THE BELL SYSTEM
TECHNICAL JOURNAL

Volume 49 March 1970 Number 3

Adaptive Delta Modulation with a One-Bit Memory

By N. S. JAYANT

(Manuscript received September 4, 1969)

We propose a delta modulator which, at every sampling instant r, adapts its step-size (for a staircase approximation to the input signal) on the basis of a comparison between the two latest channel symbols, C_r and C_{r-1}. Specifically, the ratio of the modified step-size m_r to the previous step size m_{r-1} is either $+P$ or $-Q$ depending on whether C_r and C_{r-1} are equal or not. (We recall that, in delta modulation, C_r represents the polarity of the difference, at the sampling instant r, between the input signal X_r and the latest staircase approximation to it, Y_{r-1}.)

A simulation of the delta modulator with a band-limited speech input has revealed that $PQ = 1$ and $P \simeq 1.5$ represent optimal adaptation characteristics, on the basis of signal-to-error ratios, over an important range of sampling frequencies; and that at 60 kHz, delta modulation with these adaptation parameters compares favorably with 7-bit logarithmic PCM, which reproduces speech with good telephone quality. We present several graphical results from this simulation, and include an evaluation of the effect of independent channel errors on the adaptive delta modulator.

We proceed to suggest a heuristic theory of the delta modulator which explains the optimality of the condition $PQ = 1$, and develops an upper bound of 2 for the optimum value of P.

We conclude with a summary of results from a video simulation which revealed that aforementioned optima for P and Q apply to a video signal *as well; with these optimum parameters, a useful delta-modulator output was obtained at* 10 MHz *operation.*

The results of this paper reaffirm the utility of delta modulation as a simple alternative to PCM, *particularly in systems that operate at relatively low bit-rates.*

I. INTRODUCTION

Linear (or unadaptive) delta modulators, which work with a fixed step-size for the "staircase" approximation to an input signal, have the following basic limitation. Small values of the step size introduce slope-overload distortion during bursts of large signal slope; large values of the step-size accentuate the granular noise during periods of small signal slope; and, even when the step-size is optimized, the performance of these modulators will be satisfactory only at sampling frequencies that may be undesirably high. Equivalently, one encounters important ranges of operating frequency in which the performance of conventional delta modulation falls short of the standards attainable by conventional PCM or by d-level differential PCM, of which delta modulation is a special case ($d = 2$).

With a view to employing delta modulation (which is inherently a very simple signal-processing strategy) at such relatively low operating frequencies, several types of adaptive delta modulation have been proposed.[1-3] In these schemes, the step size is changed in accordance with the time-varying slope characteristics of the input signal, as per a predetermined adaptation strategy. Such adaptation or "companding" can be either at a syllabic rate (long-term) or instantaneous (short-term).

Typical of syllabic-companding delta modulators are recently developed schemes for reproducing telephone quality speech at operating frequencies of the order of 50 kHz.[1,5,6] These systems are characterized by "continuous" adaptation of the step magnitude. Instantaneous compandors, on the other hand, usually incorporate discrete adaptations, and illustrative schemes for speech, television and Gaussian signals are given in Abate[2] and, for speech transmission, in Winkler.[3] Abate shows the capabilities of linear and exponential adaptation for speech transmission, but gives quantitative results only for specific, finite, step-size dictionaries. Likewise, Winkler's work on "High Information Delta Modulation," while providing a conceptual basis for our paper, bypasses the question of optimal adaptation. We consider in this paper, although only for a sub-class of possible schemes, the problem of optimizing the adaptation logic.

We ought to refer here to the paper entitled "Statistical Delta Modulation" by Bello, and others.[4] Philosophically, this paper treats the problem of optimizing delta modulation with a generality that exceeds the scope of our work. However, the analysis of the cited paper does not have explicit bearing on the design philosophy for the very specific, but practically important, problem of providing a time-invariant logic for step-size adaptation. The purpose of our paper is to treat the latter problem for the important case of a one-bit memory.

We begin by defining our adaptation scheme (Section II), and go on to present results from a computer simulation of the delta modulator with a speech input (Section III). The results refer to the optimization of the adaptation logic, to a comparison of the optimal delta modulator with PCM, and to an assessment of the effect of channel errors on the delta modulator. We then present a heuristic theory (Section IV) for the delta modulator and seek to explain the optimal adaptation parameters that emerged from the speech simulation. Finally, we illustrate parallel results from a video simulation (Section V) and attempt a general assessment of adaptive delta modulators (Section VI).

II. DESCRIPTION OF THE ADAPTIVE DELTA MODULATOR

In this section, we define the delta modulator with exponential adaptation and a one-bit memory, and indicate its basic performance by illustrating its response to a constant input.

2.1 *The Adaptation Logic*

The delta modulator of this paper uses instantaneous, exponential adaptation in the sense that the step-size is changed at every sampling instant by a specific factor—more precisely, by one of two specific factors. Furthermore, the adaptation logic incorporates a one-bit memory in that the immediately past channel symbol C_{r-1} is stored, and is compared with the incoming bit C_r for a decision on the new step-size m_r. Specifically, if the previous step-size is denoted by m_{r-1}, the adaptation will be of the form

$$\begin{aligned} m_r &= P \cdot m_{r-1} \quad &\text{if} \quad C_r = C_{r-1}\,; \\ m_r &= -Q \cdot m_{r-1} \quad &\text{if} \quad C_r \neq C_{r-1}\,. \end{aligned} \tag{1}$$

In this paper, we assume that P and Q are time-invariant, and note that in delta modulation, the following identity is usually assumed by definition:*

* See Ref. 7 for an example where requirement (2) is waived.

$$\operatorname{sgn} m_r = C_r = \operatorname{sgn} (X_r - Y_{r-1}) \tag{2}$$

where X_r and Y_{r-1} represent the amplitude of the input signal and that of the latest staircase approximation to it, respectively, at the sampling instant r. The sampling interval in question would be a suitably small fraction of the Nyquist interval for X. A block diagram of the modulator appears in Fig. 1.

