh \dfrac{t}{n+1}} \right\} \frac{dt}{1 + e^t}$$

$$\sim \sum_{k=1}^{\infty} \frac{(-1)^{k+1} (2^{2k} - 2)^2 B_{2k}^2}{4k(2k)!} \left(\frac{\pi}{n+1} \right)^{2k}$$

$$\approx \frac{1}{72} \left(\frac{\pi}{n+1} \right)^2 - \frac{49}{43200} \left(\frac{\pi}{n+1} \right)^4 + \ldots \tag{447}$$

and hence

$$M_n \sim \frac{2}{\pi} \log (n+1) + \frac{2}{\pi} \left(\log \frac{4}{\pi} + \gamma \right) + \frac{\pi}{18(n+1)^2} - \frac{49\pi^3}{10800(n+1)^4}$$

$$+ \ldots + \frac{(-1)^{k+1}}{\pi} \frac{(2^{2k} - 2)^2 B_{2k}^2}{k(2k)!} \left(\frac{\pi}{n+1} \right)^{2k} + \ldots \qquad (n \text{ odd}). \tag{448}$$

Now we would like to show that

$$M_n > \frac{2}{\pi}\log(n+1) + \frac{2}{\pi}\left(\log\frac{4}{\pi} + \gamma\right)$$

$$< \frac{2}{\pi}\log(n+1) + \frac{2}{\pi}\left(\log\frac{4}{\pi} + \gamma\right) + \frac{\pi}{18(n+1)^2}$$

$$> \frac{2}{\pi}\log(n+1) + \frac{2}{\pi}\left(\log\frac{4}{\pi} + \gamma\right) + \frac{\pi}{18(n+1)^2} - \frac{49\pi^3}{10800(n+1)^4}$$

$$\text{etc.,} \qquad (n \text{ odd}) \tag{449}$$

i.e., that the error in truncating the asymptotic series has the same sign as the next term of the series. To do this, we show that for $t > 0$

$$\frac{1}{t} - \frac{1}{\sinh t} > 0$$

$$< \frac{t}{6}$$

$$> \frac{t}{6} - \frac{7}{360}t^3$$

$$< \frac{t}{6} - \frac{7}{360}t^3 + \frac{31}{15120}t^5$$

$$\text{etc.} \tag{450}$$

We have (Ref. 9, p. 23)

$$\frac{t}{\sinh t} = 1 + 2t^2 \sum_{k=1}^{\infty} (-1)^k \frac{1}{t^2 + k^2\pi^2} \tag{451}$$

or

$$\frac{1}{t^2}\left\{1 - \frac{t}{\sinh t}\right\} = 2\sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{t^2 + k^2\pi^2}. \tag{452}$$

Now consider the polynomial

$$P_{2n}(t;k) = 2\frac{1 - \left(\frac{t}{ik\pi}\right)^{2n+2}}{t^2 + k^2\pi^2}. \tag{453}$$

We have

$$P_{2n}(t) \equiv \sum_{k=1}^{\infty} (-1)^{k+1} P_{2n}(t;k) \tag{454}$$

$$= \frac{1}{t^2}\left\{1 - \frac{t}{\sinh t}\right\} - 2(-1)^{n+1}t^{2n+2}\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{(k\pi)^{2n+2}(t^2 + k^2\pi^2)}$$

or

$$\frac{1}{t^2}\left\{1-\frac{t}{\sinh t}\right\}-P_{2n}(t)$$

$$=2(-1)^{n+1}t^{2n+2}\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{(k\pi)^{2n+2}(t^2+k^2\pi^2)}. \quad (455)$$

It follows that $P_{2n}(t)$ is a polynomial of degree $2n$ which agrees with the first $(2n+1)$ terms of the Taylor series of $t^{-2}\{1-t/\sinh t\}$. The sign of the difference in (455) is $(-1)^{n+1}$, as the sum is clearly positive. Then (450), and hence (449), follows from (455).

For n an even integer, the asymptotic expansion of M_n is not so readily obtained. In this case we set $\nu=(n-1)$ in (426) and keep $a=2(n+1)$. Thus

$$M_n=\frac{4}{\pi}\int_0^{\infty}\frac{e^{-t}-e^{-(n+1)t}}{1-e^{-2t}}dt$$

$$-\frac{4}{\pi}\int_0^{\infty}\frac{(1-e^{-nt})^2}{1-e^{-2t}}\frac{e^{-(n+3)t}}{1-e^{-2(n+1)t}}dt, \qquad n \text{ even.} \quad (456)$$

We have

$$2\int_0^{\infty}\frac{e^{-t}-e^{-(n+1)t}}{1-e^{-2t}}dt=\log(n+1)$$

$$+\int_0^{\infty}\{e^{-t}-e^{-(n+1)t}\}\left\{\frac{2}{1-e^{-2t}}-\frac{1}{t}\right\}dt. \quad (457)$$

Now we would like to express the second integral in (456) as an asymptotic series in $(n+1)^{-1}$. For convenience we set $e^{-t}=x$. Then

$$\frac{(1-e^{-nt})^2}{1-e^{-2(n+1)t}}=\frac{(1-x^n)^2}{1-x^{2n+2}}=\frac{\{1-x^{n+1}-x^n(1-x)\}^2}{(1-x^{n+1})(1+x^{n+1})}$$

$$=\left\{1-\frac{x^n(1-x)}{1-x^{n+1}}\right\}\left\{\frac{1-x^{n+1}}{1+x^{n+1}}-\frac{x^n(1-x)}{1+x^{n+1}}\right\}$$

$$=\frac{1-x^{n+1}}{1+x^{n+1}}-\frac{x^n(1-x)}{1+x^{n+1}}-\frac{x^n(1-x)}{1+x^{n+1}}+\frac{x^{2n}(1-x)^2}{1-x^{2n+2}}.$$

Therefore

$$\int_0^{\infty}\frac{(1-e^{-nt})^2}{1-e^{-2t}}\frac{e^{-(n+3)t}}{1-e^{-2(n+1)t}}dt=\int_0^{\infty}\frac{1-e^{-(n+1)t}}{1+e^{-(n+1)t}}\frac{e^{-(n+3)t}}{1-e^{-2t}}dt$$

$$-2\int_0^{\infty}\frac{e^{-nt}(1-e^{-t})}{1+e^{-(n+1)t}}\frac{e^{-(n+3)t}}{1-e^{-2t}}dt$$

$$+\int_0^{\infty}\frac{e^{-2nt}(1-e^{-t})^2}{1-e^{-2(n+1)t}}\frac{e^{-(n+3)t}}{1-e^{-2t}}dt$$

$$= \int_0^\infty \frac{1 - e^{-(n+1)t}}{1 + e^{(n+1)t}} \frac{dt}{e^{2t} - 1}$$

$$- 2 \int_0^\infty \frac{e^{-2(n+1)t}}{1 + e^{-(n+1)t}} \frac{dt}{e^t + 1}$$

$$+ \int_0^\infty \frac{e^{-3(n+1)t}}{1 - e^{-2(n+1)t}} \frac{1 - e^{-t}}{1 + e^{-t}} dt. \quad (458)$$

We can combine a part of the second integral with the last by noting that

$$\frac{1}{e^t + 1} = \frac{1}{2} - \frac{1}{2} \tanh \frac{t}{2}.$$

Then

$$-2 \int_0^\infty \frac{e^{-2(n+1)t}}{1 + e^{-(n+1)t}} \frac{dt}{e^t + 1} + \int_0^\infty \frac{e^{-3(n+1)t}}{1 - e^{-2(n+1)t}} \tanh \frac{t}{2} dt$$

$$= - \frac{1}{(n+1)} \int_0^\infty \frac{e^{-2t}}{1 + e^{-t}} dt + \frac{1}{(n+1)} \int_0^\infty \frac{e^{-2t}}{1 - e^{-2t}} \tanh \frac{t}{2(n+1)} dt.$$

By making use of (430) and (435) we obtain

$$M_n = \frac{2}{\pi} \log(n+1) + \frac{2}{\pi} \left(\log \frac{4}{\pi} + \gamma \right)$$

$$- \frac{2}{\pi} \int_0^\infty e^{-(n+1)t} \left\{ \frac{2}{1 - e^{-2t}} - \frac{1}{t} \right\} dt$$

$$+ \frac{2}{\pi(n+1)} \int_0^\infty \frac{1 - e^{-t}}{1 + e^t} \left\{ \frac{n+1}{t} - \frac{2}{\exp\left(\frac{2t}{n+1}\right) - 1} \right\} dt$$

$$+ \frac{4}{\pi(n+1)} \int_0^\infty \frac{e^{-2t}}{1 + e^{-t}} dt$$

$$- \frac{4}{\pi(n+1)} \int_0^\infty \frac{e^{-2t}}{1 - e^{-2t}} \tanh \frac{t}{2(n+1)} dt, \qquad n \text{ even.} \quad (459)$$

From (437), (441), and (442) we have

$$\frac{2}{e^{2t} - 1} - \frac{1}{t} = \sum_{k=1}^{\infty} 2^k \frac{B_k t^{k-1}}{k!} = -1 + \sum_{k=1}^{\infty} 2^{2k} \frac{B_{2k} t^{2k-1}}{(2k)!}, \qquad |t| < \pi \quad (460)$$

$$\frac{2}{1 - e^{-2t}} - \frac{1}{t} = 1 + \sum_{k=1}^{\infty} 2^{2k} \frac{B_{2k} t^{2k-1}}{(2k)!}, \qquad |t| < \pi \quad (461)$$

i.e.,

$$\frac{2}{1-e^{-2t}}-\frac{1}{t}-1=\frac{2}{e^{2t}-1}-\frac{1}{t}+1=\sum_{k=1}^{\infty}2^{2k}\frac{B_{2k}t^{2k-1}}{(2k)!}. \quad (462)$$

Thus

$$\frac{2}{\pi(n+1)}\int_0^\infty \frac{1-e^{-t}}{1+e^t}\left\{\frac{n+1}{t}-\frac{2}{\exp\left(\frac{2t}{n+1}\right)-1}\right\}dt$$

$$-\frac{2}{\pi}\int_0^\infty e^{-(n+1)t}\left\{\frac{2}{1-e^{-2t}}-\frac{1}{t}\right\}dt$$

$$=\frac{2}{\pi(n+1)}\int_0^\infty \frac{1-e^{-t}}{1+e^t}\left\{\frac{n+1}{t}-\frac{2}{\exp\left(\frac{2t}{n+1}\right)-1}-1\right\}dt$$

$$-\frac{2}{\pi(n+1)}\int_0^\infty e^{-t}\left\{\frac{2}{1-\exp\left(\frac{-2t}{n+1}\right)}-\frac{n+1}{t}-1\right\}dt$$

$$+\frac{2}{\pi(n+1)}\int_0^\infty\left\{\frac{1-e^{-t}}{1+e^t}-e^{-t}\right\}dt$$

$$=\frac{2}{\pi(n+1)}\int_0^\infty\left\{\frac{1-e^{-t}}{1+e^t}+e^{-t}\right\}\left\{\frac{n+1}{t}-\frac{2}{\exp\left(\frac{2t}{n+1}\right)-1}-1\right\}dt$$

$$-\frac{4}{\pi(n+1)}\int_0^\infty\frac{e^{-t}}{1+e^t}dt.$$

Then from the above and (459) we have

$$M_n=\frac{2}{\pi}\log(n+1)+\frac{2}{\pi}\left(\log\frac{4}{\pi}+\gamma\right)$$

$$+\frac{4}{\pi(n+1)}\int_0^\infty\left\{\frac{n+1}{t}-\frac{2}{\exp\left(\frac{2t}{n+1}\right)-1}-1\right\}\frac{dt}{1+e^t}$$

$$-\frac{2}{\pi(n+1)}\int_0^\infty\frac{e^{-t}}{1-e^{-t}}\tanh\frac{t}{4(n+1)}dt, \qquad n \text{ even} \quad (463)$$

which is a form suitable for asymptotic expansion. We have

$$\tanh x = \sum_{k=1}^{\infty} 2^{2k}(2^{2k}-1)B_{2k}\frac{x^{2k-1}}{(2k)!} \tag{464}$$

$$\begin{aligned}\int_0^{\infty} t^k \frac{dt}{1+e^t} &= \int_0^{\infty} t^k(e^{-t}-e^{-2t}+e^{-3t}+\ldots)\,dt \\ &= k!\left(1-\frac{1}{2^{k+1}}+\frac{1}{3^{k+1}}-\frac{1}{4^{k+1}}+\ldots\right) \\ &= k!\left(1-\frac{1}{2^k}\right)\zeta(k+1) \end{aligned} \tag{465}$$

where

$$\zeta(n) = \sum_{k=1}^{\infty}\frac{1}{k^n}, \quad n>1. \tag{466}$$

Also

$$\begin{aligned}\int_0^{\infty} \frac{e^{-t}}{1-e^{-t}} t^k\,dt &= \int_0^{\infty} t^k(e^{-t}+e^{-2t}+e^{-3t}\ldots)\,dt \\ &= k!\left(1+\frac{1}{2^{k+1}}+\frac{1}{3^{k+1}}\ldots\right) \\ &= k!\,\zeta(k+1). \end{aligned} \tag{467}$$

From (460) and (465) we have

$$\begin{aligned}&\frac{4}{\pi(n+1)}\int_0^{\infty}\left\{\frac{n+1}{t}-\frac{2}{\exp\left(\frac{2t}{n+1}\right)-1}-1\right\}\frac{dt}{1+e^t} \\ &\quad\approx -\frac{4}{\pi}\sum_{k=1}^{\infty}\frac{B_{2k}2^{2k}(2k-1)!}{(2k)!\,(n+1)^{2k}}\left(1-\frac{1}{2^{2k-1}}\right)\zeta(2k) \\ &\quad\approx -\frac{2}{\pi}\sum_{k=1}^{\infty}(2^{2k}-2)\frac{B_{2k}}{k}\frac{\zeta(2k)}{(n+1)^{2k}} \end{aligned} \tag{468}$$

and from (467) and (464)

$$\begin{aligned}&-\frac{2}{\pi(n+1)}\int_0^{\infty}\frac{e^{-t}}{1-e^{-t}}\tanh\frac{t}{4(n+1)}\,dt \\ &\quad\sim -\frac{2}{\pi}\sum_{k=1}^{\infty}2^{2k}(2^{2k}-1)\frac{B_{2k}}{(2k)!4^{2k-1}(n+1)^{2k}}(2k-1)!\,\zeta(2k) \\ &\quad\sim -\frac{2}{\pi}\sum_{k=1}^{\infty}(2-2^{1-2k})\frac{B_{2k}}{k}\frac{\zeta(2k)}{(n+1)^{2k}}. \end{aligned} \tag{469}$$

From (445) we have

$$B_{2k} = (-1)^{k+1} 2 \frac{(2k)!}{(2\pi)^{2k}} \zeta(2k). \tag{470}$$

Adding (468) and (469), using (470), we obtain

$$\begin{aligned} M_n \sim & \frac{2}{\pi} \log (n+1) + \frac{2}{\pi} \left(\log \frac{4}{\pi} + \gamma \right) \\ & - \frac{7}{36} \frac{\pi}{(n+1)^2} + \frac{127}{21600} \frac{\pi^3}{(n+1)^4} + \cdots \\ & + \frac{1}{\pi} (-1)^k (2^{4k} - 2) \frac{B_{2k}^2}{k(2k)!} \frac{\pi^{2k}}{(n+1)^{2k}} + \cdots, \qquad n \text{ even.} \end{aligned} \tag{471}$$

In the same manner as before, we can establish that

$$\begin{aligned} M_n < & \frac{2}{\pi} \log (n+1) + \frac{2}{\pi} \left(\log \frac{4}{\pi} + \gamma \right) \\ > & \frac{2}{\pi} \log (n+1) + \frac{2}{\pi} \left(\log \frac{4}{\pi} + \gamma \right) - \frac{7}{36} \frac{\pi}{(n+1)^2} \\ < & \frac{2}{\pi} \log (n+1) + \frac{2}{\pi} \left(\log \frac{4}{\pi} + \gamma \right) - \frac{7}{36} \frac{\pi}{(n+1)^2} + \frac{127}{21600} \frac{\pi^3}{(n+1)^4} \\ & \text{etc.,} \qquad \text{for } n \text{ even.} \end{aligned} \tag{472}$$

In this case we have two functions to consider in the polynomial approximation problem. First we note that

$$\frac{2}{e^{2x} - 1} + 1 = \coth x \tag{473}$$

and

$$\begin{aligned} \coth x = \frac{d}{dx} \log \sinh x = \frac{d}{dx} \log \left\{ x \prod_{k=1}^{\infty} \left(1 + \frac{x^2}{k^2 \pi^2} \right) \right\} \\ = \frac{1}{x} + 2x \sum_{k=1}^{\infty} \frac{1}{x^2 + k^2 \pi^2}. \end{aligned} \tag{474}$$

Clearly, for $x > 0$,

$$\frac{2}{e^{2x} - 1} - \frac{1}{x} < 0 \tag{475}$$

and

$$\frac{2}{e^{2x} - 1} - \frac{1}{x} + 1 = \coth x - \frac{1}{x} = 2x \sum_{k=1}^{\infty} \frac{1}{x^2 + k^2 \pi^2} > 0. \tag{476}$$

Now we have

$$\frac{\coth x - \dfrac{1}{x}}{x} = 2 \sum_{k=1}^{\infty} \frac{1}{x^2 + k^2\pi^2}. \tag{477}$$

Defining as before

$$P_{2n}(x;k) = \frac{1 - \left(\dfrac{x}{ik\pi}\right)^{2n+2}}{x^2 + k^2\pi^2}, \qquad k = 1,2,3,\ldots \tag{478}$$

and then

$$\begin{aligned} Q_{2n}(x) &\equiv 2 \sum_{k=1}^{\infty} P_{2n}(x;k) \\ &= \frac{\coth x - \dfrac{1}{x}}{x} - (-1)^{n+1} \left(\frac{x}{\pi}\right)^{2n+2} \sum_{k=1}^{\infty} \frac{2}{k^{2n+2}(x^2 + k^2\pi^2)} \end{aligned} \tag{479}$$

we have

$$\frac{\coth x - \dfrac{1}{x}}{x} - Q_{2n}(x) = (-1)^{n+1} \left(\frac{x}{\pi}\right)^{2n+2} \sum_{k=1}^{\infty} \frac{2}{k^{2n+2}(x^2 + k^2\pi^2)}. \tag{480}$$

So $Q_{2n}(x)$ is a polynomial of degree $2n$ which agrees with the first $(2n + 1)$ terms of the Taylor series of $x^{-1}\{\coth x - x^{-1}\}$ and the sign of the difference is $(-1)^{n+1}$.

For the second function we have

$$\tanh x = \frac{d}{dx} \log \cosh x = 2x \sum_{k=1}^{\infty} \frac{1}{x^2 + (2k-1)^2 \dfrac{\pi^2}{4}}. \tag{481}$$

Now we define

$$p_{2n}(x;k) = \frac{1 - \left(\dfrac{2x}{i(2k-1)\pi}\right)^{2n+2}}{x^2 + (2k-1)^2 \dfrac{\pi^2}{4}}, \qquad k = 1,2,\ldots \tag{482}$$

and

$$q_{2n}(x) = 2\sum_{k=1}^{\infty} p_{2n}(x;k)$$

$$= \frac{\tanh x}{x} - (-1)^{n+1}\left(\frac{2x}{\pi}\right)^{2n+2}\sum_{k=1}^{\infty}\frac{1}{(2k-1)^{2n+2}\left\{x^2 + (2k-1)^2\frac{\pi^2}{4}\right\}}. \tag{483}$$

Then the alternating sign of the error

$$\left\{\frac{\tanh x}{x} - q_{2n}(x)\right\}$$

follows; i.e.,

$$\begin{aligned} \tanh x &< x \\ &> x - \frac{x^3}{3} \\ &< x - \frac{x^3}{3} + \frac{2}{15}x^5 \\ &\text{etc.,} \quad \text{for } x > 0. \end{aligned} \tag{484}$$

The inequalities (472) then follow from (463), (476), (480), and (484).

Then for n even or odd we certainly have

$$M_n > \frac{2}{\pi}\log(n+1) + \frac{2}{\pi}\left(\log\frac{4}{\pi} + \gamma\right) - \frac{7}{36}\frac{\pi}{(n+1)^2} \tag{485}$$

or, giving away a little,

$$M_n > \frac{2}{\pi}\log n + \frac{2}{\pi}\left(\log\frac{4}{\pi} + \gamma\right), \qquad n \geq 1. \tag{486}$$

Obviously (486) is true for n odd since for n odd we may replace n in (486) by $(n+1)$, and for n even we consider

$$\log(n+1) - \frac{7}{72}\frac{\pi^2}{(n+1)^2} = \log n + \log\left(1 + \frac{1}{n}\right) - \frac{7}{72}\frac{\pi^2}{(n+1)^2}.$$

Now

$$\log(1+x) = \int_0^x \frac{dt}{1+t} > \int_0^x (1-t)dt = x - \frac{x^2}{2}, \quad x > 0.$$

So

$$\log\left(1+\frac{1}{n}\right)-\frac{7}{72}\frac{\pi^2}{(n+1)^2}>\frac{1}{n}-\frac{1}{2n^2}-\frac{1}{n^2}>0 \quad \text{for } n\geq 2.$$

Thus (486) is valid for $n \geq 1$.

We note that the actual computation of M_n is simplified, particularly for n odd, by making use of the identity

$$\cot\theta=\frac{1+\cos 2\theta}{\sin 2\theta}.$$

Then for $\theta = \pi/2(n+1)$,

$$M_n=\frac{2}{n+1}\sum_{\substack{1\leq k\leq n\\ k \text{ odd}}}\cot k\theta=\frac{2}{n+1}\sum_{\substack{1\leq k\leq n\\ k \text{ odd}}}\frac{1+\cos 2k\theta}{\sin 2k\theta}. \tag{487}$$

For n odd, $k = 1,3,5,\ldots,n$,

$$\cos\frac{2k\pi}{2(n+1)}=-\cos\frac{2(n+1-k)\pi}{2(n+1)}$$

$$\sin\frac{2k\pi}{2(n+1)}=\sin\frac{2(n+1-k)\pi}{2(n+1)}.$$

So

$$M_n=\frac{2}{n+1}\sum_{\substack{1\leq k\leq n\\ k \text{ odd}}}\frac{1}{\sin\dfrac{k\pi}{n+1}}, \quad n \text{ odd}. \tag{488}$$

We find

$$M_1 = 1$$

$$M_2=\frac{2}{\sqrt{3}}=\frac{2\sqrt{3}}{3}=1.1547005\ldots$$

$$M_3=\sqrt{2}=1.4142135\ldots$$

$$M_4=\frac{2}{5}\left[\frac{1}{\sin\dfrac{\pi}{5}}+\frac{2}{\sin\dfrac{2\pi}{5}}\right]=\frac{4}{5\sin\dfrac{2\pi}{5}}[1+\cos\pi/5]$$

$$=1.5216904\ldots$$

$$M_5=\frac{5}{3}=1.6666666\ldots$$

For use in the asymptotic formulas we have

$$\frac{2}{\pi} = 0.636619772367\ldots$$

$$\log\frac{4}{\pi} = 0.241564475270\ldots$$

$$\gamma = 0.577215664901\ldots$$

$$\log\frac{4}{\pi} + \gamma = 0.818780140172\ldots$$

$$\frac{2}{\pi}\left(\log\frac{4}{\pi} + \gamma\right) = 0.521251626\ldots$$

$$\frac{\pi}{18} = 0.174532925\ldots$$

$$\frac{7\pi}{36} = 0.610865238\ldots$$

$$\frac{49\pi^3}{10800} = 0.140676625\ldots$$

$$\frac{127\pi^3}{21600} = 0.182305423\ldots$$

The inequalities (449) and (472) give

$$\begin{aligned} M_1 &> 0.96252282\ldots \\ &< 1.00615605\ldots \\ &> 0.99736376\ldots \end{aligned}$$

$$\begin{aligned} M_2 &< 1.2206499\ldots \\ &> 1.1527760\ldots \\ &< 1.15502669\ldots \end{aligned}$$

$$\begin{aligned} M_3 &> 1.40379402\ldots \\ &< 1.41470233\ldots \\ &> 1.41415281\ldots \end{aligned}$$

$$\begin{aligned} M_4 &< 1.54585162\ldots \\ &> 1.52141701\ldots \\ &< 1.52170870 \end{aligned}$$

$$\begin{aligned} M_5 &> 1.66192113\ldots \\ &< 1.66676926\ldots \\ &> 1.66666071\ldots \end{aligned}$$

APPENDIX D

Proof of Theorem 5

We are given

$$w(t) = e^{z(t)} \tag{489}$$

where the Fourier transform of $z(t)$ vanishes outside $[0,\Omega]$. Also the Fourier transform of $w(t)$ vanishes over (a,b) where $0 \leq a < b$ and

$$b - a > \Omega \tag{490}$$

and we wish to show that $w(t) =$ constant.

We may write

$$w(t) = g(t) + h(t) \tag{491}$$

where the Fourier transform of $g(t)$ vanishes outside $[0,a]$ and the Fourier transform of $h(t)$ vanishes over $(-\infty,b)$. We have

$$\begin{aligned} w'(t) = z'(t)e^{z(t)} &= z'(t)\{g(t) + h(t)\} \\ &= g'(t) + h'(t). \end{aligned} \tag{492}$$

Now the Fourier transform of $z'(t)$ vanishes outside $[0,\Omega]$ and the Fourier transform of $g'(t)$ vanishes outside $[0,a]$. By Corollary 2 of Theorem 2 the Fourier transform of $z'(t)g(t)$ vanishes outside $[0,a + \Omega]$. The Fourier transform of $h'(t)$ vanishes over $(-\infty,b)$ and by Corollary 1 of Theorem 1 the Fourier transform of $z'(t)h(t)$ also vanishes over $(-\infty,b)$. Thus if $K_{a,b}$ is any kernel of L_1 satisfying

$$\begin{aligned} \int_{-\infty}^{\infty} K_{a,b}(t)e^{-i\omega t}dt &= 1, \quad 0 \leq \omega \leq a \\ &= 0, \qquad \omega \geq b \end{aligned} \tag{493}$$

we have

$$\int_{-\infty}^{\infty} w'(s)K_{a,b}(t - s)ds = z'(t)g(t) = g'(t). \tag{494}$$

Now $z(t)$ and $g(t)$ are the restrictions to the real line of entire functions of exponential type; so

$$\frac{g'(\tau)}{g(\tau)} = z'(\tau). \tag{495}$$

Hence $g(\tau)$ is zero free in the entire plane and is, therefore, of the form (Theorem 2.7.1, Ref. 6)

$$g(\tau) = Ae^{i\lambda\tau}. \tag{496}$$

Hence from (495) and (496), $z'(\tau) = i\lambda$, and since $z(\tau)$ is bounded on the

real axis it follows that $\lambda = 0$; i.e., $z(t) =$ constant and hence

$$w(t) = \text{constant}. \tag{497}$$

APPENDIX E

Proof of Theorem 6

We are given

$$w(t) = \log \{z(t)\} \tag{498}$$

where the Fourier transform of $z(t)$ vanishes outside $[0,\Omega]$ and for some positive ϵ

$$|z(t + iu)| \geq \epsilon \quad \text{for } u \geq 0$$
$$-\infty < t < \infty \tag{499}$$

Thus $w(\tau)$ is bounded and analytic in the uhp; so by Theorem 1, the Fourier transform of $w(t)$ vanishes over $(-\infty,0)$. Also we are given that the Fourier transform of $w(t)$ vanishes over (a,b) where $0 \leq a < b$ and

$$b - a > \Omega \tag{500}$$

and wish to show that $z(t) =$ constant.

We proceed as in the proof of Theorem 5 and write

$$w(t) = g(t) + h(t) \tag{501}$$

where the Fourier transform of $g(t)$ vanishes outside $[0,a]$ and the Fourier transform of $h(t)$ vanishes over $(-\infty,b)$. We have

$$w'(t) = \frac{z'(t)}{z(t)} = g'(t) + h'(t) \tag{502}$$

or

$$z'(t) = z(t)g'(t) + z(t)h'(t). \tag{503}$$

Now the Fourier transform of $h'(t)$ vanishes over $(-\infty,b)$ so by Corollary 1 of Theorem 1 the Fourier transform of $z(t)h'(t)$ also vanishes over $(-\infty,b)$. By Corollary 2 of Theorem 2, the Fourier transform of $z(t)g'(t)$ vanishes outside $[0,\alpha + \Omega]$. Since $\alpha + \Omega < b$ we conclude as in the proof of Theorem 5 that

$$z'(t) = z(t)g'(t), \quad \text{or } g'(t) = \frac{z'(t)}{z(t)} \tag{504}$$

and hence that $z(t) = Ae^{i\lambda t}$ and since $g(t)$ is bounded, $\lambda = 0$. Therefore

$$z(t) = \text{constant}. \tag{505}$$

APPENDIX F

Lower Bound on the Degree of Certain Polynomials

Suppose a polynomial of degree ν is of the form

$$P_\nu(z) = 1 + a_n z^n + a_{n+1} z^{n+1} + \ldots a_\nu z^\nu \tag{506}$$

where $|a_n| > 1$, and $P_\nu(z)$ is zero free for $|z| < 1$. Then

$$\nu \geq 2n. \tag{507}$$

To prove this assertion we assume

$$\nu < 2n. \tag{508}$$

Then assuming that $P_\nu(z)$ is zero free for $|z| < 1$, the function

$$\frac{\bar{a}_\nu + \bar{a}_{\nu-1} z + \ldots \bar{a}_n z^{\nu-n} + z^\nu}{1 + a_n z^n + a_{n+1} z^{n+1} + \ldots + a_\nu z^\nu} = f(z) \tag{509}$$

is analytic for $|z| \leq 1$ and

$$|f(e^{i\theta})| = 1, \quad -\pi \leq \theta \leq \pi. \tag{510}$$

Then

$$f(z) = \sum_0^\infty b_k z^k, \quad |z| \leq 1 \tag{511}$$

where

$$b_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(e^{i\theta}) e^{-ik\theta} d\theta. \tag{512}$$

Thus from (510) and (512),

$$|b_k| \leq 1. \tag{513}$$

However, with the assumption $\nu < 2n$ we see from (509) that

$$b_k = \bar{a}_{\nu-k} \quad \text{for } 0 \leq k \leq \nu - n < n. \tag{514}$$

In particular

$$b_{\nu-n} = \bar{a}_n. \tag{515}$$

But $|a_n| > 1$, so (515) contradicts (513) and therefore (508) is false.

REFERENCES

1. M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, New York: Dover, 1965.
2. N. Abramson, "Bandwidth and Spectra of Phase-and-Frequency Modulated Waves," IEEE Trans. Comm. Systems, *CS-11* (December 1963), pp. 407–414.
3. N. I. Achieser, *Theory of Approximation*, New York: Ungar, 1954, pp. 162–165.
4. R. D. Barnard, "On the Spectral Properties of Single-Sideband Angle-Modulated Signals," B.S.T.J., *43*, No. 9 (November 1964), pp. 2811–2838.

5. E. Bedrosian, "The Analytic Signal Representation of Modulated Waveforms," Proc. IRE, *50*, October 1962, pp. 2071–2076.
6. R. P. Boas, Jr., *Entire Functions*, New York: Academic Press, 1954.
7. R. P. Boas, Jr., "Some Theorems on Fourier Transforms and Conjugate Trigonometric Integrals," Trans. Amer. Math. Soc., *40*, 1936; pp. 287–308.
8. G. J. Foschini, "Demodulating Analytically Modulated Signals After Propagation Over Certain Channels," SIAM J. App. Math., *28*, No. 2, March 1975, pp. 282–289.
9. I. S. Gradshetyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 4th ed., New York: Academic Press, 1965.
10. R. E. Kahn and J. B. Thomas, "Bandwidth Properties and Optimum Demodulation of Single-Sideband FM," IEEE Trans. Comm. Tech., *COM-14*, No. 2, April 1966, pp. 113–117.
11. H. J. Landau, "On the Recovery of a Band-Limited Signal, After Instantaneous Companding and Subsequent Band Limiting," B.S.T.J., *39*, No. 2 (March 1960); pp. 351–364.
12. H. J. Landau and W. L. Miranker, "The Recovery of Band-Limited Signals," J. Math. Anal. and App., *2*, No. 1 (February 1961); pp. 97–104.
13. B. F. Logan, "Properties of High Pass Signals," Doctoral Thesis, Elect. Eng. Dept., Columbia University, 1965.
14. B. F. Logan, "Integrals of High-Pass Functions," Bell Telephone Laboratories, unpub. work, February 2, 1973.
15. B. F. Logan, "Hilbert Transform of a Function Having a Bounded Integral and a Bounded Derivative," Bell Telephone Laboratories, unpub. work, November 4, 1968.
16. B. F. Logan, "The Convolution Kernel $(\sin at)/\pi t$ as an Approximate Identity for L_p as $a \to \infty$," Bell Telephone Laboratories, unpub. work, January 31, 1973.
17. J. E. Mazo and J. Salz, "Spectral Properties of Single-Sideband Angle Modulation," IEEE Trans. Comm. Tech., *COM-10*, No. 1 (February 1968), pp. 52–61.
18. R. E. A. C. Paley and N. Wiener, *"Fourier Transforms in the Complex Domain,"* Vol. XIX, New York: Amer. Math. Soc. Colloq. Pub., 1934. Reprinted, Providence, R.I., 1960 (p. 8 and p. 13).
19. G. Szegö, "On Conjugate Trigonometric Integrals," Amer. Jour. Math., 65, 1943; pp. 532–536.
20. E. C. Titchmarsh, *Introduction to the Theory of the Fourier Integral*, 2nd ed., Oxford University Press, 1948, p. 96.
21. H. Voelcker, "On the Origin and Characteristics of Single Sided Angle Modulation," Correspondence IEEE Trans. Comm. Tech., *COM-13*, No. 4 (December 1965], p. 555.
22. H. Voelcker, "Demodulation of Single-Sideband Signals via Envelope Detection," IEEE Trans. Comm. Tech., *COM-14*, No. 1 (February 1966), pp. 22–30.
23. J. J. Werner, "Recovery of the Modulation Signal From Band-Limited Versions of Single-Sided Angle Modulation Signals," Doctoral Thesis, Columbia University, March 1973.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 3, March 1978
Printed in U.S.A.

Inductive Post Arrays in Rectangular Waveguide

By T. A. ABELE

(Manuscript received July 15, 1977)

Previous attempts, based on mode-matching techniques, to obtain precise data for the equivalent circuit of inductive post arrays in rectangular waveguide have consistently failed due to convergence problems. A different formulation is presented for symmetrical post arrays, which is shown to be free from this defect.

I. INTRODUCTION

Waveguide band-pass structures employing cascades of inductive posts have been built for many years. They usually contain a symmetrical arrangement of posts in each cross-section, mostly one, two or three posts. The latter arrangement, for instance, is a favorite for $\lambda/4$-coupled filters, since it strongly reduces higher order mode interaction. This allows $\lambda/4$ spacings to be used instead of the $3\lambda/4$ spacings required for single post filters, thus leading to shorter filters.

In the past all of these structures had to be designed on the basis of measured data for the equivalent circuit of the cross-sectional post arrangement, because the available theoretical calculations[1,2,8,9] are not sufficiently accurate. The obvious problem with measured data is, of course, that two errors are introduced, whose magnitudes are only poorly known: dimensional tolerances of the sample to be measured and errors in the measurement itself.

Previous attempts to obtain theoretical data based on mode-matching techniques have consistently failed due to the convergence problem typically associated with taking a finite number of unknowns out of two sets of infinitely many unknowns. This paper presents a formulation which leads to only one set of infinitely many unknowns in the case of single or double posts. It may thus be expected that, when a finite number of these is taken, no convergence problem will be encountered. One may also speculate that this will continue to be true for arrays involving three or more posts, although in these cases more than one set of infinitely many unknowns is encountered again.

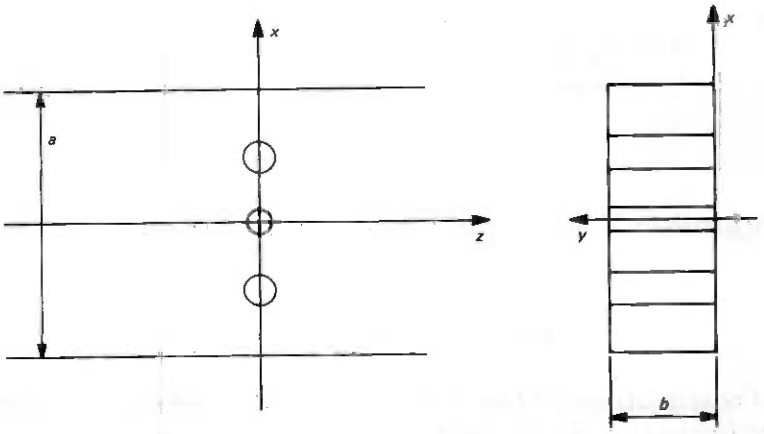


Fig. 1—Post array.

II. CONFIGURATION

We wish to determine the equivalent circuit of the array in Fig. 1 in the plane $z = 0$.

The posts are circular. They are numbered consecutively from $\mu = -M$ to $\mu = M$ with $\mu = 0$ designating the center post. The array is symmetrical with respect to the plane $z = 0$ and the plane $x = 0$. The center post may or may not be present. Each post μ has a diameter d_μ and a coordinate $x = p_\mu$ of its axis. Only dominant (TE_{10}) mode propagation is assumed. The surfaces shall be perfectly conducting.

The electric field will be calculated as the superposition of two fields; the field which exists without the posts, the unperturbed field, and the field generated by the surface currents on the posts, the perturbation field. The surface currents, or rather the coefficients of their Fourier series, are treated as unknowns, which are subsequently determined in such a way that the tangential electric field vanishes on the surface of the posts. As usual, only two special cases of excitation are studied, even and odd, since this suffices to determine the equivalent circuit.

III. UNPERTURBED FIELD

We set

$$E_{y\ \text{even}} = (e^{j\beta_g z} + e^{-j\beta_g z}) \cos \frac{\pi x}{a} \tag{1a}$$

$$E_{y\ \text{odd}} = (e^{j\beta_g z} - e^{-j\beta_g z}) \cos \frac{\pi x}{a} \tag{1b}$$

with

$$\beta_g = \left| \sqrt{\beta^2 - \frac{\pi^2}{a^2}} \right| \tag{2}$$

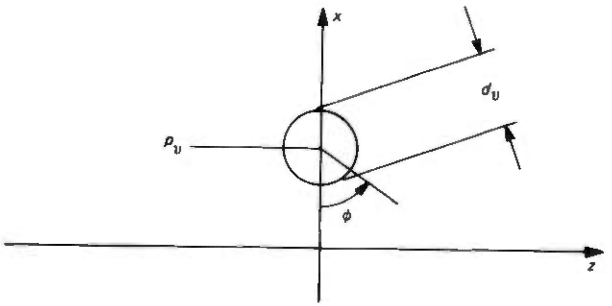


Fig. 2—Post surface.

where (dominant mode assumption)

$$\beta = \frac{2\pi}{\lambda} > \frac{\pi}{a} \tag{3}$$

λ is the wavelength in our medium. Obviously the fields in Eqs. (1) fulfill the boundary conditions everywhere except on the post surfaces.

For later use we wish to develop these fields into Fourier series on the surface of a post ν located at p_ν and of diameter d_ν.

From Fig. 2 we see that the post surface has the coordinates

$$x = p_\nu - \frac{1}{2} d_\nu \cos \phi \tag{4a}$$

$$z = \frac{1}{2} d_\nu \sin \phi \tag{4b}$$

Introduction of these expressions into Eqs. (1) and use of the well-known expansion of $e^{jz \sin\theta}$ (Ref. 5, p. 22), results in

$$E_{y \text{ even}} = \sum_{n=-\infty}^{\infty} J_n \left(\frac{1}{2}\beta d_\nu\right) e^{jn\phi} \times \left[\cos\left(\frac{\pi}{a} p_\nu - n\phi_0\right) + (-1)^n \cos\left(\frac{\pi}{a} p_\nu + n\phi_0\right)\right] \tag{5a}$$

$$E_{y \text{ odd}} = \sum_{n=-\infty}^{\infty} J_n \left(\frac{1}{2}\beta d_\nu\right) e^{jn\phi} \times \left[\cos\left(\frac{\pi}{a} p_\nu - n\phi_0\right) - (-1)^n \cos\left(\frac{\pi}{a} p_\nu + n\phi_0\right)\right] \tag{5b}$$

where

$$e^{j\phi_0} = \frac{1}{\beta}\left(\beta_g + j\frac{\pi}{a}\right) \tag{6}$$

IV. PERTURBATION FIELD

To determine the field generated by the current distributions on the posts we observe first that all of these currents are independent of y and in the direction of y. Secondly, the effect of the broad waveguide walls can be replaced, making use of the common imaging technique, by assuming that all posts are infinitely long in both y directions, and, again, have current distributions which are independent of y and in the direction of y. Thirdly, by employing the same imaging technique once more, we can replace the effect of the narrow walls by periodically repeating the array of infinitely long posts in both x directions with post locations and current distributions, which are consecutive mirror images of each other. To determine the perturbation field, we can then simply sum up the fields generated by these infinitely many and infinitely long posts, without having to worry about the boundary conditions on the waveguide walls, since they are automatically fulfilled.

From basic electromagnetic theory we get for the electric field generated by a current filament stretching in y direction from $-\infty$ to ∞, located at x_0, z_0, and of strength I_y, only the following component in y-direction

$$E_y = -\frac{j\omega\mu}{4\pi} I_y \int_{-\infty}^{\infty} \frac{e^{-j\beta\sqrt{(x-x_0)^2+y_0^2+(z-z_0)^2}}}{\sqrt{(x-x_0)^2+y_0^2+(z-z_0)^2}}\, dy_0$$

$$= -\frac{\omega\mu}{4} I_y H_0^{(2)}(\beta\sqrt{(x-x_0)^2+(z-z_0)^2}) \tag{7}$$

The latter transformation may be found in Ref. 3, p. 27. μ is the permeability of the medium.

Making use of the symmetry of our structure and summing over all currents on all post surfaces we obtain

$$E_y = -\frac{\omega\mu}{4} \sum_{\mu=-M}^{M} \int_0^{2\pi} I_\mu(\psi) \sum_{k=-\infty}^{\infty} (-1)^k H_0^{(2)}$$

$$\left(\beta\left|\sqrt{\left(z-\frac{1}{2}d_\mu \sin\psi\right)^2+\left(x-p_\mu+ka+\frac{1}{2}d_\mu\cos\psi\right)^2}\right|\right) d\psi \tag{8}$$

Figure 3 explains the quantities $I_\mu(\psi)$, d_μ, p_μ and the coordinates used.

With this expression for the perturbation field we will do two things. First we will determine its value at a large distance to obtain expressions for the elements of the dominant-mode equivalent circuit. Secondly, we will evaluate it on the post surfaces in order to be able to come up with an expression for the boundary condition for the tangential electric field there.

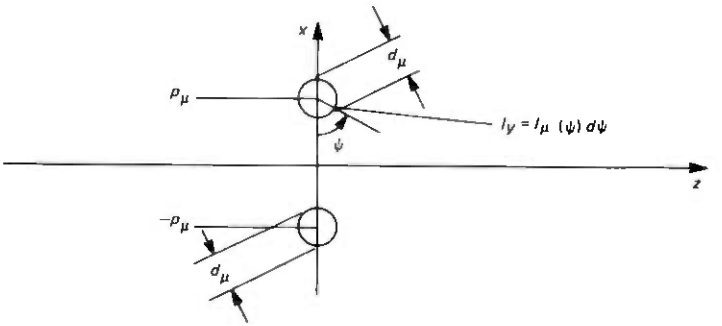


Fig. 3—Post.

V. PERTURBATION FIELD AT A LARGE DISTANCE

If we write

$$B = \frac{1}{a}\left|z - \frac{1}{2}d_\mu \sin\psi\right|$$

$$C = \frac{1}{a}\left(x - p_\mu + \frac{1}{2}d_\mu \cos\left|\right.\right)$$

$$A = \beta a$$

in Eq. (8), we can apply Eq. (34) from Appendix A to Eq. (8), which results in

$$E_y = -\omega\mu \sum_{\mu=-M}^{M} \int_0^{2\pi} I_\mu(\psi)$$

$$\left\{ \sum_{k=1}^{l} \frac{\exp - j\left|\sqrt{\beta^2 - \frac{(2k-1)^2\pi^2}{a^2}}\right| \left|z - \frac{1}{2}d_\mu \sin\psi\right|}{a\left|\sqrt{\beta^2 - \frac{(2k-1)^2\pi^2}{a^2}}\right|} \right.$$

$$\times \cos\left[\frac{(2k-1)\pi}{a}\left(x - p_\mu + \frac{1}{2}d_\mu \cos\psi\right)\right]$$

$$+ j \sum_{k=l+1}^{\infty} \frac{\exp - \left|\sqrt{\frac{(2k-1)^2\pi^2}{a^2} - \beta^2}\right| \left|z - \frac{1}{2}d_\mu \sin\psi\right|}{a\left|\sqrt{\frac{(2k-1)^2\pi^2}{a^2} - \beta^2}\right|}$$

$$\left. \times \cos\left[\frac{(2k-1)\pi}{a}\left(x - p_\mu + \frac{1}{2}d_\mu \cos\psi\right)\right]\right\} d\psi \quad (9)$$

with

$$|z| > \frac{1}{2}d_\mu, \quad \frac{(2l-1)\pi}{a} < \beta < \frac{(2l+1)\pi}{a}$$

The second of the two sums over y obviously represents the evanescent modes in the rectangular waveguide and, therefore, vanishes for large $|z|$. In accordance with our assumptions we have

$$\frac{\pi}{a} < \beta < \frac{3\pi}{a} \tag{10}$$

which means that $l = 1$ in Eq. (9). We therefore find from Eq. (9) for

$$\frac{\pi z}{a} \gg 1$$

$$E_y = -\frac{\omega\mu}{\beta_g a} \sum_{\mu=-M}^{M} \int_0^{2\pi} I_\mu(\psi) e^{-j\beta_g(z-\frac{1}{2}d_\mu \sin\psi)} \times \cos\left[\frac{\pi}{a}\left(x - p_\mu + \frac{1}{2}d_\mu \cos\psi\right)\right] d\psi$$

$$= -\frac{\omega\mu}{2\beta_g a} \sum_{\mu=-M}^{M} \int_0^{2\pi} I_\mu(\psi) \left[e^{-j\beta_g z + j\pi(x-p_\mu)/a} e^{j\frac{1}{2}\beta d_\mu \sin(\psi+\phi_0)} + e^{-j\beta_g z - j\pi(x-p_\mu)/a} e^{j\frac{1}{2}\beta d_\mu \sin(\psi-\phi_0)}\right] d\psi \tag{11}$$

where ϕ_0 is again defined by Eq. (6). Using once more the expansion already used in Eqs. (5) we obtain

$$E_y = -\frac{\omega\mu}{\beta_g a} e^{-j\beta_g z} \sum_{\mu=-M}^{M} \sum_{m=-\infty}^{\infty} J_m\left(\frac{1}{2}\beta d_\mu\right) \times \cos\left[\frac{\pi}{a}(x - p_\mu) + m\phi_0\right] \int_0^{2\pi} I_\mu(\psi) e^{jm\psi} d\psi \tag{12}$$

The inversion of the order of integration and summation employed here presents no difficulty.

Physical considerations tell us that the currents $I_\mu(\psi)$ can be developed into a Fourier series. We write in the usual manner

$$I_\mu(\psi) = \sum_{m=-\infty}^{\infty} c_{\mu,m} e^{jm\psi} \tag{13a}$$

where, of course,

$$c_{\mu,m} = \frac{1}{2\pi} \int_0^{2\pi} I_\mu(\psi) e^{-jm\psi} d\psi \tag{13b}$$

Eq. (13b), when combined with Eq. (12), results in

$$E_y = -\frac{2\pi\omega\mu}{\beta_g a} e^{-j\beta_g z} \sum_{\mu=-M}^{M} \sum_{m=-\infty}^{\infty} (-1)^m c_{\mu,m} J_m\left(\frac{1}{2}\beta d_\mu\right) \times \cos\left[\frac{\pi}{a}(x - p_\mu) - m\phi_0\right] \tag{14}$$

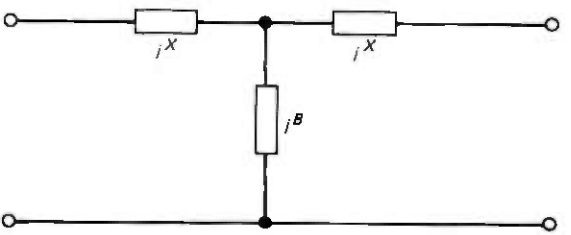


Fig. 4—Equivalent circuit.

For $x = 0$ (center of the guide) this results in

$$E_y = -\frac{2\pi\omega\mu}{\beta_g a} e^{-j\beta_g z} \sum_{\mu=-M}^{M} \sum_{m=-\infty}^{\infty} (-1)^m c_{\mu,m} J_m\left(\frac{1}{2}\beta d_\mu\right) \times \cos\left(\frac{\pi}{a} p_\mu + m\phi_0\right) \quad (15)$$

Combining this result with Eqs. (1) for $x = 0$ we find for the total (unperturbed plus perturbation) field for $\pi z/a \gg 1$

$$E_{y_{\text{odd}}^{\text{even}}} = e^{j\beta_g z} \pm e^{-j\beta_g z} - \frac{2\pi\omega\mu}{\beta_g a} e^{-j\beta_g z} \times \sum_{\mu=-M}^{M} \sum_{m=-\infty}^{\infty} (-1)^m c_{\mu,m_{\text{odd}}^{\text{even}}} J_m\left(\frac{1}{2}\beta d_\mu\right) \cos\left(\frac{\pi}{a} p_\mu + m\phi_0\right) \quad (16)$$

$c_{\mu,m_{\text{odd}}^{\text{even}}}$ is written here to distinguish between the values of $c_{\mu,m}$ for even and odd excitation. For the equivalent circuit of Fig. 4, which is valid for the plane $z = 0$, we obtain from Eq. (16) the reflection coefficient

$$\rho_{\text{odd}}^{\text{even}} = \pm 1 - \frac{2\pi\omega\mu}{\beta_g a} \times \sum_{\mu=-M}^{M} \sum_{m=-\infty}^{\infty} (-1)^m c_{\mu,m_{\text{odd}}^{\text{even}}} J_m\left(\frac{1}{2}\beta d_\mu\right) \cos\left(\frac{\pi}{a} p_\mu + m\phi_0\right) \quad (17)$$

with

$$\rho_{\text{even}} = \frac{jX + \frac{2}{jB} - 1}{jX + \frac{2}{jB} + 1} \quad (18a)$$

$$\rho_{\text{odd}} = \frac{jX - 1}{jX + 1} \quad (18b)$$

These equations permit the calculation of X and B once the values of $c_{\mu,m_{\text{odd}}^{\text{even}}}$ are known.

We note, that because of the structural symmetry with respect to $x = 0$

$$c_{-\mu,m} = (-1)^m c_{\mu,-m} \tag{19}$$

for $\mu = 0, \pm 1, \pm 2 \ldots$ Also, since we have

$$I_{\mu_{\text{odd}}^{\text{even}}}(\psi) = \pm I_{\mu_{\text{odd}}^{\text{even}}}(2\pi - \psi) \tag{20}$$

it follows that

$$c_{\mu,-m_{\text{odd}}^{\text{even}}} = \pm c_{\mu,m_{\text{odd}}^{\text{even}}} \tag{21}$$

Eqs. (19) and (21) permit reduction of $c_{\mu,m}$ for negative values of μ and/or m to those with positive values.

VI. PERTURBATION FIELD ON POST SURFACES

Referring once more to Fig. 2 we get for the perturbation field on the surface of a post ν from Eqs. (8) and (4)

$$E_y = -\frac{\omega\mu}{4} \sum_{\mu=-M}^{M} \int_0^{2\pi} I_\mu(\psi) \sum_{k=-\infty}^{\infty} (-1)^k \times H_0^{(2)} \left(\beta \left| \left(\frac{1}{2} d_\nu \sin\phi - \frac{1}{2} d_\mu \sin\psi\right)^2 + \left(p_\nu - p_\mu + ka - \frac{1}{2} d_\nu \cos\phi + \frac{1}{2} d_\mu \cos\psi\right)^2 \right|^{1/2}\right) d\psi \tag{22}$$

We wish to write for this a double Fourier series with ϕ and ψ as independent variables. This can be done with the aid of the so-called "addition" theorem (Ref. 5, p. 361) if we impose the condition that the posts do not penetrate or touch each other or the narrow walls of the waveguide. We obtain

$$\begin{aligned} E_y = -\frac{\omega\mu}{4} \sum_{\mu=-M}^{M} \int_0^{2\pi} I_\mu(\psi) \times \Bigg[& \sum_{\substack{k=0 \\ ka+p_\nu-p_\mu>0}}^{\infty} \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} (-1)^{m+k} J_m\left(\frac{1}{2}\beta d_\mu\right) J_n\left(\frac{1}{2}\beta d_\nu\right) \\ & \times H_{n+m}^{(2)}\{\beta(p_\nu - p_\mu + ka)\} e^{j(n\phi+m\psi)} \\ & + \sum_{\substack{k=0 \\ ka-p_\nu+p_\mu>0}}^{\infty} \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} J_m\left(\frac{1}{2}\beta d_\mu\right) J_n\left(\frac{1}{2}\beta d_\nu\right) \\ & \times H_{n+m}^{(2)}\{\beta(p_\mu - p_\nu + ka)\} e^{j(n\phi+m\psi)} (-1)^{n+k} \\ & + \lim_{\substack{\kappa\to 1 \\ \mu=\nu \text{ only}}} \sum_{n=-\infty}^{\infty} J_n\left(\frac{1}{2}\beta d_\nu\right) H_n^{(2)}\left(\frac{1}{2}\kappa\beta d_\nu\right) e^{jn(\phi-\psi)} \Bigg] d\psi \end{aligned} \tag{23}$$

Based on physical reasoning (summing contributions of current filaments in different order, integrating around each post before summing) we now exchange the order of summations and integration in Eq. (23) and carry out the integration. This leads to

$$E_y = -\frac{\omega\mu\pi}{2}\sum_{n=-\infty}^{\infty} J_n\left(\frac{1}{2}\beta d_\nu\right)e^{jn\phi}\left[\sum_{\mu=-M}^{M}\sum_{m=-\infty}^{\infty} J_m\left(\frac{1}{2}\beta d_\mu\right)c_{\mu,m}\right.$$
$$\times \sum_{\substack{k=-\infty \\ p_\nu - p_\nu + ka \neq 0}}^{\infty}(-1)^k H^{(2)}_{m-n}(\beta|p_\mu - p_\nu + ka|)[\operatorname{sgn}(p_\mu - p_\nu + ka)]^{n+m}$$
$$\left. + \lim_{\kappa\to 1} H_n^{(2)}\left(\frac{1}{2}\kappa\beta d_\nu\right)c_{\nu,n}\right] \quad (24)$$

With the abbreviation

$$\sum_{\substack{k=-\infty \\ A+kB\neq 0}}^{\infty}(-1)^k H_m^{(2)}(|A+kB|)\left[\operatorname{sgn}\left(\frac{A}{B}+k\right)\right]^m$$
$$= f_m(A,B) = (-1)^m f_{-m}(A,B) = (-1)^m f_m(-A,B) \quad (25)$$

we can write this as

$$E_y = -\frac{\omega\mu\pi}{2}\sum_{n=-\infty}^{\infty} J_n\left(\frac{1}{2}\beta d_\nu\right)e^{jn\phi}$$
$$\times\left[\sum_{\mu=-M}^{m}\sum_{m=-\infty}^{\infty} J_m\left(\frac{1}{2}\beta d_\mu\right)c_{\mu,m}f_{m-n}\{\beta(p_\mu - p_\nu),\beta a\}\right.$$
$$\left. + \lim_{\kappa\to 1} H_n^{(2)}\left(\frac{1}{2}\kappa\beta d_\nu\right)c_{\nu,n}\right] \quad (26)$$

If we take this result for the perturbation field, add it to the incident field Eqs. (5) and impose the condition $E_y = 0$ on the surface of the post, we get (letting $\kappa \to 1$)

$$\sum_{n=-\infty}^{\infty} J_n\left(\frac{1}{2}\beta d_\nu\right)e^{jn\phi}\left[\cos\left(\frac{\pi}{a}p_\nu - n\phi_0\right) \pm (-1)^n\cos\left(\frac{\pi}{a}p_\nu + n\phi_0\right)\right]$$
$$= \frac{\omega\mu\pi}{2}\sum_{n=-\infty}^{\infty} J_n\left(\frac{1}{2}\beta d_\nu\right)e^{jn\phi}\left[\sum_{\mu=-M}^{M}\sum_{m=-\infty}^{\infty} J_m\left(\frac{1}{2}\beta d_\mu\right)\right.$$
$$\left.\times c_{\mu,m_{\text{odd}}^{\text{even}}} f_{m-n}\{\beta(p_\mu - p_\nu),\beta a\} + H_n^{(2)}\left(\frac{1}{2}\beta d_\nu\right)c_{\nu,n_{\text{odd}}^{\text{even}}}\right] \quad (27)$$

Because of the uniqueness of Fourier series (Ref. 4, p. 186), this results in

$$\cos\left(\frac{\pi}{a}p_\nu - n\phi_0\right) \pm (-1)^n\cos\left(\frac{\pi}{a}p_\nu + n\phi_0\right)$$
$$= \frac{\omega\mu\pi}{2}\left[\sum_{\mu=-M}^{M}\sum_{m=-\infty}^{\infty} J_m\left(\frac{1}{2}\beta d_\mu\right)c_{\mu,m_{\text{odd}}^{\text{even}}} f_{m-n}\{\beta(p_\mu - p_\nu),\beta a\}\right.$$
$$\left. + H_n^{(2)}\left(\frac{1}{2}\beta d_\nu\right)c_{\nu,n_{\text{odd}}^{\text{even}}}\right] \quad (28)$$

This equation holds for $n = 0, 1, 2 \ldots$. It expresses the boundary condition on the surface of post ν for even and odd excitation. If applied to all posts $\nu = 0, \pm 1, \pm 2 \ldots \pm M$, it expresses the boundary condition on all posts. However, because of the symmetry involved, only $\nu = 0, 1, 2 \ldots M$ are needed. As before, Eqs. (19) and (21) permit reduction of $c_{\mu,m}$ for negative values of μ and/or m to those with positive values. In summary we can say that Eq. (28), if applied for $n = 0, 1, 2, \ldots$ and $\nu = 0, 1, 2 \ldots M$, will allow us to compute all of the unknown coefficients $c_{\nu,m}$. In turn, Eq. (17) will then allow us to compute the elements of the equivalent circuit, which means that our problem is solved.

Appendix A and Appendix B provide expressions suitable for the computation of $f_m(A,B)$ in Eq. (28). These alternate expressions are essential, because the defining series Eq. (25) converges very slowly, as the magnitude of $H_m^{(2)}(z)$ decreases only with $|z^{-1/2}|$ for large z. The derivation of these expressions constitutes the most difficult and laborious part of this analysis. For convenience the results are repeated below in the form most appropriate for Eq. (28). From Eqs. (40) and Eq. (41),

$$f_m(\beta p, \beta a) = \frac{4}{\pi} \tan \phi_0 \cos\left(\frac{\pi p}{a} - \frac{m\pi}{2}\right) e^{jm(\phi_0 - \pi/2)}$$

$$+ j\frac{1}{\pi} \sum_{n=3,5\ldots} \frac{\dfrac{2\lambda}{a} \cos\left(\dfrac{n\pi p}{a} - \dfrac{m\pi}{2}\right)}{\left|\sqrt{\left(\dfrac{n\lambda}{2a}\right)^2 - 1}\right| \left[\dfrac{n\lambda}{2a} + \left|\sqrt{\left(\dfrac{n\lambda}{2a}\right)^2 - 1}\right|\right]^m}$$

$$+ j\frac{1}{\pi} \sum_{n=0}^{< \frac{m}{2}} \frac{(m-n-1)!}{n!(m-2n-1)!} \times \left[\frac{\lambda}{a \sin \dfrac{\pi p}{a}}\right]^{m-2n} h_{m-2n-1}\left(\cos \frac{\pi p}{a}\right) \quad (29a)$$

$$f_{2m-1}(\beta Na, \beta a) = 0 \quad (29b)$$

$$f_{2m}(\beta Na, \beta a) = \frac{4}{\pi} \tan \phi_0 (-1)^{N+m} e^{j2m(\phi_0 - \pi/2)}$$

$$+ j\frac{1}{\pi} \sum_{n=3,5\ldots} \frac{\dfrac{2\lambda}{a}(-1)^{N+m}}{\left|\sqrt{\left(\dfrac{n\lambda}{2a}\right)^2 - 1}\right| \left[\dfrac{n\lambda}{2a} + \left|\sqrt{\left(\dfrac{n\lambda}{2a}\right)^2 - 1}\right|\right]^{2m}}$$

$$+ j\frac{1}{\pi} \sum_{n=0}^{m} \frac{2(-1)^{N+n}(m+n-1)!(2^{2n-1}-1)B_{2n}}{(m-n)!(2n)!} \left(\frac{\lambda}{a}\right)^{2n} \quad (29c)$$

From Eq. (34e) and Eq. (34f)

$$f_0(\beta p, \beta a) = \frac{4}{\pi} \frac{\cos \frac{\pi p}{a}}{\cos \phi_0} e^{j(\phi_0 - \pi/2)} + j \frac{1}{\pi} \ln \left(\cot^2 \frac{\pi p}{2a} \right)$$

$$+ j \frac{1}{\pi} \sum_{n=3,5\ldots} \frac{4 \cos \frac{n\pi p}{a}}{\left| \sqrt{\left(\frac{n\lambda}{2a}\right)^2 - 1} \right| \left[\frac{n\lambda}{2a} + \left| \sqrt{\left(\frac{n\lambda}{2a}\right)^2 - 1} \right| \right] n} \tag{29d}$$

$$f_0(\beta N a, \beta a) = \frac{4}{\pi} \frac{(-1)^N}{\cos \phi_0} e^{j(\phi_0 - \pi/2)} - (-1)^N + j \frac{2}{\pi} \left(C + \ln \frac{2a}{\lambda} \right) (-1)^N$$

$$+ j \frac{1}{\pi} \sum_{n=3,5\ldots} \frac{4(-1)^N}{\left| \sqrt{\left(\frac{n\lambda}{2a}\right)^2 - 1} \right| \left[\frac{n\lambda}{2a} + \left| \sqrt{\left(\frac{n\lambda}{2a}\right)^2 - 1} \right| \right] n} \tag{29e}$$

These expressions are valid if

$$2a > \lambda > \frac{2a}{3}$$

$$p \neq 0, \pm a, \pm 2a \ldots$$

$$m = 1, 2, 3, \ldots$$

$$N = 0, \pm 1, \pm 2 \ldots$$

The polynomials $h_m(x)$ are defined in Appendix B, Eq. (42). B_n is the nth Bernoullian number and C is Euler's constant.

VII. NUMERICAL RESULTS

A series of calculations was made to investigate the question of convergence and to ascertain that the rather involved analysis is error-free. To this end the reactances $X - 2/B$ and $-X$ [Eqs. (18)] were calculated for the cases of single, double, and triple posts with $\lambda/a = 1.2$, $p_1/a = 0.25$ and $\beta d_0 = \beta d_1 = 0.2$ and 0.4, employing increasing numbers of variables and equations. Furthermore, the computed results were compared with measured data where such data were available. Lacking a full-fledged computer program the calculations were carried out with the aid of a programmable desk calculator, except for the matrix inversion, for which a general-purpose computer program was used.

Table I summarizes the results of this work. The first observation that can be made is that, as expected, the convergence obtained for single and double posts is excellent. Three terms in the Fourier series for the post currents is all that is needed to obtain six place accuracy for the reac-

Table I

Number of posts	$\frac{p_1}{a}$	βd_0	βd_1	n_{max}	$X - \frac{2}{B}$	$-X$
1	—	0.2	—	0	1.121835970	—
				1	—	.009450748398
				2	1.121835438	—
				3	—	.009450749381
				4	1.121835438	—
				Meas.	1.12	.010
		0.4		0	.6546985053	—
				1	—	.03659710000
				2	.6546719818	—
				3	—	.03659716655
				4	.6546719813	—
				Meas.	.655	.037
2	0.25	—	0.2	0	1.121835969	—
				1	1.100680471	.009467748389
				2	1.100679708	.009467748828
				3	1.100679707	.009467799814
				4	1.100679707	.009467799806
			0.4	0	.6546985046	—
				1	.6071886432	.03659710003
				2	.6071426563	.03685942441
				3	.6071423836	.03685949132
				4	.6071423830	.03685949138
3	0.25	0.2	0.2	0	.2599117670	—
				1	.2578995041	.01871192452
				2	.2578489230	.01872800993
				3	.2578488656	.01872801335
				4	.2578488652	.01872801340
				Meas.	.255	.020
		0.4	0.4	0	.02634303527	—
				1	.02329626388	.07043107151
				2	.02238859942	.07063929772
				3	.02238444183	.07063993488
				4	.02238437265	.07063993658
				Meas.	.0265	.074

tances, which is more than enough for any technical application. The second observation is that, as was hoped, excellent convergence continues to exist for triple posts, even though in that case two sets of infinitely many unknowns are encountered instead of just one. Even for posts with susceptance values as high as $B = 20$ no more than four terms in the Fourier series are needed to obtain five-place accuracy. Presumably the analysis will converge even for four or more posts, but these arrangements are of little technical interest and thus probably not worth investigating. Finally, when comparing the computed values with measured data obtained with the aid of a very precise, computer-operated transmission measurement set,[10] sufficient agreement is found to ascertain that the analysis is free from any fundamental error.

