THE BELL SYSTEM
TECHNICAL JOURNAL
DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION
Volume 49 February 1970 Number 2

On the Interaction of Roundoff Noise and Dynamic Range in Digital Filters*

By LELAND B. JACKSON

(Manuscript received October 22, 1969)

The interaction between the roundoff-noise output from a digital filter and the associated dynamic-range limitations is investigated for the case of uncorrelated rounding errors from sample to sample and from one error source to another. The required dynamic-range constraints are derived in terms of L_p norms of the input-signal spectrum and the transfer responses to selected nodes within the filter. The concept of "transpose configurations" is introduced and is found to be quite useful in digital-filter synthesis; for although such configurations have identical transfer functions, their roundoff-noise outputs and dynamic-range limitations can be quite different, in general. Two transpose configurations for the direct form of a digital filter are used to illustrate these results.

I. INTRODUCTION

With the rapid development of digital integrated circuits in the 1960's and the potential for large-scale integration (LSI) of these circuits in the 1970's, digital signal processing has become much more than a tool for the simulation of analog systems or a technique for the implementa-

* This paper is taken in part from a thesis submitted by Leland B. Jackson in partial fulfillment of the requirements for the degree of Doctor of Science in the Department of Electrical Engineering at Stevens Institute of Technology.[1]

tion of very complex and costly one-of-a-kind systems alone. The traditional advantages of digital systems, such as high accuracy, stable parameter values, and straight-forward realization, have been supplemented through the use of integrated circuits by the additional advantages of high reliability, small circuit size, and ever-decreasing cost. As a result, it now appears that many signal processing systems which have been in the exclusive domain of analog circuits may in the future be implemented using digital circuits; while other proposed systems which could not be implemented at all because of the practical limitations of analog circuits may now be realized with digital circuits.[2]

The key element in most of these new signal-processing systems is the digital filter. The term "digital filter" here denotes a time-invariant, discrete or sampled-data filter with finite accuracy in the representation of all data and parameter values.[3-5] That is, all data and parameters within the filter are "quantized" to a finite set of allowable values with, in general, some form of error being incurred as a result of the quantization process. Implicit in this quantization is a maximum value or set of maximum values for the magnitudes of these data and parameters which, in the case of the data, is usually referred to as the "dynamic range" of the filter.

Without the above quantization effects, linear discrete filters could be implemented exactly. Of course, one very significant feature of digital signal processing is that arbitrarily high accuracy can, in fact, be maintained once the initial analog-to-digital (A-D) conversion (if any) has taken place. However, there are still practical limitations to the accuracy of any physical system, and often it is desirable to minimize the accuracy of the implementation (while still satisfying the system specifications) in order to minimize the cost of the system. Hence, a thorough understanding of quantization errors in digital filters is quite important if the full potential of digital signal processing is ever to be realized.

II. QUANTIZATION ERRORS IN DIGITAL FILTERS

The specific sources of quantization error in the implementation and operation of a digital filter are as follows:

(*i*) The filter coefficients (multiplying constants) must be quantized to some finite number of digits (usually binary digits, or bits).

(*ii*) The input samples to the filter must also be quantized to a finite number of digits.

(*iii*) The products of the multiplications (of data by coefficients)

within the filter must usually be rounded or truncated to a smaller number of digits.

(*iv*) When floating-point arithmetic is used, rounding or truncation must usually be performed before or after additions as well.

The first source of error above is deterministic and straightforward to analyze in that the filter characteristics must simply be recomputed to reflect the (small) changes in the filter coefficients due to quantizing.[6,7] However, the inclusion of coefficient quantization in the initial filter synthesis procedure in order to minimize (in some sense) the resulting filter complexity produces a complex problem in nonlinear integer programming which has only begun to be investigated.

The second source of error is often referred to as "quantization noise". It is inherent in any A-D conversion process and has been studied in great depth.[8] Hence, input quantization has not been included in our investigation, except as it relates to other error sources of interest.

The third and fourth error sources are similar to the second since they also involve quantization of the data, but they differ in two respects: (i) The data to be quantized is already digital in form, and (ii) the rounding or truncation of the data takes place at various points *within* the filter, not just at its input. To distinguish these sources of error from the input quantization noise, the resulting error processes will be referred to as "roundoff noise" (to be used generically, whether rounding or truncation is actually employed). Because of (ii), the roundoff noise is potentially much larger than the input quantization noise, and it is one of the principal factors which determine the complexity of the digital filter implementation, especially when special-purpose hardware is used.

There are three variables in the filter implementation which determine the level and character of the roundoff noise for a given input signal:

(*i*) the number of digits (bits) used to represent the data within the filter,

(*ii*) the "mode" of arithmetic employed (that is, fixed-point or floating-point), and

(*iii*) the circuit configuration of the digital filter. The number of digits in the data may be thought of as determining either the quantization step size or the dynamic range of the filter. We choose here the latter interpretation in order to have the same step size for all filters. Therefore, with this interpretation, the number of data digits does not affect the level of the roundoff noise directly, but rather it limits the maximum allowable signal level and hence the realizable signal-to-noise ratio. Data within the filter must, of course, be properly "scaled" if the

maximum signal-to-noise ratio is to be maintained without exceeding the dynamic-range limitations. Among the principal results reported here are the determination of appropriate scaling for certain important classes of input signals and the calculation of the effect of this scaling on the output roundoff noise.

The output roundoff noise from a floating-point digital filter is usually (but not always) less than that from a fixed-point filter with the same total number of data digits because of the automatic scaling provided by floating-point arithmetic.[9,10] However, since floating-point arithmetic is significantly more complex and costly to implement, most special-purpose digital filters have been, and will probably continue to be, constructed with fixed-point hardware. Hence, we have considered only fixed-point digital filters in this work although much of the analysis could be adapted to floating-point filters. Oppenheim has recently proposed another interesting mode of arithmetic for digital filter implementation, called "block-floating-point", which provides a simplified form of automatic scaling of the filter data.[11] As would be expected, the performance of block-floating-point appears to lie somewhere between those of fixed-point and of floating-point.

The third variable in the implementation of a digital filter, that of circuit configuration, is the principal factor determining the character (spectrum) of the output roundoff noise and, along with mode of the arithmetic, ultimately determines the number of data digits required to satisfy the performance specifications. In fact, the key step in the synthesis of a digital filter is the selection of an appropriate configuration for the digital circuit. There are a multitude of equivalent circuit configurations for any given linear *discrete* filter (whose transfer function is expressible as a rational fraction in z); but in the implementation of the corresponding *digital* filter, these configurations are no longer equivalent, in general, because of the effects of coefficient quantization and roundoff noise. As noted previously, the effects of coefficient quantization are deterministic and can thus be accounted for exactly as a (typically small) change in the transfer function of the discrete filter. Therefore, assuming that the coefficients for the configurations under consideration have been (or can be) quantized satisfactorily, the choice between these configurations is then determined by the level and character of their output roundoff noise. As we will show, there can be very significant differences between the roundoff-noise outputs of otherwise equivalent digital filter configurations.

The content and complexity of any analysis of roundoff noise are determined to a large extent by the assumed correlation between round-

off errors. If these errors may be assumed to be uncorrelated from sample to sample and from multiplier (or other rounding point) to multiplier, then the roundoff-noise analysis is relatively straightforward, and the results are independent of the exact nature of the input signal to the filter. If, on the other hand, uncorrelated errors may not be assumed, then the analysis is much more complex, and the results are generally dependent on the particular input signal or class of input signals. This paper is concerned exclusively with the uncorrelated-error case because this assumption seems to be valid for most filters with input signals of reasonable amplitude and spectral content. Even in this case, the inclusion of the associated dynamic-range constraints makes the analysis reasonably involved and the corresponding synthesis problem quite complex.

Although the generic term "roundoff noise" has been used to include the case of truncation as well as rounding, we actually concentrate on the rounding case. As long as the assumption of uncorrelated errors can be made, our results are applicable to either case, with the error variance for truncation being four times that for rounding. However, as the input signals become less "random", the uncorrelated-error assumption tends to break down for truncation more readily than for rounding. Hence, additional care must be exercised in applying these results to the truncation case.

III. FILTER MODEL FOR UNCORRELATED-ROUNDOFF-NOISE ANALYSIS

The analyses appearing in the literature concerning roundoff noise in digital filters usually employ the simplifying and often reasonable assumption of uncorrelated roundoff errors from sample to sample and from one error source (multiplier or other rounding point) to another.[9,12,13] This assumption is based on the intuitively plausible and experimentally supported notion that for sufficiently large and dynamic signals within the filter, the small roundoff error made at one point in the network and/or in time should have little relationship to (that is, correlation with) the roundoff error made at any other point in the network and/or time. The advantage of assuming uncorrelated errors from one sample to another is that the noise injected into the filter by each rounding operation is then "white"; while the advantage of assuming uncorrelated error sources is that the output noise power spectrum may then be computed as simply the superposition of the (filtered) noise spectra due to the separate error sources.[12] Experimental results which support the validity of this assumption, even in the case

of a single sinusoidal input, are presented in Ref. 1. In this section, we introduce the notation and develop the analysis pertaining to uncorrelated roundoff noise for later use in investigating the synthesis of digital filters.

Digital filter networks are composed of three basic elements: adders, constant multipliers, and delays. The interconnection of these elements into a particular network configuration is the key step in digital filter synthesis. For our purposes here, we need only consider the network as a directed graph, with the multipliers and delays being represented by graph branches. The branch interconnection points, or nodes, will be divided into two types: "summation nodes", which correspond to the adders and have multiple inputs and a single output, and "branch nodes", which correspond to simple "wired" interconnections that have a single input and one or more outputs.

A digital filter network may thus be represented as shown in Fig. 1. The input to and output from the filter at time $t = nT$ are denoted by $u(n)$ and $y(n)$, respectively. The corresponding output from the i^{th} branch node is denoted by $v_i(n)$; while the roundoff error introduced into the filter at the j^{th} summation node is denoted by $e_j(n)$. Since with fixed-point arithmetic, rounding is performed only after multiplications, non-zero roundoff errors are "input" to the filter only at those summation nodes which follow constant (non-integer) multiplier branches, as depicted in Fig. 2.

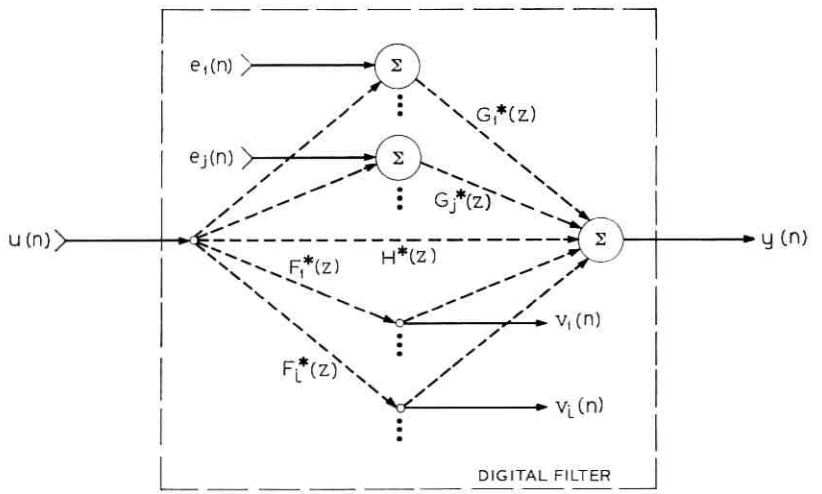


Fig. 1 — General digital filter model.

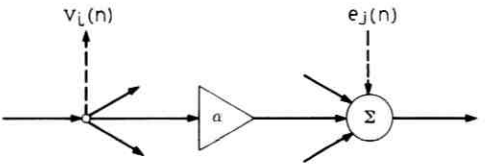


Fig. 2 — Constant multiplier with preceding branch node and succeeding summation node.

For a unit sample input to the filter at $t = 0$ and no rounding [that is, $u(0) = 1$, $u(n) = 0$ for $n \neq 0$, and $e_j(n) = 0$ for all j and n], the resulting output values $y(n)$ and $v_i(n)$ for all $n \geqq 0$ and all i are designated as $h(n)$ and $f_i(n)$, respectively. Alternatively, for a unit sample input to the j^{th} summation node and zero inputs otherwise [that is, $e_j(0) = 1$, $e_j(n) = 0$ for $n \neq 0$, and $e_k(n) = u(n) = 0$ for all n and for $k \neq j$], the resulting output values $y(n)$ for all $n \geqq 0$ are denoted by $g_j(n)$. We thus have the following transfer functions of interest, expressed in z-transform form:

From filter input to output:

$$H^*(z) = \sum_{n=0}^{\infty} h(n)z^{-n}. \tag{1}$$

From filter input to i^{th} branch-node output:

$$F_i^*(z) = \sum_{n=0}^{\infty} f_i(n)z^{-n}. \tag{2}$$

From j^{th} summation-node input to filter output:

$$G_j^*(z) = \sum_{n=0}^{\infty} g_j(n)z^{-n}. \tag{3}$$

These transfer functions are indicated in Fig. 1.

The frequency responses (Fourier transforms) corresponding to the above transfer functions are given by [3-5]

$$H(\omega) = H^*(e^{j\omega T}), \tag{4}$$

$$F_i(\omega) = F_i^*(e^{j\omega T}), \tag{5}$$

$$G_k(\omega) = G_k^*(e^{j\omega T}). \tag{6}$$

This notation will be used throughout this paper. That is, for any z-transform $A^*(z)$ which converges for $|z| = 1$, the corresponding Fourier transform is given by

$$A(\omega) = A^*(e^{j\omega T}).$$

If scaling has been included in the filter design in order to satisfy certain dynamic-range constraints, then prime marks (′) are added to denote this fact [for example, $F_i'(\omega)$, $F_j^{*\prime}(z)$].

Each error source (rounding operation) within the filter is assumed to inject white noise of uniform power-spectral density N_0. Assuming uniformly distributed rounding errors with zero mean, the variance of the roundoff noise from each error source is given by[12,13]

$$\sigma_0^2 = \Delta^2/12 \tag{7}$$

where Δ is the spacing of the quantization steps (after rounding). To eliminate the sampling period T from certain expressions of interest, we now define $N_0 = \sigma_0^2$. Hence, the variance, or total average power, corresponding to an arbitrary power-density spectrum $N(\omega)$ with no DC component (which implies a zero-mean process) is given by†

$$\sigma^2 = \frac{1}{\omega_s} \int_0^{\omega_s} N(\omega)\, d\omega \tag{8}$$

where ω_s is the radian sampling frequency given by

$$\omega_s = 2\pi/T. \tag{9}$$

Assume now that k_j error sources input to the j^{th} summation node. The spectral density of the roundoff error sequence $\{e_j(n)\}$ is then just $k_j N_0$ by our assumption of uncorrelated error sources. The total roundoff noise in the output of the filter thus has a power-density spectrum given by[12]

$$N_y(\omega) = \sigma_0^2 \sum_j k_j \mid G_j(\omega) \mid^2 \tag{10a}$$

where we have substituted σ_0^2 for N_0. If scaling has been included in the filter design, then the corresponding expression is just

$$N_y(\omega) = \sigma_0^2 \sum_j k_j' \mid G_j'(\omega) \mid^2 \tag{10b}$$

where $k_j' \geqq k_j$ to account for the additional scaling multipliers.

IV. DYNAMIC-RANGE CONSTRAINTS

The ultimate objective of the synthesis procedures to be investigated will be the minimization of some norm of $N_y(\omega)$ for a given quantization step size Δ, subject to certain "constraints". One constraint is that the

† This normalization of $N(\omega)$ is further motivated by the derivation in Section V leading to equation (30b).

specified transfer function $H^*(z)$ must be maintained. Another fundamental, but often overlooked, constraint is the finite dynamic range of the filter. Specifically, the signals $v_i(n)$ at certain branch nodes within the filter cannot be allowed to "overflow" (that is, exceed the dynamic-range limitations), at least not more than some small percentage of the time, in order to prevent severe distortion in the filter output.

Overflow constraints are required only at certain branch nodes in the digital circuit because it is only the inputs to the constant multipliers which cannot be allowed to overflow when several standard numbering systems are used (for example, one's- or two's-complement binary).[14] Specifically, in the summation of more than two numbers, if the magnitude of the correct total sum is small enough to allow its representation by the K available digits, then in these numbering systems the correct total sum will be obtained regardless of the order in which the numbers are added, even if an overflow occurs in one of the partial sums. Hence, those node outputs which correspond to partial sums comprising a larger total sum may be allowed to overflow, as long as the total sum is constrained not to overflow. This property also applies when one of the inputs to a summation node has overflowed as a result of a multiplication by a coefficient of magnitude greater than one.

Turning to the formulation of the required overflow constraints, we may easily derive an upper bound on the magnitude of the signals $v_i(n)$ for all possible input sequences $\{u(n)\}$, neglecting the (small) error signals $e_i(n)$. Assuming zero initial conditions in the filter and $e_j(n) = 0$ for all j and n, the i^{th} branch-node output $v_i(n)$ is given by

$$v_i(n) = \sum_{k=0}^{\infty} f_i(k)u(n-k), \quad \text{all} \quad n. \tag{11}$$

Therefore, given that $u(n)$ is bounded in magnitude by some number M for all n, an upper bound on the magnitude of $v_i(n)$ is given by[15]

$$|v_i(n)| \leqq M \sum_{k=0}^{\infty} |f_i(k)|, \quad \text{all} \quad n. \tag{12}$$

Thus, if the node signal $v_i(n)$ is also to be bounded in magnitude by M for all possible input sequences, the associated scaling must ensure that

$$\sum_{k=0}^{\infty} |f_i'(k)| \leqq 1. \tag{13}$$

That (13) is not only a sufficient condition to rule out overflow for all possible input sequences $\{u(n)\}$, but also a necessary condition, is easily

shown by letting $u(n) = \pm M$ for all n, with sgn $[u(n_0 - k)] =$ sgn $[f_i(k)]$ for some $n = n_0$ and all $k \geqq 0$. Then from equation (11) we see that (12) is satisfied with equality in this case, and thus (13) is a necessary condition, as well.

The norm of $f_i'(k)$ employed in (13) is not very useful in practice because of the difficulty of evaluating the indicated summation in all but the simplest cases. Also, for large classes of input signals, (12) and thus (13) are overly pessimistic. Therefore, we now derive alternate conditions on (the transform of) the scaled unit-sample response $\{f_i'(n)\}$ which ensure that for certain classes of input signals, the corresponding branch-node output $v_i(n)$ cannot overflow. The derivation of these conditions for discrete systems closely parallels the corresponding derivation for continuous systems, as given by Papoulis.[16]

An alternate expression for equation (11) in terms of z-transforms is derived as follows: Consider an (absolutely summable) deterministic input sequence $\{u(n)\}$ possessing the z-transform

$$U^*(z) = \sum_{n=-\infty}^{\infty} u(n)z^{-n}, \qquad a < |z| < b, \tag{14}$$

for some $a < 1$ and $b > 1$. Stability requires that $F_i^*(z)$, defined in equation (2), exist for all $|z| > c$ for some $c < 1$. Hence, the z-transform of $\{v_i(n)\}$ is given by[3]

$$V_i^*(z) = F_i^*(z)U^*(z), \qquad d < |z| < b, \tag{15}$$

where $d = \max(a, c)$. The inverse transform of equation (15) is given by[3]

$$v_i(n) = \frac{1}{2\pi j} \oint_\Gamma V_i^*(z)z^{n-1}\, dz \tag{16}$$

where the contour of integration Γ is contained in the region of convergence $d < |z| < b$. Since $d < 1$ and $b > 1$, let Γ be the unit circle in the z plane ($|z| = 1$), and perform the change of variables $z = e^{j\omega T}$ in equation (16). Using equation (15), the resulting equation becomes

$$v_i(n) = \frac{1}{\omega_s} \int_0^{\omega_s} F_i(\omega)U(\omega)e^{jn\omega T}\, d\omega. \tag{17}$$

The conditions to be derived from equation (17) are most easily expressed in terms of L_p norms, defined for an arbitrary periodic function $A(\cdot)$ with period ω_s by[17]

$$\| A \|_p = \left[\frac{1}{\omega_s} \int_0^{\omega_s} |A(\omega)|^p\, d\omega\right]^{1/p} \tag{18a}$$

for each real $p \geqq 1$ such that

$$\int_0^{\omega_s} | A(\omega) |^p \, d\omega < \infty .$$

It can be shown[17] that for $A(\cdot)$ continuous, the limit of equation (18a) as $p \to \infty$ exists and is given by

$$|| A ||_\infty = \max_{0 \leq \omega \leq \omega_s} | A(\omega) | . \tag{18b}$$

Assume now that $| U(\omega) |$ is bounded from above by some number M (that is, $|| U ||_\infty \leqq M$). Then, from equation (17),

$$| v_i(n) | \leqq M \frac{1}{\omega_s} \int_0^{\omega_s} | F_i(\omega) | \, d\omega$$

or

$$| v_i(n) | \leqq || F_i ||_1 \cdot || U ||_\infty . \tag{19}$$

In exactly the same manner, we may also show that

$$| v_i(n) | \leqq || F_i ||_\infty \cdot || U ||_1 . \tag{20}$$

Applying the Schwarz inequality to equation (17), on the other hand, yields that

$$| v_i(n) |^2 \leqq \frac{1}{\omega_s^2} \int_0^{\omega_s} | F_i(\omega) |^2 \, d\omega \int_0^{\omega_s} | U(\nu) |^2 \, d\nu$$

or

$$| v_i(n) | \leqq || F_i ||_2 \cdot || U ||_2 . \tag{21}$$

Note that (19), (20), and (21) are all of the form

$$| v_i(n) | \leqq || F_i ||_p \cdot || U ||_q , \qquad \left(\frac{1}{p} + \frac{1}{q} = 1 \right) \tag{22}$$

for $p, q = 1, 2$, and ∞. It can be shown[18] that (22) is true in general for all $p, q > 1$ satisfying $1/p + 1/q = 1$; and we have shown in (19) and (20) that if the L_∞ norms exist, then (22) holds for $p, q = 1$, as well. The general relation in (22) for all $p, q > 1$, is derived from Holder's inequality.

A simple, but important special case of (22) results from letting $F_i^*(z) = F_i(\omega) = 1$. Since $|| 1 ||_p = 1$ for all $p \geqq 1$, we then have simply

$$| u(n) | \leqq || U ||_q , \quad \text{all} \quad q \geqq 1. \tag{23}$$

But since (23) holds for all sequences $\{u(n)\}$, it must also be true that

$$| v_i(n) | \leqq || V_i ||_r , \quad \text{all} \quad r \geqq 1.$$

This is, in fact, the basis of (22), for Holder's inequality actually states that

$$|| V_i' ||_1 \leqq || F_i ||_p || U ||_q , \qquad \left(\frac{1}{p} + \frac{1}{q} = 1\right).$$

Therefore, the real implication of (22) is that the mean absolute value of $V_i(\omega)$ is bounded by $|| F_i ||_p || U ||_q$, and this, in turn, provides a bound on $| v_i(n) |$.

Assume, therefore, that the input transform $U(\omega)$ satisfies $|| U ||_q \leqq M$ for some $q \geqq 1$. From (23) we immediately have that $| u(n) | \leqq M$ for all n. Then, if $| v_i(n) |$ is also to be bounded by M, (22) provides a sufficient condition on the scaling to ensure this, namely

$$|| F_i' ||_p \leqq 1, \qquad (|| U ||_q \leqq M) \tag{24}$$

for $p = q/(q - 1)$. Inequality (24) is the desired condition to replace the more general, but often less useful condition given by (13).

From an engineering viewpoint, the most significant values for p and q would seem to be 1, 2, and ∞. The case $p = 1$, $q = \infty$ requires that the input transform $U(\omega)$ be everywhere bounded in magnitude by M (that is, $|| U ||_\infty \leqq M$), in which case only the L_1 norm of the scaled transfer function $F_i'(\omega)$ need satisfy (24). For an input of finite energy $E = \sum_n u^2(n)$, Parseval's identity implies that $|| U ||_2^2 = E$, and thus with $M \geqq (E)^{\frac{1}{2}}$, (24) can be satisfied for $p = q = 2$.

The case of $p = \infty$, $q = 1$ in (24) implies the most stringent condition on $F_i'(\omega)$ because from equation (18) it is evident that

$$|| F_i' ||_p \leqq || F_i' ||_\infty \tag{25}$$

for all $p \geqq 1$. It is clear, for example, that for a sinusoidal input of amplitude $A \leqq M$ and arbitrary frequency ω_0, we must have $| F_i'(\omega) | \leqq 1$ for all ω (that is, $|| F_i' ||_\infty \leqq 1$) to ensure that $| v_i(n) | \leqq M$ for all n. However, a sinusoidal input sequence $\{u(n)\}$ is not absolutely summable, and thus $U^*(z)$ as defined in equation (14) does not exist in this case. This difficulty may be circumvented, as is common in Fourier analysis, by assuming a finite sequence of length N and then passing to the limit as $N \to \infty$. The resulting (Fourier) transform of $\{u(n)\}$ is of the form

$$U_0(\omega) = \frac{A}{2} e^{j\theta}[\delta(\omega - \omega_0) + \delta(\omega - \omega_s + \omega_0)], \qquad (0 \leqq \omega \leqq \omega_s) \tag{26}$$

where $\delta(\omega)$ is the familiar Dirac delta function defined by

$$\delta(\omega) = 0, \qquad \omega \neq 0,$$
$$\int_{-\infty}^{\infty} \delta(\omega)\, d\omega = 1. \tag{27}$$

$U_0(\omega)$ is, of course, periodic in ω with period ω_s. From equations (18a), (26), and (27), we immediately have that $\| U_0 \|_1 = A \leqq M$, and thus with $p = \infty$, (24) is applicable for sinusoidal input sequences, as expected.

V. RANDOM INPUT CASE

In the case of random input sequences, (24) is not directly applicable because the z-transform $U^*(z)$ is not defined. Similar conditions may be obtained, however, by considering the discrete autocorrelation function $\varphi(\cdot)$, defined for a (wide-sense) stationary sequence $\{w(n)\}$ by

$$\varphi_w(m) = E[w(n)w(n + m)] \tag{28}$$

where $E[\cdot]$ is the expected-value operator. A z-transform $\Phi_w^*(z)$ may be defined for the sequence $\{\varphi_w(m)\}$ as in equation (14) with an inverse transform as in (16). Assuming ergodicity and a zero mean ($E[w(n)] = 0$) for $\{w(n)\}$, we immediately have from equation (28) that the variance, or total average power, of $\{w(n)\}$ is given by

$$\varphi_w(0) = E[w^2(n)] = \sigma_w^2, \tag{29}$$

and from equation (16) we also have

$$\varphi_w(0) = \frac{1}{2\pi j} \oint_\Gamma \Phi_w^*(z) z^{-1}\, dz. \tag{30a}$$

Letting Γ be the unit circle ($z = e^{j\omega T}$), equations (29) and (30a) imply that

$$\sigma_w^2 = \frac{1}{\omega_s} \int_0^{\omega_s} \Phi_w(\omega)\, d\omega. \tag{30b}$$

Hence, from equation (8) we see that $\Phi_w(\omega)$ is just the power-density spectrum of the sequence $\{w(n)\}$.

For an input sequence $\{u(n)\}$ whose autocorrelation function has the z-transform $\Phi_u^*(z)$, it is well-known that the corresponding transform for the output $\{v_i(n)\}$ is given by

$$\Phi_{v_i}^*(z) = F_i^*(z) F_i^*(z^{-1}) \Phi_u^*(z) \tag{31a}$$

or

$$\Phi_{v_i}(\omega) = | F_i(\omega) |^2 \Phi_u(\omega). \tag{31b}$$

Equations (29) through (31) imply then that

$$\sigma_{v_i}^2 = \frac{1}{\omega_s} \int_0^{\omega_s} | F_i(\omega) |^2 \Phi_u(\omega) \, d\omega. \tag{32}$$

Since equation (32) is of the same basic form as (17), a derivation similar to that leading to (22) must yield the following relations for $p, q \geqq 1$:

$$\sigma_{v_i}^2 \leqq || F_i^2 ||_p \cdot || \Phi_u ||_q , \qquad \left(\frac{1}{p} + \frac{1}{q} = 1\right) \tag{33a}$$

or, from equation (17),

$$\sigma_{v_i}^2 \leqq || F_i ||_{2p}^2 \cdot || \Phi_u ||_q , \qquad \left(\frac{1}{p} + \frac{1}{q} = 1\right). \tag{33b}$$

Two cases of (33) are of particular interest, namely

$$\sigma_{v_i}^2 \leqq || F_i ||_2^2 \cdot || \Phi_u ||_\infty \tag{34}$$

and

$$\sigma_{v_i}^2 \leqq || F_i ||_\infty^2 \cdot || \Phi_u ||_1 . \tag{35}$$

In view of equation (25), we see that (34) implies the most stringent condition on the input spectrum $\Phi_u(\omega)$, whereas (35) yields the most stringent condition on the transfer function $F_i(\omega)$. From (34) and (30b), for example, we have that if the input power-density spectrum is "white" [that is, $\Phi_u(\omega) = \sigma_u^2$ for all ω], then $\sigma_{v_i}^2 \leqq || F_i ||_2^2 \sigma_u^2$. Hence, if the input sequence $\{u(n)\}$ is a Gaussian process,[19] the node output sequence $\{v_i(n)\}$ will overflow no more (in percentage of time) than does the input, provided only that

$$|| F_i' ||_2 \leqq 1. \tag{36}$$

The inequality in (35) requires, on the other hand, that for an input sinusoid of arbitrary amplitude and frequency, $F_i'(\omega)$ must satisfy

$$|| F_i' ||_\infty \leqq 1 \tag{37}$$

to ensure against overflow, as we have seen earlier from (24).

To summarize, dynamic-range constraints of the form

$$|| F_i' ||_p \leqq 1, \qquad p \geqq 1 \tag{38}$$

have been derived for both deterministic and random inputs, where

$F_i'(\omega)$ is the (scaled) transfer response from the filter input to the i^{th} branch node and $|| \cdot ||_p$ denotes the L_p norm defined in equation (18). For a deterministic input with amplitude spectrum $U(\omega)$, (38) assumes that

$$|| U ||_q \leqq M, \qquad q = \frac{p}{p-1}, \tag{39}$$

where M is the maximum allowable signal amplitude. For a random input, on the other hand, the use of (38) requires appropriate conditions on $|| \Phi_u ||_r$, $r = p/(p-2)$ and $p \geqq 2$, where $\Phi_u(\omega)$ is the power-density spectrum of the input sequence.

The effect of (38) and (39) is to bound the mean absolute value of the amplitude spectrum at the i^{th} branch node (that is, $|| V_i ||_1$) which, in turn, bounds the peak signal amplitude at that node. The use of (38) in conjunction with (33), however, bounds only the average power at the i^{th} branch node, and thus the relationship between this average power and the peak signal amplitude at the node must also be determined in order to provide an effective dynamic-range constraint.

VI. TRANSPOSE SYSTEMS

In the evaluation of different circuit configurations for a given digital filter, a useful concept relating certain of these configurations is that of "transpose configurations". This relationship is a general property of linear graphs[20] and will be presented here in terms of a state-variable formulation.

The general state equations for a linear, time-invariant discrete system are given by[21]

$$\begin{aligned} \mathbf{x}(n+1) &= A\mathbf{x}(n) + B\mathbf{u}(n), \\ \mathbf{y}(n) &= C\mathbf{x}(n) + D\mathbf{u}(n) \end{aligned} \tag{40}$$

where $\mathbf{x}(n)$ is an N-dimensional vector describing the state of the system at time $t = nT$, $\mathbf{u}(n)$ is the corresponding J-dimensional input vector, $\mathbf{y}(n)$ is the corresponding I-dimensional output vector, and A, B, C, and D are fixed parameter matrices of the appropriate dimensions relating the input, state, and output vectors as given by equation (40). The $(N + I) \times (N + J)$ matrix S defined by

$$S = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{41}$$

provides a convenient single parameter matrix which describes the complete discrete system.

A transfer function matrix $\mathfrak{IC}_S^*(z)$ may be defined for the system (described by) S relating the input and output vector sequences $\{\mathbf{u}(n)\}$ and $\{\mathbf{y}(n)\}$ by

$$\mathbf{Y}^*(z) = \mathfrak{IC}_S^*(z)\mathbf{U}^*(z) \tag{42}$$

where $\mathbf{U}^*(z)$ and $\mathbf{Y}^*(z)$ are the vector z-transforms of $\{\mathbf{u}(n)\}$ and $\{\mathbf{y}(n)\}$, respectively. $\mathfrak{IC}_S^*(z)$ is readily shown to be given by[21]

$$\mathfrak{IC}_S^*(z) = C(zI - A)^{-1}B + D \tag{43}$$

where $(\cdot)^{-1}$ denotes the matrix inverse and I is the N-dimensional identity matrix.

Consider now a new system which is described by the parameter matrix S^t, that is,

$$S^t = \begin{bmatrix} A^t & C^t \\ B^t & D^t \end{bmatrix} \tag{44}$$

where $(\cdot)^t$ denotes the matrix transpose. From equations (41) and (43) it is easily seen that the transfer function matrix for the new system S^t is given by

$$\begin{aligned} \mathfrak{IC}_{S^t}^*(z) &= B^t(zI - A^t)^{-1}C^t + D^t \\ &= [\mathfrak{IC}_S^*(z)]^t. \end{aligned} \tag{45}$$

Thus, the transfer function matrix for the system S^t is simply the transpose of the transfer function matrix for the system S. That is, the element $H_{ij}^*(z)$ from $\mathfrak{IC}_S^*(z)$, which is the transfer function from the j^{th} input to the i^{th} output of system S, equals the element $H_{ji}^{*t}(z)$ from $\mathfrak{IC}_{S^t}^*(z)$, that is, the transfer function from the i^{th} input to the j^{th} output of S^t. Note also that while the system S has a total of J inputs and I outputs, the system S^t has I inputs and J outputs.

The concept of transpose systems will be particularly useful to us in conjunction with the digital-filter model introduced in Section III and depicted in Fig. 1. Defining the input and output vectors for the filter by

$$\mathbf{u}(n) = \begin{bmatrix} u(n) \\ e_1(n) \\ \vdots \\ e_J(n) \end{bmatrix} \quad \text{and} \quad \mathbf{y}(n) = \begin{bmatrix} y(n) \\ v_1(n) \\ \vdots \\ v_I(n) \end{bmatrix} \tag{46}$$

respectively, the transfer function matrix for the filter is given by

$$\mathfrak{C}^*(z) = \begin{bmatrix} H^*(z) & G_1^*(z) & \cdots & G_J^*(z) \\ F_1^*(z) & \text{——} & \text{——} & \text{——} \\ \vdots & \text{——} & \text{——} & \text{——} \\ F_I^*(z) & \text{——} & \text{——} & \text{——} \end{bmatrix} \tag{47}$$

where the specific expressions for the elements in other than the first row and first column are unimportant for our purposes. By equation (45), the transfer function matrix for the corresponding transpose system is then simply

$$\mathfrak{C}_t^*(z) = \begin{bmatrix} H^*(z) & F_1^*(z) & \cdots & F_I^*(z) \\ G_1^*(z) & \text{——} & \text{——} & \text{——} \\ \vdots & \text{——} & \text{——} & \text{——} \\ G_J^*(z) & \text{——} & \text{——} & \text{——} \end{bmatrix}. \tag{48}$$

Note, in particular, that the transfer function from input-1 to output-1 [that is, $H^*(z)$, the ideal transfer function from filter input to filter output] is the same for both systems.

As discussed more fully in Ref. 1, the circuit configuration realizing a given system S is not necessarily unique, and hence neither is the configuration for the transpose system S^t. However, given a particular configuration for the system S, a unique "transpose configuration", which realizes S^t, may be derived from the given configuration for S by simply reversing the direction of all branches in the given network! In particular, then, all delays and constant multipliers remain the same except for the change in direction. All summation nodes in the given configuration become branch nodes in the transpose configuration, and all branch nodes become summation nodes. Likewise, all inputs in the given configuration become outputs in the transpose configuration, and all outputs become inputs.†

That the transpose configuration defined above actually realizes the transpose system S^t is easily seen by considering the state equations in (40). The constant multiplier(s) corresponding to the element d_{ij} of the matrix D and relating the j^{th} input and the i^{th} output of the original configuration must relate the i^{th} input and the j^{th} output of the transpose

† Note that the transpose system S^t is fundamentally different from the "adjoint" system[22] because, although the signal flow is reversed in both, the transpose system does not run "backwards in time."

configuration, and thus $d_{ij} = d^t_{ji}$ for all i and j. The multiplier(s) corresponding to the element b_{ij} of B and relating the j^{th} input and the i^{th} state of the original configuration must, on the other hand, relate the i^{th} state and the j^{th} output of the transpose configuration, and thus $b_{ij} = c^t_{ji}$ for all i and j. Similarly, $c_{ij} = b^t_{ji}$ for all i and j. Finally, the multiplier(s) corresponding to a_{ij} and relating $x_j(n)$ and $x_i(n + 1)$ in the original configuration must, in the transpose configuration, relate $x_i(n)$ and $x_j(n + 1)$, and thus $a_{ij} = a^t_{ji}$ for all i and j. Therefore, the transpose configuration indeed realizes the system S^t.

VII. AN EXAMPLE: THE DIRECT FORM

To demonstrate the application of the results of the preceding sections, we now evaluate and compare the roundoff-noise outputs from two transpose configurations for a digital filter. The scaling required to satisfy the overflow constraints in (38) is derived, and the effect of this scaling on the output roundoff noise is determined.

The transfer function $H^*(z)$, defined in equation (1) and relating the input and output of the digital filter, may be expressed as a rational function in z of the form[3,4]

$$H^*(z) = \frac{\sum_{i=0}^{N} a_i z^{-i}}{1 + \sum_{i=1}^{N} b_i z^{-i}} = \frac{A^*(z)}{B^*(z)}. \tag{49}$$

Assuming that a_N and b_N are not both zero, N is referred to as the "order" of the filter. There are many different, but equivalent, forms in which equation (49) may be written, with a number of equivalent circuit configurations corresponding to each of these forms (at least two transpose configurations). Those forms such as equation (49) which require the minimum number of multiplications and additions in the general case (that is, $2N + 1$ and $2N$, respectively) are referred to as "canonical" forms. In general, however, it is necessary to add additional scaling multipliers to these canonical forms in order to satisfy the overflow constraints in (38).

The form of $H^*(z)$ given in equation (49) is often called the "direct form" of a digital filter. It has been pointed out by Kaiser[6] that use of the direct form is usually to be avoided because of the sensitivity of the roots of higher-order polynomials to small variations (that is, quantization errors) in the polynomial coefficients. The roundoff-noise outputs from the direct form can also be much larger than from other canonical

forms.[15] Nevertheless, the direct form is of theoretical interest, and it provides a convenient illustration of our results. Similar investigations for the two canonical forms most commonly employed in practice—the cascade and parallel forms—are described in Ref. 1.

Two transpose configurations which implement the direct form with scaling are shown in Figs. 3 and 4. These configurations actually realize $H^*(z)$ in the form

$$H^*(z) = \frac{K'_k \sum_{i=0}^{N} {}_k a'_i z^{-i}}{1 + \sum_{i=1}^{N} b_i z^{-i}} \tag{50}$$

where ${}_k a'_i = a_i / K'_k$, and the additional scaling multipliers K'_k, $k = 1, 2$, are required to satisfy (38) in the general case. The configuration in Fig. 3 will be designated as form 1 (that is, $k = 1$), and Fig. 4 as form 2 (that is, $k = 2$).

The branch nodes at which overflow constraints are required (because these signals input to multipliers) are indicated by (*). The dynamic-range limitations are obviously satisfied (by assumption) at the input to the filter, but for completeness, an overflow constraint is included there as indicated. The scaled transfer responses ${}_k F'_i(\omega)$ to these nodes are noted in Figs. 3 and 4, and the corresponding unscaled responses ${}_k F_i(\omega)$ apply, of course, when $K'_k = 1$.

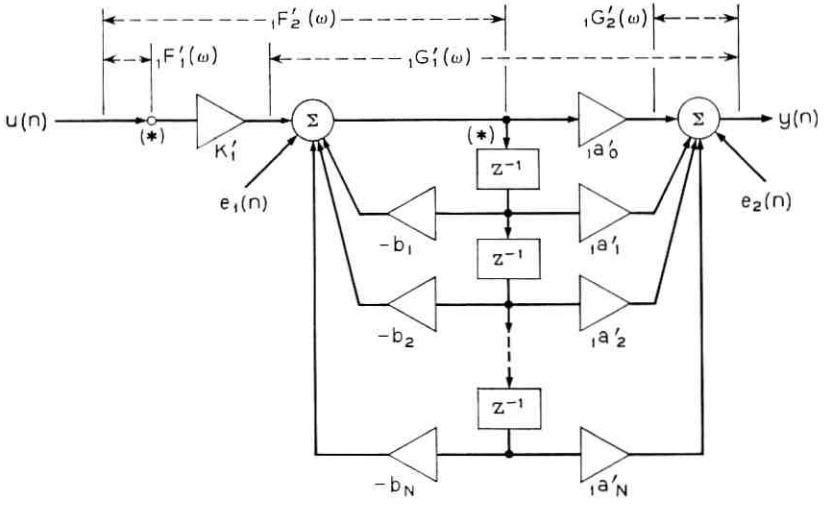


Fig. 3 — Direct form 1 with scaling.

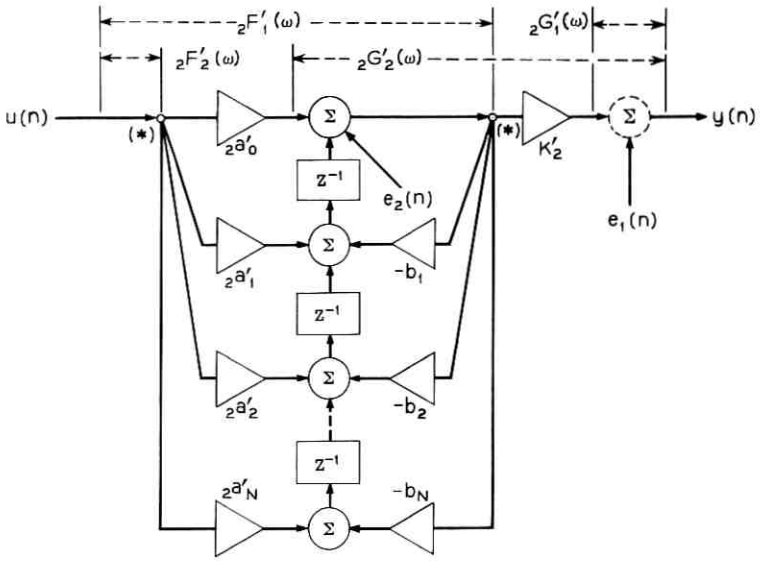


Fig. 4 — Direct form 2 with scaling.