2.2 *Simple Bounds on P and Q*

The crucial parameters of our delta modulator are the time-invariant adaptation constants P and Q. The smallest and largest allowable step-sizes are other important parameters, but we assume that their design can be treated as an independent problem; and we mention at suitable points in the paper the considerations which influence such design. We now proceed, therefore, to state two simple bounds on the adaptation parameters P and Q:

(*i*) In order to adapt to the signal during slope overload, it is necessary that

$$P > 1. \tag{3}$$

(*ii*) In order to converge to a constant input signal during a purely "hunting" situation ($m_r = -Qm_{r-1}$ with probability 1), it is necessary that

$$Q < 1. \tag{4}$$

Notice that $P = Q = 1$ represents (conventional) linear delta modulation.

The adaptation logic of Section 2.1 represents the simplest nontrivial form of discrete exponential adaptation, and the performance of this scheme will be an important lower bound for that of an "n-bit" strategy

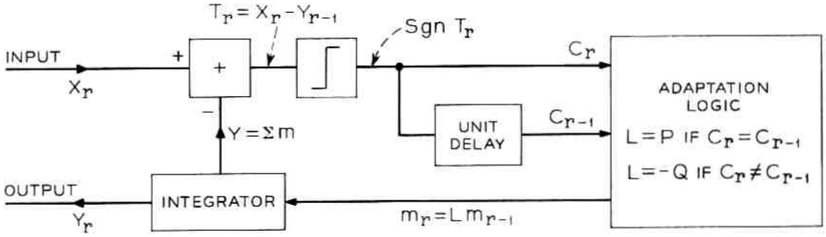

Fig. 1.— Schematic diagram of the Adaptive Delta Modulator.

($n > 2$) in which the step-size m_r is some optimal function of C_r, $C_{r-1}, \cdots, C_{r-n+1}$ and of the previous x step-sizes $m_{r-1}, m_{r-2}, \cdots, m_{r-x}$.[8]

2.3 *Step Response of the Delta Modulator*

Figure 2 shows the approximation of a step function by our adaptive delta modulator for a typical case of $P = 1.50$ and $Q = 0.66$. (These will emerge as optimum parameters later in the paper.) Step inputs of 9, 10 and 12 units have been considered for illustration, with a smallest step-size of 1 for the delta modulator.

The dependence of the "hunting" or "oscillating" characteristics on the actual magnitude of the step input is clear. We also see that during hunting, the step-size does not always assume the smallest possible value. This is an inherent feature of our adaptation logic, and emphasizes the need to make the smallest step-size as small as is practicable so that the in-band component of the noise due to hunting with nonminimal step-sizes will be tolerably low.

III. PERFORMANCE WITH A SPEECH INPUT

We describe in this section several results from a simulation of the adaptive delta modulator of Section II with a speech input. In particular, we highlight the optimization of the adaptation parameters P and

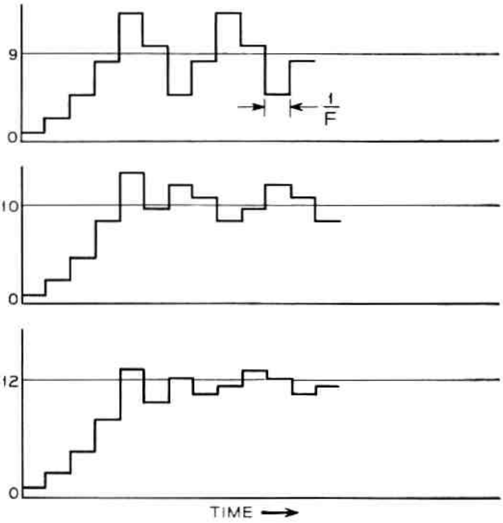

Fig. 2 — Step response of the Delta Modulator.

Q, and the relative performance of adaptive delta modulation (abbreviated henceforth as ADM) and of logarithmic PCM.

3.1 *Description of the Simulator*

The input speech signal used for the simulation was a male utterance of "Have you seen Bill?", bandlimited to 3.3 kHz, and sampled for the simulation at 20, 40, and 60 kHz. The sentence is illustrative in that it includes sounds which are known to be susceptible to slope-overload distortion.

In the computer simulation, the peak-to-peak range of the 12-bit speech signal was 4096 units. The configuration of the step-size dictionary was not predetermined, and the changes of the step-size were allowed to follow the exponential adaptation rule (*i*) of Section II. The simulation started with an initial step-size magnitude of 1 unit and it may be mentioned that step-size magnitudes as large as 380 units were typically encountered in the simulation. A histogram of utilized step-sizes for the typical case of $P = 1.50$, $Q = 0.66$ is illustrated in Fig. 3, and represents mean step magnitudes of the order of 30 units. For the special case of $P = Q = 1.0$, the constant step-size was selected to maximize a performance criterion to be defined presently. The step-size so optimized was approximately 80, 60 and 45 units for sampling frequencies of 20, 40 and 60 kHz respectively.

The simulation used an ideal integrator in the feedback loop of the delta modulator;* it also incorporated a nonrecursive low pass filter using a Fourier kernel, which was designed to have a 40 dB attenuation from 3 kHz to 3.3 kHz. Practical low pass filters may have to be sloppy in comparison, but the sharp filter was included in the simulation for a correct assessment of the modulator performance, and for comparison with Nyquist-rate PCM.

3.2 *Definition of a Signal-to-Noise Ratio G*

The basic purpose of the simulator was to study the performance of the delta modulator as a function of the adaptation parameters P and Q, and the sampling frequency F. The quality criterion which was adapted was an "objective signal-to-noise ratio" G, which was defined as the ratio of the power of the signal X_r to that of the error $E_r = X_r - Y_r$, averaged over the duration of the speech sample.

It is seen that no distinction was made between overload distortion and hunting noise in defining G. In adaptive delta modulation, instan-

* See Section 3.9 for a reference to the utility of leaky integrators for delta modulation in the presence of channel errors.

taneous companding is expected to render long bursts of one particular type of distortion very improbable; the total error power, defined as the summation of E_r^2 over the overload and hunting phases, was therefore adopted as a good measure of performance. As a matter of fact, in the absence of a better criterion, the same measure has been assumed in this paper for the nonadaptive case as well; and the credibility of the procedure has been borne out by the observation of a good correlation between the subjectively assessed quality of representative speech reproductions and the corresponding values of G.