APPENDIX A

We study the following series

$$f(z, B, C, t) = \sum_{n=-\infty}^{\infty} \frac{e^{-B\sqrt{[(2n-1)\pi+t]^2+z^2}} \cos\{[(2n-1)\pi + t]C\}}{\sqrt{[(2n-1)\pi+ t]^2 + z^2}} \quad (30)$$

with t as a real variable, $B \geq 0$ and real, C real, $\mathrm{Re}\{z\} > 0$, $\mathrm{Re}\{\sqrt{[(2n-1)\pi + t]^2 + z^2}\} > 0$. This function is even in t and also periodic in t with the period 2π. For reasons which will become apparent later we wish to develop it into a Fourier series in t

$$\frac{1}{2}a_0 + \sum_{k=1}^{\infty} a_k \cos kt \tag{31}$$

Without going into the fairly laborious detail the result is

$$f(z, B, C, t) = \frac{1}{\pi}\sum_{k=-\infty}^{\infty} (-1)^k K_0(z|\sqrt{B^2 + (C+k)^2}|) \cos kt \tag{32}$$

for the conditions stated for Eq. (30) plus either $B > 0$ or $B = 0$ and $C \neq 0, \pm 1, \pm 2 \ldots$. Setting $t = 0$ leads to

$$2\sum_{n=1}^{\infty} \frac{e^{-B\sqrt{(2n-1)^2\pi^2+z^2}} \cos[(2n-1)C]}{\sqrt{(2n-1)^2\pi^2 + z^2}} = \frac{1}{\pi}\sum_{k=-\infty}^{\infty} (-1)^k K_0(z|\sqrt{B^2 + (C+k)^2}|) \tag{33}$$

provided $\mathrm{Re}\{z\} > 0$, $\mathrm{Re}\{\sqrt{(2n-1)^2\pi^2 + z^2}\} > 0$ and either $B > 0$ and real, C real or $B = 0$, $C \neq 0, \pm 1, \pm 2 \ldots$ and real. The validity of Eq. (43) can be extended to include $\mathrm{Re}\{z\} = 0$ by analytic continuation. In doing this the points $z = 0$ and $z = \pm j(2n-1)\pi$ obviously have to be excluded, since at these points individual terms of the sums involved are not analytic. The result of the analytic continuation is for $z \to jA$

$$\sum_{k=-\infty}^{\infty} (-1)^k H_0^{(2)}(A|\sqrt{B^2 + (C+k)^2}|) = 4j\sum_{n=1}^{\infty} \frac{e^{-B\sqrt{(2n-1)^2\pi^2 - A^2}}}{\sqrt{(2n-1)^2\pi^2 - A^2}} \cos[(2n-1)\pi C] \tag{34a}$$

with

$$A > 0, \text{ real}$$

$$A \neq \pi, 3\pi, 5\pi \ldots \tag{34b}$$

$$\arg\sqrt{(2n-1)^2\pi^2 - A^2} = 0 \text{ or } \frac{\pi}{2}$$

and either

$$B > 0, \text{ real}$$

$$C \text{ real} \tag{34c}$$

or

$$B = 0$$

$$C \neq 0, \pm 1, \pm 2 \ldots, \text{ real} \tag{34d}$$

For the latter situation [Eq. (34d)] Eq. (34a) may be rewritten in the more rapidly converging form

$$\sum_{k=-\infty}^{\infty}(-1)^k H_0^{(2)}(A|C+k|) = j\frac{1}{\pi}\ln\left(\cot^2\frac{1}{2}\pi A\right) + 4j\sum_{n=1}^{\infty} \frac{A^2\cos\left[(2n-1)\pi C\right]}{(2n-1)\pi\sqrt{(2n-1)^2\pi^2 - A^2}\left[(2n-1)\pi + \sqrt{(2n-1)^2\pi^2 - A^2}\right]} \tag{34e}$$

We also need a result corresponding to Eq. (34e) for $C = N = 0, \pm 1, \pm 2 \ldots$. We observe first that in the left hand sum the term $k = -N$ has to be excluded for obvious reasons. Furthermore it is

$$\sum_{\substack{k=-\infty \\ k\neq -n}}^{\infty}(-1)^k H_0^{(2)}(A|N+K|) = (-1)^N \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty}(-1)^k H_0^{(2)}(A|k|)$$

$$= 2(-1)^N \sum_{k=1}^{\infty}(-1)^k H_0^{(2)}(Ak) \tag{35}$$

An alternate expression for the last sum is known (Ref. 6, p. 333). We get

$$\sum_{\substack{k=-\infty \\ k\neq -N}}^{\infty}(-1)^k H_0^{(2)}(A|N+k|) = (-1)^N\left\{-1 + 2j\frac{1}{\pi}\left(C + \ln\frac{A}{4\pi}\right) + 4j\sum_{n=1}^{\infty}\left[\frac{1}{\sqrt{(2n-1)^2\pi^2 - A^2}} - \frac{1}{2n\pi}\right]\right\}$$

$$= (-1)^N\left\{-1 + 2j\frac{1}{\pi}\left(C + \ln\frac{A}{\pi}\right) + 4j\sum_{n=1}^{\infty} \frac{A^2}{(2n-1)\pi\sqrt{(2n-1)^2\pi^2 - A^2}\left[(2n-1)\pi + \sqrt{(2n-1)^2\pi^2 - A^2}\right]}\right\} \tag{34f}$$

with $N = 0, \pm 1, \pm 2 \ldots$ and Eqs. (34b) in force. Note that C in this last formula is Euler's constant.

APPENDIX B

We study for $m = 1, 2 \ldots$ the series

$$f_m(z, C, t) = \sum_{n=1}^{\infty} \frac{z^m \cos\left\{\left[(2n-1)\pi + t\right]C - m\frac{\pi}{2}\right\}}{\sqrt{[(2n-1)\pi + t]^2 + z^2}\left[(2n-1)\pi + t + \sqrt{[(2n-1)\pi + t]^2 + z^2}\right]^m} \tag{36}$$

with t as a real variable with the range $-\pi \le t \le \pi$, C real $\mathrm{Re}\{z\} > 0$, $\mathrm{Re}\{\sqrt{[(2n-1)\pi + t]^2 + z^2}\} > 0$. Analogous to the situation in Appendix A we wish to develop $f_m(z, C, t) + f_m(z, C, -t)$ into a Fourier series

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos kt$$

over the range $-\pi \le t \le \pi$, thereby continuing it periodically beyond that range. Again omitting the fairly laborious detail we get for $C \neq 0, \pm 1, \pm 2 \ldots$

$$f_m(z, C, t) + f_m(z, C, -t)$$

$$= \frac{1}{8\pi}(-1)^{m+1}[4S_m(jCz)e^{jm\pi/2} - 8K_m(|C|z)(\mathrm{sgn}\ C)^m]$$

$$+ \frac{1}{4\pi}(-1)^{m+1} \sum_{k=1}^{\infty} (-1)^k [2S_m(jCz + jkz)e^{jm\pi/2}$$

$$+ 2S_m(jCz - jkz)e^{jm\pi/2} - 4K_m(|C + k|z)\{\mathrm{sgn}\ (C + k)\}^m$$

$$- 4K_m(|C - k|z)\{\mathrm{sgn}\ (C - k)\}^m] \cos kt = \frac{1}{2\pi}(-1)^m$$

$$\times \sum_{k=-\infty}^{\infty} (-1)^k [2K_m(|C + k|z)\{\mathrm{sgn}\ (C + k)\}^m$$

$$- e^{jm\pi/2} S_m(jCz + jkz)] \cos kt \quad (37a)$$

and for $C = N = 0, \pm 1, \pm 2 \ldots$

$$f_m(z, N, t) + f_m(z, N, -t)$$

$$= \frac{1}{2\pi}(-1)^m \sum_{\substack{k \neq -\infty \\ k \neq -N}}^{\infty} (-1)^k [2K_m(|N + k|z)\{\mathrm{sgn}\ (N + k)\}^m$$

$$- e^{jm\pi/2} S_m(jNz + jkz)] \cos kt + \frac{1}{m\pi}(-1)^N \cos \frac{m\pi}{2} \cos Nt \quad (37b)$$

In these equations $S_m(z)$ denotes Schlaefli's polynomial (Ref. 5, p. 313). For $t = 0$ this results in

$$\sum_{n=1}^{\infty} \frac{z^m \cos\left[(2n-1)\pi C - m\frac{\pi}{2}\right]}{\sqrt{(2n-1)^2\pi^2 + z^2}\,[(2n-1)\pi + \sqrt{(2n-1)^2\pi^2 + z^2}]^m}$$

$$= \frac{1}{4\pi}(-1)^m \sum_{k=-\infty}^{\infty} (-1)^k [2K_m(|C + k|z)\{\mathrm{sgn}\ (C + k)\}^m$$

$$- e^{jm\pi/2} S_m(jCz + jkz)] \quad (38a)$$

provided $C \neq 0, \pm 1, \pm 2 \ldots$, and in

$$\sum_{n=1}^{\infty} \frac{z^m \cos\left[(2n-1)\pi N - m\frac{\pi}{2}\right]}{\sqrt{(2n-1)^2\pi^2 + z^2}\left[(2n-1)\pi + \sqrt{(2n-1)^2\pi^2 + z^2}\right]^m}$$
$$= \frac{1}{4\pi}(-1)^m \sum_{\substack{k=-\infty \\ k\neq -N}}^{\infty} (-1)^k [2K_m(|N+k|z)\,\{\operatorname{sgn}(N+k)\}^m$$
$$- e^{jm\pi/2} S_m(jNz + jkz)] + \frac{1}{2m\pi}(-1)^N \cos\frac{m\pi}{2} \quad (38b)$$

for $N = 0, \pm 1, \pm 2 \ldots$. By working on the two series with Schlaefli's polynomials the following alternative expressions are obtained

$$\sum_{n=1}^{\infty} \frac{z^m \cos\left[(2n-1)\pi C - m\frac{\pi}{2}\right]}{\sqrt{(2n-1)^2\pi^2 + z^2}\left[(2n-1)\pi + \sqrt{(2n-1)^2\pi^2 + z^2}\right]^m}$$
$$= \frac{1}{2\pi}(-1)^m \sum_{k=-\infty}^{\infty} (-1)^k K_m(|C+k|z)[\operatorname{sgn}(C+k)]^m$$
$$+ \frac{1}{4\pi} e^{jm\pi/2} \sum_{n=0}^{<m/2} \frac{(m-n-1)!}{n!(m-2n-1)!}\left(\frac{2\pi}{jz}\right)^{m-2n}$$
$$\times \left[\frac{d^{m-2n-1}}{dx^{m-2n-1}} \frac{1}{\sin x}\right]_{x=C\pi} \quad (39a)$$

$$\sum_{\substack{k=-\infty \\ k\neq -N}}^{\infty} (-1)^k K_{2\lambda-1}(|N+k|z) \operatorname{sgn}(N+k) = 0 \quad (39b)$$

$$\sum_{n=1}^{\infty} \frac{z^{2\lambda}(-1)^{N+\lambda}}{\sqrt{(2n-1)^2\pi^2 + z^2}\left[(2n-1)\pi + \sqrt{(2n-1)^2\pi^2 + z^2}\right]^{2\lambda}}$$
$$= \frac{1}{2\pi} \sum_{\substack{k=-\infty \\ k\neq -N}}^{\infty} (-1)^k K_{2\lambda}(|N+k|z) + \frac{1}{4\lambda\pi}(-1)^{N+\lambda}$$
$$- \frac{1}{2\pi}(-1)^{N+\lambda} \sum_{k=1}^{\lambda} \frac{(\lambda+k-1)!(2^{2k-1}-1)B_{2k}}{(\lambda-k)!(2k)!}\left(\frac{2\pi}{z}\right)^{2k} \quad (39c)$$

where, again, $C \neq 0, \pm 1, \pm 2 \ldots$, real, $N = 0, \pm 1, \pm 2 \ldots$, $m = 1, 2 \ldots$, $\lambda = 1, 2 \ldots$, $\operatorname{Re}\{z\} > 0$, $\operatorname{Re}\{\sqrt{(2n-1)^2\pi^2 + z^2}\} > 0$. Following the same argumentation as in Appendix A the validity of Eqs. (39) can, by analytic continuation, be extended to include $\operatorname{Re}\{z\} = 0$ with the exception of $z = 0$ and $z = \pm j(2n-1)\pi$. The result is for $z \to jA$ with $A > 0$ and after some rearrangement

$$\sum_{k=-\infty}^{\infty} (-1)^k H_m^{(2)}(|C+k|A)[\operatorname{sgn}(C+k)]^m$$

$$= 4j \sum_{n=1}^{\infty} \frac{A^m \cos\left[(2n-1)\pi C - \frac{\pi}{2}\right]}{\sqrt{(2n-1)^2\pi^2 - A^2}\left[(2n-1)\pi + \sqrt{(2n-1)^2\pi^2 - A^2}\right]^m}$$

$$- j(-1)^m \frac{1}{\pi} \sum_{n=0}^{<\frac{m}{2}} \frac{(m-n-1)!}{n!(n-2n-1)!} \left(\frac{2\pi}{A}\right)^{m-2n}$$

$$\times \left[\frac{d^{m-2n-1}}{dx^{m-2n-1}} \frac{1}{\sin x}\right]_{x=C\pi} \tag{40a}$$

$$\sum_{\substack{k=-\infty \\ k\neq -N}}^{\infty} (-1)^k H_{2m-1}^{(2)}(|N+k|A) \operatorname{sgn}(N+k) = 0 \tag{40b}$$

$$\sum_{\substack{k=-\infty \\ k\neq -N}}^{\infty} (-1)^k H_{2m}^{(2)}(|N+k|A) = 4j(-1)^{n+m}$$

$$\times \sum_{n=1}^{\infty} \frac{A^{2m}}{\sqrt{(2n-1)^2\pi^2 - A^2}\left[(2n-1)\pi + \sqrt{(2n-1)^2\pi^2 - A^2}\right]^{2m}}$$

$$+ j(-1)^N \frac{2}{\pi} \sum_{n=0}^{m} \frac{(m+n-1)!(2^{2n-1}-1)B_{2n}(-1)^n}{(m-n)!(2n)!} \left(\frac{2\pi}{A}\right)^{2n} \tag{40c}$$

valid for $m = 1, 2 \ldots, N = 0, \pm 1, \pm 2 \ldots, C \neq 0, \pm 1, \pm 2$ and real, $A > 0$, real and $A \neq \pi, 3\pi, 5\pi \ldots$, and with $\arg\{\sqrt{(2n-1)^2\pi^2 - A^2}\} = 0$ or $\pi/2$. Eq. (40a) can be written in the following more convenient form

$$\sum_{k=-\infty}^{\infty} (-1)^k H_m^{(2)}(|C+k|A)[\operatorname{sgn}(C+k)]^m$$

$$= 4j \sum_{n=1}^{\infty} \frac{A^m \cos\left[(2n-1)\pi C - m\frac{\pi}{2}\right]}{\sqrt{(2n-1)^2\pi^2 - A^2}\left[(2n-1)\pi + \sqrt{(2n-1)^2\pi^2 - A^2}\right]^m}$$

$$+ j \frac{1}{4\pi} \sum_{n=0}^{<\frac{m}{2}} \frac{(m-n-1)!}{n!(m-2n-1)!} \left(\frac{2\pi}{A \sin C\pi}\right)^{m-2n} h_{m-2n-1}(\cos C\pi) \tag{41}$$

subject to the same restrictions as those enumerated for Eqs. (40). The polynomials $h_n(u)$ appearing in this equation are defined by

$$h_n(u) = (-1)^n \left[\sin^{n+1} x \frac{d^n}{dx^n} \frac{1}{\sin x}\right]_{\cos x = u} \tag{42}$$

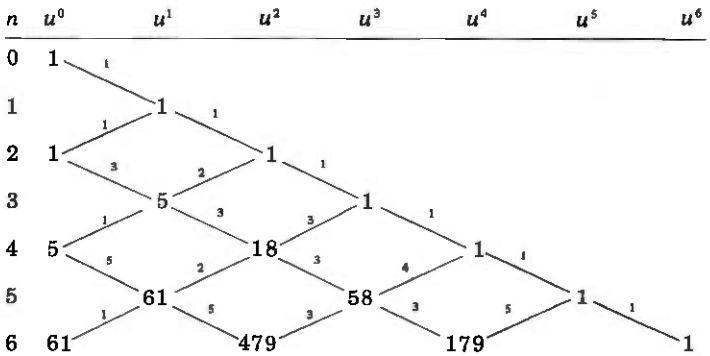
and can be shown to satisfy the following recursion formula

which begins with

$$h_n(u) = nuh_{n-1}(u) + (1-u^2)\frac{dh_{n-1}(u)}{du} \tag{43a}$$

$$h_0(u) = 1 \tag{43b}$$

It appears impossible to give a closed form expression for these polynomials, but their coefficients can easily be calculated by the following scheme which is a consequence of the recursion formula



i.e., it is

$$h_0(u) = 1$$
$$h_1(u) = u$$
$$h_2(u) = 1 + u^2$$
$$h_3(u) = 5u + u^3$$
$$h_4(u) = 5 + 18u^2 + u^4$$

It can be shown, incidentally, that the sum of all coefficients of $h_n(u)$ is equal to $n!$ and that the coefficients of u^0 and u^1 are Euler's numbers. It furthermore appears, but has not been proven, that the coefficients are equal to those in Table 7.2.2 of Ref. 7, p. 260, the "number of permutations of the first N natural numbers with t_u runs up."

REFERENCES

1. N. Marcuvitz, *Waveguide Handbook*, New York: McGraw-Hill, 1951.
2. A. M. Model' and N. S.Belevich, "Calculation of the Loaded Q-Factor of Waveguide Resonators Formed by Grid Diaphragms," *Telecommunications and Radio Engineering*, Part 2 (Radio Engineering), *18*, No. 9, 1963, pp. 15–23.
3. W. Magnus and F. Oberhettinger, *Formulas and Theorems for the Functions of Mathematical Physics*, New York: Chelsea Publishing Company, 1949.
4. E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, London: Cambridge University Press, 1927.
5. G. N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge: Cambridge University Press, 1962.
6. I. M. Ryshik and I. S. Gradstein, *Tables of Series, Products and Integrals*, Berlin: Deutscher Verlag der Wissenschaften, 1957.
7. F. N. David, M. G. Kendall, and D. E. Barton, *Symmetric Function and Allied Tables*, Cambridge: Cambridge University Press, 1966.
8. L. Lewin, *Advanced Theory of Waveguides*, London: Iliffe, 1951.
9. G. Craven and L. Lewin, "Design of Microwave Filters with Quarter Wave Couplings," Proc. IEEE, Part B, *103*, March 1956, pp. 173–177.
10. J. B. Davis, C. F. Hempstead, D. Leed, R. A. Ray, "3700 to 4200 MHz Computer Operated Measurement System for Loss, Phase, Delay and Reflection," IEEE Trans. Instrum. Meas., *IM-21*, No. 1, 1972, pp. 24–37.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 3, March 1978
Printed in U.S.A.

Measurements of Loss Due to Offset, End Separation, and Angular Misalignment in Graded Index Fibers Excited by an Incoherent Source

By T. C. CHU and A. R. McCORMICK

(Manuscript received July 20, 1977)

Transmission losses versus fiber end offset separation, and angular misalignment of graded index fibers excited by an incoherent source, have been measured in two independent experiments. The measurement setup, fiber diameter, and length were different in the two experiments, yet the measurement results are strikingly similar. The loss measurements clearly show that transverse offset is much more critical in connector and splice design than angular misalignment and end separation. Two-tenths of the fiber core radius in transverse offset alone may cause 0.5 dB loss while one fiber core radius in axial separation combined with 1° in angular misalignment may cause 0.5 dB loss.

I. INTRODUCTION

It is essential to know the transmission loss caused by misalignment of the fiber ends in designing fiber connectors and splices. Graded index fibers are important to fiberguide transmission applications that require low dispersion characteristics. The study of the transmission loss caused by misalignment of fibers having graded index profiles is thus necessary. Theoretical investigations of the loss versus offset at zero axial separation have recently been published.[1–4] Further studies of the problem—i.e., loss versus offset, end separation, and angular misalignment of graded index fibers—have been done experimentally.[5–7] This paper presents the results of two separate experiments.

II. EXPERIMENTS

The experiments were conducted independently in different laboratory locations. The first experiment (Fig. 1a) yielded the loss versus offset and end separation only. The second experiment (Fig. 1b) included angular misalignment along with end separation and offset. In both

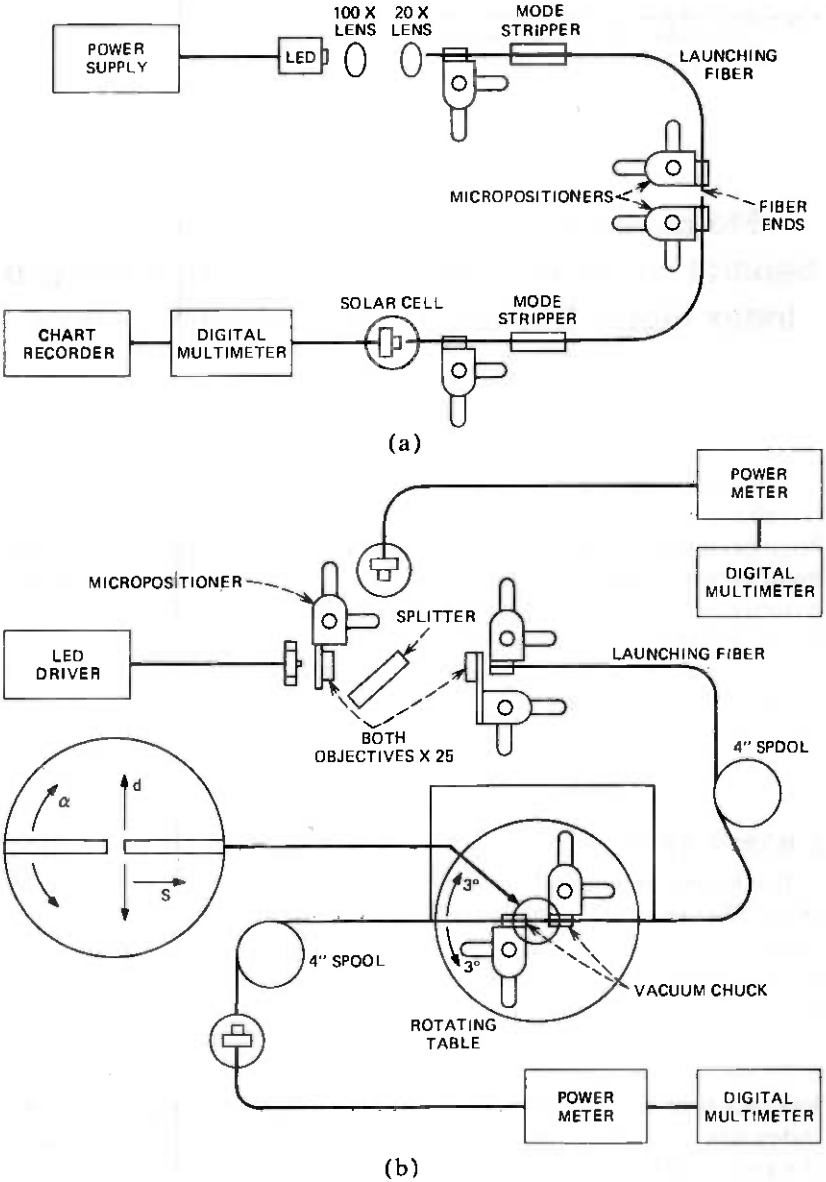


Fig. 1—(a) Coupling loss vs. fiber end misalignment measurement setup in the first experiment. (b) Coupling loss vs. fiber end misalignment measurement setup in the second experiment.

experiments, a Burrus-type LED having a 50 μm diameter emitting surface was used. The LED in the second experiment was internally modulated whereas the first was not modulated. Microscope objectives

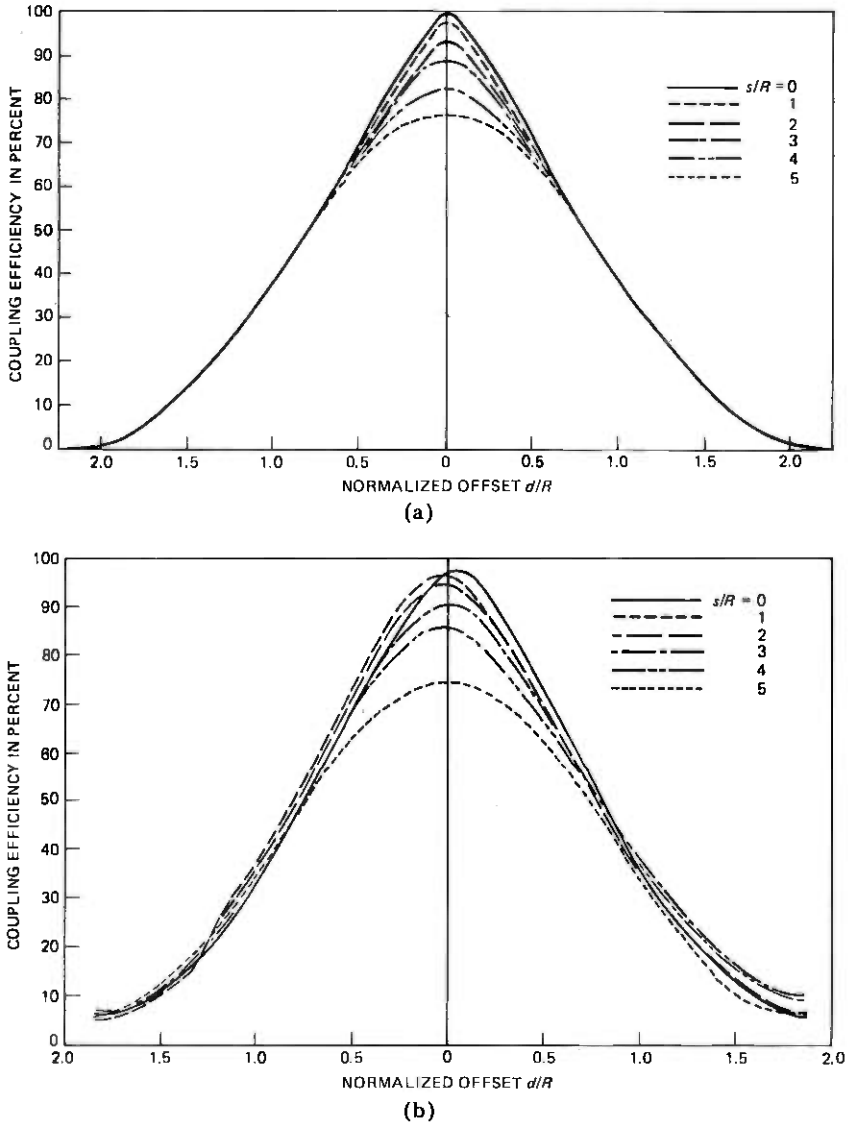


Fig. 2—(a) Coupling efficiency vs. normalized offset d/R at various separations s/R from the first experiment. (b) Coupling efficiency vs. normalized offset d/R at various separations from the second experiment.

were used to collect and focus the light into the launching fiber. Alignment was achieved by using micropositioners. In both experiments the output of the receiving fiber was detected by a power meter and monitored by a digital multimeter.

Graded index fibers were used in both experiments. The first exper-

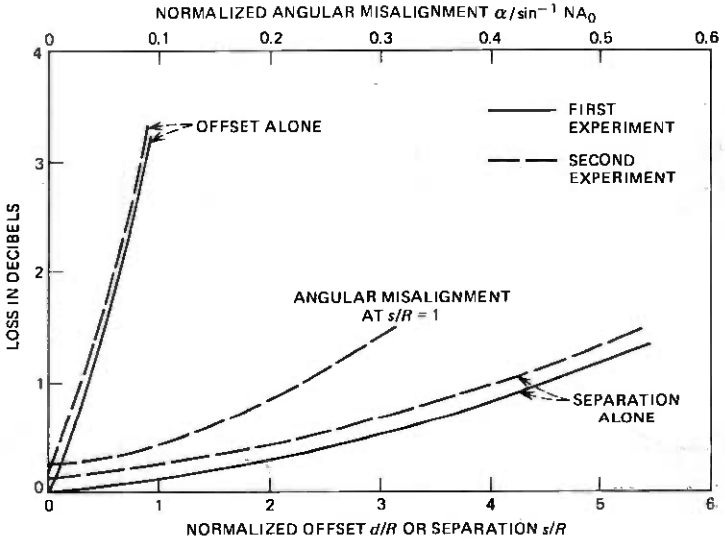


Fig. 3—Loss in dB vs. normalized offset d/R, separation s/R, and angular misalignment $\alpha°/\sin^{-1} NA_0$.

iment used a 50 μm diameter core/100 μm diameter cladding fiber while the second used a 55 μm diameter core/110 μm diameter cladding fiber. The indices of refraction of the core center and cladding of both fibers were 1.472 and 1.458 respectively. A 1.83 m length fiber was used in the first experiment and a 20 m length in the second.

In both cases the experiments began by optimizing the power output from the fibers. The fibers were then cut in the center and aligned using

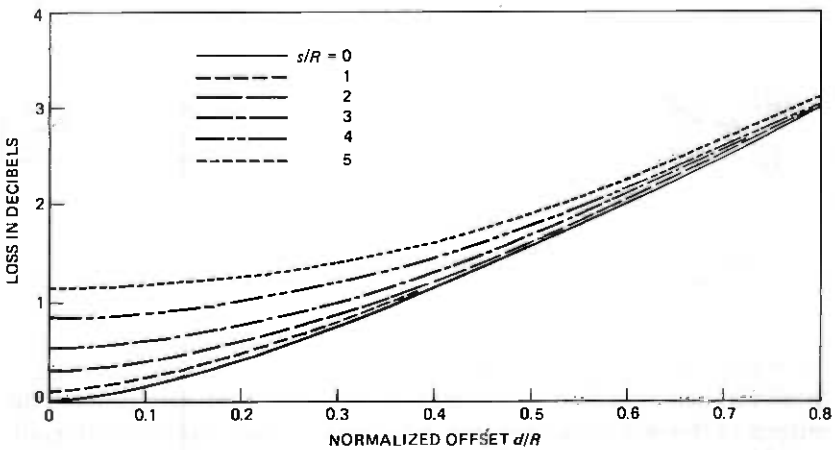


Fig. 4—Loss vs. normalized offset d/R at various normalized separations s/R from the first experiment.

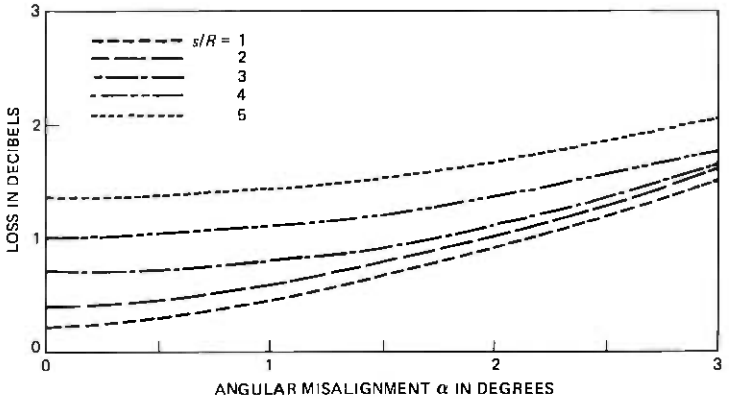


Fig. 5—Loss vs. angular misalignment α in degrees at various normalized separations s/R.

the micropositioners, and index matching fluid (glycerol) was applied to the joints. The power output in the first experiment was measured to be 0.01 dB less than the maximum power obtained before the fiber was cut. This figure was 0.07 dB in the second experiment.

The loss versus offset measurement (in both experiments) at zero separation was done by offsetting one fiber end (at the butt joint) to the

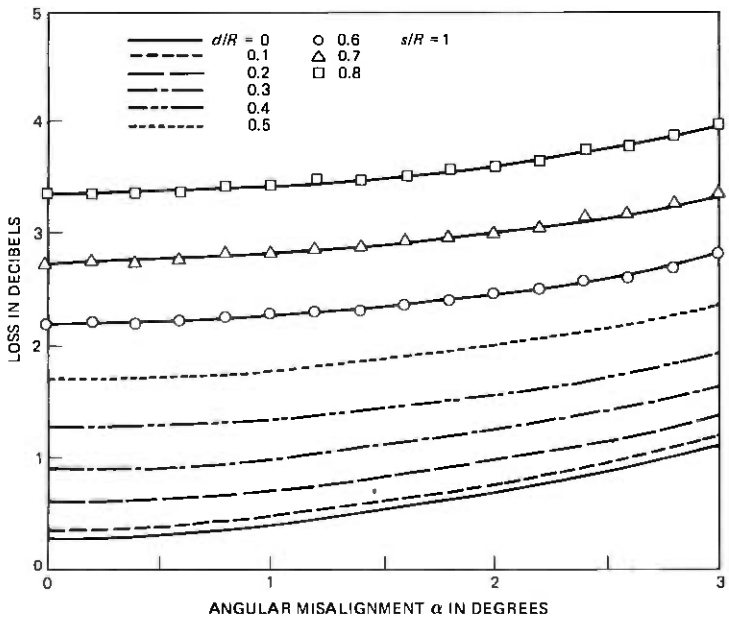


Fig. 6—Loss vs. angular misalignment α in degrees and various normalized offsets d/R at constant separation $s/R = 1$.

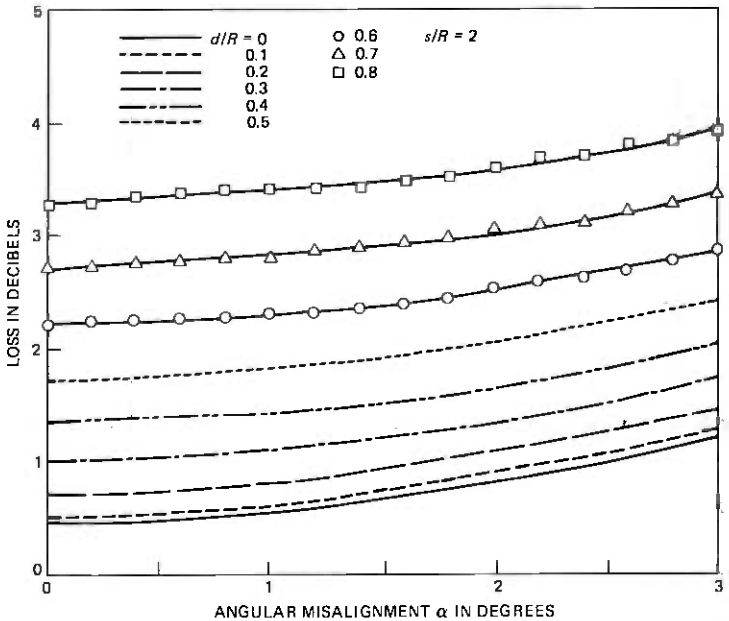


Fig. 7—Loss vs. angular misalignment α in degrees and various normalized offsets d/R at constant separation $s/R = 2$.

other by known amounts and the power output of the receiving fiber was recorded. This was repeated at normalized axial separations of 1, 2, 3, 4 and 5. The normalized separation and offset are defined as s/R and d/R, where s is the axial separation in μm, d is the offset in μm, and R is the fiber core radius in μm. The loss-versus-angular misalignment measurement (in the second experiment) began with aligning the receiving fiber with the center of rotation of the table so that the angular alignment could be changed while the axial separation and offset remained constant. The angular alignment was varied from $-3°$ to $+3°$ in increments of $0.2°$ at normalized axial separations of 1, 2, 3, 4 and 5.

III. RESULTS

The coupling efficiencies in percentage-versus-normalized offset at six normalized axial separation are shown in Fig. 2a and b (first and second experiment, respectively). The facts that the results of two experiments are very similar and the transverse offset is by far the more important parameter can be seen in Fig. 3, in which the loss-versus-normalized offset d/R at zero separation, the loss-versus-various normalized separations s/R at zero offset, and the loss-versus-normalized angular misalignment $\alpha°/\sin^{-1}\mathrm{NA}_0$ at constant separation $S/R = 1$ are plotted. Here $\mathrm{NA}_0 = \sqrt{n_1^2 - n_2^2}$ and n_1 and n_2 are the index of refraction

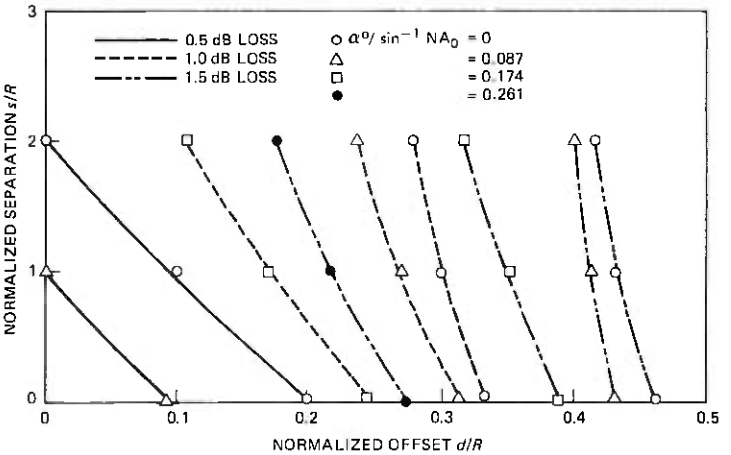


Fig. 8—Constant loss lines as the results of fiber end offset d/R, separation s/R, and angular misalignment $\alpha°/\sin^{-1} NA_0$.

of the fiber core center and cladding, respectively. The difference between the two experiments at zero offset and zero separation is due to the different amount of initial misalignment of the fiber ends after it was broken and butt-jointed. In the first experiment, the power output from the receiving fiber was 0.01 dB below the maximum power obtained before the fiber was broken; this figure was 0.07 dB in the second experiment. The designers of fiber connector or splice will be interested in the region where loss is low. Figure 4 shows the loss in dB versus small ofset ($d/R \leq 0.8$) at various separations. Figure 5 shows the loss due to angular misalignment at normalized separations of 1 through 5. Figures 6 and 7 show the loss due to angular misalignment and offsets at normalized separations of 1 and 2, respectively. Figure 8 shows constant loss curves as caused by various kinds of misalignment. As an example, consider various kinds of misalignment that all produce 0.5 dB loss: a normalized offset of 0.2 alone; a normalized separation of 2 alone; a normalized angular misalignment of 0.087 and normalized separation of 1; a normalized offset of 0.1 and normalized separation of 1. Designers of connectors will have to pay very close attention to offset, then angular misalignment and separation, respectively.

IV. CONCLUSION

Loss versus various kinds of misalignment of two ends of the same fiber has been measured in two independent experiments. The measurement setup, fiber diameter, and length were different in the two experiments, yet the measurement results are strikingly similar. Transverse offset is shown to be the most critical parameter in the design

of fiber connectors and splices. The present results provide only the minimum loss that would arise in actual fiber connectors and splices, since additional losses might be caused by other factors such as fiber diameter and index profile mismatch.

REFERENCES

1. C. M. Miller, "Transmission vs. Transverse Offset for Parabolic-Profile Fiber Splices with Unequal Core Diameters," B.S.T.J., *55*, No. 7 (September 1976), pp. 917–928.
2. D. Gloge, "Offset and Tilt Loss in Optical Fiber Splices," B.S.T.J., *55*, No. 7 (September 1976), pp. 905–916.
3. M. Young, "Geometrical Theory of Multimode Optical Fiber to Fiber Connectors," Optics Communications, *7*, No. 3 (March 1973), pp. 253–255.
4. J. S. Cook, W. L. Mammel, and R. J. Grow, "Effects of Misalignments on Coupling Efficiency of Single-Mode Optical Fiber Butt Joints," B.S.T.J., *52*, No. 8 (October 1973), pp. 1439–1448.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 3, March 1978
Printed in U.S.A.

Design and Experimental Optimization of a Canister Antenna for 18-GHz Operation

C. A. SILLER, JR., P. E. BUTZIEN, and
J. E. RICHARD

(Manuscript received May 27, 1977)

The design and experimental optimization of a canister antenna for operation in the 17.7 to 19.7 GHz frequency band are described. The canister is specifically designed to accommodate the antenna, transmitter and receiver units, and to be aesthetically innocuous in its environment. The basic antenna configuration, that of a shielded inverted periscope, is reviewed. Details of the constituent parts of the antenna, including dual mode feed, paraboloidal reflector, mirror, radome, and microwave absorber are presented. The influence that each of these items plays in determining net electrical performance is identified, including (where appropriate) steps taken to achieve electrical optimization. The antenna is shown to afford excellent sidelobe suppression, azimuthal plane cross polarization discrimination in the mid to upper 30 dB range, and a return loss of better than 23 dB. The gain efficiency is approximately 62 percent and is essentially polarization independent.

I. INTRODUCTION

The design and experimental optimization of a canister antenna for point-to-point operation in the 17.7 to 19.7 GHz frequency band are described in this paper. The antenna, which has been briefly described earlier,[1] was designed for a digital radio system (DR 18A).[2] The system will use a 274-Mb/s quaternary phase shift keying (QPSK) modulation to provide eight radio channels (seven working, one protection). Each channel provides 4032 voice circuits. Anticipated typical repeater spacings will be 2.4 to 7.2 kilometers.[3] The canister is specifically designed to accommodate the antenna, transmitter and receiver units, and to be aesthetically innocuous in its environment.

The initial antenna concept was first proposed by Crawford and Turrin[4] in 1969. Subsequently, the authors, drawing in part upon the-

Fig. 1—DR 18A repeater site located in Methuen, Massachusetts.

oretical and experimental work of colleagues at Bell Laboratories, designed an antenna for incorporation into a mast-supported, integrated canister antenna concept. A photograph of an existing repeater site in Methuen, Massachusetts, which may be characterized as "typical" in appearance, is presented in Fig. 1.

As depicted in the cut-away view in Fig. 2, the antenna is basically a shielded inverted periscope consisting of a paraboloidal reflector, feed assembly, mirror and inclined radome. The feed, the end of which is located at the focal point of the parabolic dish, illuminates the paraboloid with a spherical wave. Upon reflection, the energy is converted to a plane

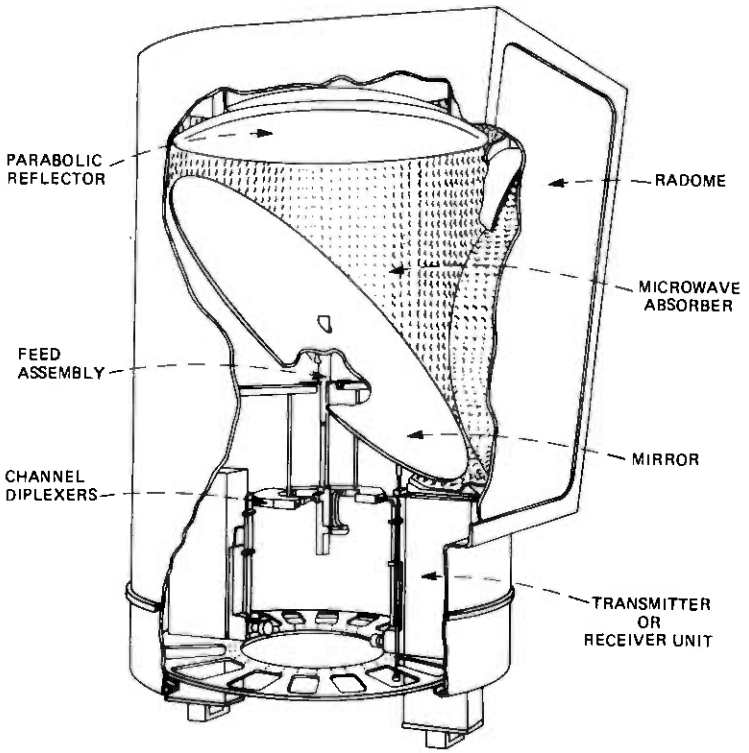


Fig. 2—Schematic depiction of canister antenna showing salient features.

wave which is subsequently incident upon a mirror nominally inclined at 45° to the paraboloidal axis. The energy is then reflected from the mirror (which is tiltable to deflect the beam up or down slightly) and exits through the circular aperture. Low wide-angle sidelobes are achieved by relying upon the absorber-lined canister to afford electromagnetic shielding and reduce edge diffraction at the aperture. It may also be noted from this figure that the canister serves as housing for the microwave networks as well as transmitter and receiver modules. This paper deals specifically with the influence of the various components of the antenna on its electrical performance and (where appropriate) steps taken to achieve electrical optimization.

Electrical objectives to be attained in the antenna design are dictated by the radio system. After a study into the regional variation of rainfall and its role in setting repeater spacings, the gain objective was established as approximately 42 dB above isotropic at the low edge of the frequency band since gain increases with frequency. This gain objective included transmission loss through the radome. The radiation pattern determines the maximum number of radio paths which may converge

at a point. With a goal of eight such paths and a desired discrimination of 60 dB, the pattern was to drop to ≤ -60 dB at 45° in azimuth. Finally, the return loss of the dish and window was to exceed 26 dB, i.e., voltage reflection ≤ 0.050.

Considerations involved in the selection of a suitable feed are reviewed. Since the influence of the feed pattern on gain cannot be divorced from the selection of a paraboloid, the inter-relationship of feed taper and dish f/D ratio is delineated by computing both illumination and spillover efficiencies to obtain net antenna efficiency. The optimum position of the feed with respect to the mirror is also explored, and the influence of the hole around the feed on antenna radiation patterns is elucidated. The wide-angle radiation suppression of the antenna is shown to be acutely dependent upon the use of broadband microwave absorber. Since antenna directivity is also affected by the presence of a radome, the tilt of the mirror, and the correct positioning of the feed, the role which these items play in influencing radiation suppression is assessed. Measured return loss and the sources which give rise to reflected power are also discussed. The paper is concluded with a summary of the electrical characteristics of the antenna design.

II. ANTENNA FEED

2.1 Properties of dual mode feeds

An objective in the design of this antenna was the attainment of circularly symmetric feed patterns for the illumination of the paraboloid. It may readily be shown that such a feed, with an illumination function characterized by

$$E_\theta = AF(\theta) \sin \phi \tag{1}$$

$$E_\phi = AF(\theta) \cos \phi \tag{2}$$

$$|E| = (E_\theta^2 + E_\phi^2)^{1/2} = AF(\theta) \tag{3}$$

where θ, ϕ are the conventional spherical coordinates, results in a completely linearly polarized field distribution after reflection from a full paraboloid. As a consequence, there are no cross polarized fields in the aperture. In addition, this feed illumination readily permits one to design for optimum polarization-independent gain.

Fields like those given in eqs. (1) and (2) are achieved by using dual mode and hybrid mode feeds. For this antenna, a dual mode feed was found to be effective, as well as simpler and less costly to produce than the hybrid mode feed. Chu[5] has noted that the observed radiation pattern of a dual mode feed is approximated closely by the H-plane pattern of an open-ended circular waveguide excited by the dominant TE_{11}° mode:

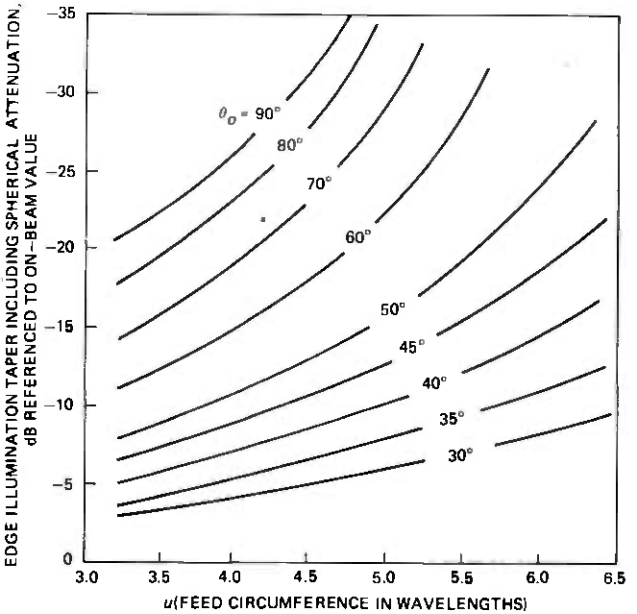


Fig. 3—Characterization of feed illumination, including spherical attenuation to a paraboloidal surface, for various values of the subtended dish half-angle, θ_o.

$$F(\theta) = [\sqrt{1-(1.841/u)^2} + \cos\theta] \cdot \frac{J_1(u \sin\theta)}{1 - \left[\frac{u \sin\theta}{1.841}\right]^2} \qquad (4)$$

where $u = \pi d/\lambda$, d is the diameter of the feed waveguide, λ is the wavelength of the applied signal, and θ is the angle measured from the axis of the feed. The region of validity for eq. (4) is $\pi \leq u \leq 2\pi$, or equivalently, the waveguide diameter is limited to one to two wavelengths.

The field intensity at the edge of a paraboloidal reflector may be computed as a function of u by using eqs. (3) and (4) and including spherical attenuation to the paraboloidal surface, $\cos^2(\theta/2)$. The latter comes from the fact that the equation of a paraboloid in spherical coordinates is $r = 2f/(1 + \cos\theta)$. The results of such a computation are presented in Fig. 3 for various values of θ_o, the subtended half-angle of the paraboloid. The use of θ_o instead of f/D is preferred by the authors because of its simpler physical interpretation. It is noted that $f/D = 1/(4\tan\theta_o/2)$, where f and D are the focal length and reflector diameter, respectively. From Fig. 3, selection of any two of the following three variables uniquely specifies the third: edge taper in decibels, feed diameter in wavelengths, and subtended half-angle in degrees. The utility of this figure arises in optimizing the gain efficiency of the antenna, the subject of the next section.

2.2 Relationship of feed taper to efficiency

The gain efficiency of a center-fed full paraboloid may be broken down into an illumination efficiency and spillover efficiency (as in Ref. 6). The concept of an illumination efficiency is well known; spillover efficiency specifies that fraction of the total power radiated by the feed which is intercepted by the paraboloidal dish. Denoting these terms by η_i and η_s, respectively, they are mathematically evaluated from the feed illumination function. Namely,

$$\eta_i = \frac{\left| \int_{\phi=0}^{2\pi} \int_{\theta=0}^{\theta_o} F(\theta) \cos^2\left(\frac{\theta}{2}\right) d\Sigma \right|^2}{A_p \int_{\theta=0}^{2\pi} \int_{\theta=0}^{\theta_o} \left| F(\theta) \cos^2\left(\frac{\theta}{2}\right) \right|^2 d\Sigma} \tag{5}$$

In eq. (5), the projected aperture area is

$$A_p = 4\pi f^2 \frac{1 - \cos\theta_o}{1 + \cos\theta_o}$$

and the integration is performed over the projected aperture of the reflector with differential area element

$$d\Sigma = \frac{f^2 \sin\theta}{\left(\cos\frac{\theta}{2}\right)^4} d\theta d\phi$$

The spillover efficiency is given by

$$\eta_s = \frac{\int_0^{\theta_o} [F(\theta)]^2 \sin\theta d\theta}{\int_0^{\pi/2} [F(\theta)]^2 \sin\theta d\theta} \tag{6}$$

The total antenna gain efficiency is expressed by the product, viz. $\eta_a = \eta_i \eta_s$. The evaluation of η_s given by eq. (6) is approximate insofar as it neglects back radiation from the feed. Radiation into the hemisphere $\pi/2 \leq \theta \leq \pi, 0 \leq \phi \leq 2\pi$ is not accurately described by eq. (4), and is so low as to contribute virtually nothing to the evaluation of η_s.

The terms η_a and η_s have been computed as a function of dish edge illumination (spherical attenuation *and* feed taper at an angle θ_o) for a variety of subtended half-angles. (Selection of a dish edge illumination and half-angle implicitly determines a feed diameter as shown in Fig. 3). Typical aperture efficiencies, expressed in decibels down from full area gain, are presented in Fig. 4. The figure shows that for a specified angle θ_o, the optimum efficiency occurs for edge tapers of -11.5 to -12.5

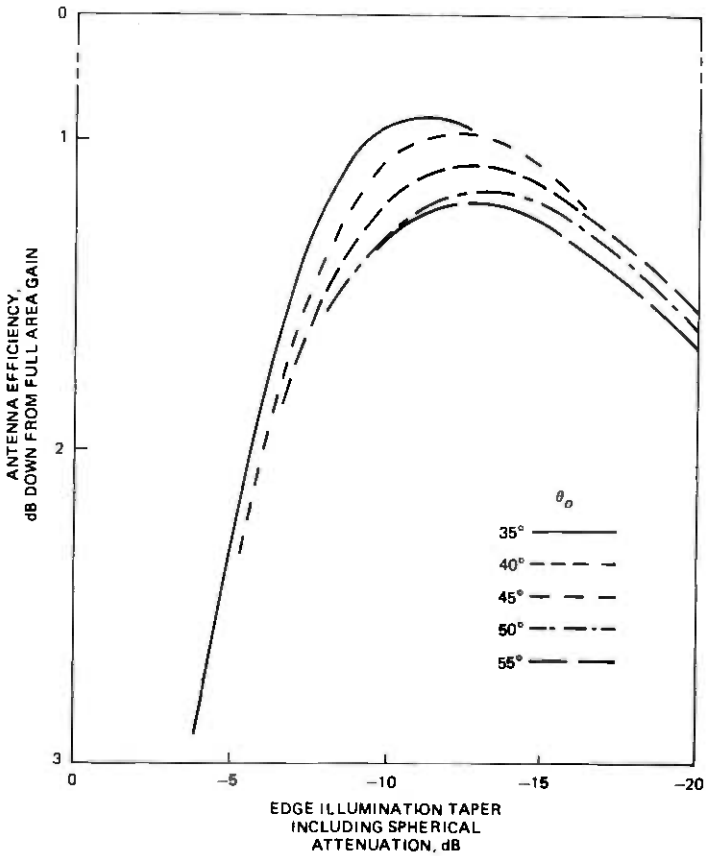


Fig. 4—Antenna efficiency as a function of edge illumination for various values of subtended dish half-angle, θ_o.

dB. Additionally, for optimum taper, the least loss in gain occurs for a half-angle of approximately 35°.

While it would appear from the above analysis that a half-angle of 35° should be selected, two additional considerations are germane. First, from an electrical standpoint, edge tapers of −11.5 to −12.5 dB for small values of θ_o necessitate relatively large feed diameters (See Fig. 3). Large feed diameters are deleterious because they can support higher order modes, the effect of which will be demonstrated subsequently. They also increase feed blockage in the antenna, thus deteriorating sidelobe suppression in the radiation pattern. In regard to this latter point, theoretical analyses of this antenna by Anderson[7] indicate that the sidelobe levels in that angular region where the imaged paraboloid would be visible to an observer are insensitive to field illumination at the edge of the dish, an observation that has been experimentally substantiated elsewhere.

within Bell Laboratories. He concluded that the level of radiation suppression is principally influenced by feed blockage. This conclusion is validated by an experimental study discussed later in the paper. The second point arguing against a 35° half-angle is aesthetic, in that such a paraboloid would require a canister more slender and tall than considered appropriate. Feed blockage is estimated to have a negligible effect on the gain of this antenna.

There is also a reason why angles greater than 45° are not especially desirable. The geometry of the antenna dictates that for angles greater than 45°, the feed must protrude above the mirror. Yet experimental evidence indicates that this, too, manifests itself as feed blockage, and can deteriorate sidelobe suppression. For these reasons, a subtended half-angle of 40° appeared reasonable.

As stated in the Introduction, the gain objective of this antenna was approximately 42 dBi at 17.7 GHz. Experience indicates that it is difficult to realize gain within 0.5 dB of theoretical, so the antenna was designed to theoretically provide a gain of 42.5 dBi. Using Fig. 4 with $\theta_o = 40°$ implies that this gain is achieved by selecting a reflector diameter of 0.813 meters and $f = 0.559$ meters. Because of limitations in available tools for spinning the dish, a focal length of 0.536 meters was actually used. For a dish of 0.813-meter diameter, this corresponds to $\theta_o = 41.4°$ or $f/D = 0.66$. Tolerances of the spun aluminum paraboloid were 0.0254 centimeters rms and a peak deviation of less than 0.0305 centimeters from the design surface. At 19.7 GHz, the highest anticipated operating frequency, these correspond to $\lambda/60$ and $\lambda/50$, respectively.

2.3 Feed designs and performance

Two dual mode configurations (as depicted in Fig. 5) of square and circular cross section with a variety of aperture sizes were considered. The principle of operation of these dual mode configurations is that of achieving a two-mode mixture of fields at the feed aperture such that the edge currents are nearly zero.[8] In Fig. 5, we may consider a TE_{11}° mode (for the case of circular cross section) propagating upward from the uniform waveguide at plane a. This wave encounters an abrupt change in cross section at plane b such that the electric field is bent to maintain a vanishing tangential component at the conducting walls. An axial component of the total electric field (E_z) is thereby generated and by plane c conversion of some of the energy to the TM_{11}° mode has been accomplished. The distance from c to d provides the proper phasing of the two modes at the feed aperture since the two modes propagate at different phase velocities. Since the two modes do travel with different phase velocities, the feed does have a bandwidth limitation. Nevertheless, the feed provides proper performance across the 17.7 to 19.7 GHz band.

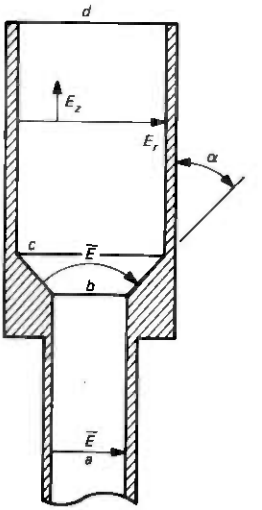


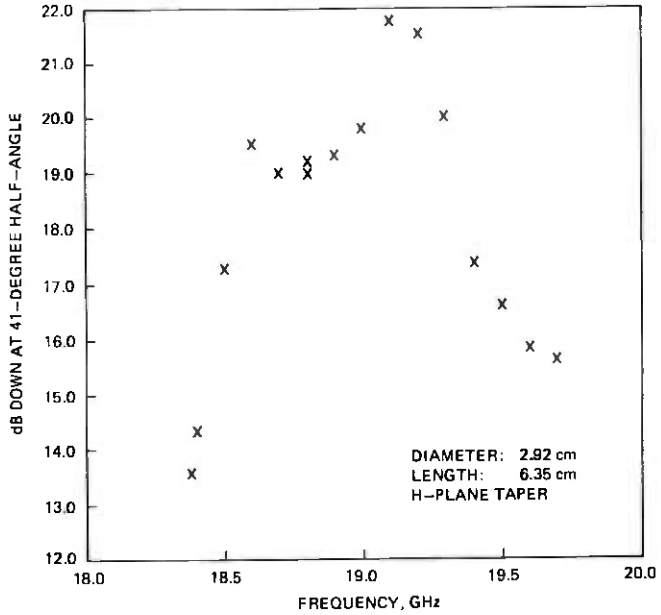
Fig. 5—Dual mode feed cross section.

Two sizes of dual mode feed were built and tested in both the round and square cross section geometries. The feeds of square cross section were found to have higher cross polarized fields in the ±45° planes than the circular feeds, and were therefore no longer considered.

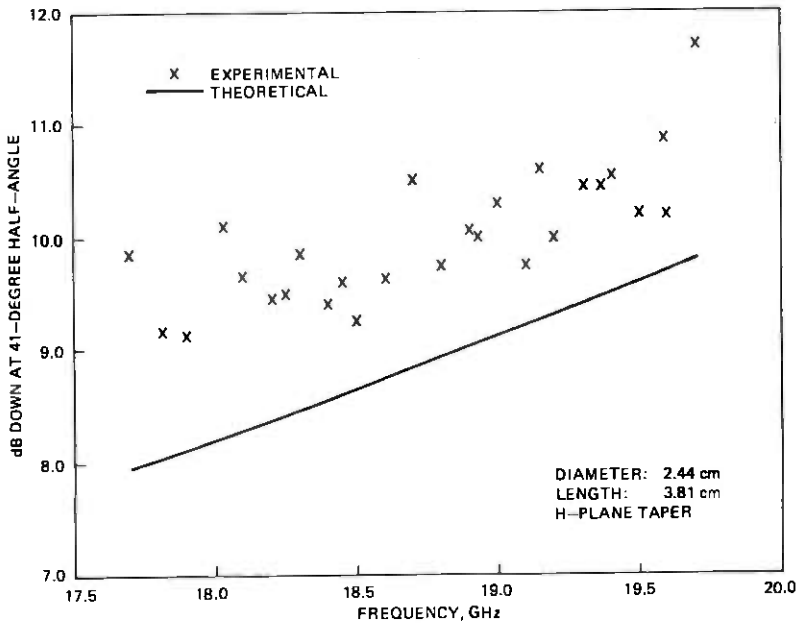
Two sizes of feed were built because the aperture diameter-to-wavelength ratio determines the illumination taper of the feed. As noted above, the canister antenna was designed with a subtended half-angle of approximately 41.4° between the feed and rim of the paraboloid. To attain optimum gain efficiency for this 0.66-f/D paraboloid fed by a dual mode feed requires an edge illumination (including spherical attenuation to the paraboloidal surface of just over 1.1 dB at 41.4°) of −12.1 dB (from Fig. 4). Using Fig. 3, such a feed requires u = 5.4 and hence has a diameter of 2.92 centimeters at 17.7 GHz. This diameter, however, allows four higher-order modes to propagate between planes c and d (Fig. 5). Therefore, a smaller version with a 2.44-centimeter diameter which just cuts off beyond the TM_{11}° mode was also built. Both of these models had cone angles of $\alpha = 30°$.

It was found that due to higher order modes (TM_{21}°, TE_{41}°, TE_{12}°, and TM_{02}°) which could propagate, the larger diameter feed produced H-plane radiation patterns which fell off much too rapidly over a large part of the frequency range of interest. This effect is depicted in Fig. 6a.

The 2.44-centimeter diameter dual mode feed was extensively tested and was subsequently chosen as the feed for this antenna. The length of the drift section was adjusted empirically to 3.81 centimeters to obtain the best match of E-plane and H-plane taper at mid-band. The feed il-

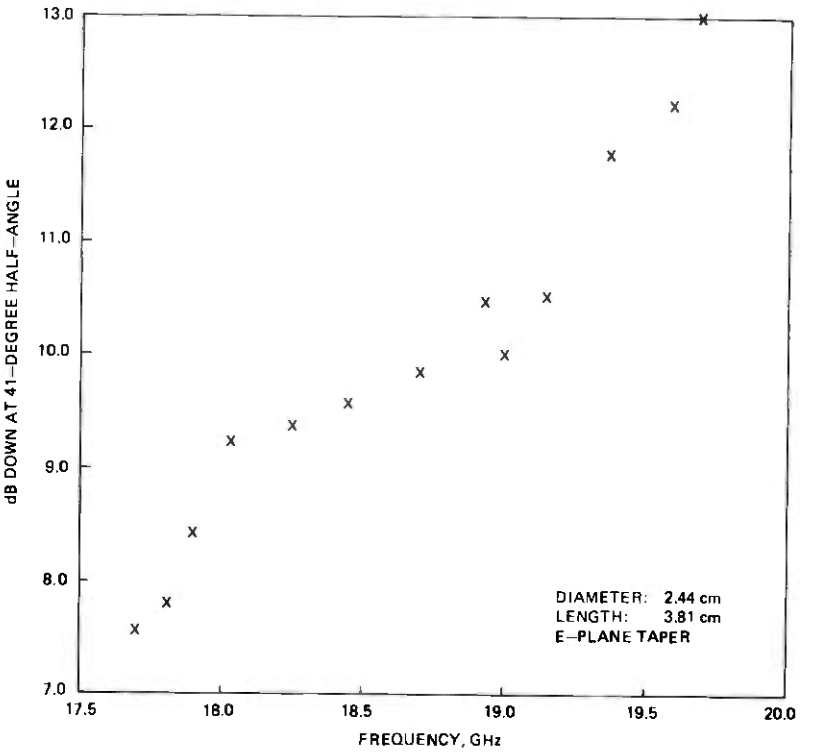


(a)



(b)

Fig. 6—Dual mode feed tapers as a function of frequency: (a) H-plane measured edge tapers of a 2.92-centimeter diameter dual mode feed; (b) H-plane measured and theoretical edge tapers of a 2.44-centimeter diameter dual mode feed; (c) E-plane measured edge tapers of the 2.44-centimeter diameter dual mode feed.

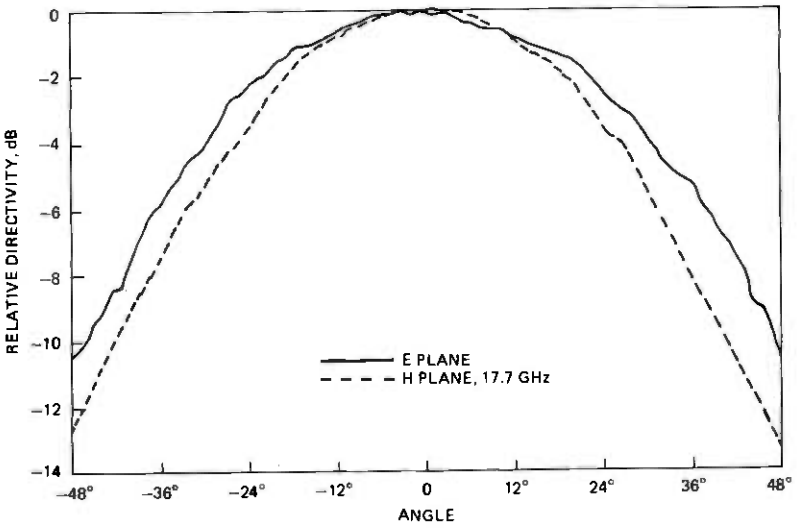


(c)

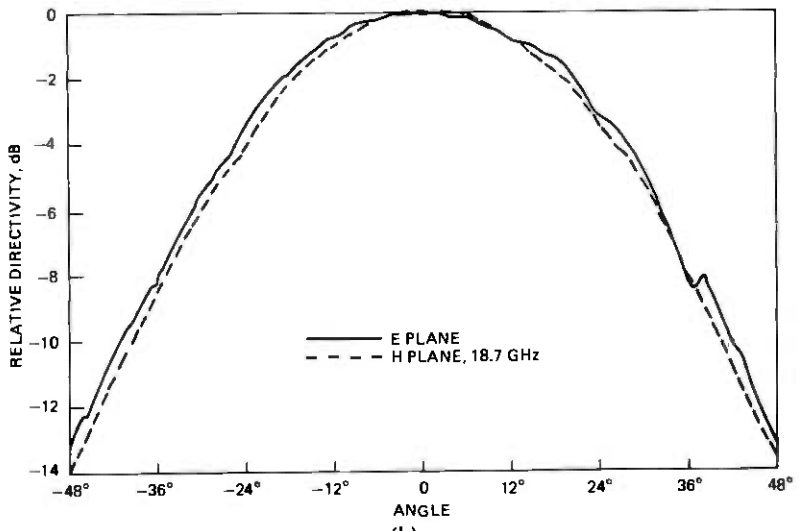
Fig. 6 (*continued*)

lumination tapers (no spherical attenuation) as a function of frequency for the H-plane and E-plane are shown in Figs. 6b and 6c, respectively.* Each point represents the power average of the edge taper at 41° on both sides of the feed radiation pattern like those shown for the E- and H-planes at 17.7 GHz, 18.7 GHz, and 19.7 GHz in Figs. 7a, 7b, and 7c, respectively. The patterns in Figs. 7a, 7b and 7c were measured on a small indoor range. Note from Figures 6b and 6c that E-plane and H-plane tapers are the same only near the design frequency, implying best illumination symmetry at that point. As the signal frequency departs from the design frequency, the illumination in the aperture of the paraboloid becomes increasingly asymmetric, an assessment of which may be obtained from these figures. For example, note from Figures 6b, 6c, and 7a that at 17.7 GHz, the H-plane taper at the edge of the dish is measured to be approximately −9.8 dB, while at this same frequency the E-plane

* As Figs. 6b and 6c show, the final antenna feed was tested at more frequencies in the H-plane than in the E-plane. This was done to assure that difficulties which were manifest in the over-moded feed (see Fig. 6a) did not occur in the smaller feed.



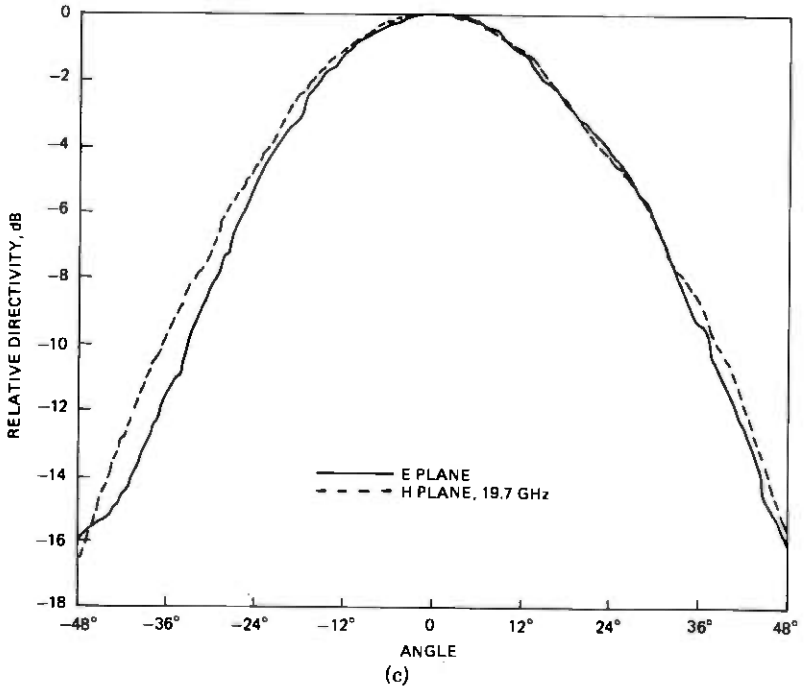
(a)



(b)

Fig. 7—E- and H-plane radiation patterns of the 2.44-centimeter diameter dual mode feed: (a) 17.7 GHz; (b) 18.7 GHz; (c) 19.7 GHz.

taper is approximately −7.6 dB. Figure 6b also contains theoretical H-plane tapers computed from eq. (4). As the data reveals, at 41° measured tapers exceed the theoretical values by approximately 1 dB. This observation is also valid for other dual mode feeds with which the authors are familiar. Note that the measured taper at 18.7 GHz is very close to the optimum (see Fig. 4).