It is intuitively clear that to preserve the greatest possible signal-to-noise ratio, the scaling should reduce the magnitude of ${}_kF_i'(\omega)$ no more than is necessary (or should increase it as much as possible, as the case may be). In other words, ${}_kF_i'(\omega)$ should satisfy

$$|| \, {}_kF_i' \, ||_p = 1. \tag{51}$$

This condition will be satisfied if the scaling factors ${}_ks_i$, defined by

$${}_kF_i'(\omega) = {}_ks_i \cdot {}_kF_i(\omega), \tag{52a}$$

are given by

$${}_ks_i = 1/|| \, {}_kF_i \, ||_p . \tag{52b}$$

It is readily seen from Figs. 3 and 4 that

$${}_1F_1'(\omega) = {}_2F_2'(\omega) = 1, \tag{53}$$

and hence equation (51) is automatically satisfied for these responses. Of more interest, however, are the responses

$${}_1F_2'(\omega) = \frac{K_1'}{B(\omega)} = K_1' \, {}_1F_2(\omega) \tag{54}$$

and

$$ {}_2F_1'(\omega) = \frac{H(\omega)}{K_2'} = \frac{{}_2F_1(\omega)}{K_2'}. \tag{55} $$

From equations (52), (54), and (55), it follows that (51) is satisfied for these configurations if (and only if)

$$ K_1' = 1/||\ 1/B\ ||_p \tag{56} $$

and

$$ K_2' = ||\ H\ ||_p\ . \tag{57} $$

The rounding-error inputs $e_j(n)$ are also shown in Figs. 3 and 4 along with the transfer responses ${}_kG_j'(\omega)$ from these inputs to the output of the filter. Note that in form 2 (Fig. 4) the error input $e_2(n)$ incorporates the roundoff errors from all of the multipliers except K_2' even though these error sources are separated by delays (z^{-1}). This is done for convenience and is possible because of the assumption of uncorrelated errors from sample to sample and source to source. The noise weights k_j' [see equation (10)] for form 1 are thus

$$ {}_1k_1' = {}_1k_2' = N + 1; \tag{58a} $$

while for form 2,

$$ {}_2k_1' = 1 \quad \text{and} \quad {}_2k_2' = 2N + 1. \tag{58b} $$

The indices i and j of the ${}_kF_i(\omega)$ and ${}_kG_j(\omega)$ have been assigned in such a way that forms 1 and 2 are related as in equations (47) and (48). That is, these unscaled responses satisfy the following equations:

$$ {}_1F_i(\omega) = {}_2G_i(\omega), \qquad i = 1, 2, \tag{59a} $$

$$ {}_1G_j(\omega) = {}_2F_j(\omega), \qquad j = 1, 2. \tag{59b} $$

Note that the scaled responses ${}_kF_i'(\omega)$ and ${}_kG_j'(\omega)$ are not related as in equation (59) because, in general, $K_1' \neq K_2'$. In particular,

$$ {}_1G_1'(\omega) = \frac{H(\omega)}{K_1'} = \left(\frac{K_2'}{K_1'}\right){}_2F_1'(\omega); \tag{60} $$

while

$$ {}_2G_2'(\omega) = \frac{K_2'}{B(\omega)} = \left(\frac{K_2'}{K_1'}\right){}_1F_2'(\omega). \tag{61} $$

However, we do have, as in equation (53), that

$$ {}_1G_2'(\omega) = {}_2G_1'(\omega) = 1. \tag{62} $$

From equations (10) and (53) through (62), the power-spectral densities of the roundoff-noise outputs from these two configurations are thus computed to be

$$ {}_1N_y(\omega) = \sigma_0^2(N+1)\left\{1 + \left|\left|\frac{1}{B}\right|\right|_p^2 |\, H(\omega)\,|^2\right\} \tag{63a} $$

and

$$ {}_2N_y(\omega) = \sigma_0^2\left\{1 + (2N+1)\,||\,H\,||_p^2 \left|\frac{1}{B(\omega)}\right|^2\right\}. \tag{63b} $$

The variances, or total average powers of the output roundoff noise from these configurations are then, from equations (8) and (18), simply

$$ ||\,{}_1N_y\,||_1 = \sigma_0^2(N+1)\left\{1 + \left|\left|\frac{1}{B}\right|\right|_p^2 ||\,H\,||_2^2\right\} \tag{64a} $$

and

$$ ||\,{}_2N_y\,||_1 = \sigma_0^2\left\{1 + (2N+1)\,||\,H\,||_p^2 \left|\left|\frac{1}{B}\right|\right|_2^2\right\}. \tag{64b} $$

The peak noise densities $||\,{}_kN_y\,||_\infty$ are, on the other hand, bounded by

$$ ||\,{}_1N_y\,||_\infty \leqq \sigma_0^2(N+1)\left\{1 + \left|\left|\frac{1}{B}\right|\right|_p^2 ||\,H\,||_\infty^2\right\} \tag{65a} $$

and

$$ ||\,{}_2N_y\,||_\infty \leqq \sigma_0\left\{1 + (2N+1)\,||\,H\,||_p^2 \left|\left|\frac{1}{B}\right|\right|_\infty^2\right\}. \tag{65b} $$

We now compare direct forms 1 and 2 on the basis of (64) and (65). Although comparisons based on bounds for $||\,{}_kN_y\,||_\infty$ as in (65) do not, of course, necessarily hold for $||\,{}_kN_y\,||_\infty$ itself, experimental results have indicated that such comparisons are quite effective qualitatively, and often quantitatively as well.[1] Consider first the expressions in equation (64) for $p = 2$ and in (65) for $p = \infty$ (that is, $||\,N_y\,||_r$, $r = 1, \infty$, for $p = r + 1$]. In these two cases, the only difference between the (a) and (b) expressions for forms 1 and 2, respectively, are the k_i', as given in equation (58). In particular, for $||\,1/B\,||_p^2||\,H\,||_p^2 \gg 1$ as is often the case, the $||\,N_y\,||_r$ for form 1 are approximately half, or 3 db less than, those for form 2. This result simply reflects the fact that only half of the noise sources in form 1 input at other than the filter output; whereas in form 2, all but one input within the filter. Hence, if the gains from these inputs to the output are large, form 1 is preferable to form 2 by up to 3 db.

For $p \neq r + 1$, however, the differences in the k_i' are of secondary importance compared with the potential differences due to the mixture of L_2 and L_∞ norms in (64) and (65). In particular, letting

$$\theta_{pq} = \left\| \frac{1}{B} \right\|_p^2 \| H \|_q^2 , \tag{66a}$$

we immediately see that if $\theta_{\infty 2} \gg \theta_{2\infty}$, then form 2 is better for $p = \infty$ while form 1 is better for $p = 2$. If, on the other hand, $\theta_{\infty 2} \ll \theta_{2\infty}$, then the opposite applies.

To gain insight into the above conditions, we rewrite equation (66a) as

$$\theta_{pq} = \left\| \frac{1}{B} \right\|_p^2 \left\| \frac{A}{B} \right\|_q^2 . \tag{66b}$$

It is then clear that the difference between $\theta_{\infty 2}$ and $\theta_{2\infty}$ is due entirely to the effect of $A(\omega)$ on the L_q norms of $A(\omega)/B(\omega)$ for $q = 2, \infty$ versus the corresponding norms of $1/B(\omega)$. In particular, $A(\omega)$ affects the L_∞ norm in $\theta_{2\infty}$. But the L_∞ norm of a function "concentrates" exclusively on the maximum absolute value of that function; whereas the L_2 norm of a function reflects the r.m.s. absolute value of that function over all argument values. Therefore, the effect of $A(\omega)$ in $\theta_{2\infty}$ results from the alteration of the maxima of $| 1/B(\omega) |$ in $| A(\omega)/B(\omega) |$; while in $\theta_{\infty 2}$, the effect concerns the difference between $| 1/B(\omega) |$ and $| A(\omega)/B(\omega) |$ over all ω.

Intuitively, one expects that the former effect is potentially much greater; that is, in many cases $A(\omega)$ should affect the L_∞ norm in $\theta_{2\infty}$ much more than the L_2 norm in $\theta_{\infty 2}$. In particular, if $| A(\omega) |$ significantly attenuates the maxima of $| 1/B(\omega) |$ [as in a band-rejection filter, for example], then $\theta_{2\infty}$ should be much smaller than $\theta_{\infty 2}$. In this case, form 2 should be used for $p = \infty$, and form 1 for $p = 2$. If, however, $| A(\omega) |$ does not provide such attenuation, then $| A(\omega) |$ must be relatively constant within the band(s) where $| 1/B(\omega) |$ is largest [by the nature of $A(\omega)$], and hence

$$\left\| \frac{A}{B} \right\|_q \approx | A(\omega_0) | \cdot \left\| \frac{1}{B} \right\|_q \tag{67}$$

where ω_0 is a frequency at or near a maximum of $| 1/B(\omega) |$. But then,

$$\theta_{pq} \approx | A(\omega_0) | \left\| \frac{1}{B} \right\|_p \left\| \frac{1}{B} \right\|_q \approx \theta_{qp} , \tag{68}$$

and the difference between direct forms 1 and 2 should be less in this case.

VIII. SUMMARY

The interaction between the roundoff-noise output from a digital filter and the associated dynamic-range limitations has been investigated for the case of uncorrelated rounding errors from sample to sample and from one error source to another. The spectrum of the output roundoff noise from fixed-point implementations was readily shown to be of the form

$$N_v(\omega) = \sigma_0^2 \sum_i k_i' \mid G_i'(\omega) \mid^2 \tag{69}$$

where the $G_i'(\omega)$ are scaled transfer responses from certain "summation nodes" in the digital circuit to the filter output. σ_0^2 is the variance of the rounding errors from each multiplier (or other rounding point), and the k_i' are integers indicating the number of error inputs to the respective summation nodes.

Defining $F_i'(\omega)$ to be the scaled transfer response from the input to the i^{th} "branch node" at which a dynamic-range constraint is required, constraints of the form

$$|| F_i' ||_p \leqq 1 \tag{70}$$

for $p \geqq 1$ were then derived, where $|| F_i' ||_p$ is the L_p norm of the response $F_i'(\omega)$. The appropriate value of p is determined by assumed conditions on the spectra of the input signals to the filter. The effect of (70) is to bound the maximum signal amplitude (for deterministic inputs) or the maximum average power (for random inputs) at the i^{th} branch node.

A state-variable description was employed to formulate the general concept of "transpose configurations" for a digital network and to illustrate the usefulness of this concept in digital-filter synthesis. A particularly important result is that for a given unscaled configuration with transpose responses $F_i(\omega)$ and $G_i(\omega)$, as described above, the responses $F_i^t(\omega)$ and $G_i^t(\omega)$ for the corresponding transpose configuration are given by

$$F_i^t(\omega) = G_i(\omega) \quad \text{and} \quad G_i^t(\omega) = F_i(\omega). \tag{71}$$

Hence, although the overall transfer functions for these two configurations are the same, their roundoff-noise outputs can be quite different, in general. The transpose configuration is obtained by simply reversing the direction of all branches in the given network configuration, and the poles and zeros of the network are thus realized in reverse order in the transpose configuration.

To illustrate these results, the roundoff-noise spectra $N_v(\omega)$ for two

transpose configurations for the direct form of a digital filter were calculated and compared. The direct form should usually be avoided in practice,[6] but it is still of theoretical interest and provides a convenient example of our general approach. Using a very natural assignment of the indices i and j for the unscaled $F_i(\omega)$ and $G_j(\omega)$, equation (69) was shown to be of the form

$$N_y(\omega) = \sigma_0^2\left\{k'_{M+1} + \sum_{j=1}^{M} k'_j \,||\, F_j \,||_p^2 \,|\, G_j(\omega) \,|^2\right\} \tag{72}$$

for these (scaled) configurations for the direct form, where M is the number of error inputs at other than the output of the filter. Hence, the variance, or total average power, of the output roundoff noise is simply

$$\sigma_y^2 = \sigma_0^2\left\{k'_{M+1} + \sum_{j=1}^{M} k'_j \,||\, F_j \,||_p^2 \,||\, G_j \,||_2^2\right\}; \tag{73}$$

while the peak spectral density $||\, N_y \,||_\infty$ is bounded by

$$||\, N_y \,||_\infty \leqq \sigma_0^2\left\{k'_{M+1} + \sum_{j=1}^{M} k'_j \,||\, F_j \,||_p^2 \,||\, G_j \,||_\infty^2\right\}. \tag{74}$$

Identical expressions to (72) through (74) can also be derived for the parallel and cascade forms of a digital filter.[1] The relationship between the noise outputs of corresponding transpose configurations is immediately indicated by (71) through (74) [although, in general, $k'_j \neq k'_i$].

REFERENCES

1. Jackson, L. B., *An Analysis of Roundoff Noise in Digital Filters,* Sc.D. Thesis, Stevens Institute of Technology, Hoboken, New Jersey (1969).
2. McDonald, H. S., "Impact of Large-Scale Integrated Circuits on Communication Equipment," *Proc. of the National Electronics Conf., 24* (December 1968), pp. 569–72.
3. Rader, C. M., and Gold, B., *Digital Processing of Signals,* New York: McGraw-Hill, 1969, pp. 1–130.
4. Kaiser, J. F., "Digital Filters," *System Analysis by Digital Computer,* New York: Wiley, 1966, pp. 218–85.
5. Rader, C. M., and Gold, B., "Digital Filter Design Techniques in the Frequency Domain," *Proc. IEEE, 55,* No. 2 (February 1967), pp. 149–71.
6. Kaiser, J. F., "Some Practical Considerations in the Realization of Linear Digital Filters," *Proc. Third Annual Allerton Conf. on Circuit and System Theory,* Monticello, Illinois, October 1965, pp. 621–33.
7. Knowles, J. B., and Olcayto, E. M., "Coefficient Accuracy and Digital Filter Response," *IEEE Trans. on Circuit Theory, CT-15,* No. 1 (March 1968), pp. 31–41.
8. Bennett, W. R., Spectra of Quantized Signals," B.S.T.J., *27,* No. 3 (July 1948), pp. 446–72.
9. Kaneko, T., and Liu, B., "Round-off Error of Floating-Point Digital Filters," *Proc. Sixth Annual Allerton Conf. on Circuit and System Theory,* Monticello, Illinois, October 1968, pp. 219–27.

10. Weinstein, C., and Oppenheim, A. V., "A Comparison of Roundoff Noise in Floating-Point and Fixed-Point Digital-Filter Realizations," *Proc. IEEE, 57,* No. 6 (June 1969), pp. 1181–3.
11. Oppenheim, A. V., "Block-Floating-Point Realization of Digital Filters," MIT Lincoln Laboratory, Technical Note 1969–19 (March 20, 1969).
12. Knowles, J. B., and Edwards, R., "Effects of a Finite-Word-Length Computer in a Sampled-Data Feedback System," *Proc. IEE, 112,* No. 6 (June 1965), pp. 1197–1207.
13. Gold, B., and Rader, C. M., "Effects of Quantization Noise in Digital Filters," *Proc. AFIPS,* 1966 SJCC, pp. 213–19.
14. Jackson, L. B., Kaiser, J. F., and McDonald, H. S., "An Approach to the Implementation of Digital Filters," *IEEE Trans. on Audio and Electroacoustics, AU-16,* No. 3 (September 1968), pp. 413–21.
15. Edwards, R., Bradley, J., and Knowles, J. B., "Comparison of Noise Performance of Programming Methods in the Realization of Digital Filters," *Proc. of the Symposium on Computer Processing in Communications, XIX,* PIB–MRI Symposia Series (1969).
16. Papoulis, A., "Limits on Bandlimited Signals," *Proc. IEEE, 55,* No. 10 (October 1967), pp. 1677–85.
17. Rice, J. R., *The Approximation of Functions,* Reading, Mass.: Addison-Wesley, 1964, pp. 4–10.
18. Bachman, G., and Naria, L., *Functional Analysis,* New York: Academic Press, 1966, pp. 110–11.
19. Davenport, W. B., Jr., and Root, W. L., *Random Signals and Noise,* New York: McGraw-Hill, 1958, pp. 154–7.
20. Mason, S. J., and Zimmerman, H. J., *Electronic Circuits, Signals and Systems,* New York: Wiley, 1960, pp. 122–3.
21. Freeman, H., *Discrete-Time Systems,* New York: Wiley, 1965, pp. 19–27.
22. Laning, J. H., Jr., and Battin, R. H., *Random Processes in Automatic Control,* New York: McGraw-Hill, 1956, pp. 239–43.

An Optimization Method for Cascaded Filters

By SHLOMO HALFIN

(Manuscript received August 1, 1969)

This paper presents a procedure for decomposing an nth *order filter into cascadable second order sections. The procedure is optimal in that it minimizes the maximal response range for the sections within the frequency band of interest. The procedure, based on a modified version of the Bottleneck Assignment Algorithm, describes methods of listing all the optimal decompositions as well as of finding a special "nested" optimal decomposition.*

I. INTRODUCTION

Let $\phi(s)$ be a transfer function

$$\phi(s) = \frac{f(s)}{g(s)}$$

where f and g are polynomials with real coefficients, and the degree of $f \leqq$ degree of g.

We consider all the decompositions of the form $\phi(s)=\phi_1(s)\phi_2(s) \cdots \phi_t(s)$ where

$$\phi_i(s) = \frac{f_i(s)}{g_i(s)} \tag{1}$$

$f_i(s)$ and $g_i(s)$ are real polynomials and the degree of f_i does not exceed the degree of g_i. The g_i are quadratic polynomials, except when the degree of g is odd; then one g_i is linear.

Let L be a passband region for ϕ, where L is a finite union of passband intervals. Then for every ϕ_i, a number $d(\phi_i)$ is defined by

$$d(\phi_i) = 20 \log_{10} \left[\frac{\underset{\omega \in [0,\infty)}{\text{Max}} \mid \phi_i(j\omega) \mid}{\underset{\omega \in L}{\text{Min}} \mid \phi_i(j\omega) \mid} \right]. \tag{2}$$

Also let

$$d = \underset{i=1,\ldots,t}{\text{Max}}\ d(\phi_i). \tag{3}$$

Then d is a function of the decomposition. We present a procedure that determines the decomposition(s) with a minimal d. E. Lueder proposed this optimality criterion.[1]

II. METHOD

First, we artificially equate the number of zeros [zeros of $f(s)$], and poles [zeros of $g(s)$], by adding a suitable number of "zeros at infinity" corresponding to constant unit polynomials. Next we make this mutual number even by adding a zero and a pole at infinity, if necessary. In this way we get, say, $2t$ zeros and $2t$ poles.

Pairing two zeros creates an f_i ; a real zero can be paired with any other real zero, while a complex zero must be paired with its conjugate in order to get a real f_i . The same is true for creation of g_i by pairing of poles.

In the following we assume that all poles, except perhaps one, are complex and therefore fixed paired. We call the real zeros which are not fixed paired *free zeros*.

Next we make all possible pairings of the free zeros. Each such pairing, together with the fixed pairing, decomposes $f(s)$ and $g(s)$:

$$f(s) = f_1(s)f_2(s) \cdots f_t(s);$$
$$g(s) = g_1(s)g_2(s) \cdots g_t(s).$$

Then we compute the matrix $D = (d_{ik})$, where the elements

$$d_{ik} = d\left(\frac{f_i}{g_k}\right) \tag{4}$$

are computed from definition (2). The element d_{ik} represents the "cost" of *matching* zero-pair i with pole-pair k.

An *assignment* is a feasible set of matchings. Using the Bottleneck Assignment Algorithm, we determine an assignment k_1 , $\cdots$, k_t for which

$$\underset{i=1,\ldots,t}{\text{Max}}\ d_{ik_i}$$

will be minimal. We call this minimum the *optimal d value* for this pairing of free zeros. Going through all the possible pairings of free zeros, we find an *optimal pairing* which yields the smallest optimal d value.

Since an optimal assignment (for a given optimal pairing) is usually

not unique, procedures for obtaining all the optimal solutions (assignments) or a *nested* solution are given. A *nested* solution is obtained by taking an optimal solution, fixing the matching with the largest d value, and then proceeding to look for an optimal assignment for the remaining $t-1$ f_i's and $t-1$ g_k's, and so on.

III. THE BOTTLENECK ASSIGNMENT ALGORITHM

This section discusses the Bottleneck Assignment Algorithm and its adaptation to the present problem.

Let $U = (u_{ij})$ be a real $t \times t$ matrix. A matching is an ordered pair of integers (i_k , j_k) $1 \leqq i_k \leqq t$, $1 \leqq j_k \leqq t$. We associate with the matching (i_k , j_k) the corresponding cell in U. The element in this cell u_{ij} is called the cost of the matching.

A set $A = \{(i_k , j_k); k = 1, \cdots , t\}$ of t matchings (cells) is called an *assignment* if in every row and in every column of U there is a cell that belongs to A. The bottleneck assignment problem looks for an assignment which minimizes the maximum of its matchings' costs.

The Gross algorithm[2], is based on the following iterative step:

(*) An assignment A and a real number α, which does not exceed all the costs of A, are given; then either a new assignment A' is con-

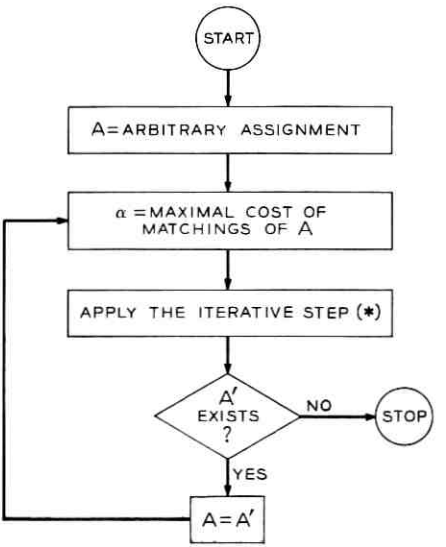


Fig. 1 — Flow chart for solving the bottleneck assignment problem.

structed, so that α exceeds more costs in A' than it does in A, or it is established that no such A' exists.

The flow chart in Fig. 1 solves the bottleneck assignment problem. This algorithm is fast and requires small memory space, since only the present assignment must be stored.

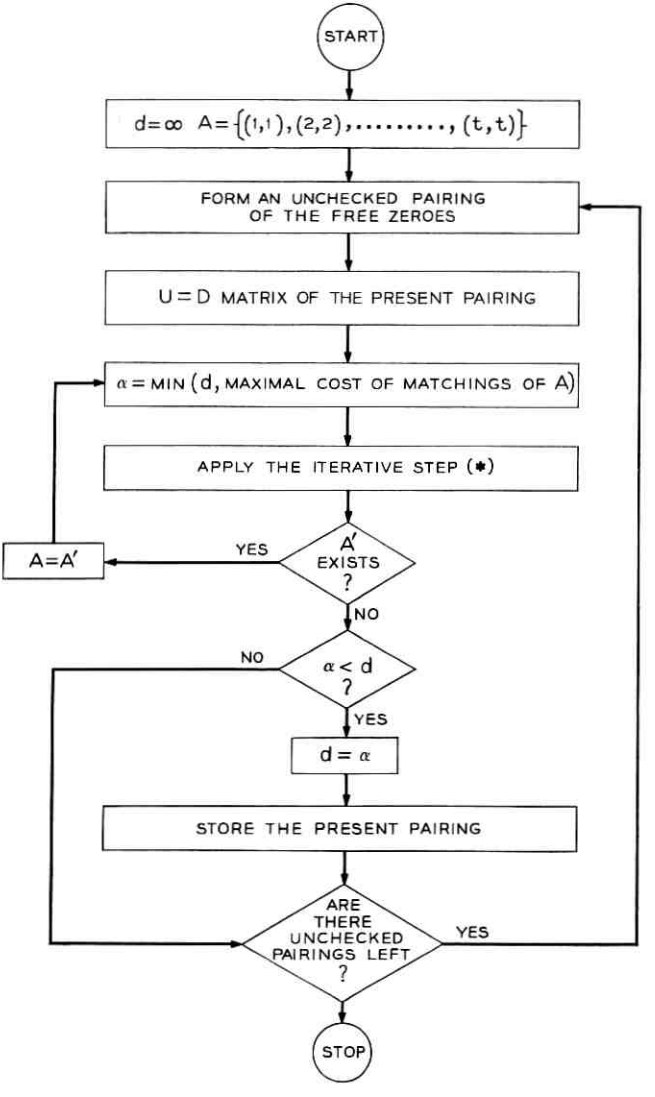


Fig. 2 — Flow chart for finding the optimal pairing and its optimal d-value.

The basic algorithm was modified to find the optimal pairing of the free zeros and its optimal d-value (Fig. 2). Note that the value of α in every iterative step (*) does not exceed the minimum of the optimal d values of the checked pairings. Thus any pairing that does not reduce the d value already obtained is immediately disregarded.

Also, for each pairing we use the optimal assignment of the preceding pairing, as an initial assignment. Thus the costs of the initial assignment matchings that correspond to the fixed paired zeros do not exceed the current d value. These procedures considerably reduce the amount of computation required for finding the optimal pairing and its optimal d value.

IV. CREATION OF THE NESTED SOLUTION

Let U denote the cost matrix which corresponds to the optimal pairing. The nested solution is created by successively applying the bottleneck assignment algorithm t times and modifying U each time in such a way that the matching with the largest cost becomes fixed and irrelevant in the further computations.

Let (i_k^*, j_k^*) be the matching with the largest cost at a certain stage. Then (i_k^*, j_k^*) becomes a part of the nested solution. U is then modified as follows:

$$U_{i_k^*,s} = \infty \quad \text{for all} \quad s \neq j_k^*;$$

$$U_{s,j_k^*} = \infty \quad \text{for all} \quad s \neq i_k^*;$$

$$U_{i_k^*,j_k^*} = 0.$$

It is easy to verify that this modification has the properties described.

V. A COMPUTATIONAL METHOD TO GENERATE ALL THE OPTIMAL ASSIGNMENTS

Let U denote again the cost matrix which corresponds to the optimal pairing, and let d^* be the optimal d value. We call a cell (i, j) *admissible* if $u_{ij} \leqq d^*$. The problem of listing all the optimal assignments then becomes the problem of listing all possible assignments that use only admissible cells. Using the flow chart of Fig. 3 can accomplish this. The number of operations can be seen to be dependent on the order of the columns of U. The dependence is quite complicated. However, a good rule of thumb for reducing the number of operations is to rearrange the columns in ascending order according to the number of their admissible cells.

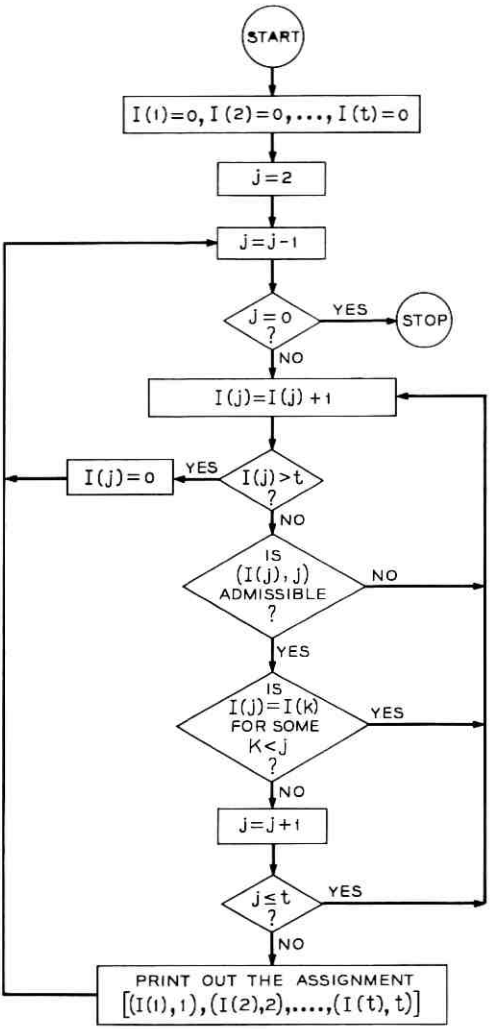


Fig. 3 — Flow chart for generating all optimal solutions.

REFERENCES

1. Lueder, E., "Cascading of RC-Active Two-Ports in Order to Minimize Inband Losses and to Avoid Distortion," Proceedings of the International Conference on Communications, Boulder, Colorado, June 9–11, 1969.
2. Gross, O., "The Bottleneck Assignment Problem," The Rand Corporation, Paper P-160, March 6, 1959.

Measured Quantizing Noise Spectrum for Single-Integration Delta-Modulation Coders

By R. R. LAANE

(Manuscript received October 14, 1969)

We give experimental verification, for idle-channel and sinusoidal inputs, of a recently developed quantizing noise theory for asymmetrical, single-integration delta-modulators.

A recent paper by Iwersen described a procedure for calculating quantizing noise for single-integration delta-modualtion coders employing unequal positive and negative integrator step sizes.[1] The purpose of this note is to provide experimental verification of the theory.

Measured quantizing noise for both idle-channel and sinusoidal inputs is given and the idle-channel noise spectrum is calculated.

Defining the positive, σ_+, and negative σ_-, integrator step sizes as

$$\sigma_+ \equiv \sigma + \epsilon$$
$$\sigma_- \equiv -\sigma + \epsilon \qquad (1)$$

where σ is the average step size, an error wave is generated by the

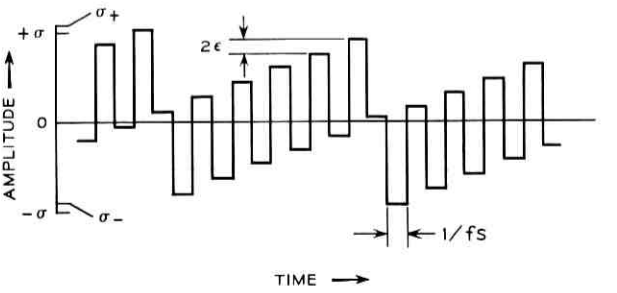


Fig. 1 — Asymmetrical integrator output for an idle-channel input, shown for $|\sigma_+| > |\sigma_-|$.

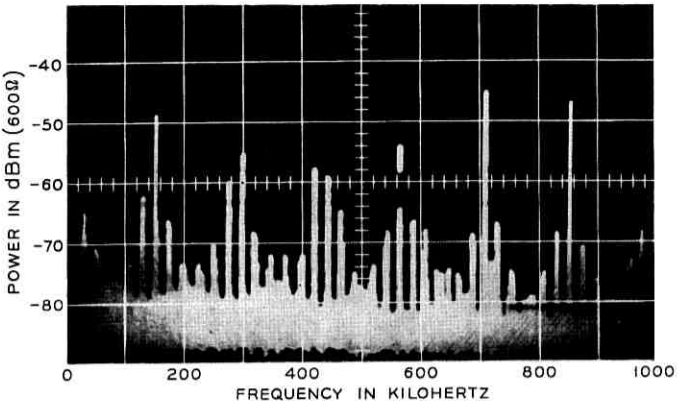


Fig. 2 — Observed idle-channel noise spectrum, $f_s = 1.56$ MHz, $\vartheta = 0.0937$.

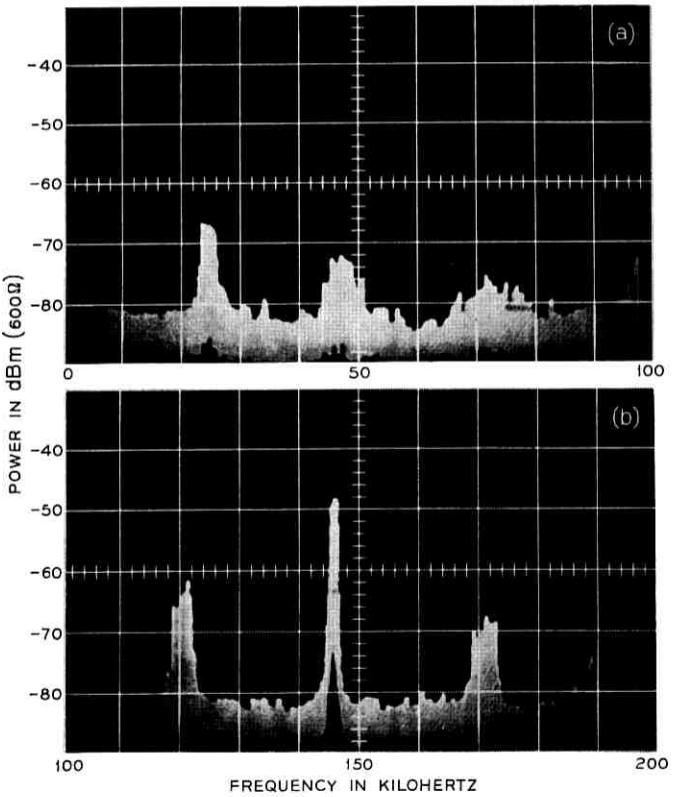


Fig. 3 — Expanded idle-channel noise spectrum: (a) 0 — 100 kHz, (b) 100 — 200 kHz.

integrator for idle-channel inputs as shown in Fig. 1. The quantizing noise spectrum resulting from the error wave is a line spectrum, and the line frequencies, f_l , for a one-sided spectrum from zero to one-half the sampling frequency are given as a function of the integer index l by[1]

$$f_l = | Q[l(1 - \vartheta)/2]f_s | \tag{2}$$

where

$$Q(\alpha) = \alpha - N(\alpha),$$

$$N(\alpha) = \text{integer nearest } \alpha$$

and ϑ is the integrator step imbalance ϵ/σ and f_s the sampling frequency.

The power at the frequency of index l is calculated from

$$P_l = 2\sigma^2/\pi^2 l^2. \tag{3}$$

The resulting noise-spectral lines will subsequently be referred to as l-lines (1-line, 2-line, 14-line, and so on).

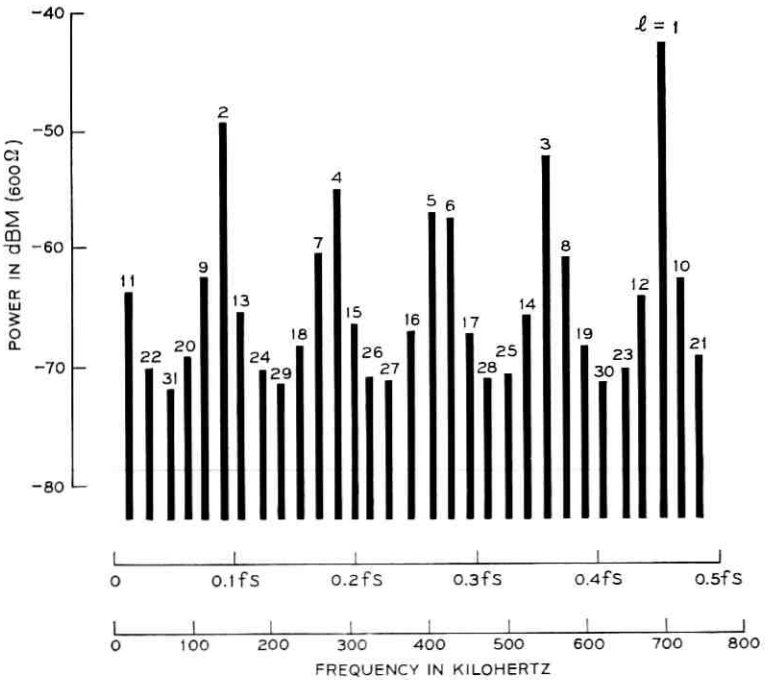


Fig. 4 — Calculated idle-channel noise spectrum, $f_s = 1.56$ MHz, $\vartheta = 0.0937$.

Measurements of the quantizing noise spectrum were made using a delta-modulation coder designed for telephone switching applications.[2] A 1.56 MHz sampling frequency and an average integrator step size of 13 millivolts were used for the measurements. Figure 2 shows the observed idle-channel spectrum of the coder for a frequency range from 0 to 1 MHz. The region near the 2-line is expanded in Figs. 3a and 3b where the noise spectrum is shown for frequencies from 0 to 100 kHz and 100 to 200 kHz respectively.

The calculated spectrum from 0 to $f_s/2$ (0 to 0.78 MHz) is shown in Fig. 4 for $\vartheta = 0.0937$. Excellent correlation can be observed with the measured spectrum in Fig. 2. For a more detailed comparison, Table I gives the calculated and measured frequencies and powers of the l-lines for the band from 0 — 200 kHz. With respect to frequency, the agreement is within experimental error. However, measured peak powers of higher order l-lines fall below the calculated values. This discrepancy is believed to be due to modulation broadening of the lines by a low-level noise input of unknown origin.

Figures 5a and 5b show the effect of sinusoidal inputs on the coder noise spectrum. As suggested by Iwersen, inputs to the coder phase-modulate the idle-channel lines and force the frequency band occupied by each l-line group, Δf, to become proportional to the slope of the input signal, $2\pi A f_0$, and to the index of the l-line,[1]

$$\Delta f \approx 2\pi l A f_0 \tag{4}$$

where A is the amplitude and f_0 the frequency of the input signal. This is illustrated in the figures where broadening of the 1-line, 2-line, 3-line and 4-line as a function of signal amplitude is clearly visible.

TABLE I—COMPARISON OF MEASURED AND CALCULATED NOISE SPECTRUM FOR 0 — 200 kHz

	Measured		Calculated	
l-line	f_l	P_l	f_l	P_l
2	146 kHz	−48 dBm	146 kHz	−48 dBm
9	121	−61	121	−61
11	24	−66	24	−63
13	171	−67	170	−64
20	98	−69	98	−66
22	48	−72	48	−69
24	194	−74	194	−70
31	73	−75	74	−72

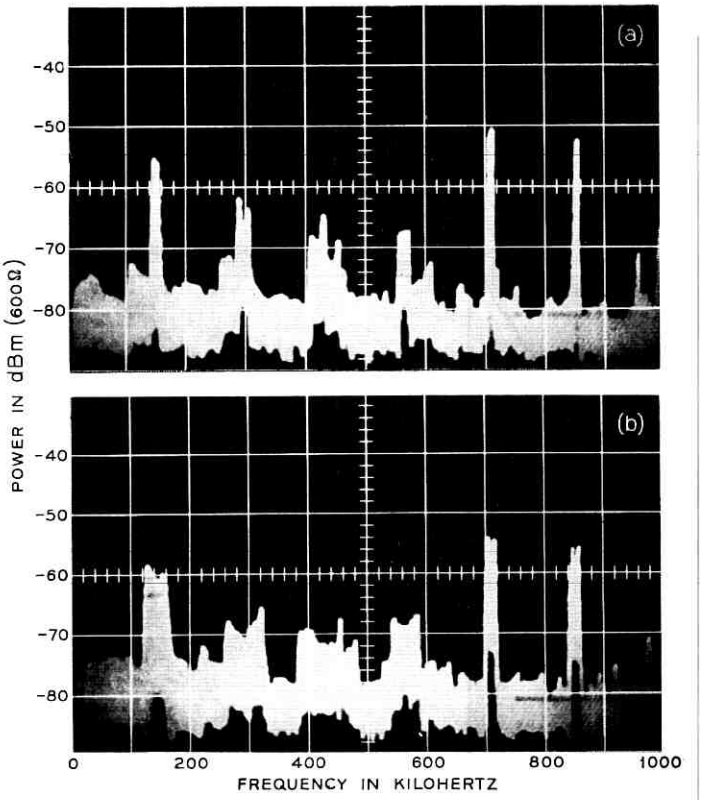


Fig. 5 — Effect of 1 kHz sinusoidal inputs on noise spectrum; (a) —40 dBm input, (b) —30 dBm input.

Additional discussion of the noise characteristics as well as a description of the design of the delta-modulation coder will be presented in a future paper.[2]

REFERENCES

1. Iwersen, J. E., "Calculated Quantizing Noise of Single-Integration Delta-Modulation Coders," B.S.T.J., *48*, No. 7 (September 1969), pp. 2359–2389.
2. Laane, R. R., and Murphy, B. T., "Delta-Modulation Coder for Telephone Transmission and Switching Applications," unpublished work.

Use of the Discrete Fourier Transform in the Measurement of Frequencies and Levels of Tones

By D. C. RIFE and G. A. VINCENT

(Manuscript received May 7, 1969)

This paper considers the application of a digital computer and discrete Fourier transform (DFT) techniques to the measurement of signals known to comprise only single-frequency tones. We discuss the use of weighting functions to improve the effective selectivity of a measurement system that estimates the frequencies and levels of tones from the coefficients of their DFT. We present three classes of weighting functions which may be used to improve the inherent accuracy of such a system. The form of the weighting functions was chosen to minimize the amount of computer memory required, without using too much computer time. Several formulas are derived for estimating the frequency and level of a tone from its DFT coefficients. We chose the formulas to minimize computation time.

Simulation results indicate that, through the use of a proper weighting function, a DFT measurement system that uses 512 samples taken at a sampling frequency of 7040 Hz can be designed so that the maximum error in the frequency estimates of two tones near 1000 Hz and separated by approximately 50 Hz is about 0.03 Hz. The corresponding maximum error in the level estimate is on the order of 0.03 dB.

I. INTRODUCTION

There have been numerous articles, in recent years, dealing with the use of the discrete Fourier transform (DFT) in the area of spectrum analysis. Much of this interest was motivated by the availability of a computational algorithm that facilitates the rapid computation of DFT coefficients by a digital computer. The algorithm is, of course, the fast Fourier transform (FFT).

We are concerned with the problem of applying DFT techniques to the measurement of the levels and frequencies of single-frequency tones,

Fig. 1 — A DFT measurement system.

particularly tones from a data set during a test. Figure 1 shows the system we have in mind. A band-limited received signal, known to comprise one or more single-frequency tones, is periodically sampled by an A-D converter. A total of N samples are taken and the DFT coefficients are computed from the samples. The computer determines which of the DFT coefficients are "large", indicating the approximate frequencies of the received tones, and then proceeds to compute accurate estimates of the frequencies and levels. Methods for achieving the first part of the procedure are well known. This paper is devoted to a consideration of how best to go about the last step in the process, the accurate estimation of the frequencies and levels of the received tones.

In data set testing, the tone measurement system would be used occasionally during a test and would have to consume a minimum amount of real time. Thus we have directed our attention toward estimation methods that use simple formulas and require a minimum amount of computer memory.

Our attention is confined to the problem of leakage, its reduction by smoothing (windowing) functions, and the development of formulas which extract tone levels and frequencies from the list of DFT coefficients. We don't discuss the important, but secondary, problems of round-off errors and other noise sources.

II. REVIEW OF DISCRETE FOURIER TRANSFORM

The definition and properties of the discrete Fourier transform are discussed in Refs. 1 and 2. The following review is to refresh the reader's memory and establish the notation that we will use later.

2.1 *Definition of Discrete Fourier Transform*

Consider an ordered set of numbers $\{X_n\}$ where $n = 0, 1, 2, \cdots, N - 1$. Following Cochran, and others,[1] we define the discrete Fourier transform (DFT) of the set $\{X_n\}$ to be another set of numbers, $\{A_K\}$, with

$$A_K = \sum_{n=0}^{N-1} X_n e^{-i2\pi nK/N}, \quad \text{all integer} \quad K. \tag{1}$$

The inverse transformation is

$$X_n = \frac{1}{N} \sum_{K=0}^{N-1} A_K e^{j2\pi nK/N}, \qquad n = 0, 1, 2, \cdots, N-1. \tag{2}$$

2.2 *Useful Properties*

Several properties of the DFT are utilized in later parts of this paper. The important properties are recorded in this section for future reference. Reference 2 provides a more complete list. Derivations are included only for results that may not be well known.

From equation (1) it is obvious that if the X_n are real, then

$$A_{-K} = A_K^* \quad (* \text{ denotes conjugate}), \tag{3}$$

$$A_{K+N} = A_K, \tag{4}$$

and

$$A_{N-K} = A_{-K} = A_K^*. \tag{5}$$

2.2.1 *Convolution*

Let

$$B_K = \sum_{n=0}^{N-1} X_n e^{-j2\pi nK/N} \tag{6}$$

and

$$C_K = \sum_{n=0}^{N-1} Y_n e^{-j2\pi nK/N}, \tag{7}$$

then

$$A_m = \sum_{n=0}^{N-1} X_n Y_n e^{-j2\pi nm/N} = \frac{1}{N} \sum_{K=0}^{N-1} B_K C_{m-K}. \tag{8}$$

In other words, if $\{B_K\}$ and $\{C_K\}$ are the DFT of $\{X_n\}$ and $\{Y_n\}$, respectively, then the DFT of $\{X_n Y_n\}$ is given by equation (8).