3.3 *Stability of the Modulator*

Preliminary studies of stability revealed the significance of the product PQ, and the adaptation was seen to be inherently unstable (that is, resulting in a step-size oscillation between limits that were independent of the input) if PQ exceeded $(1 + \epsilon)$ where ϵ is positive, and much smaller than unity. Further studies of performance therefore

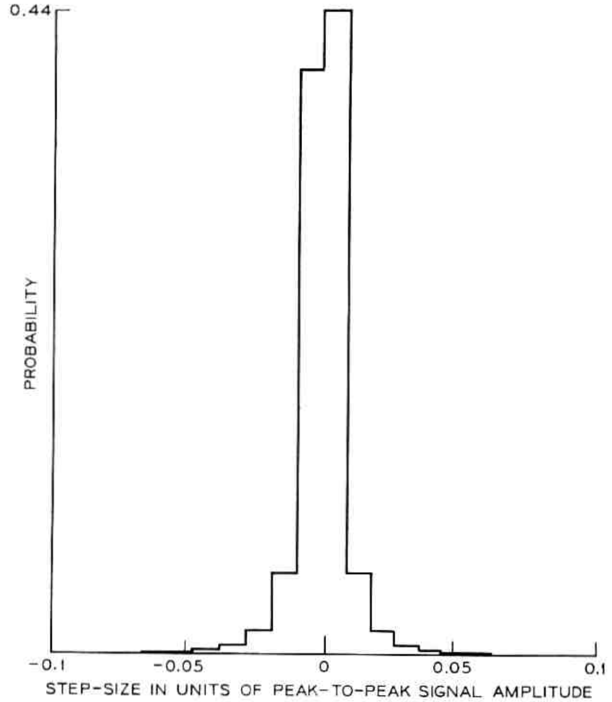

Fig. 3 — Histogram of utilized step sizes in the speech simulation. ($F = 60$ kHz)

assumed as stability condition of the form

$$PQ \leq 1. \tag{5}$$

The signal-to-noise ratio G was then studied as a function of allowable values of PQ, of P and of the sampling frequency F.

3.4 *The Dependence of G on PQ*

Using a typical value of $P = 1.6$, Fig. 4 shows the behavior of G as a function of PQ, with F as a parameter. The value of $G = -\infty$ (dB) at $PQ = 1.1$ represents an example of unstable adaptation, and the monotonic rise of G with PQ in its stable range is evident; in conjunction with the condition (5) in Section 3.3, it follows that

$$PQ = 1 \tag{6}$$

represents an optimal condition for all F; this conclusion was verified to be independent of the value of P.

Notice that (6) also represents a very desirable condition from the point of view of implementation. This is because the reciprocity of P and Q facilitates the use of a compact step-size dictionary. Finally, note that condition (6) is obviously satisfied in conventional delta modulation.

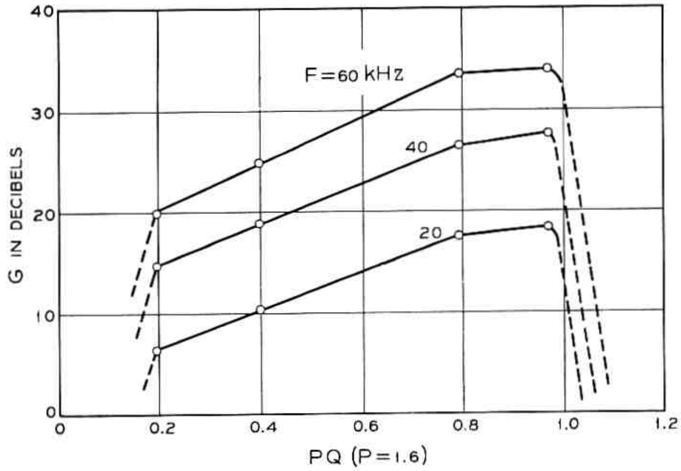

Fig. 4 — Results of the speech simulation: signal-to-error ratios as functions of (PQ).

3.5 *The Dependence of G on P*

Assuming the optimal reciprocity condition (6), the variation of G with P was investigated, and the results are given in Fig. 5. The "flat" nature of the $G - P$ curves in the region of their maxima is noteworthy,* and the fact that the optimum value

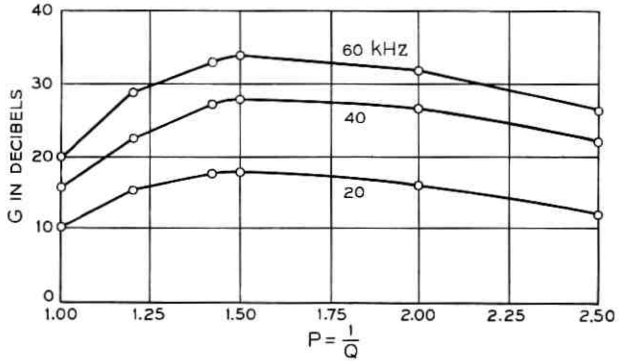

Fig. 5 — Results of the speech simulation: signal-to-error ratios as functions of P.

$$P_{\text{opt}} \simeq 1.5 \tag{7}$$

is nearly independent of F is quite striking. Furthermore, the improvement that the optimized adaptive delta modulator affords, over the conventional system ($P = 1$), is seen to be an increasing function of F; and for 60 kHz operation, the gain exceeds 10 dB.

3.6 *A Note on Implementation*

The step-sizes in our simulation were real-valued quantities which changed according to equation (1). In a practical implementation, it may be preferable to work with integer-valued step-sizes; or, equivalently, to employ a suitably discretized step-size dictionary; and to avoid the actual operation of analogue multiplication. Such multiplication could pose significant practical problems. For example, the values of the multipliers P and Q may be subject to random perturbations about their design values, and these fluctuations may be independent at the encoder and at the decoder. Preliminary simulations that incorporated such imperfect multipliers suggest that the attendant deterioration of delta-modulator performance may well justify a mandatory

* For a corresponding observation with the adaptation logic described in Ref. 8, see Fig. 12 in that reference.

use of a discretized step-size dictionary. The design procedure for such a dictionary is seen to be greatly facilitated by virtue of the reciprocity condition for PQ, and the broad optima for P. The criteria for selecting the minimum and maximum step-sizes have been mentioned elsewhere in this paper, and the intermediate (discrete) step-sizes can be chosen to fit the optimum condition (7) as closely as possible, through the range of the dictionary. A further simplification will result if the slightly suboptimal value $P = 2$ is adopted as a uniform adaptation parameter.