(c)

Fig. 7 (*continued*)

Cross polarized patterns were also measured in both the E and H principal planes and the ±45° planes. The patterns were generally room-reflection limited to better than a 40-dB dynamic range, thereby inferring the capability of measuring cross polarized fields to that level. Measurements made at more than ten frequencies within the 17.7 to 19.7-GHz band indicate that near boresight the cross polarized response approaches the upper 30 dB range in the ±45° planes. Because the antenna is circularly symmetric from a geometrical optics viewpoint, the reflector system itself adds little to cross polarization conversion except for small contributions from surface roughness and feed scattering. The antenna polarization properties are essentially set by the feed itself.

The theoretical gain efficiency of the antenna has been calculated for 41.4° and is plotted in Fig. 8 as a function of field taper at the rim of the paraboloid. Also indicated on this curve is the taper achieved with the 2.44-centimeter diameter feed at 17.7 GHz, 18.7 GHz, and 19.7 GHz (obtained from a robust linear regression of data in Figs. 6b and 6c). As this figure indicates, the feed should afford near optimum performance since the highest efficiency is predicted for the upper portion of the band where rain attenuation can be expected to be most severe.

The gain of this antenna without the radome was measured on an

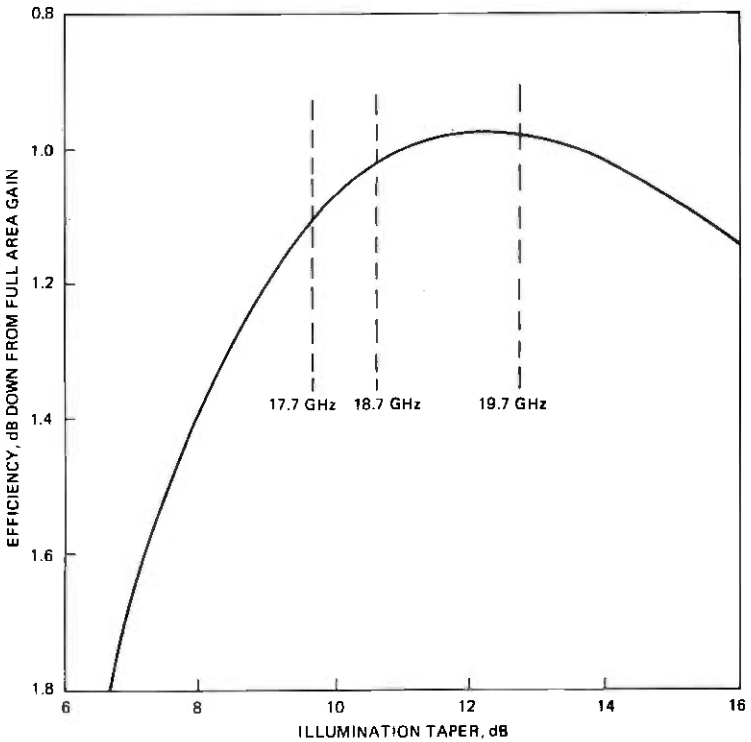


Fig. 8—Theoretical efficiency of prototype design as a function of illumination taper at edge of dish.

outdoor range using the comparison method and a standard gain horn as a reference calibration. The measured gain, relative to an isotropic radiator, and aperture efficiency are stated in Table I. Observe that higher gain for the upper portion of the frequency band has been achieved.

Experience obtained during an earlier CW experiment using this antenna concept has indicated the necessity to avoid insects and other foreign matter getting into the feed and altering electrical performance. Precipitation, on the other hand, is not expected to enter the canister since the antenna body and window are effective in this regard. In its

Table I—Measured gain (over isotropic) for 18-GHz antenna without radome

Frequency, (GHz)	Gain, (dBi)	Aperture efficiency, %
17.7	41.5	62.2
18.7	42.1	64.0
19.7	42.4	61.8

customary orientation, the feed is pointing up (see Fig. 2). A variety of feed plug configurations designed to inhibit foreign matter from falling into the feed or accumulating at the opening were fabricated and tested. Each of the plugs considered was formed of commercially-available, expanded, closed-cell, polystyrene with a relative dielectric constant $\epsilon_r \leq 1.03$ and loss tangent 0.002. Each plug was tapered on the outside and inside to inhibit the accumulation of foreign matter and minimize reflected power, respectively. The plugs were cemented in place with an electrically transparent silicone rubber adhesive and then tested for their influence on feed patterns, return loss, and antenna gain across the band. The plug configuration shown in Fig. 9 influenced the feed pattern least, was invisible in return loss and gain measurements, was simplest in design, and was therefore selected.

III. RELATIONSHIP OF FEED TO MIRROR

The topics considered in this and later sections deal primarily with effects which influence the antenna radiation characteristics. It is therefore appropriate to comment briefly on the antenna range facilities for 18-GHz measurements. Pattern and gain measurements are made on a ground reflection range. The source antenna is located close to the ground so that the field illumination across the aperture under test is uniform within one decibel (field uniformity is frequently verified by probing the field in front of the antenna). Use of a pulse transmitter (with a pulse duration of 200 nsec) and gated receiver (which samples the peak amplitude during a 50 nsec interval on the leading edge of the transmitted pulse) assures a measured response free of spurious reflections. The dynamic measurement range at 18 GHz is 70 dB.

3.1 Axial movement of feed relative to mirror

The canister antenna was designed with the flexibility of allowing feed-paraboloid translation with respect to the mirror. As shown in Figs. 10a and 10b, the feed and paraboloid may be moved in unison (providing, of course, that the feed is always kept at the focal point of the dish) with the electrical aperture of the antenna remaining fixed. Salient antenna dimensions are shown in Fig. 10c.

Moving the feed with respect to the mirror has a pronounced effect upon the antenna radiation pattern. That angular portion of the radiation pattern for which the imaged paraboloid is visible, is influenced more by feed blockage than dish edge illumination (hence the effort to set feed taper for high efficiency). Minimum feed blockage is achieved by selecting a feed with small diameter, and keeping the feed from protruding well above the mirror where it would act as a scattering obstacle. On the other hand, a feed close to the mirror offers the potential for two dele-

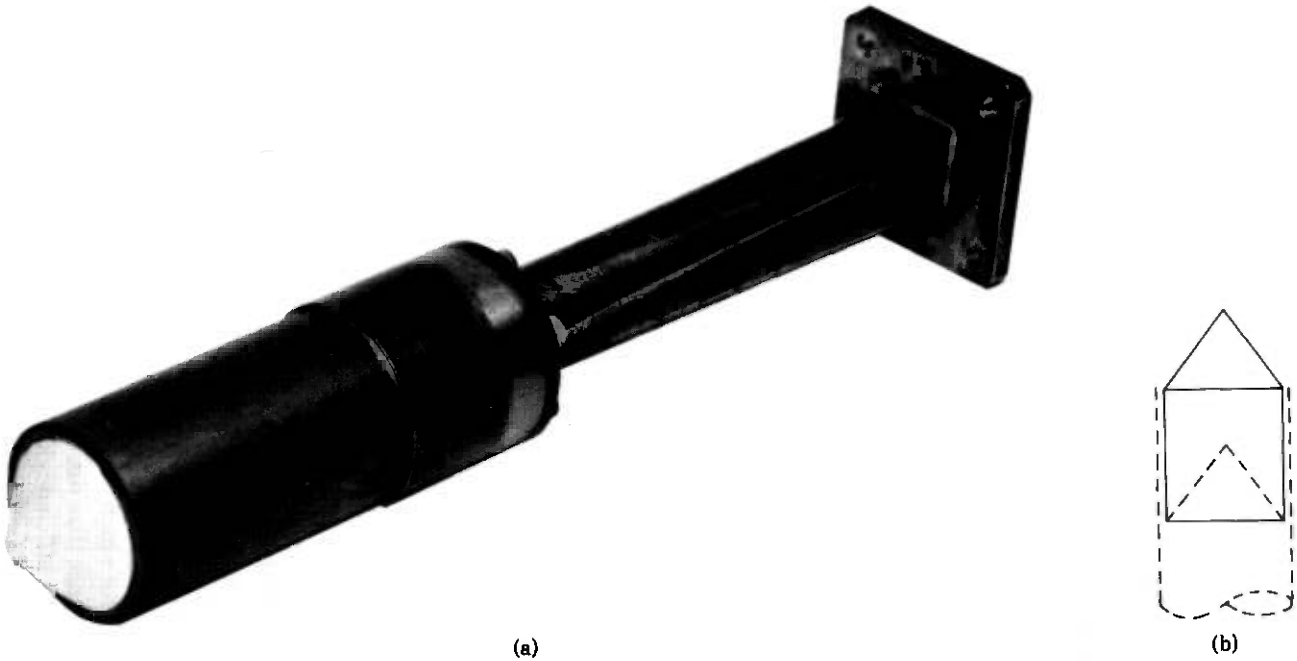


Fig. 9—Dual mode feed: (a) photograph showing insect seal; (b) diagrammatic depiction of seal.

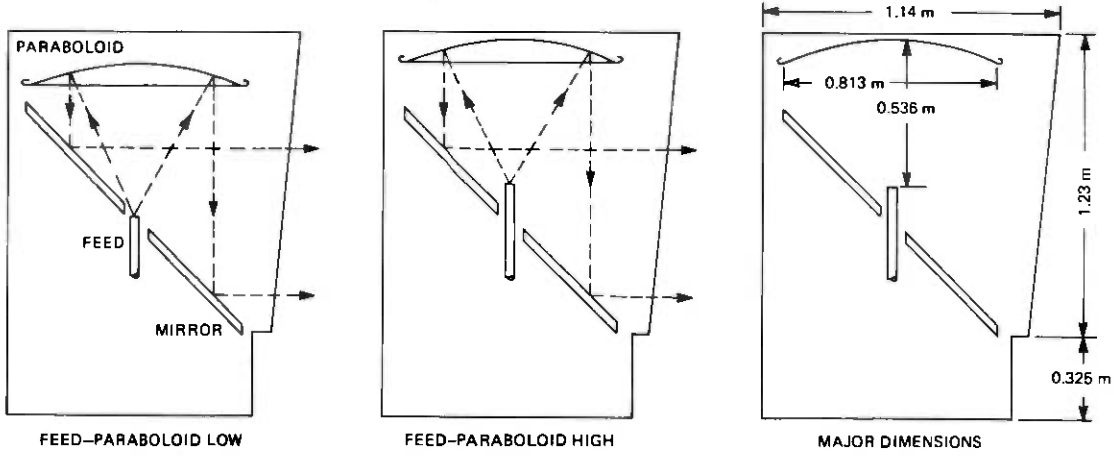


Fig. 10—Schematic depiction of antenna concept: (a) feed-paraboloid in lower-most position; (b) feed-paraboloid elevated; (c) major dimensions.

terious effects. First, a feed too low will allow grazing illumination of the mirror near the hole around the feed, and concomitant pattern degradation. Second, experiments reveal that if the feed directly illuminates the mirror with excessive energy, then the illumination on the dish will be vertically asymmetric in amplitude.[9] This latter effect would also cause pattern degradation, possibly accompanied by deteriorated cross polarization discrimination.

Optimum feed location was selected by an experimental study of the influence of feed location on antenna radiation patterns. Measurements were made at three frequencies (corresponding to the center and extremities of the 17.7 to 19.7 GHz band) for all polarization states* and differing feed locations. The results, a typical example of which is depicted in Fig. 11 as a smoothed radiation pattern,† clearly suggest that a low-profile feed affords the best performance. Therefore, the feed is placed low enough so as not to mask the horizontally directed plane wave, and yet not so low as to allow its spherical wavefront to illuminate the mirror too strongly. At its centerline, the feed axially protrudes above the top surface of the mirror approximately 2.08 centimeters.

3.2 Influence of hole surrounding feed

As shown in Fig. 2, the dual mode feed protrudes up through a hole in the mirror. The mirror itself is constructed of commercially available aluminum plate 0.953 ± 0.013 centimeter thick. The surface is flat within 0.038 centimeter peak over a 1.22-meter span. An aluminum frame is epoxied to the underside to inhibit the plate from sagging under its own weight. The hole in this mirror manifests itself as "feed blockage" and influences the antenna radiation pattern.

A typical set of radiation patterns which exhibit the influence of this hole on radiation suppression is shown in Fig. 12. The three patterns correspond to a mirror with 4.06- and 5.08-centimeter projected diameter holes, and the hole around the feed carefully closed with conducting tape. As these patterns indicate, the presence of the open annulus caused raised sidelobes in the vicinity of 24° to 36°. Data acquired from extensive experimental measurements generally suggests the benefit of making the annulus as small as possible. It is appropriate to add that the annulus cannot be made arbitrarily small since provision must be made for allowing tilt in the mirror. Therefore the hole is designed to have an elliptically shaped projection with a minor diameter no larger than

* Horizontal polarization transmitted—Horizontal polarization received (HH), Horizontal polarization transmitted–Vertical polarization received (HV), Vertical polarization transmitted—Vertical polarization received (VV), and Vertical polarization transmitted—Horizontal polarization received (VH).

† Smoothed radiation patterns are prepared by the commonly accepted practice of drawing a smooth line across the peaks in the detailed pattern, thereby forming an envelope of peaks.

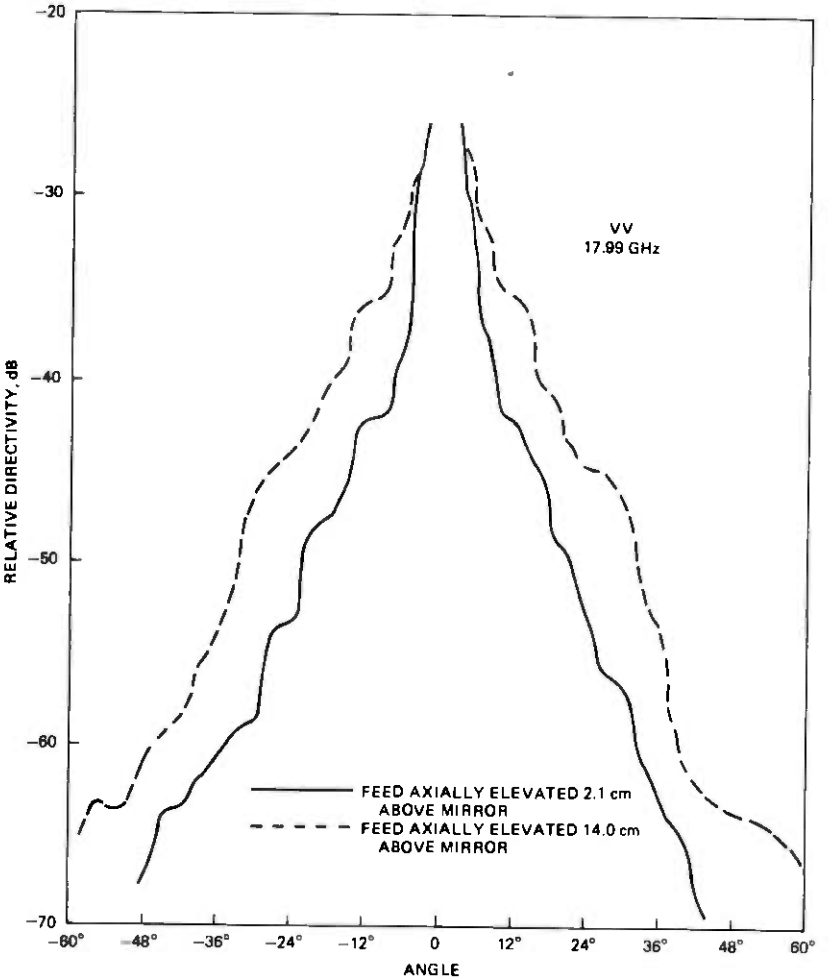


Fig. 11—Influence of feed position relative to mirror on azimuthal radiation pattern envelopes.

necessary for feed placement and orientation, and a major diameter large enough to allow reasonable tilting of the mirror.

IV. MICROWAVE ABSORBER AND ANTENNA RADOME

4.1 Absorber

The inside of the canister antenna is fully lined with microwave absorber. Experimental measurements indicate that absorber is necessary for the suppression of wide-angle sidelobe radiation.

In optimizing antenna performance, a variety of absorbers were con-

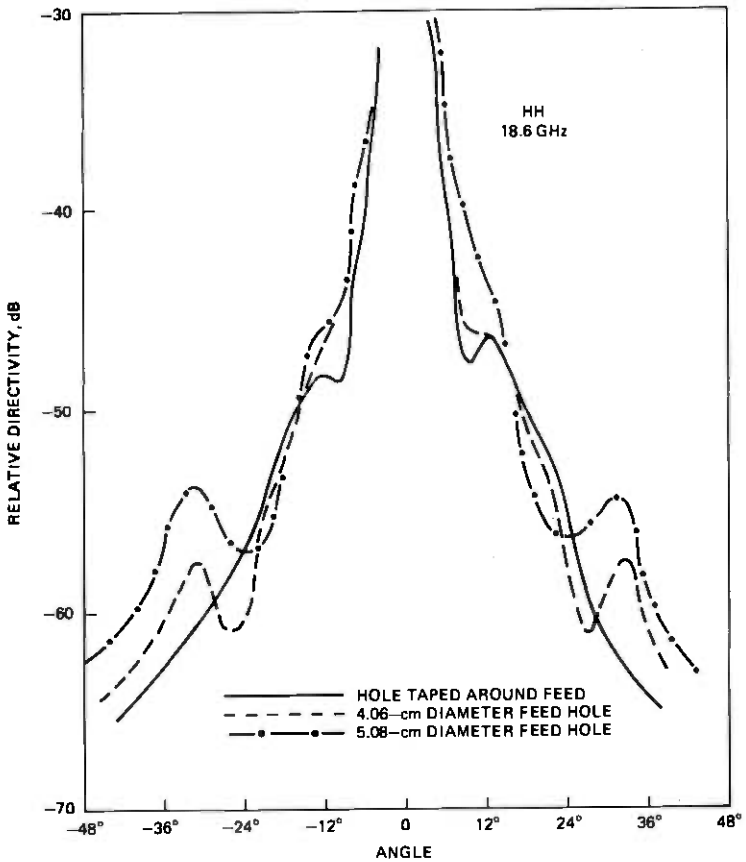


Fig. 12—Influence of feed-mirror annulus on azimuthal radiation pattern envelopes.

sidered, including 1.91-centimeter thick convoluted foam, and 5.08-centimeter thick "hair" absorber. Not only must a suitable absorber provide radiation suppression, but it must be resistant to environmental deterioration. The antenna considered in this paper will provide no absorber protection aside from inhibiting direct exposure to the outdoor elements (sunlight, rain, ice, etc.). Consequently, absorbers for this application should be heat resistant, not inclined to rapid organic decomposition, and impervious to water. Relative to this last point, while the antenna interior will not be exposed directly to rain, it is expected that condensate can form on the inside as the antenna "breathes".

It is useful to list here the respective advantages and disadvantages of the absorbers considered. Since much of the absorber inside the antenna will be subject to grazing incidence of electromagnetic fields rather than normal incidence, the convoluted surface would appear to be de-

sirable since experimentation has confirmed its excellent absorption qualities for varying angles of incidence. In contrast, "hair" has a relatively planar surface and might be expected to yield poorer performance for grazing angles. The convoluted absorber is made of an open-cell foam material while the "hair" is an open mat. Normally the "hair" absorber would have no tendency to hold water, while the convoluted foam could pick up water in a manner analogous to a sponge. To remedy this effect, the foam was covered with a sprayed Hypalon* coating approximately 0.076 millimeters thick. Hypalon in itself is an excellent coating because of its electromagnetic transparency and proven resistance to weathering. However, the coating was easily perforated by fingers during installation. Finally, "hair" was significantly less expensive than the convoluted material, or even flat foam absorbers of comparable thickness.

As a first step in assessing absorber performance, radiation patterns were made with no absorber, partial absorber, and a complete absorber lining of the antenna interior. The resulting radiation patterns using "hair" absorber are depicted in Fig. 13a. As this figure indicates, the addition of absorber dramatically reduces the wide-angle radiation of the antenna. The next step in the evaluation was to run separate patterns with the convoluted and hair absorber at three frequencies within the band for a variety of linear polarizations. The results, an example of which is shown in Fig. 13b, suggest that convoluted absorber generally affords slightly better performance. However, the differences are minor and because of the cost and weatherability advantages of the "hair," it is preferred in this application.

4.2 Radome

Radomes have been traditionally used on high performance line-of-sight antennas used in the 4-, 6-, and 11-GHz common carrier bands. These radomes are called "thin" because the 0.762- to 1.016-millimeter fiber glass membranes are only $\lambda/30$ thick at the highest frequency. At frequencies near 20 GHz, preliminary tests on the canister antenna indicate that from a pattern standpoint, a satisfactory thin radome would have to be less than approximately 0.305 millimeters in thickness. Such a thin radome would prove structurally inadequate. For that reason, attention was focused on half-wavelength radomes.

A radome used in the tests to be described was a solid laminate constructed of epoxy resin and a low-loss fiber glass called E-glass. The dielectric constant of the laminate was thought to be 4.0 so the radome was made 0.399 centimeters thick for electrical tuning at 18.7 GHz. Subsequent measurements indicated that the radome was actually tuned at approximately 18.3 GHz, inferring a dielectric constant of 4.3 (this latter

* Registered trademark of E. I. duPont de Nemours & Co., Inc.

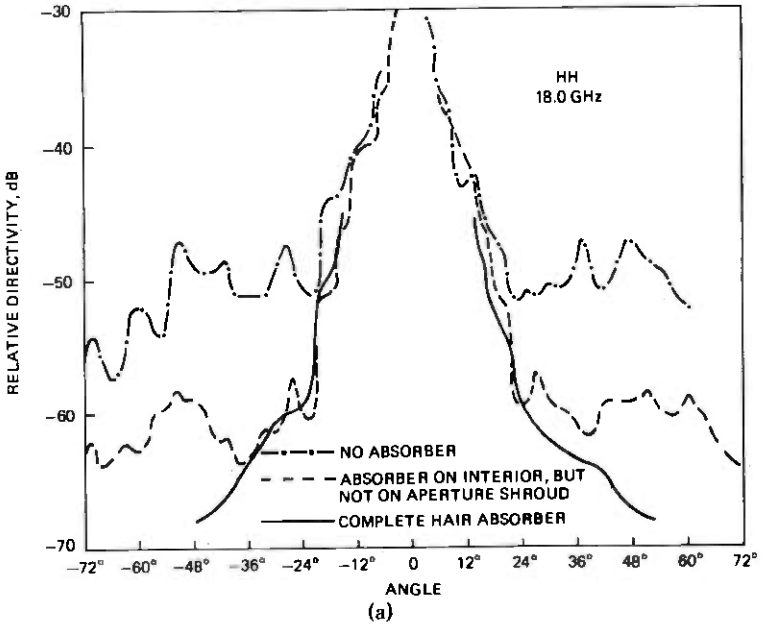


Fig. 13—Influence of microwave absorber on azimuthal radiation pattern envelopes: (a) presence or lack of absorber; (b) type of absorber.

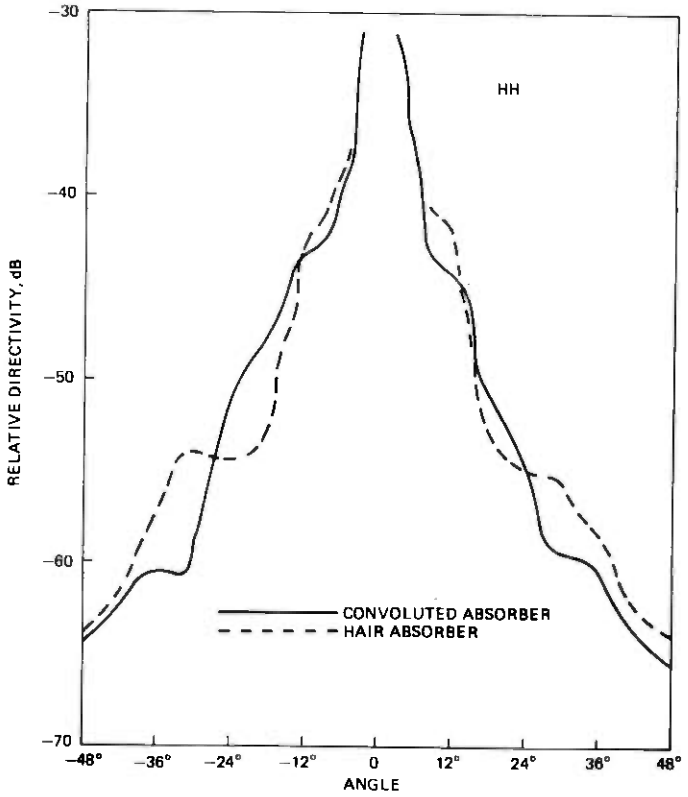
value is in accord with the commonly accepted value for this composite material). The loss tangent is 0.016.

Figures 14a and b depict representative results obtained with and without the radome on the antenna. Figure 14a indicates that measurements made near the tuned point show the radome to have little influence on the antenna radiation patterns. This is also true of the cross polarized response. Figure 14b presents measured results near the edge of the band. For this case the radome does perturb the principal polarization wide-angle radiation characteristics, though the sidelobe suppression still meets the design objectives. The common explanation of this effect is that energy is reflected back into the antenna and subsequently reradiated. Indeed, measurements made on this radome indicate that specularly reflected power is 35 dB down at 18.3 GHz, but only 13 dB to 14 dB down at 19.7 GHz.

V. MIRROR TILT AND FEED POSITION SENSITIVITY

5.1 Mirror tilt

Pairs of antennas used for point-to-point transmission are carefully oriented in elevation and azimuth to electrically point exactly at each other. For the canister antenna with relatively narrow beam width (3



(b)

Fig. 13—(*continued*)

dB beam width of 1.33°), this orientation is performed by peaking the received signal while adjusting the azimuthal and elevation angle. The azimuthal orientation is accomplished by rotating the entire antenna in its supporting cross-arm (see Fig. 1). The elevation angle is adjusted by rotating the mirror about the horizontal axis so as to tilt the beam up or down. It is readily shown that as the mirror is tilted through the angle θ_m, the beam moves $\theta_b = 2\theta_m$. The overall influence of reflector tilt on gain and pattern is covered in this section.

Reference radiation patterns were first measured with the mirror in the "nominal" 45° position (beam horizontal). Patterns were then measured with the mirror in a "tilt up" and "tilt down" position of approximately $\theta_m = \pm 1.5°$. In both cases the entire antenna was tilted 3.0° in the opposite direction to compensate for the beam tilt. This is done to maintain bore-sight beam orientation between the source and antenna. These measurements then establish the effect of the internal interaction of the plane reflector with other antenna parts. It should also

be noted that by tilting the antenna canister to compensate for reflector tilt, the patterns so obtained do not quite lie in the azimuthal plane, though the difference is slight.

As expected, it was observed that tilting the mirror causes a decrease in received signal level, but this signal is completely restored by compensatory antenna tilt. Therefore, gain changes do not occur with modest

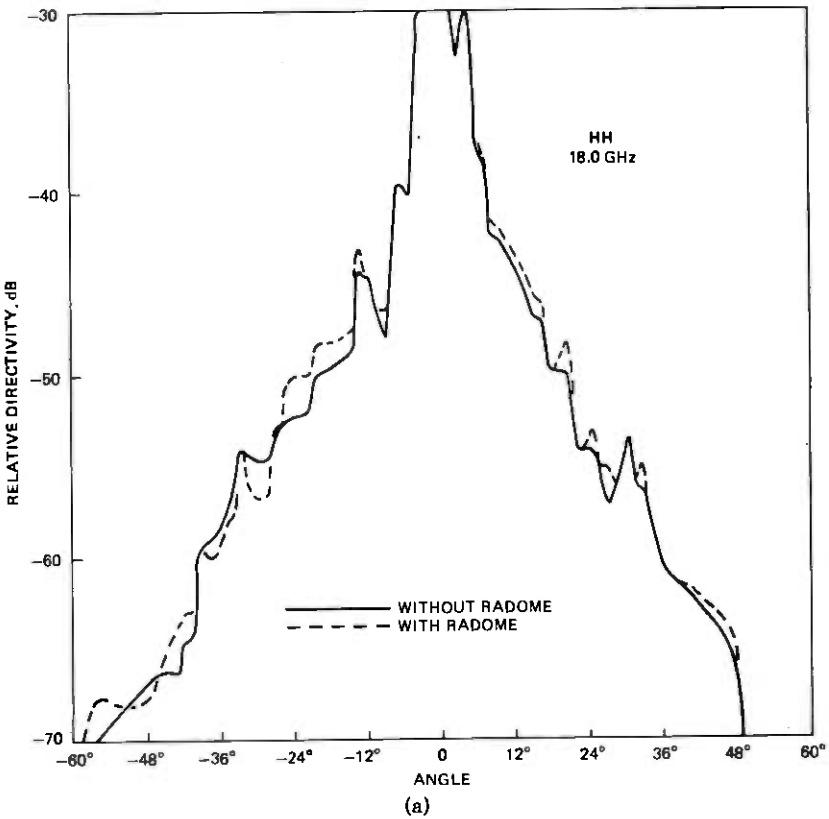
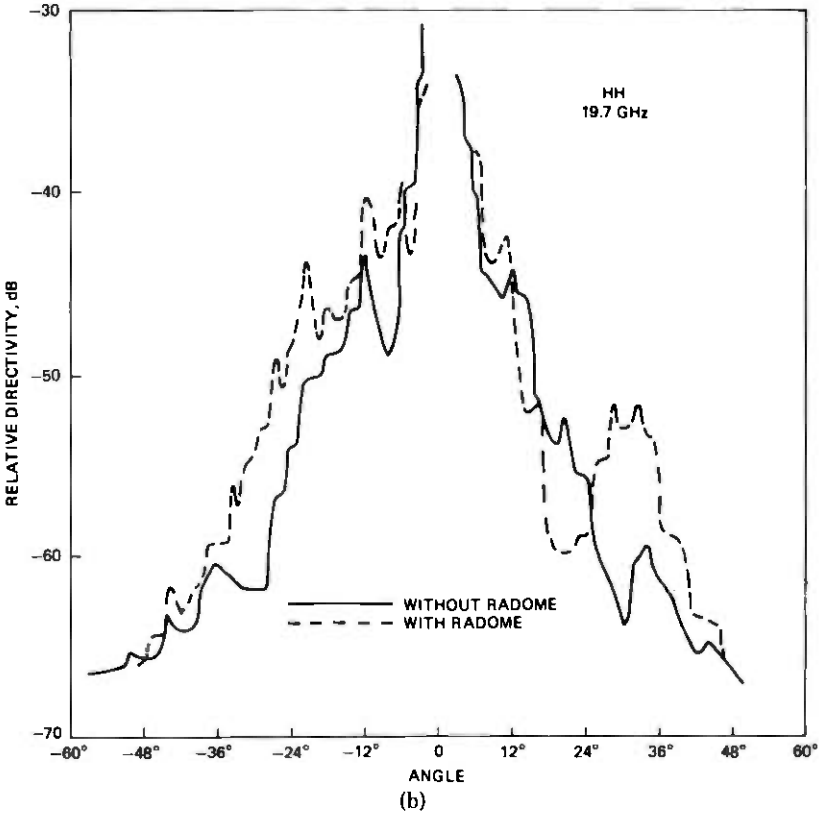


Fig. 14—Influence of half-wavelength solid laminate radome on azimuthal radiation patterns: (a) 18.0 GHz; (b) 19.7 GHz.

beam tilt. The tilt experiments were done with and without a radome on the antenna, and for both polarization states of the antenna. Tilting the mirror is found to have almost no influence on the antenna sidelobe response. At most, 3 or 4 dB changes are noted at −55 dB levels near 30°. These small changes in sidelobe level are felt to be of little consequence because they are so far down from the main beam and do not change the angle at which the patterns drop beneath −60 dB.

5.2 Feed position sensitivity

Feeds with diameters commensurate to a wavelength have phase centers located approximately in the plane of their aperture. Therefore the dual mode feed used in this antenna is positioned to have its aperture coincident with the focal point of the paraboloidal reflector. Nevertheless, it is of interest to determine the sensitivity of feed position to deg-



(b)

Fig. 14—(*continued*)

radation in antenna performance. Three conditions were examined: feed skew (feed axis inclined to paraboloidal axis of symmetry), feed linear displacement (feed and paraboloidal axis parallel, but feed axially or laterally displaced), and feed-polarizer twist. Briefly the tests show that up to 1° of feed skew, axial movement of up to 0.25 centimeter, and lateral displacement of up to 0.25 centimeter have little or no discernible influence. Feed-polarizer twist of up to 3° has no effect on gain, a minimal effect on pattern (first sidelobe increased 1 dB), but as expected, does deteriorate cross polarization discrimination.

These tests indicate that antenna performance is not overly sensitive to feed position, and allow for rather lenient support bracket tolerances which are easily maintained at minimal expense.

VI. ANTENNA RETURN LOSS

Transmitted power which is reflected back to the feed manifests itself as antenna return loss. The sources of reflected power in this antenna are: (*i*) feed mismatch including insect seal, (*ii*) parabolic dish, and (*iii*) radome. These items are individually treated below.

Since the dual mode feed has a circular cross section and the polarization diplexer which will be used with this antenna was designed in square waveguide, a suitable transition was designed to connect the two. Such a transition requires the conversion of two dominant, orthogonal, $TE_{10}^{\square}$ modes in WS-42 to two dominant, orthogonal, TE_{11}° modes in WC-50. For this purpose, a four-inch linear tapered transition with no measurable transmission loss and a return loss of better than 40 dB (reflection coefficient <0.01) was electroformed. Swept frequency return loss measurements on the feed with insect seal and linear transition were made, and are depicted in Fig. 15a. The return loss across the entire frequency range is better than 29 dB, corresponding to a VSWR ≤ 1.075 or reflection coefficient ≤ 0.035.

Return loss of the complete antenna was measured across the band of interest. Figure 15b shows this performance as a solid line for the antenna radiating into free space. The poorest return loss within the band is approximately 23 dB. A simple vector separation analysis, based on the assumption that the only contributions to the total returned power are the feed and paraboloid with radome, produced the dashed line as the contribution of the paraboloid and radome alone. Measurements indicate that with the mirror in its normal position, the contribution of the radome to total return loss is negligible since energy reflected back into the canister by the radome is not focused at the feed. This is still true with the antenna beam tilted down 3° since the reflected energy from the radome is 3° off boresight.

The resulting dish contribution of approximately 28 dB agrees well with a computed value of 26.8 dB obtained from the equation

$$\text{Return loss} = 20 \log \frac{4\pi f}{\lambda G_f} \tag{7}$$

In this equation, f is the focal length of the paraboloid (0.536 meters), and G_f is the feed gain. Feed gain has been computed as 12.9 dB at 18.7 GHz by direct integration of the normalized feed radiation pattern. The equation above is derived by determining what fraction of power radiated by the feed, W_f, is recaptured by the feed. The power density in the axial

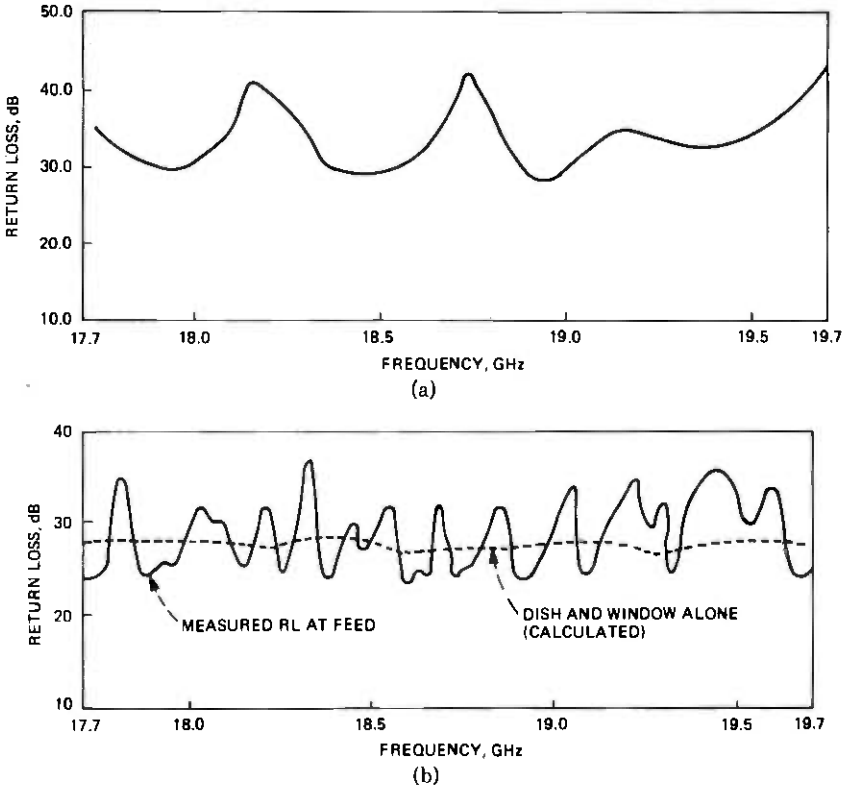


Fig. 15—Return loss performance: (a) dual mode feed alone; (b) return loss with feed in canister antenna.

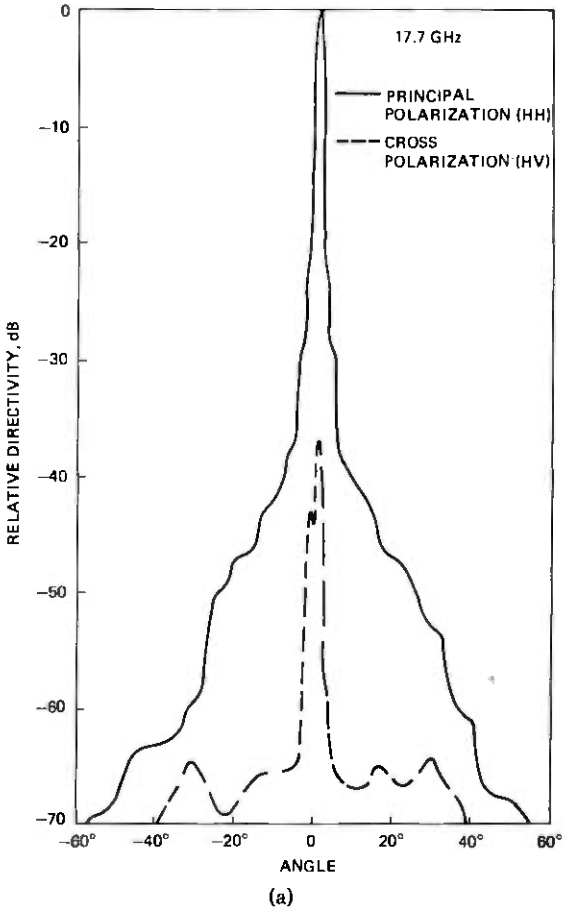
region of the dish is $P_d = W_f G_f / 4\pi f^2$. This power density is reflected back toward the feed as a plane wave and captured with effective area $A_{\text{eff}} = \lambda^2 G_f / 4\pi$. The power captured by the feed, W_c, is $W_c = A_{\text{eff}} P_d = W_f (\lambda G_f / 4\pi f)^2$, and eq. (7) follows directly.

VII. CONCLUSION AND SUMMARY OF ANTENNA CHARACTERISTICS

The design and experimental optimization of a canister antenna are reviewed in this paper. The influence of antenna feed, parabolic reflector, radome, absorber and mirror on gain, radiation pattern, and return loss is considered.

The measured gain of the antenna, virtually independent of polarization, is stated in Table I. These values of gain, measured without a radome in place, correspond to an approximate aperture efficiency of 62 percent. The loss of the solid, half-wavelength-thick radome is 0.4 dB.

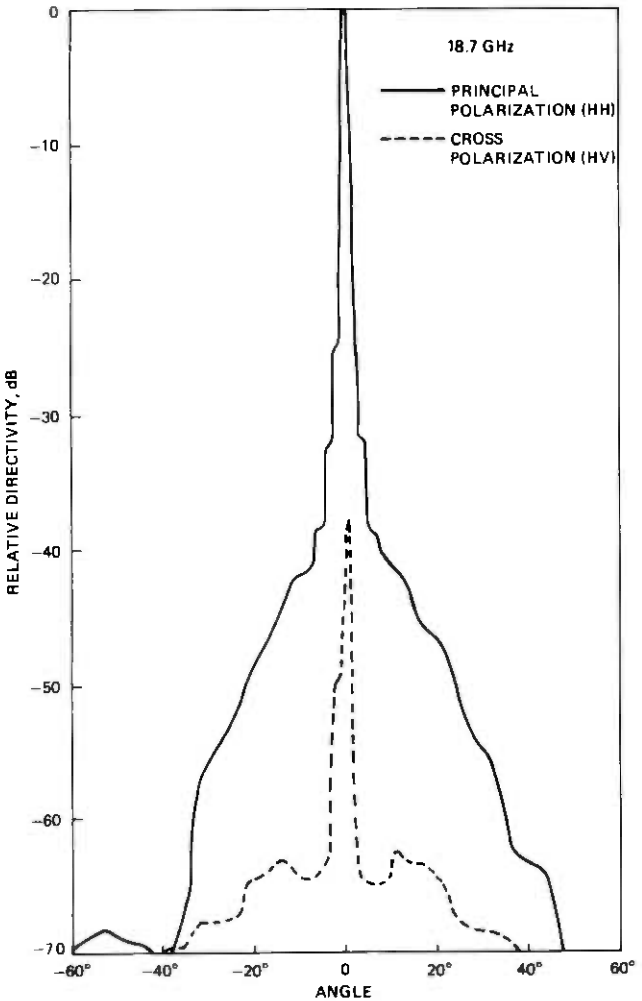
The principal and cross polarized response of the antenna is illustrated



(a)

Fig. 16—Principal (HH) and cross polarized (HV) smoothed azimuthal radiation pattern envelopes measured with radome installed: (a) 17.7 GHz; (b) 18.7 GHz; (c) 19.7 GHz.

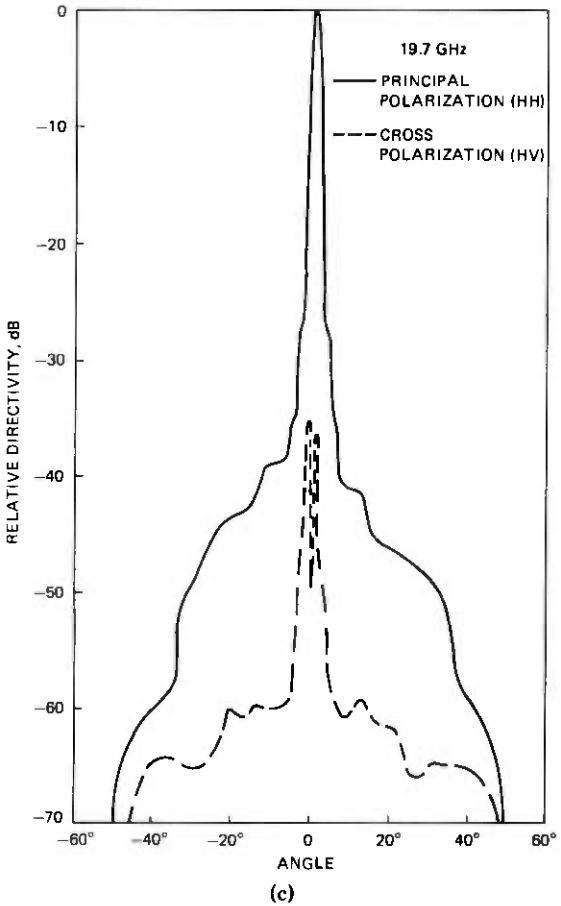
by the smoothed radiation patterns presented in Fig. 16. This figure presents the 17.7 GHz, 18.7 GHz and 19.7 GHz horizontally polarized response (HH) and the vertically polarized antenna response to a horizontally polarized transmitted signal (HV) with the half-wavelength radome on the antenna. Similar radiation patterns are obtained for the VV and VH polarizations. This is to be expected since the antenna design is essentially polarization independent provided the dual mode feed affords a balanced E and H plane illumination of the paraboloid. As noted earlier, such a balanced illumination is reasonably achieved. The principal polarization patterns include perturbation introduced by the radome. A comparison of these patterns with corresponding patterns measured without a radome reveals that sidelobe levels beneath −40 dB



(b)

Fig. 16—(*continued*)

are primarily affected, with the largest pattern deterioration occurring at the lowest radiation levels, and being most pronounced the further the frequency departs from 18.3 GHz (the "tuned" frequency for this radome). From Fig. 16, we also note that the on-axis cross polarization discrimination is in the mid to upper 30 dB range, confirming an estimate offered in an earlier paper dealing with this antenna concept.[4] The influence of mirror tilt and feed positioning is also assessed. It is shown that modest amounts of mirror tilt (e.g., ±1.5°) and inaccurate feed positioning have minimal effect upon the antenna pattern and gain.



(c)

Fig. 16—(*continued*)

Return loss measurements have been made on the complete antenna and individual components. The major contributors to reflected power are feed mismatch and the parabolic dish. The worst-case return loss of the feed with transition is approximately 29 dB. The return loss of the paraboloid is estimated to be 28 dB. With the mirror in its nominal position, the radome contribution is negligible. The poorest total return loss measured (which occurs at a different frequency than the worst-case feed contribution) within the band is approximately 23 dB.

As mentioned earlier, the antenna described in this paper was designed for an 18 GHz digital radio system. The particular antenna described herein was specifically designed to afford certain degrees of flexibility which would not be required after identification of those parameters which influence antenna performance. For example, the canister was

oversized to permit installation of a variety of absorbing materials and so the paraboloid could be translated up and down relative to the mirror. With completion of the performance characterization, certain changes were made in the final antenna design. The final design will allow beam pointing of ±8°. To accomplish this and assure adequate return loss, the radome will be fastened to the antenna in such a manner that the beam will not point within 1.5° of the normal to the radome surface. Perhaps the most significant change to be implemented in the final design is a larger paraboloidal reflector with a 0.864-meter diameter. This reflector (which necessitates a somewhat larger mirror) will afford approximately 0.5 dB more gain. This increased gain is just slightly more than the 0.4 dB loss introduced by the radome. It is expected that the antenna with a 0.864 meter dish will have a radiation pattern quite similar to that measured on the developmental model, and a total return loss of approximately 23 dB.

VIII. ACKNOWLEDGMENT

The authors are pleased to acknowledge the following Bell Laboratories colleagues and their contributions: A. B. Crawford and R. H. Turrin for their initial studies and subsequent insights into aspects of the antenna performance; I. Anderson for his work in experimental and analytical interpretation of the antenna radiation; N. R. Lampert for various aspects of the physical design of the antenna; and C. P. Bates for helpful discussions and general guidance of the activities.

REFERENCES

1. C. A. Siller, Jr., and P. E. Butzien, "A Radio Relay Antenna for Application at 18 GHz," 1974 IEEE/AP-S Symposium Program and Digest, Atlanta, Georgia, June 10–12, 1974, pp. 253–255.
2. A. C. Longton, "DR 18—A High Speed QPSK System at 18 GHz," 1976 International Conference on Communications, Conference Record, *II*, Philadelphia, June 14–16, 1976, pp. 18–14 to 18–17.
3. J. J. Kenny, "18-GHz Channelization for 275 Mb/s Transmission," Paper presented at URSI meeting, Boulder, Colorado, August 22, 1973.
4. A. B. Crawford and R. H. Turrin, "A Packaged Antenna for Short-Hop Microwave Radio Systems," B.S.T.J., *48*, No. 6 (July–August 1969), pp. 1605–1622.
5. T. S. Chu, "Maximum Power Transmission Between Two Reflector Antennas in the Fresnel Zone," B.S.T.J., *50*, No. 4, April 1971, pp. 1407–1420.
6. D. Herbison-Evans, "Optimum Paraboloid Aerial Design," Report No. 66012, Signals Research and Development Establishment, Ministry of Aviation, Christchurch, Hants, September 1966.
7. I. Anderson, unpublished memoranda.
8. R. H. Turrin, "Dual Mode Small Aperture Antennas," IEEE Trans. Ant. Prop. (Communication), *AP-15*, No. 2, March 1967, pp. 307–308.
9. R. H. Turrin, personal communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 3, March 1978
Printed in U.S.A.

Polarization Effects in Short Length, Single Mode Fibers

By V. RAMASWAMY, R. D. STANDLEY, D. SZE, and W. G. FRENCH

(Manuscript received December 14, 1976)

The ability to maintain linearly polarized output in single mode fibers is essential for utilization of polarization dependent receiver circuitry. Our measurements with long lengths of fiber (200 m) indicate that we can find input polarization angles which yield essentially linearly polarized output. However, we found that these polarization effects are greatly influenced by the presence of physical stress on the fiber such as stress due to bending, twisting, mounting, and other variations in ambient conditions. We conducted several experiments on short length fibers where special precautions were adopted to assure repeatability of the measurements. Our results indicate the existence of a general theoretical model that predicts the output polarization characteristics as a function of input polarization and fiber length. The model assumes the presence of two asynchronous, orthogonal modes, uniformly coupled over the entire fiber length. The model, however, cannot distinguish between uniformly coupled and uncoupled mode cases based on the output radiation measurements.

I. INTRODUCTION

The polarization characteristics of "single" mode optical fibers have been the subject of several previous publications.[1,2] An understanding of the polarization sensitivity of such fibers is important in assessing the applicability of polarization dependent optical circuitry.[3] An additional implication of polarization sensitivity is the introduction of delay distortion in "single" mode fibers.

In this paper, we present a theoretical model based on the propagation of two orthogonal modes; it is shown that based on the measurement of

the ellipticity of the output radiation alone, we cannot distinguish between the following cases:

(*i*) The existence of two orthogonal modes, uncoupled with different propagation constrants.

(*ii*) The existence of two orthogonal modes, with identical propagation constants and uniformly coupled by some means of periodic perturbation

(*iii*) The most general case—two orthogonal modes, uniformly coupled, but with nonidentical propagation constants.

The organization of the paper is as follows. In Section II, we summarize the theory, leaving the details to the Appendices. Section III describes the detailed experimental procedure utilized to measure the radiation ellipse as well as details of each of the measurements. In Section IV the experimental data is compared with simple theoretical results developed in Section II.

It should be noted that the assumption of synchronous uncoupled modes will not verify our data. An additional finding of importance is that the most general case indicates that for a fiber of any given length, excitation conditions exist at the input that result in linearly polarized output.

II. THEORETICAL MODEL

For completeness, we state the obvious: for synchronous uncoupled modes, the output radiation would be linearly polarized independent of the orientation of linear input polarization and fiber length. Experimentally, this is not observed and, hence, this model is ruled inappropriate.

As detailed in Appendices B and C, we assume spatially orthogonal modes whose propagation constants differ by $\Delta\beta$; we further assume that the modes are uniformly coupled with a constant coupling κ. The fiber output radiation ellipse,[4] as described in Appendix A, would possess the following characteristics. If a and b are the amplitudes of the semi-major and minor axis components of the radiation ellipse, then their ratio is

$$R = \pm \frac{a}{b} = \frac{1 \pm [1 - \sin^2 2\theta' \sin^2 2\alpha]^{1/2}}{\sin 2\theta' \sin 2\alpha} \tag{1}$$

and the orientation ψ of the major axis of the ellipse is

$$\psi = 1/2 \tan^{-1} (\tan 2\theta' \cos 2\alpha) - \eta/2 \tag{2}$$

where $\theta' = \theta + \eta/2$

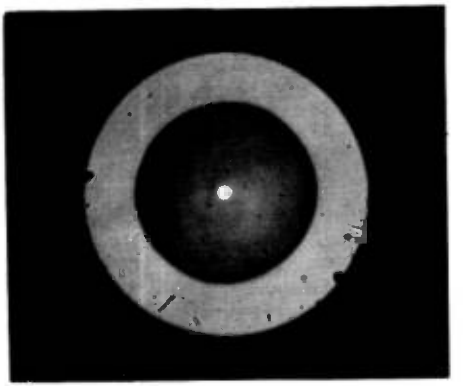


Fig. 1—Photomicrograph of single mode fibers, core diameter = 4.7 μ, outside diameter = 133 μ.

θ = input orientation of polarization with respect to x axis
$\eta = \tan^{-1} 2\kappa/\Delta\beta$
$\alpha = 1/2\sqrt{\Delta\beta^2 + 4\kappa^2}\, z$
z = fiber length

When $R = 0$ or ∞, the output is linearly polarized. From eq. (1), we see that this happens at $\theta' = \pm m\pi/2$, independent of fiber length. Thus, this model predicts that even in the presence of either phase asynchronism or mode coupling (or both), there are specific orientations of input polarization,

$$\theta = -\eta/2 \pm m\pi/2 \tag{3}$$

for which linear polarization is observed for all lengths. The measured relative phase shift $2\alpha = \sqrt{\Delta\beta^2 + 4\kappa^2}\, z$, can be thought of as due to an effective $\Delta\beta_e$ that includes the effect of coupling such that

$$\Delta\beta_e = \sqrt{\Delta\beta^2 + 4\kappa^2} \tag{4}$$

In our measurements, the reference angle $\theta' = 0$ was always selected at first by rotating the input polarization to that angle for which the output was linearly polarized; then based on the observations of the radiation ellipse at the output, we cannot distinguish between the following cases: (*i*) asynchronous, uncoupled modes, (*ii*) synchronous, coupled modes, and (*iii*) asynchronous, coupled modes. These cases are summarized in Table I.

III. EXPERIMENTAL PROCEDURE

Figure 1 shows a photomicrograph of the fiber when illuminated with white light. The fibers used had a core of pure SiO_2 and a cladding of

Table I—Summary of cases

Case	2α	$\theta' = 0$ (reference) at	$R = \frac{a}{b}$	ψ
Asynchronous ($\Delta\beta \neq 0$) Uncoupled ($\kappa = 0$)	$\Delta\beta z$	$\theta = 0$	$\frac{1 \pm \sqrt{1 - \sin^2 2\theta \sin^2 2\alpha}}{\sin 2\theta \sin 2\alpha}$	$\frac{1}{2}\tan^{-1}[\tan 2\theta \cos 2\alpha]$
Synchronous ($\Delta\beta = 0$) Uniformly coupled ($\kappa \neq 0$)	$2\kappa z$	$\theta = -\pi/4$	Same as above with θ replaced by θ'; θ rotated by $-\pi/4$	
Asynchronous $\Delta\beta \neq 0$ Uniformly coupled ($\kappa \neq 0$)	$\Delta\beta_e z = \sqrt{\Delta\beta^2 + 4\kappa^2}\, z$	$\theta = -\eta/2$	Same as above with θ replaced by θ'; now θ rotated by $-\eta/2$	

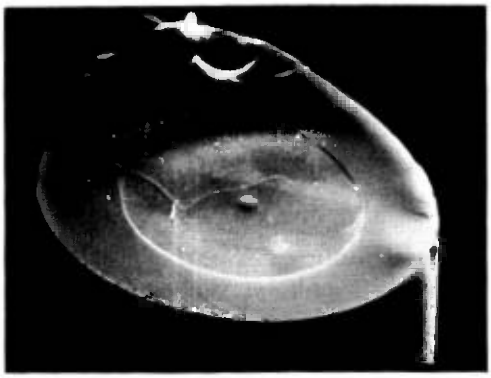


Fig. 2—Electron micrograph of etched single mode fiber.

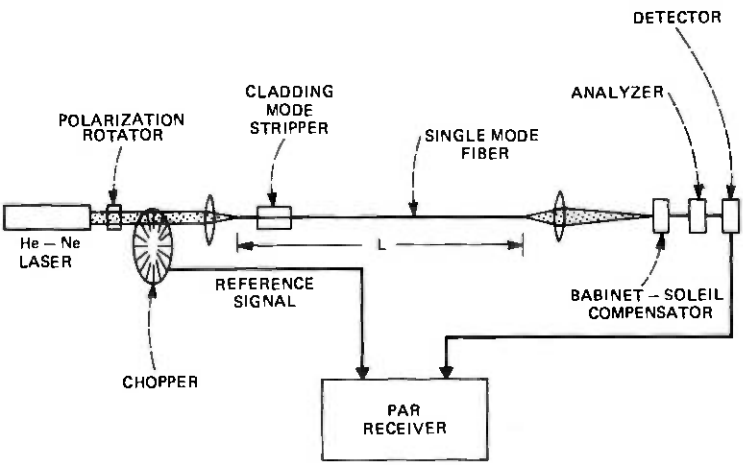


Fig. 3—Experimental set-up.

$B_2O_3 \cdot 6SiO_2$. The core and the outer diameters were 4.7 μm and 133 μm, respectively. Due to diffusion of B_2O_3 in the MCVD fabrication process, the fiber core was unintentionally graded in refractive index and the Δn was lower than expected for these compositions.[5] Figure 2 shows an electron micrograph of the etched fiber.[6] The effective core-cladding index difference was $\Delta n \approx 2 \times 10^{-3}$. Fibers of lengths usually less than 3 meters were used in our short fiber length measurements.

The experimental arrangement is shown in Fig. 3. Light from a polarized He-Ne laser is passed through a polarization rotator to facilitate rotation of the input polarization at the input of the fiber. The light was chopped to provide a reference signal at the receiver and then focused

on the fiber by means of a lens. A small section of the fiber near the input end was immersed in glycerol to remove any undesired cladding mode that might have been excited. The output light was collimated by a lens and passed through a Glan-Thomson analyzer before being detected and displayed in a PAR receiver. A Babinet-Soleil compensator could be inserted and removed, as needed, between the collimating lens and the analyzer.

Experimentally, we found the polarization effects on long and short length fibers were greatly influenced by the presence of any stress on the fiber and other changes in ambient conditions. To ensure repeatability of the measurement, it was necessary to take many precautions. The fibers were held as straight as possible by gently taping them on to a flat surface, at regular but not necessarily identical intervals to minimize inducing any stress in the fiber. The input end of the fiber was held by using a vacuum chuck except in one experiment where it was held by a mechanical holder. In that case, it was clamped first using as little pressure as possible and then taped. In all our measurements, the input polarization angle, which resulted in linearly polarized output, served as the input reference angle, i.e., $\theta' = 0$. The analyzer position oriented to measure the cross polarized component served as the reference axis for the orientation of the output ellipse. If we alter the angle θ', the output in general will become elliptically polarized. In order to obtain the phase difference δ between components parallel and perpendicular to the reference angle $\theta' = 0$, we have to orient the Babinet-Soleil compensator parallel or perpendicular to the original reference axis of the analyzer. With the plunger of the compensator adjusted to obtain linear polarization at the output, from the plunger position and the calibration of the compensator, the phase shift 2α can then be calculated. In each measurement, where input polarization is rotated with a given fixed length of the fiber, or where the length of the fiber is varied keeping the input excitation angle θ' fixed, the output analyzer was oriented parallel to the major and minor axis of the output ellipse to obtain the power ratio of these components. Orientation of the minor axis was also determined from the angular position of the analyzer.

IV. EXPERIMENTAL RESULTS

4.1 Fixed fiber length, input polarization varied

In this experiment, an input reference angle was found such that the output polarization was linearly polarized; we then measured the ellipticity of the output radiation as a function of input polarization rotation about this reference. Figures 4 and 5 compare the experimental data with the theoretical results obtained by computing $20 \log_{10} R$ from eq. (1) and evaluating the orientation of eq. (2). The best fit was for $2\alpha = 59°$, which

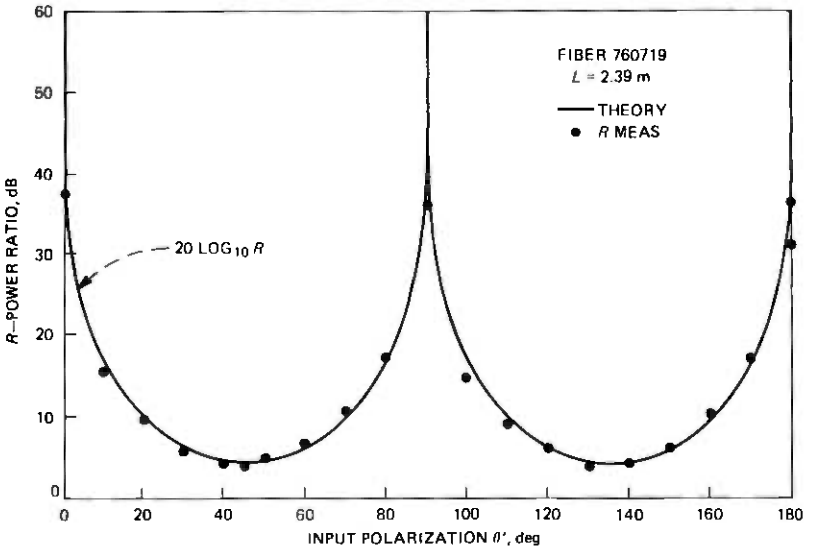


Fig. 4—Measured and theoretical power output ratio, R, of the major to minor axis components of output radiation ellipse as a function of input polarization angle (θ'). Fiber length L = 2.39 meters.

was also verified from the compensator measurement. However, this experiment cannot give a specified value for the phase retardation per unit length, i.e., $\delta/L(=\Delta\beta_c)$, since we can not determine explicitly the integral number of $\pi/2$ phase shifts included in the entire length of the fiber.

4.2 Fixed input polarization, variable length

4.2.1 On axis excitation ($\theta' = 0$)

In this experiment, after finding the reference axis, orientation and power ratio R of the output ellipse was measured; the fiber was shortened at the input end repeatedly by a small amount ($\approx$ 5 mm) and the experiment was repeated by reorienting the input polarization to obtain a linearly polarized output. From Fig. 6, we infer, as the theory predicts, essentially a linearly polarized output independent of the fiber length. The experimental limit of 38 dB indicated in the figure is the limitation imposed by the degree of polarization of the laser source. Our measurements on long lengths ($\approx$200 m) indicate the cross polarized component was down 32 dB. However, the output polarization was subject to severe variations due to physical stress and other changes in ambient conditions.

The data (not shown) indicate the input polarizer angle and the output

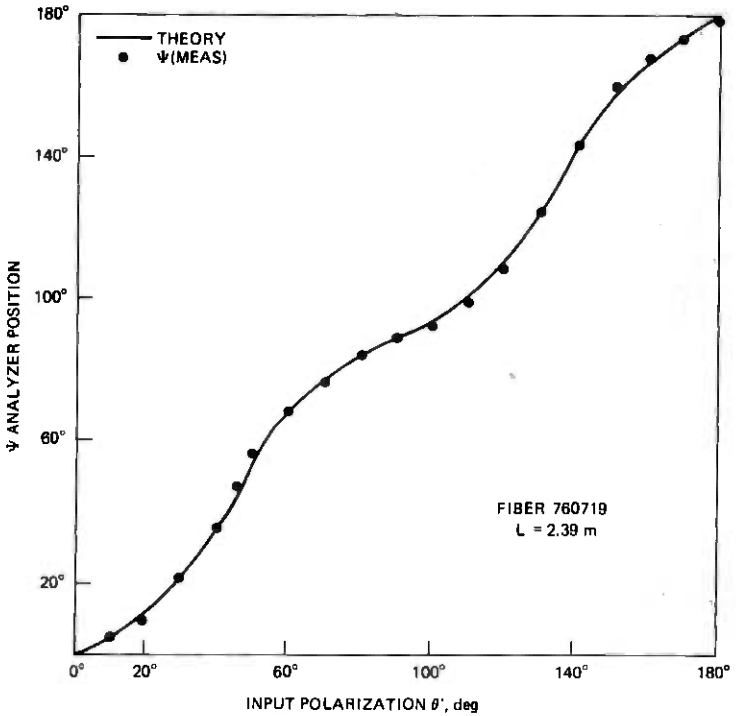


Fig. 5—Measured and estimated relative orientation of the ellipse ψ as a function of input polarization angle θ' for $L = 2.39$ meters.

analyzer angle remained essentially unchanged with deviations, within experimental errors, indicating negligible angular rotation of the fiber with length.

4.2.2 *Slightly off-axis excitation* ($\theta' \approx 6°$)

This experiment was conducted by offsetting the input polarization by an angle θ' ($\approx 6°$). The input fiber end was held mechanically; the output end was successively cut and the measurements were repeated. As expected the output was elliptically polarized; as a function of length, the polarization changes from elliptical to linear and then back to elliptical. From eq. (3) for a small angular offset, i.e., for small θ'

$$|R| \approx \frac{1}{\sin 2\alpha} \tag{5}$$

and from (4)

$$\psi = \frac{1}{2}\tan^{-1}[2\theta' \cos 2\alpha]$$
$$\approx \theta' \cos 2\alpha \tag{6}$$

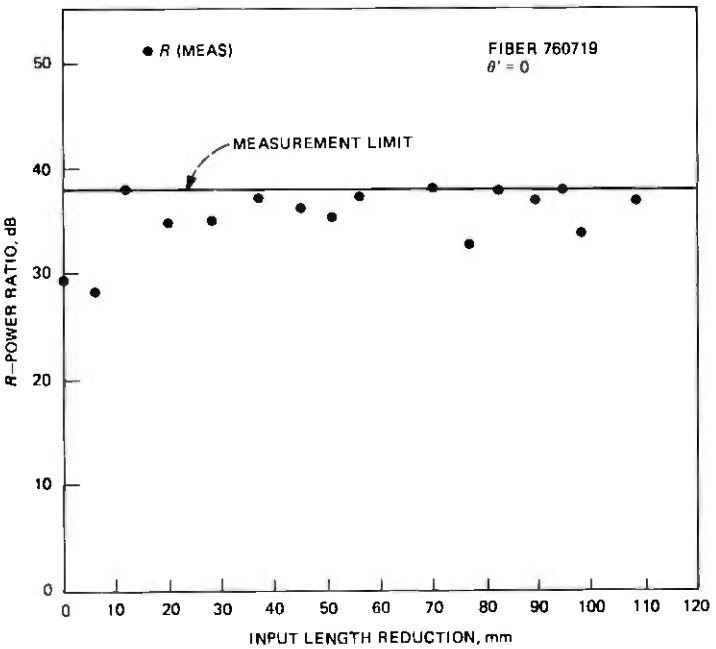


Fig. 6—Essentially linear polarization is observed at the output when input end is shortened with on axis ($\theta' = 0$) excitation.

Figure 7 shows the excellent agreement between theory and experiment. In fitting the ψ data, it was necessary to include a "twist" term of the order of 0.16°/mm. This implied a periodicity of the twist to be about 2.25 m, which is very close to the length of the fiber and was traced later to the holding and mounting arrangement at the input end of the fiber. When the fiber was allowed to lie flat and straight, and the input end was held using a vacuum chuck, the twist disappeared.