2.2.2 *Power*

It can easily be shown, for X_n and A_K defined by equations (1) and (2), that

$$\frac{1}{N} \sum_{K=0}^{N-1} A_K A_K^* = \sum_{n=0}^{N-1} X_n^2. \tag{9}$$

If the X_n are samples of some function, $f(t)$; that is, if $X_n^2 = f^2(nT/N)$,

then

$$\lim_{N\to\infty} \frac{T}{N} \sum_{n=0}^{N-1} X_n^2 = \int_0^T f^2(t)\, dt$$

if the integral exists. Thus, for large N,

$$\int_0^T f^2(t)\, dt \approx \frac{T}{N} \sum_{n=0}^{N-1} X_n^2 . \tag{10}$$

Hence, from equation (9),

$$\frac{1}{T} \int_0^T f^2(t)\, dt \approx \frac{1}{N^2} \sum_{K=0}^{N-1} | A_K |^2. \tag{11}$$

2.3 *Relationship to Fourier Transform*

The DFT of samples of a signal has a simple relationship to the regular Fourier transform of the signal. It is instructive to examine this relationship.†

Let $g(t)$ be an arbitrary function, zero for $t < 0$ and $t > T$ and continuous over $0 < t < T$. The function is allowed to be discontinuous at $t = 0$ and at $t = T$. Assume that $g(0+)$ and $g(T-)$ exist.

A well-known application of the Poisson sum formula gives[4]

$$\tfrac{1}{2}g(0+) + \tfrac{1}{2}g(T-) + \sum_{n=1}^{N-1} g\left(\frac{nT}{N}\right) = \frac{N}{T} \sum_{n=-\infty}^{\infty} G\left(\frac{2\pi nN}{T}\right) \tag{12}$$

where

$$G(u) = \int_0^T g(t)e^{-iut}\, dt. \tag{13}$$

Adopting a notation similar to that of Papoulis,[4] we define the "#" operation by

$$G^{\#}(\omega) = \frac{N}{T} \sum_{K=-\infty}^{\infty} G(\omega - K\omega_s), \tag{14}$$

where

$$\omega_s = 2\pi N/T. \tag{15}$$

Then equation (12) can be rearranged to give

$$\sum_{n=0}^{N-1} g\left(\frac{nT}{N}\right) = G^{\#}(0) + \tfrac{1}{2}[g(0+) - g(T-)], \tag{16}$$

where $g(0)$ is taken to equal $g(0+)$.

† The recent article by Bergland touches upon this subject and also contains an extensive list of references.[3]

Let $h(t)$ be any function of the sort used above for $g(t)$ with the additional property: $h(0) = h(0+)$.

Let $s(t)$ be the signal to be analyzed and define

$$f(t) = s(t)h(t). \tag{17}$$

Let $g(t) = f(t)e^{-i\omega t}$ and define

$$A(\omega) = \sum_{n=0}^{N-1} f\left(\frac{nT}{N}\right)e^{-in\omega T/N}. \tag{18}$$

Then from equation (16) and the definition in equation (14) we have

$$A(\omega) = F^*(\omega) + \tfrac{1}{2}[f(0+) - f(T-)e^{-i\omega T}], \tag{19}$$

where

$$F(\omega) = \int_0^T f(t)e^{-i\omega t}\,dt. \tag{20}$$

If $X_n = f(nT/N)$ then the A_K defined by equation (1) are given by

$$A_K = A\left(\frac{2\pi K}{T}\right). \tag{21}$$

Thus the DFT of the set $\{f(nT/N)\}$ are points along the curve described by equation (19). These points are $1/T$ Hz apart.

Observe that at $\omega = 2\pi K/T$ the term in brackets in equation (19) becomes $\frac{1}{2}[f(0+) - f(T-)]$ which is independent of K and vanishes if $f(0+) = f(T-)$.

2.4 *Weighting Functions*

If the DFT is to be taken of the set $\{s(nT/N)\}$ for $n = 0$ through $N - 1$, then $h(t)$ must be a function whose value is unity at $t = nT/N$; $n = 0, 1, \cdots, N - 1$. The function with this property that is usually taken to be $h(t)$ is the function $h_T(t)$;

$$h_T(t) = \begin{cases} 1, & 0 \leqq t < T; \\ 0, & \text{otherwise}. \end{cases} \tag{22}$$

Other weighting functions, $h(t)$, are often formed by multiplying $h_T(t)$ by a nontime-limited function. Weighting functions play a very important role in systems that use the DFT. The following paragraphs attempt to develop and present some of the pertinent theory.

From equation (19) we see the role that $F^*(\omega)$ plays in $A(\omega)$. Since $f(t) = s(t)h(t)$,

$$F(\omega) = S(\omega) * H(\omega), \tag{23}$$

where the * denotes convolution and $S(\omega)$ and $H(\omega)$ are the Fourier transforms of $s(t)$ and $h(t)$. It can be shown that, subject to the usual convergence constraints,

$$F^{\#}(\omega) = [S(\omega)*H(\omega)]^{\#} = S(\omega)*H^{\#}(\omega). \tag{24}$$

Thus $H(\omega)$, or equivalently $H^{\#}(\omega)$, plays a central role in the DFT of (weighted) samples of $s(t)$. From the development that led to equations (19) and (21), we see that, if $h(0+) = h(T-)$, the DFT of samples of $h(t)$ is a set of points taken along the periodic curve described by $H^{\#}(\omega)$. It follows, therefore, that the values of $h(nT/N)$ can be obtained from

$$h\left(\frac{nT}{N}\right) = \frac{1}{N}\sum_{K=0}^{N-1} H^{\#}\left(\frac{2\pi K}{T}\right)e^{j2\pi Kn/N}. \tag{25}$$

Also,

$$H^{\#}(\omega) = \sum_{n=0}^{N-1} h\left(\frac{nT}{N}\right)e^{-jn\omega T/N} - \tfrac{1}{2}h(0+)[1 - e^{-j\omega T}]. \tag{26}$$

Weighting in the time domain is actually done at the points $t = nT/N$; $n = 0, 1, 2, \cdots, N-1$. For every set of weights to be applied at these points there exists a continuous function with the same values at the indicated time points. Thus there is no loss of generality due to discussing weighting in terms of weighting functions, $h(t)$, that are continuous over $(0, T)$ and zero outside that interval. We have to remember, however, that if the set $\{h(nT/N)\}$ is specified, $h(t)$ is not unique. Thus, if $H^{\#}(\omega)$ is given, $h(nT/N)$ is given by equation (25), but $h(t)$ and $H(\omega)$ are not uniquely defined.

There is apparently some confusion in the literature about whether $H(\omega)$ or $H^{\#}(\omega)$ is called a weighting function (or windowing function). Blackman and Tukey,[5] for example, discuss $h(t)$ and $H(\omega)$, but when Helms[6] writes about weighting with a Dolph-Chebychev function, he is evidently referring to $H^{\#}(\omega)$. More will be said about this later. Bingham, and others, in writing about data windows (See Reference 7, Part VII) mean $h(t)$.

Observe that $H^{\#}(\omega)$ is always periodic with period ω_s, while $H(\omega)$ is not periodic. (If it were, $H^{\#}(\omega)$ would not converge properly.) Generally the $H^{\#}(\omega)$ that one uses will have a prominent main lobe about $\omega = K\omega_s$ (K is any integer, including zero) and many side lobes. For our purposes it is important to obtain a narrow main lobe and low-amplitude side lobes.

The class of $H^{\#}(\omega)$ with the minimum main-lobe width for a given

side-lobe amplitude is known as the (discrete) Dolph-Chebyshev weighting functions.[8] A convenient form, similar to the one given by Helms,[6] but changed to describe the result of weighting by sample values that peak at $T/2$ and are adjusted to cover approximately unit area as later weighting functions will do, is the following:

$$H^*(\omega) = \frac{N}{R} e^{-j\omega T/2} \cos\left[N \cos^{-1}\left(Z_0 \cos\frac{\omega T}{2N}\right)\right] \tag{27}$$

where the side-lobe amplitude, $1/R$, is related to Z_0 by

$$R = \cosh\,(N \cosh^{-1} Z_0) \tag{28}$$

and N is the same as used in equation (1).

The class of $H(\omega)$ with the minimum main-lobe width for a given side-lobe amplitude is known as the continuous Dolph-Chebyshev functions,[9] which are unrealizable. The Taylor approximations to the continuous Dolph-Chebyshev functions[9-11] are realizable, however, and provide almost the same main lobe width for a given maximum side-lobe amplitude.

The problem of choosing "good" shapes for $H^*(\omega)$ can be approached by treating $H^*(\omega)$ or by treating $H(\omega)$. Most of the well-known weighting functions are discussed in terms of $H(\omega)$ or $h(t)$.

2.5 *A Generalization*

If $h(t)$ is a function that is zero for $t < T$, and $t > T$, then it can be shown (sampling theorem) that $H(\omega)$ is given by

$$H(\omega) = Te^{-j\omega T/2} \sin\,(\omega T/2) \sum_{n=-\infty}^{\infty} \frac{C_n}{\frac{\omega T}{2} - n\pi} \tag{29}$$

and

$$TC_n = H\left(\frac{2\pi n}{T}\right). \tag{30}$$

Thus the specification of a weighting function is equivalent to the specification of the constants, C_n.

III. SELECTED WEIGHTING FUNCTIONS

3.1 *Leakage and Aliasing*

Leakage will be used here to refer to the problem of the values of $A(\omega)$ due to $\cos\,(\omega_0 t + \theta_0)$ interferring with the values of $A(\omega)$ at some

other frequency, say ω_1, where the response due to $\cos(\omega_1 t + \theta_1)$ is to be examined. Leakage, in our system, is minimized by the use of weighting functions.

Aliasing refers to the fact that in a sampled-data system tones with frequencies above $\omega_s/2$ cannot be distinguished from tones with frequencies less than $\omega_s/2$. In our system aliasing is avoided by the use of the low-pass filter (Fig. 1).

3.2 *Convolution of Weighting Functions*

The object of weighting is to produce the DFT of a weighted set of samples of the signal undergoing measurement, $s(t)$. Thus we seek to compute

$$A_K = \sum_{n=0}^{N-1} s(nt_s)h(nt_s)e^{-i2\pi nK/N}, \quad \text{for all} \quad K; \tag{31}$$

where $t_s = T/N$. A convenient way of doing the weighting is to first compute

$$B_K = \sum_{n=0}^{N-1} s(nt_s)e^{-i2\pi nK/N}, \tag{32}$$

for $0 \leqq K \leqq N - 1$. Then if the set $\{H_m\}$,

$$H_m = \sum_{n=0}^{N-1} h(nt_s)e^{-i2\pi nm/N} = H^{\#}\left(\frac{2\pi m}{T}\right), \tag{33}$$

is stored in the computer, the A_K can be computed from equation (8).

3.3 *A Special Class of Weighting Functions*

The amount of computer memory required to store the set $\{H_m\}$ will be small if $h(t)$ is a function such that $H_m = 0$ for $M < |m| \leqq N/2$ and M is a relatively small number. The $H(\omega)$ corresponding to this class can be expressed by a particular form of equation (29):

$$H(\omega) = Te^{-iX} \sin X \sum_{n=-M}^{M} \frac{C_n}{X - n\pi}, \tag{34}$$

where

$$X = \omega T/2 \tag{35}$$

and $M \ll N/2$. We have restricted our attention to the results that can be obtained with this class of weighting functions.

Most of the well-known weighting functions, such as Hanning,[5] Hamming,[5] and Taylor[10,11] are in the class defined by equation (34).

The discrete Dolph–Chebychev and the Kaiser–Bessel[12] weighting functions, however, are not.

The right side of equation (34) can be written over a common denominator to obtain the form

$$H(\omega) = Te^{-iX}\frac{\sin X}{X}\frac{P(X)}{\prod_{n=1}^{M}(X^2 - n^2\pi^2)}, \tag{36}$$

where $P(X)$ is, in general, a complex polynomial in X. We will restrict our attention to $H(\omega)$ with real C_n and $C_0 = 1$. In which case $C_{-n} = C_n$ and, if $D_n = 2C_n$, we have

$$h(t) = h_r(t)\left[1 + \sum_{n=1}^{M} D_n \cos\left(\frac{2\pi nt}{T}\right)\right]. \tag{37}$$

Equation (34) becomes

$$H(\omega) = Te^{-iX}\sin X\left[\frac{1}{X} + \sum_{n=1}^{M}\frac{D_nX}{X^2 - n^2\pi^2}\right]. \tag{38}$$

In the next few sections we will discuss three classes of weighting functions with the form of equations (37) and (38). They were chosen to provide two extreme cases of weighting and an intermediate example. Many other weighting functions in the class defined by equations (37) and (38) exist; the ones examined below provide sufficient data for our purposes.

3.4 *Class I Weighting Functions*

We first consider the class of weighting functions that provides the best possible reduction in $|H(\omega)|$ for large ω. Let this class be known as Class I.

The only part of equation (36) that can be adjusted is the polynomial, $P(X)$. Thus we must choose the coefficients, D_n, to minimize $|P(X)|$ for large X. This is done by forcing $P(X)$ to be a constant. The constant term in $P(X)$, from equation (34), is

$$P(0) = (-1)^M\pi^{2M}(M!)^2. \tag{39}$$

Hence, the desired class of weighting functions has the form, from equation (36),

$$H_M(\omega) = Te^{-iX}\frac{\sin X}{X}\frac{(-1)^M\pi^{2M}(M!)^2}{\prod_{n=1}^{M}X^2 - n^2\pi^2}. \tag{40}$$

We denote the coefficients, D_n, of this class of weighting functions as $D_I(M, n)$, making the dependence upon M explicit. From equations (36), (38), and (40) the $D_I(M, n)$ are given by

$$D_I(M, n) = \lim_{X \to n\pi} \frac{(-1)^M \pi^{2M} (M!)^2 (X^2 - n^2\pi^2)}{X^2 \prod_{K=1}^{M} X^2 - K^2\pi^2}. \tag{41}$$

Evaluation of the limit and some simplification gives

$$D_I(M, n) = \frac{2(-1)^n (M!)^2}{(M - n)!\,(M + n)!} = 2(-1)^n \prod_{K=1}^{n} \frac{M + 1 - K}{M + K}. \tag{42}$$

We denote the weighting functions that use equation (41) as $h_M(t)$. Then from equation (37)

$$h_M(t) = h_T(t)\left[1 + \sum_{n=1}^{M} D_I(M, n) \cos \frac{2\pi n t}{T}\right]. \tag{43}$$

This can easily be shown to be the same as

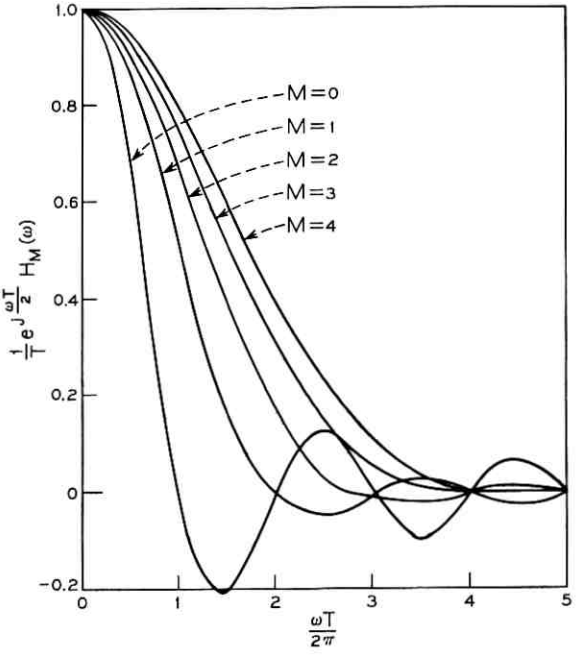


Fig. 2 — Spectra of the Class I weighting functions.

$$h_M(t) = h_T(t) \frac{4^M (M!)^2}{(2M)!} \sin^{2M}\left(\frac{\pi t}{T}\right). \tag{44}$$

Thus the Class I weighting functions are described by equations (40) and (44). The so-called *hanning* weighting[5] is equivalent to $h_1(t)$. Larsen and Singleton[13] used $h_1(t)$, $h_2(t)$, and others.

Fig. 2 shows the shape of $(1/T)e^{j\omega T/2}H_M(\omega)$ for M up to 4. In Fig. 3 we have plotted the normalized transmission of $H_M(\omega)$ (that is, 20 Log_{10} $(H_M(\omega)/T|)$. Several of the $h_M(t)$ are shown in Fig. 4 and some values of $D_I(M, n)$ have been tabulated in Table I.

3.5 *Class II Weighting Functions (Taylor)*

Class I weighting functions provide the minimum high-order side-lobe amplitude in $H(\omega)$ that is possible with a given value of M. We now turn to the class that gives the minimum main-lobe width, at the expense of higher side-lobe amplitude.

The so-called continuous Dolph–Tchebycheff weighting functions[8]

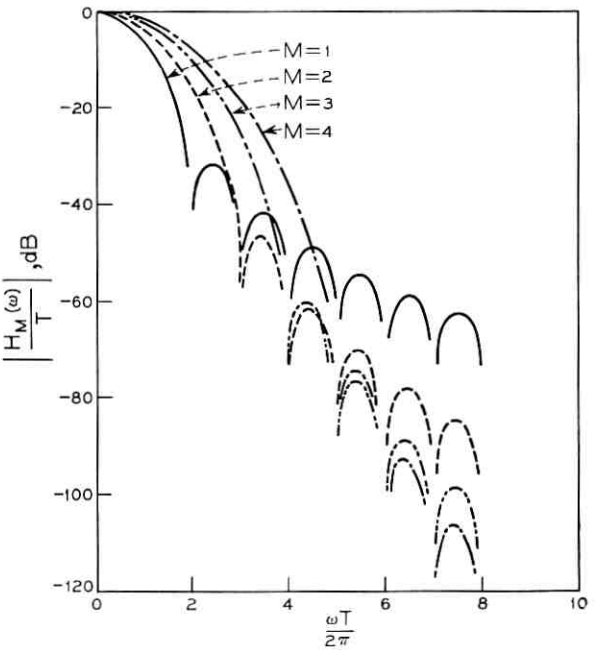


Fig. 3 — Normalized loss of the Class I weighting functions.

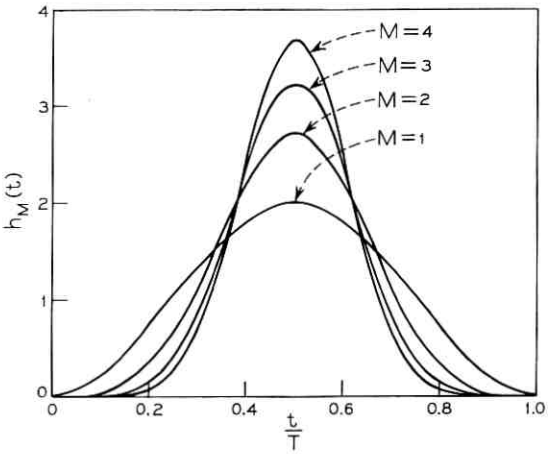


Fig. 4 — Time response of the Class I weighting functions.

provide the minimum main-lobe width in $H(\omega)$, consistent with a specified maximum side-lobe amplitude, but they are unrealizable functions. The Taylor[10] approximation to the Dolph–Tchebycheff functions provides almost the same main-lobe width and side lobes that have the specified maximum amplitude near the main lobe and then gradually decrease as ω increases.

The Taylor functions have the form given by equations (37) and (38) with the D_n dependent upon M and the maximum side-lobe amplitude, $1/R$. We will denote the D_n coefficients of Taylor weighting by $D_{II}(R, M, n)$, making the dependence upon R explicit. After adapting Taylor's equations to our situation, the D_{II}'s are given by

$$D_{II}(R, M, n) = -\frac{\prod_{K=1}^{M}\left[1 - \frac{(n/\sigma)^2}{\lambda^2 + (K - \frac{1}{2})^2}\right]}{\prod_{\substack{K=1 \\ K\neq n}}^{M}\left[1 - \left(\frac{n}{K}\right)^2\right]}, \tag{45}$$

where

$$R = \cosh(\pi\lambda) \tag{46}$$

and

$$\sigma^2 = \frac{(M + 1)^2}{\lambda^2 + (M + \frac{1}{2})^2}. \tag{47}$$

TABLE I —VALUES OF $D_I(M, n)$ FOR M UP TO 4

M	n			
	1	2	3	4
1	−1	—	—	—
2	−4/3	1/3	—	—
3	−3/2	3/5	−1/10	—
4	−8/5	4/5	−8/35	1/35

Solving equation (46) for λ gives

$$\lambda = \frac{1}{\pi} \ln [R + \sqrt{R^2 - 1}]. \tag{48}$$

We will refer to the Taylor functions described by equations (45) through (48) as Class II weighting functions and denote them by $k_M(R, t)$ and $K_M(\omega)$. References 10 and 11 give a further discussion of Taylor functions. Taylor weighting functions have the property that, if M is too small, the D's given by equations (45) will define an $H(\omega)$ whose first few side lobes have the amplitude given by equation (46), but some of the higher-order side lobes will have much higher amplitudes. Thus, for each value of desired side-lobe level, $1/R$, there is a minimum value of M that will give good side-lobe suppression.

Some minimum values of M that give good side-lobe control are listed in Table II.

Figures 5 and 6 show the shapes of a Class II weighting function with $M = 7$ and $R = 10^3$. This particular weighting function will be examined below when simulation results are compared. It will be shown there that this weighting function is useful when the received tone frequencies are very closely spaced.

3.6 *Class III Weighting Functions*

The third class of weighting functions has been chosen to have, to a

TABLE II—MINIMUM VALUES OF M FOR GOOD SIDE-LOBE CONTROL

$20 \log_{10} R$ (db)	M
36	3
42	4
48	5
54	6
60	7
66	9

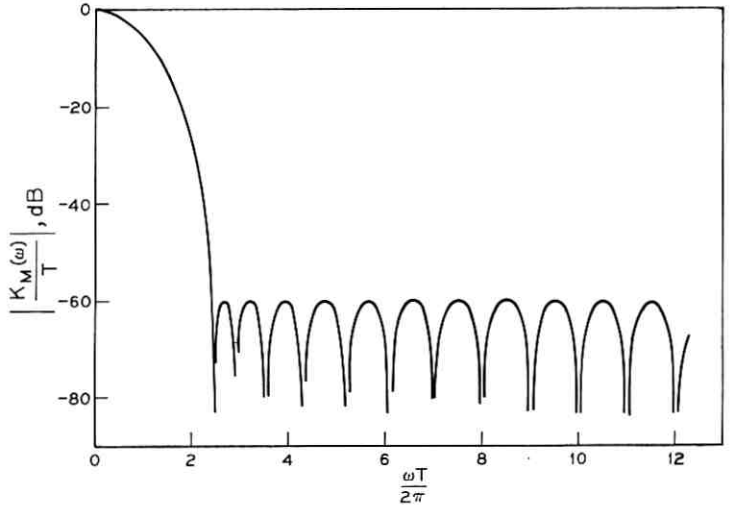


Fig. 5 — Normalized loss of the Class II weighting functions.

large extent, the desirable properties of both Class I and Class II weighting functions. That is, Class III weighting provides better resolution than Class I weighting for tones with a "small" frequency separation. Moreover, they also provide better resolution than Class II weighting of tones with a "large" frequency separation.

We will identify the Class III weighting functions by $g_M(t)$ and $G_M(\omega)$, where $g_M(t) \leftrightarrow G_M(\omega)$. The D_n coefficients for this class will be denoted by $D_{III}(M, n)$. The first member of the class is chosen as dis-

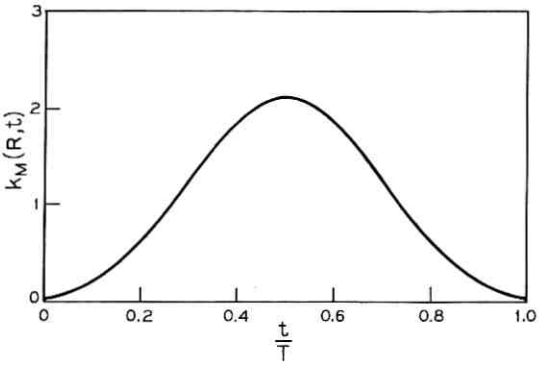


Fig. 6 — Time response of the Class II weighting functions; $M = 7$, $R = 1000$.

cussed below and the other members are obtained by operations on the first.

In order to have the high-order side lobes of $G_M(\omega)$ fall off at least as $1/\omega^2$ we must have $g_M(0+) = g_M(T-) = 0$. In terms of the D_n values this means that

$$1 + \sum_{n=1}^{M} D_{III}(M, n) = 0. \tag{49}$$

With this restriction, of course, $g_1(t)$ is the same as the Class I weighting function, $h_1(t)$. Thus the distinguishing properties of Class III weighting are determined by $g_2(t)$.

We chose the coefficients of $g_2(t)$ so that the loss of $G_2(\omega)$ reached 60 dB with as small a value of ω as possible with the side lobes of $G_2(\omega)$ never exceeding -60 dB, subject to equation (49). The D_n values for this condition are:

$$D_{III}(2, 1) = -1.19685, \qquad D_{III}(2, 2) = 0.19685.$$

The $g_2(t)$ thus defined is almost the same as Blackman's[5] proposed function, $Q_4(f)$.

The rest of the members of the Class III functions are defined in a manner similar to that used by Helms[6] for the synthesis of digital filters. We define

$$g_M(t) = \frac{h_{M-2}(t) g_2(t)}{1 + \frac{1}{2} D_{III}(2, 1) D_I(M-2, 1) + \frac{1}{2} D_{III}(2, 2) D_I(M-2, 2)}, \qquad M > 2. \tag{50}$$

The normalization in equation (50) puts $g_M(t)$ in the form of equation (37).

The Class III functions just defined have high-order side lobes, in $G_M(\omega)$, that decrease as ω^{-M}. This contrasts with $\omega^{-(M+1)}$ for Class I weighting and with ω^{-1} for Class II weighting. Thus, Class III weighting functions provide slightly narrower main-lobe width than Class I at the expense of slightly higher side lobes.

Some values of $D_{III}(M, n)$ are tabulated in Table III.

In Fig. 7 we have plotted the normalized spectra (that is, $(1/T)e^{j\omega T/2}G_M(\omega)$) of some Class III weighting functions. Fig. 8 illustrates the normalized loss provided by $G_M(\omega)$ for values of M up to 4. It is interesting to note, from Fig. 7, that $G_2(\omega)$ reaches -50 dB before any of the others, just as $H_2(\omega)$ did in Fig. 3. In Fig. 9 we have plotted $g_M(t)$ for values of M up to 4.

TABLE III—VALUES OF $D_{III}(M, n)$ FOR M UP TO 4

M	n			
	1	2	3	4
2	−1.19685	.19685	—	—
3	−1.43596	.497537	−.0615762	—
4	−1.566272	.725448	−.180645	.0179211

IV. RESPONSE TO A COSINE WAVE

We are interested in measuring the frequencies and levels of signals that comprise several sine waves. In view of this and the linearity of the DFT it is convenient to examine the properties of the DFT of samples of cos $(\omega_0 t + \theta)$.

4.1 *Basic Formulas*

Let

$$s(t) = \cos(\omega_0 t + \theta) \tag{51}$$

$$f(t) = h_T(t) \cos(\omega_0 t + \theta). \tag{52}$$

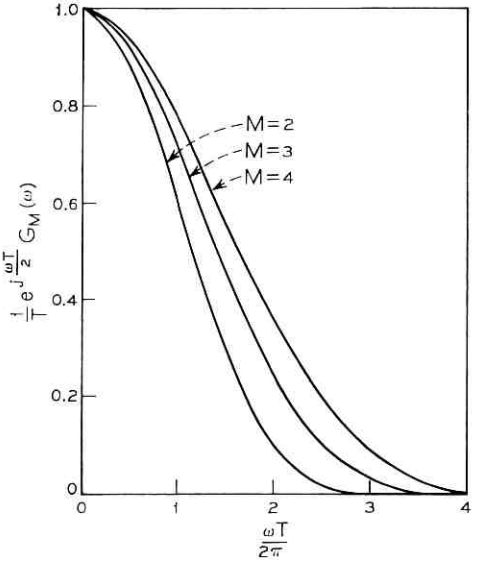


Fig. 7 — Spectra of the Class III weighting functions.

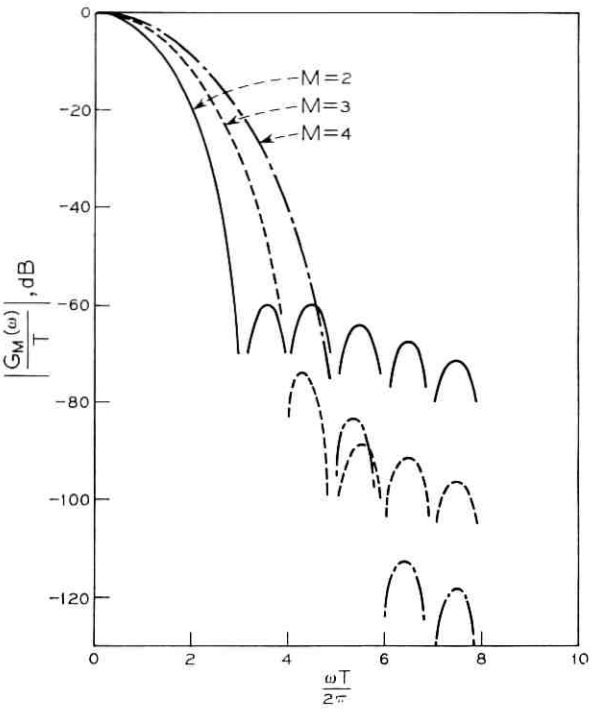


Fig. 8 — Normalized loss of the Class III weighting functions.

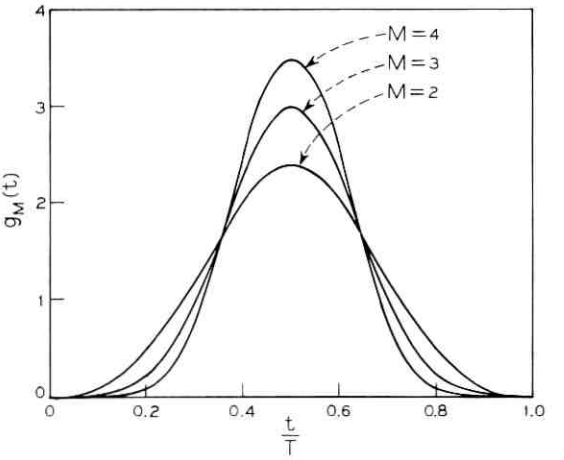


Fig. 9 — Time response of the Class III weighting functions.

and

$$H_T(\omega) = \int_{-\infty}^{\infty} h_T(t)e^{-i\omega t}\, dt \tag{53}$$

or, in the usual notation,

$$H_T(\omega) \leftrightarrow h_T(t). \tag{54}$$

Then

$$F^*(\omega) = \tfrac{1}{2}e^{i\theta}H_T^*(\omega - \omega_0) + \tfrac{1}{2}e^{-i\theta}H_T^*(\omega + \omega_0) \tag{55}$$

and from equation (19)

$$A(\omega) = \tfrac{1}{2}e^{i\theta}H_T^*(\omega - \omega_0) + \tfrac{1}{2}e^{-i\theta}H_T^*(\omega + \omega_0) + \tfrac{1}{2}[\cos\theta - \cos(\omega_0 T + \theta)e^{-i\omega T}]. \tag{56}$$

With $h_T(t)$ as defined by equation (22)

$$H_T(\omega) = Te^{-i\omega T/2}\frac{\sin(\omega T/2)}{\omega T/2}. \tag{57}$$

It is possible to put equation (57) in equation (56) and evaluate the indicated summations. The same answer can, however, be found more easily from equation (18). With equation (52) in equation (18) we have

$$A(\omega) = \sum_{n=0}^{N-1} \cos\left(\frac{\omega_0 nT}{N} + \theta\right)e^{-jn\omega T/N} \tag{58}$$

or

$$A(\omega) = \tfrac{1}{2}e^{j\theta}\sum_{n=0}^{N-1} e^{-j(\omega-\omega_0)nT/N} + \tfrac{1}{2}e^{-j\theta}\sum_{n=0}^{N-1} e^{-j(\omega+\omega_0)nT/N}. \tag{59}$$

Both sums in equation (59) are finite geometric series. Thus after rearrangement we obtain

$$A(\omega) = \tfrac{1}{2}e^{j\theta}e^{-j(N-1)z}\frac{\sin Nz}{\sin z} + \tfrac{1}{2}e^{-j\theta}e^{-j(N-1)y}\frac{\sin Ny}{\sin y} \tag{60}$$

where

$$z = \frac{(\omega - \omega_0)T}{2N} \tag{61}$$

and

$$y = \frac{(\omega + \omega_0)T}{2N}. \tag{62}$$

The formulation given by equation (56) is important from a conceptual point of view while that of equation (60) is useful for numerical evaluations.

The evaulation of equation (60) when either z or y is zero requires an examination of limits, which can be done by inspection.

4.2 *General* $A(\omega)$

The use of one of the weighting functions defined by equations (34), (37), and (38) gives rise to an $A(\omega)$ different from the one calculated in equation (60). The more generalized form of $A(\omega)$ is given by

$$A(\omega) = \tfrac{1}{2}e^{i\theta}e^{-iNz}\sin Nz \sum_{n=-M}^{M} C_n \frac{e^{i(z-n\pi/N)}}{\sin\left(z - \frac{n\pi}{N}\right)} + \tfrac{1}{2}e^{-i\theta}e^{-iNy}\sin Ny \sum_{n=-M}^{M} C_n \frac{e^{i(y-n\pi/N)}}{\sin\left(y - \frac{n\pi}{N}\right)}. \tag{63}$$

All of the simulations, to be discussed below, used this $A(\omega)$, in equation (21), to compute A_K values.

4.3 *Approximations*

We will make use of several approximations in the next section. The important ones are established here. In this section $h_T(t)$ is assumed to be the weighting function and ω is in the range $0 < \omega < \omega_s/2$.

Consider equation (19). If $| F(\omega - l\omega_s) |$ is small for $l \neq 0$; then

$$A(\omega) \cong \frac{1}{t_s} F(\omega) + \tfrac{1}{2}[f(0+) - f(T-)e^{-i\omega T}] \tag{64}$$

where

$$t_s = T/N. \tag{65}$$

Thus, from equation (21),

$$A_K \cong \frac{1}{t_s} F\left(\frac{K\omega_s}{N}\right) + \tfrac{1}{2}[f(0+) - f(T-)]. \tag{66}$$

Next consider equations (56) and (57). $| H_T(\omega) |$ is "large" only near $\omega = 0$. Thus we obtain another approximation, used when $s(t) = \cos(\omega_0 t + \theta)$,

$$F(\omega) \approx \tfrac{1}{2}e^{i\theta}H_T(\omega - \omega_0). \tag{67}$$

From equation (57) we then obtain

$$\frac{1}{t_s} F(\omega) \cong \frac{N}{2} e^{j\theta} e^{-j[(\omega-\omega_0)T/2]} \frac{\sin [(\omega - \omega_0)T/2]}{[(\omega - \omega_0)T/2]}. \tag{68}$$

Thus,

$$| A_K | \cong \frac{N}{2} \frac{\left| \sin \left(\frac{K\omega_s T}{2N} - \frac{\omega_0 T}{2} \right) \right|}{\left| \left(\frac{K\omega_s T}{2N} - \frac{\omega_0 T}{2} \right) \right|} \tag{69}$$

or, because,

$$\frac{K\omega_s T}{2N} = K\pi, \tag{70}$$

$$| A_K | \cong \frac{N}{2} \frac{\left| \sin \left(K\pi - \frac{\omega_0 T}{2} \right) \right|}{\left| \left(K\pi - \frac{\omega_0 T}{2} \right) \right|}. \tag{71}$$

From equation (71) we see that, apart from the error in the approximations, one should be able to accurately estimate the frequency and magnitude of a cosine wave from the A_K values for K near $\omega_0 T/2\pi$.

The main lobe of a $\sin X/X$ curve reaches zero at $X = \pm\pi$. Thus the main lobe of the curve, of which the A_K are points, reaches zero at $\omega = \omega_0 \pm 2\pi/T$ or, since $\omega_s/N = 2\pi/T$, at

$$\omega = \omega_0 \pm \omega_s/N.$$

The main lobe is just wide enough to contain two A_K values (estimators), except if ω_0 equals some multiple of ω_s/N. It will be shown later that two A_K values will be enough to estimate the parameters of a cosine wave.

V. FREQUENCY AND LEVEL ESTIMATION

In the preceding sections we have developed methods of computing the DFT coefficients of a known input (for simulations) and have discussed three classes of weighting functions to improve the effective selectivity of the DFT process. Our final task, which is undertaken in this section, is to determine accurate ways of processing the DFT coefficients to extract the frequencies and magnitudes present in the sampled signal.

The methods are to be useful when the real-time demands upon the

computer are important. Thus all of the methods were chosen to have simple formulas which the computer can be programmed to evaluate when it is making a measurement.

The equations we have examined fall into two classes: those that use only two estimators to calculate the frequency and level of a cosine wave and those that use many. The following paragraphs derive the most promising of these equations. The next section will present the accuracies that the various equations can achieve. We start with a formula that makes use of many estimators.

5.1 *Method 1*

The derivation of this method is somewhat involved, so we first explain the motivation behind it.

Suppose one has an $f(t)$ that is known to be given by

$$f(t) = B \cos \omega_0 t, \tag{72}$$

and one wants to determine ω_0 and B from operations on $f(t)$. One way to determine B is from

$$\begin{aligned} \lim_{T\to\infty} \frac{1}{T} \int_0^T f^2(t)\, dt &= \lim_{T\to\infty} \frac{1}{T} \int_0^T B^2/2\, dt \\ &= B^2/2 \end{aligned} \tag{73}$$

thus,

$$B^2 = 2 \lim_{T\to\infty} \frac{1}{T} \int_0^T f^2(t)\, dt. \tag{74}$$

The derivative of $f(t)$ is

$$f'(t) = -B\omega_0 \sin \omega_0 t \tag{75}$$

and

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T [f'(t)]^2\, dt = B^2\omega_0^2/2 \tag{76}$$

Thus ω_0 can be determined from

$$\omega_0^2 = \frac{\lim_{T\to\infty} \frac{1}{T} \int_0^T [f'(t)]^2\, dt}{\lim_{T\to\infty} \frac{1}{T} \int_0^T [f(t)]^2\, dt}. \tag{77}$$

The above result is the motivation for the following derivation.

Assume

$$s(t) = B \cos (\omega_0 t + \theta) \tag{78}$$

and

$$f(t) = h(t)B \cos (\omega_0 t + \theta) \tag{79}$$

where $h(t)$ is one of the weighting functions, given by

$$h(t) = \left[1 + \sum_{n=1}^{M} D_n \cos \left(\frac{2\pi n t}{T}\right)\right] h_T(t). \tag{80}$$

Let the estimators, A_K, be given by equation (31).

Define

$$P_0 = \lim_{T\to\infty} \frac{1}{T} \int_0^T [f(t)]^2 \, dt \tag{81}$$

and

$$P_1 = \lim_{T\to\infty} \frac{1}{T} \int_0^T [f'(t)]^2 \, dt. \tag{82}$$

Expansion of equation (79) gives

$$f(t) = Bh_T(t)\Big\{\cos (\omega_0 t + \theta) + \frac{1}{2} \sum_{n=1}^{M} D_n[\cos (\omega_0 t - n\omega_a t + \theta) + \cos (\omega_0 t + n\omega_a t + \theta)]\Big\} \tag{83}$$

where

$$\omega_a = 2\pi/T. \tag{84}$$

From equations (81) and (83)

$$P_0 = \frac{B^2}{2}\left\{1 + \frac{1}{2} \sum_{n=1}^{M} D_n^2\right\}. \tag{85}$$

Calculation of P_1 from equation (83) yields, assuming $n\omega_a \neq \omega_0$,

$$P_1 = \frac{B^2}{2}\left\{\omega_0^2 + \sum_{n=1}^{M} \frac{D_n^2\omega_0^2}{2} + \sum_{n=1}^{M} \frac{n^2 D_n^2 \omega_a^2}{2}\right\}. \tag{86}$$

Combining equations (85) and (86) gives

$$\omega_0^2 = \frac{P_1}{P_0} - \frac{\omega_a^2 \sum_{n=1}^{M} n^2 D_n^2}{2 + \sum_{n=1}^{M} D_n^2}. \tag{87}$$

The next step is to compute P_1/P_0 from the DFT of $f(t)$. Using the approximation

$$f(t) \approx \frac{h_T(t)}{N} \sum_{k=-N/2}^{N/2} A_K e^{jK\omega_a t} \tag{88}$$

which can be rearranged to give

$$f(t) \approx \frac{h_T(t)}{N} \left\{ A_0 + 2 \sum_{K=1}^{N/2} [\mathrm{Re}\,(A_K) \cos K\omega_a t - \mathrm{Im}\,(A_K) \sin K\omega_a t] \right\}. \tag{89}$$

From equation (89)

$$P_0 \approx \frac{1}{N^2} \left\{ A_0^2 + 2 \sum_{K=1}^{N/2} | A_K |^2 \right\} \tag{90}$$

and

$$P_1 \approx \frac{1}{N^2} \sum_{K=1}^{N/2} 2 | A_K |^2 K^2 \omega_a^2 . \tag{91}$$

Thus,

$$\frac{P_1}{P_0} \approx \frac{\omega_a^2 \sum_{K=1}^{N/2} K^2 | A_K |^2}{\frac{A_0^2}{2} + \sum_{K=1}^{N/2} | A_K |^2} . \tag{92}$$

We use equation (92) in equation (87) to obtain the final result. Denote the estimated ω_0 by $\hat{\omega}_0$. Then

$$\hat{\omega}_0^2 = \omega_a^2 \left\{ \frac{\sum_{K=1}^{N/2} K^2 | A_K |^2}{\frac{1}{2} A_0^2 + \sum_{K=1}^{N/2} | A_K |^2} - U_M \right\} \tag{93}$$

where

$$U_M = \frac{\sum_{n=1}^{M} n^2 D_n^2}{2 + \sum_{n=1}^{M} D_n^2} \tag{94}$$

By D_n we mean, of course, $D_{\mathrm{I}}(M, n)$, $D_{\mathrm{II}}(R, M, n)$ or $D_{\mathrm{III}}(M, n)$. Thus Method 1 is applicable to all three classes of weighting functions. Some values for U_M are tabulated in Table IV.

TABLE IV—VALUES OF U_M AND V_M

Class I Weighting

M	U_M	V_M
0	0	.5
1	.333333	.75
2	.571429	.972222
3	.818182	1.155
4	1.06667	1.31327

Class II Weighting

($M = 7$, $R = 1000$)

$U_7 = .357071$, $V_7 = .769710$

Class III Weighting

M	U_M	V_M
2	.45732	.8678
3	.715523	1.07833
4	.968949	1.25033

The way to use equation (93), when more than one tone is present in $s(t)$, is to use only the estimators, A_K, for $K \approx \omega_0/\omega_a$ to calculate each $\hat{\omega}_0$. The simulations below will show that this technique gives accurate results.

5.1.1 *Estimation of Level*

From equations (85) and (90), we obtain the way to estimate B.

$$\hat{B}^2 = \frac{1}{N^2} \frac{A_0^2 + 2 \sum_{K=1}^{N/2} | A_K |^2}{V_M} \tag{95}$$

where

$$V_M = \tfrac{1}{2} + \tfrac{1}{4} \sum_{n=1}^{M} D_n^2 . \tag{96}$$

Some values of V_M are tabulated in Table IV. Observe that if only the basic weighting, $h_T(t)$, is used, then U_M and V_M become zero and $\frac{1}{2}$, respectively.