3.7 *Subjective Performance*

Formal subjective tests of performance have not been carried out. However, the optimum ADM ($P = 1.50$, $Q = 0.66$) achieves very good telephone quality at 60 kHz, and the degradation at 40 kHz is very small. The ADM deteriorates in quality at 20 kHz operation, though most of the intelligibility of speech is still preserved.

3.8 *Comparison of* ADM *and Logarithmic* PCM

Table I shows the objective signal-to-noise ratio G for the optimum ADM at $F = 20$, 40 and 60 kHz; and, for n-bit logarithmic PCM at the Nyquist rate, the three values of n which provide correspondingly equal values of G. The PCM figures are due to the theory of Smith,[9] and represent average values over the significant range ($100 < \mu < 1000$) of his logarithmic-companding parameter μ. Furthermore, the PCM figures refer to the "strong-signal" or "full-load" case ($C \rightarrow 0$) in Smith's theory; inasmuch as our delta modulator could handle arbitrarily strong signals, according to equation (1), the "full-load" values for PCM performance were adopted as meaningful measures for our comparison.

It is generally accepted that 7-bit log-PCM represents a good quality of speech reproduction. It would therefore appear, from Table I, that a sampling frequency in the range of 40 to 60 kHz would be a critical figure for the employment of instantaneously companding ADM to reproduce telephone quality speech. This is an important conclusion of this paper, and follows a similar claim for a syllabic-companding delta modulator for speech at 56 kHz operation.[1]

Figure 6 replots the results of Table I, depicting G as a function of

TABLE I — COMPARISON WITH LOGARITHMIC PCM

ADM sampling rate: F(kHz)	20	40	60
ADM Performance: G(dB)	18	28	34
Equivalent log-PCM bits: n	4.7	6.3	7.3

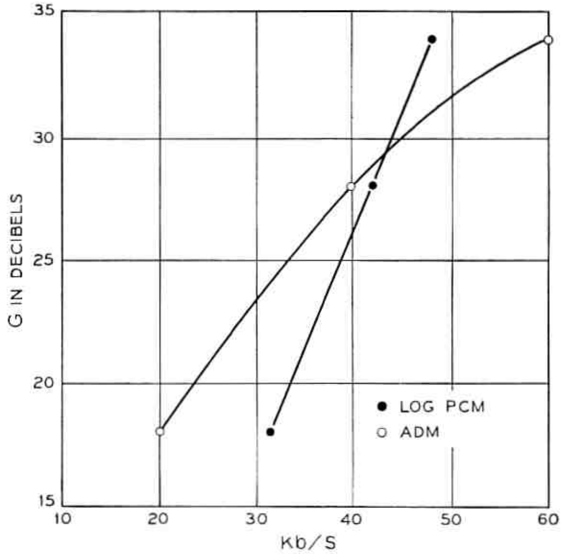

Fig. 6 — Comparison of Adaptive Delta Modulation and Logarithmic PCM: signal-to-error ratios as functions of bit-rate.

the bit-rate (product of the sampling frequency and the number of bits per channel symbol). The bit-rate is equal to the product (6600.n) for n-bit log-PCM at the Nyquist rate, and is equal to the sampling frequency F for ADM. The crossover of the curves in Fig. 6 at about 40 KBPS is significant.* It suggests an important, though narrow, range of usable bit-rates where ADM, which was conceived originally only for its simplicity, can actually excel conventional log-PCM for speech transmission.

3.9 *The Effect of Independent Channel Errors*

We conclude the section on ADM simulation with speech by presenting a qualitative discussion of the effect of independent channel errors on the performance of the delta modulator.

When such errors were first allowed in the simulation, deterioration of ADM performance was observed at error rates as low as 1 in 10^5. This was expected because of the inherent susceptibility of ADM to channel errors; every such error will have the effect of producing a long sequence of erroneous or suboptimal step-sizes which integrate in the output.

* Crossovers of this type are indicated in Ref. 2 for television signals, and in Ref. 10, for Gaussian signals with an integrated spectrum.

In order to reduce the noise-memory of the ADM, the ideal integrator in the ADM simulator of Section 3.1 was replaced by a counterpart that had a finite time constant of the order of 10 to 20 sampling intervals. Furthermore, to mitigate the instability arising out of incorrect step-size adaptation in a noisy situation, the maximum allowable step-size was limited to a suitable value* (that would not introduce noticeable slope overload in the noiseless case). As a result of these refinements, the tolerable probability of channel errors was raised from about 1 in 10^5 to a figure of the order of 1 in 10^4. In fact, the intelligibility of ADM speech was still very much preserved at error rates of the order of 1 in 10^3, but the quality of the output was affected by "clicks" that were introduced by the channel errors.

An additional parameter that has a potential for enhancing the noise-resistance of ADM would be the length of the bit-memory in the adaptation scheme. The simple adaptation used in this paper has a minimal, one-bit memory and a suitably longer memory could indeed decrease the noise-susceptibility of ADM by a useful factor. In the ultimate analysis, however, it should be clear that such noise-susceptibility is a general limitation of all classes of ADM, because of the integrator employed in these systems; and this observation will be a very important factor in the assessment of adaptive delta modulation with reference to PCM for use on specific communication channels.