From the compensator measurements after each cut, the average $\Delta\beta_e$ was evaluated to be $\approx$0.0581 radians per mm. As seen from Fig. 7, linear polarization occurs at about every 54 mm and from eq. (5), it is obvious that this occurs when $2\alpha = 0, \pi, \ldots$ etc. Therefore $2\alpha = \Delta\beta L = \pi$ with the result $\Delta\beta_e = 0.0582$ radians per mm, in agreement with the compensator measurements. This leads to an estimated value of effective index difference

$$\Delta n_e = \frac{\Delta\beta_e}{2\pi}\lambda = \frac{\lambda}{2L} = 5.86 \times 10^{-6}$$

between the modes. Thus, it seems reasonable to state that the fibers cannot be easily fabricated to such tolerances and, hence, the polarization

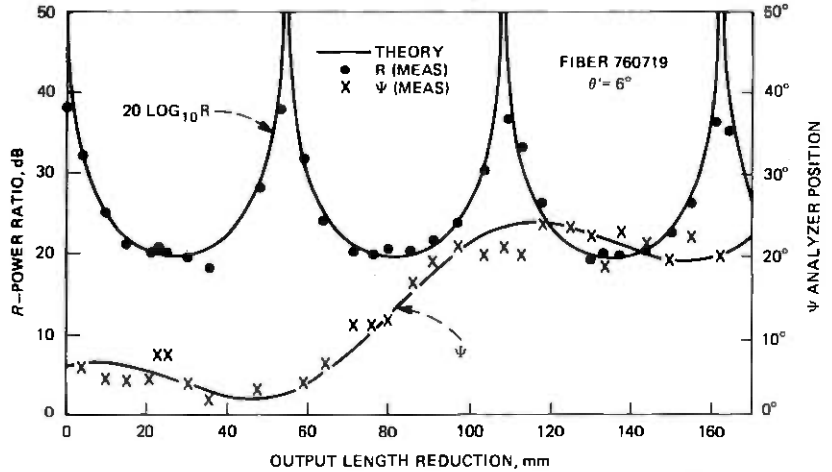


Fig. 7—Analytical predictions that R varies as $(\theta' \sin 2\alpha)^{-1}$, and ψ varies as $(\theta' \cos 2\alpha)$ as a function of length for small θ', confirmed experimentally. Period between successive linear polarization is 54 mm yielding a $\Delta\beta = 0.0582$ radians/mm.

problems observed with these experimental fibers will qualitatively be observed for all nominally circular, single mode fibers.

4.2.3 Excitation at $\theta' = \pi/4$

If the fiber was excited at 45° from the reference angle $\theta' = 0$, two modes of equal amplitudes would be excited; at the fiber output they would be out of phase by $2\alpha = \Delta\beta_e L$. Therefore in this case, the polarization will change, as a function of length, from circular to linear and then back to circular with elliptical polarization in between. For $\theta' = \pi/4$, from eq. (1) and (2),

$$|R| = \frac{1}{\tan \alpha}, \tag{9}$$

and

$$\psi = \pm\pi/4 \tag{10}$$

In eq. (4), when $\cos 2\alpha$ goes through zero, i.e., at $\alpha = \pi/4$ and $|R| = 1$ (circular polarization), ψ changes sign. Experimental results compare very well with the analysis as shown in Fig. 8.

V. CONCLUSIONS

The fibers used in these experiments were nominally circularly symmetric; however, the data on the elliptically polarized output radiation indicates that very small deviations from circular symmetry possibly exist and that these irregularities cause serious polarization effects. It

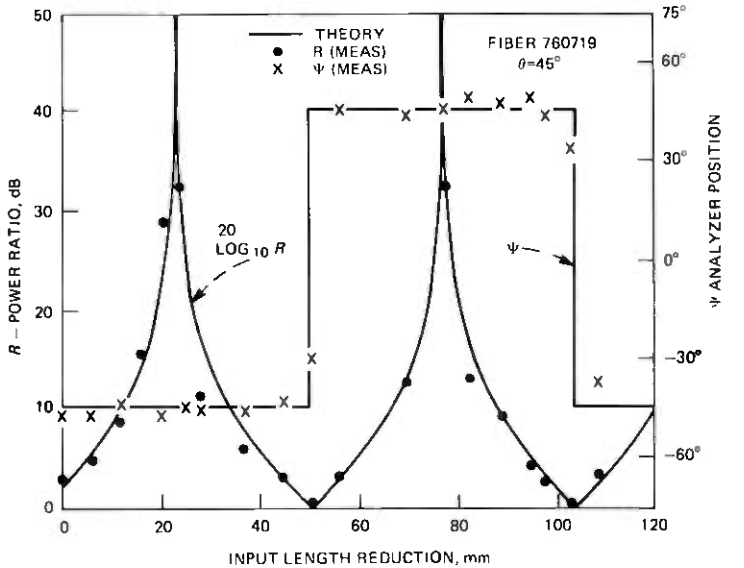


Fig. 8—Behavior of output polarization with decreasing input length. R varies as $(\tan \alpha)^{-1}$ and $\psi = \pm 45°$. The beat period is the same as before—54 mm.

appears unlikely that the fiber fabrication process can be improved to the extent that such effects can be completely eliminated.

The results indicate that input excitation conditions do exist for which linearly polarized output will be observed for a fiber of any length. However, the output polarization is subject to severe variations due to physical stress and other changes in ambient conditions. Our experiments can not determine whether any mode coupling exists. However, if any mode coupling exists, it must be uniform, at least over the few meter lengths of the fiber measured.

Polarization behavior of single mode fibers may hopefully be improved, by making the cores purposely elliptical, provided that any periodic perturbation such as "twisting" of the core is sufficiently small and does not result in strong mode coupling. Large ellipticity, however, will increase the modal dispersion. Further study is necessary to characterize the polarization behavior of elliptical core fibers and, if they hold polarization, to arrive at an optimum degree of ellipticity that would provide a balance between polarization control and dispersion effects.

VI. ACKNOWLEDGMENTS

The authors are grateful to E. A. J. Marcatili and D. Gloge for useful discussions and to P. Kaiser for drawing the fibers used in the experiment.

APPENDIX A

Orientation and Ratio of Major to Minor Axis of Vibrational Ellipse of Elliptically Polarized Waves

Consider two time-varying (S.H.) orthogonal electric field components

$$E_x = a_1 e^{j(\omega t + \delta_1)} \tag{11}$$

$$E_y = a_2 e^{j(\omega t + \delta_2)} \tag{12}$$

such that $a_1^2 + a_2^2 = 1$. The resulting electric vector, as seen from Fig. 9, traces, in general, an ellipse. The difference in the effective propagation constants between the two waves is

$$\Delta\beta = \frac{\beta_1 - \beta_2}{L} = \frac{\delta}{L} \tag{13}$$

where L is the fiber length. Whenever a_1 or a_2 is zero or when $\delta = m\pi$, where m is an integer, the output wave is linearly polarized. Furthermore, when $a_1 = a_2$, and $\delta = (2m + 1)\pi/2$, the output is circularly polarized. The ratio R of the semi-major to semi-minor axis of the ellipse can be defined as[4]

$$R = \frac{a}{b} = \tan \chi \tag{14}$$

where χ is an auxiliary angle ($-\pi/4 \leq \chi \leq \pi/4$). By choosing an angle $\phi (0 \leq \phi \leq \pi/2)$ such that

$$\tan \phi = \frac{a_2}{a_1} \tag{15}$$

the angle ψ of the orientation of the resultant ellipse, with respect to a reference axis ox, and the parameters ϕ and χ are related by[4]

$$\sin 2\chi = \sin 2\phi \sin \delta \tag{16}$$

$$\tan 2\psi = \tan 2\phi \cos \delta \tag{17}$$

In addition,

$$a_1^2 + a_2^2 = a^2 + b^2 \tag{18}$$

By eliminating χ from equations (14) and (16), and with the use of eq. (15), we find

$$R = \pm\frac{a}{b} = \frac{1 \pm \sqrt{1 - 4a_1^2 a_2^2 \sin^2 \delta}}{2a_1 a_2 \sin \delta} \tag{19}$$

and from eqs. (17) and (15),

$$\psi = \frac{1}{2} \tan^{-1}\left[\frac{2a_1 a_2}{a_1^2 - a_2^2} \cos \delta\right] \tag{20}$$

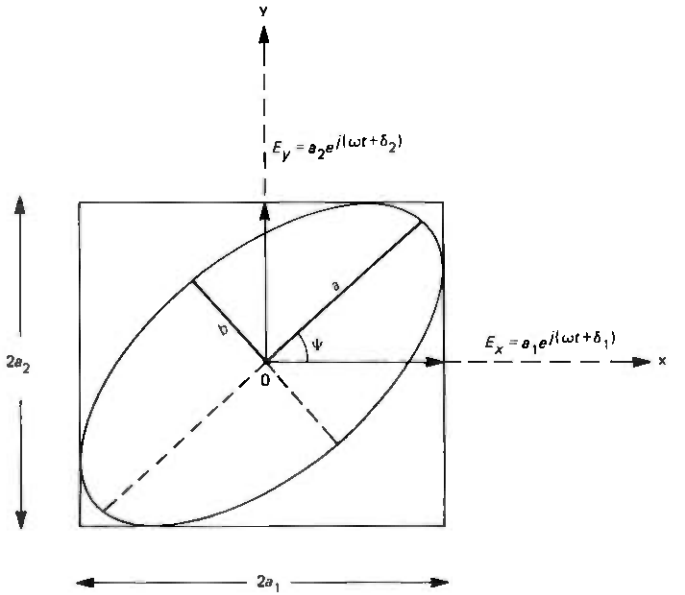


Fig. 9—Output radiation ellipse.

APPENDIX B

Relations Between Input Polarization Angle, Fiber Parameters, and Output Ellipse

We assume a cartesian co-ordinate system, as shown in Fig. 10, with ox and oy coincident with the axes of slightly elliptical core or with that of the birefringent axes of the core of a single mode fiber as the case may be. Let us also assume that a linearly polarized plane wave is incident upon the fiber at an angle θ with respect to ox; therefore the amplitudes of the excited modes with orthogonal polarization are $\cos\theta$ and $\sin\theta$ and represent the x and y component respectively.

Assuming a uniformly distributed coupling characterized by a coupling constant κ between the modes whose propagation constant differ by $\Delta\beta$, the complex field amplitudes are given by[7,8]

$$E_x = \cos\theta\cos\alpha - j\cos(\eta+\theta)\sin\alpha \tag{21}$$

$$E_y = \sin\theta\cos\alpha + j\sin(\eta+\theta)\sin\alpha \tag{22}$$

where the parameter α, proportional to the length of the fiber, is

$$\alpha = \sqrt{(\Delta\beta/2)^2 + \kappa^2}\, z \tag{23}$$

and the parameter η which relate the degree of coupling and the asynchronism between modes is given by

$$\tan\eta = 2\kappa/\Delta\beta \tag{24}$$

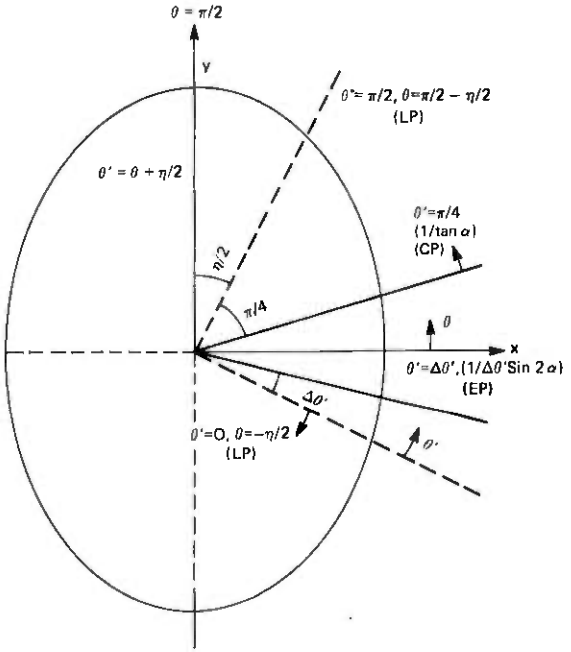


Fig. 10—Input excitation conditions and the behavior of the resulting output polarization.

For reasons indicated in the text, we exclude the case when both κ and $\Delta\beta$ are equal to zero. For the case $\kappa = 0$ and $\Delta\beta$ finite, η equals zero. If $\Delta\beta = 0$ and κ is finite, η equals $\pi/2$. When both κ and $\Delta\beta$ are finite, $0 \leq \eta \leq \pi/2$.

By rewriting eqs. (21) and (22) similar to eqs. (11) and (12), we see that

$$a_1^2 = \cos^2\theta \cos^2\alpha + \cos^2(\eta + \theta) \sin^2\alpha \tag{25}$$

$$a_2^2 = \sin^2\theta \cos^2\alpha + \sin^2(\eta + \theta) \sin^2\alpha \tag{26}$$

and

$$a_1^2 - a_2^2 = \cos 2\theta \cos^2\alpha + \cos 2(\eta + \theta) \sin^2\alpha \tag{27}$$

The phase constant δ is determined by

$$\tan\delta = \frac{2 \sin(2\theta + \eta) \tan\alpha}{\sin 2\theta - \sin 2(\theta + \eta) \tan^2\alpha} \tag{28}$$

Substituting (25) through (28) in (19) and (20) we find, after considerable trigonometric manipulation, that the ratio of the major to minor axis

of the resultant output ellipse is

$$R = \frac{1 \pm [1 - \sin^2 (2\theta + \eta) \sin^2 2\alpha]^{1/2}}{\sin (2\theta + \eta) \sin 2\alpha} \tag{29}$$

and its orientation

$$\psi = \frac{1}{2} \tan^{-1} [\tan (2\theta + \eta) \cos 2\alpha] - \frac{\eta}{2} \tag{30}$$

By appropriate choice of the sign, we see that $|R|$ in eq. (29) is bounded by 1 and ∞ or 0 and 1. Whenever $|R|$ goes to 0 or ∞, we have linear polarization. For a given length of the fiber z, therefore α, this occurs when

$$\theta = -\frac{\eta}{2} \pm m \frac{\pi}{2}, m = 0, 1, 2, \ldots \tag{31}$$

As η varies from 0 to $\pi/2$, the input angle θ at which the output is linearly polarized varies from 0 to $-\pi/4$. In the most general case, at this specific angle, phase difference δ, as seen from equation (28), equals either zero or multiples of π and is independent of α indicating that the two modes are either in or out of phase with each other at the output for all lengths. This clearly shows for a given fiber with a finite κ and $\Delta\beta$, there is always an orientation of input polarization that would result in a linear polarization at the output.

Note that in the special case $\eta = 0, \delta = 2\alpha$ and therefore, δ is dependent on length; however, at $\theta = \pm m\pi/2$, as seen from (25) and (26), one of the components goes to zero. Thus the condition for input excitation angle θ to achieve linear polarization at the output, for all values of α, is given by (31) and holds good for all values of η in the range $0 \leq |\eta| \leq \pi/2$.

APPENDIX C

General Theoretical Model Used for Observation of Output Ellipticity

In this Appendix, we consider the general case illustrated in the previous Appendix and show that two orthogonal *uniformly coupled* asynchronous modes can be resolved into a new set of orthogonal modes, but completely *uncoupled*. This implies, even if any coupling exists, as long as it is uniformly distributed, our results can be explained by assuming uncoupled, asynchronous modes. Obviously, the measured $\Delta\beta$ will now include the effects of coupling.

If no mode coupling exists between the modes, i.e., $\kappa = 0$, then from (23) and (24)

$$\alpha = \frac{\Delta\beta}{2} z \tag{32}$$

and

$$\eta = \tan^{-1}\left(\frac{2\kappa}{\Delta\beta}\right) = 0 \tag{33}$$

Under these conditions, from (21) and (22),

$$E_x = \cos\theta e^{-j\alpha} \tag{34}$$

$$E_y = \sin\theta e^{+j\alpha} \tag{35}$$

We have shown in the previous Appendix that in the most general case $0 < \eta/2 < \pm\pi/2$, there exists an input polarization angle such that $\theta = -\eta/2$, for which the output polarization remains linear for all lengths. As shown in Fig. 10, if we rotate the coordinate system by $-\eta/2$, then the angle θ' in the new coordinate system is

$$\theta' = \theta + \eta/2 \tag{36}$$

But substituting (36) in (21) and (22), and writing the amplitudes along Ox' and Oy' as

$$E'_x = E_x \cos\eta/2 - E_y \sin\eta/2 \tag{37}$$

and

$$E'_y = E_x \sin\eta/2 + E_y \cos\eta/2 \tag{38}$$

Using eqs. (34) through (38) we can easily show that

$$E'_x = \cos\theta' e^{-j\alpha} \tag{39}$$

$$E'_y = \sin\theta' e^{+j\alpha} \tag{40}$$

Eqs. (39) and (40) are identical to (34) and (35) with θ being replaced by θ'. Thus the coupled orthogonal modes can be easily resolved in terms of uncoupled orthogonal modes. Obviously, although α remains the same as in (23), it can be defined as in (32) to include the effects of the nonzero coupling as

$$\alpha = \frac{\Delta\beta_e}{2} z = \sqrt{\left(\frac{\Delta\beta}{2}\right)^2 + \kappa^2}\, z \tag{41}$$

By substituting (36) in (29) we find the ratio of major to minor axis of output ellipse is now given by

$$R = \frac{1 \pm (1 - \sin^2 2\theta' \sin^2 2\alpha)^{1/2}}{\sin 2\theta' \sin 2\alpha} \tag{42}$$

Therefore excitation at the input polarization reference angle ($\theta' = 0$) for which $R \to 0$ or ∞, the output remains linearly polarized for all lengths. Then if we vary θ' at the input, eq. (42) represents the ratio of

axes of output ellipse. The orientation of the output ellipse is now given by

$$\psi = \frac{1}{2}\tan^{-1}\left[\tan 2\theta' \cos 2\alpha\right] - \eta/2 \tag{43}$$

REFERENCES

1. F. P. Kapron, N. F. Borelli, and D. B. Keck, "Birefringence in Dielectric Optical Waveguides," IEEE J. Quant. Electronics, *8* (1972), p. 222.
2. W. Eickoff and O. Krumpholz, "Determination of The Ellipticity of Monomode Glass Fibers from Measurements of Scattered Light Intensity," Electron. Lett., *12* (1976), p. 405.
3. R. A. Steinberg and T. G. Giallorenzi, "Performance Limitations Imposed on Optical Waveguide Switches and Modulators by Polarization," Appl. Optics, *15* (1976), p. 2440.
4. M. Born and E. Wolf, *Principles of Optics*, 5th ed., Oxford: Pergamon Press.
5. G. W. Tasker, P. Kaiser, W. G. French, J. R. Simpson, H. M. Presby, and L. L. Blyler, Jr., "Low-Loss, Single-Mode Fibers with Different B_2O_3-SiO_2 Compositions," Appl. Optics, to be published.
6. H. M. Presby, R. D. Standley, J. B. MacChesney, and P. B. O'Connor, "Material Structure of Ge Doped Optical Fibers and Preforms," B.S.T.J., *54*, No. 10 (December 1975), pp. 168–1992.
7. S. E. Miller, "Coupled Wave Theory and Waveguide Applications," B.S.T.J., *33*, No. 3 (May 1954), pp. 661–719.
8. V. Ramaswamy and R. D. Standley, "A Phased, Optical Coupler Pair Switch," B.S.T.J., *55*, No. 6 (July–August 1976), pp. 767–776.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 3, March 1978
Printed in U.S.A.

Steady-State Response of a Well-Balanced Wire Pair to Distributed Interference

By W. N. BELL

(Manuscript received March 22, 1976)

We show that the longitudinal circuit defined by a well-balanced wire pair in a cable can be studied independently of the metallic circuit. The metallic circuit is then excited by the longitudinal voltage and current acting through the wire pair and terminal unbalances. Discrete parameter longitudinal circuits are defined which have the same terminal response as the distributed-parameter longitudinal circuit. These equivalent circuits are studied under an electrically short assumption, yielding simple expressions for their terminal response. The electrically short assumption enables the distributed impressed voltage which excites the longitudinal circuit to be represented by only two parameters, the "total impressed voltage" and the "center of impressed voltage." These parameters are analogous to the total mass and center of mass of a thin filament or wire. Finally, an analysis of a longitudinal circuit defined by a subscriber loop excited by a nearby power distribution system is used to derive a relationship between the short-circuit longitudinal current at the central office and the open-circuit longitudinal voltage at the telephone set. This relationship is used to estimate the distribution of short-circuit longitudinal current at the central office from a known distribution of open-circuit longitudinal voltage at the telephone set.

I. INTRODUCTION

The problem of computing the steady-state response of a multiconductor system has received considerable attention in the literature. Carson and Hoyt[1] developed the classical transmission line equations and S. O. Rice[2] developed the mathematical techniques necessary for their solution. Even so, the complexities introduced by a large number of conductors tend to limit the amount of basic understanding of fundamental problems, such as the effects of longitudinal induction and longitudinal-to-metallic conversion, that can be obtained by pursuing

the multiconductor problem. Moreover, a large number of conductors are necessarily described by a large number of parameters upon which there is often a paucity of data.

We present an in-depth analysis of the steady-state response of a well-balanced wire pair to distributed interference. This simplification of the general problem enables the analysis to continue beyond the formal solution to develop both an intuitive feel for the problem and a simple model for use in engineering applications. The effect of other pairs can be approximated in the one-pair model by a judicious choice of model parameters.

Historically, the primary emphasis has been on characterizing the longitudinal and metallic voltage at the subscriber's telephone set because of the impact of these voltages on the quality of the communication's path.[6,7] More recently, the use of electronic loop terminating equipment has generated interest in the longitudinal current at the central office. This study was motivated by the need to characterize the longitudinal current at the central office and to better understand the roles that the terminal and wire pair unbalances play in longitudinal-to-metallic conversion.

II. SUMMARY OF RESULTS

We begin with the classical transmission line equations which define the steady-state response of the wire pair when excited by a distributed impressed voltage. A transformation to the longitudinal and metallic voltages and currents is employed to study the longitudinal and metallic circuits defined by the wire pair. We show that if the wire pair and its terminations are "well-balanced," then the longitudinal circuit (LC) can be studied independently of the metallic circuit (MC). The MC is then excited by the longitudinal voltage and current acting through the wire pair and terminal unbalances. The unbalances admit the following interpretations: (*i*) longitudinal current flowing through the distributed impedance unbalance of the wire pair can be represented as a distributed series voltage generator in the MC, (*ii*) longitudinal voltage across the distributed admittance unbalance of the wire pair can be represented as a distributed shunt current generator in the MC, and (*iii*) longitudinal current through a discrete impedance unbalance in a termination can be represented as a discrete voltage generator in the termination for the MC.

The response of the LC is studied in some detail. Discrete parameter circuits are defined which have the same terminal response as the distributed parameter LC. These equivalent circuits are studied under an electrically short assumption, yielding simple expressions for their terminal response. The electrically short assumption enables the distributed

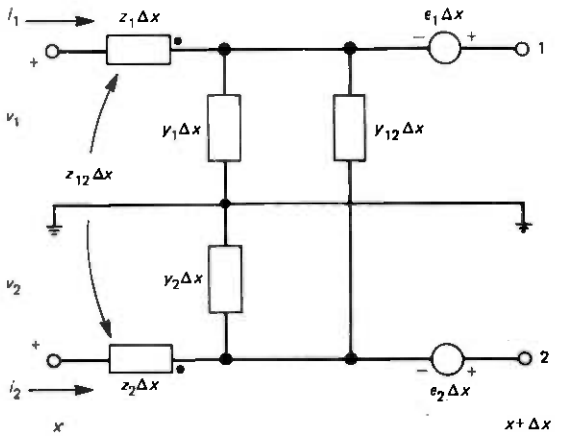


Fig. 1—Incremental circuit model.

impressed voltage to be represented by only two parameters, the total impressed voltage and the center of impressed voltage. These parameters are analogous to the total mass and center of mass of thin filament or wire. Finally, an analysis of the LC defined by a subscriber loop excited by a nearby power distribution system is used to derive a relationship between the short-circuit longitudinal current at the central office and the open-circuit longitudinal voltage at the telephone set. This relationship is used to estimate the distribution of short-circuit longitudinal current at the central office from the known distribution of open-circuit longitudinal voltage at the telephone set.

III. BASIC EQUATIONS

3.1 Transmission line equations

The transmission line equations for a wire pair in a cable excited by a distributed impressed voltage acting as a fixed radian frequency are

$$v_1'(x) = -z_1(x)i_1(x) - z_{12}i_2(x) + \epsilon_1(x)$$

$$i_1'(x) = -[y_1(x) + y_{12}]v_1(x) + y_{12}v_2(x) \qquad (1)$$

$$v_2'(x) = -z_2(x)i_2(x) - z_{12}i_1(x) + \epsilon_2(x)$$

$$i_2'(x) = -[y_2(x) + y_{12}]v_2(x) + y_{12}v_1(x). \qquad (2)$$

Boundary conditions, discussed in Section 3.2, are determined from terminations at the ends ($x = 0$ and ℓ) of the wire pair. The above equations are essentially Rice's[2] eqs. (1.1) and (1.2) except for our notation which is motivated by the incremental circuit model of Fig. 1. Notice that three types of unbalances are possible at x; an impedance

unbalance $[z_1(x) \neq z_2(x)]$, an admittance unbalance $[y_1(x) \neq y_2(x)]$ and an unbalance in the impressed voltage $[\epsilon_1(x) \neq \epsilon_2(x)]$.

The metallic and longitudinal voltages and currents are defined as follows:

$$v(x) = v_1(x) - v_2(x) = \text{metallic voltage}$$

$$i(x) = \frac{1}{2}[i_1(x) - i_2(x)] = \text{metallic current} \tag{3}$$

$$v_g(x) = \frac{1}{2}[v_1(x) + v_2(x)] = \text{longitudinal voltage}$$

$$i_g(x) = i_1(x) + i_2(x) = \text{longitudinal current.} \tag{4}$$

Transforming the transmission line equations [eqs. (1) and (2)] to the metallic and longitudinal voltages and currents yields

$$v'(x) = -zi(x) + \delta_z(x)i_g(x) + \delta_\epsilon(x)$$

$$i'(x) = -yv(x) + \delta_y(x)v_g(x) \tag{5}$$

$$v_g'(x) = -z_g i_g(x) + \delta_z(x)i(x) + \epsilon_g(x)$$

$$i_g'(x) = -y_g v_g(x) + \delta_y(x)v(x). \tag{6}$$

The parameters of eqs. (5) and (6) are defined in terms of the parameters of eqs. (1) and (2) and Fig. 1 as follows: The impedances in ohms per unit length are

$$z = z_1(x) + z_2(x) - 2z_{12}$$

$$z_g = \frac{1}{4}[z_1(x) + z_2(x) + 2z_{12}]$$

$$\delta_z = \frac{1}{2}[z_2(x) - z_1(x)]. \tag{7}$$

The admittances in mhos per unit length are

$$y = \frac{1}{4}[y_1(x) + y_2(x)] + y_{12}$$

$$y_g = y_1(x) + y_2(x)$$

$$\delta_y(x) = \frac{1}{2}[y_2(x) - y_1(x)]. \tag{8}$$

The impressed voltages in volts per unit length are

$$\epsilon_g(x) = \frac{1}{2}[\epsilon_1(x) + \epsilon_2(x)]$$

$$\delta_\epsilon(x) = \epsilon_1(x) - \epsilon_2(x). \tag{9}$$

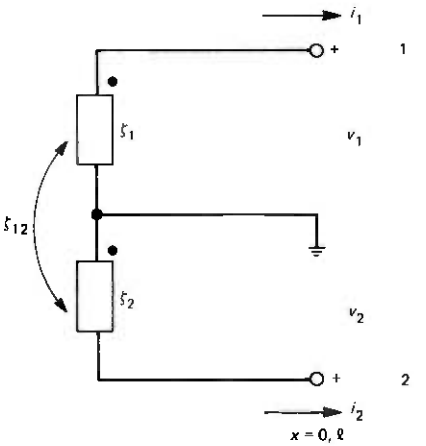


Fig. 2—Termination circuit model.

Only the unbalances, which now appear explicitly, and the impressed voltages remain x-dependent; all other parameters are assumed uniform.

3.2 Boundary conditions

The boundary conditions necessary to determine a particular solution to eqs. (1) and (2), or equivalently eqs. (5) and (6), are determined from the terminations at the ends of the wire pair. Consider the canonical passive-symmetric termination of Fig. 2. The boundary conditions for the voltages and currents relative to ground are

$$\begin{aligned} v_1 &= -\zeta_1 i_1 - \zeta_{12} i_2 \\ v_2 &= -\zeta_2 i_2 - \zeta_{12} i_1. \end{aligned} \tag{10}$$

Transforming to the metallic and longitudinal voltages and currents [eqs. (3) and (4)] yields

$$v = -\zeta i + \Delta_\zeta i_g \tag{11}$$

$$v_g = -\zeta_g i_g + \Delta_\zeta i. \tag{12}$$

The parameters of eqs. (11) and (12) are defined in terms of the parameters of eq. (10) and Fig. 2 as follows:

$$\begin{aligned} \zeta &= \zeta_1 + \zeta_2 + 2\zeta_{12} \\ \zeta_g &= \frac{1}{4}(\zeta_1 + \zeta_2 - 2\zeta_{12}) \\ \Delta_\zeta &= \frac{1}{2}(\zeta_2 - \zeta_1). \end{aligned} \tag{13}$$

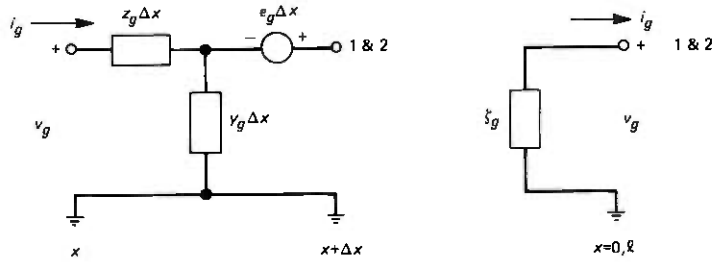


Fig. 3—Longitudinal circuit model.

3.3 Longitudinal and metallic circuits

The metallic and longitudinal voltages and currents supported by a wire pair are determined by eqs. (5) and (6) together with boundary conditions of the type given in eqs. (11) and (12) at the ends of the wire pair. Notice that if the wire pair and its terminations are perfectly balanced, then the metallic voltage and current are identically zero. Hence it follows by continuity[3] that if the wire pair and its terminations are well-balanced (i.e., the unbalances are small relative to their associated longitudinal parameters*), then the metallic voltage and current are small relative to the longitudinal voltage and current.

Under a well-balanced assumption, the second-order terms in eqs. (6) and (12) can be neglected, leaving

$$\begin{aligned} v_g'(x) &= -z_g i_g(x) + \epsilon_g(x) \\ i_g'(x) &= -y_g v_g(x) \end{aligned} \tag{14}$$

with boundary conditions of the form

$$v_g = -\zeta_g i_g. \tag{15}$$

Notice that the longitudinal voltage and current can be assumed independent of both the metallic voltage and current and the system unbalances. Moreover, the circuit models of Fig. 3 which represent the above equations define the LC.

Now assume that $v_g(x)$ and $i_g(x)$ are known in eqs. (5) and (11). Then the metallic voltage and current satisfy

$$\begin{aligned} v'(x) &= -zi(x) + \epsilon(x) \\ i'(x) &= -yv(x) + \xi(x) \end{aligned} \tag{16}$$

with boundary conditions of the form

* Cable pairs have an average resistance unbalance of 2 percent and an average capacitance unbalance to ground of 0.5 percent.

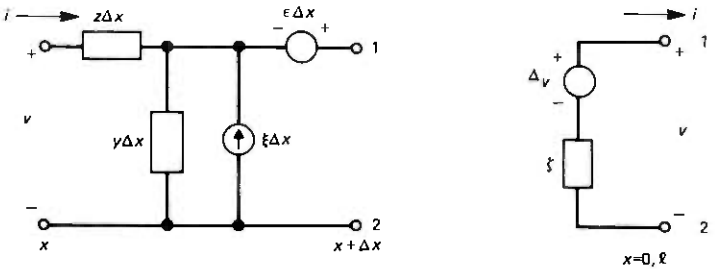


Fig. 4—Metallic circuit model.

$$v = -\zeta i + \Delta_\zeta i_g \tag{17}$$

where

$$\epsilon(x) = \delta_z(x) i_g(x) + \delta_\epsilon(x)$$
$$\xi(x) = \delta_y(x) v_g(x). \tag{18}$$

are all known. The circuit models in Fig. 4 represent the above equations and define the MC. Note that the x-dependence of the wire pair unbalances pose no analytical problems since they appear in the forcing functions $\epsilon(x)$ and $\xi(x)$. Moreover, superposition can be applied to yield a decomposition of the metallic voltage and current as a sum of terms due to each unbalance acting separately.

Let us summarize our results. A well-balanced wire pair with well-balanced terminations defines a longitudinal and a metallic circuit. The LC can be assumed independent of the MC. The MC is excited by the longitudinal voltage and current acting through the wire pair and terminal unbalances which can be interpreted as follows: longitudinal current $i_g(x)$ following through the distributed impedance unbalance of the wire pair $\delta_z(x)$ can be represented as a distributed series voltage generator $\delta_z(x) i_g(x)$ in the MC, longitudinal voltage $v_g(x)$ across the distributed admittance unbalance of the wire pair $\delta_y(x)$ can be represented as a distributed shunt current generator $\delta_y(x) v_g(x)$ in the MC, and longitudinal current i_g flowing through the discrete impedance unbalance δ_ζ in a termination can be represented as a discrete voltage generator $\Delta_\zeta i_g$ in the termination for the MC. Finally, superposition can be applied to yield a decomposition of the metallic voltage and current as a sum of terms due to each unbalance acting separately. The equations describing the LC and MC, since they have constant coefficients, are amenable to standard techniques[3] which are applied to analyze the LC in the next section.

IV. LONGITUDINAL CIRCUIT ANALYSIS

4.1 Discrete-parameter equivalent circuits

Equation (14) describing the response of the LC defined by a well-balanced wire pair is conveniently represented as a forced linear system,

$$\begin{bmatrix} v_g(x) \\ i_g(x) \end{bmatrix}' = -\begin{bmatrix} 0 & z_g \\ y_g & 0 \end{bmatrix}\begin{bmatrix} v_g(x) \\ i_g(x) \end{bmatrix} + \begin{bmatrix} \epsilon_g(x) \\ 0 \end{bmatrix}. \tag{19}$$

The boundary conditions defined by the longitudinal impedances ζ_g^0 and ζ_g^ℓ of the well-balanced terminations at the ends ($x = 0$ and ℓ) of the wire pair are, from eq. (15),

$$\begin{aligned} v_g(0) &= -\zeta_g^0 i_g(0) \\ v_g(\ell) &= \zeta_g^\ell i_g(\ell). \end{aligned} \tag{20}$$

The solution of eq. (19) is of the form

$$\begin{bmatrix} v_g(\ell) \\ i_g(\ell) \end{bmatrix} = \Phi_g(\ell)\begin{bmatrix} v_g(0) \\ i_g(0) \end{bmatrix} + \int_0^\ell \Phi_g(\ell - \xi)\begin{bmatrix} \epsilon_g(\xi) \\ 0 \end{bmatrix} d\xi. \tag{21}$$

Since the LC is assumed uniform (i.e., z_g and y_g are independent of x), the transition matrix of the LC, $\Phi_g(\xi)$, can be expressed in terms of the characteristic impedance $k_g = \sqrt{z_g/y_g}$ and propagation constant $\gamma_g = \sqrt{z_g y_g}$ of the LC,

$$\Phi_g(\xi) = \begin{bmatrix} \cosh \gamma_g \xi & -k_g \sinh \gamma_g \xi \\ -\dfrac{1}{k_g} \sinh \gamma_g \xi & \cosh \gamma_g \xi \end{bmatrix}. \tag{22}$$

Substituting the boundary conditions [eq. (20)] and the above expression for $\Phi_g(\xi)$ in eq. (21) yields after some manipulation

$$\begin{bmatrix} \zeta_g^0 \cosh \gamma_g \ell + k_g \sinh \gamma_g \ell & \zeta_g^\ell \\ \zeta_g^0 \sinh \gamma_g \ell + k_g \cosh \gamma_g \ell & -k_g \end{bmatrix}\begin{bmatrix} i_g(0) \\ i_g(\ell) \end{bmatrix} = \begin{bmatrix} \displaystyle\int_0^\ell \epsilon_g(\xi) \cosh \gamma_g(\ell - \xi) d\xi \\ \displaystyle\int_0^\ell \epsilon_g(\xi) \sinh \gamma_g(\ell - \xi) d\xi \end{bmatrix}. \tag{23}$$

The formal solution of eq. (23) for $i_g(0)$ and $i_g(\ell)$ is of the form

$$i_g = \frac{E_g}{\zeta_g + Z_g} \tag{24}$$

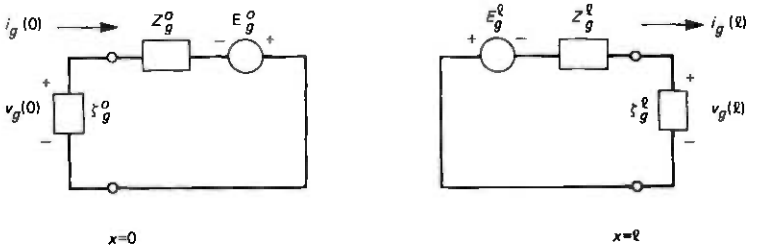


Fig. 5—Discrete-parameter longitudinal circuit models.

where E_g is the longitudinal source and Z_g is the longitudinal impedance seen looking into the LC from one end, terminated in ζ_g at the other end. These quantities define two discrete parameter circuits (see Fig. 5) which have the same terminal response as the distributed parameter circuit of Fig. 3.

General expressions for E_g and Z_g at each end of the LC are given below:

$$E_g^0 = \frac{\int_0^\ell \epsilon_g(\xi)\left[\cosh \gamma_g(\ell-\xi) + \frac{\zeta_g^\ell}{k_g}\sinh \gamma_g(\ell-\xi)\right] d\xi}{\cosh \gamma_g \ell + \frac{\zeta_g^\ell}{k_g}\sinh \gamma_g \ell} \tag{25}$$

$$E_g^\ell = \frac{\int_0^\ell \epsilon_g(\xi)\left[\cosh \gamma_g \xi + \frac{\zeta_g^0}{k_g}\sinh \gamma_g \xi\right] d\xi}{\cosh \gamma_g \ell + \frac{\zeta_g^0}{k_g}\sinh \gamma_g \ell} \tag{26}$$

$$Z_g^0 = \frac{\zeta_g^\ell + k_g \tanh \gamma_g \ell}{1 + \frac{\zeta_g^\ell}{k_g}\tanh \gamma_g \ell} \tag{27}$$

$$Z_g^\ell = \frac{\zeta_g^0 + k_g \tanh \gamma_g \ell}{1 + \frac{\zeta_g^0}{k_g}\tanh \gamma_g \ell}. \tag{28}$$

These expressions simplify if the LC is terminated in its characteristic impedance, k_g. For example, setting $\zeta_g^\ell = k_g$ in eqs. (25) and (27) yields

$$E_g^0 = \int_0^\ell \epsilon_g(\xi) e^{-\gamma_g \xi} d\xi$$

$$Z_g^0 = k_g. \tag{29}$$

Note that the voltage impressed furthest from $x = 0$ suffers the most attenuation as one would intuitively expect.

4.2 The electrically short longitudinal circuit

An LC is electrically short if its electrical length, $|\gamma_g \ell|$, is small. This is typically the case when the source of impressed voltage is a nearby power distribution system. If $|\gamma_g \ell|$ is small, then the hyperbolic functions can be approximated by the first terms of their power series expansions; $\sinh \gamma_g \ell \approx \gamma_g \ell$, $\cosh \gamma_g \ell \approx 1$, and $\tanh \gamma_g \ell \approx \gamma_g \ell$. These approximations when substituted into eqs. (25)–(28) yield approximations of the longitudinal source and longitudinal impedance,

$$\hat{E}_g^0 = E \frac{1 + \zeta_g^\ell y_g (\ell - \bar{\ell})}{1 + \zeta_g^\ell y_g \ell} \tag{30}$$

$$\hat{E}_g^\ell = E \frac{1 + \zeta_g^0 y_g \bar{\ell}}{1 + \zeta_g^0 y_g \ell} \tag{31}$$

$$\hat{Z}_g^0 = \frac{\zeta_g^\ell + z_g \ell}{1 + \zeta_g^\ell y_g \ell} \tag{32}$$

$$\hat{Z}_g^\ell = \frac{\zeta_g^0 + z_g \ell}{1 + \zeta_g^0 y_g \ell}. \tag{33}$$

The quantities E and $\bar{\ell}$ are defined as the total impressed voltage,

$$E = \int_0^\ell \epsilon_g(\xi) d\xi \tag{34}$$

and the center of impressed voltage,

$$\bar{\ell} = \frac{1}{E} \int_0^\ell \xi \epsilon_g(\xi) d\xi. \tag{35}$$

These definitions have a physical intrepretation if the impressed voltage is a real valued and nonnegative function. In this case, $E \geq 0$ and $0 \leq \bar{\ell} \leq \ell$ and the center of impressed voltage can be interpreted as that point along the LC where the impressed voltage may be concentrated without changing the terminal response of the LC. Hence a particular E and $\bar{\ell}$ define a class of equivalent impressed voltages whose canonical representative is a point source of strength E based at $\bar{\ell}$. These concepts have their mechanical analogs in the total mass and the center of mass of a thin filament or wire. Recall that the center of mass is that point along the wire where the total mass may be concentrated while preserving moments about the ends of the wire. In our case, the center of impressed voltage is that point along the LC where the impressed voltage may be concentrated while preserving the response at the ends of the LC.

V. ENGINEERING APPLICATION

5.1 The longitudinal response of an electrically short subscriber loop

The most common LC is a subscriber loop excited by a nearby power distribution system. This circuit is typically electrically short at the fundamental frequency of the impressed voltage (60 Hz). Longitudinally, a subscriber loop has a low impedance termination at the central office by virtue of the battery supply circuit, and essentially an open-circuit termination at the telephone set, assuming single-party or isolated ringer service. Consequently, the longitudinal quantities of primary interest are the short-circuit current at the central office, I_g, and the open-circuit voltage at the telephone set, V_g.

We now derive a simple relationship between these two fundamental quantities. To begin, I_g can be expressed as

$$I_g = \frac{\hat{E}_g^0}{\hat{Z}_g^0} \tag{36}$$

where $\hat{E}_g^0$ and $\hat{Z}_g^0$ are given for an arbitrary ζ_g^ℓ in eqs. (30) and (32). Letting $\zeta_g^\ell \to \infty$ in these equations yields

$$\hat{E}_g^0 = \frac{E(\ell - \bar{\ell})}{\ell} \tag{37}$$

$$\hat{Z}_g^0 = \frac{1}{y_g \ell}. \tag{38}$$

Substitution of these results into eq. (36) yields

$$I_g = E y_g (\ell - \bar{\ell}). \tag{39}$$

Similarly, V_g can be expressed as

$$V_g = \hat{E}_g^\ell \tag{40}$$

where $\hat{E}_g^\ell$ is given in eq. (31) for an arbitrary ζ_g^0. Setting $\zeta_g^0 = 0$ yields

$$V_g = E. \tag{41}$$

Hence I_g and V_g satisfy

$$I_g = V_g y_g (\ell - \bar{\ell}). \tag{42}$$

The link between V_g and I_g is the center of exposure, $\bar{\ell}$.

5.2 Estimating V_g

Historically, emphasis has been placed on characterizing the open-circuit longitudinal voltage at the telephone set because it can be converted to an interfering metallic voltage (V) across the telephone set if

unbalances are present in the loop.[7] The historical measure of loop balance[6] is

$$\text{BAL} = 20 \log \left| \frac{V_g}{V} \right| \quad \text{dB.} \tag{43}$$

The balance of the loop determines the amount of longitudinal voltage that will be converted to metallic voltage by the loop unbalances. This can be expressed mathematically as

$$N = N_g - \text{BAL.} \tag{44}$$

The metallic circuit noise (N) and the longitudinal circuit noise (N_g, usually called noise to ground) are defined as

$$N = 20 \log \left| \frac{V}{V_R} \right| \quad \text{dBrn} \tag{45}$$

$$N_g = 20 \log \left| \frac{V_g}{V_R} \right| \quad \text{dBrn.} \tag{46}$$

The reference noise (dBrn) voltage is 24.5 μV which corresponds to 1 pW across a 600 Ω resistor. The commonly used 3-type noise measuring set attenuates a longitudinal noise measurement by 40 dB. Hence 40 dB must be added to a measured value to obtain N_g in dBrn.

The distributions of noise and loop balance for Bell System loops were determined in 1964 as part of the General Loop Survey of physical and transmission characteristics of the loop plant.[4] Noise measurements were made on 1100 randomly selected loops during normal working hours at the subscriber's telephone set using both 3-kHz flat and C-message frequency weighting. C-message weighting approximates the frequency response of the telephone set and the human ear, and 3-kHz flat-weighting assigns equal weight to all frequencies in the 0 to 3 kHz band. Since the primary interferer on subscriber loops is a nearby power distribution system with a fundamental frequency of 60 Hz, a 3-kHz flat measurement is dominated by the 60-Hz component. Hence the rms voltage corresponding to N_g measured with 3-kHz flat weighting is essentially 60 Hz. The distribution of this voltage over the 1100 loops is shown in Fig. 6. The maximum voltage was 18 V. Ninety-nine percent of the loops had a measured voltage of less than 11 V. The average voltage was 1.5 V with a standard deviation of 2.1 V.

5.3 Estimating I_g

The range of longitudinal current at the central office affects both the operation of existing loop terminating equipment and the design of new equipment. In this section, we estimate the distribution of longitudinal

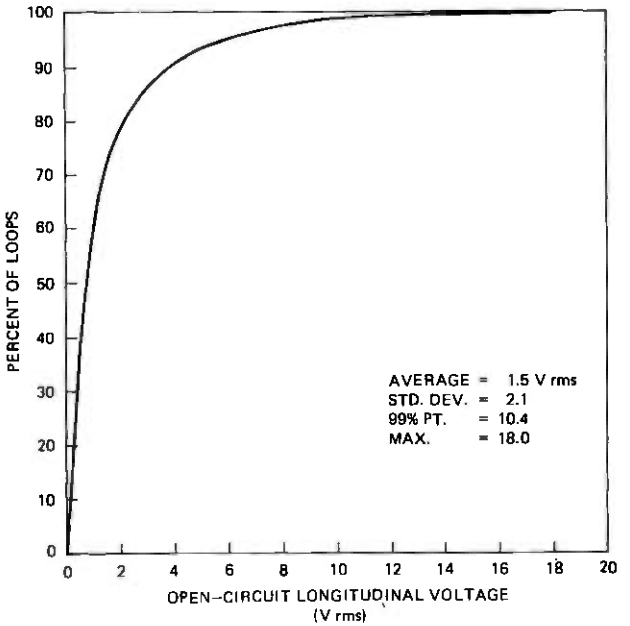


Fig. 6—Distribution of open-circuit rms longitudinal voltage at the telephone set (1964 General Loop Survey).

current using the distribution of longitudinal voltages and loop lengths determined as part of the 1964 General Loop Survey.

The basic equation [eq. (42)] relating I_g and V_g can be expressed in the form

$$|I_g| = |V_g||y_g|\ell(1 - \bar{\ell}/\ell). \tag{47}$$

The ratio $\bar{\ell}/\ell$ is a measure of where the exposure is centered along the loop. If the exposure is centered at the subscriber ($\bar{\ell}/\ell = 1$), then the short-circuit current at the central office is zero. Conversely, the current is maximum if the exposure is centered over the central office ($\bar{\ell}/\ell = 0$). In this case,

$$\max|I_g| = |V_g||y_g|\ell.$$

Equation (47) can be used to estimate the distribution of $|I_g|$ from the distributions of $|V_g|$ (see Fig. 6) and ℓ (see Fig. 2 of Ref. 4) and an assumed value of $\bar{\ell}/\ell$. The admittance of the longitudinal circuit is primarily capacitive at 0.17 μF per mi. Distributions of $|I_g|$ at 60 Hz for $\bar{\ell}/\ell = 0$, 0.5, and 0.9 are shown in Fig. 7. In the worst-case condition of $\bar{\ell}/\ell = 0$, the maximum current is 12 mA. Ninety-nine percent of all loops had a current of less than 4 mA. The average current was 0.3 mA with a

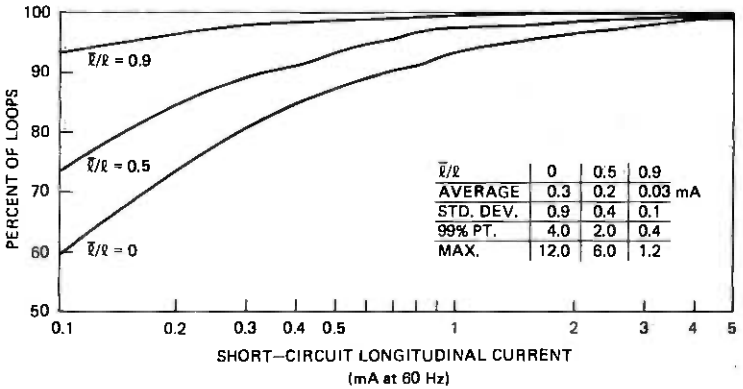


Fig. 7—Estimated distributions of short-circuit 60-Hz longitudinal current at the central office (1964 General Loop Survey).

standard deviation of 0.9 mA. If the exposures are centered at the midpoints of the loops ($\bar{\ell}/\ell = 0.5$), then the above currents are reduced by a factor of 2. If the exposures are centered near the subscriber ($\bar{\ell}/\ell = 0.9$), then the above currents are reduced by a factor of 10, i.e., the maximum current reduces to 1.2 mA and the average current reduces to 0.03 mA.

5.4 Estimating $\bar{\ell}/\ell$

The estimated distribution of the short-circuit longitudinal current at the central office is very sensitive to where the exposures are centered along the loops. An estimate of $\bar{\ell}/\ell$ can be made from simultaneous measurements of $|I_g|$ and $|V_g|$ using another form of eq. (47),

$$\bar{\ell}/\ell = 1 - \frac{|I_g|}{|V_g||y_g|\ell}. \tag{48}$$

A limited number of near-simultaneous measurements were made as part of the CO Strata Survey described in Ref. 5. Single near-simultaneous measurements of $|I_g|$ and $|V_g|$ at 60 Hz were made during normal working hours for each of the 47 test loops. These data were not discussed in Ref. 5 but were made available to us by D. N. Heirman. The ratio $\bar{\ell}/\ell$ was calculated for each test loop using eq. (48). The distribution of the calculated values over the 47 loops is shown in Fig. 8. All exposures were centered beyond the midpoints of the loops. The average value of $\bar{\ell}/\ell$ was 0.88. Hence the exposure of the average loop was centered at almost 90 percent of the loop length. These results are intuitively pleasing since the worst exposures (long aerial parallels with single-phase) are indeed

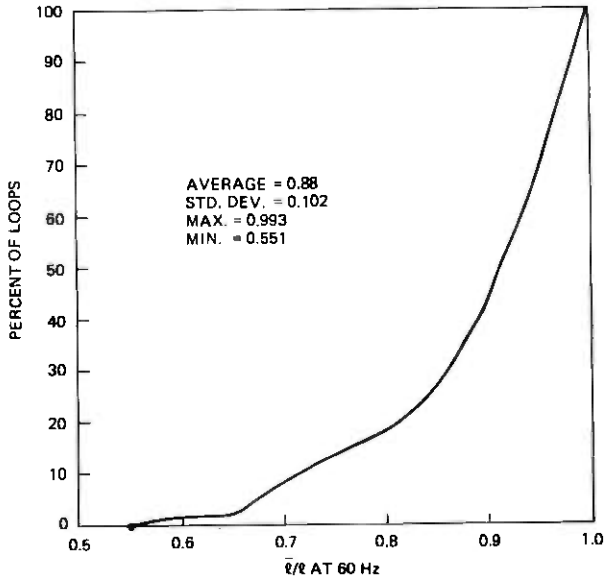


Fig. 8—Distribution of $\bar{\ell}/\ell$ at 60 Hz (Central Office Strata Survey).

over the far ends of the loops. If this distribution of $\bar{\ell}/\ell$ is representative, then the worst case condition is not $\bar{\ell}/\ell = 0$ but $\bar{\ell}/\ell = 0.5$. In this case (see Fig. 7), the maximum current at 60 Hz. is 6 mA, 99 percent of all loops had a current of less than 2 mA and the average current was 0.15 mA.

5.5 Use of the data

The data presented in this section are based on one-time measurements made during normal working hours as part of the 1964 General Loop Survey. The demand for commercial power has doubled since 1964, and one would expect the induced voltages to have increased as a result. In addition, recent surveys[5] have shown that time-of-day variation in 60-Hz induction is likely to be 2 to 1. In residential areas in particular, the induction is likely to be higher in the evening than during the day when the 1964 measurements were made. For these reasons, the reader is cautioned to use the data presented in this section with care.

VI. ACKNOWLEDGMENTS

We are grateful for the many stimulating and enlightening discussions with G. J. Foschini, E. W. Geer, Jr., D. N. Heirman, D. W. McLellan, and other colleagues at Bell Laboratories.

REFERENCES

1. J. R. Carson, and R. S. Hoyt, "Propagation of Periodic Currents Over a System of Parallel Wires," B.S.T.J., *6* (July 1927), pp. 495–545.
2. S. O. Rice, "Steady State Solutions of Transmission Line Equations," B.S.T.J., *20*, No. 2 (April 1941), pp. 131–178.
3. E. A. Coddington, and N. Levinson, *Theory of Ordinary Differential Equations*, New York: McGraw-Hill, 1955.
4. P. A. Gresh, "Physical and Transmission Characteristics of Customer Loop Plant," B.S.T.J., *48*, No. 10 (December 1969), pp. 3337–3386.
5. D. N. Heirman, "Time Variations and Harmonic Content of Inductive Interference in Urban/Suburban and Residential/Rural Telephone Plant," IEEE Trans. Commun., *Com-23*, No. 12 (December 1975).
6. A. J. Aikens and D. Lewinski, "Evaluation of Message Circuit Noise," B.S.T.J., *39*, No. 4, (July 1960), pp. 879–909.
7. D. A. Lewinski, "A New Objective for Message Circuit Noise," B.S.T.J., *43*, No. 2, (March 1964), pp. 719–740.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 3, March 1978
Printed in U.S.A.

An Inversion Technique for the Laplace Transform with Application to Approximation

By D. L. JAGERMAN

(Manuscript received July 26, 1977)

Properties of a sequence of positive operators defined by the Widder Laplace inversion formula are studied in order to obtain practical methods for the inversion of the Laplace transform, practical error formulae, and useful approximations to given functions. The approximation procedure retains essential structural characteristics of the original function, e.g., nonnegativity, monotonicity, and convexity. Thus a distribution function is approximated by distribution functions. Enhancement techniques are provided for the improvement of accuracy for a given order of approximation. The methods are illustrated by applications to renewal theory and to the covariance and recovery functions of telephone traffic theory.

I. INTRODUCTION

The Laplace transform occurs frequently in investigations of queueing theory and telephone traffic models in which it usually represents a probability distribution function. Although the mean and variance of the distribution can be readily obtained from the transform, there are many investigations in which the distribution itself is needed; in particular, good analytic and numerical approximations for the complementary distribution when the argument is large. This is the case, for example, when studying waiting times of queues, time delays of work through a computer system, and delays of message progress through data networks.

Numerical methods which have been made thus far[16-18] concentrate on accurate numerical approximation on some interval $[0, T]$, the difficulty of accurate inversion increasing with increasing T. Methods depending on Gauss-Legendre quadrature applied to the defining Laplace integral with subsequent interpolation are discussed in Ref. 19. These methods require the solution of large order linear systems whose matrices are severely ill-conditioned; thus they can bog down in meaningless

calculations. Much ingenuity has been used in specific cases to circumvent this problem. Asymptotic formulae may sometimes be used to approximate the complementary distribution for large argument; however, in many practical cases, good accuracy was obtained only when the argument was so large that the corresponding probabilities were too small to be of practical significance. One of the methods of this paper, namely, the α enhancement procedure, specifically attacks this problem by imitating the exponential decay of the original in $[T, \infty]$ while simultaneously providing accurate approximation in $[0, T]$. The transition region is sufficiently well approximated for most practical uses.

The well-known Laplace inversion formula of Widder[1,3] has not been actively used in practical work. It has been the experience of the author that an investigation of the Widder formula qua functional transformation can provide useful practical techniques for inversion and also inequalities and limit relations between the approximations and the original function. Accordingly, it is the object of this paper to study the properties of a sequence of positive operators defined by the Widder formula in order to obtain practical methods for Laplace inversion, practical error formulae, and useful approximations to given functions.

In II the Widder inversion formula is obtained and a sequence of positive operators, L_n, which form the subject of the paper, are introduced. The L_n map a function $f(t)(t \geq 0)$ to a sequence of functions $f_n(t) = L_n f$ which converge uniformly on $[0, \infty]$ to $f(t)$. This viewpoint enables one to study the approximation characteristics of the sequence $f_n(t)$, thus providing a means of approximating a given $f(t)$ besides effecting the inversion of its Laplace transform, $\tilde{f}(s)$. Several representations are given for $f_n(t)$ in terms of $f(t)$.

In III properties of the sequence $\{f_n(t)\}$ are developed which show that it possesses many desirable characteristics. In many applications it is preferable that the approximating functions globally imitate the original function in qualitative structural features rather than to the attainment of very high numerical accuracy. Thus if the original function lies between zero and one, is monotone decreasing, and is convex, then these same properties would be desired in the approximation. It is shown that the approximating sequence, $\{f_n(t)\}$, does retain those properties. A recursion relation for $f_{n+1}(t)$ in terms of $f_n(t)$, and a generating function for the sequence are also given, thus making the computation of higher approximations possibly more convenient than the direct application of the representation formulae themselves. A useful feature of the $f_n(t)$ is that, when $f(t)$ is convex, they satisfy $f_n(t) \geq f(t)$.

Part IV develops error bounds and pointwise error estimates. The results in terms of $f(t)$ reflect the use of the technique for approximation; on the other hand, the pointwise estimate of error in terms of $f_n(t)$ is especially useful for the inversion problem since then $f(t)$ is not available.

It is also shown that the successive approximations $f_0(t), f_1(t), f_2(t), \ldots$ are uniformly better for each t if $\ddot{f}(t) \geq 0$.

In practical use the initial member of the sequence, $f_0(t)$, is not an accurate approximation to $f(t)$ for t not in the neighborhoods of zero and infinity. Additionally the sequence $\{f_n(t)\}$ does not converge rapidly in n. Consequently one must go far out in the sequence to obtain adequate accuracy. Part V treats this problem. A modification, $f_{n,\alpha}(t)$, of $f_n(t)$ is introduced depending on a parameter α for which, by appropriate choice of α, $f_{0,\alpha}(t)$ is a much improved approximation to $f(t)$ than $f_0(t)$; the rapidity of convergence of the sequence $\{f_{n,\alpha}(t)\}$ is not improved over that of the unmodified sequence. However, it has been found that good accuracy is obtained by use of $f_{0,\alpha}(t)$ or $f_{1,\alpha}(t)$ as is demonstrated in the examples on covariance and recovery functions given in this paper.

In many applications, especially to complementary distribution functions, the behavior of $f(t)$ for large t must be accurately reproduced. The approximations $f_{n,\alpha}(t)$ accomplish this especially when α is related to the decrement of an exponential majorant. For functions which are exponentially small at infinity, the $f_n(t)$ do not adequately reproduce the decay of $f(t)$.

Many of the desirable features of the original method are still retained by this modification. The concept of convexity with exponent α is introduced which allows the transference of the inequality $f_n(t) \geq f(t)$ to $f_{n,\alpha}(t) \geq f(t)$. A criterion is given for deciding convexity with exponent α in terms of the transform, $\tilde{f}(s)$.

The degree of precision concept is applied to the approximation sequence in order to obtain a modified sequence, $s_n(t)$, which converges more rapidly. For sufficiently smooth functions this method is successful. The approximation $s_n(t)$ consists of a linear combination of $f_0(t), \ldots, f_n(t)$ or of $f_{0,\alpha}(t), \ldots, f_{n,\alpha}(t)$ and hence is easily applied. Its efficacy is demonstrated in the examples of this paper. Unfortunately the improvement in rapidity of convergence is so strong that the map from $f(t)$ to $s_n(t)$ is no longer positive, consequently many of the desirable structural preservation properties of the L_n are lost in favor of greater numerical accuracy.

An attempt is made to enhance the rapidity of convergence of $\{f_n(t)\}$ while simultaneously retaining the positivity of the map. This is accomplished by the construction of a new sequence, $h_n(t)$, which is also a linear combination of $f_0(t), \ldots, f_n(t)$. As is to be expected, however, the improvement is not as great as is realized with the sequence $s_n(t)$.

The pointwise error estimate developed in Part IV may be used as a correction device on $f_n(t)$ or $f_{n,\alpha}(t)$ to improve further the accuracy of computation. This, however, in the absence of an error estimate for the modification, must rely on one's understanding of the specific problem for ascertaining the reasonableness of the result.

An application of the methods of this paper to the renewal function[9] in the theory of renewal processes is made in Part VI. The remarkable accuracy of the simplest of the approximations $f_0(t)$, $f_1(t)$ is noteworthy.

Part VII presents applications of the techniques to the covariance and recovery functions of Erlang blocking models used in telephone traffic theory.[11] For the covariance function, the initial approximation, $f_{0,\alpha}(t)$, is excellent; however, in the case of the recovery function it was found that $f_{0,\alpha}(t)$, $f_{1,\alpha}(t)$ might not be considered sufficiently accurate, accordingly the linear combination, $s_1(t)$, was used. This provided sufficient enhancement of accuracy.

The generating function, $G(z,t)$, for the $f_n(t)$ can sometimes be used to obtain an explicit construction of the sequence. Some examples of this nature are treated in Part VIII.

Applications of the methods herein have been made to the complementary distributions of waiting time in $M/G/1$ queues. Also B. W. Stuck and E. Arthurs have successfully applied these techniques to the study of models of computer systems.

There are questions of an exclusively mathematical character which have not been touched upon, e.g., a semigroup interpretation and saturation phenomena. It is felt that these would be outside the essentially practical thrust of the paper. For some theorems which are applicable to the operators of this paper see Ref. 5.

A short table of operations on $f(t)$ and their corresponding maps under L_n is included to facilitate application of these methods to the construction of approximations.

II. WIDDER INVERSION—REPRESENTATIONS

Let the transform $\tilde{f}(s)$,

$$\tilde{f}(s) = \int_0^\infty e^{-su} f(u)du \tag{1}$$

exist for $s > 0$, then

$$\frac{(-1)^n}{n!} s^{n+1} \tilde{f}^{(n)}(s) = \frac{s^{n+1}}{n!} \int_0^\infty e^{-su} u^n f(u)du \tag{2}$$

in which

$$\tilde{f}^{(n)}(s) = \frac{d^n}{ds^n} \tilde{f}(s). \tag{3}$$

The function $(s^{n+1}/n!)e^{-su}u^n$ is a probability density function on $(0, \infty)$ for $s > 0, n \geq 0$ whose mean is $(n+1)/s$ and variance $(n+1)/s^2$. When

$s = (n+1)/t$, the mean and variance are t and $t^2/(n+1)$ respectively. One has:[1]

Theorem 1 (Widder). Let the transform, $\bar{f}(s)$, of $f(t)$ exist for $s > 0$, let $f(t)$ be continuous at t and bounded on $[0, \infty]$, then

$$\lim_{n\to\infty} \frac{(-1)^n}{n!} s^{n+1} \bar{f}^{(n)}(s)\Big|_{s=(n+1)/t} = f(t).$$

The convergence is uniform in every finite closed interval throughout which $f(t)$ is continuous.

Proof. Korovkin's theorem on sequences of positive functionals.[2]

The inversion theorem, in the above form, had already been stated by Feller[3] who used the law of large numbers to effect the proof. It is the purpose of this paper to study the transformation

$$L_n f = f_n = \frac{(-1)^n}{n!} s^{n+1} \bar{f}^{(n)}(s)\Big|_{s=(n+1)/t} \tag{4}$$

so that f_n may be effectively used as an approximation to f. The representation of f_n directly in terms of f is obtainable from (2); thus,

$$f_n(t) = \int_0^\infty g_n(t, u) f(u) du \tag{5}$$

$$g_n(t, u) = \frac{(n+1)^{n+1}}{n! t^{n+1}} e^{-[(n+1)/t]u} u^n \qquad n \geq 0. \tag{6}$$

Alternative forms which will be found useful are:

$$f_n(t) = \frac{n+1}{t} \int_0^\infty \psi\left(n, (n+1)\frac{u}{t}\right) f(u) du \tag{7}$$

$$f_n(t) = (n+1) \int_0^\infty \psi(n, (n+1)u) f(tu) du \tag{8}$$

$$\psi(x, a) = e^{-a} \frac{a^x}{\Gamma(x+1)} \tag{9}$$

in which $\psi(x, a)$ is the Poisson probability distribution,

$$f_n(t) = \int_0^\infty M_n\left(\frac{t}{u}\right) f(u) \frac{du}{u} \tag{10}$$

$$M_n(x) = \frac{(n+1)^{n+1}}{n!} e^{-(n+1)/x} x^{-n-1} \tag{11}$$

in which the representation is by means of convolution on the half line

(Mellin convolution), and

$$g_n(\eta) = \int_{-\infty}^{\infty} K_n(\eta) g(\eta - \xi) d\xi \tag{12}$$

$$K_n(\eta) = \frac{(n+1)^{n+1}}{n!} e^{-(n+1)\eta - (n+1)e^{-\eta}} \tag{13}$$

$$t = e^{\eta},\ u = e^{\xi},\ f(t) = g(\eta),\ f_n(t) = g_n(\eta) \tag{14}$$

in which the representation is by means of convolution on the whole real line (Fourier convolution).

The conditions of Theorem 1 are relaxed below.

Theorem 2. Let the transform, $\bar{f}(s)$, of $f(t)$ exist for $s > c$, let $f(t)$ be continuous at t and let $f(t) = 0(e^{ct})(t \to \infty)$; then,

$$\lim_{n \to \infty} f_n(t) = f(t).$$

The convergence is uniform in every finite closed interval throughout which $f(t)$ is continuous.

Proof. The representation (5) may be written as follows:

$$f_n(t) = \frac{(n+1)^{n+1}}{n! t^{n+1}} \int_0^{\infty} e^{-[(n-m+1)/t]u} u^n e^{-m/tu} f(u) du. \tag{15}$$

For all t in some finite closed interval, m may be chosen so that

$$e^{-(m/t)u} f(u) = 0(1)(u \to \infty)$$

hence Korovkin's theorem is again applicable and the conclusions follow.

III. PROPERTIES OF $f_n(t)$

Jensen's theorem applied to (5) proves

Theorem 3. $f(t)$ is convex on $(0, \infty)$

$$\Rightarrow f(t) \leq f_n(t),\ t \geq 0,\ n \geq 0.$$

The value of an approximation method is greatly enhanced when the approximating function preserves the shape of the original and coincides closely with its behavior in the neighborhoods of zero and infinity. Theorems 4, 5, 6 establish the desired properties.

Theorem 4. $a \leq f(t) \leq b \Rightarrow a \leq f_n(t) \leq b$; a, b arbitrary real.

Proof. Direct evaluation shows that

$$L_n f = f,\ f = \alpha + \beta t. \tag{16}$$

The positivity of L_n implies

$$f \leq b \Rightarrow L_n f \leq L_n b = bL_n 1 = b. \tag{17}$$

Similarly for the lower bound.

The derivatives of $f_n(t)$ may be related to those of $f(t)$ through use of (8); thus

Theorem 5. Let $f^{(r)}(t)$ be continuous and $0(e^{ct})(t \to \infty)$, then there is an m so that $f_n^{(r)}(t)$ exists and is continuous for $n \geq m$ and

$$f_n^{(r)}(t) = (n+1) \int_0^\infty \psi(n, (n+1)u) u^r f^{(r)}(tu) du.$$

One may set $m = 0$ if $f^{(r)}(t) = 0(1)(t \to \infty)$.