5.2 *Method 2*

The preceding formulas for estimating the frequency and level of a cosine wave from its DFT use more than two estimators. The next few paragraphs establish formulas that require only two estimators. The formulas apply only to the Class I weighting functions, $h_M(t)$.

We start by recalling the approximations given by equations (64) and (67). With $H_M(\omega)$ substituted for $H_T(\omega)$ in equation (67) we have

$$A(\omega) \approx \frac{1}{2t_s} e^{j\theta} H_M(\omega - \omega_0). \tag{97}$$

From equation (40)

$$H_M\left(\frac{2X}{T}\right) = T e^{-iX} \frac{\sin X}{X} \frac{(-1)^M \pi^{2M} (M!)^2}{\prod_{n=1}^{M} (X^2 - n^2\pi^2)} \tag{98}$$

where $X = \omega T/2$. If $s(t) = B \cos(\omega_0 t + \theta)$ then, from equations (97) and (98),

$$| A(\omega) | \approx \left| (-1)^M (M!)^2 \frac{BN}{2} \frac{\sin \pi v}{\pi \prod_{n=-M}^{M} (v + n)} \right| \tag{99}$$

where

$$v = (\omega_0 - \omega)/\omega_a . \tag{100}$$

Suppose the largest† estimator is A_l and its largest immediate neighbor A_m. Of course $m - l = \pm 1$. Define

$$\alpha = m - l = \pm 1. \tag{101}$$

Let

$$a_1 = | A_l | \tag{102}$$

and

$$a_2 = | A_m | = | A_{l+\alpha} |. \tag{103}$$

Define

$$u = \frac{\omega_0}{\omega_a} - l; \quad -1/2 \leqq u \leqq 1/2. \tag{104}$$

Then from equation (99)

$$a_1 \approx (-1)^M (M!)^2 \frac{BN}{2} \frac{\sin \pi u}{\pi \prod_{n=-M}^{M} (u + n)} \tag{105}$$

† By largest we mean $| A_l | \geq | A_K |$ for $K \neq l$.

and

$$a_2 \approx (-1)^M (M!)^2 \frac{BN}{2} \frac{\sin \pi(u - \alpha)}{\pi \prod_{n=-M}^{M} (u + n - \alpha)} \tag{106}$$

since $\alpha = \pm 1$, equation (106) is the same as

$$a_2 \approx -(-1)^M (M!)^2 \frac{BN}{2} \frac{\sin \pi u}{\pi \prod_{n=-M-\alpha}^{M-\alpha} (u + n)} . \tag{107}$$

Division of equation (105) by equation (107) gives

$$\frac{a_1}{a_2} \approx -\frac{u - \alpha(M + 1)}{u + \alpha M} . \tag{108}$$

Define

$$u_1 = \frac{a_2(M + 1) - a_1 M}{a_1 + a_2} \tag{109}$$

then from equation (108) an estimate of u is

$$\hat{u} = \alpha u_1 . \tag{110}$$

Hence, the estimate of ω_0 is given by

$$\hat{\omega}_0 = \omega_a(l + \hat{u}). \tag{111}$$

From equation (105) the estimate of B is

$$\hat{B} = \frac{a_1 2\pi(-1)^M \prod_{n=-M}^{M} (\hat{u} - n)}{N(M!)^2 \sin \pi\hat{u}} , \tag{112}$$

where $\hat{u}$ is defined by equation (110).

Another version of equation (112), better for machine computation, is

$$\hat{B} = \frac{2a_1\pi\hat{u}}{N \sin (\pi\hat{u})} \prod_{n=1}^{M} 1 - \left(\frac{\hat{u}}{n}\right)^2 . \tag{113}$$

Method 2 with $M = 0$ is essentially the same as was derived by Penhune and Martin[14] to solve a radar problem.

5.3 *Method 3*

The simplicity of the estimation equations of Method 2 led us to extend this method to include any class of weighting functions described, in general form, by equations (37) and (38). We will refer to this more general method as Method 3.

From equations (109) and (110) we see that if a Class I weighting function is used, one way to obtain an estimate of u is to use the function

$$u_1 = \frac{Ca_2 - Da_1}{Ea_2 + a_1} \tag{114}$$

in equation (110). There are three degrees of freedom in the bilinear form and we chose to express them in the manner shown in equation (114).

Method 3 is simply the application of equation (114) to other classes of weighting functions. We obtained values for the coefficients in equation (114), for several weighting functions by:

(i) Computing a_1 and a_2, using equations (21), (63), (102), and (103), with $\Theta = 0$ and many values for ω_0 near $\omega_s/4$.

(ii) Computing the corresponding values of u from equation (104).

(iii) Choosing values for C, D, and E such that equation (114) gave a good fit to the data computed in the first two steps. It turns out that the curve described by equation (114) is satisfactory if it fits the computed data exactly at $u = 0$, $\frac{1}{4}$, and $\frac{1}{2}$.

In this manner we obtained the following coefficient values:

Class II weighting, $M = 7$, $R = 1000$;
$C = 1.96339$, $D = 1.01643$, $E = 0.893534$.

Class III weighting, $M = 2$; $C = 2.56919$,
$D = 1.5374$, $E = 1.06345$. For $M = 3$;
$C = 3.6020$, $D = 2.5862$, $E = 1.0317$.

Using an approximation similar to equation (71), but extended to include weighting functions described by equation (38), we obtain an estimate for B,

$$\hat{B} = \frac{2\pi a_1}{N \sin(\pi u_1)\left[\frac{1}{u_1} + \sum_{n=1}^{M} \frac{D_n u_1}{u_1^2 - n^2}\right]}. \tag{115}$$

Method 3 can be used with weighting functions specified only in terms of $H^*(\omega)$ as well as those given in terms of $H(\omega)$.

VI. COMPARISON OF ACCURACIES

In the preceding paragraphs we have derived several formulas that produce estimates of ω_0 and B from A_K values (estimators). We now turn to a comparison of these formulas on the basis of accuracy. The

accuracies we will compare do not include any possible computation or nonlinearity errors or other accuracy limitations that may be present in a DFT analysis system. Our accuracy comparisons include only the effects of leakage.

The estimators used in the simulations were generated by using the function $A(\omega)$, described by equation (63), in equation (21). This is equivalent to applying the weighting by multiplication in the time domain or by convolution in the frequency domain. The use of equation (63) in simulations greatly reduces computation time. All of the simulations used $N = 512$ and $f_s = N/T = 7040$ Hz.

All of the estimation methods presented above use approximations. In this section we shall demonstrate just how good the approximations are.

Consider the case where a tone of frequency f_0, angle θ_0, and amplitude B_0 is being measured while another tone, at frequency f_1, angle θ_1, and level B_1, is also being received. The presence of f_1 will affect the accuracy of any estimate one makes of f_0 or B_0 (due to leakage). The size of the errors in the estimates of f_0 and B_0 will depend upon which formula (method) is used and upon the values of f_0, B_0, θ_0, f_1, B_1, and θ_1. The combination of parameters that causes one method to give the worst estimates will, in general, not be the combination that causes another method to be at its worst. Thus it is difficult to compare methods.

We have compared the three methods on the basis of the worst estimates each will make when θ_0 and f_0 are confined to a specified range of values (for example, $990 \leqq f_0 \leqq 1003.75$ Hz and $0 \leqq \theta_0 \leqq 360$ degrees) and B_0, B_1, f_1, and θ_1 are fixed (for example, $B_0 = B_1 = 1$ and $\theta_1 = 0$ degrees).

Notice that if f_1 is equal to some multiple of $1/T$ then its A_K will be very small except for K near Tf_1. Thus such an f_1 cannot cause much error in any of the three methods of estimation of f_0, which use A_K values. For this reason we have fixed f_1 at a value that is an odd multiple of $\frac{1}{2}T$.

Tables V and VI illustrate how inaccurate frequency and level esti-

TABLE V—POOREST ESTIMATES WITH INTERFERENCE SEPARATION IN THE RANGE 55 ± 6.88 Hz, NO SPECIAL WEIGHTING

Method	Frequency Error, Hz	Magnitude Error, dB
1†	6.82	.596
2	7.37	.580

† Method 1 using only six estimators, those from $l - 2\alpha$ to $l + 3\alpha$.

TABLE VI—POOREST ESTIMATES WITH INTERFERENCE SEPARATION IN THE RANGE 178.96 ± 6.88 Hz, NO SPECIAL WEIGHTING

Method	Frequency Error, Hz	Magnitude Error, dB
1†	3.08	.0628
2	3.75	.165

† Method 1 using only six estimators.

mates are when no special weighting ($M = 0$) is used. The interfering tones corresponding to the simulations described in Tables V and VI were located at 1051.88 and 1175.63 Hz respectively. The frequency and magnitude error entries in these, and subsequent, tables give absolute values only.

The simulation results presented in Tables VII and VIII indicate that substantial improvement in the accuracy of frequency and magnitude estimates can be achieved when weighting is used. The data in Table VII shows that when the two tones are separated by a "small" frequency difference accurate frequency and level estimates can be obtained by using Method 3 with Class II or Class III weighting. Table VIII indicates that as the frequency separation increases Method 2 with Class I weighting is better. The accuracy of estimates made on closely spaced tones can, of course, always be improved by increasing N and T while keeping the ratio N/T constant.

It is interesting to note from Figs. 3 and 8 that, for a given value of

TABLE VII—POOREST ESTIMATES WITH INTERFERENCE SEPARATION IN THE RANGE 55 ± 6.88 Hz

Class I Weighting			
Method	M	Frequency Error, Hz	Magnitude Error, dB
1	1	.89	.11
1	2	3.32	.47
1	3	6.16	.96
2	1	.513	.12
2	2	.104	.11
2	3	.409	.5
Class II Weighting, R = 1000			
Method	M	Frequency Error, Hz	Magnitude Error, dB
1	7	1.14	.120
3	7	.0651	9.35E-3
Class III Weighting			
Method	M	Frequency Error, Hz	Magnitude Error, dB
1	2	2.11	.225
1	3	5.12	.821
3	2	.034	.026
3	3	.149	.0655

TABLE VIII—POOREST ESTIMATES WITH INTERERENCE SEPARATION IN THE RANGE 178.96 ± 6.88 Hz

Class I Weighting

Method	M	Frequency Error, Hz	Magnitude Error, dB
1	1	1.67E-3	2.79E-4
1	2	3.98E-4	5.25E-5
1	3	4.55E-2	4.70E-3
2	1	5.73E-3	1.42E-3
2	2	1.47E-4	2.19E-5
2	3	2.13E-5	1.55E-6

Class II Weighting, R = 1000

Method	M	Frequency Error, Hz	Magnitude Error, dB
1	7	5.56E-3	2.05E-5
3	7	8.09E-2	1.07E-2

Class III Weighting

Method	M	Frequency Error, Hz	Magnitude Error, dB
1	2	9.10E-5	1.19E-5
1	3	1.85E-2	1.90E-3
3	2	3.35E-3	3.24E-3
3	3	5.94E-4	9.62E-6

M, there is not a great deal of difference between the weighting contributed by $H_M(\omega)$ or $G_M(\omega)$. However, from Table VII it is obvious that the use of $G_2(\omega)$, when the tones are close together, will yield much more accurate estimates than Class I weighting.

In equation (101) the "pointer", α, was defined. The value of α is used by the system to determine whether the frequency being measured is above or below the frequency of the largest estimator, A_l. Our simulation studies showed that under certain circumstances α, as calculated by equation (101) will point in the wrong direction. In general this happens when the contributions to A_{l-1} and A_{l+1} due to the interference is equal to or greater than the difference in the contributions to the same estimators due to the tone being measured. Thus if $|A_{l-1}| \approx |A_{l+1}|$, a small difference in their magnitudes can change α. In our simulations this effect only caused trouble when f_0 and f_1 were separated by less than half the width of the main lobe of the weighting function, $H(\omega)$.

Since we have fixed $B_0 = B_1 = 1$ in the simulations we have ignored the adverse effects of "large" level differences on the accuracies of the various methods. Leakage from an interfering tone with a high level, relative to the tone of interest, would certainly tend to reduce the accuracy provided by any of the three methods, no matter which weighting is used.

VII. CONCLUSIONS

The discrete-tone measurement system we have been discussing is particularly well suited to systems that involve computer-controlled testing or measurement, provided the real time needed for the computations is available. Two advantages are:

(i) The only interface hardware is the A-D converter (with its lowpass filter).

(ii) The system is capable of measuring many received frequencies and levels during the same computation time.

For a given number of samples taken at a given sampling rate, the accuracy of the system can be significantly improved through the use of some type of weighting function. We have examined three classes of such smoothing functions and have developed formulas which permit the extraction of received signal frequencies and levels from the DFT coefficients. The results of system simulations, presented in Tables V through VIII, show that the inherent accuracy of the described system can, through the proper use of weighting functions and estimation methods, be made satisfactory for many applications.

With Method 2 considered to be a special case of Method 3, the tables show that the best estimation method, for all of the weighting functions examined, is Method 3.

The tables also show that there is no "best" weighting function. The weighting to be used for any particular application should be selected only after a consideration of the expected tone frequencies, the relative levels, the measurement accuracy desired, and the desired value for N. The sampling frequency, N/T, should be more than twice the highest frequency to be measured.

It is interesting to observe that the Taylor (Class II) weighting function used in the simulation is, *for the situations simulated*, not significantly better than the Class III weighting, $G_2(t)$, which is essentially that proposed by Blackman.[5] There may be other situations, however, when the nearly optimum main-lobe width of the Taylor functions is useful.

If the system could tolerate the relatively large amount of computer memory required, then the discrete Dolph–Chebychev functions described by equation (27) could provide some advantages.

VIII. ACKNOWLEDGMENTS

We are indebted to Messrs. D. R. Johnson, B. R. Saltzberg, and R. A. Smith for helpful discussions.

REFERENCES

1. Cochran, W. T., and others, "What is the Fast Fourier Transforms?," *IEEE Trans. Audio and Electroacoustics, AU-15,* No. 2 (June 1967), pp. 45–55.
2. Gentleman, W. M. and Sande, G., "Fast Fourier Transforms—for Fun and Profit," *American Federation of Information Processing Societies Proc., 29,* Washington, D. C.: Spartan, 1966, pp. 563–578.
3. Bergland, G. D., "A Guided Tour of the Fast Fourier Transform," *IEEE Spectrum, 6,* No. 7 (July 1969), pp. 41–52.
4. Papoulis, A., *The Fourier Integral and Its Applications,* New York: McGraw-Hill, 1962, pp. 48, 49.
5. Blackman, R. B. and Tukey, J. W., *The Measurement of Power Spectra,* New York: Dover, 1958, pp. 95–100.
6. Helms, H. D., "Nonrecursive Digital Filters: Design Methods for Achieving Specifications on Frequency Response," *IEEE Trans., AU–16,* No. 3 (September 1968), pp. 336–342.
7. Bingham, C., Godfrey, M. D., and Tukey, J. W., "Modern Techniques of Power Spectrum Estimation," *IEEE Trans. Audio and Electroacoustics, AU-15,* No. 2 (June 1967), pp. 56–66.
8. Dolph, C. L., "A Current Distribution for Broadside Arrays which Optimizes the Relationship between Beam Width and Side-Lobe Level," *Proc. IRE, 34,* No. 6 (June 1946), pp. 335–348.
9. Cook, C. E. and Bernfeld, M., *Radar Signals—An Introduction to Theory and Application,* New York: Academic Press, 1967, pp. 178–182.
10. Taylor, T. T., "Design of Line-Source Antennas for Narrow Beamwidth and Low Side Lobes," *IRE Trans., AP-3,* No. 1 (January 1955), pp. 16–28.
11. Klander, J. R., and others, "Theory and Design of Chirp Radars," B.S.T.J., *39,* No. 4 (July 1960), pp. 782–790.
12. Kuo, F. F. and Kaiser, J. F., *Systems Analysis by Digital Computer,* New York: Wiley, 1966, Chapter 7.
13. Larsen, A. G. and Singleton, R. C., "Real-Time Spectral Analysis on a Small General-Purpose Computer," *1967 Fall Joint Computer Conf., AFIPS Proc.,* Washington, D. C.: Spartan, 1967, pp. 665–674.
14. Penhune, J. P. and Martin, L. R., "Determination of Doppler Velocity and Ballistic Coefficient from Coherent Radar Data," ESD-TDR-65-41, Technical Rept. 378, M.I.T. Lincoln Labs., Lexington, Mass., pp. 11–53.

Reed-Contact Switch Series for the I.F. Band

By M. B. PURVIS and R. W. KORDOS

(Manuscript received October 17, 1968)

A series of switches using a miniature dry-reed sealed contact in a cable switch configuration has been developed to provide switching capability from dc to 100 MHz. We present a description of the development, performance characteristics, and mechanical design features.

I. INTRODUCTION

The nationwide network of transmission facilities is not only growing in number of routes and capacity but also in terms of service capability and administrative flexibility. Within the network there are usually alternate routes for providing service between two points. Interconnection between points may ultimately be controlled by a remote, centralized, real time machine that contains an accurate map of the state of the network.

The broadband restoration system, for example, can detect failures, make routine maintenance checks and report to a regional control center where an alternate route between the two points is selected. The control center then remotely operates the wideband switch at each junction of the route to effect a restoration of service.

One component group needed to implement these systems is a family of wideband switches capable of meeting the transmission requirements of low insertion loss, high isolation loss, high crosstalk loss, and having an impedance well matched to the 75-ohm system impedance.

The 266B (8 × 8 matrix), 274A (1 × 8), and 273B (1 × 2) switches have been developed to meet these requirements with low operate power, small size, and moderate cost. All of these codes use 237-type miniature dry-reed sealed contacts in a cable-switch[1] arrangement to provide an extremely high isolation loss in the open state and a low insertion loss and good impedance match in the closed state. Appropriate matrix configurations are achieved by interconnecting the cable switches with

stripline networks designed to provide good system performance from dc to 100 MHz. The requirements, performance characteristics, mechanical design features, and a description of the development of these new wideband switching matrices are presented in this paper.

II. REQUIREMENTS

The restoration transmission requirements for an 8 × 8 matrix of 64 crosspoints used to interconnect 75-ohm coaxial transmission paths over the frequency range of dc to 100 MHz are:

Insertion loss (closed-contact loss)	0.6-dB maximum
Isolation loss (open-contact loss)	95-dB minimum
Crosstalk loss	95-dB minimum
Return loss	28-dB minimum

Transmission requirements for the 1 × 2 matrix and the 1 × 8 matrix are identical to those of the 8 × 8 matrix except for the crosstalk loss requirement, which is not pertinent in single-level matrices. Where more than one switch is enclosed in the same housing, the crosstalk requirement will apply between switches.

Speed of switching is not a stringent requirement, and operation in the millisecond range is satisfactory. Compact matrix size and moderate manufacturing cost are additional features required for practical application in the restoration switching systems.

The switch arrays, 8 × 8, 1 × 8, and 1 × 2 are illustrated schematically in Fig. 1. Photographs of the three switch types are shown in Fig. 2. The major elements of the design in a transmission sense are the coaxial crosspoint developed from the cable switch, the input/output circuit boards, and the coaxial jacks. From the schematic diagram, one can see that in the 8 × 8 and 1 × 8 designs the closure of any crosspoint leaves seven open crosspoints connected by stubs on each associated circuit board or "tree." Because of the length of these stubs, the structural considerations become important design parameters that seriously affect performance. The design considerations in each of these elements are presented in the sections that follow.

III. CABLE SWITCH THEORY

An extremely high isolation-loss requirement in the megahertz range normally precludes the use of conventional electromechanical switches,

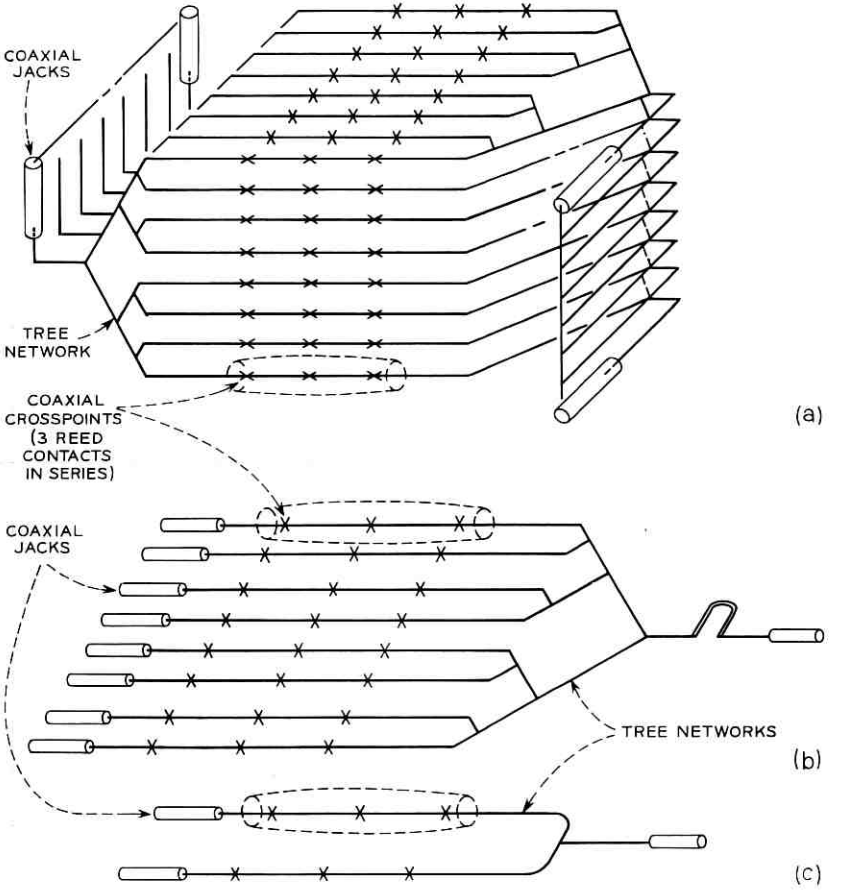


Fig. 1 — Array schematic: (a) 8 × 8, (b) 1 × 8, (c) 1 × 2.

such as the wire spring and flat spring relay, the crossbar switch, and the ferreed, because of their generally large open-contact capacitance of between 0.1 and 1.0 pF. However, an arrangement of two or more conventional switching elements connected in series by a length(s) of low-loss transmission cable is particularly well suited for operation in broadband switching applications where extremely high isolation is required. This broadband switching crosspoint is called the cable switch.

Applying conventional lumped constant analysis to a string of open switches (that is, serially connected switches with substantially zero transmission paths between them) produces the following conventional and well known voltage divider approximation expression:

$$| V_{out}/V_{in} | = \omega CR/K + 1 \quad (1)$$

when $\omega CR \ll 1$ and where

ω = angular frequency,
C = open switch capacitance,
$K + 1$ = number of switches, and
R = the load impedance.

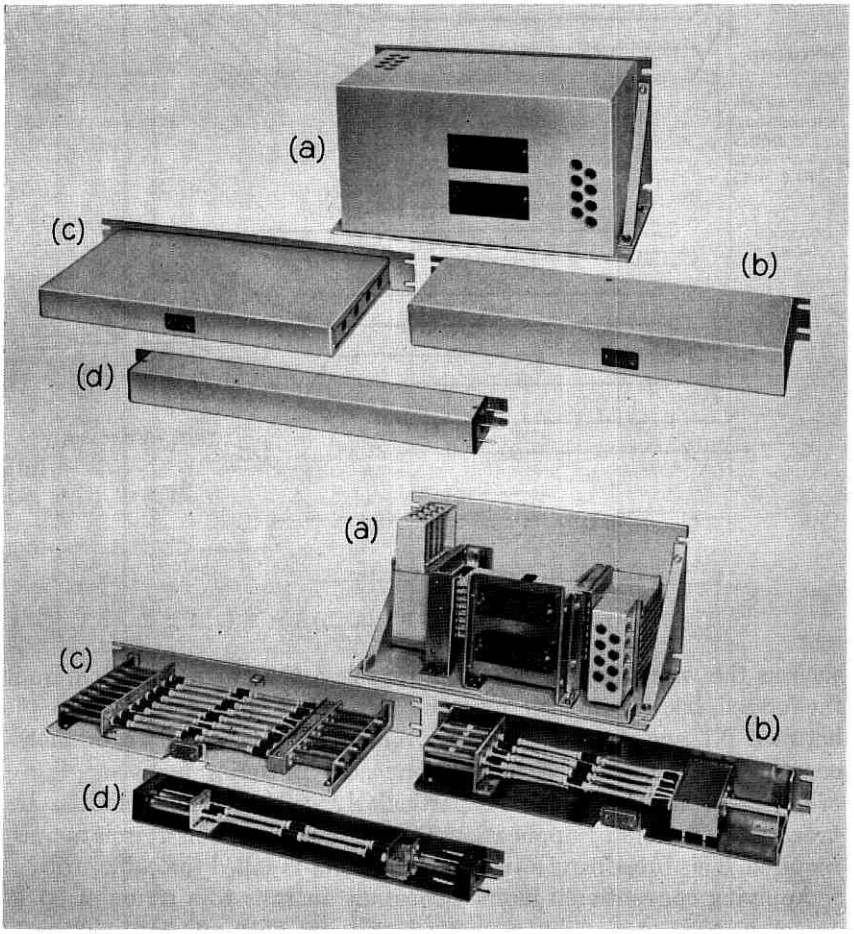


Fig. 2 — Photograph of switch arrays: (a) 8 × 8, (b) 1 × 8, (c) 1 × 2 (4 switches per package), (d) 1 × 2 (1 switch).

In other words, each switch added to the string reduces the ratio by adding one to the denominator.

The above expression does not apply, however, when coaxial cables are connected between the switches. In particular, for $(K + 1)$ switches with K pieces of identical lossless coaxial cables interconnecting them, the following expression applies (see Appendix):

$$\left| \frac{V_{out}}{V_{in}} \right| = \frac{\omega CR}{A_K + (A - \omega CZ_0 \sin \beta d) \sum_{n=0}^{K-1} A^{K-1-n} A_n}$$

$$\left| \frac{V_{out}}{V_{in}} \right| = \frac{\omega CR}{\sum_{n=0}^{K} A^{K-n} A_n} \tag{2}$$

for practical components where the values of A_n are given in the following table:

n	A_n
0	1
1	A
2	$2A^2 - 1$
3	$4A^3 - 3A$
⋮	⋮
K	$2AA_{K-1} - A_{K-2}$

and $A = \cos \beta d + \sin \beta d / 2\omega CZ_0$ for lossless lines where: $\beta = \omega(\epsilon_R \mu_R / c)^{\frac{1}{2}}$ (phase constant).

ϵ_R = relative dielectric constant of coaxial cable,
μ_R = relative permeability constant of coaxial cable,
c = velocity of light,
d = distance between contacts from switch to switch,
Z_0 = characteristic impedance of coaxial cable in ohms, and the remaining symbols have the same meanings as in equation (1).

When the length of transmission line, d, between two switching elements equals one-quarter wavelength, equation (2) indicates that the isolation loss in dB of the overall switch is twice the isolation loss of the individual switching element. However, a plot of equation (2) for the

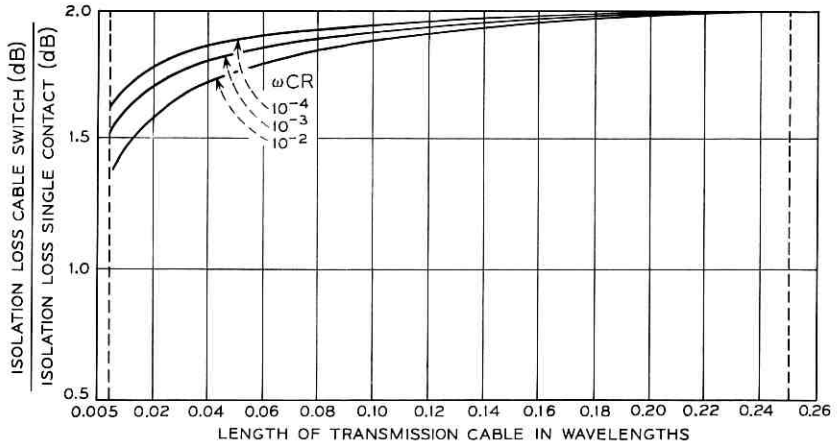


Fig. 3 — Isolation loss improvement for the cable switch. Multiplying factor for the isolation loss in dB of a two-element cable switch as a function of cable length between switching elements. ($Z_0 = R$).

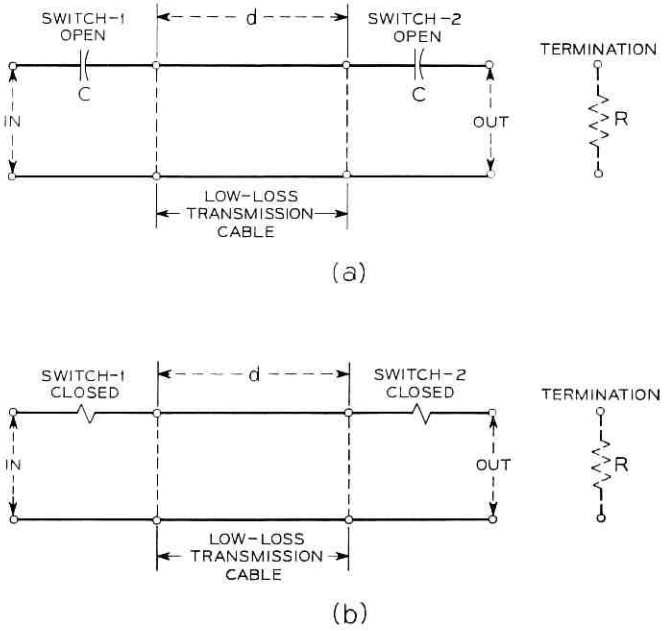


Fig. 4 — Equivalent circuit for the cable switch: (a) open, (b) closed.

specific case of two switching elements (Fig. 3) shows that greater than one and one-half times the isolation loss in dB of a single switching element can generally be realized with a length of cable of only 0.005 wavelength. In addition, for short lengths of transmission cable the increase in isolation loss of the cable switch over that of a single switching element is relatively independent of frequency. This results in an extremely broad frequency bandwidth of operation.

As equation (2) indicates, a further increase in isolation loss can be obtained by adding more cable sections and switching elements to the structure. This, of course, increases the insertion loss as well as the physical length of the cable switch. Alternately, a choice of Z_0 less than the system impedance will result in a further increase in isolation loss. However, in practice Z_0 is chosen equal to the system impedance in order to avoid an impedance mismatch between the switch input and the termination when the switching elements are closed.

Figure 4 shows a schematic representation of the open and closed

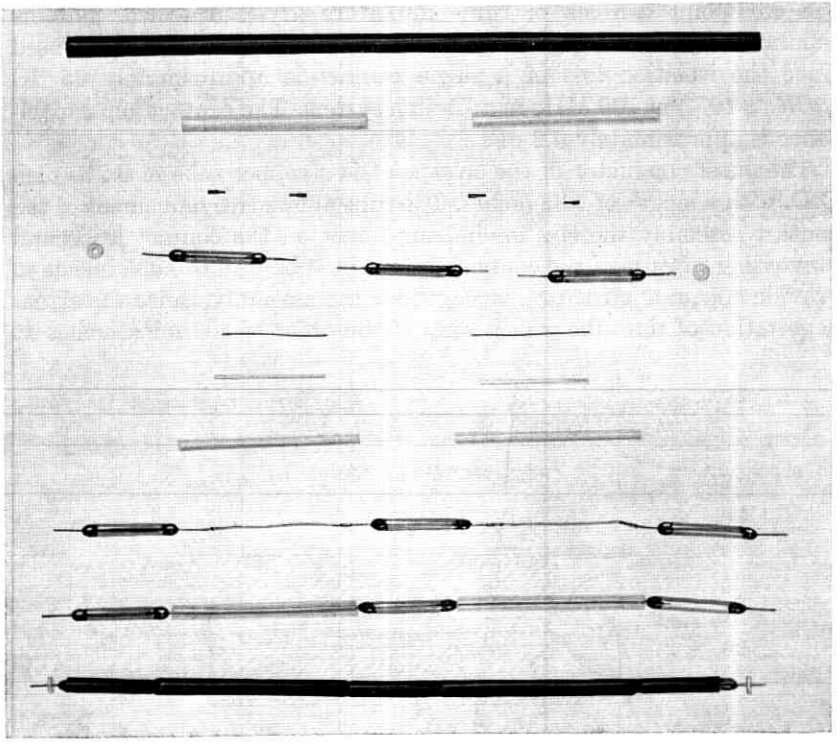


Fig. 5. — Crosspoint elements and assembly.

cable switch in which each switching element is approximated by a small closed-contact resistance. For the closed state, the insertion loss of the cable switch is about equal to the combined loss of the individual switching elements since the loss of the short transmission cable is negligible, and the impedance of the cable is generally equal to the system impedance. This factor of two in the insertion loss for a two-element cable switch is the only penalty in achieving the marked improvement in isolation loss desired.

The cable switch concept reduces crosstalk between various signal paths in the switching matrix because whenever the switch structure is effectively shielded, crosstalk will be defined in terms of the isolation loss of the individual crosspoints.

IV. COAXIAL CROSSPOINT DESIGN

The crosspoint assembly and its piece parts are shown in Fig. 5. A cross-sectional view of the assembly is shown in Fig. 6. As can be seen, the crosspoint consists of three miniature dry-reed sealed contacts separated by short lengths of coaxial line. Three contacts are used since the isolation loss of a single contact is approximately 45 dB ($\omega CR \approx 10^{-2}$) at 100 MHz in a 75-ohm system. The contact gap capacitance is approximately 0.2 pF.

The outer conductor of the crosspoint is a copper tube of 0.210 inch O.D. with a length of 7.82 inches. The tube allows free movement of the contact assembly thereby minimizing forces on the contact leads and preventing rupture of the contact seals. The tube wall is 0.005 inches to provide adequate structural strength for the assembly. Since the signal penetration of the tube wall is only of the order of 300 microinches at

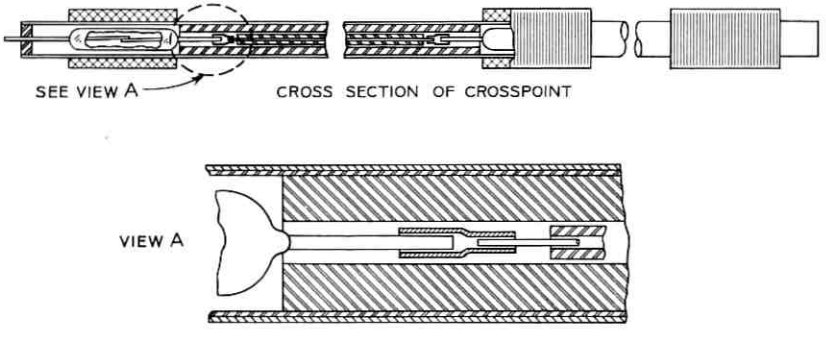


Fig. 6 — Cross-section of crosspoint assembly.

100 MHz, the integrity of the coaxial transmission line is preserved, and there is no crosstalk coupling to the control windings.

The tube length provides approximately three inches between the contact make points with two inches of coaxial line between the glass bottles of the contacts. Since this line segment is loaded with a molded polypropylene sleeve whose dielectric constant is 2.3, the equivalent electrical length is about three inches or 0.03 wavelength at 100 MHz. Equation (2) indicates that the open circuit isolation loss in dB of this three-switch crosspoint approaches 2.3 times that of a single contact.

The diameter of the center conductor and the dielectric material between the contacts proved to be one of the most easily changed variables, and a wide range of diameters and materials were evaluated. The diameter of the copper center conductor that gave the best return loss in the 8 × 8 matrix was found to be 0.010 inches (Z_0 = 120 ohms). The higher impedance (with respect to 75 ohms) of the center conductor section is required to offset the capacitance of the tree networks as discussed in Section 5.1.

The yield strength of the annealed copper conductor between contacts is reached with 1.25-lbs force so that any stresses applied to the crosspoint will be absorbed by the center conductor thereby protecting the contact seal.

In initial switch models, the center conductor was wrapped around

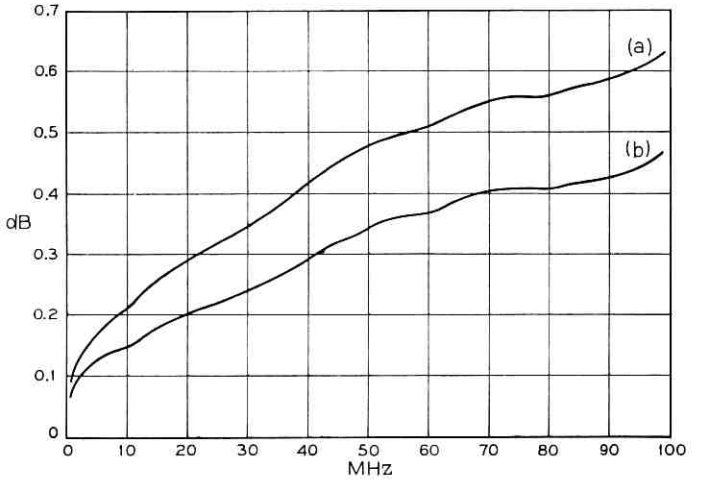


Fig. 7 — Crosspoint insertion loss: (a) with 237B contacts, (b) with 237G contacts.

the contact leads and soldered. Although performance was adequate, experience has indicated the use of the connecting sleeve, Figs. 5 and 6, between the contact and the center conductor facilitates manufacture.

A dielectric end plug fits over the contact lead and into the copper tube. This plug supports the lead during the assembly process. A heat shrinkable polyethelene tube is shrunk over the entire assembly to provide mechanical stability in manufacture.

The insertion loss of the 237B contacts was found to be of the order of 0.15 dB at 100 MHz. This high loss characteristic occurs because only a relatively short length of the contact blade is plated with gold and silver. To provide a continuous surface conductor along the blade with a better conductivity than the nickel-iron blade alloy, a barrel plating process was developed by the Western Electric Company, Allentown Works. These barrel-plated blades are assembled in a recently coded contact, the 237G. The leads in this contact are solder dipped for 0.1 inch to provide easy assembly and a good bond with the

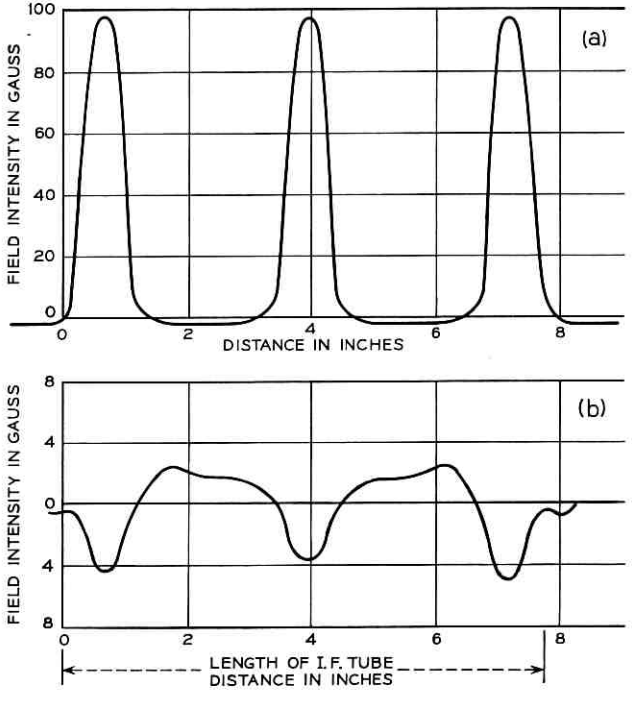


Fig. 8 — Magnetic flux distribution along the axis of a crosspoint: (a) operated, (b) interference from diagonally operated crosspoints in an 8 × 8 array.

sleeve. The insertion loss characteristics of the cable switch when made from the standard 237B contact and the new 237G contact are contrasted in Fig. 7.

Each contact is driven by a 300 turn, 5 layer, 32 gage coil. The coils of a given crosspoint are series connected giving an overall resistance of

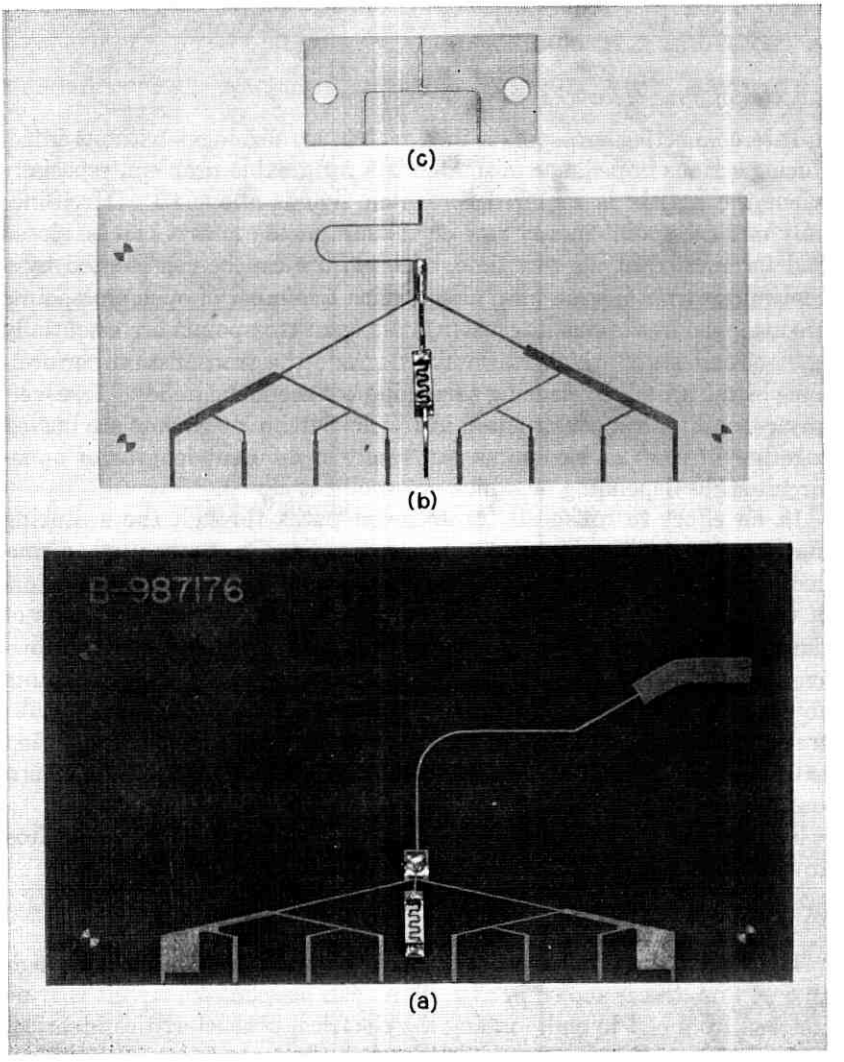


Fig. 9 — Tree networks: (a) 8 × 8, (b) 1 × 8, (c) 1 × 2.

about 11 ohms. Figure 8a shows the flux distribution along the axis of an operated crosspoint. Figure 8b shows the interfering flux from diagonal crosspoints in the 8 × 8 array. Current reversal in alternate columns of the array provides an interfering flux cancellation so the resulting flux level is well below the contact operate levels when driven at the design level of 4.5 ± 1 volt.

V. CROSSPOINT INTERCONNECTION

5.1 *Use of Tree Networks*

Interconnecting crosspoints in a matrix arrangement without introducing serious impedance mismatch is a formidable problem, especially when the matrix is required to switch signals above 10 MHz. Since only one crosspoint in any row or column on the matrix can be closed and terminated at a given time, each closure can be represented by a continuous transmission path along which a number of open crosspoints are attached at various intervals. These open crosspoints are essentially open-circuit stubs which can easily degrade the transmission performance of the switch by causing severe impedance mismatches. Moreover, any general matrix construction in which switching elements are bussed together in rows and columns can result in different lengths of open-circuit stubs depending on which crosspoint is closed.