It may not be out of place to comment on the effect of transmitter-receiver mistracking on delta-modulator performance. In general, "mistracking" would characterize a situation where the step-size sequence in the receiver tracks that at the transmitter only in polarities and adaptation ratios—as determined by the transmitted binary sequence—but not in actual step-size magnitudes. Typically, this can be a result of some kind of an asynchronous operation. Thus, for example, the receiver may be switched on at a random instant in time, with the transmitter already in operation; the step-size in the decorder will then be different, in general, from that in the transmitter at that time instant. It would appear, now, that the effect of such mistracking would be akin to that of a random channel error occurring at the time instant in question; for, as in the case of such an error, the "decoding failure" due to asynchronous operation can be traced to a single point in time, although it propagates in the decoder output in the form of a long sequence of suboptimal step-sizes. In other words, we expect that the

* Specifically, the maximum step-size was limited to $0.05D$, where D was the dynamic range of the input speech; in the original simulation of the noiseless case, step-sizes as large as $0.10D$ had been encountered.

effect of mistracking—as that of a channel error—will be perceived as a transient in the decoder output, and the extent of such decoding failure will again depend, among other things, on (*i*) the time constant of the integrator employed at the decoder, and (*ii*) the maximum and minimum allowable step-sizes, which provide "locking points" for an asynchronous transmitter-receiver pair.

IV. A THEORY OF THE DELTA MODULATOR

We have mentioned in Section III that the optimal adaptation equations (6) and (7) were nearly invariant with respect to the sampling frequency. We shall see later, by virtue of the simulation in Section V, that these equations also hold good for a video input. These observations suggested the possibility of a fundamental and general explanation for the observed optima of P and Q. The purpose of this section is to provide such an explanation. Specifically, we propose a heuristic statistical model for the adaptive delta modulator, and go on to explain the reciprocity between the optimum values of P and Q. We also develop an upper bound of 2 for the optimum value of P.

4.1 *The Model*

Our statistical model is based on assumptions that are backed by computer simulation and physical appeal. We believe that the resulting theory provides a simplified, but useful, description of our delta modulator. The following are our tacit assumptions:

(*i*) The signal gradient $s_r = X_r - X_{r-1}$ is a random variable with a probability density function that is symmetrical about a mean value of zero.

(*ii*) In the optimal modulator, the "dynamic range" of the distribution of $|X_r|$, which denotes the signal magnitude, is much greater than the "dynamic range" of $|m_r| = |Y_r - Y_{r-1}|$, which denotes the random step-size in the staircase approximation to X.

(*iii*) With optimal adaptation, the probabilities of "P" type and "$-Q$" type adaptations of the step-size are equal. If we denote these probabilities by p and q respectively, we assume that

$$p = q = 0.5. \tag{8}$$

Assumption (*iii*) would appear to be the strongest. It is also the most crucial part of our model. In essence, the assumption states that, with optimal adaptation, overload and hunting situations are equally likely. In other words, the best adaptation logic is one which, by definition, is neither over-slow nor over-fast, but optimal in an average sense—as expressed in equation (8)—for the given input signal.

4.2 *The Optimum Value of PQ*

Consider the ratio $R(N)$ of the magnitude of m_{U+N} (the step size at the sampling instant $U + N$) to that of the step size m_U at the sampling instant U. Let the number of "P" type and "$-Q$" type adaptations of the step size m in the interval N be Np_0 and Nq_0 respectively, so that

$$R(N) = \frac{|\, m_{U+N} \,|}{|\, m_U \,|} = P^{Np_0} \cdot Q^{Nq_0} = (PQ^{q_0/p_0})^{Np_0}. \tag{9}$$

Note that, for the "most typical" sequence of step-sizes, as $N \to \infty$, p_0 and q_0 tend to the probabilities p and q. Furthermore, we have said that, for optimal adaptation, $p = q$. We can therefore define, for the optimal case, a "most typical" asymptotic value $R^M(\infty)$ for $R(N)$ as follows:

$$R^M_{\text{opt}}(\infty) = \lim_{N\to\infty} R^M_{\text{opt}}(N) = \lim_{N\to\infty} (P_{\text{opt}} Q_{\text{opt}})^{Np}. \tag{10}$$

We will now postulate an optimality criterion which will insist that the asymptotic ratio defined in equation (10) be finite and non-zero;* and because $Np \to \infty$ when $N \to \infty$, a necessary and sufficient condition for such stability will be given by

$$(PQ)_{\text{opt}} = 1. \tag{11}$$

Note that this condition applies only to the optimal system defined by equation (8).

The next two sections of this article are devoted to the derivation of a lower bound on the optimum value of Q_{opt}. By virtue of equation (11) such a bound on Q_{opt} will implicate a reciprocal bound on P_{opt}.

4.3 *Minimization of Mean Square Error*

We will adopt minimum mean square error as a criterion of optimality, and employ the notation

$$\text{Min}\,\langle E_r^2\rangle \to \text{Min}\,\langle (X_r - Y_r)^2\rangle \tag{12}$$

$$\begin{aligned} &\to \text{Min}\,\langle (X_{r-1} + s_r - Y_{r-1} - m_r)^2\rangle \\ &\to \text{Min}\,\langle (E_{r-1} + s_r - L_r m_{r-1})^2\rangle \qquad (13) \\ &\to \underset{L_r}{\text{Min}} \left\langle \left(\frac{E_{r-1} + s_r}{m_{r-1}} - L_r\right)^2 \cdot m_{r-1}^2 \right\rangle. \end{aligned}$$

* Clearly, the idea is to prevent the tendency of the step-size m either to increase beyond bounds or to decay; and the formulation in equation (10) provides a tractable way of expressing this idea.

The method of optimization that will be adopted in the sequel is equivalent to carrying out the above minimization for every specific value of m_{r-1}. Therefore, we may write

$$\text{Min}\ \langle E_r^2 \rangle \rightarrow \underset{L_r}{\text{Min}} \left\langle \left(\frac{E_{r-1} - s_r}{m_{r-1}} - L_r \right)^2 \right\rangle. \tag{14}$$

We note that, given the polarity of the adaptation parameter L_r, the magnitude of L_r is time-invariant, and hence that $| L_r |$ is independent of E, s, or m. Therefore, it can be seen that the minimization of $\langle E_r^2 \rangle$ is equivalent to the following optimization of L_r:

$$\text{Min}\ \langle E_r^2 \rangle \rightarrow \left[L_{\text{opt}} = \left\langle \frac{E_{r-1} + s_r}{m_{r-1}} \right\rangle, \quad \text{given sgn} \quad (L_r) \right]^* . \tag{15}$$

[The above optimization of L_r has the following physical meaning. In the optimal system, the step size m_r at every sampling instant is designed so that, on the average, the resulting value of Y_r tends to that of the input X_r. In other words, the value of m_r attempts to compensate, at every sampling instant, for the corresponding "lag" of the staircase signal, as expressed by the quantity $X_r - Y_{r-1}$.