Proof. For m sufficiently large, the integral of the theorem converges uniformly in t; hence the representation (8) may be differentiated under the integral sign r times. If $f^{(r)}(t)$ is bounded then $m = 0$ is a permissible choice since one still has uniform convergence.

Corollary 1. $f^{(r)} \geq 0 \Rightarrow f_n^{(r)} \geq 0, n \geq m.$

Proof. This follows from the positivity of the kernel.

The above corollary implies that if f has a continuous derivative and is monotone then f_n is monotone, and if f has a continuous second derivative and is convex then f_n is convex. A stronger structural result will be obtained in Theorem 6. One also has that if f is completely or absolutely monotonic then f_n is completely or absolutely monotonic respectively.

Corollary 2. $f_n^{(r)}(0+) = \lambda_{n,r} f^{(r)}(0+), n \geq m,$

$$\lambda_{n,r} = \frac{\Gamma(n+r+1)}{n!(n+1)^r}.$$

In particular

$$f_n(0+) = f(0+) \qquad n \geq m$$
$$\dot{f}_n(0+) = \dot{f}(0+) \qquad n \geq m.$$

Proof. Define $\lambda_{n,r}$ by

$$\lambda_{n,r} = (n+1) \int_0^\infty \psi(n, (n+1)u) u^r du$$

then evaluation of the integral yields the formula stated. Since the operator is bounded, the limit statements follow. Also one has $\lambda_{n,0} = \lambda_{n,1} = 1$.

Corollary 3. Let $f^{(r)}(\infty) < \infty$; then

$$f_n^{(r)}(\infty) = \lambda_{n,r} f^{(r)}(\infty) \qquad n \geq 0$$

In particular

$$f_n(\infty) = f(\infty) \qquad n \geq 0,$$
$$\dot{f}_n(\infty) = \dot{f}(\infty) \qquad n \geq 0.$$

Proof. Dominated convergence allows the interchange of limit and integral.

The following concepts will be needed to establish further structural properties.

For an arbitrary sequence in $(-\infty, \infty)$, $-\infty < t_1, t_2, \ldots, t_\ell < \infty$, the number of changes of sign is called the variation of the sequence and will be indicated by $v(t_1, t_2, \ldots, t_\ell)$; thus

$$v(3, -1, 0, 2, -2) = 3 \tag{18}$$

$$v(1, 2, 4, 6) = 0. \tag{19}$$

One sets $v(0, 0, \ldots, 0) = -1$. Let $f(t)$ be defined on $(0, \infty)$, and let $0 < t_1 < t_2 < \ldots < t_\ell < \infty$ be an arbitrary, ordered sequence in $(0, \infty)$, the quantity $\sup v(f(t_1), f(t_2), \ldots, f(t_\ell))$, in which the supremum is taken over all sequences, i.e., for all choices of $(t_1, t_2, \ldots, t_\ell)$ and for all $\ell \geq 1$, is called the variation of f and will be indicated by $v(f)$. A transformation L on a given class of f will be called variation diminishing if and only if $v(Lf) \leq v(f)$ for every f in its domain. The definition used here is adopted from Hirschman and Widder.[4]

Let $\phi(\eta)$ be a frequency function on $(-\infty, \infty)$, that is,

$$\phi(\eta) \geq 0, \int_{-\infty}^{\infty} \phi(\eta) d\eta = 1 \tag{20}$$

and let

$$\tilde{\phi}(s) = \int_{-\infty}^{\infty} e^{-s\eta} \phi(\eta) d\eta. \tag{21}$$

Define $E(s)$ by

$$E(s) = \tilde{\phi}(s)^{-1}. \tag{22}$$

Then a theorem of Schoenberg[4] states that the transformation

$$Tg = \int_{-\infty}^{\infty} \phi(\xi) g(\eta - \xi) d\xi \qquad g \epsilon BC(-\infty, \infty) \tag{23}$$

is variation diminishing if and only if

$$E(s) = e^{-Cs^2+bs} \prod_k \left(1 - \frac{s}{a_k}\right) e^{s/a_k} \tag{24}$$

$$C \geq 0,\ b,\ a_k \text{ real},\ \sum_k 1/a_k^2 < \infty.$$

The designation $g \epsilon BC(-\infty, \infty)$ means that $g(\eta)$ is bounded and continuous on $(-\infty, \infty)$. It may be observed that the mean of ϕ is b and the variance

$$2C + \sum_k 1/a_k^2$$

The Laplace transform of $K_n(\eta)$ (13),

$$\bar{K}_n(s) = \int_{-\infty}^{\infty} e^{-s\eta} K_n(\eta) d\eta \tag{25}$$

is

$$\bar{K}_n(s) = \frac{\Gamma(n+s+1)}{n!(n+1)^s}. \tag{26}$$

This may be written in the following forms

$$\bar{K}_n(s) = \frac{\Gamma(1+s)}{(n+1)^s} \prod_{k=1}^{n} \left(1 + \frac{s}{k}\right) \tag{27}$$

$$\bar{K}_n(s)^{-1} = e^{s\nu_n} \prod_{k=n+1}^{\infty} \left(1 + \frac{s}{k}\right) e^{-s/k} \tag{28}$$

$$\nu_n = ln\ (n+1) + \gamma - \sum_{j=1}^{n} \frac{1}{j} \tag{29}$$

in which $\gamma = 0.5772157$ is Euler's constant. Thus by Schoenberg's theorem, the transformation $T_n g = g_n$, $g \epsilon BC(-\infty, \infty)$, defined in (12) is variation diminishing. The mean and variance of K_n are respectively ν_n as given above, and σ_n^2 given by

$$\sigma_n^2 = \frac{\pi^2}{6} - \sum_{j=1}^{n} \frac{1}{j^2}. \tag{30}$$

Since the map $t = e^{\eta}$ is monotone, the following theorem has been established.

Theorem 6. The transformation L_n defined on $f \epsilon BC(0, \infty)$ is variation diminishing, i.e.,

$$v(f_n) \leq v(f).$$

Corollary. f_n does not cross any straight line more often than f. In par-

ticular, if f is monotone then f_n is monotone, and if f or $-f$ is convex then f_n or $-f_n$ is convex.

Proof. From (16) and Theorem 6,

$$v(L_n(f - \alpha - \beta t)) = v(L_n f - \alpha - \beta t) \leq v(f - \alpha - \beta t). \tag{31}$$

Clearly f is monotone $\Leftrightarrow v(f - \alpha) \leq 1$ for arbitrary α, and f or $-f$ is convex $\Leftrightarrow v(f - \alpha - \beta t) \leq 2$ for arbitrary α, β.

It is clear from (8) that the approximating sequence to $f(at)(a > 0)$ is $f_n(at)$; this may be expressed in a more illuminating way as follows. Define the operator A by

$$Af(t) = f(at) \qquad a > 0 \tag{32}$$

then A and L_n commute; thus

$$L_n A f = A L_n f \tag{33}$$

Hence the eigenfunctions of A, which are t^r, are also the eigenfunctions of L_n. In fact one easily obtains

$$L_n t^r = \lambda_{n,r} t^r, \; r \geq 0, \; n \geq 0. \tag{34}$$

It may be observed that if L_n is defined by (8) instead of by (4) then (34) remains valid even for $r < 0$ provided n is large enough.

Other operations with the same eigenfunctions will also commute with L_n. Of importance in discussing the convergence of $f_n^{(r)}(t)$ is the operator

$$\theta \equiv t \frac{d}{dt} \tag{35}$$

One has

Theorem 7. The operators $L_n (n \geq 0), \theta$ commute; thus

$$L_n(\theta^r f) = \theta^r f_n.$$

It is assumed that the rth derivative of f exists and is continuous on $(0, \infty)$.

Proof. It is observed that the eigenfunctions of θ are t^r; alternatively the result follows directly from (8).

Corollary. Proceeding inductively, one can now establish that if $f^{(r)}$ is continuous and $0(e^{ct})(t \to \infty)$ then

$$\lim_{n \to \infty} f_n^{(r)} = f^{(r)}$$

In addition to shape preserving properties, another way of assessing

the adequacy of an approximation process is by comparison of moments; the rth moment of $f(t)$ is here taken to be $\int_0^\infty t^r f(t)dt$. The Mellin transform, $\bar{f}(s)$, given by

$$\bar{f}(s) = \int_0^\infty t^{s-1} f(t)dt \tag{36}$$

is the appropriate tool. Since the transform of $M_n(s)$, eq. (11), is

$$\bar{M}_n(s) = \frac{(n+1)^s \Gamma(n-s+1)}{n!} \tag{37}$$

one has from (10)

$$\bar{f}_n(s) = \frac{(n+1)^s \Gamma(n-s+1)}{n!} \bar{f}(s). \tag{38}$$

For the following, it is convenient to use the factorial symbol

$$n^{(0)} = 1, n^{(r)} = n(n-1)\ldots(n-r+1), r > 0 \text{ (integral)}. \tag{39}$$

The following theorem may now be stated.

Theorem 8. Let the rth moment of $f(t)$ exist ($r \geq 0$, integral); then the rth moment of $f_n(t)$ exists for $n > r$ and one has

$$\int_0^\infty t^r f_n(t)dt = \frac{(n+1)^{r+1}}{n^{(r+1)}} \int_0^\infty t^r f(t)dt.$$

Special cases are

$$\int_0^\infty f_n(t)dt = \frac{n+1}{n} \int_0^\infty f(t)dt \qquad n \geq 1 \tag{40}$$

$$\int_0^\infty t f_n(t)dt = \frac{(n+1)^2}{n(n-1)} \int_0^\infty t f(t)dt \qquad n \geq 2. \tag{41}$$

When $\int_0^\infty t^{-1} f(t)dt$ exists, an interesting special case of (38) occurs for $s = 0$; thus

$$\int_0^\infty t^{-1} f_n(t)dt = \int_0^\infty t^{-1} f(t)dt. \tag{42}$$

Formulae (40), (41) may be used to ensure equality of moments. Thus if it is required that the zeroth order moments agree, then, according to (40), one may use as the approximating sequence $nf_n(t)/(n+1)$. If the zeroth and first moments are to agree simultaneously, one may use a linear combination of $f_n(t)$ and $f_m(t)$; for example, $\frac{3}{2}f_3(t) - \frac{2}{3}f_2(t)$.

Another set of moment relations may be obtained from (7) involving sums of $f_n((n+1)t)$. These are given in

Theorem 9. Let the rth absolute moment of $f(t)$ exist; then

$$\sum_{n=0}^{\infty} n^{(r)} f_n((n+1)t) = t^{-r-1} \int_0^{\infty} u^r f(u)du.$$

Proof. One has

$$\sum_{n=0}^{\infty} n^{(r)} \psi(n, a) = a^r \tag{43}$$

and, from (7),

$$f_n((n+1)t) = t^{-1} \int_0^{\infty} \psi\left(n, \frac{u}{t}\right) f(u)du. \tag{44}$$

Multiplication of both sides of (44) by $n^{(r)}$ and summing, the result follows on interchanging summation and integration. Dominated convergence justifies the interchange.

Another property of $f_n(t)$ as a function of n is given in:

Theorem 10. Let $f(t) \geq 0$ on $(0, \infty)$, $0(e^{ct})(t \to \infty)$, then there is an m so that

$$f_{n+1}(t) \leq e^{-1} \left(\frac{n+2}{n+1}\right)^{n+2} f_n(t), \; t \geq 0, n \geq m.$$

Proof. Since

$$(n+2)\psi(n+1, (n+2)u) = ue^{-u} \left(\frac{n+2}{n+1}\right)^{n+2} (n+1)\psi(n, (n+1)u) \tag{45}$$

one may write

$$f_{n+1}(t) = (n+1) \int_0^{\infty} \psi(n, (n+1)u) ue^{-u} \left(\frac{n+2}{n+1}\right)^{n+2} f(tu)du. \tag{46}$$

Observing that $ue^{-u} \leq e^{-1}$, the inequality follows.

Corollary. $f(t) \geq 0 \Rightarrow \dfrac{f_n(t)}{n+1}$ is monotone decreasing in n for all $t \geq 0$, $n \geq m$.

Proof. One has

$$\frac{f_{n+1}(t)}{n+2} \leq \frac{f_n(t)}{n+1} e^{-1} \left(\frac{n+2}{n+1}\right)^{n+1}. \tag{47}$$

The result follows since

$$e^{-1}\left(\frac{n+2}{n+1}\right)^{n+1} \leq 1. \tag{48}$$

A stronger monotonicity property of $f_n(t)$ is stated in Theorem 21.

A useful recurrence relation allowing one to compute the members of the sequence $f_n(t)$ successively starting with $f_0(t)$ is given in the following theorem.

Theorem 11.

$$f_{n+1}(t) = f_n\left(\frac{n+1}{n+2}t\right) + \frac{t}{n+2}\dot{f}_n\left(\frac{n+1}{n+2}t\right), n \geq 0, t \geq 0.$$

Proof. Define $\hat{f}_n(s)$ by

$$\hat{f}_n(s) = \frac{(-1)^n}{n!}s^{n+1}\tilde{f}^{(n)}(s) \tag{49}$$

then, by (4),

$$f_n(t) = \hat{f}_n(s)\Big|_{s=(n+1)/t} \tag{50}$$

One has

$$s\frac{d}{ds}\hat{f}_n(s) = (n+1)\hat{f}_n(s) - (n+1)\hat{f}_{n+1}(s) \tag{51}$$

$$\hat{f}_{n+1}(s) = \hat{f}_n(s) - \frac{1}{n+1}s\frac{d}{ds}\hat{f}_n(s). \tag{52}$$

Thus

$$f_{n+1}(t) = \left[\hat{f}_n(s) - \frac{1}{n+1}s\frac{d}{ds}\hat{f}_n(s)\right]_{s=(n+2)/t} \tag{53}$$

The recurrence relation is now obtained on performing the substitution for s.

A useful alternative method of presenting the structure of the entire sequence $\{f_n\}_0^\infty$ in terms of f_0 is by means of a generating function. This is given in the theorem below.

Theorem 12. $f(t)$ is bounded on $(0, \infty) \Rightarrow$

$$\sum_{n=0}^{\infty} z^n f_n((n+1)t) = \frac{1}{1-z}f_0\left(\frac{t}{1-z}\right).$$

The series is convergent for $|z| < 1$ and analytically continuable in the half plane $\mathrm{Re}\, z < 1$.

Proof. The formula follows from (7) after interchange of summation and integration. The series is clearly convergent for $|z| < 1$ while dominated convergence justifies the interchange for Re $z < 1$.

The case $r = 0$ of Theorem 9 provides the following corollary.

Corollary.

$$f(t)\epsilon L(0, \infty) \Rightarrow \lim_{z \to 1-} \frac{1}{1-z} f_0\left(\frac{t}{1-z}\right) = t^{-1} \int_0^\infty f(u)du.$$

The Mellin transform, $\bar{f}(s)$, of $f(t)$ may be directly obtained from its Laplace transform, $\tilde{f}(s)$, by use, for example, of (38) for $n = 0$; thus

$$\bar{f}(s) = \frac{\tilde{f}_0(s)}{\Gamma(1-s)}. \tag{54}$$

Accordingly one may now write (38) in the form

$$\bar{f}_n(s) = \frac{(n+1)^s \Gamma(n+1-s)}{n!\Gamma(1-s)} \bar{f}_0(s) \tag{55}$$

or, equivalently,

$$\bar{f}_n(s) = (n+1)^s \binom{n-s}{n} \bar{f}_0(s). \tag{56}$$

At times (55) or (56) provides a convenient alternative to Theorems 11 and 12 when $f_n(t)$ is required as a function of n.

The range of applicability of the Jensen inequality of Theorem 3 may be extended by use of Mellin or Fourier convolution. A sequence $\{f_n(t)\}_{n=0}^\infty$ will be called an approximation sequence if there is an $f(t)$ so that $f_n(t) = L_n f(t)$. Let $*$ designate Mellin convolution; then

Theorem 13. $f_n * g$ is the approximation sequence for $f * g$.

Proof. One has

$$\overline{L_n(f*g)} = \bar{M}_n \bar{f}\bar{g} = \overline{L_n f g} \tag{57}$$

thus,

$$L_n(f*g) = (L_n f)*g = f_n * g. \tag{58}$$

The converse of Theorem 3 is also true.

Theorem 14. $f_n(t) \geq f(t)$ for all $n \geq 0, t \geq 0, f(t)$ is bounded on $(0, \infty) \Rightarrow$ $f(t)$ is convex on $(0, \infty)$.

Proof. The result follows from Theorem 8 of Karlin and Ziegler.[5]

Corollary. f convex on $(0, \infty), g \geq 0$ on $(0, \infty) \Rightarrow f * g$ convex on $(0, \infty)$.

Proof. One has from Theorem 3

$$f_n \geq f \tag{59}$$

and, since $g \geq 0$,

$$f_n * g \geq f * g. \tag{60}$$

Since, by Theorem 13, $f_n * g$ is the approximation sequence to $f * g$, application of Theorem 14 proves the corollary.

It may be observed that the inequality of (60) remains valid when $*$ is interpreted as Fourier convolution although, in this case, $f_n * g$ is not the approximation sequence for $f * g$.

Another set of convexity results may be obtained from (8) by considering logarithmic convexity.

Theorem 15. If $f(t)$ is log-convex on $t \geq 0$, then $f_n(t)$ is log-convex for $n \geq 0$, $t \geq 0$.

Proof. Equation (8) and the additivity of log-convex functions.[6]

Further one may state the following inequalities.

Theorem 16. If $f(t)$ is log-convex on $t \geq 0$ then

$$f(t) \leq e^{L_n \ell n f(t)} \leq f_n(t).$$

Proof. The inequality on the left follows from Theorem 3 applied to $\ell n f(t)$; the one on the right is a consequence of the geometric mean-arithmetic mean inequality.

IV. ERROR ESTIMATION

Error estimates take different forms depending on the class of functions for which they are intended and whether or not they are bounds or pointwise estimates. From a practical point of view the pointwise estimate is the most useful provided it may be easily evaluated in terms of the approximation itself. The next three theorems provide error bounds for different function classes; the fourth theorem provides an approximate formula for the pointwise evaluation of error, while (112) does the same but in terms of $f_n(t)$. The error of approximation, $\epsilon_n(t;f)$, is defined by

$$\epsilon_n(t;f) = f_n(t) - f(t) \tag{61}$$

Theorem 17. Let $\dot{f}(t)$ be continuous on $(0, \infty)$; then

$$|\epsilon_n(t;f)| \leq \frac{t}{\sqrt{n+1}} \sup_{t>0} |\dot{f}(t)|.$$

Proof. One has

$$\epsilon_n(t;f) = \int_0^\infty g_n(t, u)|f(u) - f(t)|du \tag{62}$$

$$|\epsilon_n(t;f)| \leq \int_0^\infty g_n(t, u)|f(u) - f(t)|du \tag{63}$$

$$|\epsilon_n(t;f)| \leq \sup_{t>0} |\dot{f}(t)| \cdot \int_0^\infty g_n(t, u)|u - t|du \tag{64}$$

$$|\epsilon_n(t;f)| \leq \sup_{t>0} |\dot{f}(t)| \cdot \left\{ \int_0^\infty g_n(t, u)(u - t)^2 du \right\}^{1/2} \tag{65}$$

$$|\epsilon_n(t;f)| \leq \frac{t}{\sqrt{n+1}} \sup_{t>0} |\dot{f}(t)|. \tag{66}$$

The last inequality follows because the mean and variance of $g_n(t, u)$ are t and $t^2/(n + 1)$ respectively.

Theorem 18. Let $\ddot{f}(t)$ be continuous on $(0, \infty)$; then

$$|\epsilon_n(t;f)| \leq \frac{t^2}{2n+2} \sup_{t>0} |\ddot{f}(t)|.$$

Proof. The Taylor expansion of $f(u)$ about t has the form

$$f(u) = f(t) + (t - u)\dot{f}(t) + \frac{1}{2}(t - u)^2 \ddot{f}(\xi) \tag{67}$$

in which ξ lies between t and u. Thus

$$\epsilon_n(t;f) = \frac{1}{2} \frac{t^2}{n+1} \ddot{f}(\xi), \ \xi \epsilon (0, \infty). \tag{68}$$

The inequality of the theorem now follows.

The next theorem provides an error bound which is uniform for $t \epsilon [0, \infty]$. For this purpose the absolute first moment of $K_n(\eta)$ (13), α_n, is needed; thus

$$\alpha_n = \int_{-\infty}^\infty K_n(\eta)|\eta|d\eta. \tag{69}$$

Theorem 19. Let $\dot{f}(t)$ be continuous on $(0, \infty)$; then

$$|\epsilon_n(t;f)| \leq \alpha_n \sup_{t>0} |t\dot{f}(t)|.$$

Proof. One has from (12)

$$\epsilon_n(e^\eta;f) = \int_{-\infty}^\infty K_n(\eta - \xi)|g(\xi) - g(\eta)|d\xi \tag{70}$$

$$|\epsilon_n(e^{\eta};f)| \leq \sup_{-\infty<\eta<\infty} |g'(\eta)| \cdot \int_{-\infty}^{\infty} K_n(\eta-\xi)|\eta-\xi|d\xi \tag{71}$$

$$|\epsilon_n(t;f)| \leq \alpha_n \sup_{t>0} |t\dot{f}(t)|.$$

Corollary. The convergence of $f_n(t)$ to $f(t)$ $(n \to \infty)$ is uniform for $t \epsilon [0, \infty]$.

Proof. It is necessary to show that

$$\lim_{n\to\infty} \alpha_n = 0.$$

The expression for α_n (69) is rewritten as follows

$$\alpha_n = \rho_n \int_{-\infty}^{\infty} K(\eta)^{n+1}|\eta|d\eta \tag{72}$$

$$\rho_n = \frac{1}{n!}\left(\frac{n+1}{e}\right)^{n+1}, K(\eta) = e^{1-\eta-e^{-\eta}}. \tag{73}$$

Use of the power series expansion for $e^{-\eta}$ yields

$$\alpha_n \sim \rho_n \int_{-\infty}^{\infty} e^{-[(n+1)/2]\eta^2}|\eta|d\eta = \frac{2}{en!}\left(\frac{n+1}{e}\right)^n. \tag{74}$$

Stirling's formula now shows that

$$\alpha_n \approx \sqrt{\frac{2}{\pi n}}. \tag{75}$$

Some numerical values of α_n are $\alpha_0 = 1.0160$, $\alpha_1 = 0.6388$, $\alpha_2 = 0.5006$, $\alpha_3 = 0.4247$, $\alpha_4 = 0.3751$, $\alpha_5 = 0.3396$, $\alpha_6 = 0.3126$, $\alpha_7 = 0.2911$; the asymptotic formula (75) is sufficiently accurate for $n > 7$.

To continue the study of $\epsilon_n(t;f)$, it is useful to obtain an explicit formula of Peano type, that is an integral transform of $\ddot{f}$.
Let

$$x_+ = x \qquad x \geq 0 \tag{76}$$

$$= 0 \qquad x \leq 0 \tag{77}$$

then the Taylor expansion of $f(t)$ with remainder is

$$f(t) = f(0) + \dot{f}(0)t + \int_0^{\infty} (t-v)_+\ddot{f}(v)dv. \tag{78}$$

From (5) and (16), one has

$$f_n(t) = f(0) + \dot{f}(0)t + \int_0^{\infty} \ddot{f}(v)dv \int_v^{\infty} g_n(t,u)(u-v)du. \tag{79}$$

Thus

$$\epsilon_n(t, f) = \int_0^\infty E_n(t, v)\ddot{f}(v)dv \tag{80}$$

$$E_n(t, v) = \int_v^\infty g_n(t, v)(u - v)du - (t - v)_+. \tag{81}$$

The kernel $E_n(t, v)$ (the Peano kernel for error representation) is, clearly,

$$E_n(t, v) = L_n(t - v)_+ - (t - v)_+. \tag{82}$$

The explicit evaluation of the kernel may be most simply carried out by means of (24) since the Laplace transform of $(t - v)_+$ is e^{-sv}/s^2. Let

$$S_n(x) = \sum_{j=0}^{n} \frac{x^j}{j!} \tag{83}$$

and

$$\ell_n(a) = e^{-(n+1)a}[S_n((n + 1)a) - aS_{n-1}((n + 1)a)] \tag{84}$$

then

$$L_n(t - v)_+ = t\ell_n(a) \qquad a = v/t \tag{85}$$

$$E_n(t, v) = t[\ell_n(a) - (1 - a)_+]. \tag{86}$$

In particular, one has

$$E_0(t, v) = te^{-v/t} - (t - v)_+ \tag{87}$$

$$E_1(t, v) = (t + v)e^{-2v/t} - (t - v)_+. \tag{88}$$

Since $(t - v)_+$ is a convex function of t for each v, (82) and Theorem 3 establish

$$E_n(t, v) \geq 0 \text{ for all } t \geq 0, v \geq 0. \tag{89}$$

The moments of the kernel, $E_n(t, v)$, may be obtained by substituting the functions $f(t) = t^r$ $(r \geq 2)$ into (80), and using (34) and (61) for evaluation of $\epsilon_n(t;t^r)$; the following is obtained:

$$\int_0^\infty v^r E_n(t, v)dv = \frac{\lambda_{n,r+2} - 1}{(r + 1)(r + 2)} t^{r+2} \qquad r \geq 0. \tag{90}$$

In particular

$$\int_0^\infty E_n(t, v)dv = \frac{t^2}{2} \frac{1}{n + 1} \tag{91}$$

$$\int_0^\infty vE_n(t, v)dv = \frac{t^3}{6} \frac{3n + 5}{(n + 1)^2}. \tag{92}$$

One may now obtain an approximate evaluation of $\epsilon_n(t;f)$.

Theorem 20. Let $\ddot{f}(t)$ be continuous on $(0, \infty)$ and $0(e^{ct})(t \to \infty)$; then there is an m so that

$$\epsilon_n(t;f) \approx \frac{t^2}{2n+2} \ddot{f}\left(t \frac{3n+5}{3n+3}\right), n \geq m;$$

also if $\ddot{f}(t)$ is convex, then the approximation is a lower bound.

Proof. The one point Gaussian quadrature formula for $\int_0^\infty E_n(t, v) f(v) dv$ is of the form $Af(\alpha)$ in which the constants, A, α, are determined by requiring the quadrature to be exact for all linear functions. Use of (91), (92) now yields the formula of the theorem. The inequality follows from the nonnegativity of $E_n(t, v)$ (89) and Jensen's inequality.

Since by the Corollary to Theorem 7, $\ddot{f}_n(t)$ approximates $\ddot{f}(t)$, in practice the required value of $\ddot{f}(t)$ is approximated either from the analytic form of $f_n(t)$ or numerically from a table or curve already computed for $f_n(t)$.

At this point another property of the sequence $\{f_n(t)\}_0^\infty$ can be proved.

Theorem 21. Let $\ddot{f}(t) \geq 0$, continuous on $(0, \infty)$, and $0(e^{ct})(t \to \infty)$; then there is an m so that

$$f_{n+1}(t) \leq f_n(t) \text{ for all } t \geq 0, n \geq m.$$

Proof. Clearly the monotonic decreasing character of $f_n(t)$ as a function of n will hold if $\epsilon_n(t;f)$ has the same property. The nonnegativity of $\ddot{f}(t)$ and (80) shows that the result is implied if $E_n(t, v)$ is monotonically decreasing in n; in turn, by (86), this will follow if $\ell_n(a)$ is monotonically decreasing in n for each $a \geq 0$. From (84), by direct calculation,

$$\frac{d}{da} \ell_n(a) = -e^{-(n+1)a} S_n((n+1)a) \tag{93}$$

$$\frac{d^2}{da^2} \ell_n(a) = \frac{(n+1)^{n+1}}{n!} a^n e^{-(n+1)a}. \tag{94}$$

Let

$$h_n(a) = \ell_{n-1}(a) - \ell_n(a) \qquad n \geq 1 \tag{95}$$

then, from (94),

$$\frac{d^2}{da^2} h_n(a) = r_n(a) \frac{d^2}{da^2} \ell_{n-1}(a) \tag{96}$$

$$r_n(a) = 1 - \left(1 + \frac{1}{n}\right)^{n+1} a e^{-a}. \tag{97}$$

It is clear from (94) that the sign of

$$\frac{d^2}{da^2} h_n(a)$$

is the same as that of $r_n(a)$. There exist two points $0 < a_0(n) < a_1(n)$ with the following properties:

$$r_n(a) \geq 0 \qquad 0 \leq a \leq a_0(n) \tag{98}$$

$$r_n(a) < 0 \qquad a_0(n) < a < a_1(n) \tag{99}$$

$$r_n(a) \geq 0 \qquad a \geq a_1(n). \tag{100}$$

Since

$$\ell_n(0) = 1, \frac{d}{da} \ell_n(0) = -1, \quad n \geq 0 \tag{101}$$

it follows that

$$h_n(0) = 0, \frac{d}{da} h_n(0) = 0, \quad n \geq 1. \tag{102}$$

One has the following integral representations for $h_n(a)$:

$$h_n(a) = \int_0^a db \int_0^b r_n(c) \frac{d^2}{dc^2} \ell_{n-1}(c) dc, \tag{103}$$

$$h_n(a) = \int_a^\infty db \int_b^\infty r_n(c) \frac{d^2}{dc^2} \ell_{n-1}(c) dc. \tag{104}$$

Thus (98) and (103) imply

$$h_n(a) \geq 0 \qquad 0 \leq a \leq a_0(n); \tag{105}$$

similarly (100) and (104) imply

$$h_n(a) \geq 0 \qquad a \geq a_1(n). \tag{106}$$

The function $h_n(a)$ cannot be negative in $(a_0(n), a_1(n))$ since then it would have at least one local minimum; however, (99) shows that in $(a_0(n), a_1(n))$

$$\frac{d^2}{da^2} h_n(a) < 0$$

which is a contradiction. Thus

$$h_n(a) \geq 0, a \geq 0, n \geq 1 \tag{107}$$

and the theorem is proved.

It is possible to estimate conveniently $\epsilon_n(t;f)$ directly from $f_n(t)$ if $\dot{f}_n(t)$ and $\ddot{f}_n(t)$ are readily obtainable, at least possibly numerically from values of $f_n(t)$. From (38) one has

$$\bar{f}(s) = \overline{M}_n(s)^{-1}\bar{f}_n(s). \tag{108}$$

Expansion of $\Gamma(n+1-s)$ into a power series in s and substitution into (37) provides the following series

$$\overline{M}_n(s) = 1 + \nu_n s + \frac{\sigma_n^2 + \nu_n^2}{2} s^2 + \dots \tag{109}$$

$$\overline{M}_n(s)^{-1} = 1 - \nu_n s + \frac{\nu_n^2 - \sigma_n^2}{2} s^2 + \dots \tag{110}$$

Thus

$$\bar{\epsilon}_n(s;f) = \left[\nu_n s + \frac{\sigma_n^2 - \nu_n^2}{2} s^2 + \dots \right] \bar{f}_n \tag{111}$$

$$\epsilon_n(t;f) \approx -\nu_n \theta f_n(t) + \frac{\sigma_n^2 - \nu_n^2}{2} \theta^2 f_n(t). \tag{112}$$

To facilitate the use of (112) some values of the coefficients are given in Table I.

The following readily obtained asymptotic formulas may be used for values of n beyond the table:

$$\begin{aligned} \nu_n &\approx \frac{1}{2(n+1)} + \frac{1}{12(n+1)^2} \\ \sigma_n^2 &\approx \frac{1}{n+1} + \frac{1}{2(n+1)^2} \\ \frac{\sigma_n^2 - \nu_n^2}{2} &\approx \frac{1}{2(n+1)} + \frac{1}{8(n+1)^2}. \end{aligned} \tag{113}$$

V. ENHANCEMENT OF ACCURACY

The excellent behavior of the operator L_n in constructing approximations to a given $f(t)$ which preserve its structural properties and its limiting values and which provide inequalities exacts a penalty in the form of slow convergence. A high value of n is required to attain high

Table I — Coefficients

n	ν_n	σ_n^2	$(\sigma_n^2 - \nu_n^2)/2$
0	0.5772	1.6449	0.6559
1	0.2704	0.6449	0.2859
2	0.1758	0.3949	0.1820
3	0.1302	0.2838	0.1334
4	0.1033	0.2213	0.1053
5	0.0856	0.1813	0.0870
6	0.0731	0.1535	0.0741
7	0.0638	0.1331	0.0645
8	0.0566	0.1175	0.0572

numerical accuracy. In many practical problems, fortunately, very high accuracy is not needed; notwithstanding, the value of n required may still be inconveniently high. Considering that one starts with the Laplace transform, $\tilde{f}(s)$, of $f(t)$ and uses (4), or constructs $f_0(t)$ and uses the recursion of Theorem 11, a high value of n implies obtaining a correspondingly high order of derivative of $\tilde{f}(s)$ or of $f_0(t)$ which can be a time-consuming operation. Thus it would be useful to modify the basic approximation, $L_n f$, while still preserving many of its original characteristics so that the accuracy for a given value of n may be increased.

In many cases the transform, $\tilde{f}(s)$, has the property that for some $\alpha > 0$, $\tilde{f}(s-\alpha)$ converges for $s > 0$. This property is used to construct a new approximation, $f_{n,\alpha}(t)$, defined by

$$f_{n,\alpha}(t) = e^{-\alpha t} L_n(e^{\alpha t} f(t)) \tag{114}$$

and, correspondingly, a new operator

$$L_{n,\alpha} f = f_{n,\alpha}. \tag{115}$$

The following theorem permits the construction of $f_{n,\alpha}(t)$ directly from $f_n(t)$.

Theorem 22.

$$f_{n,\alpha}(t) = \frac{e^{-\alpha t}}{\left(1 - \dfrac{\alpha t}{n+1}\right)^{n+1}} f_n\left(\frac{t}{1 - \dfrac{\alpha t}{n+1}}\right).$$

Proof. From (5) and (6), one has

$$L_n f = \frac{(n+1)^{n+1}}{n! t^{n+1}} \int_0^\infty e^{-[(n+1)/t]u} u^n f(u)\, du \tag{116}$$

$$L_n(e^{\alpha t} f) = \frac{(n+1)^{n+1}}{n! t^{n+1}} \int_0^\infty e^{-[(n+1)/t]u + \alpha u} u^n f(u)\, du. \tag{117}$$

Thus

$$L_n f \Big|_{\frac{t}{1-\alpha t/(n+1)}}$$

$$= \left(1 - \frac{\alpha t}{n+1}\right)^{n+1} \frac{(n+1)^{n+1}}{n! t^{n+1}} \int_0^\infty e^{-[(n+1)/t]u + \alpha u} u^n f(u)\, du. \tag{118}$$

Comparison of (118) with (117) shows that

$$L_n(e^{\alpha t} f) = \frac{1}{\left(1 - \dfrac{\alpha t}{n+1}\right)^{n+1}} f_n\left(\frac{t}{1 - \dfrac{\alpha t}{n+1}}\right) \tag{119}$$

hence, the result follows from (114).

The approximations $f_{n,\alpha}(t)$ satisfy theorems similar to those proved for $f_n(t)$; however, modifications are required. Only Theorem 3 will be discussed. A function $f(t)$ will be said to be convex over an interval with exponent α if and only if $e^{\alpha t}f(t)$ is convex over the same interval.

Theorem 23. If $f(t)$ is convex on $(0, \infty)$ with exponent α then

$$f(t) \leq f_{n,\alpha}(t).$$

Proof. One has from Theorem 3,

$$e^{\alpha t}f(t) \leq L_n(e^{\alpha t}f(t)). \tag{120}$$

The result now follows from the nonvanishing of $e^{\alpha t}$ and (114).

The error of approximation by $f_{n,\alpha}(t)$ will be designated $\epsilon_{n,\alpha}(t;f)$ and defined by

$$\epsilon_{n,\alpha}(t;f) = f_{n,\alpha}(t) - f(t). \tag{121}$$

Clearly

$$\epsilon_{n,\alpha}(t;f) = e^{-\alpha t}\epsilon_n(t;e^{\alpha t}f). \tag{122}$$

also, if the condition of Theorem 21 is satisfied, one has

$$\epsilon_{n,\alpha}(t;f) \geq 0. \tag{123}$$

One of the useful aspects of the approximation, $f_{n,\alpha}(t)$, is that it more accurately reflects the asymptotic behavior of $f(t)(t \to \infty)$ than $f_n(t)$ does for a given value of n. In the later applications this will be an important characteristic.

Clearly, ordinary convexity corresponds to convexity with exponent zero; however, the following theorem relates convexity with exponent α to log-convexity.

Theorem 24. Let $\ddot{f}(t)$ be continuous on some interval I; then $f(t)$ is log-convex on I if and only if it is convex with exponent α on I for all α.

Proof. One has

$$\ell n\ (e^{\alpha t}f(t)) = \alpha t + \ell nf(t) \tag{124}$$

hence $e^{\alpha t}f(t)$ is convex with exponent α on I for all α if $f(t)$ is log-convex on I. The derivative condition for convexity with exponent α on I is

$$\ddot{f}(t) + 2\alpha\dot{f}(t) + \alpha^2 f(t) \geq 0 \text{ on } I \tag{125}$$

and the derivative condition for log-convexity on I is

$$f(t)\ddot{f}(t) - \dot{f}(t)^2 \geq 0 \text{ on } I. \tag{126}$$

The choice

$$\alpha = -\dot{f}(t)/f(t) \tag{127}$$

which is always possible since $f(t) > 0$ on I, in (125) verifies (126).

Convexity with exponent α and, hence, by Theorem 24, log-convexity may be decided by means of the Laplace transform and the use of the Hausdorff-Bernstein theorem.[7]

Theorem 25. Let $\ddot{f}(t)$ be continuous on $(0, \infty)$, then $f(t)$ is convex with exponent α on $(0, \infty)$ if and only if

$$(s + \alpha)^2\tilde{f}(s) - (s + 2\alpha)f(0+) - \dot{f}(0+)$$

is completely monotonic in s on $(0, \infty)$ and is absolutely convergent on $s > 0$.

Proof. The expression cited is the Laplace transform of

$$e^{-\alpha t}\frac{d^2}{dt^2}(e^{\alpha t}f(t))$$

whose nonnegativity is the necessary and sufficient condition for convexity of $f(t)$ with exponent α. The Hausdorff-Bernstein theorem now completes the proof.

It may be observed that the quantities $f(0+)$, $\dot{f}(0+)$ are obtainable from

$$\lim_{s\to\infty} s\tilde{f}(s) = f(0+) \tag{128}$$

$$\lim_{s\to\infty} \{s^2\tilde{f}(s) - sf(0+)\} = \dot{f}(0+). \tag{129}$$

Another method of enhancement is related to the concept of "degree of precision." An approximation operator T, i.e., $Tf \approx f$, in which the functions t^r, suitably restricted to an appropriate interval ($r \geq 0$, integral), are in its domain, is said to have degree of precision k if $Tt^r = t^r$ for $0 \leq r \leq k$ and $Tt^r \neq t^r$ for $r = k + 1$. Thus the singular operators L_n studied here have degree of precision one.

The enhancement method consists of the following: coefficients $\delta_j (0 \leq j \leq k - 1)$ are determined by the moment conditions

$$\sum_{j=0}^{k-1} \delta_j L_j(t^r) = t^r \qquad 0 \leq r \leq k \tag{130}$$

and, accordingly, the linear combination

$$s_{k-1}(t) = \sum_{j=0}^{k-1} \delta_j f_j(t) \tag{131}$$

is now taken to approximate $f(t)$. Clearly the map from f to s_{k-1} has degree of precision k, however, unfortunately, it is not positive. If f is

sufficiently smooth there will result a significant improvement in accuracy over the use of f_{k-1} alone. The system (130) may be expressed in terms of $\lambda_{n,r}$ as follows

$$\sum_{j=0}^{k-1} \delta_j \lambda_{j,r} = 1 \qquad 1 \le r \le k. \tag{132}$$

Accordingly special cases of (131) are

$$s_1 = -f_0 + 2f_1 \tag{133}$$

$$s_2 = \frac{1}{2} f_0 - 4f_1 + \frac{9}{2} f_2. \tag{134}$$

A method of enhancement consisting of a linear combination of the $f_n(t)$ similar to (131) which, however, retains the positivity of the map will now be constructed. The accuracy attained will usually not be as great as that of (131) for a given set of values $\{f_j(t)\}_0^n$. The new sequence will be designated $h_n(t)$ and is defined by

$$h_n(t) = \sum_{j=0}^{n} \rho_j f_j(t). \tag{135}$$

Define $W_n(u)$ by

$$W_n(u) = \sum_{j=0}^{n} \rho_j (j+1) \psi(j, (j+1)u) \tag{136}$$

then the coefficients, ρ_j, are constrained by

$$\sum_{j=0}^{n} \rho_j = 1, \; W_n(u) \ge 0 \text{ for all } u \ge 0. \tag{137}$$

Theorem 26.

$$|h_n(t) - f(t)| \le t \sup_{t>0} |\dot{f}(t)| \left\{ \sum_{j=0}^{n} \frac{\rho_j}{j+1} \right\}^{1/2}.$$

Proof. From (8), one has

$$h_n(t) = \int_0^\infty W_n(u) f(tu) du. \tag{138}$$

Also, from (137),

$$\int_0^\infty W_n(u) du = 1 \tag{139}$$

hence

$$h_n(t) - f(t) = \int_0^\infty W_n(u)[f(tu) - f(t)] du. \tag{140}$$

The nonnegativity of $W_n(u)$ now permits the following inequality

$$|h_n(t) - f(t)| \leq \int_0^\infty W_n(u)|f(tu) - f(t)|du; \tag{141}$$

hence,

$$|h_n(t) - f(t)| \leq t \sup_{t>0} |\dot{f}(t)| \int_0^\infty W_n(u)|u - 1|du. \tag{142}$$

The Cauchy-Schwartz inequality applied to (142) yields

$$|h_n(t) - f(t)| \leq t \sup_{t>0} |\dot{f}(t)| \left\{ \int_0^\infty W_n(u)(u-1)^2 du \right\}^{1/2}. \tag{143}$$

Evaluation of the integral in (143) provides the inequality of the theorem.

Consider the sum, S, defined by

$$S = \sum_{j=0}^{n} \frac{\rho_j}{j+1} \tag{144}$$

then, in order to obtain the best approximation, the ρ_j must be chosen to minimize S besides satisfying the conditions of (137). One has

$$e^u W_n(u) = \sum_{j=0}^{n} \frac{(j+1)^{j+1}}{j!} \rho_j e^{-ju} u^j. \tag{145}$$

Let

$$z = e^{1-u} u \tag{146}$$

then the constraint of (137) may be written

$$\sum_{j=0}^{n} \frac{(j+1)^{j+1}}{j!} e^{-j} \rho_j z^j \geq 0 \qquad 0 \leq z \leq 1. \tag{147}$$

Define the polynomials $P(x)$ by (147) with $z = (x+1)/2$; then one has

$$P(x) = \sum_{j=0}^{n} x^j \left[\sum_{k=j}^{n} \frac{(k+1)^{k+1}}{k!} \binom{k}{j} (2e)^{-k} \rho_k \right] \tag{148}$$

and

$$P(x) \geq 0 \qquad -1 \leq x \leq 1. \tag{149}$$

The cosine polynomial $P(\cos\theta)$ is now obtained and written in the Fourier form

$$P(\cos\theta) = \frac{1}{2} a_0 + \sum_{j=1}^{n} a_j \cos j\theta \tag{150}$$

in which the a_j are obtained from

$$a_j = \frac{1}{\pi} \int_{-\pi}^{\pi} P(\cos\theta) \cos j\theta d\theta. \tag{151}$$

The nonnegativity of $P(\cos\theta)$ implies the following representation[8]

$$P(\cos\theta) = |h(\theta)|^2 \tag{152}$$

$$h(\theta) = x_0 + x_1 e^{i\theta} + \ldots + x_n e^{in\theta} \tag{153}$$

in which the coefficients $x_0, \ldots, x_n$ are real; thus

$$a_j = 2 \sum_{\nu=0}^{n-j} x_\nu x_{\nu+j} \qquad 0 \leq j \leq n. \tag{154}$$

When (154) is solved for the ρ_j in terms of $x_0, \ldots, x_n$, the problem of minimizing S subject to $\rho_0 + \ldots + \rho_n = 1$ becomes that of minimizing a quadratic form relative to another quadratic form.

The optimum ρ_j have been obtained for the case $n = 2$; the result is

$$h_2(t) = 0.146993 f_0(t) - 0.944260 f_1(t) + 1.797267 f_2(t) \tag{155}$$

with $S = 0.273952$. Thus, from Theorem 26,

$$|h_2(t) - f(t)| \leq 0.523404 t \sup_{t>0} |\dot{f}(t)|. \tag{156}$$

The estimate of $\epsilon_n(t;f)$ in (112) may be effectively used to reduce error. One may take as an approximation to $f(t)$ the following

$$f(t) \approx f_n(t) + \nu_n \theta f_n(t) - \frac{\sigma_n^2 - \nu_n^2}{2} \theta^2 f_n(t). \tag{157}$$

In order to improve $f_{n,\alpha}(t)$, the approximate calculation of $\epsilon_{n,\alpha}(t;f)$ proceeds by use of (122). The practical use of (112), (157) uses difference quotients to evaluate $\theta f_n(t)$, $\theta^2 f_n(t)$ from the values already obtained for $f_n(t)$. Thus, let $h > 0$ be the distance between consecutive values of t for which $f_n(t)$ is calculated; then

$$\theta f_n(t) \approx t \frac{f_n(t+h) - f_n(t-h)}{2h} \tag{158}$$

$$\theta^2 f_n(t) \approx \theta f_n(t) + t^2 \frac{f_n(t+h) - 2f_n(t) + f_n(t-h)}{h^2}. \tag{159}$$

The following comment should prove useful in reduction of error. If a function $g(t)$ is known which approximates $f(t)$, for example, the leading term of an asymptotic expansion for $f(t)$, then one may use

$$f(t) \approx g(t) + L_n(f - g). \tag{160}$$

Evidently an appropriate $g(t)$ should always be sought before constructing practical approximations to $f(t)$.

VI. THE RENEWAL FUNCTION

In this section some of the preceding theory will be applied to obtaining approximations for the renewal function, $M(t)$, of a renewal stream.[9] Let $A(t)$, with $A(0+) = 0$, be the interarrival time distribution and $\hat{A}(s)$, given by

$$\hat{A}(s) = \int_0^\infty e^{-st} dA(t) \tag{161}$$

its Laplace-Stieltjes transform, then

$$\tilde{M}(s) = \frac{1}{s} \frac{\hat{A}(s)}{1 - \hat{A}(s)}. \tag{162}$$

The sequence of approximations, $M_n(t)$, may now be constructed from

$$M_0(t) = \frac{1}{1 - \hat{A}(1/t)} - 1. \tag{163}$$

In particular one has

$$M_1(t) = \frac{1}{1 - \hat{A}(2/t)} - 1 - \frac{2}{t} \frac{\hat{A}'(2/t)}{[1 - \hat{A}(2/t)]^2} \tag{164}$$

in which

$$\hat{A}'(s) = \frac{d}{ds} \hat{A}(s). \tag{165}$$

Let λ be the arrival rate, and σ^2 the variance of interarrival time, that is,

$$\lambda^{-1} = \int_0^\infty t dA(t) \tag{166}$$

$$\sigma^2 = \int_0^\infty t^2 dA(t) - \lambda^{-2} \tag{167}$$

then evaluation of the contribution of $\tilde{M}(s)$ at $s = 0$ provides the term

$$\lambda t + \frac{\sigma^2 \lambda^2 - 1}{2}. \tag{168}$$

Thus one may introduce a new function, $f(t)$, by

$$M(t) = \lambda t + \frac{\sigma^2 \lambda^2 - 1}{2} + f(t) \tag{169}$$

with

$$f_0(t) = \frac{\hat{A}(1/t)}{1 - \hat{A}(1/t)} - \lambda t - \frac{\sigma^2\lambda^2 - 1}{2}. \tag{170}$$

Since linear functions are invariants of the operators L_n, there is no reduction of error when approximating $f(t)$ by $f_n(t)$ over approximating $M(t)$ by $M_n(t)$; however, often $f(t)$ is exponentially dominated and the enhancement technique of Theorem 22 is applicable.

The following example will be considered:

$$A(t) = \operatorname{erf}\sqrt{\frac{t}{2}} \qquad \hat{A}(s) = \frac{1}{\sqrt{1+2s}} \tag{171}$$

$$\tilde{M}(s) = \frac{1}{s}\frac{1}{\sqrt{1+2s}-1}. \tag{172}$$

Thus,

$$M_0(t) = \frac{1}{\sqrt{1+2/t}-1} \tag{173}$$

$$M_1(t) = \frac{1}{\sqrt{1+4/t}-1} + \frac{2}{t}\frac{1}{\sqrt{1+4/t}(\sqrt{1+4/t}-1)^2}. \tag{174}$$

Since $\lambda = 1$, $\sigma^2 = 2$, one has

$$M(t) = t + \frac{1}{2} + f(t)$$

$$f_1(t) = \frac{1}{\sqrt{1+4/t}-1} + \frac{2}{t}\frac{1}{\sqrt{1+4/t}(\sqrt{1+4/t}-1)^2} - t - \frac{1}{2}. \tag{175}$$

The α transformation of Theorem 22 may be applied to $f_1(t)$. Assuming $f_1(t)$ to be ultimately of one sign, the singularity farthest to the right of $\tilde{f}(s)$, namely $-1/2$, coincides with the abscissa of convergence; hence, $\alpha = 1/2$. Table II compares the approximations for $M(t)$ given by $M_1(t)$, the enhancement procedure of (133), and $t + 1/2 + f_{1,1/2}(t)$ with more accurate values obtained from the exact solution

$$M(t) = \sum_{n=0}^{\infty} \frac{1}{\Gamma\left(n+\frac{1}{2}\right)} \int_0^{t/2} e^{-u} u^{n-1/2} du. \tag{176}$$

Since $M(t) \approx t + 1/2 (t \to \infty)$, the accuracy increases with increasing t. This is characteristic of the applications to the renewal function.

An example will now be considered in which the interarrival time distribution is a mixture of exponentials since this is of frequent practical use. Accordingly let

Table II — Comparison of approximations

t	$M_1(t)$	$s_1(t)$	$t+\frac{1}{2}+f_{1,\frac{1}{2}}(t)$	$M(t)$
0	0	0	0	0
0.5	0.83333	0.85764	0.84297	0.86007
1	1.39443	1.42283	1.41014	1.42466
2	2.44338	2.47255	2.46303	2.47161
5	5.48142	5.50480	5.49568	5.49718
10	10.49350	10.50977	10.49980	10.49989

$$A(t) = 1 - \frac{7}{10}e^{-t} - \frac{3}{10}e^{-2t}. \tag{177}$$

Then,

$$\hat{A}(s) = \frac{7}{10}\frac{1}{s+1} + \frac{6}{10}\frac{1}{s+2} \tag{178}$$

$$\tilde{M}(s) = \frac{1}{s^2}\frac{20+13s}{17+10s}. \tag{179}$$

Also one has

$$M(t) = \frac{20}{17}t + \frac{21}{289} + f(t) \tag{180}$$

$$f_0(t) = -\frac{210}{289}\frac{1}{17t+10}. \tag{181}$$

Application of Theorem 12 provides

$$f_1(t) = -\frac{8400}{289}\frac{1}{(17t+20)^2} \tag{182}$$

$$f_2(t) = -\frac{5.67\times 10^5}{289}\frac{1}{(17t+30)^3}. \tag{183}$$

The exact solution for this simple example is

$$M(t) = \frac{20}{17}t + \frac{21}{289}\left(1 - e^{-(17/10)t}\right). \tag{184}$$

Table III compares calculations from $M_2(t)$ and the enhancement procedures of (134) and (155) with the exact value. The α enhancement procedure with $\alpha = 1.7$ was not used because it produces the exact result.

This example shows the operation of the enhancement procedures (134) and (155); clearly, $s_2(t)$ is very accurate since the constraint of positivity of the approximation operators is discarded in its construction.

Table III — Comparison of calculations

t	$M_2(t)$	$s_2(t)$	$h_2(t)$	$M(t)$.
0	0	0	0	0
0.5	0.62652	0.62967	0.62712	0.62984
1	1.23024	1.23559	1.23127	1.23586
2	2.41812	2.42353	2.41913	2.42318
5	5.95373	5.95595	5.95407	5.95500
10	11.83713	11.83747	11.83710	11.83737

VII. THE COVARIANCE AND RECOVERY FUNCTIONS

The study of errors in switch count and continuous scan observational methods in telephone traffic engineering is facilitated by use of the covariance function of the number of busy trunks in the Erlang blocking system.[11] Specifically let x_t be the number of trunks busy at time t in an equilibrium $M/M/C$ blocking system with unit mean holding time and offered load of a erlangs, then the covariance function, $R(t)$, is

$$R(t) = E(x_0 x_t) - (Ex_0)^2. \tag{185}$$

In order to express the Laplace transform, $\tilde{R}(s)$, it is necessary to introduce the Poisson-Charlier polynomials[10,13] which may be obtained from

$$G_j(x, a) = \sum_{\nu=0}^{j} (-1)^{j-\nu} \binom{j}{\nu} \nu! a^{-\nu} \binom{x}{\nu}. \tag{186}$$

They satisfy the following recurrence

$$G_{j+1}(x, a) = \frac{x - j - a}{a} G_j(x, a) - \frac{j}{a} G_{j-1}(x, a) \tag{187}$$

$$G_0(x, a) = 1 \qquad G_1(x, a) = \frac{x}{a} - 1.$$

Also needed is the function $\alpha_j(x, a)$ given by

$$\alpha_j(x, a) = \frac{G_{j-1}(x, a)}{G_j(x, a)} \tag{188}$$

which satisfies the first order recurrence

$$\alpha_{j+1}(x, a)^{-1} = \frac{x - j - a}{a} - \frac{j}{a} \alpha_j(x, a) \tag{189}$$

$$\alpha_1(x, a) = \left(\frac{x}{a} - 1\right)^{-1}.$$

The zeros of $G_j(x, a)$ are all positive and simple; in particular, the zeros of $G_j(-s - 1, a)$ as a function of s are designated r_i and ordered by

$$r_j < r_{j-1} < \ldots < r_1 < 0. \tag{190}$$

In the approximation to be developed, r_1 will be the dominant root.

The Erlang loss function,[10] $B(c, a)$, given by

$$B(c, a) = \frac{a^c}{c!} \Big/ \sum_{j=0}^{c} \frac{a^j}{j!}$$

gives the probability that all servers are busy. In the formulae below it will be designated simply by B. The mean number of servers busy, μ, is

$$\mu = a(1 - B) \tag{191}$$

and the variance, σ^2, of the number of busy servers is

$$\sigma^2 = \mu - a(c - \mu)B. \tag{192}$$

The Laplace transform, $\tilde{R}(s)$, and the covariance, $R(t)$, are[11]

$$\tilde{R}(s) = \frac{\sigma^2 + \mu^2}{1 + s} + \frac{a\mu}{s(1 + s)} - \frac{acB}{(1 + s)^2} - \frac{\mu^2}{s} + \frac{acB}{s(1 + s)^2} \alpha_c(-s - 1, a) \tag{193}$$

$$R(t) = \sum_{j=1}^{c} A_j e^{r_j t} \qquad A_j = -\frac{a^2 B}{r_j(1 + r_j)^2} \prod_{i \neq j} \left(1 - \frac{1}{r_j - r_i}\right) \tag{194}$$

The approximation $R_0(t)$ is

$$R_0(t) = \frac{\sigma^2 + \mu^2}{1 + t} + \frac{a\mu t}{1 + t} - \mu^2 - \frac{acBt}{(1 + t)^2} + \frac{acBt^2}{(1 + t)^2} \alpha_c\left(-\frac{1}{t} - 1, a\right). \tag{195a}$$

Since the dominant root is r_1 one may choose α satisfying $0 \leq \alpha \leq -r_1$ to obtain

$$R_{0,\alpha}(t) = \sigma^2 e^{-\alpha t} g(t)$$

$$g(t) = \frac{1 + \mu^2/\sigma^2}{1 + (1 - \alpha)t} + \frac{a\mu t/\sigma^2}{(1 - \alpha t)(1 + (1 - \alpha)t)} - \frac{\mu^2/\sigma^2}{1 - \alpha t} - \frac{acBt/\sigma^2}{(1 + (1 - \alpha)t)^2} + \frac{acBt^2/\sigma^2}{(1 - \alpha t)(1 + (1 - \alpha)t)^2} \alpha_c\left(-\frac{1}{t} + \alpha - 1, a\right). \tag{195b}$$

It is known that the zeros r_j are separated by at least one so that $1 - 1/(r_j - r_i) > 0$, and hence A_j is positive for each j; thus $R(t)$ is log-convex.

Accordingly, the following inequality is valid (Theorem 23):

$$R(t) \leq R_{0,\alpha}(t). \tag{196}$$

In order to facilitate the use of $R_{0,\alpha}(t)$ an accurate upper bound for r_1 is needed to provide a suitable choice for α. Such a bound is available in Ref. 12. Thus let

$$\ell = \sum_{\nu=1}^{c} \frac{1}{\nu} c^{(\nu)} a^{-\nu} \tag{197}$$

$$m = \ell^2 - 2 \sum_{\nu=2}^{c} \frac{1}{\nu} c^{(\nu)} a^{-\nu} \sum_{j=1}^{\nu-1} \frac{1}{j} \tag{198}$$

then

$$r_1 \leq -\frac{c}{\ell + \sqrt{(c-1)(cm - \ell^2)}} - 1. \tag{199}$$

To illustrate the practical performance of (195b) and (199), calculations were made for the cases $a = 4, 8, 12$ and $c = 8$ corresponding to medium, heavy, and very heavy loads respectively. The corresponding equilibrium blocking probabilities are $B(8, 4) = 0.030420$, $B(8, 8) = 0.235570$, $B(8, 12) = 0.422655$. Table IV compares the exact and approximate values. Figures 1(a), 1(b), and 1(c) compare the corresponding curves.

Table IV — Comparison of exact and approximate values

	$a = 4$		$a = 8$		$a = 12$	
t	$R(t)$	$R_{0,\alpha}(t)$	$R(t)$	$R_{0,\alpha}(t)$	$T(t)$	$R_{0,\alpha}(t)$
0	3.377	3.377	2.564	2.564	1.492	1.492
0.4	2.143	2.145	1.075	1.091	0.312	0.331
0.8	1.365	1.367	0.474	0.483	0.075	0.079
1.2	0.870	0.872	0.212	0.216	0.019	0.020
1.6	0.555	0.556	0.095	0.097	0.005	0.005
2.0	0.354	0.355	0.043	0.043	0.001	0.001
2.4	0.226	0.227	0.019	0.019		
2.8	0.144	0.145	0.009	0.009		
3.2	0.092	0.092	0.004	0.004		
3.6	0.059	0.059	0.002	0.002		
4.0	0.038	0.038	0.001	0.001		

The quality of approximation of (199) may be seen from the following values of α used in (195b) compared to the exact r_1 values.

a	$-r_1$	α
4	1.1218	1.1215
8	2.0000	1.9730
12	3.4778	3.3415

The transition probabilities $P_{ij}(t)$—the probability j trunks are busy at time t given i trunks are busy at time zero—may all be obtained from

the transition probability $P_{cc}(t)$[11]; this probability as a function of time is called the recovery function. It may be used in a similar manner to the covariance function, $R(t)$, for the study of errors in scan measurement techniques[14]; additionally it is especially important in the analysis of telephone retrial models.

The Laplace transform, $\tilde{P}_{cc}(s)$, and the recovery function, $P_{cc}(t)$, are[11]

$$\tilde{P}_{cc}(s) = \frac{1}{s} + \frac{c}{as}\alpha_c(-s-1, a) \tag{200}$$

$$P_{cc}(t) = B - \sum_{j=1}^{c} B_j e^{r_j t} \tag{201}$$

$$B_j = \frac{1}{r_j} \prod_{i \neq j} \left(1 - \frac{1}{r_j - r_i}\right). \tag{202}$$

As for the covariance function, $B = B(c, a)$, and $r_j (1 \leq j \leq c)$ are the roots of $G_c(-s-1, a)$ as a function of s.

In order to apply the α enhancement procedure, the function

$$f(t) = P_{cc}(t) - B \tag{203}$$

is considered whose Laplace transform is

$$\tilde{f}(s) = \frac{1}{s}\left[1 - B + \frac{c}{a}\alpha_c(-s-1, a)\right]. \tag{204}$$

It may be observed from (201) and (203) that $f(t)$ is log-convex, hence the approximations obtained will constitute upper bounds. In order to demonstrate the operation of the approximations, the functions $f_{0,\alpha}(t)$, $f_{1,\alpha}(t)$, and

$$s_1(t) = 2f_{1,\alpha}(t) - f_{0,\alpha}(t) \tag{205}$$

were constructed; they are

$$f_{0,\alpha}(t) = \frac{e^{-\alpha t}}{1 - \alpha t}\left[1 - B + \frac{c}{a}\alpha_c\left(-\frac{1}{t} + \alpha - 1, a\right)\right], \tag{206}$$

$$f_{1,\alpha}(t) = \frac{e^{-\alpha t}}{\left(1 - \frac{\alpha t}{2}\right)^2}\left[1 - B + \frac{c}{a}\alpha_c\left(-\frac{2}{t} + \alpha - 1, a\right)\right] + \frac{e^{-\alpha t}}{1 - \frac{\alpha t}{2}}\frac{2c}{at}\alpha_c'\left(-\frac{2}{t} + \alpha - 1, a\right). \tag{207}$$

The prime on $\alpha_c(x, a)$ indicates differentiation with respect to x. The

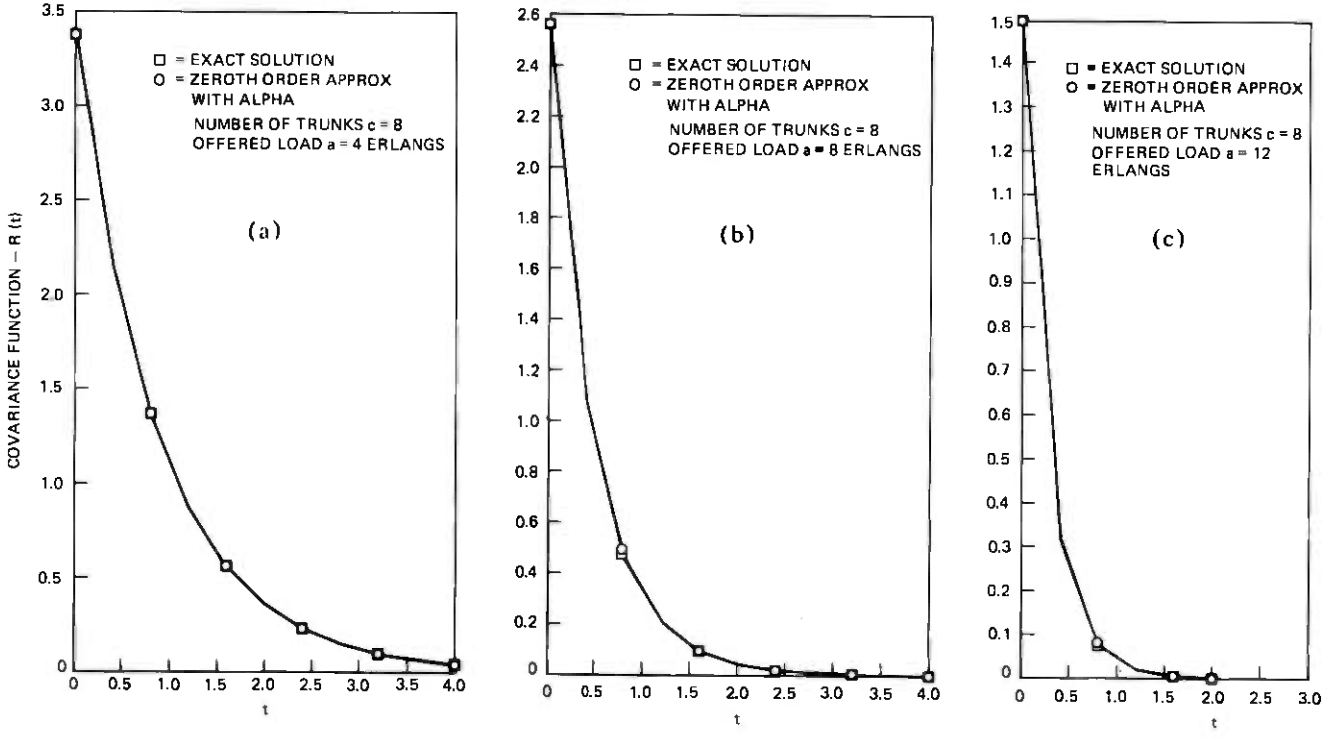


Fig. 1—Covariance function.

following recurrence relation is obtained from (189);

$$\alpha'_{j+1}(x, a) = \frac{1}{a}\,[j\alpha'_j(x, a) - 1]\alpha_{j+1}(x, a)^2 \tag{208}$$

$$\alpha'_1(x, a) = -\frac{1}{a}\,\alpha_1(x, a)^2.$$

Since $s_1(t)$ does not correspond to a positive operator, it does not provide a bound for $P_{cc}(t)$; one has, however,

$$P_{cc}(t) \le B + f_{1,\alpha}(t) \le B + f_{0,\alpha}(t). \tag{209}$$

The first inequality follows from Theorem 23 and the second inequality from Theorem 21.

The same cases as for the covariance function were treated. Tables V, VI, and VII compare the exact and approximate values. Figures 2(a), 2(b), and 2(c) compare the corresponding curves.