In an effort to make all transmission paths through the switching matrix appear alike electrically, tree networks, Fig. 1, are used to connect groups of crosspoints to the various input and output connectors of the matrix. Through use of tree networks the rows and columns of the matrix are formed so that no long open-circuited stubs exist. However, short stubs still exist at the positions where the open crosspoints are connected to the closed transmission path. Open-circuit stubs are equivalent to small capacitors short circuiting the closed transmission path and, unless carefully designed, usually preclude meeting the return loss requirement of 28 dB at frequencies above 50 MHz.

The tree networks for the 8 × 8, the 1 × 8 and the 1 × 2 matrices are shown in Fig. 9.

5.2 *Tree Network for* 8 × 8 *Matrix*

The equivalent circuit of a single closed transmission path through an 8 × 8 matrix is shown in Fig. 10. As can be seen, the circuit is symmetrical from end to end, making the electrical characteristics identical for both directions of transmission through the matrix.

The open-circuit stubs represented by short capacitors C_1, C_2, and

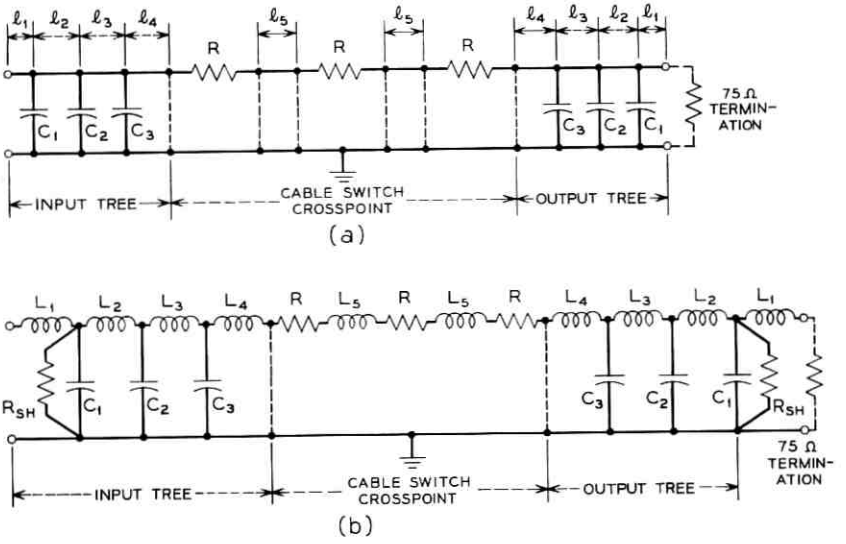


Fig. 10 — Equivalent circuit of a closed connection through an 8 × 8 matrix: (a) untuned circuit (R = contact resistance of the 237G contact; C_1, C_2, C_3 = equivalent capacitance of open-circuited stubs; l_1, l_2, l_3, l_4, l_5 = 75Ω transmission-line lengths), (b) low-frequency approximation of the tuned circuit. (R_{SH} = shunt resistance to adjust real part of input impedance, L_1, L_2, L_3, L_4, L_5 = equivalent inductance of high-impedance transmission lines.)

C_3 in Fig. 10a can be effectively tuned out by

(*i*) adjusting the lengths of l_1, l_2, l_3 and l_4 and increasing their characteristic impedance to a value higher than 75 ohms, and

(*ii*) increasing the impedance of l_5, the two inch coaxial line between the 237G contacts, above 75 ohms.

The low-frequency equivalent circuit of this tuned network is shown in Fig. 10b, where the high-impedance transmission lines are approximated by series inductors.

While this analysis provides physical insight into the factors affecting

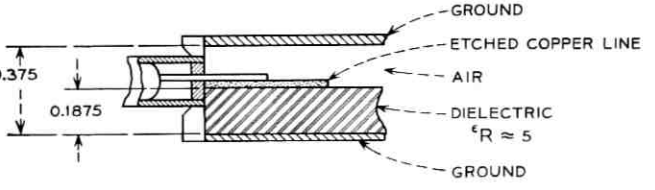


Fig. 11 — Cross-section of tree network for an 8 × 8 matrix.

the design, it cannot provide a solution in closed form to the design problem. The organization of the 8 × 8 matrix was dictated by such factors as minimization of the stub lengths by minimizing the distance between crosspoints, symmetry in the array, interfering flux from adjacent crosspoint coils, and the availability of suitable materials for design.

The 0.375-inch dimension between crosspoints is a reasonable lower limit for mechanical assembly of the crosspoint array since the O.D. of the driving coils is approximately 0.31 inches. Circuit board material standard thickness is $\frac{3}{16}$ inch so that the buildup of the strip line tree circuits is readily accomplished. Flux interference was not a problem as seen in Fig. 8. The question became, then, whether utilizing the insights given by the above circuit analysis, the system design requirements could be met for the adopted physical configurations.

Initial efforts on the tree structure followed conventional stripline technology in which a planar center conductor is positioned between parallel ground planes by dielectric layers. Dielectric materials with a low loss and a uniform dielectric constant over the frequency band

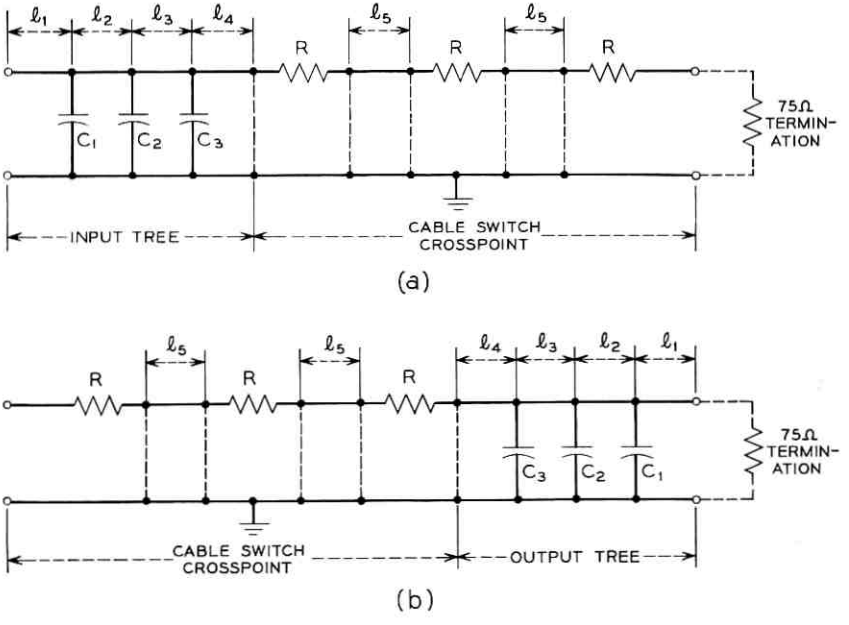


Fig. 12 — Equivalent circuit of a closed connection through a 1 × 8 matrix: (a) 1-to-8 direction, (b) 8-to-1 direction

include the polyolefins and polyphenylene oxides. These are widely used for microwave (300 MHz to 300 GHz) printed circuits. However, the frequency band of concern is well below that for which these more expensive materials are designed, and their generally poorer mechanical stability and peel strength suggested an alternate approach be taken.

Epoxy glass has relatively better mechanical properties and peel strength and is more economical than the above materials. It also has considerably higher loss and dielectric constant.

The compromise solution, Fig. 11, is the use of $\frac{3}{16}$-inch-thick epoxy glass board with one-ounce copper and an air space above the circuit to lower the effective overall dielectric constant. The impedance levels and capacitance of the various branches of the tree were adjusted experimentally with the pattern shown in Fig. 9 resulting.

A shunt resistance, as seen in Fig. 10b, was added at the input and output of the circuit to compensate for the series resistance of the tree and crosspoint. The resistor used, Fig. 9, is the 257A type ceramic with evaporated tantalum nitride film element. It is mounted directly on the board by its leads.

The ground plane spring is formed from a single stamped part of five thousandth-inch beryllium copper over-plated with fifty millionths of hard gold for corrosion resistance and sealing to the end plates and side rails of the switch assembly. The thickness of the material guarantees a minimum of 100 dB crosstalk loss through the circuit boards in the assembled switch. The spring's rolled edges compress when slid into the side rail slots while the front slotted edge seats firmly to the face of the switch. The circuit board assembly is forcibly held in position by the bracket which seats against the rear slotted edge. The result is a well defined geometry insensitive to the temperature ranges to which the switch will be exposed and efficiently sealed against crosstalk.

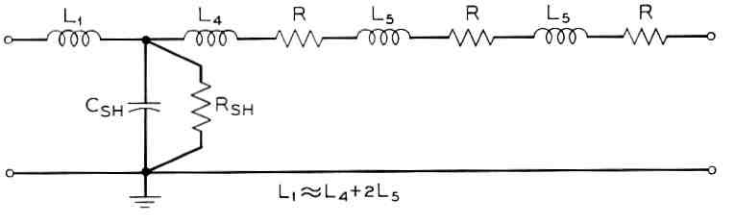


Fig. 13 — Low-frequency approximation of a tuned path through a 1 × 8 matrix. (L_1, L_4, L_5 = equivalent inductance of high impedance transmission lines; C_{SH} = equivalent lumped capacitance of open-circuited stubs; R = contact resistance of the 237G contact; R_{SH} = shunt resistance to adjust real-part of input impedance.)

5.3 *Tree Network for the 1 × 8 Matrix*

The tree network for the 1 × 8 matrix is shown in Fig. 9. The tree differs from that used in the 8 × 8 matrix because the 1 × 8 matrix is physically asymmetric from end to end. In such an array the tree is used to either connect one input to one of eight possible outputs (1-to-8) or to connect one of eight possible inputs to one output (8-to-1). Fig. 12a shows the equivalent circuit of a closed transmission path through the matrix in the 1-to-8 direction, and Fig. 12b shows the equivalent circuit for the 8-to-1 direction. The only possible way to approach a good input impedance match for this matrix for both directions of operation is to:

(*i*) minimize the lengths l_2 and l_3 so that C_1, C_2, and C_3 can be approximated by a single shunt capacitance, C_{SH},
(*ii*) adjust the length of l_1 and the characteristic impedance of l_1, l_4, and l_5 such that the low-frequency equivalent series impedance of l_1 equals $l_4 + 2l_5$, and
(*iii*) add a shunt resistance across C_{SH} to compensate for the series resistance of the tree and the crosspoint.

As can be seen from the low-frequency equivalent circuit of the resulting network, shown in Fig. 13, the input impedance is essentially the same when viewed from either end with the other end terminated in the 75-ohm system impedance.

The circuit board for the 1 × 8 matrix is shown in Fig. 9. In order to achieve the higher stripline impedance, a thinner epoxy glass board is used, and the ground plane spacing is increased as shown in Fig. 14.

5.4 *Tree Networks for the 1 × 2 Matrix*

The tree network for the 1 × 2 matrix is also shown in Fig. 9. The analysis parallels that presented for the 1 × 8 matrix. The simpler

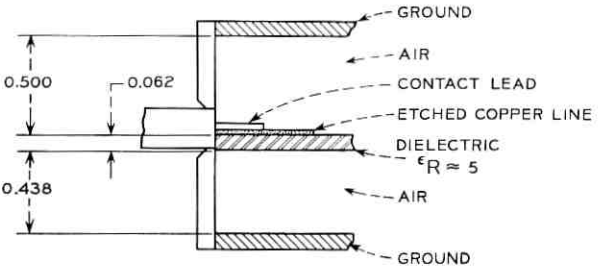


Fig. 14 — Cross-section of tree network for 1 × 8 and 1 × 2 matrices.

nature of the matrix was confirmed by the relative ease of obtaining an experimental solution to the design problems.

VI. CONNECTORS

Standard 478A and 477B jacks accommodating type 728A coaxial cable are used for signal interface. The 478A jack flange was modified to allow closer placement of the input jacks on the 8 × 8 array.

The crosspoint control windings for each crosspoint are individually terminated in standard commerically available connectors allowing wide latitude for control circuit design.

VII. STRUCTURE

Since the structure carries the signal ground, the integrity of the structure at each crosspoint is vital to the maintenance of high return loss. A failure of a tube joint will bring the return loss for that crosspoint to 10 dB even though all other joints are structurally sound. The structure must also be tight to prevent crosstalk. Gaps at the flanges of the connector, between the tree circuits and the tube face, or along the rail at the spring will quickly raise the crosstalk above the minimum required limit of 95 dB below the signal level.

The tube array in each switch code is fixed to the base on one end and pinned on the other to provide axial freedom for thermal expansion. This degree of freedom is sufficient to protect the solder joints between the tubes and the end plates and those between the contact assembly and the circuit boards.

The switch assembly has been vibration tested over the range of 5 to 500 Hz. Resonant points were found for the structure in early models and modifications were made to provide a stiffer structure at those frequencies. Tests of the switch models shown in Fig. 2 led to the inclusion of shipping blocks in the 8 × 8 switch to provide damping of the pinned end of the tube assembly during transport. The switches otherwise withstand anticipated shock and vibration in normal shipping and installation.

The switch has been cycled over the temperature range of 40°F to 140°F at relative humidities to saturation without incident.

VIII. PERFORMANCE

Switch performance has been measured at three Bell Laboratories locations: Merrimack Valley at North Andover, Massachusetts; Holmdel, New Jersey; and Columbus, Ohio. The switch has also been

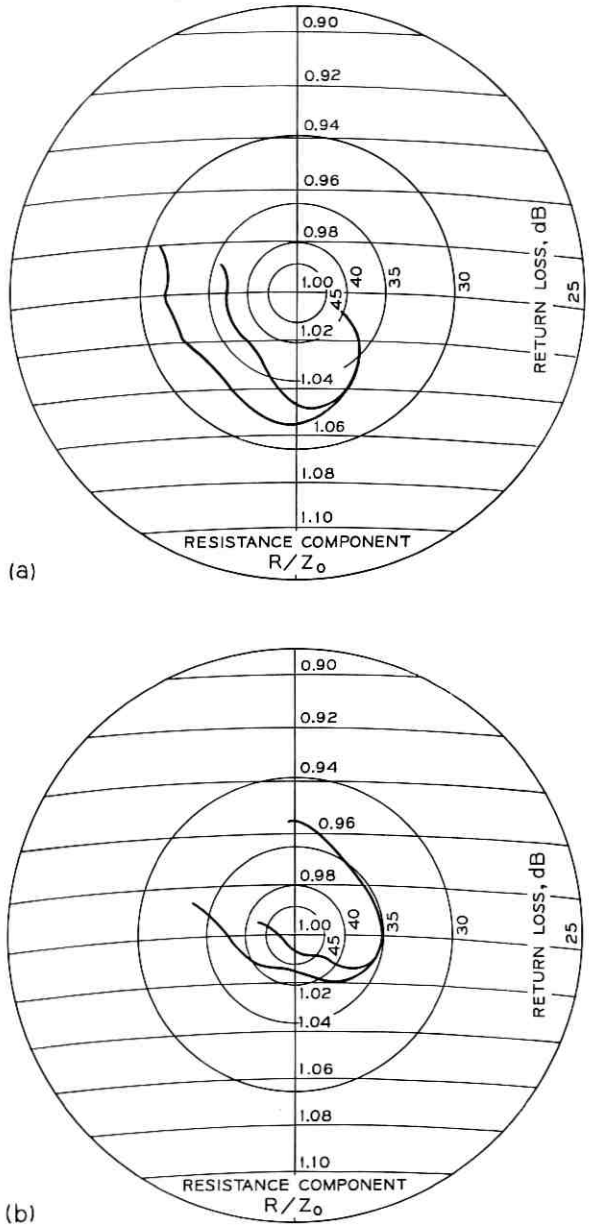


Fig. 15 — Return loss for a single crosspoint in an 8 × 8 array: (a) 8250 Ω resistor, (b) 2740 Ω resistor.

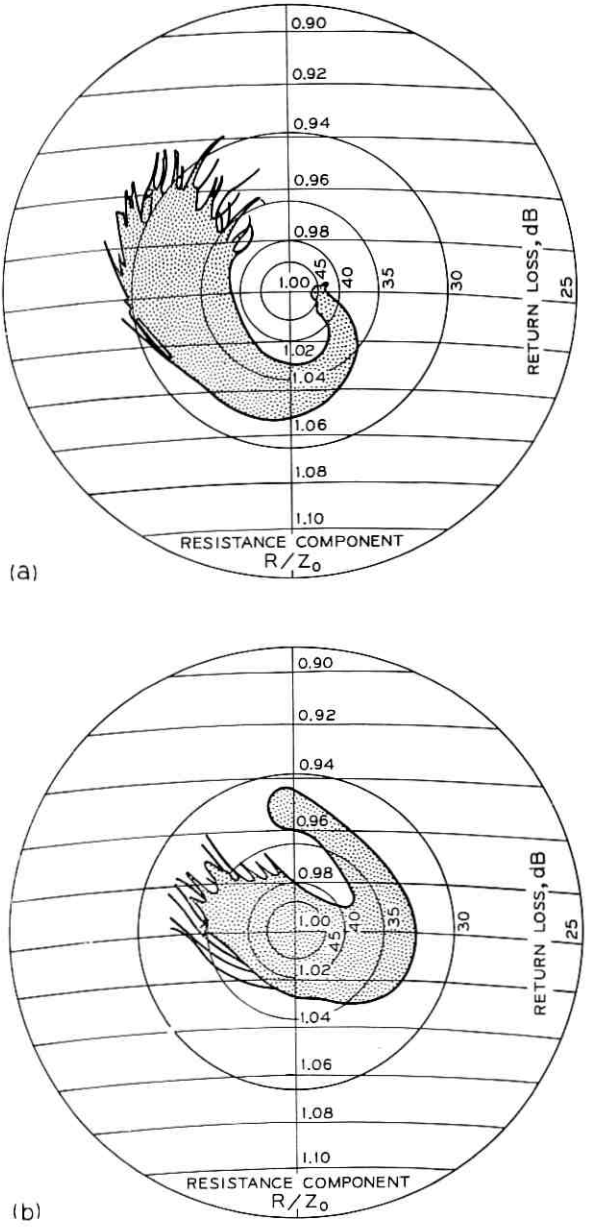


Fig. 16 — Return loss for the 8 × 8 array from dc to 100 MHz: (a) with 8200-ohm shunt resistance, (b) with 2740-ohm shunt resistance.

measured at the Western Electric Works in Kansas City where production has been scheduled.

Return loss results on a given piece of apparatus are reproducible to within 1 dB over the range from 28 to 48 dB. Insertion loss is reproducible to 0.02 dB. 75-ohm attenuators are used in bridge circuits as calibration standards for the measurements program.

8.1 *Return Loss*

The return loss for a single crosspoint of an 8 × 8 array is shown in Figure 15. The presentation is in the form of a Smith chart. The trace is swept from 100 kHz to 100 MHz with the output of the switch terminated in 75 ohms. Two curves are obtained for each crosspoint by testing each end as the input. Since the physical structure is mechanically symmetric, performance should be the same. The electrical asymmetry observed is the result of mechanical tolerances of switch elements with respect to the tuned circuit performance when seen from opposite ends.

The return loss characteristic for two 8 × 8 arrays is shown in Figure 16. The envelope of performance of all sixty-four crosspoints swept

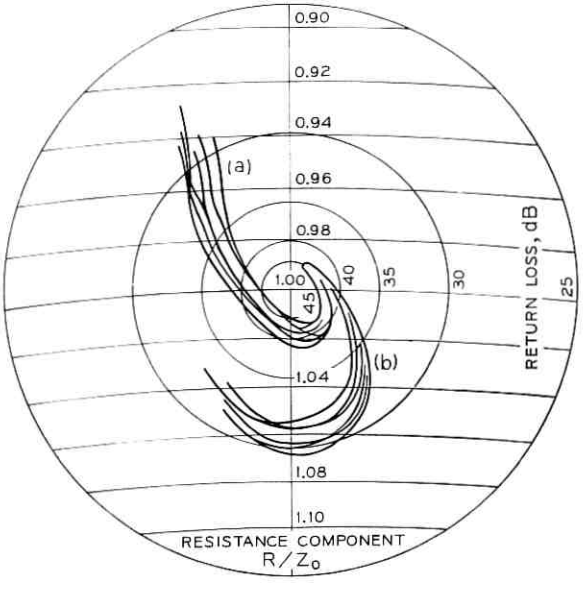


Fig. 17 — Return loss for the 1 × 8 array from dc to 100 MHz: (a) 1-to-8 direction, (b) 8-to-1 direction.

from 100 kHz to 100 MHz from both ends, that is, 128 curves on each graph, is shown. The return loss is obviously better at the lower frequencies where the geometric factors are a smaller fraction of a wave length. No one crosspoint defines the envelope of the plot for more than a fraction of the frequency swept. The scatter at the higher frequencies is a function of both switch geometry and mechanical tolerances. The return loss of the switch with 8200-ohm shunt resistance is seen to be better than 30 dB over most of the band. The improved performance at the higher frequencies obtained with the 2740-ohm resistance is at the expense of the performance at the lower frequencies. In either case the performance limit of 28 dB is met.

The return loss for the 1 × 8 array is shown in Figure 17. The eight traces for either end as input indicates the design compromises required to meet the return loss objective. The switch could be optimized for either input at the expense of the return loss for the opposite end input.

The return loss for the 1 × 2 switch is shown in Figure 18. No shunt resistor was used in this design since it readily meets the requirements for either end as input.

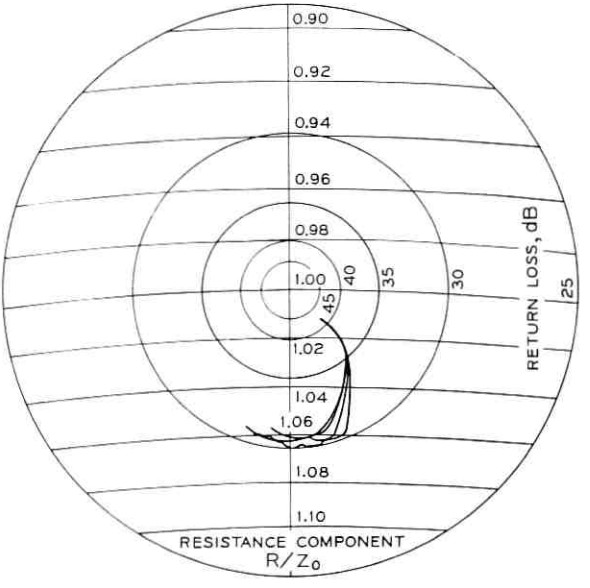


Fig. 18 — Return loss for the 1 × 2 array from dc to 100 MHz in 1-to-2 direction and in 2-to-1 direction.

8.2 *Insertion Loss*

The insertion loss for the 8 × 8 array is shown in Fig. 19. Only the two crosspoints are shown which represent the upper and lower limits of the insertion loss for the switch. Fig. 19b indicates the loss for the switch with shunt resistors of 8200 ohms. Fig. 19a indicates the loss for the switch with 2740-ohm shunt resistors. The higher insertion loss of the second switch is seen to be the penalty for the improved return loss as shown in Figs. 16a and 16b.

Figures 20a and 20b indicate the insertion loss for the 1 × 8 and 1 × 2 arrays respectively.

8.3 *Isolation and Crosstalk Loss*

These losses were found to exceed 105 dB across the band on all codes of switches. The crosstalk loss is at least as good as the isolation loss since the results of crosstalk loss are essentially the same as for isolation loss. There are no significant differences between near end, far end, terminated and unterminated crosstalk observations.

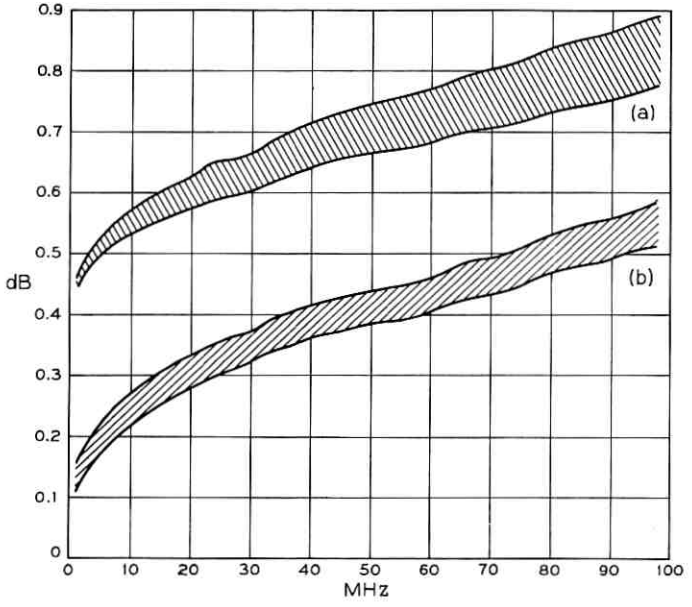


Fig. 19 — Insertion loss for the 8 × 8 array from dc to 100 MHz: (a) with 2740-ohm shunt resistance, (b) with 8200-ohm shunt resistance.

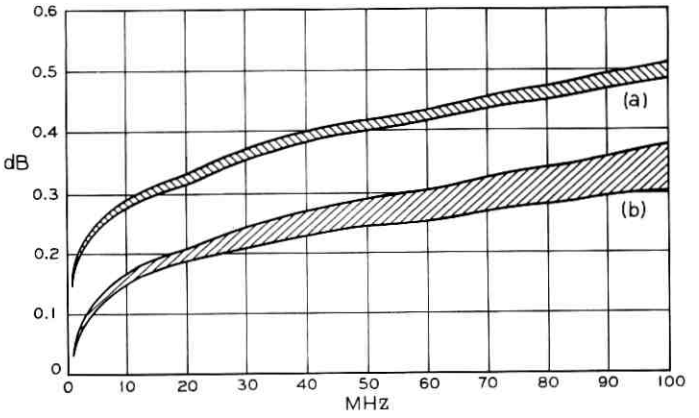


Fig. 20 — Insertion loss from dc to 100 MHz: (a) 1 × 8 array, (b) 1 × 2 array.

IX. SUMMARY

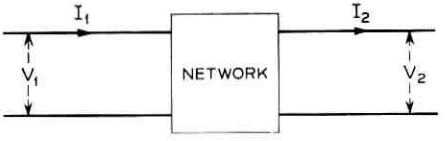
A series of switches designed to meet Bell System requirements across the dc to 100 MHz band have been designed and models built. Tests show the performance characteristics over the band meet system requirements and that the physical structures are mechanically satisfactory for the environmental conditions anticipated in transport and installation.

APPENDIX

Derivation of the V_{out}/V_{in} Relationship for the Cable Switch

A.1 *Network Transfer Matrix*

Any network can be described in terms of an *ABCD* transfer matrix:



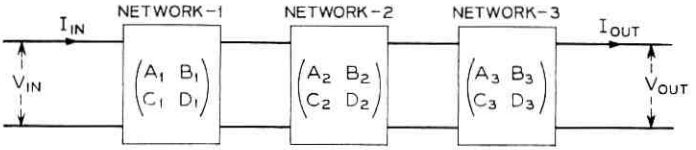
V_1 = input voltage
I_1 = input current
V_2 = output voltage
I_2 = output current

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} V_2 \\ I_2 \end{bmatrix} \tag{3}$$

where $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ is the *ABCD* transfer matrix.

A.2 *Overall Matrix*

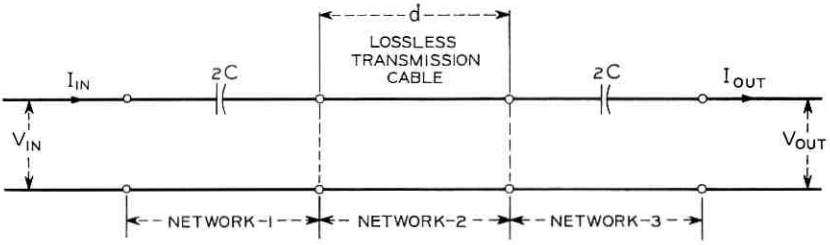
The overall *ABCD* matrix of a series of networks in tandem is equal to the matrix multiplication of the individual network *ABCD* matrices:



$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}_{\text{overall}} = \begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix} \cdot \begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix} \cdot \begin{pmatrix} A_3 & B_3 \\ C_3 & D_3 \end{pmatrix} \cdot \tag{4}$$

A.3 *Equivalent Matrix*

The equivalent circuit for *one* of K identical sections of the cable switch is given by:



Now, the *ABCD* matrices for the 3 networks in this circuit are:

Networks-1 and 3

$$\begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix} = \begin{pmatrix} A_3 & B_3 \\ C_3 & D_3 \end{pmatrix} = \begin{pmatrix} 1 & 1/j\omega 2C \\ 0 & 1 \end{pmatrix} \tag{5}$$

Network-2

$$\begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix} = \begin{pmatrix} \cos \beta d & jZ_0 \sin \beta d \\ j \sin \beta d/Z_0 & \cos \beta d \end{pmatrix} \tag{6}$$

where $\beta = 2\pi/\lambda$
λ = wavelength
d = length of transmission line
Z_0 = characteristic impedance of the transmission line.

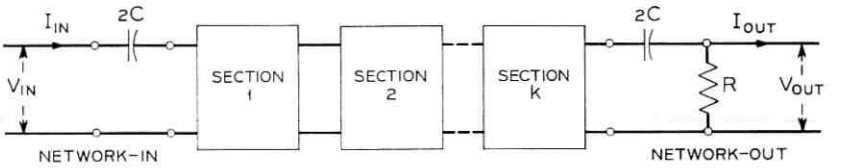
The *ABCD* matrix of one of the K identical sections can be obtained by matrix multiplication:

$$\begin{bmatrix} A_{1s} & B_{1s} \\ C_{1s} & D_{1s} \end{bmatrix} = \begin{bmatrix} \cos \beta d + \dfrac{\sin \beta d}{\omega 2CZ_0} & \dfrac{\cos \beta d}{j\omega C} + \dfrac{\sin \beta d}{j(\omega 2C)^2 Z_0} + jZ_0 \sin \beta d \\ j\dfrac{\sin \beta d}{Z_0} & \cos \beta d + \dfrac{\sin \beta d}{\omega 2CZ_0} \end{bmatrix} \tag{7}$$

Note: $A_{1s} = D_{1s}$.

A.4 *Switch Circuit*

The equivalent circuit for the *entire* cable switch is given by:



The *ABCD* matrices for the networks in this circuit are:

Network-OUT

$$\begin{bmatrix} A_{\text{out}} & B_{\text{out}} \\ C_{\text{out}} & D_{\text{out}} \end{bmatrix} = \begin{bmatrix} 1 + 1/j\omega 2CR & 1/j\omega 2C \\ 1/R & 1 \end{bmatrix} \tag{8}$$

Network of K identical sections

$$\begin{bmatrix} A_K & B_K \\ C_K & D_K \end{bmatrix} = \begin{bmatrix} A_{1s} & B_{1s} \\ C_{1s} & D_{1s} \end{bmatrix}^K \tag{9}$$

Note: $A_K = D_K$.

By matrix multiplication:

n	A_n	B_n	C_n
0	1	$B_{1s}(0)$	$C_{1s}(0)$
1	A_{1s}	$B_{1s}(1)$	$C_{1s}(1)$
2	$2A_{1s}^2 - 1$	$B_{1s}(2A_{1s})$	$C_{1s}(2A_{1s})$
3	$4A_{1s}^3 - 3A_{1s}$	$B_{1s}(4A_{1s}^2 - 1)$	$C_{1s}(4A_{1s}^2 - 1)$
⋮	⋮	⋮	⋮
K	$2A_{1s}A_{K-1} - A_{K-2}$	$B_{1s}\sum_{n=0}^{K-1} A_{1s}^{K-1-n}A_n$	$C_{1s}\sum_{n=0}^{K-1} A_{1s}^{K-1-n}A_n$

(10)

Network-IN

$$\begin{bmatrix} A_{\text{in}} & B_{\text{in}} \\ C_{\text{in}} & D_{\text{in}} \end{bmatrix} = \begin{bmatrix} 1 & 1/j\omega 2C \\ 0 & 1 \end{bmatrix}. \tag{11}$$

The overall $ABCD$ matrix for the cable switch can be obtained by matrix multiplication:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}_{\text{overall}} = \begin{bmatrix} A_K\left(1+\frac{1}{j\omega CR}\right)+\frac{B_K}{R}+\frac{C_K}{j\omega 2C}\left(1+\frac{1}{j\omega 2CR}\right) & \frac{A_K}{j\omega C}+B_K-\frac{C_K}{(\omega 2C)^2} \\ \frac{A_K}{R}+C_K\left(1+\frac{1}{j\omega 2CR}\right) & A_K+\frac{C_K}{j\omega 2C} \end{bmatrix}. \tag{12}$$

A.5 *Special Case*: $I_{\text{out}} = 0$

For the case where $I_{\text{out}} = 0$:

$$\left.\frac{V_{\text{out}}}{V_{\text{in}}}\right|_{I_{\text{out}}=0} = 1/A_{\text{overall}}.$$

Therefore, for the cable switch:

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{1}{A_K\left(1+\frac{1}{j\omega CR}\right)+\frac{B_K}{R}+\frac{C_K}{j\omega 2C}\left(1+\frac{1}{j\omega 2CR}\right)}. \tag{13}$$

Since $\omega CR \ll 1$ for practical values of C and R,

$$\frac{V_{\text{out}}}{V_{\text{in}}} \approx \frac{j\omega CR}{A_K + jB_K\omega C + \frac{C_K}{j\omega 4C}}. \tag{14}$$

Substituting the expressions for B_K and C_K in (10) and the expressions for B_{1s} and C_{1s} in (7) into (14) gives:

$$\frac{V_{\text{out}}}{V_{\text{in}}} = \frac{j\omega CR}{A_K + \left(\cos\beta d + \frac{\sin\beta d}{\omega 2CZ_0} - \omega CZ_0\sin\beta d\right)\sum_{n=0}^{K-1} A_{1s}^{K-1-n}A_n}$$

$$\left|\frac{V_{\text{out}}}{V_{\text{in}}}\right| = \frac{\omega CR}{A_K + (A_{1s} - \omega CZ_0\sin\beta d)\sum_{n=0}^{K-1} A_{1s}^{K-1-n}A_n}. \tag{15}$$

Since $A_{1s} \gg \omega CZ_0 \sin\beta d$ for practical values of C and Z_0,

$$\left|\frac{V_{\text{out}}}{V_{\text{in}}}\right| = \frac{\omega CR}{\sum_{n=0}^{K} A^{K-n}A_n}. \tag{16}$$

REFERENCE

1. Church, D. S., and Kordos, R. W., "Coaxial Cable Switch," United States Patent No. 3,355,684, November 28, 1967.

A Lumped-Circuit Study of Basic Oscillator Behavior

By N. D. KENYON

(Manuscript received August 27, 1969)

This paper presents an experimental study of the oscillations set up in a circuit consisting of a negative conductance and a multiple-resonant load. Its purpose was to verify that such a circuit can account for many of the irregular phenomena commonly observed during tuning of practical microwave solid-state oscillators; such effects as discontinuous frequency changes, low circuit Q-factors, power variations, spurious oscillations and noise conditions are all readily reproduced in a simple low-frequency analogue. There is close correspondence with a first-order analysis.

I. INTRODUCTION

In the course of routine locking-bandwidth measurements on IMPATT diodes in the 50–60 GHz range, some observations were made that could not be accounted for by the simple theory[1-4] of injection locked oscillators. In particular, locking ranges of about 100 MHz for −40 dB injected power were measured, indicating an extremely low effective circuit Q-factor (≈ 5). In addition this range was asymmetrical about the free-running frequency, and was not exactly proportional to the injected voltage. On occasions there was at one end of the locking range a hysteresis between locking and unlocking conditions.

Other phenomena commonly observed during tuning experiments on solid-state oscillators include the following:

(*i*) A discontinuous change in frequency (here referred to as a "jump") as a parameter is varied (bias current, perhaps, or a tuning stub). If the tuning is reversed the jump occurs at a displaced frequency ("hysteresis").

(*ii*) In the neighborhood of a jump, the hitherto single line spectrum may acquire sidebands at displacements of order 0.1 to 1 percent.

(*iii*) Under some circuit conditions a broadband noisy output may be obtained.

To explain these effects the passive circuit "seen" from the diode terminals must be treated as more complex than the simple resonant circuit normally assumed. Kurokawa[5] has analyzed the case in which the active element is as simple as possible, but is connected to a passive load impedance of general form $Z(\omega)$; it is found that most of the observations can be accounted for by ascribing certain patterns to the locus $Z(\omega)$. The conditions are summarized in Section II.

Since the form of $Z(\omega)$ for a packaged diode in a waveguide structure can be very complicated, is difficult to determine precisely, and more difficult still to design, a lumped-circuit approach at low frequency was used to verify the theoretical predictions. The frequency of 350 kHz was chosen so that measurements would be relatively unhindered by stray capacitance effects, but that spectra could be analyzed with sufficient resolution.

II. SUMMARY OF ANALYTICAL PREDICTIONS

When a negative conductance $-\bar{G} + j\bar{B}$ is connected to a load $G + jB$, the voltage amplitude A and frequency ω of the resulting equilibrium oscillation are determined by

$$\bar{G}(A) = G(\omega) \tag{1a}$$

$$-\bar{B}(A) = B(\omega). \tag{1b}$$

These equations define the intersections in the complex impedance plane of the loci $\bar{G}(A) - j\bar{B}(A) = -\bar{Y}$ and $G(\omega) + jB(\omega) = Y(\omega)$, these being referred to as the "device line" and "load line" respectively. For a single-tuned parallel-resonant circuit, $Y(\omega)$ is a straight vertical line; for multiple-tuned circuits it may acquire bends and loops. From Kurokawa's analysis of the latter situation, the following predictions emerge:

(*i*) Let θ be defined as in Fig. 1. Stable oscillation at the point P is only possible if $0 < \theta < \pi$. Moreover, whenever the condition $\theta \to 0$ obtains, the noise of the oscillations will greatly increase.

(*ii*) In Fig. 2, the line PT is the device line or, if that is not straight, PT is the tangent to the device line at P. The point T lies on a line drawn through the frequency points $\omega_0 \pm \Delta\omega$ on the load line. If the resistive component G' of PT satisfies

$$G' = -\tfrac{1}{2}A_0 \frac{\partial \bar{G}}{\partial A}$$

then spurious oscillations will grow at $\omega_0 \pm \Delta\omega$. If spurious oscillations

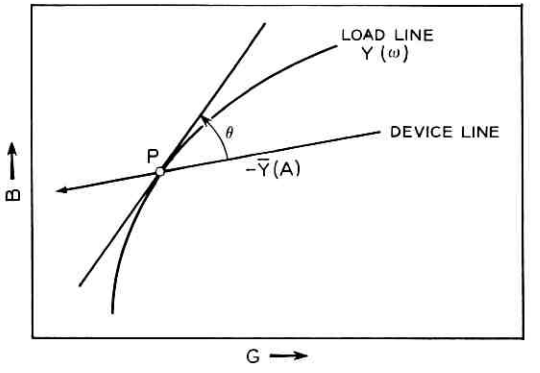


Fig. 1 — Definition of θ at oscillation point.

are not small, the values of ω_0 , A_0 , and so on will be affected, and the first order condition above will not apply.

(iii) Figure 3 shows an injected signal of small amplitude a_0 at a frequency ω_s close to, but not coincident with, the intersection of device and load lines at ω_0 . The perpendicular d is constructed from ω_s to the device line (which is here assumed straight); locking occurs if the length of d is not greater than a_0/A_0 . There are additional requirements for stability of locked oscillation, the principal one being that the angle β be less than $\pi/2$.

(iv) The relationship between injected current, oscillation amplitude and locking range for a near-horizontal device line ($\partial\bar{B}/\partial\bar{G} \simeq 0$) is determined by $| \Delta B | = a_0/A_0$, ΔB being the susceptance change from ω_0 to unlocking frequency. Thus if a_0 and A_0 are held constant the ex-

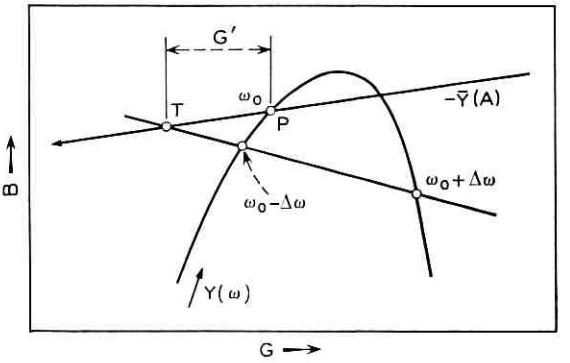


Fig. 2 — Definition of G′ for spurious sidebands.

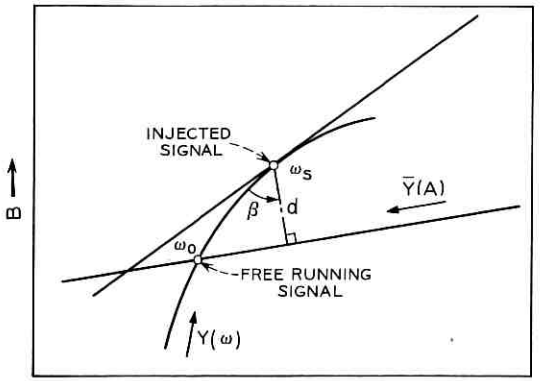


Fig. 3 — Injection locking conditions.

tremities of the locking range will be those for which $B(\omega) - B(\omega_0) = \pm a_0/A_0$.

If the device line is not horizontal, but has a small gradient g, then ΔB must be corrected by the factor $(1 - g\ \Delta G/\Delta B)$.

(*v*) The theory can be applied equally well to the case of a negative impedance $-\bar{R}(A) + j\bar{X}(A)$ simply by reading current for voltage and vice versa, and substituting R, X, Z, and so on, for G, B, Y. The circuit used must then of course have large impedance far from resonance, to inhibit undesired oscillation.

III. ACTIVE ELEMENTS

Both negative conductances and negative resistances were used, taking the circuit forms shown in Fig. 4. The negative conductance is

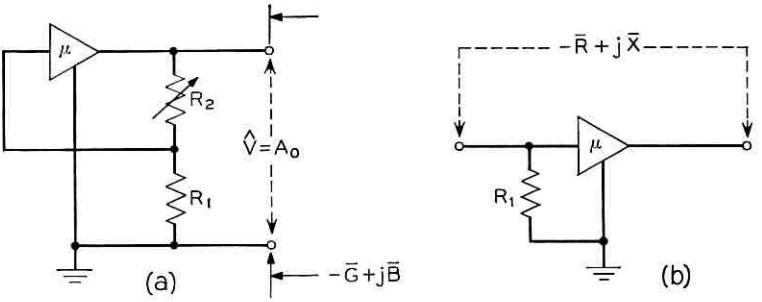


Fig. 4 — Negative admittance and impedance circuits.

readily seen to be

$$\bar{Y} = \frac{-\mu R_1 + (R_1 + R_2 + R_3)}{(R_1 + R_2)R_3}$$

where R_3 is the amplifier output impedance and μ the voltage gain. If there is a phase shift $(2n\pi + \varphi)$ across the amplifier, this equation contains a complex μ:

$$\mu = \mu_0(1 + j\varphi)$$

which introduces an imaginary component $\bar{B}$ into $\bar{Y}$. The amplitude A of oscillation depends on the saturation behavior of the amplifier $\mu(A)$: as A grows, $\mu(A)$ (and therefore $\bar{G}$) declines until the equilibrium condition (1a) is established. At the same time the frequency settles such that the total circuit susceptance is zero (1b).

The theory leading to the predictions of Section II assumes that the device characteristic $\bar{Y}(A)$ is independent of frequency. This is so here if $\mu(A)$ is not frequency dependent and the circuit is free of parasitic capacitance.