This random lag $(X_r - Y_{r-1})$ has two distinct components. The first component is given by the random error E_{r-1} (an overload or undershoot) arising out of the "instantaneous" suboptimality of the previous step m_{r-1}; the second component of the lag is the signal gradient s_r, which is the amount by which the signal X will have deviated after the delta-modulator integrated its previous step m_{r-1}. Our optimization procedure is tantamount to estimating the expected value of the sum of these two components of the lag—E_{r-1} and s_r—with respect to the value of m_{r-1}.]

4.4 *An Upper Bound for* $P_{\text{opt}} = 1/Q_{\text{opt}}$

As mentioned earlier, in view of the reciprocity that has been developed for the values of P_{opt} and Q_{opt}, we can now restrict the optimization procedure to that of optimizing the value of Q on the basis of equation (15):

$$-Q_{\text{opt}} = \left\langle \frac{E_{r-1} + s_r}{m_{r-1}} \right\rangle, \quad \text{given that sgn} \quad (L_r) = -1; \tag{16}$$

$$= \left\langle \frac{E_{r-1}}{m_{r-1}} \right\rangle + \left\langle \frac{s_r}{m_{r-1}} \right\rangle, \quad \text{given that sgn} \quad (L_r) = -1; \tag{17}$$

* We have utilized the well known statistical result: If A is a random variable and B is a parameter that is statistically independent of A, the expectation $\langle (A - B)^2 \rangle$ has a minimum at $B_{\text{opt}} = \langle A \rangle$.

$$= Q_1 + Q_2 , \quad \text{given that sgn} \quad (L_r) = -1 \tag{18}$$

where Q_1 and Q_2 obviously refer to conditional expectations of the ratios in equation (17).

Figure 7 depicts a situation where sgn (L_r) is negative, and illustrates the random variables in (17). The problem will be to evaluate Q_1 and Q_2 with reference to Fig. 7. Notice at the outset that in the figure

$$\left.\begin{aligned} E_{r-1} &< 0 \\ m_{r-1} &> 0 \end{aligned}\right\}. \tag{19}$$

In what follows, we will denote the probability density functions of E_{r-1} , s_r , and of the signal amplitude X_{r-1} by $f_E(\ )$, $f_s(\ )$, and $f_X(\ )$ respectively.

4.4.1 *Evaluation of* Q_1 :

Let us first note the following equivalence of events:

$$\{E = e\} \leftrightarrow \{X = Y + e\}. \tag{20}$$

Notice next the following constraint for the overshoot error E_{r-1} :

$$-m_{r-1} < E_{r-1} < 0. \tag{21}$$

In other words, allowable values of E fall in the interval $(-m_{r-1} , 0)$. We can now invoke assumption (*ii*) in Section 4.1, (which says that the "dynamic range" of the step-magnitude $| m |$ is much smaller than that of the signal amplitude $| X |$) to make the approximation

$$f_X(Y_{r-1} + e_1) \simeq f_X(Y_{r-1} + e_2) \tag{22}$$

where e_1 and e_2 are two values of E within the "small" permissible range $(-m_{r-1} , 0)$ for E. In writing (22), we have approximated $f_X(\ )$ in the "narrow" range—from $Y + e_1$ to $Y + e_2$—by a constant function. In other words, the distribution of the overshoot error can be assumed to be uniform in the allowable range of E:

$$f_E(e) = \frac{1}{m_{r-1}} ; \qquad -m_{r-1} < e < 0. \tag{23}$$

Obviously then, the expected value of the ratio of the overshoot error E_{r-1} to the step-size m_{r-1} is given by

$$Q_1 = \int_e \frac{e}{m_{r-1}} \cdot f_E(e) \, de = \int_{-m_{r-1}}^{0} \frac{e}{m_{r-1}} \cdot \frac{1}{m_{r-1}} \, de = -0.5. \tag{24}$$

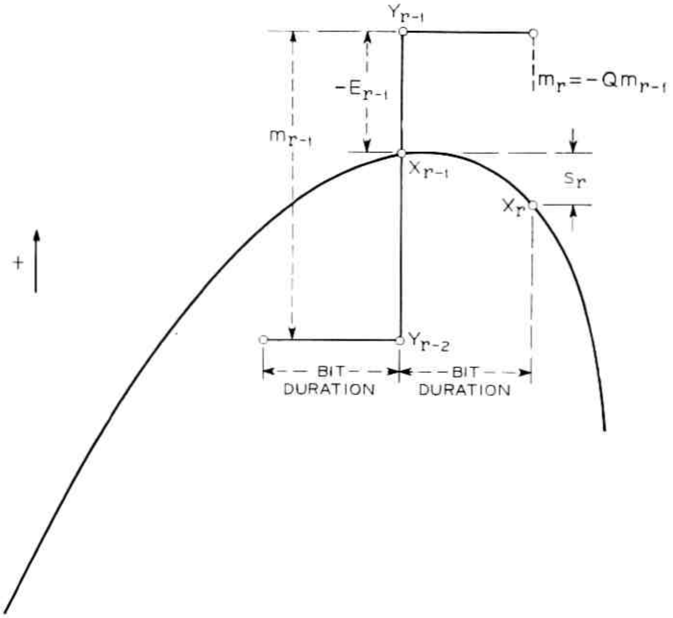

Fig. 7 — Illustration of a reversal of step-polarity.

4.4.2 *Evaluation of Q_2* :

As a requirement for the reversal of step polarity in Fig. 7, one notes the constraint

$$E_{r-1} + s_r < 0. \tag{25}$$

Hence, allowable values of the signal gradient s_r have to lie in the range

$$-\infty < s_r < -E_{r-1} . \tag{26}$$

Notice that, by virtue of (19), the upper bound for s_r in (26) is positive.