Table V — $a = 4$

t	$P_{cc}(t)$	$B + f_{0,\alpha}(t)$	$B + f_{1,\alpha}(t)$	$s_1(t)$
0	1.0000	1.0000	1.0000	1.0000
0.1	0.5178	0.5907	0.5597	0.5287
0.2	0.3304	0.4280	0.3856	0.3432
0.3	0.2380	0.3335	0.2901	0.2468
0.4	0.1844	0.2703	0.2296	0.1889
0.5	0.1497	0.2249	0.1880	0.1511
0.6	0.1256	0.1907	0.1578	0.1248
0.7	0.1080	0.1641	0.1350	0.1059
0.8	0.0947	0.1429	0.1174	0.0918
0.9	0.0844	0.1258	0.1034	0.0810
1.0	0.0762	0.1118	0.0922	0.0726

Table VI — $a = 8$

t	$P_{cc}(t)$	$B + f_{0,\alpha}(t)$	$B + f_{1,\alpha}(t)$	$s_1(t)$
0	1.0000	1.0000	1.0000	1.0000
0.1	0.5756	0.6335	0.6088	0.5842
0.2	0.4379	0.5005	0.4725	0.4445
0.3	0.3727	0.4256	0.4006	0.3756
0.4	0.3347	0.3770	0.3561	0.3352
0.5	0.3099	0.3432	0.3262	0.3092
0.6	0.2927	0.3187	0.3050	0.2913
0.7	0.2802	0.3005	0.2895	0.2786
0.8	0.2708	0.2866	0.2779	0.2692
0.9	0.2637	0.2760	0.2690	0.2621
1.0	0.2581	0.2677	0.2622	0.2567

This example will be used to show the operation of the error estimate (112). Using the increment $h = 0.1$, (158) and (159) were used to obtain $\theta[e^{\alpha t}f_{0,\alpha}(t)]$, $\theta^2[e^{\alpha t}f_{0,\alpha}(t)]$ and $\theta[e^{\alpha t}f_{1,\alpha}(t)]$, $\theta^2[e^{\alpha t}f_{1,\alpha}(t)]$ at $t = 0.5$. Equation (122) was used to estimate $\epsilon_{0,\alpha}(t)$, $\epsilon_{1,\alpha}(t)$. The error in $s_1(t)$ was

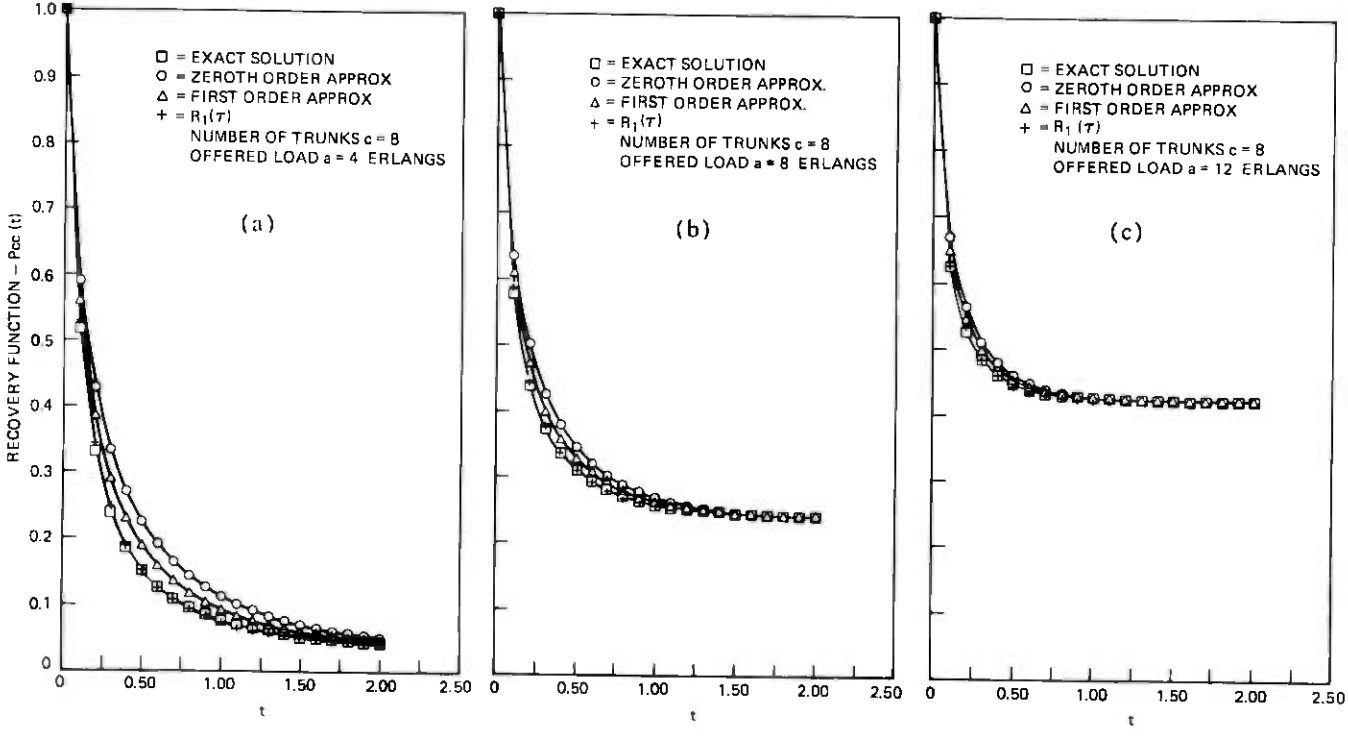


Fig. 2—Recovery function.

Table VII — $a = 12$

t	$P_c(t)$	$B + f_{0,\alpha}(t)$	$B + f_{1,\alpha}(t)$	$s_1(t)$
0	1.0000	1.0000	1.0000	1.0000
0.1	0.6245	0.6679	0.6493	0.6307
0.2	0.5255	0.5629	0.5458	0.5286
0.3	0.4834	0.5097	0.4969	0.4840
0.4	0.4611	0.4789	0.4698	0.4607
0.5	0.4479	0.4598	0.4535	0.4472
0.6	0.4396	0.4476	0.4427	0.4378
0.7	0.4342	0.4396	0.4366	0.4336
0.8	0.4306	0.4342	0.4322	0.4301
0.9	0.4282	0.4306	0.4292	0.4278
1.0	0.4265	0.4281	0.4272	0.4262

estimated by $2\epsilon_{1,\alpha}(t) - \epsilon_{0,\alpha}(t)$ in which the estimates for $\epsilon_{0,\alpha}(t)$, $\epsilon_{1,\alpha}(t)$ were used. The results obtained are given in Table VIII.

Table VIII — Error estimates at $t = 0.5$

a	$\epsilon_{0,\alpha}$	Estimate	$\epsilon_{1,\alpha}$	Estimate	$s_1 - f$	Estimate
4	0.0752	0.0714	0.0383	0.0372	0.0014	0.0030
8	0.0333	0.0317	0.0163	0.0160	−0.0008	0.0002
12	0.0120	0.0113	0.0056	0.0001	−0.0007	−0.0111

VIII. SOME APPLICATIONS OF THEOREM 12

The generating function, which will be designated $G(z, t)$, of Theorem 12, namely,

$$G(z, t) = \frac{1}{1 - z} f_0\left(\frac{t}{1 - z}\right) \tag{210}$$

may sometimes be used to obtain explicitly the form of $f_n(t)$. The following are some examples.

For $f(t) = \cos t$, one has $f_0(t) = 1/(1 + t^2)$, hence

$$G(z, t) = \frac{1 - z}{(1 - z)^2 + t^2}. \tag{211}$$

The generating function for the Chebyshev polynomials, $T_n(t)$, of first kind is[15]

$$\frac{1 - tz}{1 - 2tz + z^2} = \sum_{n=0}^{\infty} T_n(t) z^n \tag{212}$$

hence

$$G(z, t) = \sum_{n=0}^{\infty} z^n (1 + t^2)^{-(n+1)/2} T_n\left(\frac{1}{\sqrt{1 + t^2}}\right). \tag{213}$$

One now obtains

$$f_n(t) = L_n \cos t = \left[1 + \left(\frac{t}{n+1}\right)^2\right]^{-(n+1)/2} T_n\left[\frac{1}{\sqrt{1 + \left(\frac{t}{n+1}\right)^2}}\right] \tag{214}$$

For $f(t) = \sin t$, one has $f_0(t) = t/(1 + t^2)$, hence

$$G(z, t) = \frac{t}{(1 - z)^2 + t^2}. \tag{215}$$

The generating function for the Chebyshev polynomials, $U_n(t)$, of second kind is

$$\frac{1}{1 - 2tz + z^2} = \sum_{n=0}^{\infty} U_n(t) z^n \tag{216}$$

hence

$$G(z, t) = \sum_{n=0}^{\infty} z^n (1 + t^2)^{-(n+1)/2} t U_n\left(\frac{1}{\sqrt{1 + t^2}}\right). \tag{217a}$$

Thus

$$f_n(t) = L_n \sin t = \frac{t}{n+1}\left[1 + \left(\frac{t}{n+1}\right)^2\right]^{-(n+1)/2} \times U_n\left[\frac{1}{\sqrt{1 + \left(\frac{t}{n+1}\right)^2}}\right]. \tag{217b}$$

The Bessel functions provide additional interesting relations with orthogonal polynomials. For $f(t) = J_0(t)$, one has $f_0(t) = 1/\sqrt{1 + t^2}$, and

$$G(z, t) = \frac{1}{\sqrt{(1 - z)^2 + t^2}}. \tag{218}$$

The Legendre polynomials, $P_n(t)$, are generated by

$$\frac{1}{\sqrt{1 - 2tz + z^2}} = \sum_{n=0}^{\infty} P_n(t) z^n, \tag{219}$$

hence

$$f_n(t) = L_n J_0(t) = \left[1 + \left(\frac{t}{n+1}\right)^2\right]^{-(n+1)/2} P_n\left[\frac{1}{\sqrt{1 + \left(\frac{t}{n+1}\right)^2}}\right]. \tag{220}$$

By the substitution of it for t, one derives immediately

$$f_n(t) = L_n I_0(t) = \left[1 - \left(\frac{t}{n+1}\right)^2\right]^{-(n+1)/2} \cdot P_n\left[\frac{1}{\sqrt{1 - \left(\frac{t}{n+1}\right)^2}}\right]. \quad (221)$$

Since $I_0(t)$ is convex, one also has

$$I_0(t) \leq L_n I_0(t) \quad (222)$$

for sufficiently large n.

As another example relating to Bessel functions, consider $f(t) = J_0(2\sqrt{t})$, then $f_0(t) = e^{-t}$ and

$$G(z, t) = \frac{1}{1-z} e^{-t/(1-z)}. \quad (223)$$

The generating function for the Laguerre polynomials, $L_n(t)$, is

$$\frac{1}{1-z} e^{-tz/(1-z)} = \sum_{n=0}^{\infty} L_n(t) z^n \quad (224)$$

hence

$$f_n(t) = L_n J_0(2\sqrt{t}) = e^{-t/(n+1)} L_n\left(\frac{t}{n+1}\right). \quad (225)$$

IX. SUMMARY

The methods of this paper have been found particularly useful in analyzing complex queueing phenomena whose Laplace transform representations are quite often implicitly defined. The error estimate of (112) has been found especially useful. Its computation is numerically effected by use of (158) and (159).

It would be desirable to have an effective method of estimating the α parameter of (114) directly from $f_n(t)$. In fact a method of this type which yields a rough evaluation has been devised and will be reported in a later paper. Of interest also would be further elaboration of the way structural properties of $f(t)$ are reflected in $f_{n,\alpha}(t)$.

The investigation of linear combinations of iterates, L_n^r, of the operators L_n may prove useful in providing additional enhancement methods. Especially, further investigation is needed concerning enhancement

methods which preserve the positivity of the approximation process.

The isolated result of Theorem 16, which shows that $\exp(L_n \ell_n f(t))$ is a better approximation to $f(t)$ than $f_n(t)$ when $f(t)$ is log-convex, should be examined with the purpose of the possible construction of nonlinear approximation methods exploiting this structural characteristic.

X. ACKNOWLEDGMENTS

I should like to acknowledge the helpful discussions with B. W. Stuck and E. Arthurs during their application of these methods to the analysis of flow distributions for computer models, and to acknowledge gratefully the excellent computer programming work of Rosemary Harris and Darlene Shearer.

APPENDIX

Operations

$f(t)$	$f_0(t)$ or $f_n(t)$
$\frac{1}{t}f(t)$	$\frac{1}{t}\int_0^t f_0(x)\frac{dx}{x}$
$f(at),\ a > 0$	$f_n(at)$
$e^{\alpha t}f(t)$	$\left(1-\frac{\alpha t}{n+1}\right)^{-n-1} f_n\left(\frac{t}{1-\alpha t/(n+1)}\right)$
$\int_0^t f(x)dx$	$\frac{t}{n+1}\sum_{j=0}^{n} f_j\left(t\frac{j+1}{n+1}\right)$
$\dot{f}(t)$	$\frac{f_0(t)-f(0)}{t},\ \frac{n+1}{t}\left[f_n(t)-f_{n-1}\left(\frac{n}{n+1}t\right)\right]\ (n \geq 1)$
$\ddot{f}(t)$	$\frac{f_0(t)-f(0)-t\dot{f}(0)}{t^2},\ 4\frac{f_1(t)-2f_0(t/2)+f(0)}{t^2}$
	$(n+1)^2\ \frac{f_n(t)-2f_{n-1}\left(\frac{n}{n+1}t\right)+f_{n-2}\left(\frac{n-1}{n+1}t\right)}{t^2}\ (n \geq$
$tf(t)$	$tf_0(t)+t^2\dot{f}_0(t)$
$t\dot{f}(t)$	$t\dot{f}_n(t)$
$\frac{1}{t}\int_0^t f(x)dx$	$\frac{1}{t}\int_0^t f_n(x)dx$
$f(t)*h(t)$(Mellin)	$f_n(t)*h(t)$

REFERENCES

1. R. E. Paley and N. Weiner, *Fourier Transforms in the Complex Domain*, American Mathematical Society, Vol. XIX, Colloquium Publications, 1934.
2. P. P. Korovkin, *Linear Operations and Approximation Theory*, New York: Gordon and Breach, 1960.
3. W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. II, John Wiley & Sons.
4. I. I. Hirschman and D. V. Widder, *The Convolution Transform*, Princeton University Press, 1955.
5. S. Karlin and Z. Ziegler, "Iteration of Positive Approximation Operators," Journal of Approximation Theory, *3*, 1970, pp. 310–339.
6. E. Artin, *The Gamma Function*, Holt, Rinehart and Winston.
7. J. A. Shohat and J. D. Tamarkin, "The Problem of Moments," Mathematical Surveys No. 1, American Mathematical Society, 1943.
8. G. Pólya and G. Szegö, *Aufgaben und Lehrsätze aus der Analysis*, Vol. 2, Dover, 1945.
9. D. R. Cox, *Renewal Theory*, New York, John Wiley and Sons.
10. D. L. Jagerman, "Some Properties of the Erlang Loss Function," B.S.T.J., *53*, No. 3 (March 1974).
11. V. E. Beneš, *Mathematical Theory of Connecting Networks and Telephone Traffic*, Academic Press, 1965.
12. D. L. Jagerman, "Nonstationary Blocking in Telephone Traffic," B.S.T.J., *54*, No. 3 (March 1975).
13. C. Jordon, *Calculus of Finite Differences*, New York: Chelsea, 1947, Chap. 8.
14. A. Descloux, "On the Accuracy of Loss Estimates," B.S.T.J., *44*, No. 6 (July-August), 1965.
15. J. Todd, *Introduction to the Constructive Theory of Functions*, New York: Academic Press, 1963.
16. R. Piessens, "A Bibliography on Numerical Inversion of the Laplace Transform and its Applications," J. Comp. Appl. Math., *1*, 1975, 115–128.
17. R. Piessens and N. D. P. Dang, "A Bibliography on Numerical Inversion of the Laplace Transform and Applications: A Supplement," J. Comp. Appl. Math., *2*, No. 3, 1976, pp. 225–228.
18. M. Eisenberg, "A Program for the Analysis of a Class of Electronic Switching Systems—Appendix A," MM-71-3423-3.
19. R. Bellman, R. E. Kalaba, J. A. Lockett, *Numerical Inversion of the Laplace Transform*, New York: American Elsevier Pub. Co., 1966.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 3, March 1978
Printed in U.S.A.

Observations of Errors and Error Rates on T1 Digital Repeatered Lines

By M. B. BRILLIANT

(Manuscript received July 29, 1977)

Measurements of errors on T1 repeatered lines were made at five Bell System offices during 1973 and 1974. They included an informal survey of error rates and error-free seconds on lines in service, and detailed recordings, normally of 24 hours duration on selected lines. The detailed recordings show the existence of at least two distinct error mechanisms, differing significantly in diurnal variation of error rate, distribution of intervals between errors, and dependence on the transmitted bit pattern. It was found that certain T1 lines made errors when in service, driven by D1 channel banks, but not when driven by a pseudorandom test signal.

I. INTRODUCTION

The T1 repeatered line is a short-haul digital transmission system using cable pairs to transmit binary information at a rate of 1.544×10^6 bits per second.[1] T1 lines have been in use since 1962, primarily in conjunction with D1 channel banks to provide a carrier system for voice channels. In the design of the T1 line, the distribution of error rates was an essential design parameter.[2] However, since the field trial of an experimental prototype system, in which the error performance of one worst-case repeatered line was briefly reported,* it has become clear that more knowledge is needed on the subject of the error performance of working T1 lines.

Measurements of timing jitter and errors on T1 lines were made at five Bell System offices during 1973–74. The error measurements were of two types: a survey of error rates and error-free seconds; and detailed recordings of the error process, normally for 24 hours duration, on selected T1 lines. The jitter measurements will be reported elsewhere.[4]

The survey consisted of error rate measurements (determined by

* See pp. 95–96, Ref. 3.

counting bipolar violations on lines in service) on 2594 T1 lines, and error-free seconds measurements on 1640 of these lines. The primary purpose of the survey was to select particular lines, known to be making errors, for detailed recordings of the error process. Such recordings were made on 89 lines. However, many of these recordings showed few or no errors. Recordings of 37 of these lines were analyzed to characterize the following properties: diurnal variation of error rate, distribution of the intervals between successive errors, and sensitivity to the pattern of bits transmitted on the T1 line. Not all these properties could be analyzed for every line; there are 22 lines for which all three properties have been characterized.

Most of the lines in the survey were terminated with D1 channel banks; some had D2 banks. All of the detailed recordings were made on lines with D1 banks. Since the pulse stream format generated by a D1 channel bank is different from that generated by other equipment (for example, in regard to density of ones), some specific results of these measurements are valid only for the D1 environment.

II. ERRORS AND THEIR MEASUREMENT

The digital format on the T1 line is bipolar: that is, the absence of a pulse in any position represents a binary "zero," while a pulse of either polarity represents a binary "one," and consecutive pulses, regardless of the number of intervening zeros, have opposite polarity. The pulses become attenuated and distorted in transmission along the cable pair and are reconstructed by repeaters located at intervals of nominally 6000 feet along the line. An "error" is an incorrect reconstruction in any one pulse position: a pulse where originally no pulse was sent, or a blank where a pulse should be.

If one looks only at the presence or absence of pulses—the binary ones and zeros—errors cannot be detected unless one knows what was sent. But by looking at pulse polarity, one can detect violations of the bipolar format: the occurrence of two consecutive pulses (with or without intervening blank spaces) having the same polarity. Any single isolated error—either the omission or insertion of one pulse—always results in exactly one "bipolar violation" (BPV). On the other hand, multiple errors can combine so that a bipolar violation does not occur, while the reversal of the polarity of a pulse would create two bipolar violations without any errors, and, in any case, the occurrence of a bipolar violation does not define precisely where an error occurred or whether it was an insertion or deletion. In practice, however, the rate of occurrence of bipolar violations is generally close to the error rate. Both are quoted as numbers representing the ratio of the number of events (errors or bipolar violations) to the number of bits transmitted.

Another measure of performance, "percent error-free seconds," is used to specify objectives for the Digital Data System.[5] To define percent error-free seconds, the measurement period is divided into 1-second intervals from an arbitrary starting point, and the 1-second intervals in which errors occur—"error seconds"—are counted. In practice, this count is estimated by counting the intervals in which bipolar violations occur. The remaining intervals are "error-free seconds."

The surveys reported here were based on measurement of lines in service, carrying pulse streams that were unknown to us (except that they obeyed the constraints imposed by the D1 channel bank format). Therefore, although the terms "error rate" and "error-free seconds" are loosely used in referring to the results of the survey, all the survey results and measurements are actually in terms of bipolar violations rather than errors.

On the other hand, the detailed recordings of selected lines were originally planned to record true errors on T1 lines removed from service and carrying a known pseudorandom bit stream. As the program developed we also included recordings of bipolar violations on T1 lines carrying unknown pulse streams from D1 channel banks. Therefore, in referring to the detailed recordings, "errors" and "bipolar violations" are not the same; these terms will be used more strictly in that context than in referring to the survey results.

III. THE SURVEY OF ERROR RATE AND ERROR-FREE SECONDS

3.1 Survey procedure

The scope of the survey is summarized in Table I. The activity at the first office visited was a pilot run for the rest of the program. The survey activity at this office consisted of 1-minute error counts using simple bipolar violation counters. After the pilot project a device was built to count bipolar violations for 54-minute intervals on 16 lines at a time, in conjunction with a PDP-11/20 computer (which was used primarily for the jitter measurements and the detailed error recordings). This device was used in the remaining four offices. After the survey at the second office, the software was modified to obtain counts of error seconds concurrently with the error counts. Error-second counts were thus made at the last three offices surveyed.

Since the original purpose of the survey at each office was to select T1 lines for extended error tests, general information about the population of lines surveyed was not recorded. Some information, however, is available about each office as a whole. Thus, for example, it is known that the lines were all two-cable at office 3, all one-cable at office 5, and mixed at the other offices. (A two-cable T1 system is one in which opposite directions of transmission are segregated in separate cables.)

Table I — Scope of the survey

Office	Survey dates	Number of lines	Data recorded
1	3/1/73 to 4/1/73	461	1-min BPV
2	9/20/73 to 9/27/73	493	54-min BPV
3	10/31/73 to 11/9/73	561	54-min BPV and error-seconds
4	11/29/73 to 1/9/74	792	54-min BPV and error-seconds
5	1/21/74 to 1/24/74	287	54-min BPV and error-seconds
Total		2594	

The survey was not based on random sampling. Measurements began at one end of the office repeater bay lineup and continued either until the other end of the lineup was reached (at offices 3, 4, and 5) or until the expiration of the time allotted for activity at that office (at offices 1 and 2). Lines that were out of service for any reason were skipped over and excluded from the survey. Measurements for the survey were made only on business days between the hours of 8:00 a.m. and 6:00 p.m.

At each office repeater included in the survey, a bipolar violation detector was plugged into the receive monitor jack to detect errors in the pulse stream from the distant office. Each office repeater was counted as a separate T1 line. A through system from one distant office to another via the office surveyed was thus effectively split into two parts and measured as though it consisted of two systems, each connecting the survey office with one of the distant offices. One consequence is that there are more short T1 lines (especially single-span lines) in the survey than in the plant.

Error rate measurements at office 1 were made for 1-minute intervals on one line at a time using one of two different instruments. One was a Philco-Sierra 314A T1 error detecting set, which has a built-in counter with a maximum counting rate of 10 per second. The other was a Western Electric J98710G-2 L3 error detecting set connected to a General Radio 1192 counter, which counted bipolar violations as fast as they could occur on the T1 line.

At the other offices, errors were counted by a 16-line bipolar violation detector using a PDP-11/20 computer. The counting program had two versions, as indicated previously. Version 1, used only at office 2, counted only errors. Version 2, used at the last three offices surveyed, counted error seconds as well as errors. The counting rate of this equipment was subject to the following limitations.

(*i*) (Version 2 only)—If the count in any one second exceeded 1544 errors on all 16 lines combined (corresponding to 10^{-3} error rate on any

one line) the counting program was terminated, so that its use of computer time would not interfere with concurrent detailed recording or jitter measurement on another line.

(*ii*) Maximum short-term (fractional-second) counting rate on all lines combined was about 10,000 per second, because the computer took about 95 μs to process each error counted (about 60 μs in Version 1).

(*iii*) Resolution varied from normally about 30 μs to several hundred microseconds depending on error activity. Errors closer together than this on the same T1 line would be counted as a single error.

3.2 Survey data processing

The survey data was recorded manually at the test site and later transcribed to punched cards. The survey data consisted of groups of up to 16 lines, monitored for 54 minutes per group, except at office 1 where each line was monitored for 1 minute. Twenty lines that lost signal at some time during the 54-minute test were eliminated. Error counts were converted directly to error rates by dividing by the number of T1 line bits expected in the monitoring interval at the nominal line rate. Computer programs were used to obtain statistical summaries of three quantities for each line (where available): the percent of seconds with error, the error rate, and a clustering factor. These are all represented graphically by cumulative distribution functions in Figs. 1 to 6.

The clustering factor is not simply the ratio of the number of errors to the number of error seconds. This would give large overestimates of clustering at high error rates (above about 10^{-6}) because the number of error seconds in 54 minutes is bounded. The calculated clustering factor therefore compares the actual count of error seconds with the number of error seconds that would be expected if the errors counted had occurred randomly (Poisson process model). This factor gives the same result at low error rates, but approaches unity at high error rates.

For a 54-minute test, the error rate (strictly, the bipolar violation rate) is

$$\mathrm{ER} = \frac{\text{bipolar violation count}}{1.544 \times 10^6 \times 60 \times 54}$$

The fraction of seconds that contain errors is

$$\mathrm{ESF} = \frac{\text{error second count}}{60 \times 54}$$

The percent error-free seconds is simply

$$\mathrm{PEFS} = 100 \times (1 - \mathrm{ESF})$$

Given an error rate ER, a Poisson process model could be used to predict

an error second fraction

$$\mathrm{ESF}' = 1 - \exp(-1.544 \times 10^6 \times \mathrm{ER})$$

The clustering factor is defined as

$$\mathrm{CF} = \frac{\mathrm{ESF}'}{\mathrm{ESF}}$$

The graphs were arranged so that the horizontal axis consisted of a logarithmic scale and the vertical axis consisted of a normal probability scale; thus, a lognormal distribution would be plotted as a straight line. Percent of seconds with error and BPV rate graphs were based on total population, i.e., lines with and without errors. The lowest point plotted corresponds to the percent of lines that had either no errors or one error. Clustering factor graphs were based solely on T1 lines with errors. To avoid step functions appearing on the graphs, a continuous curve was created by plotting only the upper "corners" of the step function.

3.3 Results of the survey

Figure 1 shows the distribution of error rate at each office. Each point on a given distribution can be interpreted as a statement that some percent of the lines had error rates equal to or less than (better than) a certain error rate. For example, about 98.8 percent of lines at office 2 had error rates equal to or better than 10^{-6}. Because of the counting rate limitations described previously, the true error rate may be somewhat higher than the measured error rate for some lines. On the average, it is probably not more than twice the measured error rate, based on estimates of clustering from an earlier survey[6] as well as results from our detailed error recordings. Thus, for the same example, we would estimate that at office 2 less than 98.8 percent of lines had better than 10^{-6} error rate, but probably more than 98.8 percent had better than 2×10^{-6} error rate. That is, the true error rate curve might be farther to the right, by less than a factor of 2 in error rate.

Figure 2 shows the aggregate distributions of error rate compared with previous surveys. All the surveys refer to lines terminated predominantly or entirely at D1 channel banks. Two curves are shown for our survey because measurements of error rates below 10^{-8} are not available for office 1. However, the two curves are very close together.

The previous surveys had suggested a possibility that T1 error performance was changing. The latter of these surveys,[6] in 1971, showed error rates about 10 times as high as the earlier survey,[7] in 1966. A survey at one site in 1968 showed a distribution (not shown in Fig. 2) lying between the other two.[8] However, the results of our survey show such large differences between offices that the difference between the 1966 and 1971

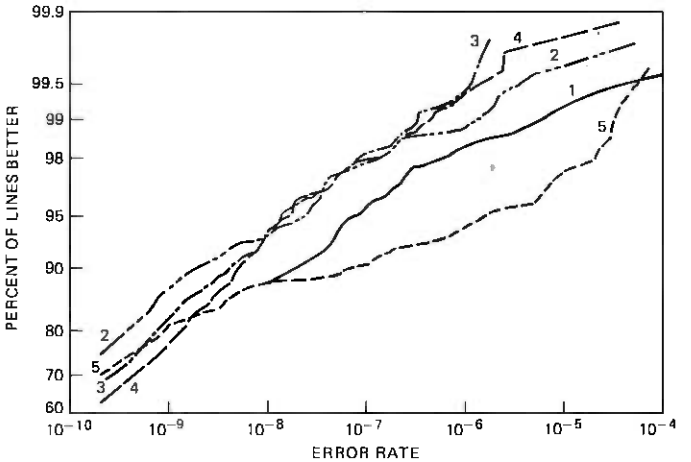


Fig. 1—Error rate distributions for individual offices.

surveys cannot be said to show a significant trend over time. Qualitative observations suggest that the important difference between offices is in maintenance organization and effort, which tends for practical reasons to be correlated with (although not fully determined by) both office size and T-carrier network size.

Offices 2, 3, and 4 in our survey had error rate distributions very close to the 1966 survey. Even allowing for a possible underestimate of a factor

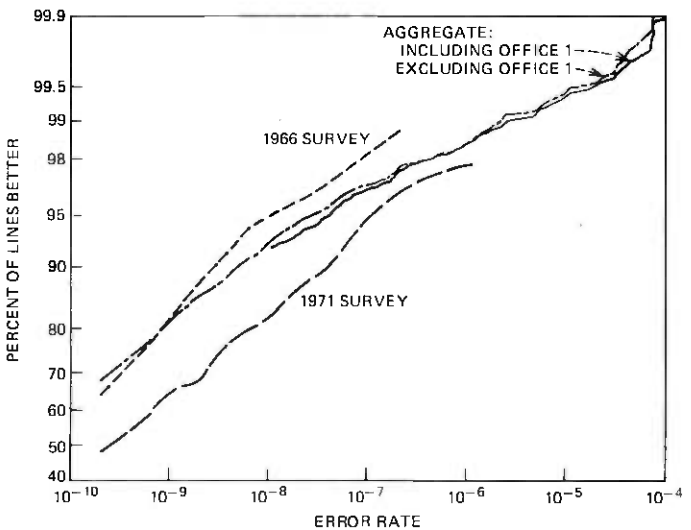


Fig. 2—Aggregate error rate distributions. Results of previous surveys are shown for comparison.

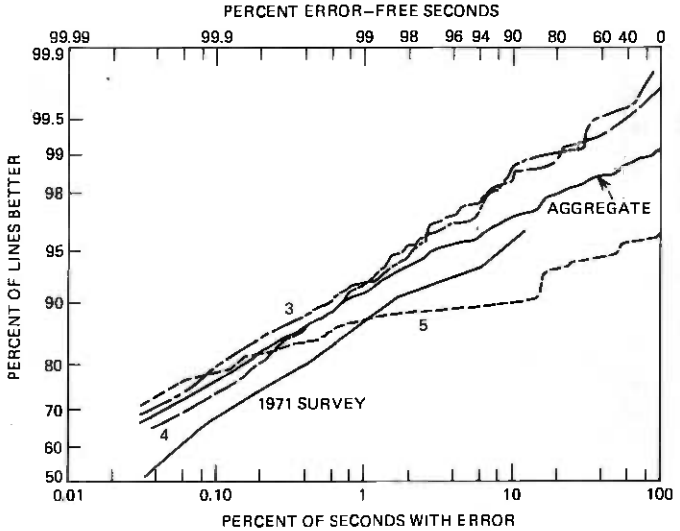


Fig. 3—Distributions of percent seconds with error, compared with the aggregate of the previous survey.

of 2, they are still much better than the 1971 error rate distribution, and much closer to the 1966 curve than to the 1971 curve. These three offices are all large offices in large metropolitan networks, as were the offices in the 1966 survey. In the 1971 survey, on the other hand, two of the three offices were suburban offices.

The office showing the poorest distribution in our survey was also the smallest. This building housed two step-by-step switching machines, and a combined maintenance force was responsible for both switches and T-carrier equipment. This distribution was noticeably affected by the existence of high error rates (between 10^{-6} and 10^{-4}) on 12 lines in the same cable (4 percent of the sample), apparently due to a single cause; omitting these lines would have moved the curve closer to the others.

Error-free seconds results are plotted in Fig. 3 in terms of percent seconds with error (which is 100 minus percent error-free seconds), so that the distribution curves are similar to the error rate distributions. These results are not affected by the counting-rate limitations of the measurement apparatus. The best error-seconds distribution in our survey (office 3) is better than the aggregate of the 1971 survey for about a factor of 4 in percent seconds with error (see Fig. 3); our aggregate is better by a factor of 2. The scatter plot in Fig. 4 indicates that most lines had about the same number of errors as seconds with error (except at high error rates, where the number of seconds with error approaches the number of seconds in 54 minutes). However, clustering can occur at all levels. Figs. 5 and 6 show the distribution of clustering factor (calculated

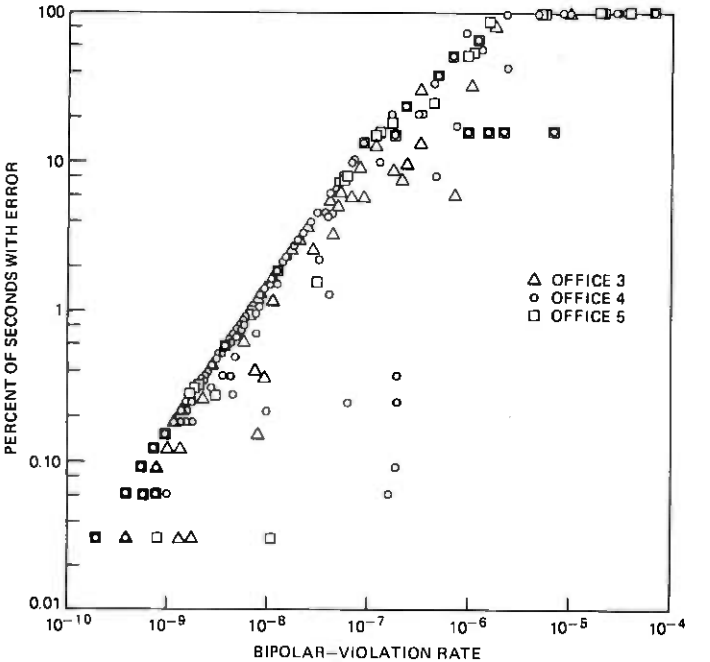


Fig. 4—Scatter plot of percent seconds with error vs. error rate.

so as to avoid false indications of clustering at high error rates, as described in the preceding section). Table II shows the mean and standard deviation of the clustering factor for each office and for the aggregate. In general the standard deviation is much larger than the mean, except at office 3. Overall, the mean is about two events per second-with-error.

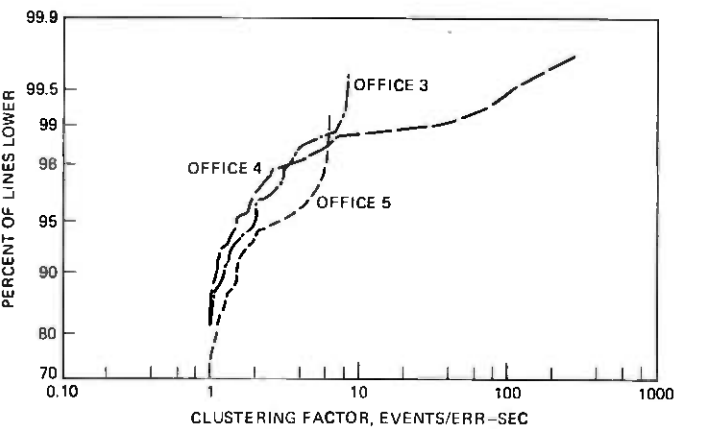


Fig. 5—Distributions of clustering factor for individual offices.

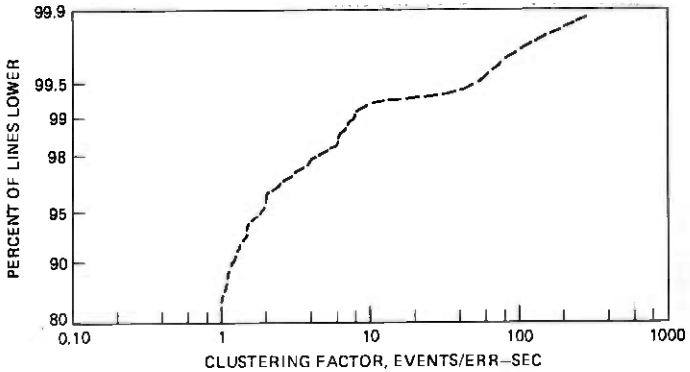


Fig. 6—Aggregate distribution of clustering factor.

Table II — Statistics of clustering factor

Office	Mean	Standard deviation
3	1.20	0.98
4	3.13	22.9
5	1.70	4.93
Aggregate	2.26	16.5

The difference between clustering factors at different offices should probably not be regarded as significant. The unusual statistics for office 4 are due to just four lines with unusually high clustering factors (1 percent of the lines with errors). At office 5, the sample was small, and it is known that several lines could be disturbed by the same condition (although the 12 lines with high error rate in the same cable, mentioned earlier, did not have high clustering factors).

IV. DETAILED RECORDINGS OF ERRORS AND BIPOLAR VIOLATIONS

4.1 Experimental procedure

This phase of the measurement program was designed to obtain complete and continuous recordings of errors on T1 lines for periods of 24 hours (or over a weekend), recording at what time, and on which bit, each individual error occurred. We made no effort to find or fix the cause of the errors.

As originally conceived, the basic procedure was to remove a line from service and apply a test signal with the line looped back to the test site, so that the received signal could be directly compared with the transmitted signal. The test apparatus is shown schematically in Fig. 7. The test signal was a pseudorandom linear shift-register sequence[9] with a period of 1,048,575 bits, generated by a 20-bit shift register, as shown

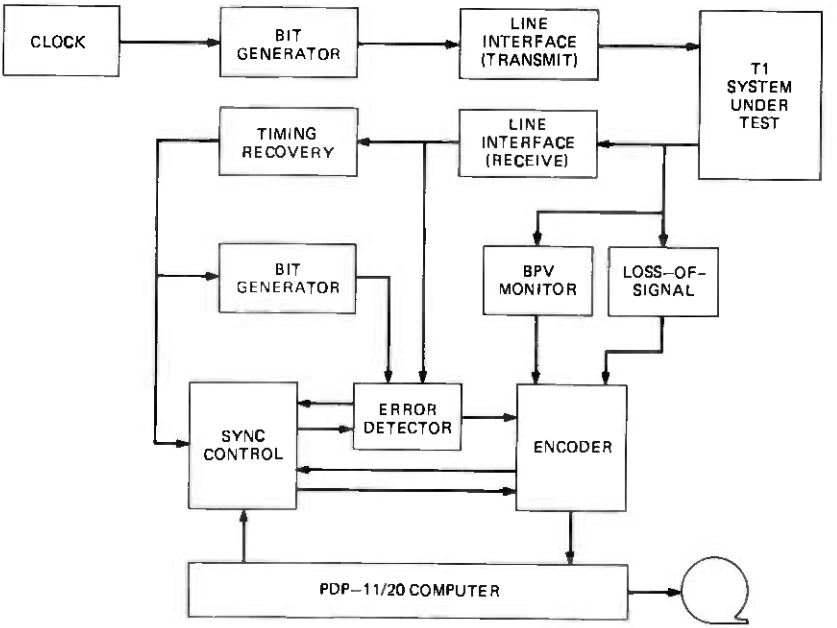


Fig. 7—Error measurement system block diagram.

in Fig. 8. The bit stream returned via the looped T1 line was compared, bit by bit, with the output of a similar sequence generator to detect errors on the line. The occurrence of each error was recorded on magnetic tape. A Digital Equipment Corporation PDP-11 computer controlled test sequencing and formatted the output to the tape.

In most cases the entire T1 line was not removed from service. A T1 line from one office to another generally passes through several intermediate offices, which divide it into spans of the order of 5 miles long. Spare spans are normally provided. The least troublesome procedure

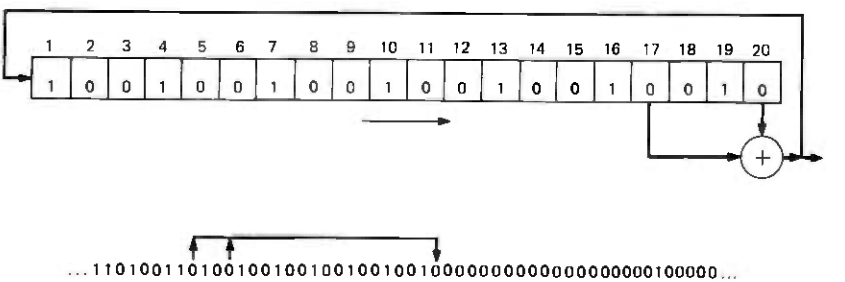


Fig. 8—Pseudorandom sequence generator. The digits in the shift register move from left to right, and hence appear in the output as read from right to left. Generation of the first of the string of 19 zeros is illustrated.

to obtain a length of T1 line for testing is to patch service on one span onto a spare. This procedure requires patching only at the test site and one adjacent office, and does not involve any other offices.

In the earliest measurements at office 1, lines were chosen at random to record both jitter and errors. Since most of the lines so chosen did not make any errors, the bipolar violation survey was introduced as a means of selecting lines for error test that were known to make errors, continuing to use random selection for the jitter measurements. We reasoned that if bipolar violations were observed in the pulse stream coming from the D1 channel bank at the distant terminal office, errors must have been occurring somewhere on the line, and there was a reasonable probability that these errors would be in the span adjacent to the test site.

This procedure resulted in a reasonable yield of error observations at office 1 and office 2. However, it was not successful at office 3. In an effort to improve the yield at this office, lines were chosen that terminated at the next office, so that looping one span would necessarily include the part of the line where the bipolar violations originated. But when these lines were looped and the pseudorandom test signal was applied, errors either did not occur at all, or occurred only occasionally, at rates far lower than the BPV rates observed in service in the survey. Such cases had been observed occasionally at offices 1 and 2. In addition, at office 4, we were able to loop T1 lines back to the test site at distant terminal offices, and again we observed that most lines that showed bipolar violations when in service did not make errors when carrying the pseudorandom test stream. In most cases we verified that the bipolar violations did not originate in the channel bank.

The error recording equipment was therefore modified, late in the measurement program, so as to be able to record bipolar violations in the same manner as errors. The procedure was modified to include pretests that would indicate systematically whether the line should be tested using bit generators and the error detector to record errors ("error run"), as shown in Fig. 9a, or restored to service and tested as in Fig. 9b, to record bipolar violations in service ("BPV run").

The modified procedure included two pretests designed to investigate the effect of the signal on the error performance, using the configuration shown in Fig. 9c. The T1 span was out of service and looped at the next office as in Fig. 9a, but use was made of the signal from the D1 channel bank, which was now routed to its destination via the spare. The dashed lines show alternative connections. One pretest recorded BPVs on the spare. In the next pretest, a signal tapped from the spare was fed into the looped span, and BPVs were recorded on the signal coming back from the looped span. The significant result was that in many cases, BPVs were not seen on the spare, but were seen on the signal that came from the spare via the looped span, although errors did not occur when the signal

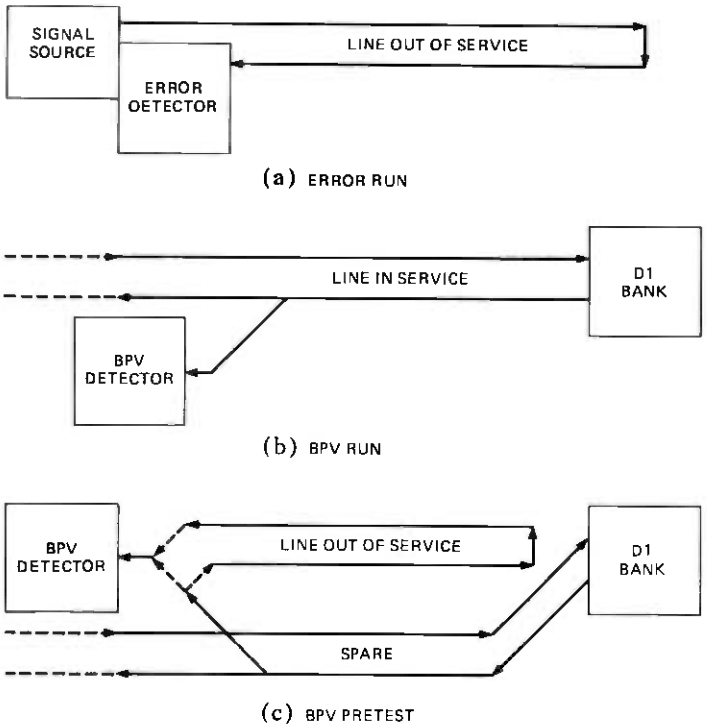


Fig. 9—Error and BPV test configurations: (a) error run, (b) BPV run, (c) BPV pretest.

from the bit generator was transmitted on the looped span (as in Fig. 9a). That is, errors would, or would not, occur on the looped span, depending on the signal fed into it.

Each continuous recording of errors or bipolar violations on a particular line or span is referred to as a "run." Normally the duration of a run is about 24 hours. Longer runs occurred when a run was successfully started on a Friday, since in that case the run was allowed to continue over the weekend. Shorter runs occurred when a line under test turned out to have few or no errors in a period of a few hours, or had an abnormally high error rate; the operator would then terminate the run and select another line for test. Each run was recorded as a distinct file on magnetic tape.

A plot of error rate versus time was derived for each run as a whole. For further analysis, "samples" of up to 1600 consecutive errors were extracted. At low error rates, a "sample" could include a whole 24-hour run. Where possible, separate samples were taken when the plot of error rate versus time showed different conditions occurring at different times. These samples were analyzed to study the "error-free interval" distri-

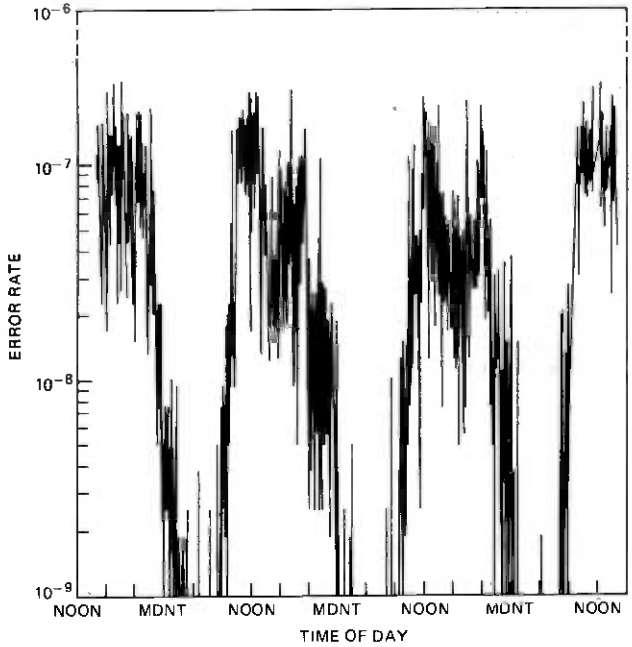


Fig. 10—Normal diurnal variation of error rate, recorded over a weekend, in an error run.

butions (distributions of intervals between successive errors) and the dependence of error probability on the content of the bit stream.

4.2 Diurnal variation of error rate

Since most of the error and BPV runs lasted about 24 hours, diurnal variation of error rate could be observed, but could not be entirely separated from long-term and short-term variation. Even when a run lasted over a weekend, diurnal components cannot be fully separated because a weekend does not consist of identical diurnal cycles. However, some distinct types of diurnal variation were identified.

In roughly one-third of the error and BPV runs of 1 day or longer, the error rate had a broad maximum during the working day (with short-term variations superimposed), falling off gradually during the evening to a minimum error rate at about 4:00 a.m., and rising again to its workday level about 8:00 a.m. Figure 10 shows such a pattern observed over a weekend. We refer to this pattern as "normal" diurnal variation.

Most of the remaining error runs, and a few of the BPV runs, showed essentially steady error rates over the day, as in Fig. 11. Irregular variations, on the other hand, were observed in most of the remaining BPV

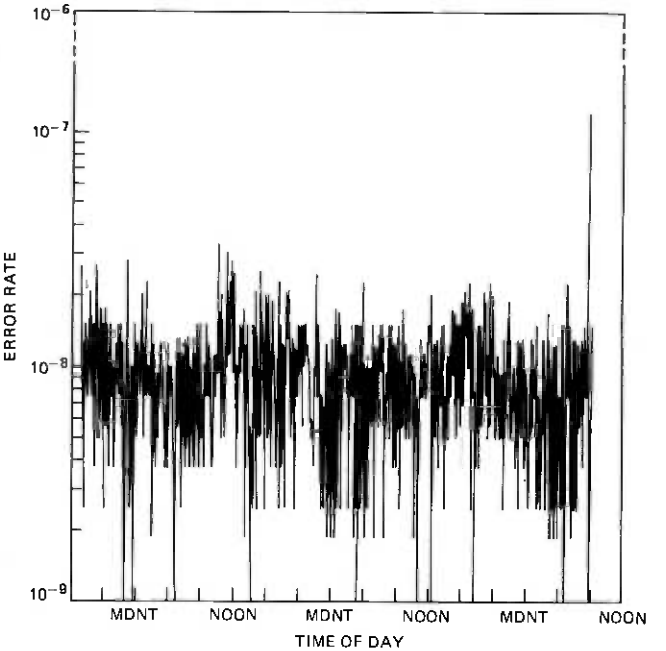


Fig. 11—Steady error rate, recorded over a weekend, in a BPV run. Each plotted point is computed from the number of errors in a cell about 5 minutes long, containing an average of about 5 errors; hence the point-to-point variation is entirely accounted for by assuming errors occurring independently at random (Poisson process), except at the very end of the run.

runs, and a few error runs. Fig. 12 shows distinctly irregular variation, with large nonperiodic changes. However, any variation that did not fit a recognizable pattern was classified as irregular.

The predominance of irregular variation in the BPV runs (as contrasted with steady error rates predominantly in the error runs) may be due to the variability in the content of the bit stream from the channel bank. As described in a later section, the lines with steady or irregularly varying error rates tended to be sensitive to the bit pattern on the lines. The invariably repeating pseudorandom sequence in the error runs would tend in such cases to give error rates that were constant with time; a pulse stream with variable content, such as the channel bank output, would result in varying error rates.

Two T1 lines (one BPV run, one error run) showed two distinct error rates during the diurnal cycle, with abrupt transitions from one error rate to the other, as in Fig. 13. This pattern has been classified as "exaggerated" diurnal variation.

Table III lists the total number of error and BPV runs with each type of diurnal variation. Only runs that actually extended overnight are

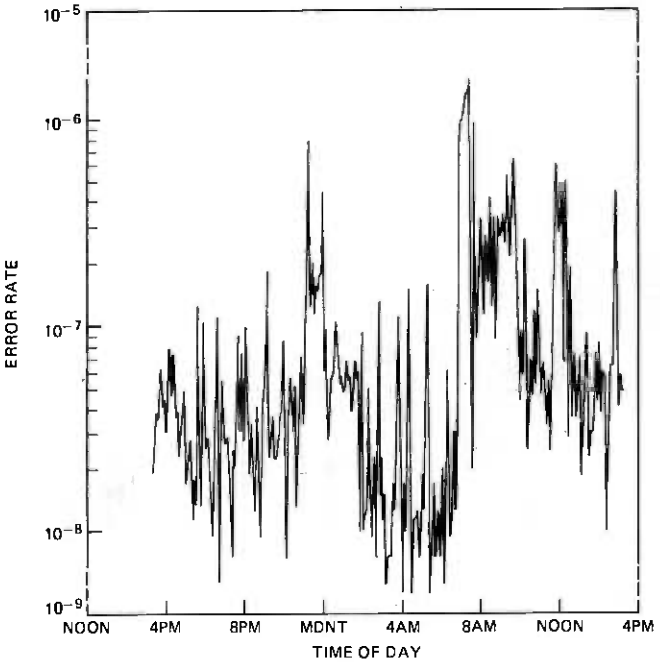


Fig. 12.—Irregular variation of error rate, recorded for 24 hours, in a BPV run.

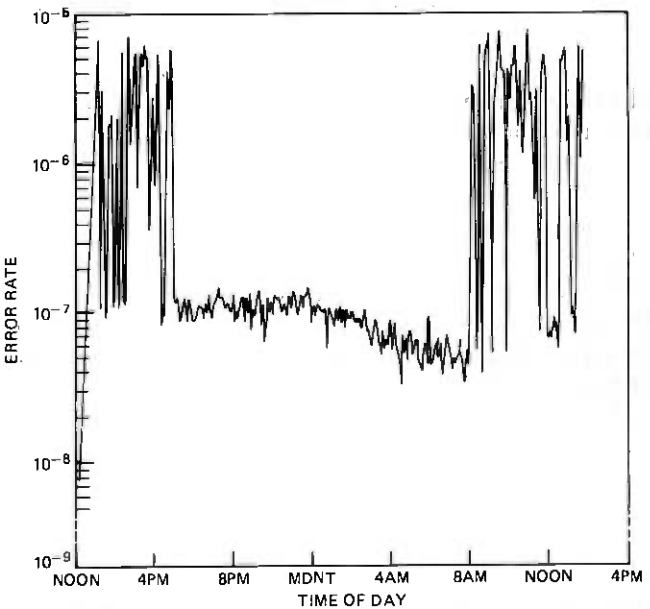


Fig. 13—Exaggerated diurnal variation, recorded for 24 hours, in a BPV run.

Table III — Frequency of occurrence of different types of diurnal variation

Variation type	Error-run lines	BPV-run lines	Total lines
Normal	6	4	10
Steady	6	1	7
Irregular	3	5	8
Exaggerated	1	1	2
Total	16	11	27

shown; shorter runs are excluded because the error rate minimum near 4:00 a.m. was an essential feature for identification of the "normal" type.

In the diurnal variation curves in Figs. 10 through 13, each point plotted represents the average error rate in a 5-minute cell. Since the cell boundaries were defined by the starting times of the data records on the tape, which occurred irregularly, the cells are not exactly 5 minutes long. However, the error rate for each cell is correct, being determined by dividing the number of errors by the actual number of bits in the cell.

4.3 Distribution of intervals between errors

Distributions of "error-free intervals" (the intervals between successive errors, in time units equal to the reciprocal of the bit rate) were plotted for every error sample. Three principal types were observed, identified as unimodal, bimodal, and trimodal. The same types were observed in both BPV runs and error runs. Table IV shows the number of lines observed with each type of distribution.

A typical unimodal distribution is shown in Fig. 14. This curve may be interpreted (except for the numbers on the vertical axis) as the probability density function of the logarithm of the interval, considering the interval as a continuous random variable. The curve was actually derived by setting up a logarithmic interval axis, dividing it uniformly into 40 cells per decade, and plotting the number of intervals that fell

Table IV — Frequency of occurrence of different types of distribution of the intervals between successive errors or BPVs

Distribution type	Error-run lines	BPV-run lines	Total lines
Unimodal	2	4	6
Nearly unimodal	6	5	11
Bimodal	5	2	7
Trimodal	6	3	9
Total	19	14	33

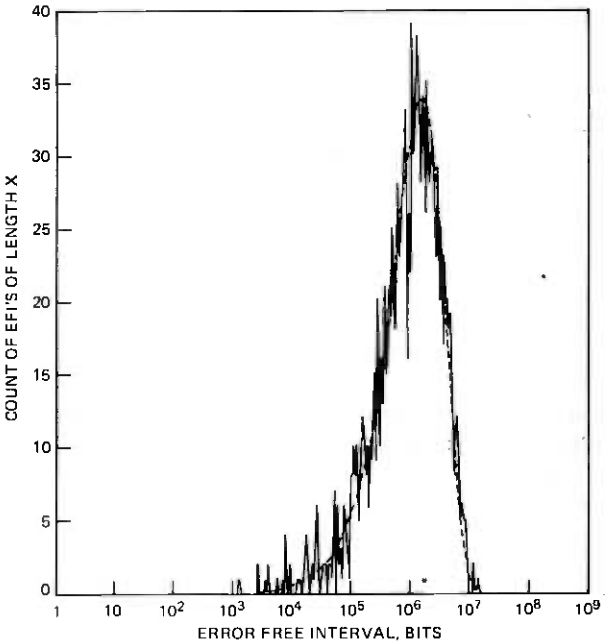


Fig. 14—Unimodal distribution of "error-free intervals," observed in a BPV run.

in each cell. The presence of only one mode indicates the absence of clustering. The asterisk just above the horizontal axis, in the vicinity of the mode, is plotted at the reciprocal of the mean error rate, that is, at the mean interval between errors.

Also shown in Fig. 14, as a dashed curve, is the probability density function of an exponential distribution with the same mean, scaled to fit the same logarithmic horizontal axis and to represent the same total number of errors on the vertical axis. The apparent fit suggests that the errors occurred independently (that is, as a Poisson process). As will be shown later, errors on this line were not independent, but depended strongly on the immediate pattern of bits on the line. However, there is a clear indication of large-scale independence: error-sensitive patterns occurred regularly, but the occurrence of an actual error in each pattern was independent of previous occurrences.

Fig. 15 shows a phenomenon that (fortunately for service) did not happen very often, or last very long: a very high error rate. The distribution shows a very low mean interval between errors. For the shortest intervals, the number of occurrences of each discrete interval are resolved in the plot. Clearly the errors are not completely independent (as in Bernoulli trials), because the count would then be highest at 1 and would decrease with increasing interval length. But the independent-error

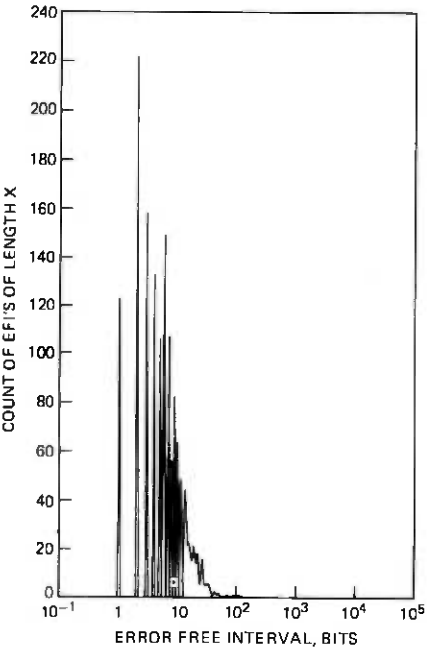


Fig. 15—Distribution of "error-free intervals" at very high error rate, observed in an error run.

model is conceptually useful as a point of departure. These distributions are not included in Table IV.

The bimodal distribution shown in Fig. 16 indicates error clustering. The mode at the right represents the intervals between clusters, and its shape again suggests large-scale independence (Poisson process) as in Fig. 14, but in this case also the occurrence of errors actually depended on local bit patterns. (The small amplitude of this mode may be deceptive; it actually includes about one-fifth of the total number of intervals.) The mode at the left represents the intervals between errors within a cluster, and, as in Fig. 15, errors are not quite independent within clusters. However, the general appearance is not very different from what one would expect from a combination of two exponential waiting-time distributions.

A substantial number of lines had interval distributions intermediate between the unimodal and bimodal types. In these lines, most of the intervals were grouped around a single mode at the right, with less than 10 percent of the intervals (sometimes only one) defining another mode at the left. These distributions are referred to as "nearly unimodal."

The trimodal distribution in Fig. 17 indicates two levels of clustering, that is, clusters within superclusters. The mode at the right represents

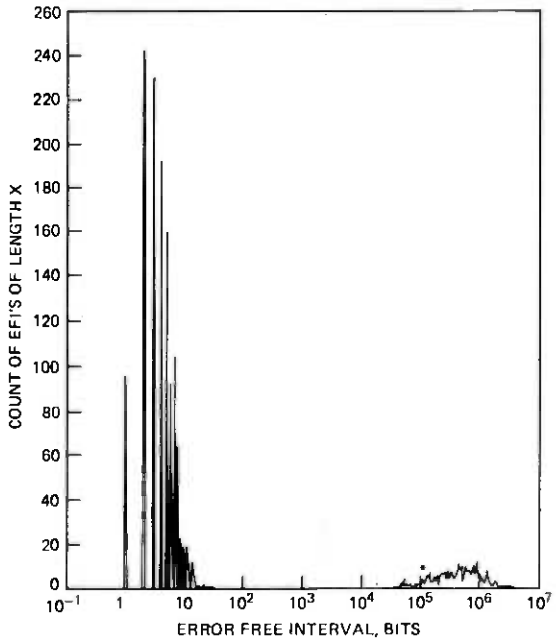


Fig. 16—Bimodal distribution of "error-free intervals," observed in an error run.

intervals between superclusters, the mode in the middle represents intervals between clusters within superclusters, and the mode on the left represents intervals between errors in a cluster. Actually, in this terminology, a "cluster" or "supercluster" could consist of a single error, since the average clustering factor was not large (usually about two to three errors per supercluster) and errors often occurred in isolation. On the other hand, it was not uncommon to find superclusters containing as many as 30 errors, occurring both alone and in clusters within the same supercluster.

Whenever a trimodal distribution was observed in an overnight run (permitting identification of the diurnal variation type), the line showed "normal" diurnal variation. Conversely, most of the "normal" diurnal variation runs showed trimodal interval distributions. In these cases, the interval distribution remained trimodal throughout the diurnal cycle. The middle mode, which indicates structure within the supercluster, remained usually in the same place, extending from about 100 to 500 bits, while the right mode, which described the occurrence of superclusters, moved left or right as the error rate on the line went up or down (respectively) with time of day.

A number of distributions classed as bimodal (or nearly unimodal) differed in appearance from Fig. 16 because of other features. In two

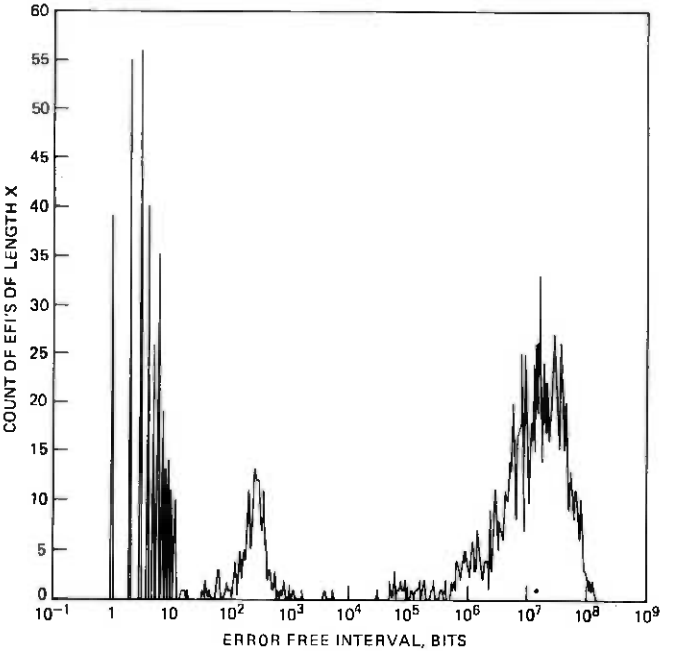


Fig. 17—Trimodal distribution of "error-free intervals," observed in a BPV run.

runs, discussed in Section 4.5, only certain values of short interval occurred. In a BPV pretest for an error run, multiples of about 193 (the D1 frame duration) predominated in the long intervals. In another sample, representing a burst of relatively high error rate in a BPV run, the long intervals are almost all multiples of about 20,000 bits ($1/75$ second). In most of these cases, longer and shorter intervals are interspersed, indicating error clustering. On the other hand, we obtained a sample from one error run that looks much like Fig. 11, but actually represents a transition from a low error rate, as in Fig. 14, to a very high error rate, as in Fig. 15; all the long intervals came first, followed by the short ones.

4.4 Effects of local bit patterns

On most lines, the probability of error depended on the local bit pattern being transmitted. The probability of error on any given line was usually not the same for ones as for zeros (bias effect), and also depended on the values of preceding and following bits (intersymbol interference effect). The pattern sensitivity varied greatly from line to line in both error and BPV runs. Pattern sensitivity was categorized as high, medium, or low for each line; Table V shows the number of lines observed in each category.

Table V — Frequency of occurrence of different levels of pattern sensitivity of probability of error or BPV

Pattern sensitivity	Error-run lines	BPV-run lines	Total lines
High (≥50.0)	7	3	10
Medium	0	8	8
Low (≤4.0)	7	3	10
Totals	14	14	28

Pattern sensitivity was directly observable in the error runs (except at office 2) because the bit stream on the line was a known pseudorandom sequence. (At office 2, a broken wire forced the second bit of every 16-bit data word on every tape to zero; the evaluation of error rate, diurnal variation, and the interval distribution was not seriously affected by the consequent loss of data, but pattern dependence could not be determined because identification of bit positions in the sequence was ambiguous.) In the error runs, pattern sensitivity fell clearly into two groups: very high and very low (except for short periods of very high error rate, which showed no discernible pattern sensitivity at all). In the BPV runs, evidence of pattern sensitivity appeared as a consequence of the periodic structure of the D1 channel bank frame. The bit stream is organized into frames of 193 bits at 8000 frames per second, in which one bit (the framing bit) is alternately one and zero in successive frames to enable the receiving channel bank to determine the frame phase. The remaining 192 bits consist of 24 words of 8 bits each, one word for each channel, in which one bit carries signaling and the other seven represent the voice by pulse code modulation (PCM). Pattern dependence could be inferred by relating the occurrence of bipolar violations to the periodicities of the D1 channel bank frame, and indications of pattern dependence presumably depended on the content of the channel bank output.

In the error runs, bias effects were characterized by identifying each error as an insertion (error in a zero) or a deletion (error in a one) and computing the percent of errors that were deletions. The percent deletions observed on 14 T1 lines varied from 0 to 100 percent, in a manner consistent with a uniform distribution.

Preliminary analysis suggested that intersymbol interference effects were usually confined to the bit following the error and the 6 bits preceding it. These bits, together with the bit in error, comprise an 8-bit string with the error in the seventh bit. For each error sample, a tally was then made of the number of times each of the 256 possible 8-bit strings occurred with an error in the seventh bit. Figs. 18 and 19 show the results of two such tallies, representing examples of low and high sensitivity respectively. (In these figures, the seventh bit is shown as transmitted; the slash through it indicates that this was not the value received.)

As a numerical measure of pattern sensitivity, the parameter

$$D = \frac{256}{N^2} \sum_{i=1}^{256} N_i^2 - \frac{255}{N} - 1$$

was computed for each error sample, where N_i is the number of times the ith 8-bit pattern occurred with an error in its seventh bit, and N is the total number of errors. This parameter is related to chi-square as it would be evaluated to test the null hypothesis of equal probabilities for all 256 patterns. However, while chi-square is useful primarily as a measure of statistical significance, D is designed to be nearly unaffected by the sample size for strongly pattern dependent errors, so that it measures pattern sensitivity as a property of the T1 line. The largest possible value of D, which would be attained if all errors occurred in the same pattern, is nearly 255 (actually, 255 − 255/N). If errors were independent of pattern, so that all 256 patterns were equally likely, D would have zero mean (based on a binomial distribution, $p = N/256$, for N_i) and a standard deviation of approximately $22.6/N$ (based on the related chi-square distribution, valid for large N, and verified by enumeration as roughly correct for $N = 2$ and $N = 3$).

The values of D fell into two groups: high (50 or above) and low (2.2 or below). The only exception is one sample that has an intermediate value because it spans a transition from one type to the other. However, in almost all cases the value of D is much larger than $22.6/N$, and hence significantly different from zero. The exceptions are of two types: small samples, and very high error rates (above 10^{-3}). Pattern sensitivity showed no consistent relation to percent deletions.

In the BPV runs, the bit patterns on the line were unknown. However, it could be presumed that some bit patterns might tend to recur periodically either at the 193-bit frame period or at the 8-bit word period within the frame. Accordingly, in each BPV sample each bit was assigned a frame position from 1 to 193, starting arbitrarily at the beginning of the sample, and a tally was made of the number of bipolar violations found in each of the 193 positions. These tallies were analyzed both numerically and graphically.