The negative impedance of Fig. 4(b) is

$$\bar{Z} \cong -\mu R_1 + R_3 .$$

By making R_1 low, $\bar{Z}$ is confined to a few hundred ohms. The behavior of $\bar{Z}(A)$ is very similar to $\bar{Y}(A)$ above, but the circuit has the disadvantage that no point in the oscillating loop can be grounded; thus large errors arose when injection-locking characteristics were measured. We confine our attention here to the negative admittance circuit.

IV. METHOD OF MEASUREMENT

The amplifier used was a C-Cor 1319F transistor video amplifier, with a small signal gain of 40 dB into 50 ohms, and a bandwidth of 15 MHz. The characteristic $\bar{Y}(A)$ for a given feedback resistor was established by direct measurement, not by calculation from μ, R_1, R_2, R_3.

Figure 5 shows schematically the essentials of the experiment. Apart from the active $\bar{Y}$ and passive $Y(\omega)$ networks, there is an oscilloscope, an injection signal source, a monitoring circuit for the frequency spectrum, and a bridge for passive admittance measurements on $Y(\omega)$.

Monitoring equipment is hung on to a current pick-up lead, which has a negligible loading effect on the circuit. From a counter the frequency is known to 0.01 kHz, oscillations usually being stable to this degree. The wave analyzer determines to the same accuracy the fre-

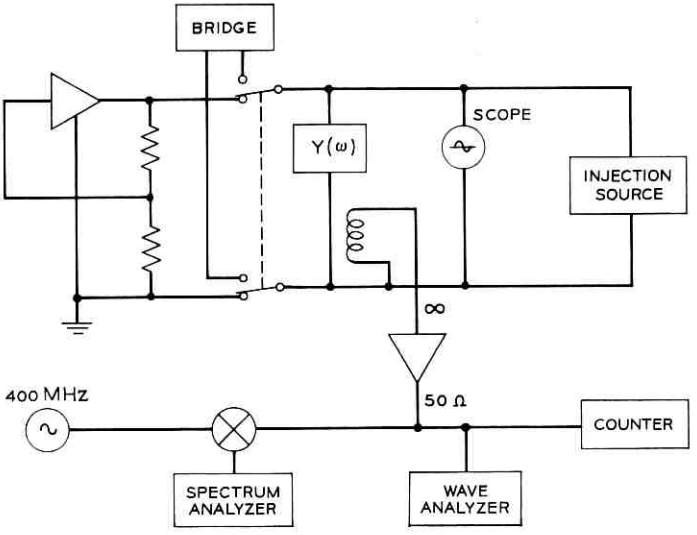


Fig. 5 — Schematic of low-frequency experiment.

quency of sidebands. Finally a convenient display of the spectrum is provided by up-converting into a spectrum analyzer.

To make passive measurements the negative admittance was removed and the RF bridge connected by a short cable to the points shown. The wave analyzer was employed as a very sensitive null detector (to −90 dB).

The arrangement of Fig. 6 was convenient for giving an oscilloscope display of the admittance.

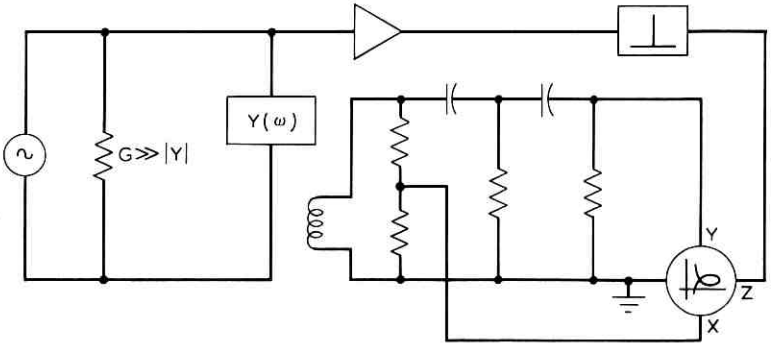


Fig. 6 — Admittance display circuit.

The oscilloscope beam is driven in a circular path of radius proportional to the current through $Y(\omega)$. The spot is brightened by a 2 ns pulse at the instant of maximum voltage. The ac voltage amplitude is approximately constant. The circuit is obviously of limited bandwidth and was not used for quantitative measurements.

V. MEASUREMENTS

5.1 *Characterization*

The device line $-\bar{Y}(A)$ was established from oscillations with a single-tuned parallel-resonant circuit (Fig. 7, inset). Since the imaginary component $\bar{B}$ was small, the frequency was close to the resonant frequency, and the latter was varied by changing L. The oscillation frequency ω_0 and amplitude A were measured, and then the bridge was used to determine the admittance $Y(\omega_0)$: this quantity is plotted di-

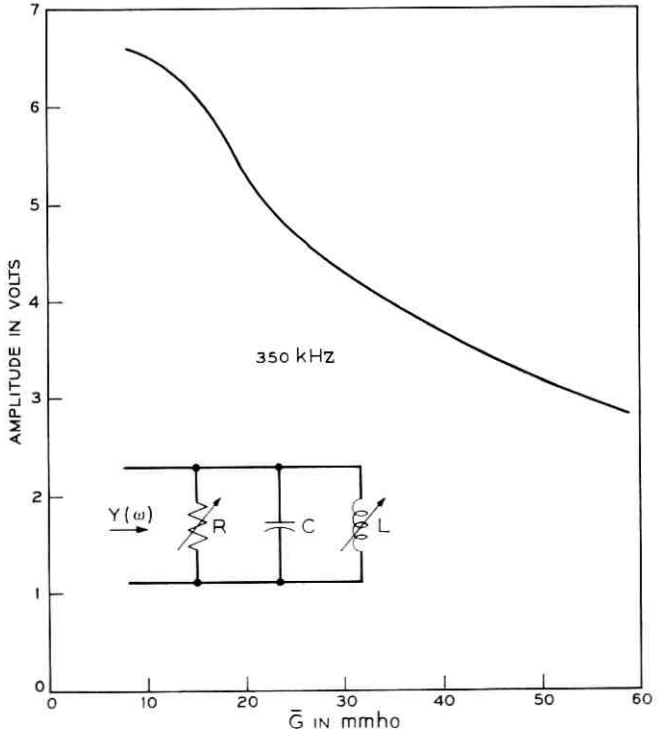


Fig. 7 — Conductance saturation curve at **350 kHz**.

rectly as the device line, since $-\bar{Y}(A) = \bar{Y}(\omega_0)$ at equilibrium. Figure 7 gives the conductance-saturation characteristic at 350 kHz, these amplitude points being then superposed on the device line plot of Fig. 8. The variation of these characteristics over the range 310–390 kHz were found to be quite small. The device line is seen to be neither steep nor sharply curved: thus, the negative admittance is a good approximation to the idealized one assumed in the analysis.[5]

It may be worth noting at this point that tuning of the shunt inductance L is roughly equivalent to a complete vertical displacement of the locus $Y(\omega)$ in the complex admittance plane. The device line in this experiment remains stationary, and the frequency of oscillation and amplitude vary according to the changing point of intersection. However similar phenomenological observations are to be expected if it is the device line that undergoes a vertical displacement—this commonly occurs for bias current changes of solid-state microwave sources. In succeeding paragraphs it is assumed that only the relative vertical relationship of device and load lines is significant, and that this can be changed at will by tuning of L.

5.2 *Double Resonant Circuit*

The addition of a series resonant circuit L_2, C_2, R across the parallel L_1, C_1 provides a number of interesting situations (Figs. 9, 11, 13). The resonant frequency is the same for both circuits.

Defining $Q_1 = \omega_0 C_1 R$, $Q_2 = \omega_0 L_2/R$, and $b = B/R$, we have

$$\partial b/\partial\omega\,|_{\omega_0} = 2C_1R - 2L_2/R = (2/\omega_0)(Q_1 - Q_2).$$

The susceptance of the series circuit tends to compensate that of the original parallel circuit in the neighborhood of resonance, and the combination becomes broader band than either of the two; the effective

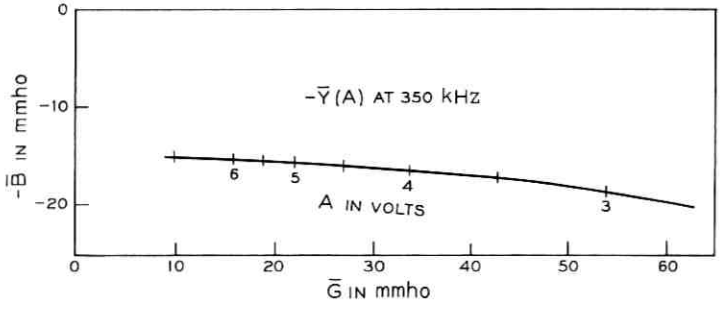


Fig. 8 — Device line plot at 350 kHz.

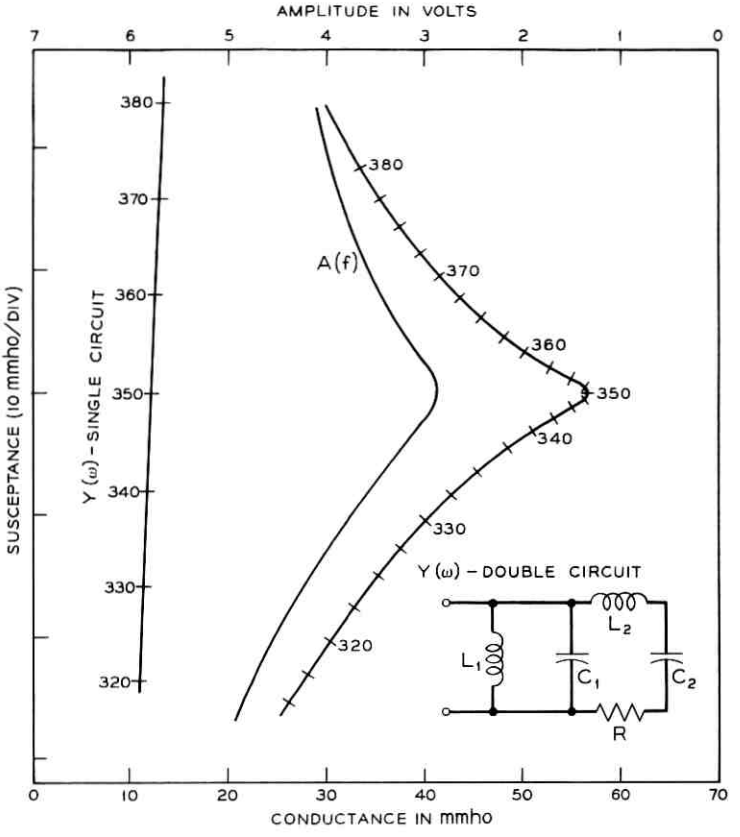


Fig. 9 — Double-tuned circuit behaviour.

Q at ω_0 is the difference $Q_1 - Q_2$, which may obviously be made as low as desired. We shall consider in turn the three cases $Q_1 > Q_2$, $Q_1 = Q_2$, $Q_1 < Q_2$.

5.2.1 *Case 1:* $Q_1 > Q_2$

Figure 9 shows a typical measured $Y(\omega)$ locus for this case. We impose the further condition that at no point is the device line steeper than this load line so that the requirement $0 < \theta < \pi$ (See i in Section II) is met. The measured amplitude at each frequency is also plotted in Fig. 9, and corresponds to the appropriate points of the device characterization, including the general decline with increasing frequency. By comparison with the single-resonant case the frequency points are

relatively crowded at the center of the locus, and there is considerable variation of G in this region. When this circuit is connected to the device, the oscillations generated vary continuously in frequency and amplitude as L_1 is tuned, and there are no unstable points.

The small-signal locking behavior of this kind of circuit was investigated. A variable-frequency source of constant short-circuit current (5 mA peak) injected a driving signal into the terminals of $\bar{Y}$ (See Fig. 6). Oscillation amplitude at 350 kHz was determined, together with the frequencies ω_1 , ω_2 at which the oscillator fell out of synchronism with the driver. Then the negative admittance was removed, and the change of susceptance $2\Delta B$ of the passive circuit between ω_1 and ω_2 was accurately determined (by the addition of known capacitances across the terminals to maintain bridge balance). This procedure was repeated for various $(Q_1 - Q_2)$ by changing R.

The results appearing in Table I show good consistency of ΔB with a/A, except for the last two cases. These have critically low Q and the locking phenomenon is affected by excessive noise and drifting. Figure 10 shows susceptance and frequency ranges as a function of injected current level for a particular value of R and A. The $\Delta B - a$ relationship is linear, as expected, while the locking range falls off at higher levels. Over the linear region the voltage-gain $\times$ fractional-bandwidth product is constant, as for a single-tuned circuit, namely,

$$g \times b = 2/(Q_1 - Q_2).$$

We have thus confirmed that the locking range of an oscillator for a specific circuit is $\omega_1 - \omega_2$, where

$$| B(\omega_1) - B(\omega_2) | = 2a/A.$$

[For microwave circuits the right-hand side would be better expressed

TABLE I

$2\Delta f$ (kHz)	A (Volts)	a/A (mmho)	ΔB (mmho)	
1.95	5.8	0.85	0.82	
2.6	5.3	0.94	0.96	
3.6	4.9	1.02	1.05	
5.5	4.7	1.08	1.08	
7.1	4.6	1.08	1.10	
9.7	4.5	1.12	1.06	
16.6	4.5	1.11	0.99	$(Q_1 = Q_2)$
19.0	4.5	1.11	0.94	$(Q_1 < Q_2)$

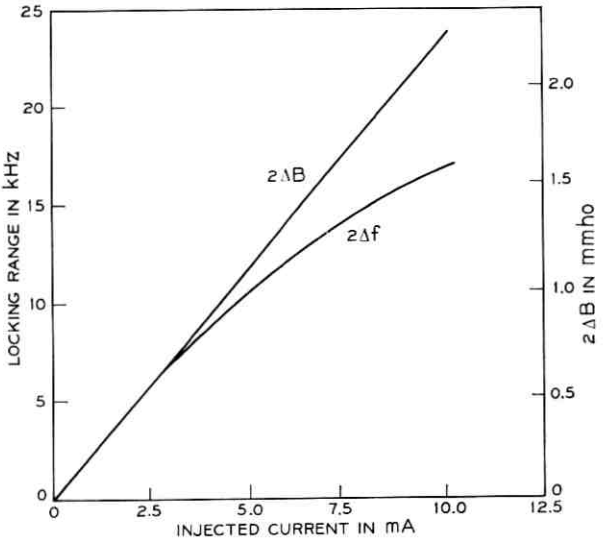


Fig. 10 — Locking range vs. driving signal in broadband circuit.

as $(4/R)(P_i/P_0)^{\frac{1}{2}}$, where P_0 is the output power and P_i the available driver power.] Moreover the combination of two tuned circuits of similiar Q can give a very low effective Q and correspondingly large locking gain-bandwidth product.

5.2.2 *Case 2:* $Q_1 \cong Q_2$

Under this condition a cusp appears on the admittance locus: Fig. 11(a) compares this with the $Q_2 = 0$ case (—markers are at 5 kHz

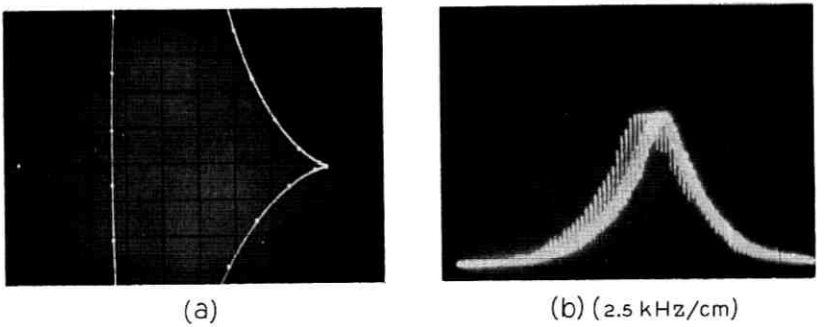
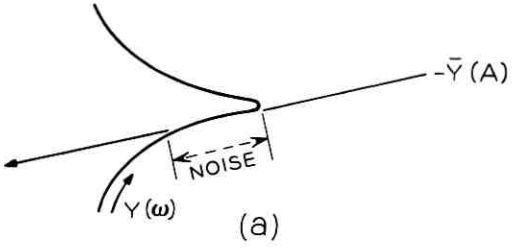


Fig. 11 — Circuit admittance for $Q_2 = 0$ and $Q_2 = Q_1$ cases.

intervals). Oscillations right at the point of the cusp are very noisy (Fig. 11b). When Q_1 is slightly greater than Q_2, the load line may be parallel to the device line ($\theta = 0$) over a considerable band, and the extremely noisy broadband output results (Fig. 12).

5.2.3 *Case 3:* $Q_2 > Q_1$

A loop appears in the admittance locus (Fig. 13a). Part *i* in Section II states that if the device line has the slope indicated by the dotted line, the region between the points A and C (at which the device line would be tangent) has $0 > \theta > -\pi$ and is therefore unstable. As the



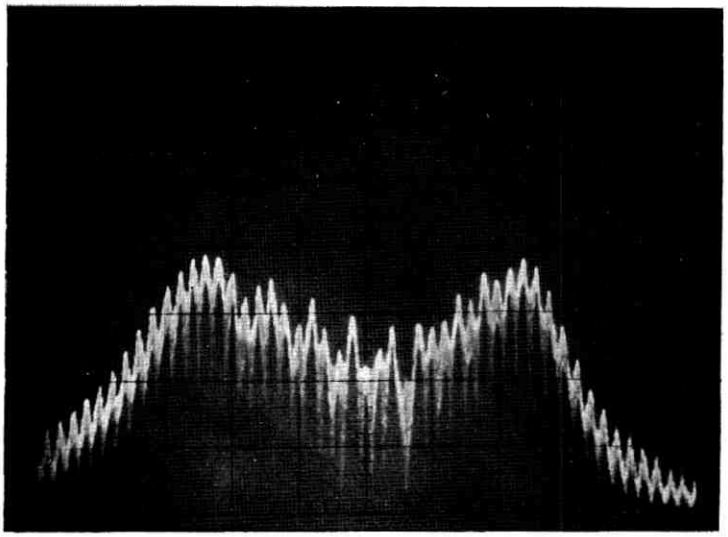


Fig. 12 — Broadband noise conditions.

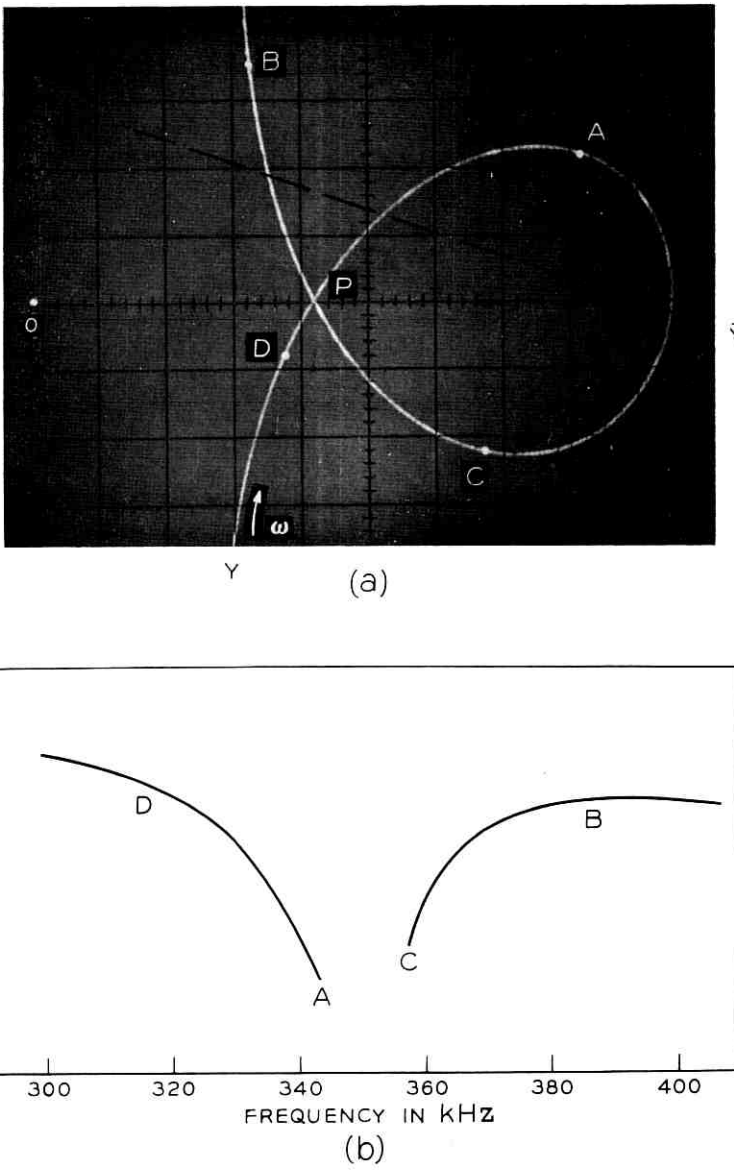


Fig. 13 — Admittance loop for $Q_2 > Q_1$.

circuit is tuned (by decreasing L_1) the frequency moves through D to A, at which it jumps to B; tuning the other way, the frequency moves down to C, then jumps further down still to D. Fig. 13(b) shows the behavior of the amplitude. There are two regions of overlap DA, BC in which oscillation is entirely stable at either of two points, depending on the way in which this oscillation was set up. By accurate measurements of the slope of $Y(\omega)$ at the jump frequencies, and the slope $\partial\bar{B}/\partial\bar{G}$ of the device line at the appropriate values of A and f, it was established that the jumps AB and CD occurred very close to the points of tangency of the device line to the $Y(\omega)$ locus.

If a large parallel conductance G_p is added to $Y(\omega)$, all oscillation ceases (Fig. 14). Reduction of G_p is now equivalent to a leftward displacement of $Y(\omega)$; at the dotted position oscillation of low amplitude is initiated. Clearly this oscillation can never begin at any point on the loop, except the cross-over point P. If a low-amplitude signal is initiated at P, it is possible for the spectrum to contain both frequencies, though usually one will predominate.

5.3 *Spurious Oscillations*

The condition of part *ii* in Section II requires that the line joining points $\omega_0 \pm \Delta\omega$ should intersect the tangent to the device line through ω_0 at a point to the *left* of A_0. Considering Fig. 15(a) (shaded part unstable) we see that for the double-resonant circuit such intersections invariably occur to the *right*. In practice, noisy sidebands were only observed in such a circuit when the device-line had been given a sharp curvature.

Figure 15(b) shows that the sideband condition will be fulfilled if

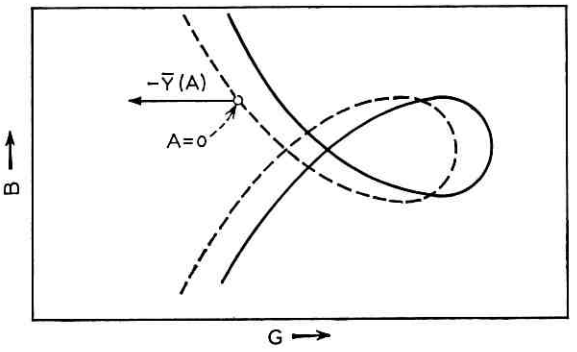


Fig. 14 — Signal initiation.

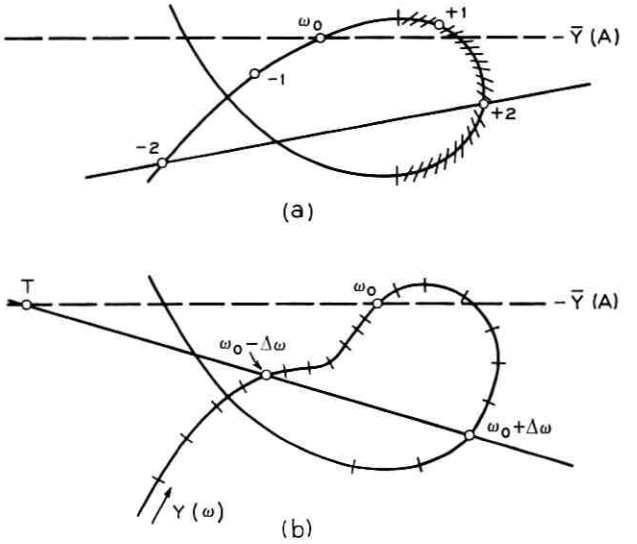


Fig. 15 — Spurious sideband conditions.

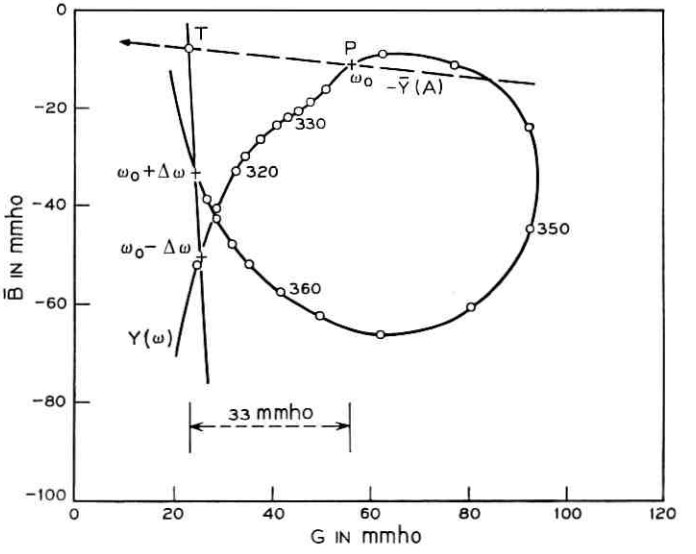
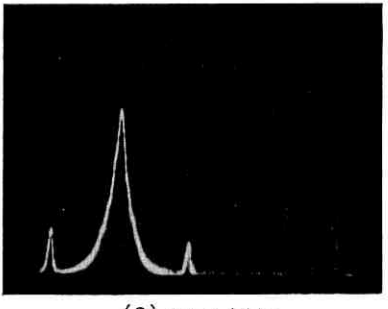


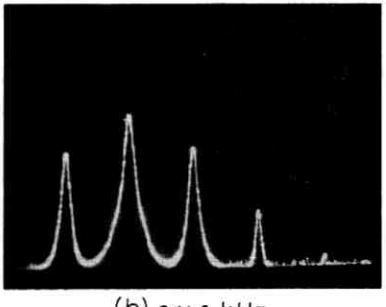
Fig. 16 — Sideband measurements.

the frequency points to the left of ω_0 can be crowded together. This was achieved by the addition of another series resonant circuit of intermediate Q and tuned to about 330 kHz. It was then very easy to get spurious sidebands over an appreciable range of frequencies. A typical case is shown in Fig. 16. As L_1 is reduced the frequency climbs to 343.5 and jumps to 380; between 340 and 343.5 there are noise sidebands, beginning at 340 with very small amplitude, progressing through greater amplitudes and additional sidebands at $\omega_0 \pm n\Delta\omega$; before the jump, the first sidebands were almost the same amplitude as the principal oscillation, and interstitial components at $\omega_0 \pm n\Delta\omega/2$ and so on also appeared (Fig. 17).

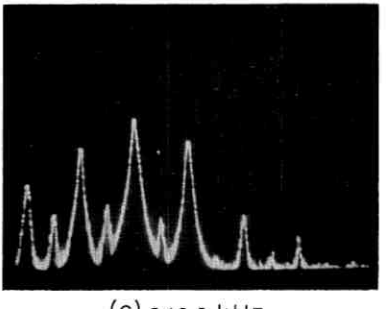
Accurate measurements were made in the case of very small sidebands, for which the small signal theory may be expected to apply. In Fig. 16 the admittance $Y(\omega)$ is shown accurate to within 1 mmho (relatively) on each scale. The line through $\omega_0 \pm \Delta\omega$ is thus subject to little uncertainty. The measured device-line slope at ω_0 gives an



(a) 339.7 kHz



(b) 341.5 kHz



(c) 343.5 kHz

Fig. 17 — Oscillation spectra.

intersection at T as shown, and PT $\cong$ 33 mmho; the measured value of $(-\frac{1}{2}A_0\ \partial\bar{G}/\partial A)$ was 39 mmho, indicating that the intersection should have been more to the left. Other similar measurements gave discrepancies of the same order and of both signs. Obviously only a small error in the amplitude or device slope measurements will move the calculated intersection considerably. Moreover it is not known how the weak function $\bar{Y}(\omega)$ alters the theoretical sideband condition.

5.4 *Further Observations*

Hysteresis between locking-in and unlocking frequencies was observed even with double-tuned circuits, usually where the load-line is most inclined to the vertical. The effect is ascribed to the second-order effect of amplitude variations, and is such that more power is required to pull-in an unlocked oscillator than to maintain locking at the same frequency.

Low-frequency switching between two states was found for some critical adjustments of a triple-tuned circuit in which two loop-type resonances interfere. The time-constant was associated with that of build up and decay of oscillations in the circuit.

The oscillator could be pulled by a small injected signal, when the frequencies were widely different but the circuit admittances similar, as by the cross-over of a loop.

Where spurious sidebands were present, locking of the whole pattern occurred whenever the driver was close to any of the sidebands.

VI. CONCLUSION

Measurements on a low-frequency lumped circuit model have substantiated earlier theory[5] concerning the interaction of a negative admittance and a multi-resonant circuit. The model described is not capable of simulating devices whose characteristics are more complicated than the simple form $\bar{Y}(A)$. For example, it is not easy to duplicate with lumped elements the observations which have been made on microwave oscillators, involving generation of the *same* frequency at two different diode bias currents, or utilizing second-harmonic tuning or sub-harmonic pumping.

However the experiment shows that many of the complex phenomena associated with tuning of solid-state microwave oscillators may be reproduced under these simplified conditions, and that therefore it is very frequently the microwave circuit, rather than the device, which is at fault. In addition, the practicability of broad-banding circuits for deviator and locking-amplifier applications has been demonstrated.

VII. ACKNOWLEDGMENT

The author wishes to acknowledge the advice and encouragement of K. Kurokawa in this work.

REFERENCES

1. Van der Pol, B., "Forced Oscillations in a Circuit with Nonlinear Resistance," *Phil. Mag., 3* (1927), pp. 65–80.
2. Minorsky, N., *Nonlinear Oscillations,* Princeton, N. J.: D. Van Nostrand Co., Inc., 1962.
3. Khokhlov, R. V., "A Method of Analysis in the Theory of Sinusoidal Self-Oscillations," *IRE Trans. Circuit Theory, CT-7,* No. 4 (December 1960), p. 398.
4. Adler, R., "A Study of Locking Phenomena in Oscillators," *Proc. IRE, 34,* No. 6 (June 1946), p. 351.
5. Kurokawa, K., "Some Basic Characteristics of Broadband Negative Resistance Oscillator Circuits," *B.S.T.J., 48,* No. 6 (July 1969), pp. 1937–1955.

Radiation Losses of Tapered Dielectric Slab Waveguides

By DIETRICH MARCUSE

(Manuscript received November 10, 1969)

In this paper we calculate radiation losses of a single mode dielectric slab waveguide for TE and TM modes. The theory is based on the determination of the radiation losses of one abrupt step. We obtain the losses of arbitrarily deformed waveguides by regarding the arbitrary deformations as a succession of infinitely many infinitesimal steps. This method yields the same results as a very different method presented earlier. It allows us to calculate the losses of TM modes that were hard to obtain by the earlier method.

The radiation losses of single mode slab waveguides with abrupt steps of a 2:1 ratio are surprisingly low and can be kept below 1 percent by dimensioning the guide properly. The loss advantage of linear tapers becomes noticeable only when the tapers are very long. An optimized taper changes more rapidly in its wider portion and becomes more gradual in its narrow part.

I. INTRODUCTION

The study of radiation losses of dielectric waveguides, which has been described in three earlier papers,[1-3] has been extended to cover abrupt steps in a single mode waveguide as well as continuous tapers. The mathematical theory of radiation losses caused by a step in the waveguide is used to compute the losses caused by tapers by regarding the taper as a succession of infinitely many infinitesimal steps. This method can also be used to rederive the equations for a dielectric slab waveguide with small wall distortions presented earlier.[1] Both the earlier method and the derivation based on small steps lead to identical results. The perturbation theory used in Ref. 1 was not very well suited for calculating the losses of TM modes. The step method is equally applicable to TM and TE modes and allows us to derive for TM modes

the corresponding expressions which for TE modes were presented in Ref. 1.

The radiation losses of steps and tapers are surprisingly small. A step which changes the thickness of a dominant mode slab waveguide to one half of its original value causes a loss of only about 1 percent for TE modes and about 2 percent for TM modes if operated at favorable frequencies. The losses of tapers are even smaller and can be made as small as desired for sufficiently long tapers.

Comparison of the radiation losses of slab waveguides with round and rectangular waveguides (to be published) shows that the slab waveguide losses are exceptionally low. The losses caused by steps in circular waveguides are higher by an order of magnitude.

II. THE MODES OF THE SLAB WAVEGUIDE

We state briefly the TE and TM modes of the dielectric slab waveguide. For simplicity we assume that all the fields are independent of one spatial coordinate so that we can write symbolically

$$\frac{\partial}{\partial y} = 0. \tag{1}$$

Incidentally, it is only because we limit the discussion to cases where equation (1) applies that it is possible to speak of transverse electric (TE) and transverse magnetic (TM) modes. In the general case the modes are hybrids and possess longitudinal E as well as H components. The modes of the dielectric slab waveguide consist of a finite set of guided modes and a continuum of radiation modes. The slab geometry is shown in Fig. 1.

2.1 *TE Modes*

The field components E_x, E_z and H_y vanish. The remaining components of the magnetic field can be obtained from E_y

$$H_x = -\frac{i}{\omega\mu}\frac{\partial E_y}{\partial z} = -\frac{\beta}{\omega\mu}E_y \tag{2}$$

$$H_z = \frac{i}{\omega\mu}\frac{\partial E_y}{\partial x}. \tag{3}$$

The dependence of the field component on the length coordinate z and on the time t is given by

$$e^{i(\omega t-\beta z)}. \tag{4}$$

This factor will be omitted from the following equations.

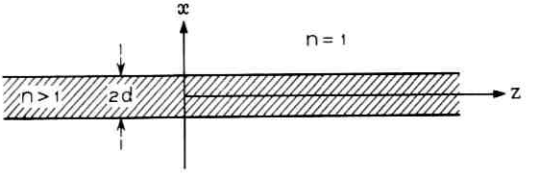


Fig. 1 — Dielectric slab waveguide.

2.1.1 *Even Guided Modes*

$$\left.\begin{aligned} E_y &= A_e \cos \kappa x && |\, x \,| \leqq d \\ E_y &= A_e e^{\gamma d} \cos \kappa d e^{-\gamma |x|} && |\, x \,| \geqq d \end{aligned}\right\}. \tag{5}$$

The coefficient A_e is related to the power P carried by the mode by the following equation

$$A_e = \left\{\frac{2\omega\mu_0 P}{\beta_0 d + \dfrac{\beta_0}{\gamma}}\right\}^{\frac{1}{2}}. \tag{6}$$

The relation between κ, γ and β_0 is given by

$$\kappa = [(nk)^2 - \beta_0^2]^{\frac{1}{2}}, \tag{7a}$$

$$\gamma = [\beta_0^2 - k^2]^{\frac{1}{2}}, \tag{7b}$$

$$k = \omega(\epsilon_0\mu_0)^{\frac{1}{2}}. \tag{8}$$

n is the index of refraction of the dielectric slab. The index of the surrounding medium is taken to be $n = 1$. The eigenvalue equation for the determination of β_0 is

$$\tan \kappa d = \frac{\gamma}{\kappa}. \tag{9}$$

A few numerical values for β_0 are shown in Table I. The TE modes are power orthogonal. With the power flow P in z-direction (per unit length of y) we have

$$P\delta_{nm} = \frac{\beta_n}{\omega\mu} \int_0^\infty E_{yn} E_{ym}^* \, dx. \tag{10}$$

2.1.2 *Even Radiation Modes*

$$\left.\begin{aligned} E_y &= B_e \cos \sigma x && |\, x \,| \leqq d \\ E_y &= C_e e^{i\rho|x|} + C_e^* e^{-i\rho|x|} && |\, x \,| \geqq d \end{aligned}\right\}. \tag{11}$$

TABLE I—SOME NUMERICAL VALUES OF β_0

kd	n	TE Mode $\beta_0 d$	TM Mode $\beta_0 d$
2.5	1.01	2.50271	2.50263
5.0		5.01550	5.01519
10.0		10.06061	10.06016
20.0		20.16711	20.16680
0.25	1.432	0.25781	0.25207
0.5		0.54916	0.51677
1.0		1.21972	1.12809
1.5		1.93825	1.84210
2.0		2.66839	2.58934
3.0		4.13075	4.08131

Propagation constants of TE and TM modes

(The asterisk indicates the complex conjugate value) with

$$\sigma = [(nk)^2 - \beta^2]^{\frac{1}{2}}, \tag{12}$$

$$\rho = [k^2 - \beta^2]^{\frac{1}{2}}, \tag{13}$$

$$C_e = \tfrac{1}{2}B_e \exp(-i\rho d)\left(\cos \sigma d + i\frac{\sigma}{\rho}\sin \sigma d\right), \tag{14}$$

$$B_e = \left\{\frac{2\rho^2\omega\mu P}{\pi\beta(\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 \sigma d)}\right\}^{\frac{1}{2}}. \tag{15}$$

The power orthogonality of the radiation modes can be expressed by the equation

$$P\delta(\rho - \rho') = \frac{\beta}{\omega\mu}\int_0^\infty E_y(x, \rho)E_y^*(x, \rho')\,dx. \tag{16}$$

P is the power flowing per unit length (in y-direction) in the z-direction.

The odd TE modes have been listed in Ref. 1 (together with the even TE modes). Since we are limiting the discussion of TE modes to symmetrical tapers excited by an even mode we will not need the odd TE modes in this paper.

2.2 *TM Modes*

With the restriction imposed by equation (1) the only nonvanishing components of the TM modes are H_y,

$$E_x = \frac{i}{\omega\epsilon}\frac{\partial H_y}{\partial z}, \tag{17}$$

$$E_z = -\frac{i}{\omega\epsilon}\frac{\partial H_y}{\partial x}. \tag{18}$$

We have no occasion to use the odd guided TM modes, therefore only the even guided modes will be listed.

2.2.1 *Even Guided Modes*

$$\left.\begin{aligned} H_y &= A_e \cos \kappa x && \text{for} \quad |x| \leqq d \\ H_y &= A_e e^{\gamma d} \cos \kappa d e^{-\gamma|x|} && \text{for} \quad |x| \geqq d \end{aligned}\right\}. \tag{19}$$

The amplitude constant is related to the power P carried by the mode

$$A_e = \left\{\frac{\gamma}{\beta_0} \frac{2\omega\epsilon P}{\frac{n^2 k^2}{\beta_0^2 + n^2\gamma^2} + \gamma d}\right\}^{\frac{1}{2}}. \tag{20}$$

The constants κ and β are related to β_0 by equations (7a) and (7b). The eigenvalue β_0 of the even guided TM modes is obtained as a solution of the eigenvalue equation

$$\tan \kappa d = n^2 \frac{\gamma}{\kappa}. \tag{21}$$

A few numerical values for β_0 are shown in Table I. The power orthogonality of the guided TM modes can be expressed by

$$P\delta_{nm} = \frac{\beta_n}{\omega} \int_0^\infty \frac{1}{\epsilon} H_{yn} H_{ym}^* \, dx = \frac{\omega}{\beta_n} \int_0^\infty \epsilon E_{xn} E_{xm}^* \, dx. \tag{22}$$

2.2.2 *Even Radiation Modes*

$$\left.\begin{aligned} H_y &= B_e \cos \sigma x && |x| \leqq d \\ H_y &= C_e e^{i\rho|x|} + C_e^* e^{-i\rho|x|} && |x| \geqq d \end{aligned}\right\} \tag{23}$$

with ρ and γ given by equations (12) and (13) and with

$$C_e = \tfrac{1}{2} B_e \left(\cos \sigma d + \frac{i}{n^2} \frac{\sigma}{\rho} \sin \sigma d\right) e^{-i\rho d}. \tag{24}$$

The amplitude B_e is given by

$$B_e = \rho \left\{\frac{2\omega\epsilon P}{\pi\beta\left(n^2\rho^2 \cos^2 \sigma d + \frac{\sigma^2}{n^2} \sin^2 \sigma d\right)}\right\}^{\frac{1}{2}}. \tag{25}$$

2.2.3 *Odd Radiation Modes*

$$\left.\begin{aligned} H_y &= B_0 \sin \sigma x && \text{for} \quad |x| \leqq d \\ H_y &= \frac{x}{|x|} \{C_0 e^{i\rho|x|} + C_0^* e^{-i\rho|x|}\} && \text{for} \quad |x| \geqq d \end{aligned}\right\} \tag{26}$$

with

$$C_0 = \tfrac{1}{2}B_0 e^{-i\rho d}\left(\sin \sigma d - \frac{i}{n^2}\frac{\sigma}{\rho}\cos \sigma d\right) \tag{27}$$

and

$$B_0 = \rho\left\{\frac{2\omega\epsilon P}{\pi\beta\left(n^2\rho^2 \sin^2 \sigma d + \frac{\sigma^2}{n^2}\cos^2 \sigma d\right)}\right\}^{\frac{1}{2}}. \tag{28}$$

The power orthogonality of the radiation modes is expressed as

$$P\delta_{e0}\,\delta(\rho - \rho') = \frac{\beta}{2\omega}\int_{-\infty}^{\infty}\frac{1}{\epsilon}\,H_{\nu e}(x,\,\rho)H^*_{\nu 0}(x,\,\rho')\,dx. \tag{29}$$

All the modes are orthogonal among each other. The amount of power P carried by each mode is normalized to the same value. The actual power carried by the field is determined by the expansion coefficients.

III. TE MODE RADIATION LOSS

Prior to discussing the radiation losses of a waveguide taper we calculate the losses of an abrupt step in the dielectric slab waveguide. We limit our investigation to the case that only the lowest order guided mode of each type exists. These modes do not experience a cut-off and can exist on waveguides with vanishingly small thickness. The steps are considered to be sufficiently small to keep the guide dimensions below the point where a second guided TE or TM mode becomes possible.

The geometry of the step is shown in Fig. 2. The loss problem is solved by assuming that one guided (TE or TM) mode is incident on the step. The discontinuity in the waveguide causes a reflected mode as well as forward and backward traveling radiation modes to occur. The unknown amplitudes of these modes are determined by requiring

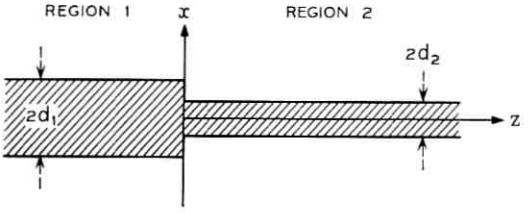


Fig. 2 — Abrupt step in a dielectric slab waveguide.

that the transverse field components are continuous at the step. For TE modes we get the following equations:

$$E_y^{(i)} + a_r E_y^{(r)} + \int_0^\infty q_r(\rho) E_y^{(r)}(\rho)\, d\rho = c_t E_y^{(t)} + \int_0^\infty q_t(\rho) E_y^{(t)}(\rho)\, d\rho, \tag{30}$$

$$H_x^{(i)} + a_r H_x^{(r)} + \int_0^\infty q_r(\rho) H_x^{(r)}(\rho)\, d\rho = c_t H_x^{(t)} + \int_0^\infty q_t(\rho) H_x^{(t)}(\rho)\, d\rho. \tag{31}$$

The superscripts i, r and t indicate incident, reflected and transmitted waves. The field components whose ρ dependence is explicitly shown are radiation modes, the other field components belong to guided modes.