Before proceeding to evaluate the expected value of s_r, we shall comment on the use of the unconditional density function of s_r in the ensuing analysis. With a one-bit memory, the polarity of m_{r-2} is unknown. Equivalently, it can be seen that there is no constraint on the gradient s_{r-1} analogous to that on s_r in (26). This means that, with a one-bit memory, one cannot develop any conditional distributions for the future signal gradient s_r, and the use of the unconditional density function $f_s(\ )$ will therefore be valid. Consequently, using (26) and (19) and the zero-mean assumption (*i*) for s_r in Section 4.1,

$$Q_2 = \int_x \frac{x}{m_{r-1}} \cdot f_s(x)\, dx = \frac{1}{m_{r-1}} \int_{-\infty}^{-E_{r-1}} x \cdot f_s(x)\, dx; \tag{27}$$

$$= \frac{1}{m_{r-1}} \left[\int_{-\infty}^{\infty} x f_s(x)\, dx - \int_{-E_{r-1}}^{\infty} x f_s(x)\, dx \right]; \tag{28}$$

$$= \frac{1}{m_{r-1}} [0 - \epsilon]; \qquad \epsilon > 0. \tag{29}$$

In other words, during an overshoot situation, the expected value of the future signal gradient is negative with respect to the present step m_r; and this is a consequence of a finite positive bound $-E_{r-1}$ (26) on the symmetrically distributed random variable s_r.

Utilizing equations (24) and (29) we can rewrite (18) in the form

$$-Q_{\text{opt}} = -0.5 - \delta; \qquad \delta > 0. \tag{30}$$

Finally, utilizing the simple upper bound (4) of 1 for Q_{opt}, we may write

$$\begin{gathered} 0 < \delta < 0.5, \\ 0.5 < Q_{\text{opt}} < 1.0 \end{gathered} \tag{31}$$

and, by virtue of (11),

$$1.0 < P_{\text{opt}} = \frac{1}{Q_{\text{opt}}} < 2.0. \tag{32}$$

4.5 *Evaluation of the Theory*

Table II presents the values of important adaptation parameters obtained in a 60 kHz speech simulation of optimum delta modulation and compares them with the predictions of our theory. The comparison is good, and is particularly so with reference to the critical parameter p of assumption (*iii*).

We believe, in retrospect, that the heuristic theory of this section

TABLE II—CHARACTERISTICS OF AN OPTIMUM DELTA MODULATOR

Parameter	$(PQ)_{\text{opt}}$	P_{opt}	p	δ	Q_1
Theoretical Value	1	$1 < P_{\text{opt}} < 2$	0.50	$0 < \delta < 0.5$	−0.5
Value from Speech Simulation	1	1.5	0.47	0.12	−0.55

provides a simple understanding of adaptive delta modulation characterized by exponential adaption and a one-bit memory. The theory is still insufficient, however, and unanswered problems include an explicit derivation for the signal-to-error ratio and the question of analyzing the noise performance of adaptive delta modulation.

V. RESULTS FROM A VIDEO SIMULATION

We devote this section to a cursory presentation of results obtained from a simulation of the ADM with a video signal in a format that may be appropriate for communication purposes. The picture frame was made up of 250 scan lines, and a resolution of about 275 picture elements per line. The picture elements were 10-bit samples; therefore, assuming a scan rate of 30 frames/second, we were employing a 20 megabit/sec (MBPS) original. The simulator used an ideal integrator in the feedback loop and incorporated a digital low pass filter with a sharp cut-off at 1 MHz.

An important finding of the simulation was that optimum values of the adaptation parameters P and Q were still nearly equal to 1.5 and 0.66, which were values encountered in the speech simulation. Furthermore, as with speech, these optima of P and Q were nearly independent of the sampling frequency. Also, the optimized ADM performed significantly better than the unadaptive ($P = Q = 1$) encoder with an optimized step-size; at 10 MHz operation, for example, the performance

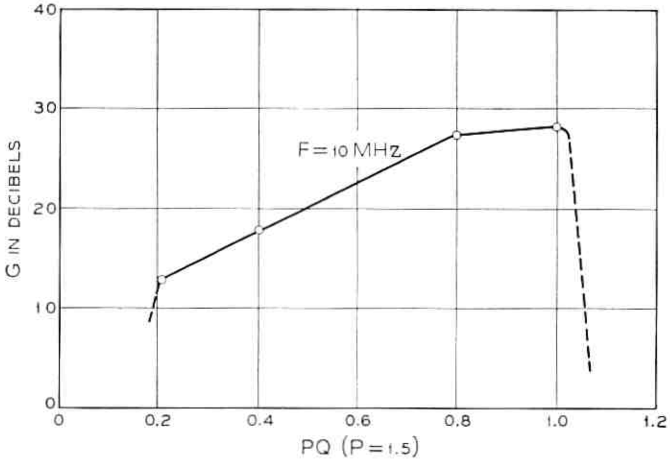

Fig. 8 — Results of the video simulation: signal-to-error ratio as a function of (PQ).

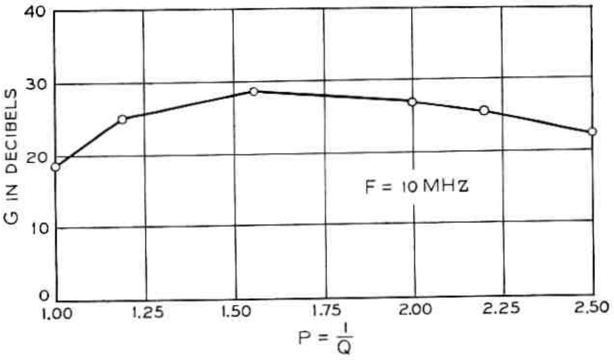

Fig. 9 — Results of the video simulation: signal-to-error ratio as a function of P.

gain, using the criterion of Section 3.2, was nearly 10 dB. We have provided, in Figs. 8 and 9, signal-to-error ratio curves that demonstrate the delta-modulator performance at 10 MHz, as a function of P and Q; the function G represents a signal-to-error ratio as averaged over the "active" or "picture" portion of the video frame.