As a general numerical measure of pattern sensitivity for BPV runs, the parameter

$$D' = \frac{193}{N^2} \sum_{i=1}^{193} N_i^2 - \frac{192}{N} - 1$$

was computed for each BPV sample, where N_i is the number of BPVs that occurred in the ith position in the 193-bit frame. This parameter has similar properties to the parameter D computed for the error samples. Specifically, its largest possible value is $192(1 - 1/N)$, and its standard deviation, for random errors, is $19.6/N$.

```
27 01101081   11 11010181   8 01010080   6 00010081   5 01100081   4 01100170   3 11000081   2 11010171
26 10101080   11 11100080   8 01010081   6 00100081   5 01100170   4 01111081   3 11000170   2 11011181
25 01101080   10 00000181   8 01011181   6 00100180   5 01111080   4 10000080   3 11001070   2 11101070
25 10101081   10 00110081   8 01100180   6 00100181   5 10001170   4 10010170   3 11001071   2 11101071
24 11101080   10 01101170   8 01110081   6 00110180   5 10010081   4 10100081   3 11010071   2 11111171
22 01010181   10 10010171   8 01110180   6 00111080   5 10010170   4 10110171   3 11011171   1 00000171
20 11001081   10 10101180   8 01111171   6 01000081   5 10011081   4 10111081   3 11100070   1 00100170
20 11101081   10 10110071   8 10100070   6 01010181   5 10100080   4 11000071   3 11100071   1 01001070
19 00001071   10 10110180   8 10101171   6 01010170   5 10110080   4 11001180   3 11100170   1 01101070
19 00101080   10 10111080   8 10101170   6 01011071   5 10111171   4 11001171   2 00000080   1 01101071
19 00101081   10 11001170   8 10111181   6 01110080   5 11000080   4 11010080   2 00001070   1 01110070
19 10001081   10 11110080   8 11100081   6 10000081   5 11001070   4 11011180   2 00001071   1 10000071
18 00001170   9 00000071   8 11101180   6 10001181   5 11001180   4 11011170   2 00010070   1 10001170
17 00001080   9 00011081   8 11110181   6 10010180   5 11001181   4 11111071   2 00100171   1 10001070
17 10000181   9 00100070   7 00000071   6 10100171   5 11010081   3 00001181   2 00101070   1 10001071
15 10001080   9 00101180   7 00010080   6 10111170   5 11010070   3 00010171   2 00110070   1 10010071
15 10001171   9 00110080   7 00010180   6 11010180   5 11011070   3 00100080   2 00110181   1 10101070
14 01011080   9 00111080   7 00011070   6 11100180   5 11111171   3 00110071   2 00111070   1 10111070
14 11101171   9 01001171   7 00111170   6 11101181   4 00000170   3 01000071   2 00111080   1 11011071
13 11011080   9 01011180   7 01000180   6 11110071   4 00010071   3 01000171   2 01001071   1 11110070
13 11111081   9 01100181   7 01000170   6 11111181   4 00011071   3 01010070   2 01010071   1 11110180
12 00101170   9 01101180   7 01110181   5 00000070   4 00011181   3 01011071   2 01011070   1 11111180
12 00101171   9 10011080   7 01111180   5 00001180   4 00011170   3 01011171   2 01101181
12 01101171   9 10011181   7 01111171   5 00010181   4 00100071   3 01100171   2 01110071
12 10011171   9 10100180   7 10001180   5 00011080   4 00110171   3 01110170   2 01110171
12 11001080   9 10100181   7 10011170   5 00111071   4 00111081   3 10000180   2 01111070
12 11101170   9 10110181   7 10100071   5 00111171   4 00111181   3 10010080   2 01111071
11 00001171   9 11011071   7 11100181   5 01000080   4 01000070   3 10011180   2 10000070
11 01001080   8 00011180   7 11110081   5 01001181   4 01001180   3 10110071   2 10000171
11 10101171   8 00011171   7 11110170   5 01001170   4 01010180   3 10110170   2 10010171
11 11001181   8 00101171   7 11111080   5 01010171   4 01100080   3 10111071   2 10101071
11 11001171   8 01000181   6 00000180   5 01011170   4 01100171   3 10111180   2 11010170
```

Fig. 18—Computer printout of occurrences of errors within 8-bit patterns showing low pattern sensitivity in an error run ($D = 0.5$).

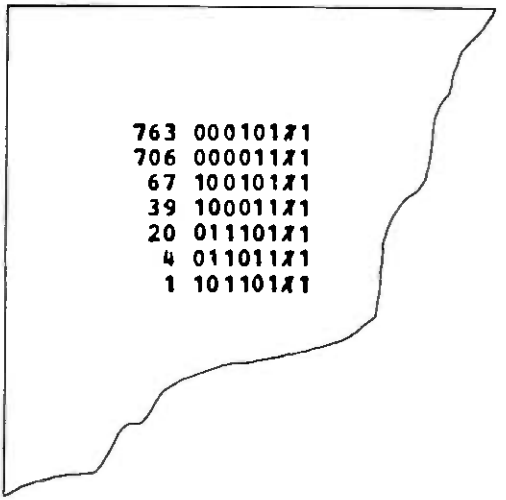


Fig. 19—Like Fig. 18, but showing high pattern sensitivity (D = 107.5).

The values of D' are significantly different from zero in most cases, and vary as widely as the values of D for the error samples. However, they do not fall clearly into two distinct groups; about half fell into the "medium" range, between 2.2 and 50, in which D never fell. This may be explained by the fact that dependence of BPV rate on frame position depends not only on the error mechanism on the T1 line but also on the properties of the transmitted bit stream. The presence of frame periodicity in the bipolar violations actually depends on two factors: the probability of a bipolar violation must depend on the bit pattern, and recurrent patterns must exist in the D1 bank frame.

It should be noted that bipolar violations are necessarily pattern-dependent, even if the errors causing them are not, because a bipolar violation can be detected only when a one is received. In the D1 channel bank format, some frame positions are more likely to contain ones than others; hence, bipolar violations will always be frame-position dependent to some extent. This factor may account for some of the observed frame-position dependence where this dependence is weak. However, it cannot account for the observed cases of strong frame-position dependence, because there are constraints in the channel bank that prevent long strings of zeros (as described in Section 4.5).

Figures 20 through 22 respectively show examples of weak (D' = 0.4), medium (D' = 11.7), and strong (D' = 108.5) dependence of bipolar violations on bit position in the frame. These figures represent 3-dimensional bar graphs, the height of each bar representing the number of bipolar violations at each frame position. The lower left corner represents the first bit in the frame, the first 8 bits are laid out from left to

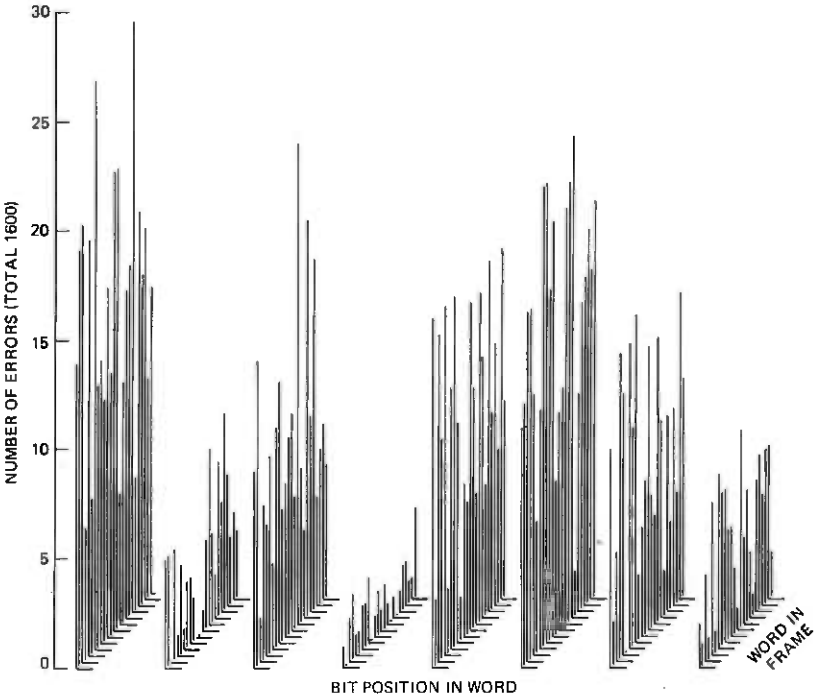


Fig. 20—Three-dimensional bar graph showing weak dependence of bipolar violation probability on position in the D1 bank frame in a BPV run ($D' = 0.4$). Frame alignment is hypothetical.

right, successive 8-bit words are laid out behind the first, and the 193rd bit is in the far left corner of the base of the diagram. Each row running from front to back thus represents a given bit position in each word. In these figures, the hypothetical starting point of the frame was shifted to the position that gave the figure the most regular appearance (by maximizing a parameter similar to D', but based on the number of bipolar violations in each position in the word), on the presumption that such a choice probably approximated alignment with the actual channel bank frame being transmitted.

Figure 20 shows an example of relatively weak position dependence; bipolar violations occur in all positions, but their probability clearly depends on the position of the bit in the word. An example of medium position dependence is shown in Fig. 21. Fig. 22, where most of the bipolar violations occurred in one position in the frame, shows strong position dependence. Fig. 22 is an extreme case; the same line at another time of day showed an intermediate pattern in which no single bit position had a majority of the BPVs. Figure 22 necessarily indicates high sensitivity to some recurrent pattern. Figures 20 and 21, however, could

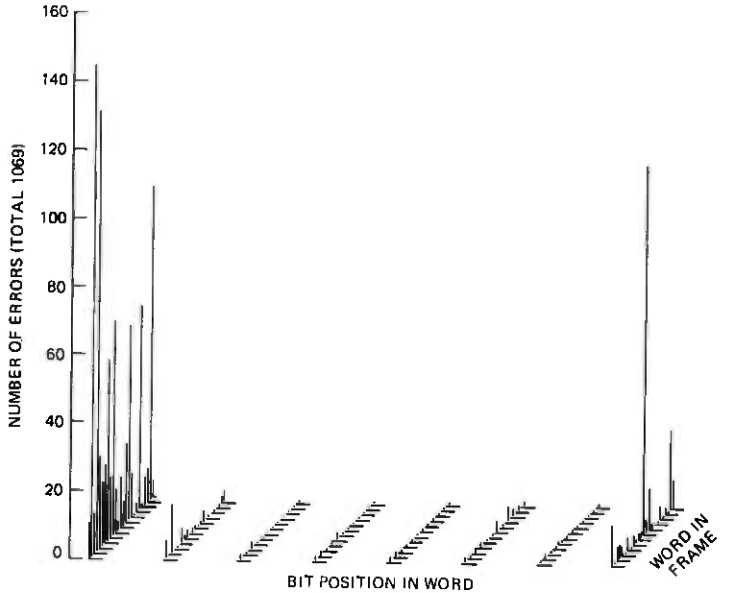


Fig. 21—Like Fig. 20, but showing medium position dependence ($D' = 11.7$).

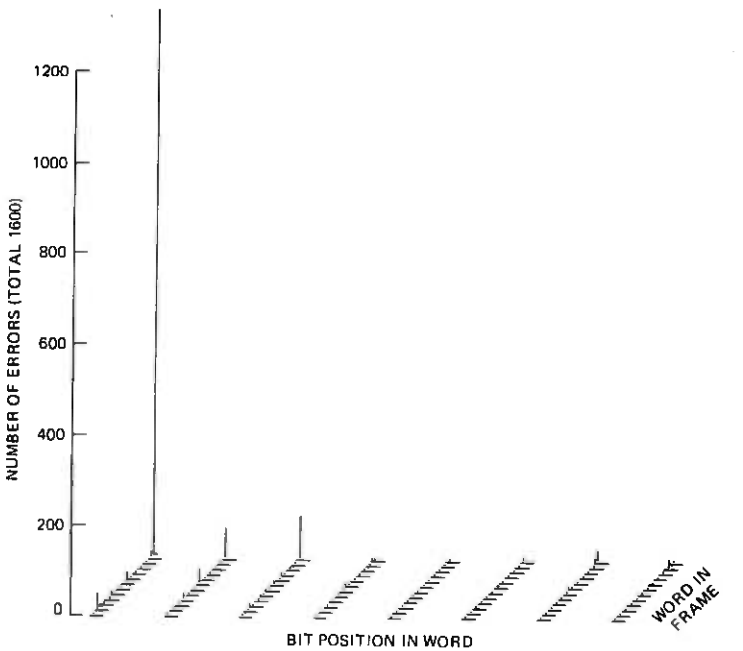


Fig. 22—Like Fig. 20, but showing strong position dependence ($D' = 108.5$).

be explained either by lower pattern sensitivity (possibly none at all), or by high sensitivity to patterns that occur less regularly. Correlation with other properties of the error process, discussed in Section 4.6, suggests that the former explanation is applicable to cases such as Fig. 20, and the latter to cases such as Fig. 21; but such conjectures are speculative.

4.5 Effects of density of ones in the bit stream

The pseudorandom sequence does not meet the constraints on minimum ones density for bit streams transmitted on T1 lines. These constraints are intended to ensure that there is always sufficient energy in the timing tank circuit in each T1 line repeater. This tank circuit gets an input pulse whenever a one is received, and if the density of ones is too low the repeater timing becomes inaccurate and errors are more likely to occur.

It was, therefore, expected that any given T1 line, driven by the pseudorandom sequence generator, would be more prone to make errors than when driven by a D1 channel bank, which does meet these constraints. This expectation was not fulfilled. On different lines, the error rate for the pseudorandom sequence might be greater than, equal to, or less than the error rate for the D1 bank output. In most cases, errors occurred only when the signal source was a channel bank; sometimes, however, errors occurred only when the signal was supplied by the pseudorandom generator. This does not necessarily mean that T1 lines in general have higher error probability when driven by a D1 channel bank, because lines were usually selected for test on the basis of errors observed with a channel bank as the signal source, so that the sample of lines is biased.

In addition, it was expected that errors would be more likely to occur in those parts of the pseudorandom sequence where the constraints were most severely violated. As described below, two of the lines tested showed this tendency to a remarkable degree, but the rest showed no such tendency at all.

The pseudorandom sequence differs from the output of a channel bank both in maximum length of strings of zeros and in average ones density. The D1 channel bank must have a one in each word; the longest string of zeros that it can generate is 15. The pseudorandom sequence, on the other hand, contains eight strings of more than 15 zeros, including one of 19 zeros, in each repeated frame of a million bits (actually 1,048,575 bits, about 0.68 seconds).

The average ones density is nominally one-half for both, but in the channel bank output it varies from one bank to another, while in the pseudorandom sequence it varies over the million-bit frame. In the D1 channel bank, the ones density depends on how the analog-to-digital

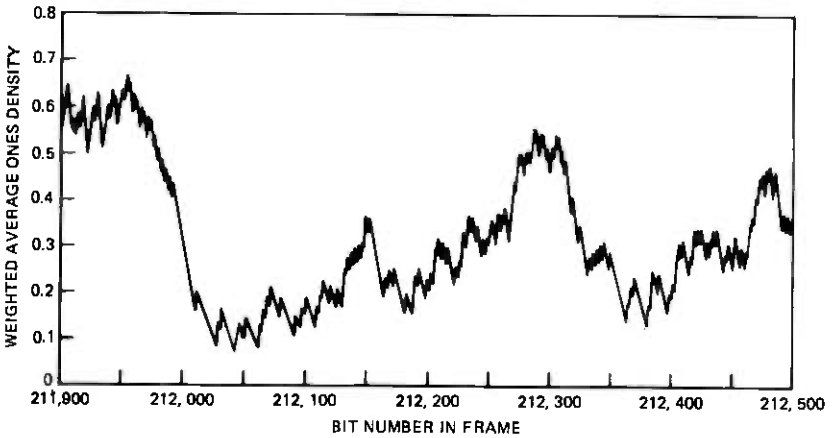


Fig. 23—Effective ones density for a timing tank with $Q = 60$, as a function of bit position in the pseudorandom sequence, in the vicinity of the string of 19 zeros, showing the location of the lowest effective density of ones in the sequence.

encoders (of which there are two in each bank) are adjusted to encode the zero level. The intended zero code is 1000000, which leads to a low ones density. However, with a slight misadjustment, zero might be encoded as the next lower code level, 0111111, which leads to a high ones density. The ones density is also affected by the signaling bit (a one for the on-hook condition) and by the encoding of noise and speech. One-second measurements on D1 channel banks in service have shown ones densities ranging from about one-fourth to three-fourths, with little variation over the day for any given bank.[10] Other observations have shown ones densities as low as about one-eighth on T1 lines in service.[11]

In the pseudorandom sequence, the average ones density over the million-bit frame is almost exactly one-half (actually 524,288/1,048,575). The variation of ones density within the frame can be evaluated as a weighted average of the past bits, using exponential weighting with a time constant of Q/π bits (about 32 bits for the typical timing-tank Q of 100). This average is theoretically proportional to the amplitude of ringing in a repeater timing tank circuit having the specified Q.

Figures 23 through 25 show, for three values of Q, the variations of ones density following the string of 19 zeros. (Since the ones density increases at each one and decreases at each zero, the strings of zeros can be identified by their downward slope. The string of 19 zeros starts at the 211,993rd bit after the framing bit. The framing bit itself was the last in the string of 20 consecutive ones.) The ones density has an absolute minimum of about one-fourth in this segment, but the exact value and location of the minimum depends on the timing tank Q.

Only two of the T1 lines that were tested had a noticeable tendency

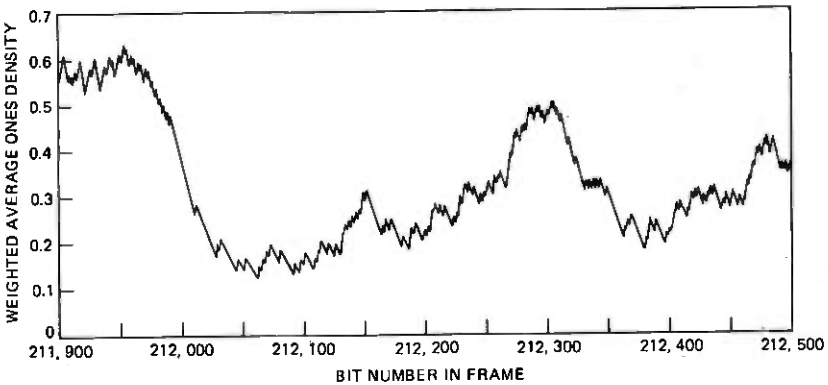


Fig. 24—Like Fig. 23, but for timing tank $Q = 100$. Note the change in the position of the minimum effective ones density as well as the change in minimum value.

to make errors in this part of the frame. One of these lines had an error rate of about 4×10^{-8} for the channel bank pulse stream (55-minute test), but had a steady error rate of only 3×10^{-9} for the pseudorandom sequence. On this line, errors occurred most often on the sixth one after the run of 19 zeros, on the first or second zero after that, or the eleventh one after the 19-zero sequence. These would be the locations of the lowest timing tank amplitudes assuming a Q of about 100, as in Fig. 24.

The other line, recorded at office 1, was a maintenance spare modified (as a stress test) by changing the line buildout (LBO: a circuit installed to equalize the loss of repeater sections of different lengths) in the office repeater. As installed, with an 836A LBO, the line made no errors. Substitutions of 836B through 836E had no apparent effect, an 836F resulted in an error rate of 2.5×10^{-6} (several errors per frame), and with an 936G the errors occurred too fast to be recorded. With the 836F LBO, errors

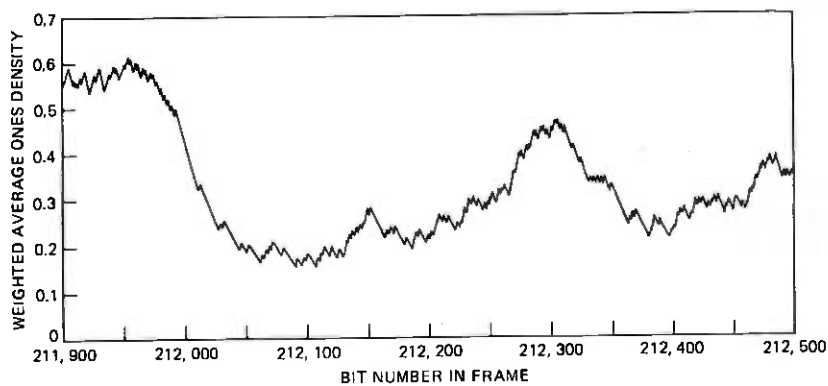


Fig. 25—Like Fig. 23, but for timing tank $Q = 140$.

Table VI — Frequency of occurrence of combinations of characteristics of errors of BPVs on 22 T1 lines*

Distribution type	Diurnal variation type			
	Normal	Steady	Irregular	Exaggerated
Unimodal	*m*	*Hh*	*m*	*h*
Nearly unimodal	*m*	*HH*	*Lmmm*	
Bimodal		*H*	*m*	*L*
Trimodal	*LLLLLl*			

* The letter indicates run type and pattern sensitivity: capital = error run, lower case = BPV run; pattern sensitivity is high (*H*/*h*), medium (*m*), or low (*L*/*l*).

occurred most often on the fourth and sixth ones after the string of 19 zeros, and sometimes on the two zeros after the seventh one. The fourth and sixth ones would just follow the two lowest values of effective ones density if the timing tank Q were about 60, as in Fig. 23.

4.6 Association between properties of the error process

Table VI suggests that the different characteristics of the error process tended to occur in certain typical combinations. The Appendix describes how the statistical significance of the apparent tendencies was verified. The most consistent combination was normal diurnal variation with a trimodal interval distribution (remaining trimodal as the error rate varied) and weak pattern dependence. Every line with a trimodal interval distribution had normal diurnal variation. Conversely, nearly every line with normal diurnal variation had a trimodal interval distribution. Every line with a trimodal interval distribution also had weak pattern dependence. This combination occurred in both error and BPV runs.

Another frequent combination was steady error rate with strong pattern dependence and a unimodal or nearly unimodal interval distribution. Every line with a steady error rate had strong pattern dependence. Every line with strong pattern dependence (except the two lines that made errors only at the minimum of ones density in the pseudorandom sequence) had a unimodal or nearly unimodal interval distribution.

Other combinations were less consistent. A substantial number of lines had irregular variation of error rate, most of them in BPV runs; these lines had either unimodal or bimodal interval distributions and usually showed an intermediate degree of pattern dependence. Both the irregular variation and the ambiguous indication of pattern dependence might be explained by the variability of the channel bank output bit stream.

These observations indicate that at least two different error mechanisms can be observed. In one, the error rate varies with traffic, but depends very little on the bit pattern on the line, and two levels of error clustering occur as indicated by the trimodal interval distribution. In

the other, the probability of error is very much dependent on the pattern of bits on the line, the error rate remains steady as long as the same bit stream is repetitively transmitted, and there is little or no error clustering. It might be conjectured that errors are related to switching noise in the former type, and to intersymbol interference in the latter.

Occasional short intervals of very high error rate occurred in several lines. Samples during these intervals usually showed measured error rates above 10^{-3}, and sometimes above 0.1, with little or no pattern dependence, and about 50 percent deletions in the error runs. It is clear that in these events the errors were quite unaffected by the pattern of bits on the line, but the cause is unknown.

The great difference between the error rate in the pseudorandom sequence and the error rate in the channel bank bit stream is unexplained. It probably is not due to short term effects of ones density (on the time scale of variations in the timing tank output), as discussed in Section 4.5. On lines that show strong pattern dependence, BPV runs and error runs might be expected to give different results. But it is not clear why a line can show BPVs in service with very little pattern dependence (evidenced both directly by the distribution of errors over the frame, and indirectly by normal diurnal variation), and yet transmit the pseudo-random sequence without error.

4.7 Practical implications of the detailed recording results

The clearest general implication that can be drawn from the detailed error recordings is that error rate measurements on digital facilities should be interpreted with caution. For many possible reasons, the error rate measured on a T1 line may be quite different from the error rate experienced when the line is used for communication.

An error rate measurement made at night, especially in the early morning hours, can be misleading because many lines consistently have much lower error rates at that time than during the business day. Furthermore, since some lines have irregularly varying error rates, a single measurement during the business day would not necessarily show the typical performance of a line. It would seem that continuous monitoring, or at least frequent sampling, would be required to evaluate the error rate performance of an individual line.

It may also be misleading to test a line with a special test signal, or with any signal that is different from the communication signal that it normally carries. The measurement program showed conclusively that on many lines the error rate for a pseudorandom test sequence is quite different from the error rate for the D1 channel bank output carried by the line in service. Hence the error rates for a digital data signal, or for another channel bank output (either a different type, or a different unit

of the same type), might be different from the error rates measured either with the line in service, or driven by a "quasirandom signal source" (a single generator used in telephone central offices, which generates a sequence similar to our pseudorandom sequence, except that ones are inserted to avoid long runs of zeros). However, the properties of the signal that affect the error rate remain unidentified.

Other implications are less clear. The relationships observed among diurnal variation, pattern sensitivity, and error-free internal distribution, suggest the existence of at least two clearly distinct causes of errors on T1 lines. It has been suggested that these are the same as the two major sources of error considered theoretically by Cravis and Crater:[2] The errors characterized by normal diurnal variation and low pattern dependence would be attributed to impulse noise originating in switching machines, which can be strong enough, when it occurs, to cause errors regardless of the bit pattern; the errors characterized by high pattern dependence would be attributed to crosstalk, which would cause errors mainly in those patterns in which intersymbol interference was most severe. This appears plausible as a tentative hypothesis. However, no attempt was made during the measurement program to identify the causes of the errors on individual lines, because of the extensive effort that would have been involved.

V. CONCLUSIONS

The T1 error measurement program of 1973–74 has resulted in both a survey of the population of T1 lines and some detailed observations of the error process.

The survey is in some ways less detailed, less systematic, and less accurate than previous (unpublished) surveys, but is probably overall the best overview we have of the error performance of T1 lines (when terminated predominantly with D1 channel banks). The large differences between offices show that a trend in the distribution of error rates on T1 lines, as a function of either time, or growth over time, cannot be inferred from comparison of surveys taken at different offices at different times. In other respects our results are consistent with previous surveys. Detailed conclusions were discussed in Section 3.3.

The detailed error recordings have shown a few remarkable results. The fact that many T1 lines can make errors in service without making any errors on a pseudorandom signal is both unexpected and unexplained. Another notable result is the observable existence of at least two different patterns of errors: one with a seemingly traffic-related diurnal variation, two levels of error clustering, and less sensitivity to the bit patterns on the line; the other with simpler patterns of variation and clustering, but more sensitivity to bit patterns. Another result is the fact that errors are always dependent to some extent on the bit patterns

on the line, except at very high error rates. Detailed results of the intensive error measurements were discussed in Section 4.6.

VI. ACKNOWLEDGMENTS

The author entered the jitter and error measurement experiment after it began, when the measurements at office 2 were being made. The error measurement program was begun by C. A. Richardson, who designed the experiment, built the error measurement system (and the 16-line "BPV box" used for surveys after office 1) and produced the first graphical output and tentative conclusions from the data from office 1. P. C. Lopiparo and T. C. Spang worked with him and continued the experiment after that point. G. Zaccaria was responsible for the real-time monitor software for the PDP-11 computer. B. R. Wittig programmed the analysis of the survey data. A. K. Jain provided suggestions and critical comments on the statistical evaluation of the detailed recording results. The cooperation of the many telephone operating company personnel who worked with us at the test sites, and of the many Bell Laboratories members who operated the equipment, is gratefully appreciated.

APPENDIX

Statistical Significance of Apparent Association of Error Characteristics

The statistical significance of the frequencies of particular combinations in Table VI was verified by a test based on the chi-square test for independence in contingency tables. In order to conclude that Table VI shows an actual tendency toward certain combinations we had to determine that such an apparent tendency could not easily have occurred by chance. Two problems were met in the testing procedure. First, the null hypothesis to be tested must not be the hypothesis that all four dimensions in Table VI are independent, because the apparent tendency for certain characteristics to be associated with BPV runs is explainable as a property of the measurement technique; the test must allow for this and determine whether any further interdependence can be deduced from the data. Second, the sample size is so small, and the number of cells so large, that the chi-square distribution is not applicable.

The first problem was not to be solved by simply combining the error runs with the BPV runs, because this would leave medium pattern sensitivity apparently associated with irregular diurnal variation, actually because both are associated with BPV runs. Instead, the dependence of pattern sensitivity on run type was allowed for by considering the five observed combinations of these variables as a single variable. The dependence of diurnal variation on run type was then removed from the table by combining the steady and irregular types into one. The resulting 3-way contingency table is shown in Table VII.

Table VII — Three-way contingency table derived from Table VI and used in statistical testing of the significance of the results*

Run type and pattern sensitivity	Distribution type					Estimated probability
	Unimodal	Nearly unimodal	Bimodal	Trimodal	Total	
error, high	1*V*	2*V*	1*V*	0	4	0.182
error, low	0	1*V*	1*X*	5*N*	7	0.318
BPV, high	1*X*, 1*V*	0	0	0	2	0.091
BPV, medium	1*N*, 1*V*	1*N*, 3*V*	1*V*	0	7	0.318
BPV, low	0	0	0	2*N*	2	0.091
Total	5	7	3	7	22	
Estimated probability	0.227	0.318	0.136	0.318		1.0
Total of $N = 9$, $V = 11$, $X = 2$						
Estimated probability of $N = .409$, $V = .500$, $X = .091$						

* N = normal diurnal variation, V = variable (steady or irregular), X = exaggerated.

The estimated probabilities for each type in Table VII were set equal to the relative frequencies. As in an ordinary chi-square test for independence, the probability of each of the 60 possible combinations of types was derived (by a computer program) by multiplying the corresponding type probabilities. For example, the probability that a line would combine unimodal distribution type ($P = 0.2$), normal diurnal variation ($P = 0.45$), and "error high" run type and pattern sensitivity ($P = 0.2$), is $0.2 \times 0.2 \times 0.45 = 0.018$. The expected number of occurrences of each combination, in 22 lines, is 22 times the corresponding probability. The statistic called chi-square, measuring the deviation of the observed numbers, o_i, from the expected numbers, e_i, is derived as

$$\chi^2 = \sum_{i=1}^{60} (e_i - o_i)^2/e_i$$

For the actual observed numbers, chi-square was 88.04.

In a large sample this value would be compared with critical values based on a chi-square distribution with 50 degrees of freedom (the number of combinations of types, minus one, minus the number of independent probability values—not counting the ones determined by the constraint that probabilities must add up to 1—that were estimated from the observed data). If we did this we would conclude that a value of chi-square as large as 88.04 would occur by chance, if the three properties were independent, with a probability of about 0.01 percent, and that the deviation was therefore certainly significant. However, this reasoning is not valid, because with a sample as small as 22 the chi-square statistic does not have, even approximately, a chi-square distribution.

A Monte Carlo method was therefore used to estimate the relevant distribution. The model that was simulated on a computer considered

three independent properties (diurnal variation, waiting time distribution, and run-type combined with pattern sensitivity), each property having 22 perdetermined outcomes distributed according to the type totals shown in Table VII. To implement each trial, a random-number generator assigned one of the 22 outcomes for each property (by drawing without replacement) to each of 22 lines. In 10,000 trials, 62 trials had values of the chi-square statistic greater than 88.04.

We could thus estimate that, under the null hypothesis, the probability that the chi-square statistic would exceed 88.04 is 0.0062, but this estimate has some uncertainty that we cannot easily estimate or allow for. The result, therefore, was interpreted by the following reasoning. If the null hypothesis is true—that is, if the three properties examined are independent—the single field experiment and the 10,000 Monte Carlo trials are replications of the same experiment, and the value of chi-square for the field experiment was among the 63 highest values in 10,001 trials. This would be an event with probability 0.0063, or 0.63 percent. This is small enough to consider that the deviation from independence is statistically significant.

A similar procedure was followed with another statistic, $\max |o_i - e_i|$. This had a value of 4.09 for the field trial, which was equaled in 3 out of 10,000 computer trials (and never exceeded). Under the null hypothesis this would be an event with probability 0.0004, or 0.04 percent, showing even more clearly a significant deviation from independence.

Having concluded that independence of the properties does not account satisfactorily for the observations, we are then justified (by Occam's razor) in adopting the simplest hypothesis that does account for them, which is that we have observed two different types of error process, each having different probabilities for the properties.

REFERENCES

1. K. E. Fultz and D. B. Penick, "The T1 Carrier System," B.S.T.J., *44*, No. 7 (September 1965), pp. 1405–1451.
2. H. Cravis and T. V. Crater, "Engineering of T1 Carrier System Repeatered Lines," B.S.T.J., *42*, No. 2 (March 1963), pp. 431–486.
3. J. S. Mayo, "A Bipolar Repeater for Pulse Code Modulation Signals," B.S.T.J., *41*, No. 1 (January 1962), pp. 25–97.
4. P. C. Lopiparo, to be prepared.
5. J. J. Mahoney, Jr., J. J. Mansell, and R. C. Matlack, "Digital Data System: User's View of the Network," B.S.T.J., *54*, No. 5 (May–June, 1975), pp. 833–844.
6. A. K. Reilly, unpublished work.
7. J. E. Kessler, unpublished work.
8. E. W. Geer, Jr., unpublished work.
9. S. W. Golomb, *Shift Register Sequences*, San Francisco: Holden-Day, 1967.
10. H. Hagen, unpublished work.
11. R. F. Ewald, unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 3, March 1978
Printed in U.S.A.

On the Stability of Higher Order Digital Filters Which Use Saturation Arithmetic

By J. E. MAZO

(Manuscript received July 15, 1977)

The device of using "saturation arithmetic" to cope with adder overflow in recursive digital filters has, for a number of years now, been known to yield stable operation when the filter is of second order and is linearly stable. Mitra has recently given examples to show that this happy situation does not prevail for higher order filters. Here we investigate conditions on the filter coefficients which would guarantee stability for higher order filters using saturation arithmetic. We are only able to give sufficient conditions for stability. These conditions in their simplest form can be written as linear inequalities involving the coefficients of the filter.

I. INTRODUCTION AND SUMMARY

We shall be concerned with real nth order nonlinear difference equations of the form

$$y(k+n) = f\left[\sum_{i=1}^{n} a_i y(k+n-i)\right], \quad k = 0, 1, 2, \ldots \tag{1}$$

The variables $y(\cdot)$ and the coefficients a_i are real. The initial conditions $y(j)$, $j = 0, 1, \ldots, (n-1)$ are arbitrary, subject only to the important condition $|y(j)| \leq 1$. The function $f(\cdot)$ will be assumed here to have the form given in eq. (2):

$$\begin{aligned} f(x) &= x, \ |x| \leq 1 \\ f(x) &= \operatorname{sgn} x, \ |x| > 1 \end{aligned} \tag{2}$$

This function models a method of handling overflow in the practical implementation of digital filters and in that literature is referred to as "saturation arithmetic." An important unsolved problem is the asymptotic stability of this undriven system. Specifically we would like

to describe the region in "a-space" (i.e. $\{a_i\}$, $i = 1, \ldots, n$) for which

$$\lim_{n \to \infty} y_n = 0$$

for any initial conditions. We always assume that the a_i's are already restricted so that linear stability holds. That is, if $f(\cdot)$ were replaced by the identity function the system would be stable. This is equivalent to all roots λ_i of the characteristic equation

$$c(z) \equiv z^n - \sum_{i=1}^{n} a_i z^{n-i} = 0 \tag{3}$$

satisfying $|\lambda_i| < 1$.* Since $\{a_i\}$ are real, the complex λ_i occur in conjugate pairs.

The case $n = 2$ has special importance to a certain strategy for implementing digital filters and has been considered earlier.[1-3] It was shown for this case that eq. (1) is stable for saturation arithmetic whenever the system is linearly stable. Recently Mitra[4] has shown by example that a result of this generality does not hold for any $n > 2$. This surprising development has regenerated the author's interest in the problem in its own right. In addition, direct implementation of digital filters of the form [eq. (1)] for $n > 2$ is now of interest, so questions of stability must be answered. Saturation arithmetic seems to be an experimentally favored procedure at the moment.

Our main results, Theorems I and II, provide sufficient conditions that a given set of coefficients $\{a_i\}_1^n$ yield a stable filter with saturation arithmetic. Both theorems require one to test if a pair of linear inequalities in a set of variables can be satisfied when the latter lie in a hypercube of (at most) dimension n. A finite algorithm which is sufficient to decide this question is given in Section IV. While we feel that Theorem II will give more powerful results (i.e., determine a larger stability region), Theorem I allows one to list a set of linear inequalities in the a_i, which, if any one is satisfied, would guarantee stability. This result is given as Corollary I.

II. SOME LINEAR AND NONLINEAR THEORY

If one were concerned with the linear version of eq. (1) [$f(\cdot)$ equal to the identity function], the solutions would be

$$y(k) = \sum_{i=1}^{n} k_i \lambda_i^k \quad k = 0, 1, \ldots \tag{4}$$

* Under this condition, stability of eq. (1) with $|y_i| \leq 1$ and

$$\sum_{1}^{m} |a_i| < 1$$

is trivial, since the system is then linear.

where the λ_i are the characteristic roots of eq. (3), assumed to be distinct in writing the solution eq. (4). The constants k_i are determined by initial conditions, or equivalently, the values initially stored in the registers. Suppose we wish to solve for the k_i in terms of $y(0), \ldots, y(n-1)$. This clearly requires the inversion of the matrix

$$V(\lambda) = \begin{bmatrix} 1 & \cdots & 1 \\ \lambda_1 & & \lambda_n \\ \lambda_1^2 & & \lambda_n^2 \\ \vdots & & \vdots \\ \lambda_1^{n-1} & & \lambda_n^{n-1} \end{bmatrix}$$

called a Vandermonde matrix. A brief discussion of these matrices is given in the Appendix, as well as some notation we shall use related to them.

The linear difference equation can also be written in matrix form if we introduce the n-vector†

$$Y(k) = \begin{pmatrix} y(k) \\ y(k+1) \\ \vdots \\ y(k+n-1) \end{pmatrix}, \quad k = 0, 1, 2, \ldots \tag{5}$$

or, in component form $y_i(k) = y(k+i-1), i = 1, \ldots, n$. The "time" argument is indicated by the discrete index k. The equation-of-motion is then

$$Y(k+1) = AY(k) \tag{6}$$

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ & & \vdots & & \\ 0 & 0 & \cdots & 0 & 1 \\ a_n & a_{n-1} & \cdots & a_2 & a_1 \end{bmatrix} \tag{7}$$

Expanding the determinant of the matrix $A - \lambda I$ (I being the identity matrix) by the last row, we obtain (except for a sign) the polynomial eq. (3), and, not surprisingly, the eigenvalues of A are the roots λ_i mentioned earlier.

If these eigenvalues are distinct, a well-known theorem of algebra guarantees that there exists a nonsingular matrix P such that

$$P^{-1}AP = \Lambda \tag{8}$$

where Λ is simply the diagonal matrix of eigenvalues of A.

† Vectors and matrices will be denoted by capital letters.

Theorem: If the eigenvalues of A are distinct, then

$$V^{-1}AV = \Lambda \tag{9}$$

where V is the Vandermonde matrix formed from the roots of the characteristic eq. (3).

The choice $P = V$ is not claimed to be unique.

Proof:

$$\sum_{j=1}^{n} a_{ij}v_{jk} = \sum_{j=1}^{n} a_{ij}\lambda_k^{j-1} = \begin{cases} \lambda_k^i & i \le n-1 \\ \sum_{j=1}^{n} a_{n+1-j}\lambda_k^{j-1} = \lambda_k^n & \text{for } i = n \end{cases}$$

The second line in the last member makes explicit use of the fact that λ_k is a root of the characteristic equation. Thus

$$\sum_{i=1}^{n} v_{\ell i}^{-1}\left(\sum_{j=1}^{n} a_{ij}v_{jk}\right) = \sum_{i=1}^{n} v_{\ell i}^{-1}(\lambda_k v_{ik}) = \lambda_k \delta_{\ell k}$$

as was to be shown.

One reason for wishing to diagonalize A in the linear case is the simple form that the equation of motion takes. If we multiply eq. (6) by V^{-1} and perform the standard trick of inserting $I = VV^{-1}$ after the A in eq. (6), we obtain

$$Z(k+1) = \Lambda Z(k) \tag{10}$$

where

$$Z(k) = V^{-1}Y(k) \tag{11}$$

Since Λ is diagonal, the solution for the ith component of Z is simply

$$z_i(k) = \lambda_i^k z_i(0) \tag{12}$$

Turning now to nonlinear problems, we wish to summarize some results from Liapunov stability theory,[7] without proofs, and without complete generality.*

We are concerned with an autonomous (time independent) nonlinear difference equation

$$Y(k+1) = F[Y(k)] \tag{13}$$

where F is a nonlinear (or linear) vector function of the vector $Y(k)$.

* A. N. Willson was the first to explicitly apply Liapunov theory to the present problem for $n = 2$.[3] He has also attacked other stability questions for $n = 2$ with these methods in Ref. 8.

If

$$\lim_{k \to \infty} Y(k) = 0$$

then the system is called asymptotically stable.

Theorem (*Liapunov*): If there exists a (strictly) positive definite quadratic form w[Y], such that, for any allowed n-vector Y, $w[F(Y)] - w[Y] < 0$ (strict inequality) then the system is asymptotically stable.†

In other words, if we can find a positive definite quadratic form (of the state variables of the system) which is always decreasing as the motion proceeds, then the motion must proceed to the origin. The function $w[\cdot]$ is called a Liapunov function.

In terms of $f(\cdot)$ and A the function $F(\ )$ is determined by

$$[F(Y)]_i = [AY]_i \qquad i = 1, \ldots, (n-1)$$

$$[F(Y)]_n = \begin{cases} [AY]_n, & \text{if } |[AY]_n| \leq 1 \\ \operatorname{sgn}[AY]_n, & \text{otherwise} \end{cases} \tag{14}$$

As a simple example of a Liapunov function consider the linear case with nondegenerate eigenvalues, and again set $Z = V^{-1}Y$. Choose

$$w = \sum_{i=1}^{n} |z_i|^2 \tag{15}$$

the z_i being regarded as functions of the $y(j)$, $j = 0, \ldots, n-1$. We know that when $Y \to AY$ we have $Z \to \Lambda Z$ and so

$$w \to \sum_{i=1}^{n} |\lambda_i|^2 |z_i|^2$$

Since we assume $|\lambda_i| < 1$, strict decrease of w is assured.

III. A SPECIAL LIAPUNOV FUNCTION

For the nonlinear problem eq. (1), we shall, for a first pass, choose a $w[\cdot]$ whose form is inspired by the one just described. Noting from the Appendix

$$[V_M^{-1}Y]_i = \frac{(-1)^i \upsilon^{(i)}(\lambda)}{\upsilon(\lambda)} \sum_{y=1}^{n} (-1)^j p_{n-j}^{(i)}(\lambda) y_j$$

$$= \frac{(-1)^{i+n} \upsilon^{(i)}(\lambda)}{\upsilon(\lambda)} \sum_{\ell=0}^{n-1} (-1)^\ell p_\ell^{(i)}(\lambda) y(n-\ell-1) \tag{16}$$

we shall single out the functionals

† For our problem any vector Y is allowed that has components $|y_i| \leq 1$. We identify Y with $Y(0)$ and so $y_i = y(i-1)$, $i = 1, \ldots, n$.

$$x_i = \sum_{\ell=0}^{n-1} (-1)^\ell p_\ell^{(i)}(\lambda) y_{n-\ell} = \sum_{\ell=0}^{n-1} (-1)^\ell p_\ell^{(i)}(\lambda) y(n-\ell-1) \quad (17)$$

for special attention. Clearly the equation of motion for the Z variables (10) implies that under the substitution $Y \to AY$ we have $x_i \to \lambda_i x_i$. We choose

$$w[Y] = \sum_{i=1}^{n} |x_i|^2 \quad (18)$$

where the x_i, via eq. (17), are regarded as functions of y_i, $i = 1, \ldots, n$, the components of Y. Of course we always have $|y_i| \leq 1$.

In order to investigate the consequences of the (sufficient) stability condition

$$w[F(Y)] - w[Y] < 0 \quad (19)$$

we note that under $Y \to AY$ we have $x_i \to x_i^{(L)}$ (L stands for linear) where

$$x_i^{(L)} = \sum_{\ell=0}^{n-1} (-1)^\ell p_\ell^{(i)}(\lambda) y(n-\ell) \quad (20)$$

and

$$y(n) = \sum_{\ell=1}^{n} a_\ell y(n-\ell) \quad (21)$$

Finally the nonlinear (NL) "version" of eq. (20) is

$$x_i^{(NL)} = \begin{cases} x_i^{(L)} \text{ if } |y(n)| \leq 1 \\ \operatorname{sgn} y(n) + \sum_{1}^{n-1} (-1)^\ell p_\ell^{(i)}(\lambda) y(n-\ell) \qquad \text{if } |y(n)| > 1 \end{cases} \quad (22)$$

where we made use of the definition of $F(\cdot)$, and the fact that $p_0^{(\ell)}(\cdot) = 1$. Then

$$w[F(Y)] \equiv \sum_{i=1}^{n} |x_i^{(\mathrm{NL})}|^2 \quad (23)$$

We have already noted that if $|y(n)| \leq 1$ we have a linear iteration $[F(Y) = AY]$ and

$$w[F(Y)] - w[Y] = w[AY] - w[Y]$$

$$= \sum_{i=1}^{n} |\lambda_i|^2 |x_i|^2 - \sum_{i=1}^{n} |x_i|^2 < 0 \quad (24)$$

Hence we will only be concerned with $x_i^{(\mathrm{NL})}$ when $|y(n)| > 1$. In this case,

$w[F(Y)]$ is a function of $y(1), y(2), \ldots, y(n-1)$, while the condition $y(n) > 1$ involves $y(0)$ as well.*

We shall not use the stability condition in the form of eq. (19), but instead note that since

$$\sum_{i=1}^{n} |x_i^{(L)}|^2 = \sum |\gamma_i|^2 |x_i|^2 < \sum |x_i|^2 \tag{25}$$

the condition

$$\sum_{i=1}^{n} |x_i^{(\mathrm{NL})}|^2 - \sum_{i=1}^{n} |x_i^{(L)}|^2 < 0 \text{ when } y(n) > 1 \tag{26}$$

is sufficient for stability. At the price of losing some power in the method we shall see momentarily that we have gained considerably in analytic simplicity. For convenience define

$$c_i = -\sum_{\ell=1}^{n-1} (-1)^{\ell} y(n-\ell) p_{\ell}^{(i)}(\lambda) \tag{27}$$

so that [when $y(n) > 1$]

$$\begin{aligned} x_i^{(L)} &= y(n) - c_i \\ x_i^{(\mathrm{NL})} &= 1 - c_i \end{aligned} \tag{28}$$

Then

$$\begin{aligned} \sum_{1}^{n} |x_i^{(L)}|^2 - \sum_{1}^{n} |x_i^{\mathrm{NL}}|^2 &= \sum_{i=1}^{n} (y(n)-1)(y(n)+1-c_i-c_i^*) \\ &= (y(n)-1)\left[n(y(n)+1) - 2\sum_{1}^{n} c_i\right] \end{aligned} \tag{29}$$

We have used the fact that complex λ_i occur in conjugate pairs and so

$$\sum_{i} p_{\ell}^{(i)}(\lambda) = \sum_{i} p_{\ell}^{(i)*}(\lambda)$$

Thus if the inequalities

$$\begin{aligned} 2\sum_{i=1}^{n} c_i - n(1+y(n)) &> 0 \\ y(n) &> 1 \end{aligned} \tag{30}$$

have no solution in the n-cube $|y(i)| \leq 1, i = 0, 1, \ldots, n-1$ the nonlinear equation (1) is stable. More explicitly by using the definition of the c_i [eq. (27)] Lemma I in the Appendix coupled with (67) to express $\Sigma_{i=1}^{n} c_i$ in terms of the a_i, and using finally eq. (21), we have:

* By symmetry of the problem, $|y(n)| > 1$ is here and henceforth replaced with $y(n) > 1$.

Theorem I: If the two inequalities

$$\sum_{\ell=1}^{n} (n - 2\ell)a_\ell y(n - \ell) \geq n \tag{31}$$

$$\sum_{\ell=1}^{n} a_\ell y(n - \ell) > 1, \tag{32}$$

cannot be simultaneously satisfied by some set of $y(i)$ in the cube $|y(i)| \leq 1$, $i = 0, 1, \ldots, n - 1$, then the system (1) is stable for saturation arithmetic.

We note that if the inequalities are satisfied, no conclusion is drawn.

A systematic algorithm for checking the above inequalities is given in the next section. Here we deduce the simple:

Corollary I: If the coefficients a_ℓ satisfy

$$\sum_{\ell=1}^{n} |a_\ell||n - 2\ell| < n \tag{33}$$

then eq. (1) is stable. If the coefficients a_1 satisfy

$$\sum_{\ell=1}^{n} |a_\ell||k - \ell| < k \tag{34}$$

for at least one k, $[n/2] < k \leq n$, eq. (1) is stable.*

The first inequality follows immediately from the first inequality in Theorem I plus the fact that $|y(i)| \leq 1$, all i. The remaining inequalities follow from the observation that in eq. (31), the coefficients of a_ℓ have the same sign for $\ell \leq [n/2]$, whereas for $\ell > [n/2]$ they have opposite signs. Thus if eqs. (31–32) have a simultaneous solution, so do

$$\sum_{1 \leq \ell \leq [n/2]} |a_\ell|(n - 2\ell) + \sum_{\ell > [n/2]}^{n} (n - 2\ell)a_\ell y(n - \ell) \geq n \tag{35}$$

$$\sum_{\ell \leq [n/2]} |a_\ell| + \sum_{\ell > [n/2]} a_\ell y(n - \ell) > 1 \tag{36}$$

where the above is obtained by setting $y(n - \ell) = \operatorname{sgn} a_\ell$, $1 \leq \ell \leq [n/2]$. If, for $k > [n/2]$, we multiply the second inequality by $(2k - n) > 0$ and add the result to the first we obtain

$$\sum_{\ell \leq [n/2]} |a_\ell|(k - \ell) + \sum_{\ell > [n/2]} (k - \ell)a_\ell y(n - \ell) \geq k \tag{37}$$

which, if $|y(i)| \leq 1$, cannot possibly be satisfied when

* The notation $[x]$ denotes the integer part of x.

$$\sum_{\ell=1}^{n} |a_\ell||k-\ell| < k \tag{38}$$

There is one important point to be noted here. Our results have only been proven for the case of nondegenerate eigenvalues. For degenerate eigenvalues, the Liapunov function that we have chosen in this section is not strictly positive definite. We have in fact constructed Liapunov functions which are specifically designed to handle the degenerate case. Using them, we have proven that the conclusions are still true for degenerate eigenvalues. We believe that the form of the results, being simple conditions on the a_l's, will allow the reader to readily accept that they are true in general. Since our proof of the extension is long and out of proportion to its importance, we have chosen to omit it.

IV. HYPERPLANE ALGORITHM

Let $\{b_i\}_{i=1}^k$ and $\{c_i\}_{i=1}^k$, ξ and η denote fixed constants. We wish to determine when it is possible to simultaneously satisfy the inequalities

$$\sum_{i=1}^{k} z_i b_i \geq \xi$$

$$\sum_{i=1}^{k} z_i c_i \geq \eta$$

$$|z_i| \leq 1 \quad i = 1, 2, \ldots, k \tag{39}$$

The dimensionality of the problem is immediately reduced if sgn b_j = sgn c_j for some j since we may immediately take z_j = sgn b_j. It is important to note that we assume this to have been done and therefore assume $b_i c_i < 0$, $1 \leq i \leq k$.

Lemma: If the simultaneous inequalities eq. (39) are satisfied then there exists $\tilde{z}_i$, and a j, $1 \leq j \leq k$, such that

$$\sum_{i=1}^{k} \tilde{z}_i b_i \geq \xi$$

$$\sum_{i=1}^{k} \tilde{z}_i c_i \geq \eta$$

$$|\tilde{z}_j| \leq 1, \ |\tilde{z}_\ell| = 1 \text{ all } \ell \neq j \tag{40}$$

In other words all but possibly one of the coordinates may be given values ± 1.

This is geometrically evident if $k = 2$. If $k > 2$ one need only consider the z_i variables two at a time, always applying the Lemma for $k = 2$. Eventually all but perhaps one of the z_i will have value ± 1.

Continuing with the description of the algorithm, let E be any k-vector

having components $\epsilon_\ell = \pm 1$; there are 2^k such E vectors. Choose a j and let's test if indeed it is the j of the Lemma. Let $\bar{z}_\ell = \epsilon_\ell$, $\ell \neq j$. Then if eq. (40) has this solution we have

$$\frac{\xi - \Sigma' b_\ell \epsilon_\ell}{b_j} \leq \bar{z}_j \leq \frac{\eta - \Sigma' \epsilon_\ell c_\ell}{c_j} \quad \text{if } b_j > 0 \tag{41a}$$

or

$$\frac{\eta - \Sigma' \epsilon_\ell c_\ell}{c_j} \leq \bar{z}_j \leq \frac{\xi - \Sigma' b_\ell \epsilon_\ell}{b_j} \quad \text{if } b_j < 0 \tag{41b}$$

If these inequalities are consistent (i.e., the upper bound is at least as big as the lower bound) and if they can be satisfied by some $\bar{z}_j$, $|\bar{z}_j| \leq 1$ we are done—the simultaneous inequalities are satisfied. If not, try another E vector, or another j. For a given j there are 2^{k-1} E vectors to try. Hence after at most $k \cdot 2^{k-1}$ such attempts we have exhausted all things that need to be checked, and checking the inequalities is, it has turned out, a finite procedure.

This procedure is not only applicable to Theorem I, but also to Theorem II occurring in Section V.

V. ANOTHER LIAPUNOV FUNCTION

We have already noted that the entire sequence $\{y_i\}_0^\infty$ is determined by the first n elements. For our second choice of the Liapunov function we chose the expression for the energy in the remainder of the sequence for the linear problem:

$$w[Y] = \sum_n^\infty y_k^2 \quad \text{(linear case)} \tag{42}$$

The right member is regarded as a positive definite quadratic form in $Y(0)$. In the linear case we also have, numerically,

$$w[AY] = \sum_{n+1}^\infty y_k^2 \tag{43}$$

which is smaller than $w[Y]$ by $y^2(n)$. Thus this $w[\cdot]$ doesn't necessarily decrease after every iteration and thus it is not strictly a Liapunov function. However after at most n iterations it must decrease (unless all $y_i = 0$) and the effect will be the same. We shall have stability if we can show whenever $y(n) > 1$, that

$$w[F(Y)] - w[Y] < 0 \tag{44}$$

or, equivalently

$$w[F(Y)] - w[AY] < y_n^2 \tag{45}$$

We begin by writing down the generating function for the sequence $y(n), y(n+1), \ldots,$ when linear theory holds. By definition

$$H(z) \equiv \sum_{k=0}^{\infty} y(n+k)z^k \tag{46}$$

Using standard linear techniques we calculate from eq. (1) and its initial conditions

$$H(z) = \frac{\sum_{j=1}^{n} y(n-j) \sum_{s=0}^{n-j} a_{j+s} z^s}{-\sum_{s=0}^{n} a_s z^s} \tag{47}$$

where we have arbitrarily defined $a_0 = -1$. We note the characteristic polynomial

$$c(z) = -\sum_{i=0}^{n} a_i z^{n-i}$$

has the same modulus as the denominator of $H(z)$ when $|z| = 1$ (since the a_i are real).

We next note that

$$\sum_{k=n}^{\infty} y_k^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(z = e^{i\theta})|^2 d\theta \tag{48}$$

To express this as a quadratic form introduce, for $0 \leq |s-t| \leq (n-1)$ the integrals ($z = e^{i\theta}$)

$$I_{st} \equiv \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{z^{s-t}}{\left| \sum_{s=0}^{n} a_s z^s \right|^2} d\theta = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\cos(s-t)\theta}{\left| \sum_{s=0}^{n} a_s z^s \right|^2} d\theta \tag{49}$$

Also, as a contour integral around the unit circle, we have

$$I_{st} = \frac{1}{2\pi i} \oint \frac{z^{n+s-t-1}}{(\Sigma a_s z^s)(\Sigma a_t z^{n-t})} dz \tag{50}$$

The first form of the integrals shows they are real and $I_{st} = I_{ts}$. Also introduce the real symmetric matrix

$$H_{jk} = \sum_{s=0}^{n-j} \sum_{t=0}^{n-k} a_{j+s} a_{k+t} I_{st}, \quad j, k = 1, \ldots, n \tag{51}$$

Note for $j = k$, $H_{jj} > 0$ since the right side of eq. (51) is then a Teoplitz form with positive spectral function. Then with these notations simply using eq. (47) to expand the integral in eq. (48) yields

$$w[Y] = \sum_{k=n}^{\infty} y_k^2 = \sum_{j,k=1}^{n} y(n-j)y(n-k)H_{jk} \tag{52}$$

or

$$w[AY] = \sum_{k=n+1}^{\infty} y_k^2 = \sum_{j,k=1}^{n} y(n+1-j)y(n+1-k)H_{jk} \tag{53}$$

When $y_n > 1$, $w[F(Y)]$ has the same form as eq. (53) except that y_n is replaced by unity. Thus in evaluating $w[F(Y)] - w[AY]$ [by using eq. (53)] most of the quadratic terms will cancel. Doing this and a few minor manipulations, the criterion for stability will read that we must have

$$[y(n)+1]H_{11} + 2\sum_{j=2}^{n} H_{1j}y(n+1-j) \geq -\frac{y^2(n)}{y(n)-1} \tag{54}$$

whenever

$$y(n) = \sum_{j=1}^{n} a_j y(n-j) > 1 \tag{55}$$

If we define $H_{1,n+1} = 0$, we may write this as:

Theorem II: If there are no simultaneous solutions to the inequalities

$$\sum_{i=1}^{n} [-a_i H_{11} - 2H_{1,i+1}]y(n-i) \geq H_{11} + \frac{y^2(n)}{y(n)-1} \tag{56}$$

$$y(n) = \sum_{j=1}^{n} a_j y(n-j) > 1 \tag{57}$$

$$|y(i)| \leq 1 \quad i = 0, 1, \ldots, n-1 \tag{58}$$

then the difference eq. (1) is stable with saturation arithmetic.

To convert this into a hyperplane problem (discussed in Section IV) the nonlinear term $y^2(n)/(y(n)-1)$ may be dropped or replaced by the value four (since

$$\frac{x^2}{x-1} \geq 4$$

when $x \geq 1$). If we drop this term, the resulting stability has the physical interpretation that at any time the energy in the remaining tail of the nonlinear response is less than or equal to the corresponding energy for the linear problem, regardless of the previous state. That is, if measured by energy, the nonlinear undriven response dies off at least as fast as the linear response for any initial conditions.

We also note that whenever the Liapunov function $w = Y^t HY$ [eq. (52)] drops to the value unity the system behaves as a linear one from there on (no future y_k will exceed unity). Several bounds for this quantity may be given. Since

$$\left|\sum_{s=0}^{\infty} a_s z^s\right|^2 = \prod_{i=1}^{n} |1 - z\lambda_i|^2 \geq \prod_{i=1}^{n} (1 - |\lambda_i|)^2$$

we have

$$w = \frac{1}{2\pi}\int_{-\pi}^{\infty} d\theta |H(z)|^2 \leq \max_{z=e^{i\theta}} |H(z)|^2$$

$$= \max_{z=e^{i\theta}} \left| \frac{\sum_{j=1}^{n} y(n-j) \sum_{s=0}^{n-j} a_{j+s} z^s}{-\sum_{0}^{n} a_s z^s} \right|^2$$

$$\leq \frac{\left| \sum_{j=1}^{n} |y(n-j)| \sum_{s=0}^{n-j} |a_{j+s}| \right|^2}{\prod_{i=1}^{n} (1 - |\lambda_i|)^2} \leq \frac{\left| \sum_{j=1}^{n} |a_j| \right|^2 \left| \sum_{i=0}^{n-1} |y(i)| \right|^2}{\prod_{i=1}^{n} (1 - |\lambda_i|)^2} \tag{59}$$

VI. EXAMPLES

The few simple examples of this section will shed light on the two methods we have given. When $n = 2$ the two inequalities of Corollary I are simply $|a_2| < 1$, $|a_1| < 2$. Since linear stability implies $|a_1| = |\lambda_1 + \lambda_2| < 2$, $|a_2| = |\lambda_1\lambda_2| < 1$, we see that for $n = 2$ linear stability implies stability with saturation arithmetic. We have already mentioned this is not true for $n > 2$. Mitra has constructed a counter example using the degenerate case $\lambda_1 = \lambda_2 = \ldots = \lambda_n \equiv \gamma$. For $n = 3$ he finds oscillations if $|\gamma| \geq 0.858$, although if $|\gamma|$ is smaller than this stability is not implied. If we consider the second inequality of Corollary I for $n = 3, k = 2, a_1 = 3\gamma, a_2 = -3\gamma^2, a_3 = \gamma^3$ we have stability if

$$3|\gamma| + |\gamma|^3 < 2$$

or $|\gamma| < 0.596$. No better result is obtained for this case by a complete use of Theorem I.

On the other hand if we apply the criterion of Theorem II (neglecting the nonlinear term) with the algorithm of Section IV, we find stability if $|\gamma| < 0.71$. Insignificant improvement would be obtained here if we had also included the nonlinear term. The application of Theorem II to the present case was sufficiently simple so that the calculation could be done by hand. The integrals were done exactly to give

$$H_{11} = \frac{\lambda^2}{(1-\lambda^2)^5}[9 - 9\lambda^2 + 10\lambda^4 - 5\lambda^6 + \lambda^8]$$

$$H_{12} = \frac{-3\lambda^3}{(1-\lambda^2)^5}(3 + \lambda^2)$$

$$H_{13} = \frac{3\lambda^4}{(1-\lambda^2)^5}(1 + \lambda^2) \tag{60}$$

Clearly in general the integrals would have to be done numerically. The fact that the nonlinear term did not contribute significantly is due to the combination of facts that at the critical value for λ, $y(n = 3)$ is not extremely close to one and also H_{11} is large, due to its denominator.

It should be pointed out that

$$\sum_{n}^{\infty} y_k^2$$

can be expressed using the solution eq. (4) for the linear problem as

$$\sum_{n}^{\infty} y_k^2 = Y^+(V^{-1})^\dagger \Gamma V^{-1} Y \tag{61}$$

where

$$\Gamma_{ij} = \frac{(\gamma_i^*)^n \lambda_j^n}{1 - \gamma_i^* \gamma_j} \tag{62}$$

However this explicit form is only true for the nondegenerate case and, although all limits exist as $\lambda_1 \to \lambda_2$, etc., the expression would probably not be suitable for numerical computation when eigenvalues are close to being degenerate.

Another example of limiting misbehavior being only apparent is that in a similar manner one could compute that

$$H(z) = \sum_{i=1}^{n} \frac{1}{1 - z\lambda_i} (V^{-1}Y)_i \tag{63}$$

Individual terms in this expression are badly behaved as, for example, if $\lambda_1 \approx \lambda_2$, but the alternate form eq. (47) shows everything is well behaved in the limit.

If we return to Corollary I applied to $n = 3$ when $(\lambda_1, \lambda_2, \lambda_3) = \rho(i, -i, 1)$, $0 < \rho < 1$, we see that the inequality

$$|a_1| + |a_3| < 2$$

is sufficient to guarantee that for filter poles in these relative position saturation arithmetic will give a stable filter for any ρ, all the way out to the boundary of linear stability.

Based on these examples we feel that Theorem II is the more powerful method although the simpler Corollary I can yield considerable information for particular cases.

Finally, we note that an important investigation on the present problem has just been completed by Mitra,[9] resulting in different stability criteria from those presented here. Mitra's results will give a polynomial type criterion for absence of *periodic* oscillations. These results in themselves do not prove stability in that Ref. 9 does not preclude unending periodic outputs with no input. However, we take the

liberty of mentioning that Mitra has extended the proof to include stability and thus the results of Ref. 9 may be taken as proving the same type of stability as discussed here. Another comparison with Ref. 9 involves the size of the stability region in "tap-space" which the two methods give. Neither Mitra's criterion nor ours can claim to describe the largest stability region. Also it does not even seem possible at this stage to give theoretical arguments to decide if one of the methods is always superior in this respect. However, several examples indicate that the region determined by Mitra's criterion is larger. Assuming this to be the case in general, an effective practical procedure would be to first test for stability using our simple Corollary I, and if this fails, apply Mitra's polynomial test.

APPENDIX

The Vandermonde Matrix and Symmetric Polynomials

The α denote an ordered set of n complex members α_i, i.e., $\alpha_1, \alpha_2, \ldots, \alpha_n$. By the Vandermonde matrix $V(\alpha)$, we shall mean the matrix

$$V(\alpha) = \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ \alpha_1 & \alpha_2 & \alpha_3 & \ldots & \alpha_n \\ \alpha_1^2 & \alpha_2^2 & \alpha_3^2 & \ldots & \alpha_n^2 \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \alpha_1^{n-1} & \alpha_2^{n-1} & \alpha_3^{n-1} & \ldots & \alpha_n^{n-1} \end{bmatrix} \tag{64}$$

This can be written $[V(\alpha)]_{ij} = \alpha_j^{i-1}$, $i, j = 1, \ldots, n$. We let $v(\alpha) = \det V(\alpha)$, and it is known[5] that

$$v(\alpha) = \Pi(\alpha_j - \alpha_i) \tag{65}$$

where the product extends over all i, j satisfying

$$1 \leq i < j \leq n$$

If $\alpha_i \neq \alpha_j$ for $i \neq j$ then the inverse of $V(\alpha)$ exists, and is known. Before giving its structure, we wish to list some facts concerning some special symmetric polynomials.[6]

Definition: The ℓth elementary symmetric function of n-variables ($\ell = 1, 2, \ldots, n$) is the sum of all formally distinct products of the variables taken ℓ at a time. We also define $p_0 \equiv 1$.

For example if $n = 3$ we have

$$p_0(\alpha) = 1$$

$$p_1(\alpha) = \alpha_1 + \alpha_2 + \alpha_3$$

$$p_2(\alpha) = \alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3$$

$$p_3(\alpha) = \alpha_1\alpha_2\alpha_3 \tag{66}$$

A well known theorem of algebra states that any symmetric polynomial in all the α's can be uniquely written as a *polynomial* in the quantities $p_i(\alpha)$, $i = 0, \ldots, n$.

Note that in the characteristic polynomial eq. (3) we have

$$a_i = (-1)^{i+1} p_i(\lambda) \tag{67}$$

where the λ_i, $i = 1, \ldots, n$, are the roots of eq. (3). Thus the theorem just stated says that any symmetric polynomial in the roots of a polynomial can be expressed as a polynomial in the coefficients of the equation (rather than a complicated function as would be required to express an individual root).

We shall use the notation $p_\ell^{(j)}(\alpha)$, $\ell = 0, \ldots, n-1$, to denote the ℓth elementary symmetric function formed from the $(n-1)$ ordered variables $\alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_n$. Likewise $v^{(i)}(\alpha)$ denotes the determinant of the corresponding $(n-1) \times (n-1)$ Vandermonde matrix.

Theorem:

$$[V^{-1}(\alpha)]_{ij} = \frac{(-1)^{i+j} v^{(i)}(\alpha)}{v(\alpha)} p_{n-j}^{(i)}(\alpha) \tag{68}$$

$$i, j = 1, \ldots, n$$

Proof: We have

$$q_{kj} \triangleq \sum_{i=1}^{n} [V(\alpha)]_{ki} [V^{-1}(\alpha)]_{ij} = \sum_{i=1}^{n} \alpha_i^{k-1} \frac{(-1)^{i+j} v^{(i)}(\alpha) p_{n-j}^{(i)}(\alpha)}{v(\alpha)} \tag{69}$$

From eq. (65) and the definition of $v^{(i)}(\alpha)$ we see that

$$\frac{v^{(i)}(\alpha)}{v(\alpha)} = \frac{1}{(-1)^{n-1} \prod_{\ell \neq i} (\alpha_i - \alpha_\ell)} \tag{70}$$

and hence eq. (69) is

$$q_{kj} = \sum_{i=1}^{n} \frac{\alpha_i^{k-1} p_{n-j}^{(i)}(\alpha)(-1)^{n-j}}{\prod_{\ell \neq i} (\alpha_i - \alpha_\ell)} \tag{71}$$

Form

$$q(x) \triangleq \sum_{j=1}^{n} q_{kj} x^j$$

and note that*

$$\sum_{j=1}^{n} p_{n-j}^{(i)}(\alpha)(-1)^{n-j} x^j = \sum_{\ell \neq i} (x - \alpha_\ell) \tag{72}$$

Thus

$$\sum_{j=1}^{n} q_{kj} x^{j-1} = \sum_{i=1}^{n} \alpha_i^{k-1} \frac{\prod_{\ell \neq i} (x - \alpha_\ell)}{\prod_{\ell \neq i} (\alpha_i - \alpha_\ell)} \tag{73}$$

has value α_m^{k-1} when $x = \alpha_m$, $m = 1, \ldots, n$. From this it follows that $q(x) = x^{k-1}$, so that $q_{kj} = \delta_{kj}$, which we were to prove.