There are two ways to compute the radiation losses. We can calculate the coefficients c_t and a_r of the transmitted and reflected guided mode and calculate the radiated power loss from

$$\frac{\Delta P}{P} = 1 - |c_t|^2 - |a_r|^2 \tag{32}$$

or we can calculate the coefficients q_t and q_r and obtain the radiation losses from

$$\frac{\Delta P}{P} = \int_{-k}^{0} |q_r|^2 \frac{|\beta|}{\rho}\, d\beta + \int_0^k |q_t|^2 \frac{\beta}{\rho}\, d\beta. \tag{33}$$

Both methods should, of course, lead to the same result.

It is impossible to obtain exact solutions of equations (30) and (31); a comparison of both methods (32) and (33) allows an estimate of the validity of the approximations that are used to solve these equations.

We obtain approximate solutions by the following argument. Since all modes of the same waveguide section are orthogonal we can use the orthogonality of the modes to isolate c_t on the right hand side of equations (30) and (31). We get for TE modes from (30)

$$c_t = \frac{\beta_2}{\omega\mu P}(1 + a_r) \int_0^\infty E_y^{(i)} E_y^{(t)*}\, dx \tag{34a}$$

and from equation (31)

$$c_t = \frac{\beta_1}{\omega\mu P}(1 - a_r) \int_0^\infty E_y^{(i)} E_y^{(t)*}\, dx. \tag{34b}$$

The coefficient q_r was neglected. For large steps the radiation is scattered predominantly in forward direction so that q_r is indeed small. If the step height is small the fields $E_y^{(i)}$ and $E_y^{(r)}(\rho)$ become more nearly orthogonal so that q_r again does not contribute very much to equations (34a) and (34b). The propagation constant β_2 belongs to the guided mode on the waveguide to the right of the step while β_1 belongs to the guided mode to the left of the step. Because of the different waveguide size these propagation constants are not the same.

Equations (34a) and (34b) allow the determination of c_t and a_r

$$c_t = \frac{2\beta_1\beta_2}{\beta_1 + \beta_2}\frac{1}{\omega\mu P}\int_0^\infty E_y^{(i)}E_y^{(t)*}\,dx, \tag{35}$$

$$a_r = \frac{\beta_1 - \beta_2}{\beta_1 + \beta_2}. \tag{36}$$

The integral can be evaluated with the help of equation (5) so that we obtain

$$c_t = \frac{4(n^2 - 1)\beta_1\beta_2 k^2 \cos\kappa_2 d_2}{\left[\left(\beta_1 d_1 + \frac{\beta_1}{\gamma_1}\right)\left(\beta_2 d_2 + \frac{\beta_2}{\gamma_2}\right)\right]^{\frac{1}{2}}(\beta_1 + \beta_2)^2(\beta_1 - \beta_2)(\kappa_1^2 + \gamma_2^2)}$$
$$\cdot[\gamma_2 \cos\kappa_1 d_2 - \kappa_1 \sin\kappa_1 d_2 + (\gamma_1 - \gamma_2)\cos\kappa_1 d_1 e^{-\gamma_2(d_1 - d_2)}]. \tag{37}$$

The determination of q_r and q_t is not quite as simple. The functions $E_y^{(r)}(\rho)$ and $E_y^{(t)}(\rho)$ belong to different waveguides and are not orthogonal. For large steps with predominantly forward scattering q_r may again be negligible but this is certainly not true for small steps. We would need different approximations for large and small steps. To avoid this difficulty we consider only small steps and construct large steps and waveguide tapers as a succession of small steps. For infinitesimal steps the modes $E_y^{(r)}$ and $E_y^{(t)}$ are very nearly orthonormal and reflected guided modes can be neglected. Using the orthogonality of the modes we obtain

$$q_t(\rho) = \tfrac{1}{2}(\beta_0 + \beta)I \tag{38}$$

and

$$q_r(\rho) = \tfrac{1}{2}(\beta_0 - \beta)I \tag{39}$$

with

$$I = \frac{1}{\omega\mu P}\int_0^\infty E_y^{(i)}E_y^{(t)*}(\rho)\,dx. \tag{40}$$

The expression I does not depend on the sign of β, we therefore obtain

$q_r(\rho)$ from $q_t(\rho)$ by reversing the sign of the propagation constant β of the radiation mode. We may drop the subscript r and t and obtain after integration

$$q(\rho) = -(n^2 - 1)k^2 \cdot \frac{\rho \cos \kappa d \cos \sigma d \, \Delta d}{(\pi)^{\frac{1}{2}}(\beta_0 - \beta)\left[|\beta| \left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)(\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 \sigma d)\right]^{\frac{1}{2}}}. \quad (41)$$

The difference $\Delta d = d_2 - d_1$ is assumed to be small. Because of the relation between $q_r(\rho)$ and $q_t(\rho)$ we can write equation (33) more simply

$$\frac{\Delta P}{P} = \int_{-k}^{k} |q(\rho)|^2 \frac{|\beta|}{\rho} d\beta. \quad (42)$$

IV. APPLICATION TO TAPERS

Equation (41) can immediately be extended to apply to symmetrical waveguide distortions of arbitrary shape. We assume that the shape of the waveguide wall is described by the function $f(z)$ as shown in Fig. 3. We can then write

$$\Delta d = \frac{df}{dz} dz. \quad (43)$$

The amplitude $q(\rho)$ was calculated for a small step at $z = 0$. Locating the step at z the guided wave arrives there with the phase $e^{-i\beta_0 z}$ instead of with phase zero as assumed in equation (41). The radiation mode was also referred to $z = 0$. Referring it to a step at z adds the phase factor $e^{i\beta z}$ to equation (41) because the amplitude B of the radiation mode enters equation (41) with its complex conjugate value. A step at z would be described by an expression like (41) with an additional phase factor

$$e^{-i(\beta_0 - \beta)z}. \quad (44)$$

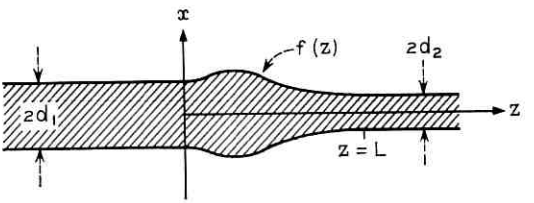


Fig. 3 — A symmetrical wall distortion (symmetrical taper) of a dielectric slab waveguide.

It must be assumed that β_0 (but not β) is a function of z if the guide thickness is changing.

The total radiation loss of a section of waveguide (for example a taper) of length L is given by equation (42) with

$$q(\rho) = -(n^2 - 1)k^2 \cdot \int_0^L \frac{\rho \cos \kappa d \cos \sigma d e^{-i(\beta_0 - \beta)z} \frac{df}{dz}}{(\beta_0 - \beta)\left[\pi \mid \beta \mid \left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)(\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 \sigma d)\right]^{\frac{1}{2}}} dz. \tag{45}$$

Except for the restriction to symmetrical waveguides, equation (45) describes the same problem as treated in Ref. 1. In fact, we can obtain equation (57) of Ref. 1 by a partial integration. The formulation of Ref. 1 applies to the case that the thickness of the waveguide at $z = 0$ and $z = L$ is very nearly the same. The function $f(z)$ deviates so little from the half thickness d of the perfect waveguide that β_0, κ and γ can be assumed to be independent of d. With these assumptions, we obtain as a result of a partial integration

$$q(\rho) = \frac{(n^2 - 1)k^2 \rho \cos \kappa d \cos \sigma d \varphi(\beta)}{i\left[\pi \mid \beta \mid \left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)(\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 \sigma d)\right]^{\frac{1}{2}}} \tag{46}$$

with

$$\varphi(\beta) = \int_0^L f(z) e^{-i(\beta_0 - \beta)z} dz. \tag{47}$$

The agreement with equation (57), Ref. 1, is perfect if we keep in mind that the functions describing the upper and lower side of the waveguide are now identical except for a minus sign and that the function $f(z) - d$ of Ref. 1 is now redefined and replaced by $f(z)$.

The fact that equation (45) is identical to the theory of Ref. 1 proves the validity of our method of continuous steps.

4.1 *TM Mode Radiation Loss*

The radiation losses of the lowest order guided TM mode at a symmetrical step in the dielectric slab waveguide can be calculated from equations (30) and (31) by changing the subscript x to y and y to x.

The c_t coefficient for the lowest order (dominant) even TM mode is

$$c_t = \frac{2I_1 I_2}{I_1 + I_2} \tag{48}$$

and the a_r coefficient is

$$a_r = \frac{I_1 - I_2}{I_1 + I_2}, \tag{49}$$

with

$$I_1 = \frac{(n^2 - 1)\beta_1 \cos \kappa_2 d_2}{(\beta_2^2 - \beta_1^2)(\kappa_1^2 + \gamma_2^2)} \left[\frac{4\gamma_1\gamma_2}{\beta_1\beta_2 \left(\frac{n^2k^2}{\beta_1^2 + n^2\gamma_1^2} + \gamma_1 d_1 \right) \left(\frac{n^2k^2}{\beta_2^2 + n^2\gamma_2^2} + \gamma_2 d_2 \right)} \right]^{\frac{1}{2}}$$
$$\cdot \{\kappa_1 k^2 \sin \kappa_1 d_2 - \gamma_2(\kappa_1^2 + \beta_2^2) \cos \kappa_1 d_2$$
$$+ [\gamma_2(n^2k^2 + \beta_2^2 - \beta_1^2) - n^2\gamma_1 k^2] e^{-\gamma_2(d_1 - d_2)} \cos \kappa_1 d_1\}, \tag{50}$$

and

$$I_2 = \frac{(n^2 - 1)\beta_2 \cos \kappa_2 d_2}{(\beta_2^2 - \beta_1^2)(\kappa_1^2 + \gamma_2^2)} \left[\frac{4\gamma_1\gamma_2}{\beta_1\beta_2 \left(\frac{n^2k^2}{\beta_1^2 + n^2\gamma_1^2} + \gamma_1 d_1 \right) \left(\frac{n^2k^2}{\beta_2^2 + n^2\gamma_2^2} + \gamma_2 d_2 \right)} \right]^{\frac{1}{2}}$$
$$\cdot \{\kappa_1(k^2 + \beta_1^2 - \beta_2^2) \sin \kappa_1 d_2 - n^2\gamma_2 k^2 \cos \kappa_1 d_2$$
$$+ n^2[(\gamma_2 - \gamma_1)k^2 + \gamma_1(\beta_2^2 - \beta_1^2)] e^{-\gamma_2(d_1 - d_2)} \cos \kappa_1 d_1\}. \tag{51}$$

The corresponding expression for the TE modes, equation (37) is apparently considerably simpler.

The expression for the radiation loss of TM modes on a dielectric waveguide of arbitraty shape is obtained from

$$\frac{\Delta P}{P} = \int_{-k}^{k} \{ | q_e(\rho) |^2 + | q_0(\rho) |^2 \} \frac{| \beta |}{\rho} d\beta \tag{52}$$

with the coefficient of the even radiation modes

$$q_e(\rho) =$$
$$-\int_0^L \frac{(n^2 - 1)\rho\gamma^{\frac{1}{2}}(\beta_0\beta \cos \sigma d + \gamma\sigma \sin \sigma d) \cos \kappa d e^{-i(\beta_0 - \beta)z} \left(\frac{\partial f}{\partial z} - \frac{\partial h}{\partial z} \right)}{2(\beta_0 - \beta) \left\{ \pi\beta_0 | \beta | \left(\frac{n^2k^2}{\beta_0^2 + n^2\gamma^2} + \gamma d \right) \left(n^2\rho^2 \cos^2 \sigma d + \frac{\sigma^2}{n^2} \sin^2 \sigma d \right) \right\}^{\frac{1}{2}}} dz \tag{53}$$

and the coefficient for the odd radiation modes

$$q_0(\rho) =$$
$$-\int_0^L \frac{(n^2 - 1)\rho\gamma^{\frac{1}{2}}(\beta_0\beta \sin \sigma d - \gamma\sigma \cos \sigma d) \cos \kappa d e^{-i(\beta_0 - \beta)z} \left(\frac{\partial f}{\partial z} + \frac{\partial h}{\partial z} \right)}{2(\beta_0 - \beta) \left\{ \pi\beta_0 | \beta | \left(\frac{n^2k^2}{\beta_0^2 + n^2\gamma^2} + \gamma d \right) \left(n^2\rho^2 \sin^2 \sigma d + \frac{\sigma^2}{n^2} \cos^2 \sigma d \right) \right\}^{\frac{1}{2}}} dz. \tag{54}$$

The restriction to symmetrical waveguides was dropped so that equations (52) through (54) hold for waveguides of arbitrary shapes as shown in Fig. 4. A comparison of equations (53) and (45) shows immediately how equation (45) could be generalized to an arbitrary waveguide shape. Corresponding expressions for the odd TE radiation modes could immediately be constructed by a comparison of equation (61), Ref. 1, with equation (54). The function $h(z)$ describes the shape of the dielectric slab waveguide at the lower air-dielectric interface. The theory of dielectric slab waveguides with rough wall, as presented in Ref. 1, was limited to TE modes. The same procedure which lead from equations (45) to (46) allows us to derive the TM-mode radiation loss equations for waveguides with rough walls.

$$q_e(\rho) = \frac{(n^2 - 1)\rho\gamma^{\frac{1}{2}}(\beta_0\beta \cos \sigma d + \gamma\sigma \sin \sigma d) \cos \kappa d[\varphi(\beta) - \psi(\beta)]}{2i\left\{\pi\beta_0 \mid \beta \mid \left(\frac{n^2k^2}{\beta_0^2 + n^2\gamma^2} + \gamma d\right)\left(n^2\rho^2 \cos^2 \sigma d + \frac{\sigma^2}{n^2} \sin^2 \sigma d\right)\right\}^{\frac{1}{2}}}, \tag{55}$$

and

$$q_0(\rho) = \frac{(n^2 - 1)\rho\gamma^{\frac{1}{2}}(\beta_0\beta \sin \sigma d - \gamma\sigma \cos \sigma d) \cos \kappa d[\varphi(\beta) + \psi(\beta)]}{2i\left\{\pi\beta_0 \mid \beta \mid \left(\frac{n^2k^2}{\beta_0^2 + n^2\gamma^2} + \gamma d\right)\left(n^2\rho^2 \sin^2 \sigma d + \frac{\sigma^2}{n^2} \cos^2 \sigma d\right)\right\}}. \tag{56}$$

The Fourier component $\varphi(\beta)$ is given by equation (47). The corresponding Fourier component $\psi(\beta)$ follows from equation (47) by replacing $f(z)$ with $h(z)$.

V. NUMERICAL RESULTS

The radiation losses caused by a symmetrical step with the ratio $d_2/d_1 = 0.5$ for $n = 1.01$ are shown in Fig. 5. The solid curves are obtained from equation (42) with the help of equations (45) and (53)

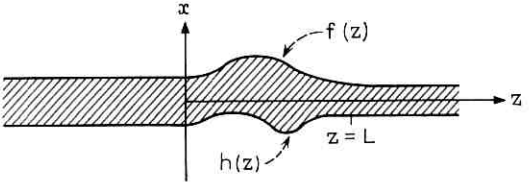


Fig. 4 — An asymmetrical wall distortion of the slab waveguide.

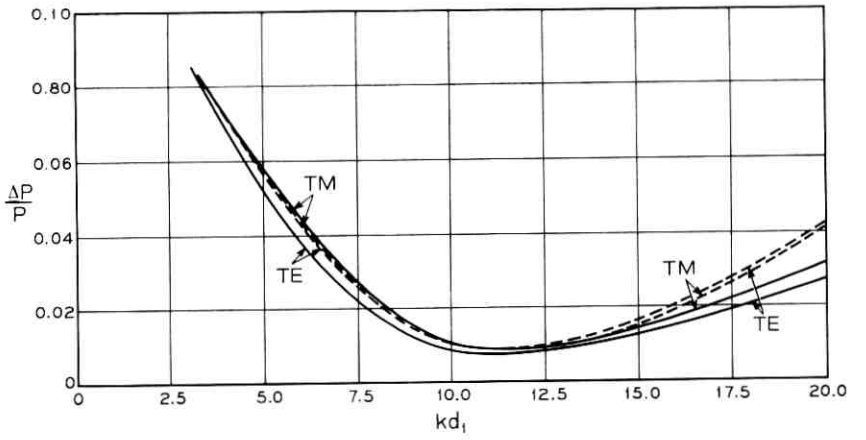


Fig. 5 — TE and TM mode losses caused by a step in the slab waveguide. Solid line calculated from (42), (52) dotted line calculated from (32). $n = 1.01$, $d_2/d_1 = 0.5$.

by approximating the step with a steep linear taper of length $L/d_1 = 1$. For very short tapers, the radiation loss is independent of the length of the taper. The dotted curves were obtained from equations (32) and (37) for TE modes and equations (48) through (51) for TM modes. The agreement between the results obtained by the two different methods is quite good. It is also apparent that TE modes and TM modes suffer very nearly the same losses in this case. It is surprising how low the radiation losses are in the region of $kd_1 = 11$. Both modes pass this considerable step with a power loss of less than 1 percent. For $kd_1 > 20$ the larger portion of the waveguide can support more than one guided mode. This is the reason why the loss curves were not extended past this point. Both the TE as well as TM modes show minimum loss values for particular values of kd_1 suggesting the possibility of optimizing waveguide steps.

Fig. 6 shows the radiation losses of the even, lowest order TE and TM mode for a step on a single mode waveguide with $n = 1.432$. The TE and TM mode losses are quite different for this waveguide with high dielectric constant. The fact that for TE as well as TM modes there is an increasing discrepancy between the two methods of calculation for increasing values of kd, with the dotted curve for the TM modes even becoming negative, may indicate that the solid curves are more reliable. For small values of kd, the agreement between the two methods becomes quite good. The losses of the TM mode are generally higher than the TE mode loss. However, even in this case the TE mode loss can be

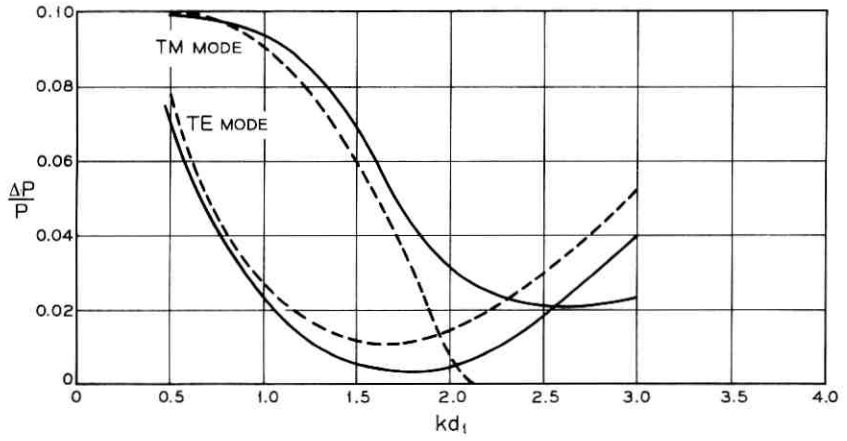


Fig. 6 — Same as Fig. 5. $n = 1.432$.

made approximately 1 percent while the TM mode loss can be as low as 2 percent if the step is used at its optimum point of operation. For $kd_1 > 3$ the larger waveguide section ceases to be single mode.

The dependence of the TE-mode radiation losses on the ratio of the width d_2/d_1 of the guide on either side of the step is shown in Fig. 7. This curve was computed from equations (32), (36) and (37). The dielectric constant of the waveguide material was chosen as $n = 1.01$

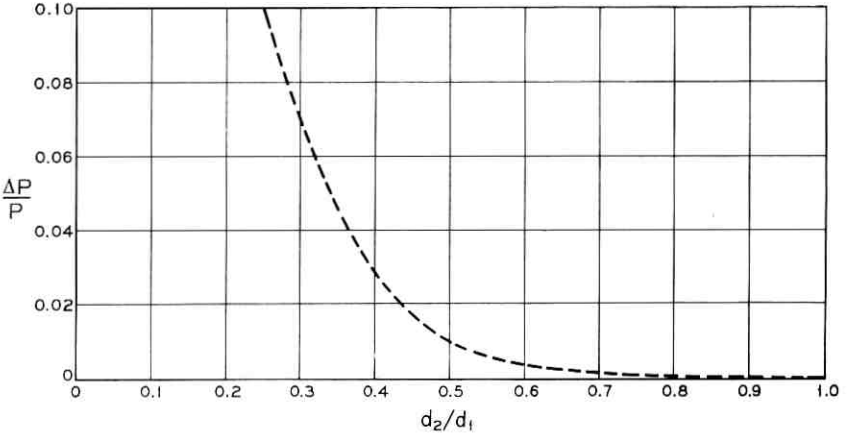


Fig. 7 — Step loss of TE mode as a function of the ratio d_2/d_1. $n = 1.01$, $kd_1 = 10.0$.

and $kd_1 = 10$ was used. It is apparent that the radiation losses increase rapidly as $d_2/d_1 \to 0$.

So far we have discussed the radiation losses of abrupt steps. The reduction of the TM-mode losses as the step is changed into a taper is seen in Fig. 8. This figure was calculated from equations (52) and (53) for a ratio of $d_2/d_1 = 0.5$ of the straight guide sections that are connected by a symmetrical linear taper. It is apparent that the linear taper needs to be quite long before a substantial improvement of the radiation loss is obtained. The actual length of an effective taper need not be very large. The length of the taper is represented in Fig. 8 as the ratio of its actual length to the half width d_1 of the thicker waveguide section. Extrapolating the result of Fig. 8 to a value of $L/d_1 = 100$ appears to lead to a loss reduction to approximately 1/10 of the loss of the abrupt step. With $\lambda = 1\mu$ we find that $kd_1 = 1$ corresponds to $d_1 = 0.16\ \mu$ so that $L/d_1 = 100$ corresponds to $L = 16\ \mu$.

It appears that there are more effective shapes than linear tapers. Equations (45), (53) and (54) show that the loss of a taper is essentially determined by two factors, the magnitude of the derivatives df/dz and dh/dz and the value of $\beta_0 - \beta$. Rapid oscillations of the function $\exp[i(\beta_0 - \beta)z]$ cause the value of the integral to be small. The largest value of β is $\beta = k$. The worst value appearing in the argument of the exponential function is, therefore, $\beta_0 - k$. The propagation constant of the guided mode depends on the width of the waveguide and is therefore a function of z. The optimum taper, that is intended to connect two

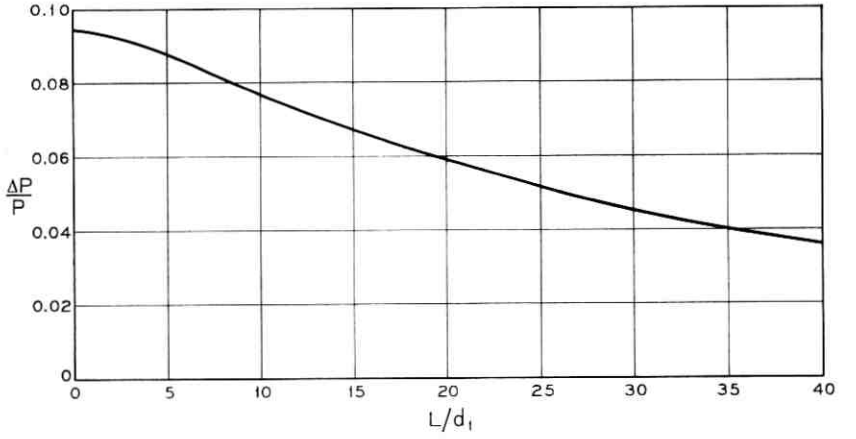


Fig. 8 — TM mode radiation loss as a function of the length L of the taper. $n = 1.432$, $kd_1 = 1.0$, $d_2/d_1 = 0.5$.

different waveguides in a given length, would attempt to use larger values of df/dz and dh/dz on the wide part of the taper where $\beta_0 - k$ is still larger and provide smaller values of these derivatives on its narrow part where $\beta_0 - k$ is smaller. A linear taper radiates more on its narrower portion where the field is less tightly guided. An optimum taper would attempt to distribute the radiation loss uniformly over the length of the taper.

VI. RANDOM WALL DISTORTION

In Ref. 1 we computed the losses of the lowest order guided TE mode that is caused by random distortions of one of the two waveguide walls. For the sake of completeness we include here the corresponding formula for TM modes which can be immediately obtained from the theory presented in Ref. 1 and our present equations (55) and (56). The ensemble average of the relative power loss of the lowest order even TM mode (caused by the distortion of one wall by a random process whose correlation function is a simple exponential function, equation (85) of Ref. 1) with r.m.s. deviation A and correlation length B is given by

$$\left\langle \frac{\Delta P}{P} \right\rangle_{av} = \frac{A^2 \gamma L (n^2 - 1)^2}{2\pi B \beta_0} \int_{-k}^{k} \frac{\rho \cos^2 \kappa_0 d}{\left[(\beta_0 - \beta)^2 + \frac{1}{B^2}\right]\left[\frac{n^2 k^2}{\beta_0^2 + n^2\gamma^2} + \gamma d\right]} \cdot \left\{ \frac{(\beta_0 \beta \cos \sigma d + \gamma\sigma \sin \sigma d)^2}{n^2 \rho^2 \cos^2 \sigma d + \frac{\sigma^2}{n^2} \sin^2 \sigma d} + \frac{(\beta_0 \beta \sin \sigma d - \gamma\sigma \cos \sigma d)^2}{n^2 \rho^2 \sin^2 \sigma d + \frac{\sigma^2}{n^2} \cos^2 \sigma d} \right\} d\beta. \tag{57}$$

The radiation loss that is obtained from this equation is shown in Figs. 9 and 10, by the solid lines. The dotted curves are reproduced from Ref. 1 and give the loss of the TE mode for comparison. The curves labeled $\Delta P^-/\Delta P^+$ show the ratio of backward to forward scattered power. The conclusion to be drawn from these curves is that the TM mode losses caused by small random wall perturbation are very nearly the same as for TE modes. Neither type of mode seems to offer a distinct advantage.

The radiation losses of slab waveguides with random wall distortions are representative of the losses of round waveguides with similar wall distortions. However, the radiation losses of slab waveguide tapers are considerably lower than those of round waveguides. (A discussion of round waveguides will be published.)

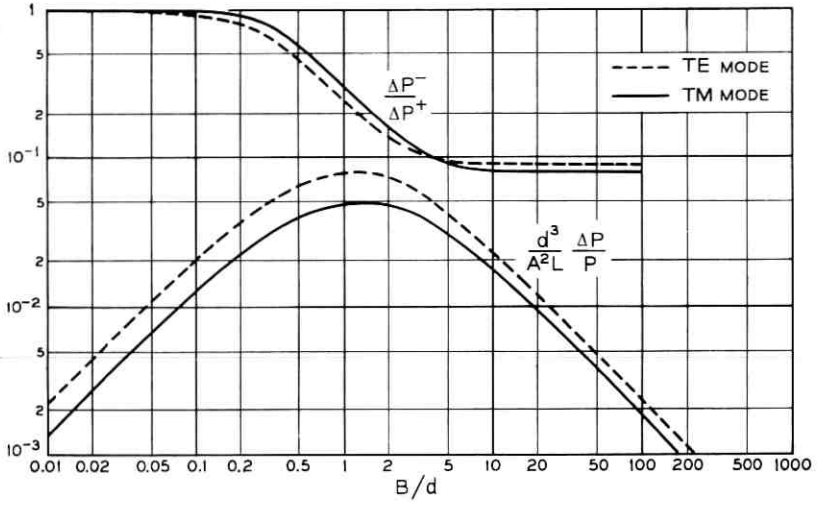


Fig. 9 — Comparison of TM loss (solid line) and TE loss (dotted line) caused by a random distortion of one waveguide wall. B = correlation length, A = r.m.s. wall distortion, d = half width of slab, L = length of distorted guide section. $n = 1.5$, $kd = 1.3$.

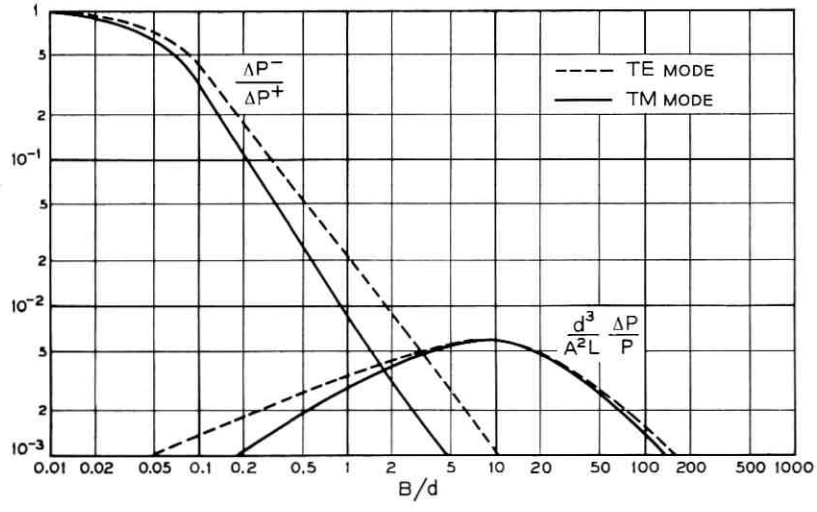


Fig. 10 — Same as Fig. 9. $n = 1.01$, $kd = 8.0$.

VII. CONCLUSION

We have derived radiation loss formulae for the dominant mode dielectric slab waveguide. The losses for steps and tapers in the waveguide were calculated for TE as well as TM modes. The theory of radiation losses for random wall imperfections, that was developed earlier for TE modes, was extended to TM modes.

The radiation losses of abrupt steps with a 2:1 ratio were found to be surprisingly low (a few percent). The advantage of gradual linear tapers over abrupt steps becomes appreciable only if the taper is much longer than the width of the slab.

The losses of steps and tapers of the slab waveguide are exceptionally low. Dielectric waveguides with round and rectangular cross sections have considerably highest losses. However, the method of describing waveguide distortions as successions of abrupt steps is applicable to all dielectric waveguides and simplifies their treatment considerably.

REFERENCES

1. Marcuse, D., "Mode Conversion Caused by Surface Imperfections of a Dielectric Slab Waveguide," B.S.T.J., *48*, No. 10 (December 1969), pp. 3187–3216.
2. Marcuse, D., "Radiation Losses of Dielectric Waveguides in Terms of the Power Spectrum of the Wall Distortion Function," B.S.T.J., *48*, No. 10 (December 1969), pp. 3233–3242.
3. Marcuse, D., and Derosier, R. M., "Mode Conversion Caused by Diameter Changes of a Round Dielectric Waveguide," B.S.T.J., *48*, No. 10 (December 1969), pp. 3217–3232.

An Efficient Heuristic Procedure for Partitioning Graphs

By B. W. KERNIGHAN and S. LIN

(Manuscript received September 30, 1969)

We consider the problem of partitioning the nodes of a graph with costs on its edges into subsets of given sizes *so as to minimize the sum of the costs on all edges cut. This problem arises in several physical situations—for example, in assigning the components of electronic circuits to circuit boards to minimize the number of connections between boards.*

This paper presents a heuristic method for partitioning arbitrary graphs which is both effective in finding optimal partitions, and fast enough to be practical in solving large problems.

I. INTRODUCTION

1.1 *Definition of the Problem*

This paper deals with the following combinatorial problem: given a graph G with costs on its edges, partition the nodes of G into subsets no larger than a *given maximum size*, so as to minimize the total cost of the edges cut.

One important practical example of this problem is placing the components of an electronic circuit onto printed circuit cards or substrates, so as to minimize the number of connections between cards. The components are the nodes of the graph, and the circuit connections are the edges. There is some maximum number of components which may be placed on any card. Since connections between cards have high cost compared to connections within a board, the object is to minimize the number of interconnections between cards.

This partitioning problem also arises naturally in an attempt to improve the paging properties of programs for use in computers with paged memory organization. A program (at least statically) can be thought of as a set of connected entities. The entities might be subroutines, or procedure blocks, or single instruction and data items, depending on viewpoint and the level of detail required. The connections

between the entities might represent possible flow or transfer of control, or references from one entity to another. The problem is to assign the objects to "pages" of a given size so as to minimize the number of references between objects which lie on different pages.

To pose the partitioning problem mathematically, we shall need the following definitions. Let G be a graph of n nodes, of sizes (weights) $w_i > 0, i = 1, \cdots, n$. Let p be a positive number, such that $0 < w_i \leqq p$ for all i. Let $C = (c_{ij}), i, j = 1, \cdots, n$ be a weighted connectivity matrix describing the edges of G.

Let k be a positive integer. A *k-way partition* of G is a set of nonempty, pairwise disjoint subsets of G, $v_1, \cdots, v_k$ such that $\bigcup_{i=1}^{k} v_i = G$. A partition is *admissible* if

$$| v_i | \leqq p \quad \text{for all} \quad i,$$

where the symbol $| x |$ stands for the *size* of a set x, and equals the sum of the sizes of all the elements of x. The *cost* of a partition is the summation of c_{ij} over all i and j such that i and j are in different subsets. The cost is thus the sum of all external costs in the partition.

The partitioning problem we consider here is to find a minimal-cost admissible partition of G.

There are three other problems which are equivalent to this one. First, minimizing external cost is equivalent to maximizing internal cost because the total cost of all edges is constant. Further, by changing the signs of all c_{ij}'s, we can maximize external cost, or minimize internal cost.

1.2 *Exact Solutions*

A strictly exhaustive procedure for finding the minimal cost partition is often out of the question. To see this suppose that G has n nodes of size 1 to be partitioned into k subsets of size p, where $kp = n$. Then there are $\binom{n}{p}$ ways of choosing the first subset, $\binom{n-p}{p}$ ways for the second, and so on. Since the ordering of the subsets is immaterial, the number of cases is

$$\frac{1}{k!}\binom{n}{p}\binom{n-p}{p}\cdots\binom{2p}{p}\binom{p}{p}.$$

For most values of n, k, and p, this expression yields a very large number; for example, for $n = 40$ and $p = 10$ ($k = 4$), it is greater than 10^{20}.

Formally the problem could also be solved as an integer linear programming problem, with a large number of constraint equations necessary to express the uniformity of the partition.

Because it seems likely that any direct approach to finding an optimal

solution will require an inordinate amount of computation, we turn to an examination of heuristics. Heuristic methods can produce good solutions (possibly even an optimal solution) quickly. Often in practical applications, several good solutions are of more value than one optimal one.

The first and foremost consideration in developing heuristics for combinatorial problems of this type is finding a procedure that is powerful and yet sufficiently fast to be practical. A process whose running time grows exponentially or factorially with the number of vertices of the graph is not likely to be practical. In most cases, a growth rate of more than the square of the number of vertices is still not too practical. (If the running time of a procedure grows as $f(n)$, where n is the number of vertices involved, we shall refer to it as an $f(n)$-procedure.)

1.3 *False Starts*

To point out a few pitfalls, we mention some unsuccessful attempts at heuristic solutions to the partitioning problem.

1.3.1 *Random Solutions*

One tactic is simply to generate random solutions, keeping the best seen to date, and terminating after some predetermined time or value is reached. This is quite fast, although actually an n^2-procedure. Unfortunately, this approach is unsatisfactory for problems of even moderate size, since there are generally few optimal or near-optimal solutions, which thus appear randomly with very low probabilities. Experience with 2-way partitions for a class of 0–1 matrices of size 32×32, for example, has indicated that there are typically 3 to 5 optimal partitions, out of a total of $\frac{1}{2}\binom{32}{16}$ partitions, giving a probability of success on any trial of less than 10^{-7}.

1.3.2 *Max Flow-Min Cut*

Another partitioning method is the Ford and Fulkerson max flow-min cut algorithm[1]. The graph is treated as a network in which edge costs correspond to maximum flow capacities between pairs of nodes. A cut is a separation of the nodes into two disjoint subsets. The max flow-min cut theorem states that the maximal flow values between any pair of nodes is equal to the minimal cut capacity of all cuts which separate the two nodes. In our terminology, a cut is a 2-way partition, and the cut capacity is the cost of the partition. The Ford and Fulkerson algorithm finds a cut with maximal flow, which is thus a minimal cost cut; this represents a minimum cost partition of the graph into two subsets of unspecified sizes.

There are several difficulties involved in using the Ford and Fulkerson algorithm for our partitioning problem. The most severe of these is the fact that the algorithm has no provision for constraining the sizes of the resultant subsets, and there seems to be no obvious way to extend it to include this. Thus if flow methods are used to perform a split, then further processing is necessary to make the resulting subsets the correct size. If the subsets are greatly different in size, then use of this algorithm will have produced essentially no benefit. Hence in spite of its theoretical elegance, the Ford and Fulkerson algorithm is not suitable for this application. (Note however, that since it does find the minimal cost unconstrained 2-way partition, the value it produces is a lower bound for solutions produced by any method.)

1.3.3 *Clustering*

A class of much more intuitive methods is based on identifying "natural clusters" in the given cost matrix—that is, groups of nodes which are strongly connected in some sense. For example, one can use very simple heuristics for building up clusters, based on collecting together elements corresponding to large values in the cost matrix. But again these methods do not in general include much provision for satisfying constraints on the sizes of the subsets, nor do they provide for systematic assignment of "stragglers" (nodes which do not obviously belong to any particular subset).

1.3.4 λ-*Opting*

Lin, working on the Traveling Salesman Problem, [See Ref. 2] categorized a set of methods of improving given solutions by rearranging single links, double links, triplets, and in general, λ links. He referred to a change involving the movement of λ links as a λ-change. If a configuration of the system is reached in which no λ-change can be made which results in a decrease in cost, the configuration is said to be "λ-opt."

For the partitioning problem, an analogous operation is the interchange of groups of λ points between a pair of sets. Thus a 1-change is the exchange of a single point in one set with a single point in another set. A configuration is then said to be "1-opt" if there exists no interchange of two points which decreases the cost of the partition. Experiments to evaluate 1-opting for 2-way partitions of 0-1 matrices (32×32) within which about one-half of the elements were nonzero, show that apparently optimal values can be achieved in about 10 percent of the trials; values within 1 or 2 of the optimal can be achieved in about 75 percent of cases.

It appears fruitless to extend λ beyond 1 (1-opting is already an n^2-procedure), or to extend 1-opting experiments to partitions into more than two subsets, since more powerful methods have been developed. These methods are the topic of the next sections.

II. TWO-WAY UNIFORM PARTITIONS

2.1 *Introduction*

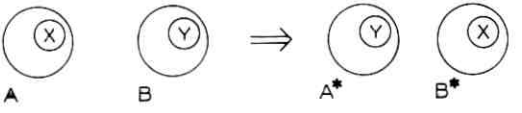
The simplest partitioning problem which still contains all the significant features of larger problems is that of finding a minimal-cost partition of a given graph of $2n$ vertices (of equal size) into two subsets of n vertices each. The solution of the 2-way partitioning problem is the subject of this section. The solution provides the basis for solving more general partitioning problems. In Section 2.6, we discuss 2-way partitions into sets of unequal size.

Let S be a set of $2n$ points, with an associated cost matrix $C = (c_{ij})$, $i, j = 1, \cdots, 2n$. We assume without loss of generality that C is a symmetric matrix, and that $c_{ii} = 0$ for all i. There is no assumption about nonnegativity of the c_{ij}'s. We wish to partition S into two sets A and B, each with n points, such that the "external cost" $T = \sum_{A \times B} c_{ab}$ is minimized.

In essence, the method is this: starting with any arbitrary partition A, B of S, try to decrease the initial external cost T by a series of interchanges of subsets of A and B; the subsets are chosen by an algorithm to be described. When no further improvement is possible, the resulting partition A', B' is locally minimum with respect to the algorithm. We shall indicate that the resulting partition has a fairly high probability of being a globally minimum partition.

This process can then be repeated with the generation of another arbitrary starting partition A, B, and so on, to obtain as many locally minimum partitions as we desire.

Given S and (c_{ij}), suppose A^*, B^* is a minimum cost 2-way partition. Let A, B be any arbitrary 2-way partition. Then clearly there are subsets $X \subset A$, $Y \subset B$ with $|X| = |Y| \leqq n/2$ such that interchanging X and Y produces A^* and B^* as shown below.



$$A^* = A - X + Y$$
$$B^* = B - Y + X$$

The problem is to identify X and Y from A and B, without considering all possible choices. The process we describe finds X and Y approximately, by sequentially identifying their elements.

Let us define for each $a \in A$, an *external cost* E_a by

$$E_a = \sum_{y \in B} c_{ay}$$

and an *internal cost* I_a by

$$I_a = \sum_{x \in A} c_{ax} .$$

Similarly, define E_b , I_b for each $b \in B$. Let $D_z = E_z - I_z$ for all $z \in S$; D_z is the difference between external and internal costs.

Lemma 1: *Consider any $a \in A$, $b \in B$. If a and b are interchanged, the gain (that is, the reduction in cost) is precisely $D_a + D_b - 2c_{ab}$.*

Proof: Let z be the total cost due to all connections between A and B that do not involve a or b. Then

$$T = z + E_a + E_b - c_{ab} .$$

Exchange a and b; let T' be the new cost. We obtain

$$T' = z + I_a + I_b + c_{ab}$$

and so

$$\begin{aligned} \text{gain} &= \text{old cost} - \text{new cost} = T - T' \\ &= D_a + D_b - 2c_{ab} . \end{aligned}$$

2.2 *Phase 1 Optimization Algorithm*

In this subsection we present the algorithm for 2-way partitioning.

First, compute the D values for all elements of S. Second, choose $a_i \in A$, $b_j \in B$ such that

$$g_1 = D_{a_i} + D_{b_j} - 2c_{a_i b_j}$$

is maximum; a_i and b_j correspond to the largest possible gain from a single interchange. (We will return shortly to a discussion of how to select a_i and b_j quickly.) Set a_i and b_j aside temporarily, and call them a_1' and b_1' , respectively.

Third, recalculate the D values for the elements of $A - \{a_i\}$ and for $B - \{b_j\}$, by

$$\begin{aligned} D_x' &= D_x + 2c_{xa_i} - 2c_{xb_j} \qquad x \in A - \{a_i\}, \\ D_y' &= D_y + 2c_{yb_j} - 2c_{ya_i} \qquad y \in B - \{b_j\}. \end{aligned}$$

The correctness of these expressions is easily verified: the edge (x, a_i) is counted as internal in D_x, and it is to be external in D'_x, so c_{xa_i} must be added twice to make this correct. Similarly, c_{xb_i} must be subtracted twice to convert (x, b_i) from external to internal.

Now repeat the second step, choosing a pair a'_2, b'_2 from $A - \{a'_1\}$ and $B - \{b'_1\}$ such that $g_2 = D_{a_2'} + D_{b_2'} - 2c_{a_2'b_2'}$ is maximum (a'_1 and b'_1 are *not* considered in this choice). Thus g_2 is the additional gain when the points a'_2 and b'_2 are exchanged as well as a'_1 and b'_1; this additional gain is maximum, given the previous choices. Set a'_2 and b'_2 aside also.

Continue until all nodes have been exhausted, identifying $(a'_3, b'_3), \cdots, (a'_n, b'_n)$, and the corresponding maximum gains $g_3, \cdots, g_n$. As each (a', b') pair is identified, it is removed from contention for further choices so the size of the sets being considered decreases by 1 each time an (a', b') is selected.

If $X = a'_1, a'_2, \cdots, a'_k$, $Y = b'_1, b'_2, \cdots, b'_k$, then the decrease in cost when the sets X and Y are interchanged is precisely $g_1 + g_2 + \cdots + g_k$. Of course $\sum_1^n g_i = 0$. Note that some of the g_i's are negative, unless all are zero.