We leave it to the reader to convince himself of the following:

Lemma I:

$$\sum_{\ell=1}^{n} p_{n-j}^{(\ell)}(\alpha) = j p_{n-j}(\alpha) \quad j > 0 \tag{74}$$

Lemma II:

$$p_j(\lambda) - p_j^{(i)}(\lambda) = \lambda_i p_{j-1}^{(i)}(\lambda) \quad j = 1, 2, \ldots, n-1 \tag{75}$$

REFERENCES

1. P. M. Ebert, J. E. Mazo, and M. G. Taylor, "Overflow Oscillations in Digital Filters," B.S.T.J. *48*, No. 9 (November 1969), pp. 2999–3020.
2. I. W. Sandberg, "A Theorem Concerning Limit Cycles in Digital Filters," Proc. 7th Annual Allerton Conf. Circuits and Systems Theory, pp. 63–68, 1969.
3. A. N. Willson, Jr., "Limit Cycles Due to Adder Overflow in Digital Filters," IEEE Trans. Circuit Theory, *CT-19*, No. 4, pp. 342–346, 1972.
4. D. Mitra, "Large Amplitude, Self-Sustained Oscillations in Difference Equations Which Describe Digital Filter Sections Using Saturation Arithmetic," IEEE Trans. Acoustics, Speech and Signal Processing, *ASSP-25 (1977)*, No. 2, pp. 134–143.
5. R. Bellman, *Introduction to Matrix Analysis*, McGraw-Hill, 2nd ed., 1970, p. 193.
6. M. Bocher, *Introduction to Higher Algebra*, Macmillan, 1936, Chapter 18.
7. J. LaSalle and S. Lefschetz, *Stability by Liapunov's Direct Method*, Academic Press, 1961.
8. A. N. Willson, Jr., "Some Effects of Quantization and Adder Overflow on the Forced Response of Digital Filters," B.S.T.J. *51*, No. 4 (April 1972), pp. 863–867.
9. D. Mitra, "Criteria for Determining if a High Order Filter using Saturation Arithmetic is Free of Overflow Oscillations," B.S.T.J., *56*, No. 9 (November 1977), pp. 1679–1700.

* The right member of (72) is a polynomial expressed directly in terms of roots, the left member, via (67), the polynomial expressed directly in terms of coefficients.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 3, March 1978
Printed in U.S.A.

Sequentially Companded Modulation for Low-Clock-Rate Speech Codec Applications

By S. V. AHAMED

(Manuscript received October 7, 1977)

The expansion of the step size of a speech codec may be arranged to change as the number of identical consecutive bits starts to increase. This technique causes the codec to respond partially to a fewer number of identical consecutive bits and more dramatically to larger numbers. In contrast to a typical exponential expansion of the step size, the proposed technique, in addition, expands the exponent. For speech, two distinct advantages have been observed: (i) the improvement of higher frequency audio frequency response at the same clock rate and (ii) the reduction of idle channel noise. In practice we have found that three- and four-bit companding will suffice for a typical 24 kHz, ADM codec. The proposed companding appears to be an acceptable choice between two-, three-, and four-bit companding which leads to better frequency response but worse noise, and four-bit companding which leads to both worse frequency and noise responses.

I. INTRODUCTION

The many distinct advantages of companding to encompass the dynamic range of speech signals are well documented. In most cases, a simple law is used repeatedly to arrive at the companded step size. For instance, in the 37.7 kHz, ADM SLC-40 codec,[1] a nonlinear function of the frequency of occurrence of four identical consecutive bits forces an expansion of the step size. In the two-bit companding described in Ref. 2, the expansion of step size follows a geometric progression. Three bit companding described in Ref. 3, again depends on a base-two geometric series for increasing the step size, when two consecutive bits are the same, and again on the same series with a base-half for decreasing the step size when the bits are of opposite polarity. Most of these systems perform adequately at higher (typically above 32 kHz) clock rates. However, when the clock rate is decreased, the simple fixed rules of companding either offer an unacceptable quantization noise at lower step sizes, or make the

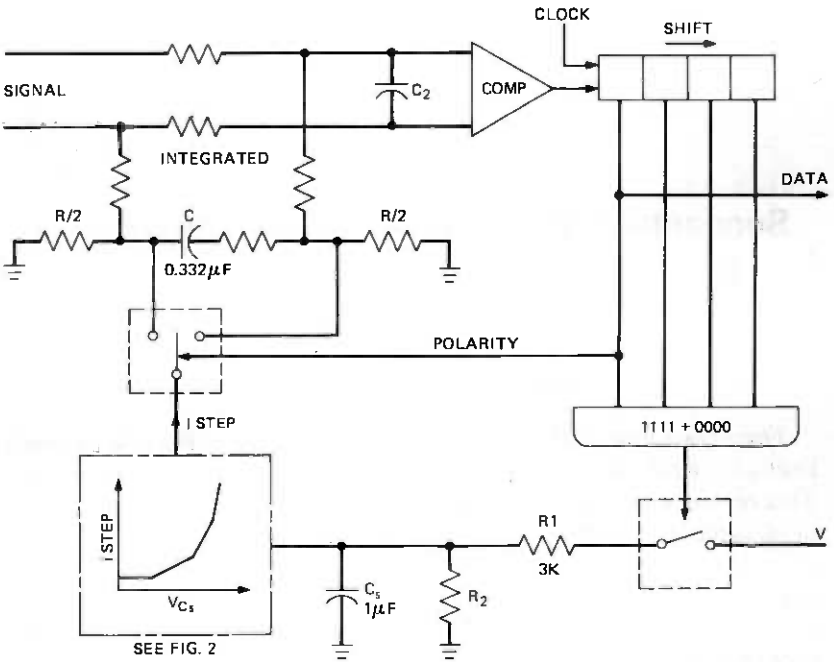


Fig. 1—Block diagram of encoder (see Ref. 4, Fig. 7).

message so choppy at larger step sizes that it becomes unintelligible. A happy compromise between the two is simply impossible. For this reason, we have investigated the problem by having a series of weightings attached to identical consecutive bit patterns of two, three, and four bits. When the bit pattern reverses, the decay of the step size is effected by an RC circuit with a time constant of about 9 msec.

Sequential companding proposes to utilize the binary sequence of data for companding the step size in a multiplicity of modes at different instants of time in a gradual way, whereas conventional encoding translates the companding information more drastically after a critical threshold has been reached. The multiple use of bit stream to convey companding information enhances the effective usage of bits at the same bit rate, or achieves the same quality of speech at a lower bit rate.

II. PERIPHERAL CIRCUIT FOR TESTING SEQUENTIAL COMPANDING

An existing ADM codec[1] has been used to test the principle of sequential companding. The encoder has a double integration feedback loop with the main pole at 235 Hz and the secondary pole at 2870 Hz. Figure 1 is a block diagram of the encoder. Four bit companding is effected by a logic circuit which forces an incremental charge on a 1 μF capacitor through a 3 kΩ resistance. The duration for which the charging

takes place equals the time during which the logic circuitry senses four consecutive ones or zeros at the output of the comparator, which in turn senses the difference between the incoming speech signal and the voltage across the feedback loop. When the four bits sensed are not all identical, the 1 μF capacitor is allowed to discharge to a 9 kΩ resistance. This combination yields two time constants; one for charging (attack) of about 3 msec, and one for a discharging (decay) of about 9 msec. The voltage across the 1 μF capacitor dictates the step size. A nonlinear circuit generates a step current whose magnitude depends upon the voltage across

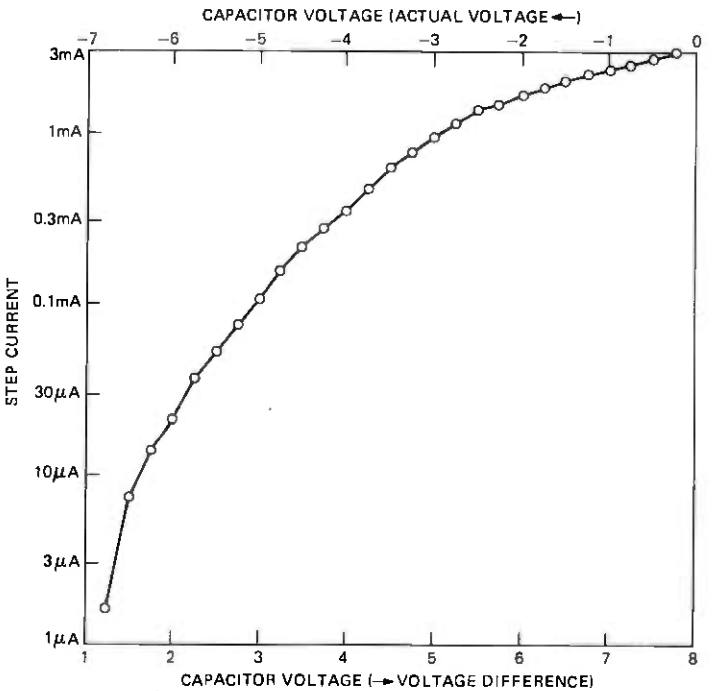


Fig. 2—Compander capacitor voltage and step current relationship.

the 1 μF capacitor. Figure 2 depicts the 1 μF capacitor voltage and the compander current. When there is no companding at all, the voltage across the capacitor becomes quite low (about 1.6 V) and the minimum step size of the 10 μA is reached. As the voltage reaches about 6.5 volts, the current step size reaches about 2 mA yielding about 46 dB range for the step size. The compander current is used to accumulate or deplete a charge on a final 0.33 μF integrator capacitor and it is the voltage across this capacitor which yields the original voice frequency signal after a low pass filter with a 3 dB loss at about 2000 Hz. The final capacitor has

about 1.2 kΩ in its discharge path. When the current step size is about 10 μA, the voltage swings between ±0.58 mV across the integrator capacitor at a clock frequency of 24 kHz.

III. ESSENTIAL DIFFERENCES BETWEEN CONVENTIONALLY COMPANDED AND SEQUENTIALLY COMPANDED CODECS

3.1 The idle channel noise

The character of the idle channel noise is totally different in the sequentially companded codecs. Whereas in the conventionally companded encoder, the bit pattern generated by the encoder during a silent period is nondeterministic and depends on the comparator characteristics, the sequentially companded encoder generates a sort of restless limit cycle in which the climax occurs when the three consecutive ones are produced, and the step size increases very slightly followed by pairs of zeros and pairs of ones for a few cycles. Meanwhile, the step size starts to decay due to lack of any companding, the paired zero-ones gradually vanish to a single zero and one combination and then the whole cycle repeats. This constitutes a semistable limit cycle and has been photographed in Figure 3a. The top trace shows the audio output from the decoder. The central trace is the integrator voltage in the decoder. The last trace is the clock at 24 kHz which also triggers the oscilloscopic sweep. In contrast a similar oscillogram (Fig. 3b) for a conventionally companded codec shows a total absence of any pattern during the silence.

These two pictures also forecast the difference in character of the two idle channel noises. The sequentially companded decoder presents a component of frequency at about 190–200 Hz which is considerably quieter than random noise generated by the conventionally companded ADM. But the higher frequency channel noises are less than those in the conventionally companded ADM coded. Psychologically the effect of such a low-frequency steady low-level signal appears to be more tolerable than the randomly varying noise produced by the latter.

Sequentially companded codecs also have one additional feature to enhance the signal to idle channel noise ratio. The frequency of companding in the three and four bit compander is higher than that in the four bit compander alone. Hence, the signal strength at the input to encoder can be higher at the same input frequency. The study of the four bit conventional codec suggests a 60 mV occasional peak at the input level to the codec. This value is appropriate and is consistent with an average voltage across to the integrator capacitor with the compander switch being functional (Ref. 4) for about 10 percent of the time. However, for a sequentially companded codec, the compander switch would be functional almost twice as frequently. However, the charging rate of

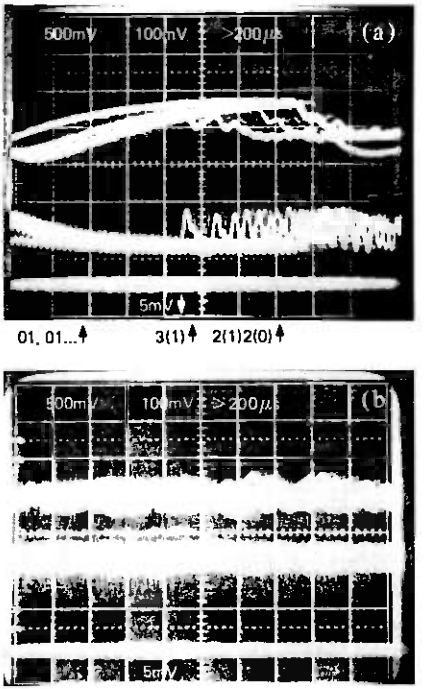


Fig. 3—Idle channel noise characteristics of (a) sequentially companded ADM codec and (b) conventionally companded ADM codecs.

the three bit compander is less than the charging rate of the four bit compander. Further, the switch for the four bit works in unison with the three bit switch. Hence, the average signal level for the sequentially companded codec tends to be on the order of 160–220 mV. At this level the sequentially companded codec at 24 kHz "sees" a frequency of 4 kHz the same way as a conventionally companded codec at 37.7 kHz would see a frequency of 4.7 kHz. But since the resistances in the charging paths are different, the responses would also be slightly different. Further, the response to the lower frequency from the sequentially companded codec should become nominally better* since both three- and four-bit companding can take place simultaneously.

The signal to noise implication of this difference of behavior between the two codecs is that whereas the noise remains about the same for the sequentially companded codec, the signal level increases about two to three times bringing down the signal to idle channel noise ratio considerably. This result has to be observed consistently during normal functioning of the codec.

* Experimentally we have seen very little difference at low audio frequencies which tend to pass through the telephone network.

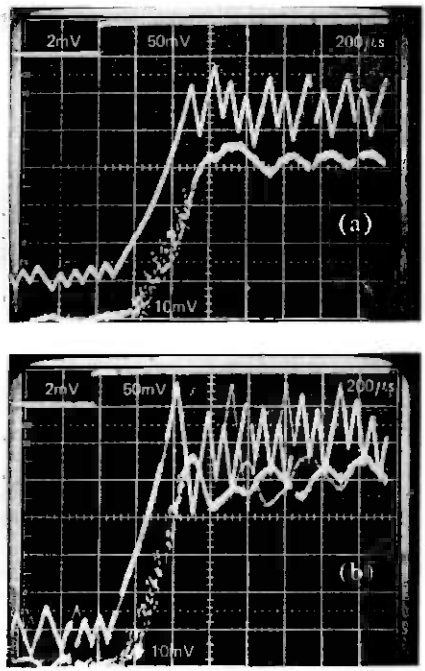


Fig. 4—Unit function response of (a) conventionally companded codec and (b) sequentially companded codec for a 50 mV step function.

3.2 Unit function response

This response indicates the rapidity with which the codec can respond to changes in the input level. It also brings about the high frequency response of the codec. During the normal operation the capacity of the sequentially companded codec to respond rapidly is reflected by lower slope overload noise. When a 50 mV step is imposed on the encoder, the decoder responses for the conventionally and sequentially companded codecs are presented in Figures 4a and b and output voltages are also tabulated in Table 1A. Similar response to a 360 mV surge is shown in Figs. 5a and b and the output voltages are presented in Table IB.

3.3 Higher audio frequency response

An audio frequency of 3149 Hz (nonsynchronous with the 24 kHz clock) is chosen for the comparison of performance of the two codecs. Figures 6a and b represent two spectrograms generated at the output. In the response from the conventionally companded codec (Fig. 6a), the peak at the input frequency is surrounded by a large number of cluttered peaks with rapidly changing (indicated by the density of broken patterns and smears in the figure) tones. Each smear is an audible change in tone

Table IA — Differences in responses to 50 mV step function at encoder

Elapsed time, μsec	Conventionally companded codec, mV	Sequentially companded codec, mV
200	2.5	3.0
300	10.0	15.0
400	20.0	29.0
500	32.0	47.0
600	43.5	55.5
800	44.0	48.0
1000	44.0	50.0

Table IB — Differences in responses to 360 mV step function at encoder

160	19	35
320	50	80
480	110	180
640	205	300
800	320	370
960	320	370

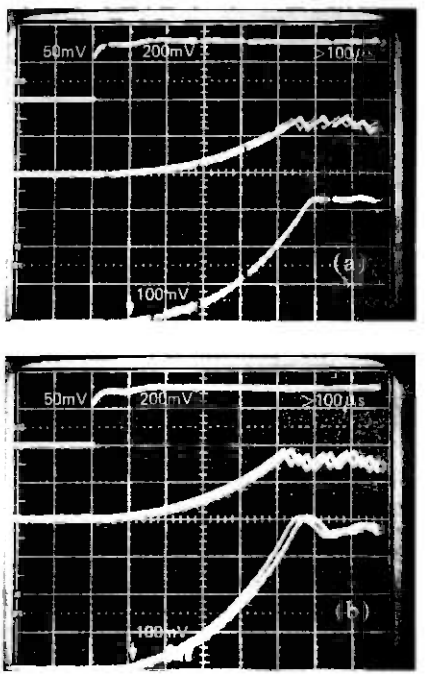


Fig. 5—Unit function responses of (a) conventionally companded and (b) sequentially companded codecs for a 360 mV input surge.

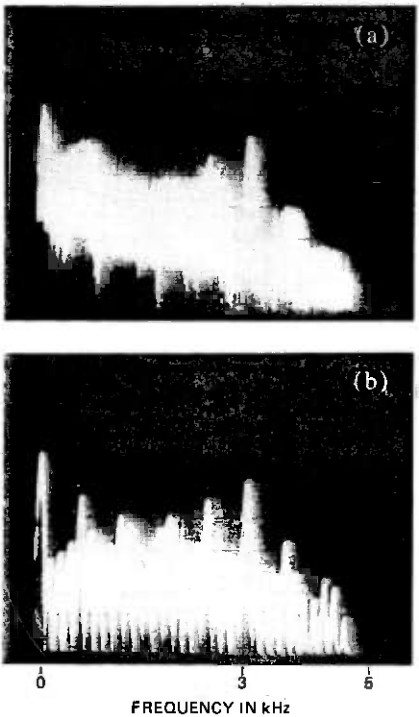


Fig. 6—Typical responses at about 3000 Hz from (a) conventionally companded and (b) sequentially companded ADM codes.

which cannot be missed by a listener. In contrast the response from the sequentially companded codec is cleaner and better formed with a fewer number of breaks and smears. Listening to the output tones from the two codecs also confirms this result.

IV. COMPUTED SIGNAL-TO-NOISE RATIOS

4.1 Validation of the computer model

Computerized models of the conventionally and sequentially companded codecs have been developed to compute the S/N ratios. The model of the codec programmed for a Nova 800 minicomputer is a general purpose version of a typical ADM codec whose companding can be changed by input variables. The same model serves to compute the binary sequence, the output wave shapes, the signal to noise ratios, etc., by altering the data to the minicomputer in a conversational mode of communication between the operator and the machine. For the encoder model an ideal comparator is used. Hence, the additional noise generated by the imperfection of the comparator is absent from the computed re-

Table II — Computed and measured S/N ratios

Frequency	Measured* (Ref. 4), dB	Computed,* dB
300	38.5	40.0
800	32.5	32.5
1600	22.5	22.0

* The clock rate is 37.7 kHz.

Table III — 24 kHz steady-state S/N ratios

Audio frequency, Hz*	Conventional companding		Sequential companding	
	Bits	S/N	Bits	S/N
2900	3	7.9		
	4†	5.7	3, 4	8.35
2490	3	3.7		
	4†	2.1	3, 4	3.9
1660	3	16.0		
	4†	13.0	3, 4	15.3
830	3	25.1		
	4†	23.2	3, 4	29.4
415	3	31.6		
	4†	31.0	3, 4	31.8

* These numbers are chosen to be irrational fractions of to clock rate at 24 kHz to avoid a limit cycle condition.
† Presently used charging resistance = 3.01 kΩ.

sults. This leads to the assertion that the computed S/N ratios would be the upper bound for the measured S/N ratio. These values (published in Ref. 4) have been used to validate the model at different sine wave inputs and have been presented in Table II.

From the computational models it also becomes evident that the S/N ratio is by no means a constant but a time varying quantity. Whereas the S/N ratio is measured as a time-average over a period (typically between 0.5 to 3 seconds), the computed S/N ratios are averaged over much shorter intervals and, hence, one would expect some difference between the computed and the measured values. Nonetheless, since the time constant is the same for all computations, the cross comparison of the computed values would still be a valid relative measure of their performances. To reduce the effect of transient variations of the S/N ratios, the computed value is the average of 60 such ratios at 60 consecutive clock cycles, each ratio being calculated as the moving average of twenty adjoining ratios around each clock cycle. Further, a series of such computations are made over numerous input frequency cycles and an average number is derived.

4.2 Effect of sequential companding on steady-state sinusoidal inputs

Table III lists the values of computed signal-to-noise ratios for conventionally companded and sequentially companded codecs at 24 kHz.

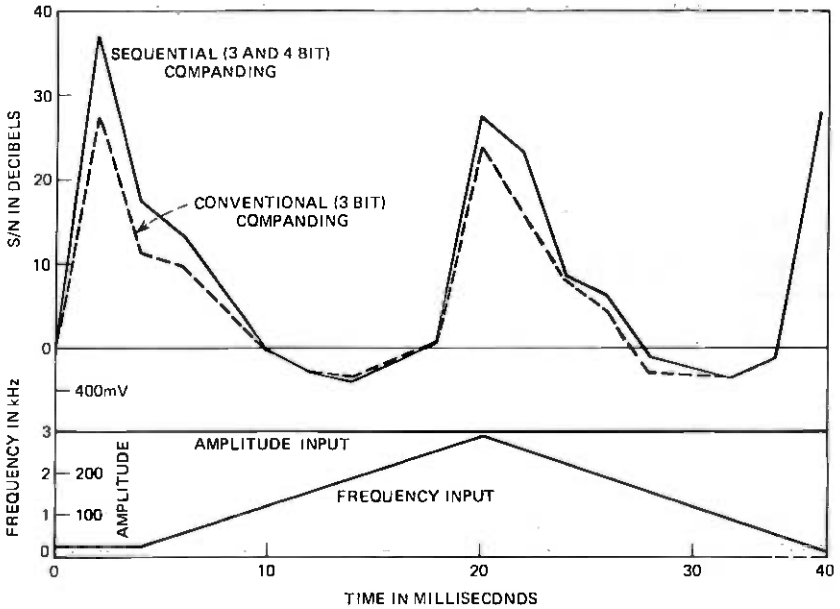


Fig. 7—S/N ratios for rapidly varying frequencies.

The input parameters have been held substantially the same for all the cases presented. However, the values of the charging resistances* for three-bit companding has been computationally optimized as 5.2 kΩ, whereas its corresponding value for four-bit companding has been retained as 3.01 kΩ as it exists in current applications. In case of the sequential companding their values† have been optimized as 5.2 kΩ for three bits and 4.36 kΩ for four bits, even though any other resistance values in their proximities will perform adequately.

4.3 Effects of sequential companding on frequency modulated sine wave inputs

Steady tones are rarely encountered in telephone conversations. Rapidly varying frequencies are, however, typical and for this reason we have studied the responses of conventional and sequential coding at 24 kHz clock rate when the input signal has a frequency which changes gradually from 250 Hz to 2900 Hz and back to 250 Hz within 30–50 milliseconds. Such changes are well perceived by the listener and the average signal to noise ratios indicate the relative faithfulness with which

* This charging resistance controls the voltage on the step size capacitor, which in turn controls the current step for charging or depleting the final integrator.

† Experimental determination (Section III) for best subjective listening seems to indicate that 5.2 kΩ and 3.01 kΩ are the desirable values for three- and four-bit companding.

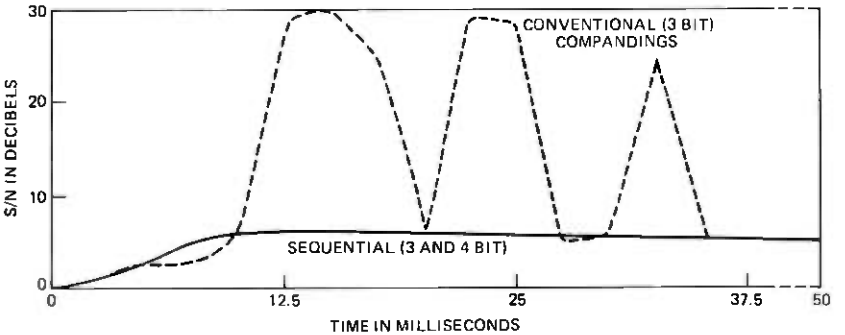


Fig. 8—Square-wave response to 1660 Hz signal.

the codecs can follow the change in frequencies and thus retain the original message characteristics.

The response to such a frequency characteristic is presented in Fig. 7. The average S/N ratio during the entire cycle is 10.18 dB for the sequential and 8.29 dB for the conventional encoding. The averages of positive S/N ratios are 17.95 dB for sequential and 14.91 dB for conventional encoding.

To signify the uneven behavior of the conventional encoding further, square wave at 1660 Hz was presented at the encoder. The results are shown in Fig. 8. The effectiveness of the sequential companding in following rapidly changing input is also illustrated in Fig. 9. Computed S/N ratios are plotted when a tone burst signal at 1660 Hz is presented to the two types of the encoders. To meet the rapidity of response of the three- and four-bit companding, the charging resistance (see Sec. 4.2) of the

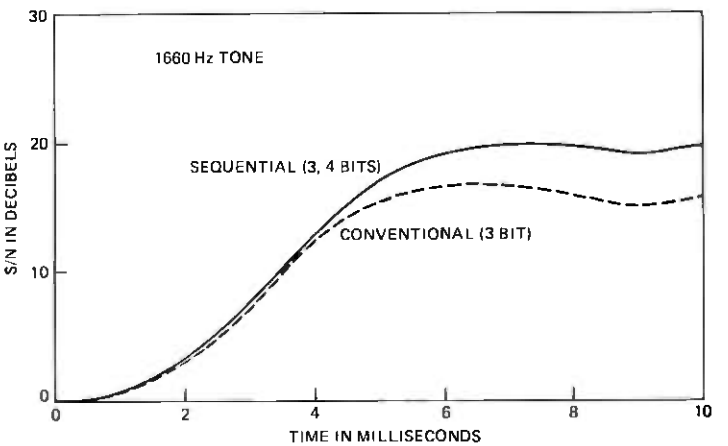


Fig. 9—1660 Hz tone burst response.

three-bit conventional companding is adjusted from 5.2 kΩ to 4.5 kΩ. However, the steady-state S/N ratio attained by the three- and four-bit sequential companding is about 18.7 dB as against 16.3 dB for the conventional companding. We have not been successful in achieving a rapid response and also a consistently high steady state S/N ratio from conventional companding to match the performance of the sequential companding. To quantify this assertion we have roughly the same rapidity of response from both codings at 830 Hz. However, the steady-state S/N ratio of sequential coding is 4.3 (29.4 vs. 25.1) dB higher than conventional coding. Complementarily, when the steady-state S/N ratios are approximately the same at 250 Hz, the sequential companding S/N ratio is roughly 9.5 (37.0 vs. 27.5) dB better than conventional companding S/N ratio 2.3 msec after the tone burst.

V. DISCUSSION OF THE DIFFERENCES

5.1 Experimental results

Consider an ideal case where a series of ones is presented at the input data of the ADM decoder. The integrator voltage responds to an increasing step current. If the discharge of this integrator capacitor is ignored for the present, then the voltage across the integrator is directly proportional to the increasing current step. Qualitatively this voltage may be represented by curves A–E of Fig. 10. At 48 kHz the change in step size after the eighth "one" is $h'i'$, whereas the step size after the fourth "one"at 24 kHz (corresponding to the same lapsed interval of time) is only hj. If the codec had three- and four-bit companding, the change in integrator voltage would have been $h''j''$ which is greater than hj due to two reasons: (i) the additional companding step at d and (ii) the simultaneous action of both the three- and four-bit companding at f''. A simple three-bit companding would have had a slower response as depicted by the curve D.

Again consider the influence of the resistances in charging paths of the step size capacitor if there are two resistances R_1 and R' in the charging paths energized by the three-bit and four-bit companding, then the slope of the curve B can be adjusted to any desired value. Different values of these resistances yield different ranges of these curves with one essential, vital difference. When the resistance is too low, the step size becomes too coarse leading to crackle within the word every time the polarity of a bit changes after a series of ones or zeros, and it is this direction in which a real compromise must be sought while adjusting the value of the resistances to minimize the slope overload noise.

In essence, the sequentially companded ADM codec behavior differs from that of a conventionally companded codec to the extent that an additional factor of nonlinearity is imposed in its response. The con-

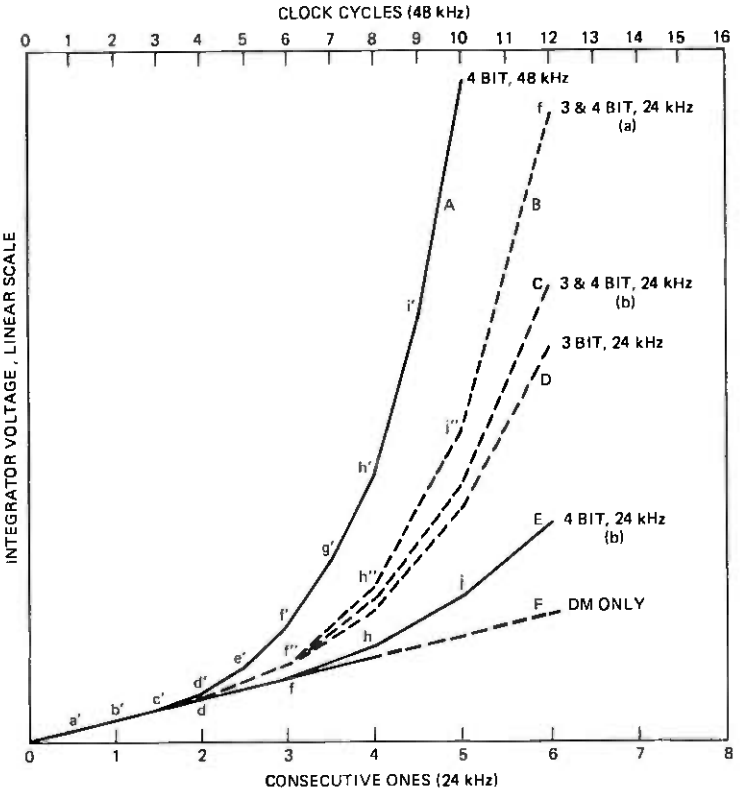


Fig. 10—Idealized responses of ADM codecs.

ventionally companded ADM codec responds by the inherent nonlinearity between the step size capacitor voltage and the current step, whereas the sequentially companded codec expands the step size capacitor voltage itself with a nonlinear character, and this nonlinearity rides along the nonlinear relation of Fig. 2 at all sizes of the current steps making the codec more responsive and more sensitive especially at lower clock rates. The validity of this assertion is demonstrated in experimental results in Figs. 4, 5, and 6.

5.2 Computational results

The steady-state sinusoidal response of a sequentially companded codec can be matched by a conventionally companded codec operating at a lower number of bits of the sequentially companded codec. However, the rapidity of such a response cannot be achieved by conventional companding which can also yield a consistently higher steady state S/N ratio. The computational verification of this assertion is indicated in Fig.

9 due to consistantly lower S/N ratio from the conventionally companded codec during the cycle of frequency variation. This characteristic is reflected by crisper sounding words from sequentially companded codecs.

Sequentially companded codecs seem to reach and hold a steady state more quickly and more effectively as shown in Fig. 8. The large swing of the S/N ratio when reaching the steady state by the conventionally companded codec indicates its inadequacy to yield steady tones essential for MFKP system and *TOUCH-TONE*® signals.

VI. CONCLUSIONS

Both experimental and computational results confirm that sequential companding reduces the response time from the codec even though conventionally companded codecs can compete well in the steady state response for sine wave excitations. Further, the tones generated from sequentially companded codecs tend to have fewer breaks and thus are steadier and yield higher S/N ratios at higher audio frequencies.

Experimental results indicate that the idle channel noise from sequentially companded codecs is considerably lower than the idle channel noise from conventionally companded codecs during actual message transmission. The intelligibility of the message is also better due to crisply formed words and lower background noise.

Computational results indicate that when the frequency centers around 800 Hz and when a small frequency modulation is embedded, then the S/N ratio from the sequentially companded codecs is about 2–3 dB better than an optimally designed conventional codec.

VII. ACKNOWLEDGMENT

The author acknowledges the assistance of Craig Thompson in fabrication of the sequentially companded ADM.

REFERENCES

1. S. J. Brolin and G. E. Harrington, "The SLC-40 Digital Carrier Subscriber System," IEEE Intercon Conference Record 1975, 81, pp. 1–5.
2. N. S. Jayant, "Adaptive Delta Modulation with 1 bit Memory," B.S.T.J., *49*, No. 3 (March 1970), pp. 321–342.
3. M. R. Winkler, "High Information Delta Modulation" IEEE Intl. Conv. Rec., 11, No. 8, 1963, pp. 260–265.
4. R. J. Canniff, "Signal Processing in SLC-40, a 40 Channel Rules Subscriber Carrier," ICC 75 Conference Record, *Vol.* 3, pp. 40–7 to 40–11.
5. S. V. Ahamed, "The Nature and Use of Limit Cycles in Stabilizing the Behavior of Semideterministic Systems," internal Bell System communication.
6. Bell Telephone Laboratories, "Transmission Systems for Communications," fourth edition, Western Electric Company, Inc., Winston Salem, North Carolina, February 1970.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 3, March 1978
Printed in U.S.A.

Gradient Encoding for Low-Bit-Rate Stored Speech Applications

By S. V. AHAMED

(Manuscript received October 7, 1977)

In stored speech applications, the waveform of the message is completely specified and can be effectively used to reduce the bit rate at which the message may be synthesized. In gradient encoding, we propose to match the gradient of the output wave of the differential decoder with the required gradient between discrete clock cycles. When the required gradient is very steep the bit pattern selected maximizes the rate of change of the decoder voltage, otherwise appropriate bits of opposite polarity are inserted to match the amplitude of the decoder voltage with the required voltage at the discrete clock cycles. The performances of gradient encoding and conventional encoding are presented as corresponding signal-to-noise ratios under different inputs and circuit conditions. Further, our preliminary results indicate that gradient encoding can lead to comparable quality of speech at about half the bit rate of the conventional encoding between 32 to 24 kbaud.

I. INTRODUCTION

For stored speech application, one of the ways of generating efficient binary data is tree encoding, which examines and verifies the sequences of a prespecified number of bits by varying the data in every possible combination and selecting the one that yields the best signal-to-noise ratio. This way of exhaustive searching for the best bit pattern demands a large number of computations, and the number of computations expands geometrically as the number of bits in the tree (i.e., the number of sequential bits chosen to explore the range of variation of the decoder voltage) increases. This leads to a further uncertainty about whether the number of bits chosen is satisfactory or not for any given section of speech.

To circumvent this problem we have chosen to seek an alternative algorithm and proceed on a variable length of speech waveform determined by the gradient around the section under investigation. This is accomplished by realizing that all speech wave shapes consist of peaks and valleys, and the duration between these successive extrema should guide the duration of the computation, and that the gradient between them should guide the fragmentation of the compute-and-match procedure attempted during the relaxation* of bits for the intervals between the peaks and valleys or valleys and peaks.

II. THE BASIC APPROACH

When the locations of the peaks and valleys contained in any segment of speech have been determined, the synthesis of the optimal bit sequences may be routinely and systematically determined as follows:

(*i*) Determine the change in amplitude required from the differential decoder and the interval for the change.

(*ii*) Determine the best the decoder can accomplish by forcing a sequence of zeros (for peak to valley fit) or ones (for a valley to peak fit).

(*iii*) If the decoder can exceed the required change, halve the interval for the computation and evaluate the decoder performance by stuffing zeros or ones during half the interval.

(*iv*) Proceed to repeat (*ii*) and (*iii*) until one of the following occurs: (*a*) The decoder performance comes to within a very tight tolerance level of what the original speech wave called for. (*b*) The interval for computation has collapsed to one clock cycle of the decoder and if so choose a (0) or (1) that minimizes the error at the end of that particular clock cycle.

(*v*) When *iv*(*a*) or *iv*(*b*) are completed, update the new peak as the last point processed under *iv*(*a*) or *iv*(*b*) if the search pattern is progressing from a peak to a valley and retain the same valley or update the new valley as the last point processed under *iv*(*a*) or *iv*)*b*) if the search pattern is progressing from valley to peak and retain the same peak.

(*vi*) When the binary bits during the interval have been synthesized, proceed to the next section of the speech wave shape—i.e., to the next peak-valley or valley-peak pair.

III. DIFFERENCE BETWEEN CONVENTIONAL ENCODING AND GRADIENT ENCODING

Conventional encoding ignores the *a priori* information about the location of the next extreme point and can make large errors in achieving the best performance from a decoder. Gradient encoding ignores the

* In this context relaxation implies a systematic iterative selection.

basic premise of conventional encoding by forcing the next bit to be of opposite polarity if the present decoder voltage has exceeded the input signal, and retains a slight error at the present position in an overall attempt to do its best to reach the target extremity of the wave shape. When targets become very far apart, then the intermediate ranges of fit start to shrink and minimize the error at the intermediate points. In an extreme case of absolute silence, gradient encoding and conventional encoding converge in the bit selection of alternate zeros and ones.

The anticipatory characteristics of gradient encoding also prepare the decoder for sudden peaks in wave shapes by sending a series of identical bits before arriving at the peak, so that the extreme point is within a predetermined range of error. This is totally absent in conventional encoding.

IV. PERFORMANCE OF GRADIENT ENCODING—ALGEBRAIC WAVES

4.1 Summed sine waves

Figure 1a indicates the performance of the conventional ADM encoding technique when a sampling frequency of 12 kHz has been used to excite the encoder which is following an input wave generated as the sum of two sine waves, one at 400 Hz with an amplitude of 80 mV, and the other at 1200 Hz with an amplitude of 100 mV. In contrast, Fig. 1b indicates the performance of the gradient encoder technique under the same conditions. In Fig. 1a it can be seen that the point 3 on the dotted line being very close to 3 on the full line can materially change the next bit polarity and thus change the ensuing bit pattern; whereas in Fig. 1b the gradient encoding tolerates errors at 3, 4, 5 and 6 in order to match the segment 2–6 on the decoder curve (dotted line) as closely with the section 2–6 of the input, (solid line) and concentrates the transition of 010 near the peak where it should logically be placed. Other such variations are also noticeable by comparing 1a and 1b.

4.2 Interrupted sine wave

The anticipatory character of gradient encoding is evident by comparing Figs. 2a and 2b. The input to the two types of encoders is a sine wave at 1200 Hz at 100 mV interrupted at a frequency of 400 Hz. In Fig 2a it can be noted that the encoder starts to respond by a series of fixed values only after the input wave has actually been presented at the encoder, whereas the gradient encoder in Fig 2b starts to process the bits prior to actual impact of the wave, and adjust its bits accordingly.

V. SIGNAL-TO-NOISE RATIO ANALYSIS

5.1 Program description

The performances of conventional and gradient encoders have been modeled on a Nova 800 minicomputer by a sequence of machine language

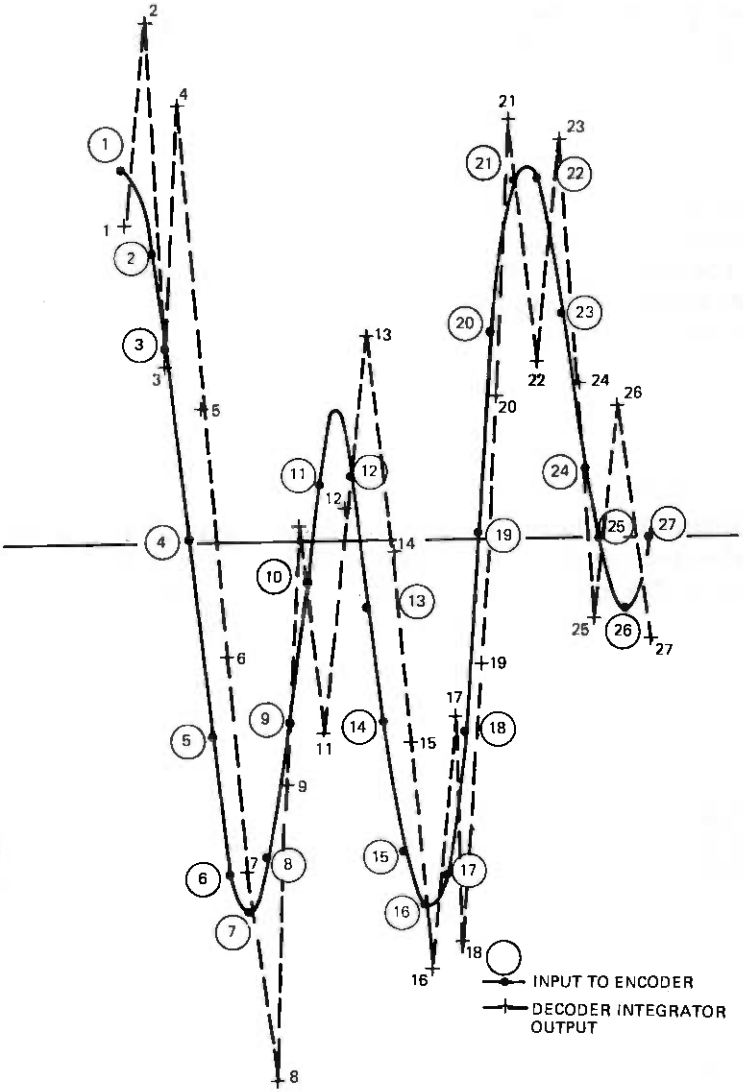


Fig. 1—(a) Conventional encoding.

and Fortran programs. The numerical computations are confined to a block of replaceable Fortran programs and the data handling from disc is performed by a set of machine language subroutines. Both communicate with the operator in a conversational mode and the circuit parameters are input controlled. It is thus possible to compare the relative performance of the two types of coding schemes under any set of conditions. The results are presented in the following sections.

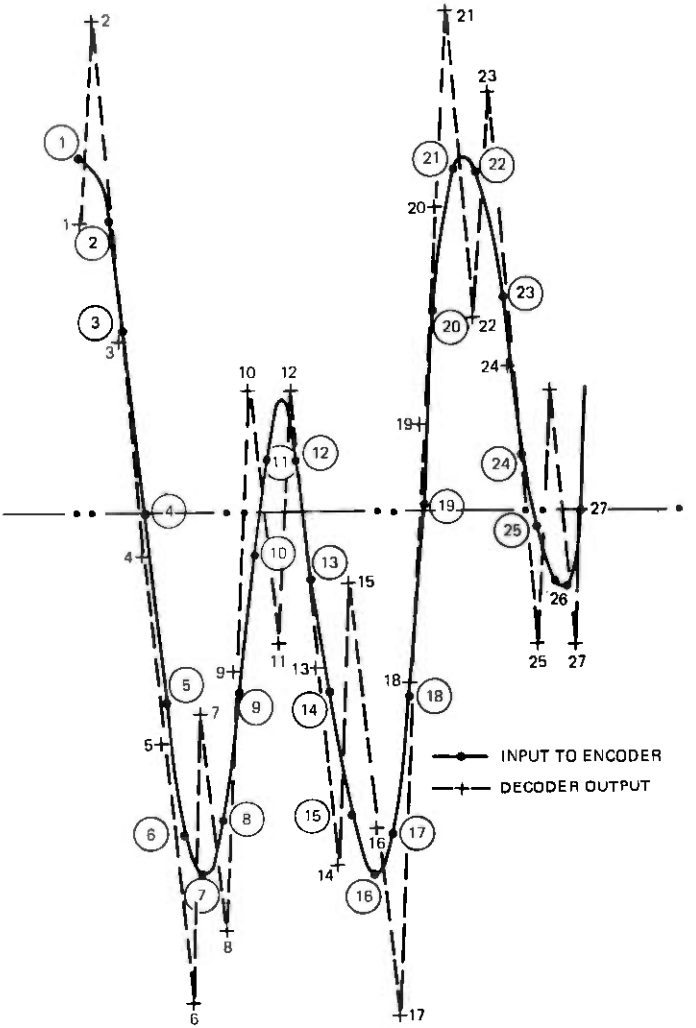


Fig. 1—(b) Gradient encoding.

5.2 Cross comparison of performance at the same clock rate

Four clock rates (32, 24, 16, and 12 kHz) are chosen to compare the performance with the ADM codec described in Ref. 1. The charging time constants of the step size capacitor and the levels have been optimized to achieve the best S/N ratios with different number of bits for companding (see Section II, Ref. 1). However, in the case of the 4 bit companding the charging time constant has been retained as 3 msecs as it

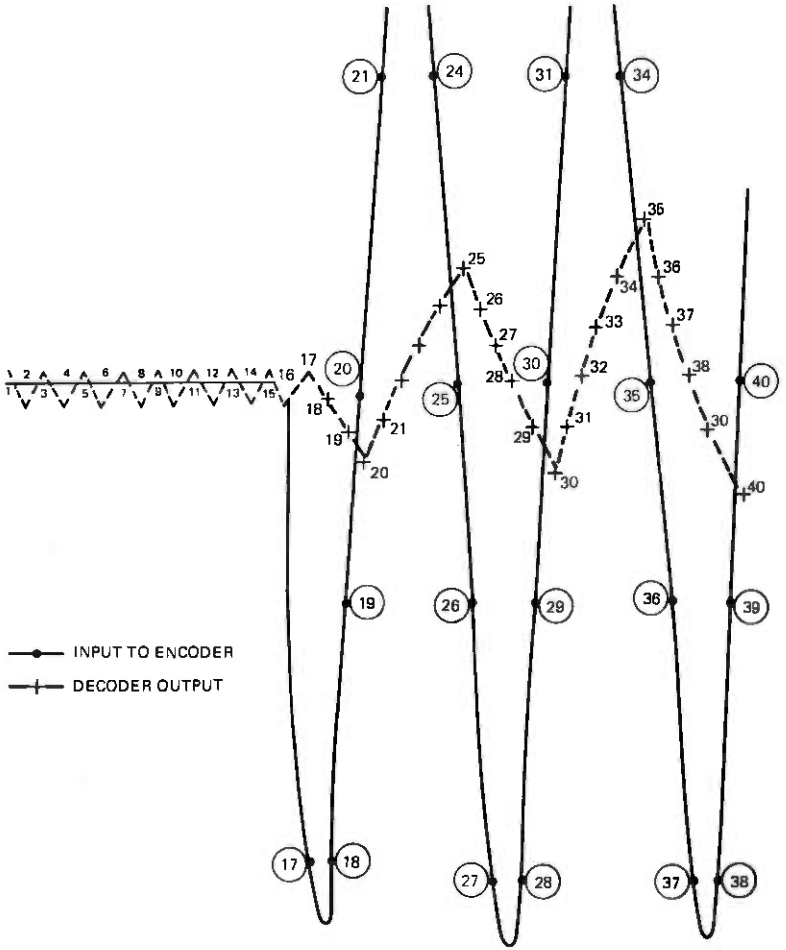


Fig. 2—(a) Conventional encoder.

currently exists. Table I contains the computed S/N ratios for 32 and 24 kbit per second clock rates and Table II contains the result from 16 and 12 kbits per second simulations.

5.3 Half-rate gradient encoding performance

Tables III and IV compare the performances of the 16 and 12 kbit per second gradient encoding against 32 and 24 kbits per second conventional encoding respectively. The time constants and levels have again been optimized to yield the best performance from the codec.

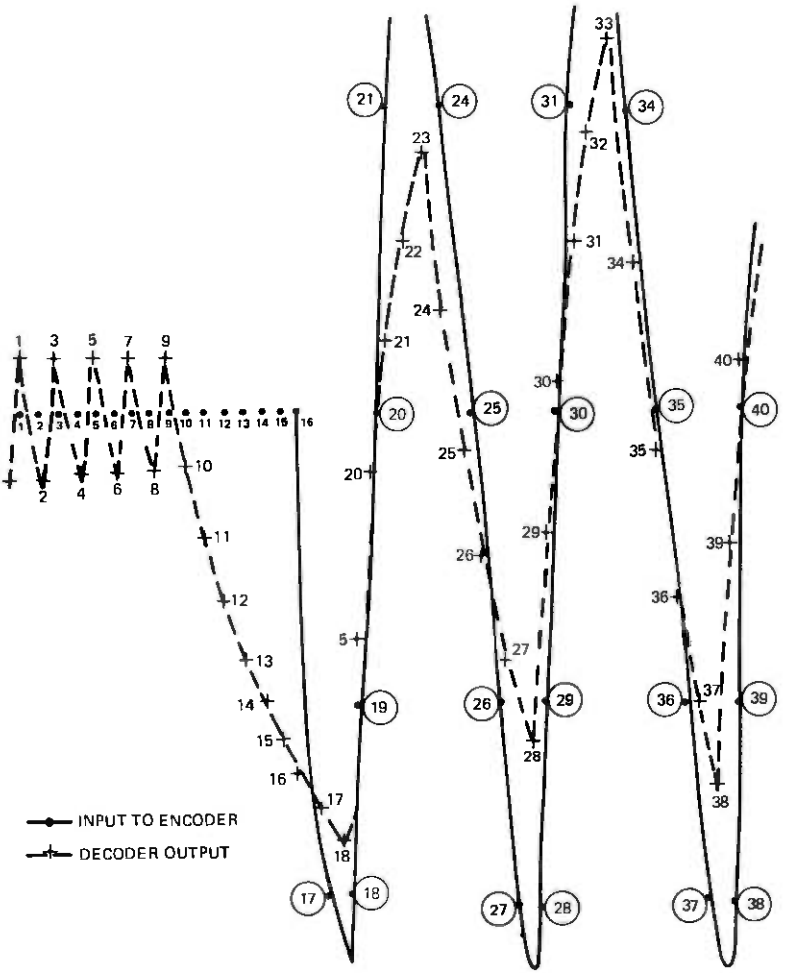


Fig. 2—(b) Gradient encoder.

VI. PERFORMANCE WITH SPEECH

Gradient encoding outperforms conventional encoding at the same bit rate. As the bit rate is reduced for gradient encoding a region of indifference is encountered between 50 to 60 percent of the rate of conventional encoding. The attack time constant (i.e., the product of the resistance for charging the step size capacitor and its value) starts to influence the higher frequency response but enhances the signal to noise ratio at the lower end of the audio frequency response and vice versa. A compromise is necessary to achieve the best response over the range

Table I — Computed *S/N* ratios at 32 and 24 kHz

Clock:		32 KHz		24 KHz	
Coding:		Conventional	Gradient	Conventional	Gradient
Audio frequency, Hz (sine waves)	No. of bits	*S/N*, dB	*S/N*, dB	*S/N*, dB	*S/N*, dB
5700	3	—*	10.2	—	
	4	—	0.8	—	
4700	3	—	12.8	—	
	4	—	13.5	—	
3950	3	9.0	8.0	—	13.8
	4	7.2	10.3	—	2.5
3150	3	12.0	14.7	—	12.3
	4	1.4	13.3	—	12.4
2900	3	9.7	15.8	7.9	10.0
	4	6.4	15.6	5.7	9.2
2490	3	11.9	20.3	3.7	16.6
	4	12.3	15.6	2.1	12.0
1660	3	19.6	24.5	16.0	19.3
	4	16.0	21.4	13.0	17.0
830	3	29.7	35.2	25.1	30.0
	4	26.0	32.3	23.2	26.6
415	3	38.5	40.7	31.6	33.0
	4	34.0	39.1	31.0	34.1

* — indicates near-zero values

of importance for telephone conversation. However, since the quantization noise in gradient encoding is scarcely present due to the optimization of the selected bit pattern, the region of indifference between gradient encoding and conventional encoding tends to be biased in favor of the former at a slight expense of the higher audio frequency response. Informal subjective testing has indicated that the 12 kbit per sec, 2 bit companded, gradient-encoded speech is comparable with the 24 kbit per sec, 4 bit companded, conventionally encoded speech.* However, the 24 kbit per sec sequentially companded speech[1] shows a favorable margin of performance over the 12 kbit per sec gradient-encoded speech.

The computation time depends on the bit rate and concentration of peaks and valleys. Lower frequency wave forms demand more computations in order to perform intermediate compute-and-match attempts. Higher frequencies on the other hand are adequately fitted by fewer overall gradient matching trials. When long telephone announcements

* This improvement has been made possible because gradient encoding does not attempt to maximize the signal to noise ratio but instead, matches the extremities of the wave shape. When the incoming wave shape offers a large cyclic change in the transition (due to change in pitch) of the gradient between and peak-valley or valley-peak pair together with a steep gradient between the points, gradient encoding ignores the cyclic variation in the gradient whereas an attempt to maximize the S/N (as it is done in tree encoding) tries to accommodate the cyclic change and can lead to a perceptually poorer quality of speech. To this extent gradient encoding outperforms tree encoding.

Table II — Computed *S/N* ratios at 16 kbits and 12 kbits/sec

Bit rate:		16 kbits/sec		12 kbits/sec	
Coding:		Conventional	Gradient	Conventional	Gradient
Audio frequency, Hz	Companding No. of bits	*S/N*, dB	*S/N*, dB	*S/N*, dB	*S/N*, dB
3150	2	—*	7.2	—	3.4
	3	—	2.8	—	—
2900	2	—	9.2	—	5.3
	3	—	4.0	—	0.3
2490	2	—	13.9	—	5.9
	3	—	7.3	—	1.0
1660	2	2.1	18.0	1.6	11.7
	3	0.92	19.0	—	7.5
830	2	19.5	28.5	18.4	26.0
	3	21.2	28.4	17.8	27.3
415	2	30.4	30.6	19.9	30.0
	3	32.2	30.2	27.5	29.3

*— indicates near-zero values

Table III — Comparison of 12 kbits/sec gradient and 24 kbits/sec conventional coding

Bit rate:	12 kbits/sec		24 kbits/sec	
Coding:	Gradient		Conventional	
Audio frequency, Hz	Companding No. of bits	*S/N*, dB†	Companding No. of bits*	*S/N*, dB
3150	2	3.4	3	—
	3	—‡	4	—
2900	2	5.3	3	7.9
	3	.32	4	5.7
2490	2	5.9	3	3.7
	3	1.0	4	2.1
1660	2	11.7	3	16.0
	3	7.5	4	13.0
830	2	26.0	3	25.1
	3	27.6	4	23.2
415	2	30.0	3	31.6
	3	29.3	4	31.0

* 2-bit companding at 24 kbits/sec leads to extremely noisy silence periods.
† Changing the time constants of the attack circuit (see Section II, Ref. 1) changes the distribution of *S/N* ratios between the low and high audio frequencies. For instance with a 66 percent time constant the values of the *S/N* ratios are 4.8, 9.4, 9.6, 13.4, 21.6, 23.5 dB from 3150 to 415 Hz respectively with 2-bit companding.
‡ — indicates near-zero values

are synthesized we have noticed a one-third second (corresponding to 256 sixteen-bit words at 12 kbaud) of speech occasionally demanding as long as 20 minutes of Nova-800 minicomputer CPU time. This particular machine has an 800-nsec cycle time and hardware floating point multiply-divide facility. Conversely other one-third second speech

Table IV — Comparison of 16 kbits/sec gradient and 32 kbits/sec conventional coding

Bit rate:	16 kbits/sec		32 kbits/sec	
Coding:	Gradient		Conventional	
Audio frequency, Hz	Companding No. of bits	*S/N*, db	Companding No. of bits	*S/N*, dB
3950	2	6.1	3	9.0
	3	1.0	4	7.2
3150	2	4.7	3	12.0
	3	2.8	4	1.4
2900	2	6.5	3	9.7
	3	4.0	4	6.4
2490	2	8.5	3	11.9
	3	7.3	4	12.3
1660	2	15.3	3	19.6
	3	19.0	4	16.0
830	2	27.7	3	29.7
	3	28.4	4	26.0
415	2	36.8	3	38.5
	3	30.2	4	34.0

segments are synthesized in as little as four minutes. Averaged over three and a half minutes of speech synthesis, the computational time is roughly half an hour per second of real time speech signifying two thirds billion arithmetic operations† for every second of message. Stated alternatively one may expect one third million numerical functions between a typical peak and valley of the speech waveshape.

The computations during the silence periods are not trivial since gradient encoding is always alert to the incidence of the next peak (or valley). During the interval the dispersion of zeros and ones alternately is limited to that period which is too long to prepare the decoder for the next peak (or valley).

VII. REAL TIME IMPLEMENTATION

The real time implementation of gradient encoding is feasible in two distinct ways: (*i*) by a multiplicity of decoder circuits with feedback paths, each one being excited by a bit pattern of zeros or ones over a finite intervals and then selecting the pattern of the decoder which yields the waveform closest* to the waveform of the original speech or (*ii*) by one decoder circuit whose internal timing has been hastened dramatically by decreasing all the time constants in the circuit accordingly, and then

† This includes the modeling of all nonlinearities as they exist in the codec, the algebraic representation of most circuit elements, all the address computations, changing the synchronization rates between the scanning A/D converter and the codec clock rate, etc.

* Such as to maximize the S/N.

choosing that bit pattern which yields the waveform closest to that of the incoming speech. The former technique proposes that the data from the encoder transmitted to the final real time decoder is selected as the data of that particular real time decoder whose output came closest to that of the incoming speech. The latter technique proposes that the data of the encoder transmitted to the final real time decoder is selected as that binary combination of bits which had brought the waveform of the faster non-real-time decoder closest to that of the original speech.

Both of these techniques explore every branch of tree encoding to determine which one of the binary sequences yields the best performance. Our preliminary estimations show that eight decoder circuit for implementing technique (i) and an accelerated clock rate at about 100 kHz for a 12 kbaud data rate would be a reasonable compromise between complexity of the encoder design and optimality of bit configuration from the encoder design and optimality of bit configuration from the encoder. Such an arrangement is expected to enhance overall signal to noise ratio by 2 to 4 dB during the transmission of speech and the accelerated decoder circuits are completely capable of responding at about 100 kHz. Further tree encoding with 3-bit look-ahead option achieves most of the advantages obtained by these encoding schemes.

VIII. CONCLUSION

The success of the gradient encoding lies in the complete knowledge of the incoming speech waveshape. This prior information has been employed to optimize the bit pattern and thus reduce the bit rate. Attempts to reduce the storage requirement by one-half appear to have achieved an encouraging degree of success. The combined effects of sequential companding and gradient encoding are being investigated for bit rate reductions ranging between 60 and 66 percent for the same quality of speech.

REFERENCE

1. S. V. Ahamed, "Sequentially Companded ADM for Low Clock Rate Speech Coder Applications," B.S.T.J., this issue.

Contributors to This Issue

Thomas A. Abele, Dipl.-Ing. degree in electrical engineering and Ph.D. Ing. degree in electrical engineering in 1958 and 1961, respectively, from the Institute of Technology, Aachen, Germany. From 1958 to 1962 he was engaged in teaching and research at the Institute for High Frequency Techniques, Institute of Technology, Aachen, Germany. He then joined Bell Laboratories, where he has been concerned with the development and characterization of transmission components, first as a member of technical staff, then as a supervisor, and, since 1968, as head of a department. In 1973, while on a leave of absence, he spent a year as professor for microwave engineering at the Institute of Technology, Aachen, Germany. Senior Member, IEEE.

Syed V. Ahamed, B.E., 1957, University of Mysore; M.E., 1958, Indian Institute of Science; Ph.D., 1962, University of Manchester, U.K.; Post Doctoral Research Fellow, 1963, University of Delaware; Assistant Professor, 1964, University of Colorado; Bell Laboratories, 1966—. At Bell Laboratories, Mr. Ahamed has worked in computer-aided engineering analysis and design of electromagnetic components. He has designed and implemented minicomputer software and hardware interfacing. He has applied algebraic analysis to the design of domain circuits and investigated computer aids to the design of bubble circuits. He has investigated new varactor designs for microwave power in the c-band. He has developed hardware and software interfacing for audio frequency codecs. Since 1975, he has been optimizing codec designs, encoding techniques, and speech encoded data storage and manipulation by minicomputers.

W. N. Bell, B.S., 1967, Pratt Institute; M.S., 1969 (mathematics) Stevens Institute of Technology; Bell Laboratories, 1967–1973; New Jersey Bell, 1973–1975; Bell Laboratories, 1975—. Mr. Bell has worked on problems of crosstalk and inductive interference in telephone cables. He is presently a member of the Loop Topology and Methods Department and is concerned with loop plant utilization and construction budget analysis.

Martin B. Brilliant, B. A., 1955, Washington and Jefferson College; S.B., S.M., 1955, ScD., 1958, Massachusetts Institute of Technology; Bell Laboratories, 1955 and 1966 —. Mr. Brilliant has also held positions with the Air Force Cambridge Research Center; National Company, Inc.; Hazeltine Research Corporation; University of Kansas; and Booz Allen Applied Research, Inc. At Bell Laboratories, he worked in 1955 on a transistor pulse generator for the Electronic Central Office. Since 1966 he has been concerned with systems engineering problems in integrated digital switching and transmission and in network objectives. Member, AAAS.

Paul E. Butzien, B.S. (Electrical Engineering), 1961, Newark College of Engineering, and M.S. (E.E.), 1962, New York University; Pacific Telephone and Telegraph (Portland, Oregon), 1948–1953; Bell Laboratories, 1953—. Mr. Butzien has done research in electron tubes, superconductivity, and other cryogenic and high vacuum device studies, solid state microwave amplifiers, and phased array antenna elements. More recently he has been working on exploratory antenna measurements and is presently employing a data-gathering system of his design to study the spatial radiation characteristics of the Bell System pyramidal horn-reflector antenna. Member, IEEE.

T. C. Chu, B.S., 1964, Cheng Kung University (Taiwan); M. S., 1967, Syracuse University; Ph.D., 1971, Aerospace Engineering, Cornell University; Bell Laboratories, 1972—. Mr. Chu has worked the T4M digital transmission system, optical fiber connector and splice designs, and fiberguide transmission system. Member, OSA, Sigma Xi.

William G. French, B.A., 1965, University of California, Riverside; Ph.D., 1969, University of Wisconsin; Bell Laboratories, 1969—. Mr. French has worked on fundamental studies of glass as well as glass purification techniques and the development of low-loss optical fiber materials. His present interests are concerned with vapor deposition methods for the fabrication of low-loss fibers with low dispersion characteristics. Member of the Optical Society of America, American Chemical Society, and American Ceramic Society.

D. L. Jagerman, B.E.E., 1949, Cooper Union; M.S., 1954, and Ph.D. (mathematics), 1962, New York University; Bell Laboratories, 1963—. Mr. Jagerman has been engaged in mathematical research on quadrature, interpolation, and approximation theory. Also he has done research on the theory of widths and entropy with application to the storage and transmission of information. His work for the past several years concerns the theory of queueing systems applied to telephone traffic and computer flow time problems.

R. H. Knerr, B.S.E.E., 1960, Technical University, Aachen, Germany; Dipl. Ing., 1962, National Polytechnical Inst. (ENSEEHT), Toulouse, France; Ph.D. (E.E.), 1968, Lehigh University; Bell Laboratories, 1968—. Mr. Knerr has been engaged in R&D on microwave ferrite devices, IMPATT diode amplifiers, bipolar transistor amplifiers, GaAs FET microwave power amplifiers, and GaAs FET low-noise amplifiers. He is currently concerned with low-cost microwave receiver techniques. Senior member, IEEE; chairman IEEE-S-MTT Technical Committee on Microwave and Millimeter Wave Integrated Circuits; member IEEE-S-MTT Administrative Committee.

Benjamin F. Logan, Jr., B.S. (Electrical Engineering), 1946, Texas Technological College; M.S., 1951, Massachusetts Institute of Technology; Eng.D.Sc. (Electrical Engineering), 1965, Columbia University; Bell Laboratories, 1956—. While at MIT, Mr. Logan was a research assistant in the Research Laboratory of Electronics, investigating characteristics of high-power electrical discharge lamps. Also at MIT he engaged in analog computer development at the Dynamic Analysis and Control Laboratory. From 1955 to 1956 he worked for Hycon-Eastern, Inc., where he was concerned with the design of airborne power supplies. He joined Bell Laboratories as a member of the Visual and Acoustics Research Department, where he was concerned with the processing of speech signals. Currently, he is a member of the Mathematical Research Department. Member, Sigma Xi, Tau Beta Pi.

J. E. Mazo, B.S. (Physics), 1958, Massachusetts Institute of Technology; M.S. (Physics), 1960, and Ph.D. Department of Physics, University of Indiana, 1963–1964; Bell Laboratories, 1964—. At the University of Indiana, Mr. Mazo worked on studies of scattering theory. At Bell Laboratories, he has been concerned with problems in data transmission and is now working in the Mathematical Research Center. Member, American Physical Society, IEEE.

A. R. McCormick, RCA T3, 1972; Bell Labs, 1972—. Mr. McCormick has worked on FT3 optical transmission systems, phase locked loops with injection, optical connectors and optical communication systems. Member, IEEE.

V. Ramaswamy, B.Sc., 1957, Madras University, India; D.M.I.T., Madras Institute of Technology, Chromepet, Madras, India; M.S., 1962, and Ph.D., 1969, Northwestern University; Zenith Radio Corporation, 1962–65; Bell Laboratories, 1969—. His previous work has included microwave components, diode parametric amplifiers and wave propagation in semiconductor plasmas. His present research interests are thin-film optical waveguides and devices and polarization effects in single mode optical fibers. Member, Sigma Xi, IEEE.

J. E. Richard, A. Eng., 1962, Franklin Institute of Boston; Bell Laboratories, 1962—. Mr. Richard has performed RF propagation studies for a number of applications including mobile radio telephone, the New York-to-Washington high-speed train mobile telephone system, and early phases of the High Capacity Mobile Telephone System. Since 1971, he has been engaged in microwave antenna measurements at the Merrimack Valley Laboratory in Massachusetts.

Curtis A. Siller, Jr., B.S.E.E., 1966, M.S., 1967, and Ph.D., 1969, University of Tennessee; Bell Laboratories, 1969—. Mr. Siller has done exploratory work in linear antenna theory; designed, theoretically analyzed, and participated in the experimental testing of large-aperture microwave antennas; and assessed antenna system performance as it impacts on the terrestrial radio network. Most recently he has been investigating the relationship of environmental siting factors to antenna performance degradation, studying the effect of coating antenna radomes with hydrophobic materials, and pursuing methods to ameliorate signal impairment during periods of multipath fading. Member, Phi Eta Sigma, Eta Kappa Nu, Tau Beta Pi, Phi Kappa Phi, and Sigma Xi.

R. D. Standley, B.S., 1957, University of Illinois; M.S., 1960, Rutgers University; Ph.D., 1966, Illinois Institute of Technology; USASRDL, Ft. Monmouth, N.J., 1957–1960; IIT Research Institute, Chicago, 1960–1966; Bell Laboratories, 1966—. Mr. Standley has been engaged in research projects involving microwave, millimeter wave, and optical components. He is presently concerned with electron beam lithography as applied to fabrication of integrated optic devices. Member, IEEE, Sigma Tau, Sigma Xi.

C. B. Swan, B.Sc., 1954, University of New Brunswick, Canada; M.A.Sc., 1959, University of Toronto; Ph.D., 1963, University of Toronto; Bell Laboratories 1962—. From 1962 to 1968 Mr. Swan studied the interaction of microwaves with ionized gases and developed varactor diode harmonic generators and IMPATT diode oscillators. Since 1969 he has supervised the Microwave Integrated Circuit and Device Group at the Allentown, Pennsylvania, Laboratories. Senior member, IEEE; member, Association of Professional Engineers of Ontario.

Deborah Y. Sze will receive her A.B. from the Division of Engineering and Applied Physics of Harvard University in 1978. She is a student member of the Harvard Society of Engineers and Scientists and of the National Society of Professional Engineers. She was a Summer Research Associate at Bell Laboratories in 1976.