Choose k to maximize the partial sum $\sum_{i=1}^k g_i = G$. Now if $G > 0$, a reduction in cost of value G can be made by interchanging X and Y. After this is done, the resulting partition is treated as the initial partition, and the procedure is repeated from the first step.

If $G = 0$, we have arrived at a locally optimum partition, which we shall call a *phase 1 optimal partition*. We now have the choice of repeating with another starting partition, or of trying to improve the phase 1 optimal partition. We shall discuss the latter option shortly. Figure 1 is a flowchart for the phase 1 optimization procedure.

2.3 *Effectiveness of the Procedure*

One general approach to solving problems such as this one is to find the *best* exchange involving say λ pairs of points, for some λ specified in advance[2]. The difficulty encountered is that use of a small value of λ is not sufficient to identify good exchanges, but the computational effort required grows rapidly as λ increases.

The procedure we have described *sequentially* finds an approximation to the best exchange of λ pairs. λ is not specified in advance, but rather is chosen to make the improvement as large as possible. This technique sacrifices a certain amount of power for a considerable gain in speed.

Since we construct a sequence of gains g_i, $i = 1, \cdots, n$, and find the

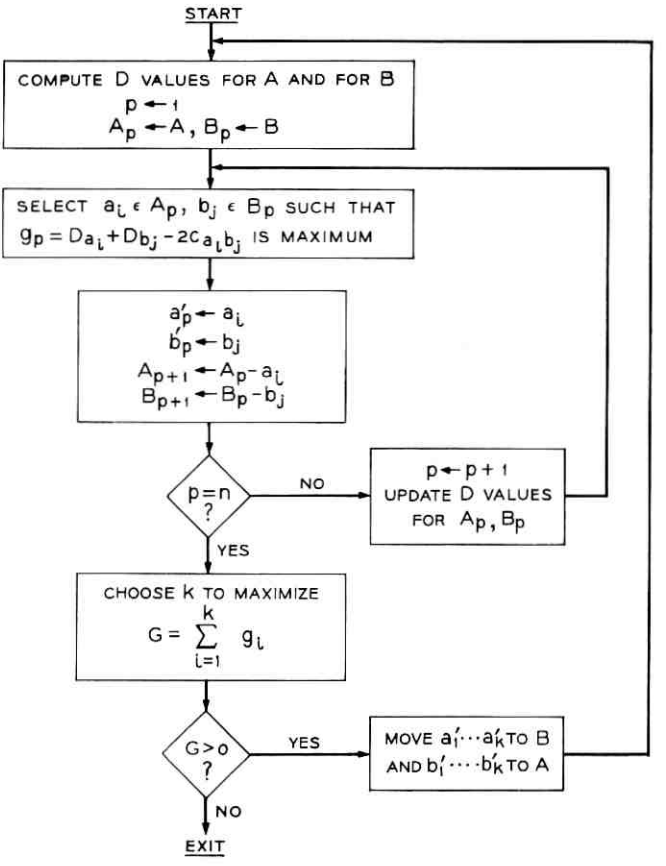


Fig. 1 — Flowchart of phase 1 optimization procedure.

maximum partial sum, *the process does not terminate immediately when some g_i is negative.* This means that the process can sequentially identify sets for which the exchange of only a few elements would actually increase the cost, while the interchange of the entire sets produces a net gain.

Numerous experiments have been performed to evaluate the procedure on different types of cost matrices. The matrices used have included (*i*) 0-1 matrices, with density of nonzero elements ranging from 5 percent to 50 percent, (*ii*) integer matrices with elements uniformly distributed on $[0, k]$, $k = 2, \cdots, 10$, (*iii*) matrices with clusters of known sizes and binding strength. Results on all of these matrices have

been similar, so we shall only summarize them here. A more extended discussion may be found in Ref. 3.

A useful measure of the power of a heuristic procedure is the probability that it finds an optimal solution in a single trial. Suppose that p is the probability that a phase 1 optimal solution found using a random starting partition is globally optimal. We have examined the behavior of this probability as the size of the matrices involved is varied. Experiments show p is around 0.5 for matrices of size 30×30, 0.2 to 0.3 for 60×60, and 0.05 to 0.1 for 120×120. The functional behavior of p is approximately $p(n) = 2^{-n/30}$.

These values are derived primarily from 0-1 matrices having about 50 percent 1's (randomly placed). Experiments on matrices with lower densities of 1's yield larger variances, but substantially identical mean values for p.

2.4 *Running Time of the Procedure*

Let us define a *pass* to be the operations involved in making one cycle of identification of $(a_1', b_1'), \cdots, (a_n', b_n')$, and selection of sets X and Y to be exchanged. The total time for a pass can be estimated this way. First, the computation of the D values initially is an n^2-procedure, since for each element of S, all the other elements of S must be considered. The time required for updating the D values is proportional to the number of values to be updated, so the total updating time in one pass grows as

$$(n - 1) + (n - 2) + \cdots + 2 + 1$$

which is proportional to n^2.

The dominant time factor is the selection of the next pair a_i', b_i' to be exchanged. The method we have used to perform this searching is to sort the D values so that

$$D_{a_1} \geqq D_{a_2} \geqq \cdots \geqq D_{a_n}$$

and

$$D_{b_1} \geqq D_{b_2} \geqq \cdots \geqq D_{b_n} .$$

When sorting is used, only a few likely contenders for a maximum gain need be considered. This is because when scanning down the set of D_a's and D_b's, if a pair D_{a_i}, D_{b_j} is found whose sum does not exceed the maximum improvement seen so far in this pass, then there cannot be another pair a_k, b_l with $k \geqq i$, $l \geqq j$, with a greater gain, (assuming $c_{ij} \geqq 0$) and so the scanning can be terminated. Thus the next pair

for interchange is found rapidly. Sorting is an $n \log n$ operation, so in this method, the total time required to sort D values in a pass will be approximately

$$n \log n + (n - 1) \log (n - 1) + \cdots + 2 \log 2$$

which grows as $n^2 \log n$.

To reduce the time for selection of an (a, b) pair, it is possible to use techniques which are faster than sorting, but which do not necessarily always give the maximum gain at each stage. For example, one method is to scan for the largest D_a and the largest D_b, and use the corresponding a and b as the next interchange. This method is essentially linear-time and would probably be implemented as part of the recomputation of the D values. It is best suited for sparse matrices, where the probability that $c_{ab} > 0$ is small. A slight extension, involving negligible extra cost, is to save the largest two or three D_a's and D_b's, so that if the largest pair does not give the maximum gain (because c_{ab} is too large), then another can be tried. Experience indicates that three values are sufficient in virtually all cases, even for matrices with a relatively high percentage of nonzero entries. Use of this method reduces running time by about 30 percent in the present implementation, with very small degradation of power.

The number of passes required before a phase 1 optimal partition is achieved is small. On all matrix sizes tested at the time of writing (up to 360 points), it has been almost always from 2 to 4 passes. On the basis of this experimental evidence, the number of passes is not strongly dependent on the value of n.

From the foregoing observations, it is possible to estimate the total running time of the procedure. If we use a method which sorts the D values at each stage (time proportional to $n^2 \log n$), then the running time should grow as $n^2 \log n$. If a fast-scan method is used, and the number of passes is constant, the running time should have an n^2 growth rate; this is a lower bound.

For comparison, examination of all pairs of sets X and Y, and evaluation of the costs would require time proportional to

$$\begin{aligned} n^2 \sum_{k=1}^{n/2} \binom{n}{k}^2 &\sim \frac{n^2}{2} \sum_{k=0}^{n} \binom{n}{k}^2 \\ &= \frac{n^2}{2} \binom{2n}{n} \\ &\sim \frac{n^2}{2} 4^n \left(\frac{1}{\pi n}\right)^{\frac{1}{2}} \end{aligned}$$

for large n. This function grows as $n^{3/2}4^n$.

Running times have been plotted in Fig. 2. The observed times have an apparent growth rate of about $n^{2.4}$, which is reasonably close to n^2. Although on the logarithmic plot this curve is close to linear over the range $n = 20$ to $n = 130$, it may actually be $n^2 \log n$; insufficient data is available to check this. All times are based on an implementation in FORTRAN G on an IBM System 360 Model 65.

2.5 *Improving the Phase 1 Optimal Partition*

In this section, we discuss a method which might be used to improve the partition produced by the phase 1 procedure, which may not be globally optimum. The method suggested in this section is based heavily on experimental evidence, although there are quite plausible reasons for performing the particular set of operations. The basic idea is to perturb the locally optimal solution in what we hope is an enlightened manner, so that an iteration of the process on the perturbed solution will yield a further reduction in the total cost. If this tactic fails, nothing has been lost except some computation time, since the best solution seen so far is always saved.

Computer results for problems with up to 64 points suggest that whenever a phase 1 optimal solution is not globally optimal, $| X | = | Y | \approx n/2$. Roughly, this implies that if $| X |$ and $| Y |$ had been small

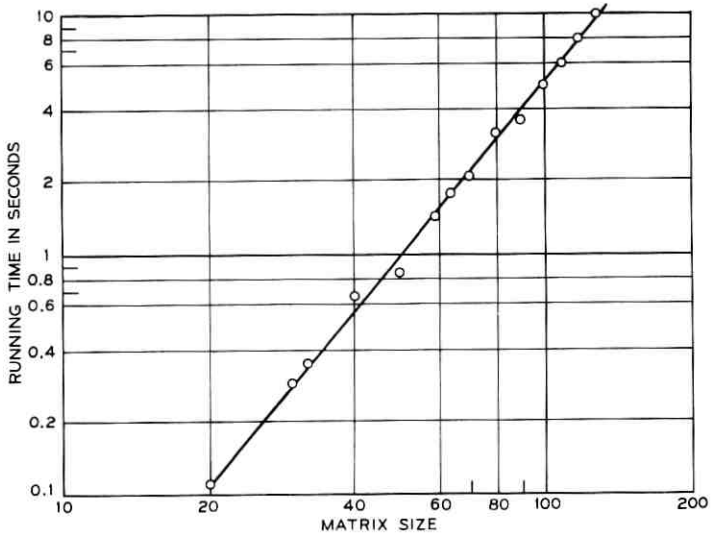


Fig. 2 — Running time.

compared to $n/2$, they would have been found by the process; it is only larger sets which are not identified all the time.

A successful heuristic to find the correct X and Y in this case is to find a phase 1 optimal partition for each of the sets A and B, say $A \rightarrow [A_1, A_2]$ and $B \rightarrow [B_1, B_2]$. (That is, find near-optimal partitions of A and of B separately.) Recombine the 4 sets into 2, say $A_0 = A_1 \cup B_1$ and $B_0 = A_2 \cup B_2$, and continue with phase 1 optimization. If our expectation is correct, the new X and Y will be small, and thus readily identified by the phase 1 process.

When A is split into A_1, A_2 and B into B_1, B_2 there are two ways in which the smaller sets can be recombined. A series of tests was made on matrices of moderate size (up to 64 $\times$ 64), in which both possible recombinations were done, generating three phase 1 optimal values for each starting partition. For matrices of size 32 $\times$ 32, the apparent optimal value was observed at least once in each triple of values, for a large number of cases. With matrices of size 64 $\times$ 64, there were occasional failures.

It might be noted that the extra time involved for the recombination approach is three times that required to do a completely new partition from a random start, assuming an n^2-procedure.

It is possible to estimate whether a particular improvement tactic is profitable or not in the following way. Suppose that some method increases the probability of finding an optimal partition from p to p', while it increases the running time from t to t'. Then in a fixed amount of time, it is possible to do k trials of the basic procedure, and kt/t' trials of the improved method. The corresponding probabilities of achieving an optimal solution are $1 - (1 - p)^k$ and $1 - (1 - p')^{kt/t'}$ respectively. The improved method is then desirable if the second expression is greater than the first; by simple manipulation, this condition becomes

$$1 - p' < (1 - p)^{t'/t}.$$

On the basis of the numerical values in this section, it may be useful to try the recombination method.

2.6 *Partitioning into Unequal-Sized Sets*

It is simple to modify the procedure to partition a set S with n elements into two sets of specified sizes n_1 and $n_2 (n_1 + n_2 = n)$. Assume $n_1 < n_2$. Then restrict the maximum number of pairs that can be exchanged in one pass of the procedure to n_1. All other operations are performed on all elements of each set. (The starting partition is into two sets, of n_1 and n_2 elements respectively.)

Suppose we wish to partition S into two sets, such that there are at least n_1 elements and at most n_2 elements in each subset; $n_1 + n_2 = n$, but they are not specified further.

The procedure is easily modified to handle this sort of constraint by the addition of "dummy" elements. These are elements which have no connections whatsoever; that is, they have zero entries in the cost matrix wherever they appear. Add $2n_2 - n$ dummies so S has $2n_2$ elements, and perform the procedure on it. The resulting partition will assign the dummy elements to the two subsets so as to minimize the external cost; at this point the dummies are discarded, leaving a partition into two subsets that satisfy the size constraints given.

2.7 *Elements of Unequal Sizes*

We have made the assumption so far that the elements (vertices) of the graph are all of the same size. This requirement may be relaxed to a large extent by converting any node of size $k > 1$ to a cluster of k nodes of size 1, bound together by edges of appropriately high cost. The size of the problem will obviously increase proportionally to the value of k, so it may be necessary to sacrifice some accuracy to keep the number of generated nodes within reasonable bounds.

III. MULTIPLE-WAY PARTITIONS

3.1 *Reduction to 2-Way Partitioning Problem*

So far, the discussion has been concerned exclusively with the basic problem of performing a 2-way partition on a set of $2n$ objects. In this section we extend the technique to perform k-way partitions on a set of kn objects, using the 2-way procedure as a tool.

The essential idea is to start with some partition into k sets of size n and by repeated application of the 2-way partitioning procedure to pairs of subsets, make the partition as close as possible to being pairwise optimal. (Section 3.2 treats the question of what starting sets to use.) Of course pairwise optimality is only a necessary condition for global optimality. There may be situations where some complex interchange of three or more items from three or more subsets is required to reduce a pairwise optimal solution to globally optimum; at the moment, no reasonable method for identifying such sets is known.

There are $\binom{k}{2}$ pairs of subsets to consider, so the time for one pass through all pairs is (assuming an n^2-procedure) $\binom{k}{2}n^2 \approx (kn)^2/2 =$ (number of points)$^2/2$. In general, more passes than this will actually be required, since when two sets are made optimal, this may change their optimality with respect to other sets.

Experience indicates that the number of passes is small and the process converges quickly. For example, our algorithm selects (i, j) as the next pair of sets to be optimized, where either i or j has been changed since the last time the pair (i, j) was selected. Using this selection process, the average number of passes through each pair of sets is a slowly growing function of both k and n. For matrices of size 100 or less and $k < 6$, the number of passes has been less than 5. [The average number of passes is computed as the average number of pairs considered to reach pairwise optimality, normalized by $\binom{k}{2}$.]

In any particular trial, there is a correlation between the number of pairs selected and the quality of the final partition. To get a better solution requires more work.

Convergence is rapid: two passes account for more than 95 percent of the improvement in most cases; the remaining passes contribute only small further reductions. Let $p(n, k)$ be the proportion of minimum cost solutions found for a particular n and k. For k fixed and small compared to n, the functional behavior of $p(n, k)$ is similar to the case $k = 2$, but the actual values are lower. Roughly, we observe $p(n, k + 1) \approx \frac{1}{2}p(n, k)$ for k in the range 2–4, and n up to 100, with considerable variation depending on the matrix being tested. For instance, for matrices of size about 40, $p(40, 2) \approx 0.4$, $p(42, 3) \approx 0.2$, and $p(40, 4) \approx 0.1$.

Another interesting question is measurement of how close to optimum the partitions found are. The solutions obtained by pairwise optimization have values concentrated in a narrow range. In almost all cases, the largest value found by the procedure is within 4–5 percent of the smallest. As another measure, if c is the mean cost of random partitions and b is the cost of the best partition observed, then virtually all partitions found have values v such that

$$v - b \leqq 0.1(c - b).$$

For instance, one test case was a series of 4-way partitions of a 0–1 matrix of size 80. This matrix had 1278 nonzero entries (a density of 0.2), corresponding to 639 edges in the graph. The mean value of randomly chosen partitions was 480.6. Twenty-four partitions of this matrix were found using the method described above. The lowest value encountered was 352 (1 time), the highest 365 (1 time); the mean value was 359.5, the median 360.

3.2 *Starting Partition*

In this subsection we discuss various methods of generating good starting partitions, based on modifications of the basic procedure.

The primary reason for choosing good starting partitions is that this particular form of preprocessing reduces the amount of work required to make the system pairwise optimal. It may also make the probability of an optimal solution higher, although this tendency is very difficult to evaluate.

Several methods for finding good multi-way starting partitions which are based on repeated application of the procedure itself have been investigated. The essential idea is to generate a k-way starting partition by first forming an r-way partition, then an s-way partition on each of the resulting subsets, and so on, up to t-way. (Here $k = rs \cdots t$.) The partitions found this way will in general be better than those which are completely arbitrary. A pairwise optimization stage is applied to the final set of subsets.

For example, if k is a power of 2, then perform a 2-way split, then a 2-way split on each of these subsets, and so on until the desired size of subsets is found.

This general approach is prone to the following difficulty: the first split divides the original set into r subsets by trying to make the internal connections in each subset as large as possible. Obviously this may conflict directly with the next stage, which is to try to divide each subset further. Carried to several levels, it can lead to a relatively poor overall solution. In experiments with 4-way partitions of matrices of sizes up to 64×64, this method yields optimal solutions approximately as often as does starting with a 4-way partition in the first place. In addition, this method will be effective if the matrix happens to have natural clusters of approximately the correct size (that is, equal to the final subset size).

A second method which can be used is to partition the set of kn elements into a set of n and a set of $(k - 1)n$, using the slightly modified version of the basic procedure discussed in the first part of Section 2.6. The set of n elements is set aside, and the next n elements from the remaining $(k - 1)n$ are identified. This continues until k subsets have been formed; again the pairwise optimization technique is used to improve on this partition.

This method can make an error in the identification of the first set which will bias the choice of the second, and so on; the effect is most severe for the case where k is large, so each set is small.

The method of breaking off subsets sequentially has another potential flaw: regardless of the starting configuration, it will identify approximately the same set each time it is used on a particular problem, and hence little is gained by using it twice on one cost matrix. However

variations in the order of performing pairwise optimizations can still produce different final partitions in general.

Limited computational experience with sequential break-off followed by pairwise optimization suggests that it yields solutions which are on the average at least as good as (and sometimes slightly better than) those provided by pairwise optimization applied to an arbitrary k-way starting partition. Pairwise optimization yields the optimum with a higher probability, however, because it is less susceptible to error caused by a bad choice made early. For instance, in tests on the 80 point matrix mentioned previously, sequential break off yielded 4-way solutions with a mean value of 358.6, but the lowest value found was 355. (The highest was 363.) These may be compared to 359.5, 352 and 365 for the standard partitioning method.

Running time for the sequential break-off method is lower than for straight pairwise optimization.

Insufficient data is available for a direct comparison between sequential break off and the method of repeated subdivision.

In all cases, the original process, be it a completely random generation of some initial configuration, or the production of a good starting partition, is followed by a pairwise optimizing phase. It is unlikely that using better starting partitions will lead to worse results than random starts, on the average. Whether the possible improvement in results and running times will justify the extra computational effort required to generate the starting partition depends on the characteristics of the particular class of matrices being studied.

Some limited experiments were performed to compare the present procedure with a multi-dimensional scaling technique[4], on a Boolean matrix of 316 points, with about 1400 nonzero entries. The results indicated that the procedure identifies clusters well, even when no attempt is made to provide a good starting partition.

3.3 *Expansion Factor*

The introduction of dummy elements was mentioned in Section 2.6 as a method of handling partitioning into subsets of unequal sizes. This can be viewed equally well as a means of introducing "slack" into a solution, in an attempt to get a lower overall cost by allowing "expansion." That is, so far we have treated the problem of finding a partition with a constraint on the sizes of the subsets, *and* on the number of subsets, since given kn points, we have tried to find the best partitions into exactly k subsets of n points each. Suppose we now relax this second constraint by permitting the addition of dummy elements to in-

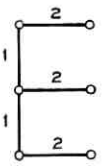


Fig. 3 — Cost reduction by expansion.

crease the size of the problem, and attempt to find the best solution involving *any number* (greater than or equal to k) of subsets, with *at most* n points in each. This solution with k or greater subsets will in general have a lower cost than the constrained solution.

Figure 3 shows an example in which introducing slack permits a lower overall cost. Assume n is 3 and all nodes are size 1. The vertical edges have cost 1 and the horizontal ones cost 2. Any partition into 2 equal subsets has a cost of at least 3, but there is an obvious partition into 3 subsets with cost 2. Any nontrivial partition into 4 or more subsets has a cost greater than 2, so 3 subsets represents the optimal expansion. It is possible to find the minimal cost solution and the corresponding optimal amount of expansion as follows. Suppose the problem has kn points to be partitioned into k sets of n points each. Starting with no slack (kn points), the optimal assignment is found. Then n dummies, enough to create one extra subset, are added, making a $(k + 1)n$ problem, and so on. Eventually, one subset is produced which consists entirely of dummies. When this occurs, we take the partition with this set of dummies removed as our optimum solution.

REFERENCES

1. Ford, L. R., and Fulkerson, D. R., *Flows in Networks,* Princeton, New Jersey: Princeton University Press, 1962, p. 11.
2. Lin, S., "Computer Solutions of the Traveling Salesman Problem," B.S.T.J., *44,* No. 10 (December 1965), pp. 2245–2269.
3. Kernighan, B. W., "Some Graph Partitioning Problems Related to Program Segmentation," Ph.D. Thesis, Princeton University, January 1969, pp. 74–126.
4. Kruskal, J. B., Multi-Dimensional Scaling by Optimizing Goodness of Fit to a Non-Metric Hypothesis," *Psychometrika, 29,* No. 1 (March 1964), pp. 1–27, and No. 2 (June 1964), pp. 115–129.

Laser Speckle Pattern— A Narrowband Noise Model

By CHRISTOPH B. BURCKHARDT

(Manuscript received September 29, 1969)

We represent by an electrical model the imaging of a one-dimensional coherently illuminated and diffusely reflecting surface by an optical system with a rectangular aperture. We then obtain the statistical properties of the image intensity from the statistical properties of the square of the envelope of a narrowband noise signal in the electrical model. The analysis is simple because use can be made of results known in communication theory. The results agree with those obtained in a direct way.

I. INTRODUCTION

The speckle pattern in the image of a coherently illuminated and diffusely reflecting object has been analyzed by Enloe[1]. Enloe's results show a remarkable similarity to some results occurring in the theory of narrowband noise (See Ref. 2, pp. 397 ff.). This similarity prompts the question whether Enloe's results can be derived by the use of an electrical model analogy involving narrowband noise.* In this paper we will show that this is indeed possible for the special case of a one-dimensional subject and an optical system with a rectangular aperture. The analysis is simple because we can use results known in communication theory and the results agree with those of Enloe[1].

II. THE OPTICAL MODEL

The optical model is shown in Fig. 1. In plane P_1 there is a coherently illuminated row of scatterers which scatter with random phase (a one-dimensional diffusely reflecting surface). At a distance d in plane P_2 there is a lens. In front of the lens there is a rectangular aperture with an amplitude transmission $H(x_2, y_2)$. The distance d is large compared to the focal length f of the lens and therefore to a good degree of approxima-

* Such an analogy was already suggested by Rigden and Gordon.[3]

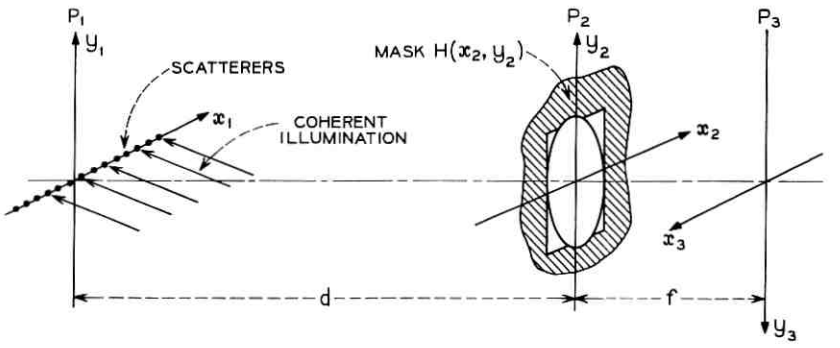


Fig. 1 — Optical model.

tion the image of the scatterers in plane P_1 is situated in plane P_3 at a distance f from the lens. Enloe computed the intensity as well as the autocorrelation of the intensity in the image plane.[1]

As a help in understanding the following electrical model we make the following comments regarding the optical model. Suppose that the instantaneous value of the electric field $e(x_1, t)$ in the object plane P_1 is given by

$$e(x_1, t) = a\left(\frac{x_1}{\lambda d}\right) \exp(-2\pi j\nu_0 t). \tag{1}$$

$e(x_1, t)$ is zero for $y_1 \neq 0$. The time-independent phasor is

$$a\left(\frac{x_1}{\lambda d}\right) = a(\alpha_1), \tag{2}$$

where we have introduced the spatial frequency coordinate

$$\alpha_1 = \frac{x_1}{\lambda d}. \tag{3}$$

(We write a( ) as a function of the spatial frequency $x_1/\lambda d$ and not of x_1 in order to simplify the following computation.) For later use we also introduce

$$\beta_1 = \frac{y_1}{\lambda d}. \tag{4}$$

Since the lens is situated in the far field of the object, the phasor $A(x_2, y_2)$ in plane P_2 in front of the aperture is the Fourier transform of $a(\alpha_1)$,

$$A(x_2, y_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(\alpha_1) \exp(2\pi j\alpha_1 x_2) \exp(2\pi j\beta_1 y_2)\, d\alpha_1\, d\beta_1 . \tag{5}$$

Since $a(\alpha_1)$ is zero for $\beta_1 \neq 0$ the above two-dimensional Fourier transform operation reduces to the one-dimensional Fourier transform operation

$$A(x_2 , y_2) = \int_{-\infty}^{\infty} a(\alpha_1) \exp (2\pi j \alpha_1 x_2) \, d\alpha_1 . \tag{6}$$

It is seen that $A(x_2 , y_2)$ is a function of x_2 only and we will therefore write $A(x_2)$. The phasor behind the aperture, $B(x_2 , y_2)$ is obtained as the product of $A(x_2)$ and the aperture transmission function $H(x_2 , y_2)$

$$B(x_2 , y_2) = A(x_2) \cdot H(x_2 , y_2). \tag{7}$$

As was mentioned we assume a rectangular aperture function. We assume that $H(x_2 , y_2)$ can be written as

$$H(x_2 , y_2) = H_1(x_2) \cdot H_2(y_2). \tag{8}$$

Since our object is one-dimensional we are only interested in the image intensity at $y_3 = 0$. The image intensity at $y_3 = 0$ is not influenced by $H_2(y_2)$ except for a factor which remains constant over all x_3. For equation (7) we can therefore write

$$B(x_2) = A(x_2) \cdot H_1(x_2) \tag{9}$$

with the understanding that $B(x_2)$ does vary in the y_2 direction but that this variation is of no interest to us. So much for the optical model. We will now present the electrical model.

III. THE ELECTRICAL MODEL

The electrical model is shown in Fig. 2. The electric field in the object plane is scanned by a detector whose output voltage is proportional to the instantaneous value of the electric field in the object plane. (In the optical region there are no such detectors available. This should not present any conceptual difficulties since we can scale up our model to a longer wavelength where there are such detectors.)

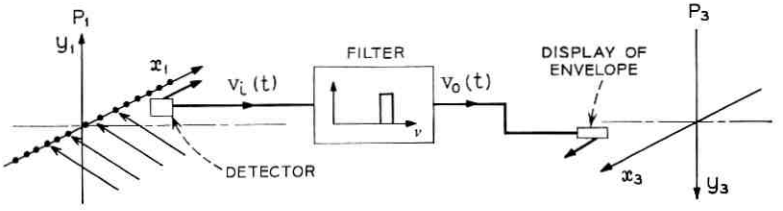


Fig. 2 — Electrical narrowband noise model.

The electric field $e(x_1, t)$ in plane P_1 is given by equation (1) and the output voltage $v_i(t)$ of the scanning detector is

$$v_i(t) = c_1 a\left(\frac{v_1 t}{\lambda d}\right) \exp(-2\pi j \nu_0 t). \tag{10}$$

v_1 is the scanning velocity and we have

$$x_1 = v_1 t. \tag{11}$$

c_1 denotes a constant of proportionality which is of no interest. The Fourier transform of $v_i(t)$ of equation (10), $FT[v_i(t)]$ is given by

$$\begin{aligned} FT[v_i(t)] &= c_1 \frac{\lambda d}{v_1} A\left[\frac{\lambda d}{v_1}(\nu - \nu_0)\right] \\ &= c_2 A\left[\frac{\lambda d}{v_1}(\nu - \nu_0)\right]. \end{aligned} \tag{12}$$

$A(\nu)$ is the Fourier transform of $a(t)$. c_2 is a constant of proportionality. The Fourier transform or spectrum of $v_i(t)$ is centered at ν_0 as shown in Fig. 3. In the optical model the spectrum $A(x_2)$ is multiplied by the aperture transmission function $H_1(x_2)$ in plane P_2. See equation (9). In order to simulate this in our electrical model we have to pass the voltage $v_i(t)$ through a temporal frequency filter with the frequency response $H_1[(\lambda d/v_1)(\nu - \nu_0)]$. The spectrum $B[(\lambda d/v_1)(\nu - \nu_0)]$ at the output of the filter is then given by

$$B\left[\frac{\lambda d}{v_1}(\nu - \nu_0)\right] = A\left[\frac{\lambda d}{v_1}(\nu - \nu_0)\right] H_1\left[\frac{\lambda d}{v_1}(\nu - \nu_0)\right]. \tag{13}$$

The filter $H_1[(\lambda d/v_1)(\nu - \nu_0)]$ is shown in the electrical model of Fig. 2 and its frequency response is sketched in Fig. 3. (The frequency response $H_1[(\lambda d/v_1)(\nu - \nu_0)]$ in the electrical model corresponding to the rectangular aperture transmission function of the optical system can only be

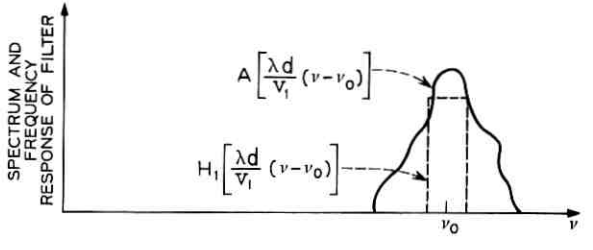


Fig. 3 — Spectrum of the electrical signal and frequency response of the filter.

realized with a time lag, but this need not disturb us.) At the output of the filter we have the voltage $v_o(t)$

$$v_o(t) = c_3 b\left(\frac{v_1 t}{\lambda d}\right) \exp(-2\pi j \nu_0 t), \tag{14}$$

where $b(t)$ is the Fourier transform of $B(\nu)$. c_3 again is a constant of proportionality of no interest. The output device scans the image plane P_3 with a velocity v_2 and

$$\frac{v_2}{v_1} = \frac{f}{d} \tag{15}$$

because in the optical model the image is demagnified by a factor d/f with respect to the object. The inversion of the optical image with respect to the object is accounted for by inverting the coordinate system in the image plane P_3, see Fig. 1. We can now write for equation (14)

$$\begin{aligned} v_o(t) &= c_3 b\left(\frac{v_2 t}{\lambda f}\right) \exp(-2\pi j \nu_0 t) \\ &= c_3 b\left(\frac{x_3}{\lambda f}\right) \exp(-2\pi j \nu_0 t). \end{aligned} \tag{16}$$

In the optical model we detect the intensity in the image plane, that is, the square of the absolute value of the phasor. Therefore in our electrical model the output device displays $| b(x_3/\lambda f) |^2$.

The voltage $v_o(t)$ at the output of the filter is narrowband compared to ν_0. We now consider $v_o(t)$ as a narrowband noise voltage. Another equivalent representation is (see, for example, Ref. 2, p. 397)

$$v_o(t) = N_c(t) \cos 2\pi\nu_0 t + N_s(t) \sin 2\pi\nu_0 t, \tag{17}$$

where

$$\begin{aligned} N_c(t) &= \mathrm{Re}\left[b\left(\frac{v_2 t}{\lambda f}\right)\right], \\ N_s(t) &= \mathrm{Im}\left[b\left(\frac{v_2 t}{\lambda f}\right)\right]. \end{aligned} \tag{18}$$

Re [ ] means "real part of" and Im [ ] means "imaginary part of". We now assume that the filter is so narrowband that its impulse response is much wider than the time over which $v_i(t)$ shows any appreciable correlation. (This assumption is the equivalent of Enloe's assumption that the optical spread function is much wider than the average distance between the images of scatterers with independent phase.) If this

assumption holds, the voltage $v_o(t)$ results from many independent values of $v_i(t)$. According to the central limit theorem $v_o(t)$ and therefore $N_c(t)$ and $N_s(t)$ show a Gaussian distribution. The voltage $v_o(t)$ therefore has the statistical properties of narrowband Gaussian noise which are discussed for example in Ref. 2, pp. 397 ff. As was mentioned, the square of the envelope corresponds to the optical intensity. The envelope $E(t)$ is obtained as

$$\begin{aligned} E(t) &= [N_c^2(t) + N_s^2(t)]^{\frac{1}{2}} \\ &= \left(\left\{\mathrm{Re}\left[b\left(\frac{v_2 t}{\lambda f}\right)\right]\right\}^2 + \left\{\mathrm{Im}\left[b\left(\frac{v_2 t}{\lambda f}\right)\right]\right\}^2\right)^{\frac{1}{2}}. \end{aligned} \tag{19}$$

According to Ref. 2, equation (9.18), $E(t)$ has a Rayleigh probability distribution density $W(E)$

$$W(E) = \frac{E \exp(-E^2/2\psi)}{\psi}. \tag{20}$$

The average value of the square of the envelope $\langle E^2\rangle_{av}$ which is equal to the average value of the optical intensity $\langle I\rangle_{av}$ is given by [Ref. 2, Eq. (9.21a)]

$$\langle E^2\rangle_{av} = \langle I\rangle_{av} = 2\psi. \tag{21}$$

We are now also interested in the autocorrelation $R(s)$ of the intensity I

$$\begin{aligned} R(s) &= \langle I(x_3)I(x_3 + s)\rangle_{av} \\ &= \langle E^2(x_3)E^2(x_3 + s)\rangle_{av}. \end{aligned} \tag{22}$$

According to Ref. 2, equation (9.24), $R(s)$ is given by the following expression

$$\begin{aligned} R(s) &= 4\psi^2[1 + k_0^2(s)] \\ &= \langle I^2\rangle_{av}[1 + k_0^2(s)], \end{aligned} \tag{23}$$

where we have used equation (21). As explained in Ref. 2, [equations (9.10b and 9.12b)] $k_0(s)$ is the autocorrelation of the spread function corresponding to a filter with the frequency response $H_1((\lambda d/v_1)\nu)$. This is the frequency response of the filter in our electrical model (see Fig. 2), *but centered at zero frequency*. The frequency response $H_1((\lambda d/v_1)\nu)$ is shown in Fig. 4. $k_0(s)$ is normalized to one at $s = 0$. (The above discussion holds when the frequency response of the filter is symmetrical about the origin as is true for our case.) The spread function is the

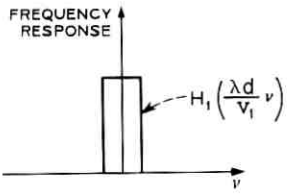


Fig. 4 — Frequency response of the filter $H_1[(\lambda d/v_1)\nu]$.

Fourier transform of the frequency response and so we have

$$FT\left[H_1\left(\frac{\lambda d}{v_1}\nu\right)\right] = c_4 h\left(\frac{v_2 t}{\lambda f}\right) \tag{24}$$

where $H_1(\nu)$ and $h(t)$ are Fourier transform pairs. c_4 is a constant of no interest. Therefore we obtain for $k_0(s)$

$$\begin{aligned} k_0(s) &= \frac{1}{\rho(0)} \int_{-\infty}^{+\infty} h^*\left(\frac{v_2 t}{\lambda f}\right) h\left(\frac{v_2 t}{\lambda f} + \frac{v_2 \tau}{\lambda f}\right) dt \\ &= \frac{\rho\left(\frac{v_2 \tau}{\lambda f}\right)}{\rho(0)} = \frac{\rho\left(\frac{s}{\lambda f}\right)}{\rho(0)}, \end{aligned} \tag{25}$$

where

$$\rho(u) = \int h^*(t) h(t + u)\, dt. \tag{26}$$

Using equations (23) and (25) we finally obtain for $R(s)$

$$R(s) = \langle I^2 \rangle_{av} \left[1 + \frac{\rho^2\left(\frac{s}{\lambda f}\right)}{\rho^2(0)}\right]. \tag{27}$$

This last equation is the same as Enloe's equation (14) if the last term in that equation is disregarded. The last term in Enloe's equation (14) vanishes if the spread function of the filter is much broader than the average distance between images of scatterers with independent phase. The power spectrum follows from the autocorrelation function in the same way as in Enloe's analysis and this derivation will not be repeated here.

In summary: We have derived the statistical properties of the intensity in the image of a one-dimensional coherently illuminated and diffusely reflecting subject with the help of a narrowband noise model. Use was made of results known in communications theory.

REFERENCES

1. Enloe, L. H., "Noise-like Structure in the Image of Diffusely Reflecting Objects in Coherent Illumination," B.S.T.J., *46*, No. 7 (September 1967), pp. 1479–1489.
2. Middletown, D., *An Introduction to Statistical Communication Theory,* New York: McGraw-Hill Book Co., 1960.
3. Rigden, J. D., and Gordon, E. I., "The Granularity of Scattered Optical Maser Light," *Proc. I.R.E., 50,* No. 11 (November 1962), pp. 2367–2368.

Contributors to This Issue

CHRISTOPH B. BURCKHARDT, Dipl.-Ing., 1959, Dr. sc. techn., 1963, Swiss Federal Institute of Technology; Bell Telephone Laboratories, 1963—. Initially Mr. Burckhardt was engaged in the analysis of varactor frequency multipliers. Since 1965, he has been working in holography. Member, IEEE, Optical Society of America.

SHLOMO HALFIN, M.Sc., 1958, and Ph.D., 1962, The Hebrew University of Jerusalem (Israel); Bell Telephone Laboratories, 1968—. Mr. Halfin is working on theoretical problems in the area of mathematical programming and on the development of special purpose optimization algorithms. Member, American Mathematical Society, Operations Research Society of America, and the Society for Industrial and Applied Mathematics.

LELAND B. JACKSON, S.B. and S.M., 1963, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1961–62, 1966—. Mr. Jackson was first associated with Bell Laboratories under the M.I.T. cooperative program in electrical engineering. Since 1966 he has been primarily concerned with the analysis and synthesis of digital filters and related systems. He has completed the requirements for the Sc.D. degree from Stevens Institute of Technology. Member, IEEE, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.

NORMAN D. KENYON, B.A., 1963, M.A., 1966, and Ph.D., 1967, Cambridge University, England; Bell Telephone Laboratories, 1968—. Mr. Kenyon is currently working on circuit structures for millimeter-wave IMPATT diodes. Member, IEEE.

BRIAN W. KERNIGHAN, B.A.Sc., 1964, University of Toronto; Ph.D., 1969, Princeton University; Bell Telephone Laboratories, 1969—. Mr. Kernighan is working primarily on graph-partitioning and its applications and extensions. Member, Association for Computing Machinery.

RONALD W. KORDOS, B.E.E., 1957, University of Detroit; M.S.E.E., 1959, Northeastern University; Bell Telephone Laboratories, 1957—. Mr. Kordos has engaged in the design of microwave ferrite devices,

including field-displacement and resonance isolators for several microwave radio relay systems, and a passive power limiter for the *Telstar*® Satellite. In 1964, he transferred to the Columbus Laboratory and was engaged in the development of broadband switching matrices and in a study of microstrip interconnection in high-speed integrated circuits. Since 1968, he has been a member of the Radio Transmission Laboratory at Merrimack Valley involved in the development of microwave integrated circuits. Member, Tau Beta Pi, Eta Kappa Nu.

R. R. Laane, B.S.E.E., 1962, University of Illinois; M.S.E.E., 1964, New York University; Bell Telephone Laboratories, 1962—. Mr. Laane has worked on the application of super-conductive switches to data processing systems and has investigated the application of optical processing techniques to data processing systems. Since 1967, he has been engaged in exploratory work on the application of semiconductor devices to telephone switching networks and on analog-to-digital conversion techniques. He is presently also working on solid-state Picturephone(R) switching networks. Member, IEEE.

Shen Lin, B.S. (summa cum laude), 1951, University of the Philippines; M.A., 1953, Ph.D., 1963, Ohio State University; Bell Telephone Laboratories, 1963—. He has worked in the field of Turing machine theory, combinatorial analysis and number theory. At present, he is working on applications of computers in various optimization and number-theoretic problems. Member, AAAS, American Mathematical Society, Mathematical Association of America, SIAM, Phi Kappa Phi, Sigma Pi Sigma, Pi Mu Epsilon.

Dietrich Marcuse, Diplom Vorpruefung, 1952, Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–57; Bell Telephone Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research, studying coaxial cable and circular waveguide transmission. At Bell Telephone Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He spent one year (1966–1967) on leave of absence from Bell Telephone Laboratories at the University of Utah where he wrote a book on quantum electronics. He is presently working on the transmission aspect of a light communications system. Member, IEEE, Optical Society of America.

MERTON B. PURVIS, B.S., 1944, M.S., 1949, Iowa State College; Ph.D., 1954, Pennsylvania State University; Bell Telephone Laboratories, 1955—. Mr. Purvis worked initially on development of network cooling, optics and photographic aspects of a large computer memory for the Morris Electronic Switching System. He subsequently engaged in research on the magnetic characteristics of ferreed switches used in electronic switching, development work on switches for the 60–90 MHz band and exploratory work on fluidic networks. He currently heads the Apparatus Design Department dealing with relays, crossbar switches and miscellaneous apparatus. Member, American Society of Mechanical Engineers, Sigma Xi, Pi Tau Sigma, American Society for the Advancement of Science.

DAVID C. RIFE, B.S.E.E., 1960, University of Washington; M.E.E., 1962, New York University; Bell Telephone Laboratories 1960—. Mr. Rife has worked on data carrier systems, automatic calling units and data test equipment. Since 1963 he has been supervising the development of data station testing systems. He is working on his Ph.D. at the Polytechnic Institute of Brooklyn. Member, IEEE, Phi Beta Kappa, Tau Beta Pi.

GEORGE A. VINCENT, B.S.E.E., 1966, Newark College of Engineering; M. E. (Electrical), 1968, Stevens Institute of Technology; Bell Telephone Laboratories, 1966—. Mr. Vincent is engaged in the development of data test equipment. He is presently involved in the development of a computer controlled Automatic Data Test Center. Member, IEEE, Tau Beta Pi, Eta Kappa Nu.

Erratum

On page 3440 of the December 1969 *Bell System Technical Journal,* Reference 6, "Comparison of An Energy Density Antenna System with Predetection Combining Systems for Mobile Radio" by W. C. Y. Lee, was mistakenly listed as an unpublished work. That article was published in the *IEEE Trans. on Communication Technology, 17,* No. 2 (April 1969), pp. 277–284.

Erratum

On page 3440 of the December 1969 *Bell System Technical Journal*, Reference 6, "Comparison of An Energy Density Antenna System with Predetection Combining Systems for Mobile Radio" by W. C. Y. Lee, was mistakenly listed as an unpublished work. That article was published in the *IEEE Trans. on Communication Technology, 17*, No. 2 (April 1969), pp. 277–284.