Other sampling rates used in the simulation were 5 and 20 MHz. The performance of the modulator at 5 MHz was unsatisfactory, while the picture reproduction at 20 MHz was very acceptable. The capabilities and limitations of our scheme were best revealed in the 10 MHz simulation. In Fig. 10, we compare the output of the 10 MHz ADM, corresponding to a single frame of video input, with the 20 MBPS PCM original. The 10 MBPS ADM picture can be said to constitute a useful output; but it is not indistinguishable from the original. One notices, for example, the inadequate reproduction of the stripes on the dress of the subject.* This is attributable to the inability of the coder to follow sudden changes of input signal level; and would manifest, in the ADM version of a moving scene, as a corresponding twinkle.

The processing of moving scenes as well as the accumulation of subjective performance measures, were topics that were beyond the scope of our simulation. But such studies represent important prerequisites for a correct assessment of our delta modulator for general video application.

* Interested readers may obtain glossy prints of Fig. 10 from the author at Bell Telephone Laboratories, Murray Hill, New Jersey.

VI. CONCLUSION

We have presented a very simple form of discrete adaptive delta modulation, characterized by the use of a one-bit memory and by exponential adaptations of the step-size. We have discussed optimization procedures for such a device, and demonstrated the applicability of the modulator to audio- and video-signal reproduction at practically useful operating frequencies, such as 60 kHz for audio and 10 MHz for video. It is well known that conventional (linear) delta modulators are inefficient at such frequencies. Though our ADM can be practically important in its own right, we reiterate that the performance of our adaptation logic is to be regarded as a lower bound on the performance of more sophisticated schemes[7]—in particular, of adaptations that employ more than a one-bit memory, or of those which exploit very specific statistics of the signal to be encoded.

We have also afforded, in this paper, a comparative evaluation of adaptive delta modulation and of PCM in the contexts of Fig. 6 (for speech signals) and Fig. 10 (for video signals). It is an important conclusion from the aforecited illustrations—and from Fig. 15 in Ref. 2—that there are ranges of bit-rates, in both speech and picture systems, where ADM performance is competitive with that of PCM; this constitutes a nontrivial observation in that the original conception of delta

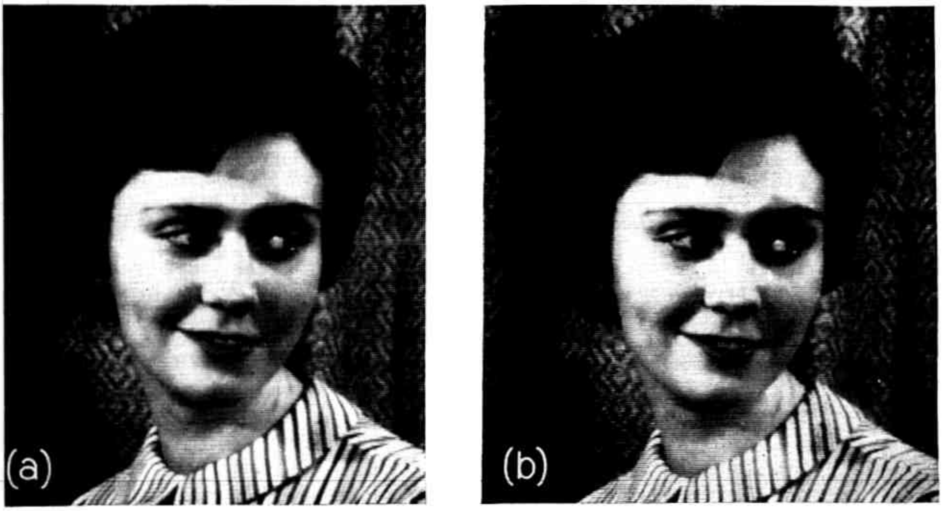

Fig. 10 — Results of a video simulation: (a) 20 MBPS PCM original (b) 10 MBPS ADM output.

modulation was very much that of an inferior, though useful, alternative to PCM. The noise-susceptibility of delta modulation could however delimit its utility for specific noisy channels. On the other hand, a simple adaptive delta modulator would appear to have an edge over conventional/differential PCM in systems characterized by relatively noise-protected channels, in low bit-rate applications, and in systems where simplicity of implementation is a critical matter.

VII. ACKNOWLEDGMENT

The author thanks Mr. J. L. Flanagan for suggesting the general topic of the paper, Mr. B. S. Atal for providing the low-pass filter used in the speech simulation, and Mr. C. A. Sjursen for processing the video tapes. The author is indebted, for important comments on the manuscript, to Messrs. B. J. Bunin, J. C. Candy, T. V. Crater and J. O. Limb.

REFERENCES

1. Tomozawa, A., and Kaneko, H., "Companded Delta Modulation for Telephone Transmission," IEEE Trans. on Communication Technology, *COM-16,* No. 1 (February 1968), pp. 149–157.
2. Abate, J. E., "Linear and Adaptive Delta Modulation," Proc. IEEE, *55,* No. 3 (March 1967), pp. 298–307.
3. Winkler, M. K., "High Information Delta Modulation," IEEE International Convention Record, pt. 8 (1963), pp. 260–265.
4. Bello, P. A., Lincoln, R. N., and Gish, H., "Statistical Delta Modulation," Proc. IEEE, *55,* No. 3 (March 1967), pp. 308–318.
5. Greefkes, J. A., and de Jager, F., "Continuous Delta Modulation," Philips Research Reports, *23,* No. 2 (April 1968), pp. 233–246.
6. Greefkes, J. A., "Digitally Controlled Code Modulation," Philips Research Laboratories (Eindhoven) Report, m.s. 6045, 1969.
7. Candy, J. C., "Refinement of a Delta Modulator," Proceedings of the Symposium on Picture Bandwidth Compression, MIT Press, 1970.
8. Bosworth, R. H., and Candy, J. C., "A Companded One-Bit Coder for PICTUREPHONE® Transmission," B.S.T.J., *48,* No. 5 (May-June 1969), pp. 1459–1479.
9. Smith, B., "Instantaneous Companding of Quantized Signals," B.S.T.J., *36,* No. 3 (May 1957), pp. 653–709.
10. O'Neal, J. B., Jr., "A Bound on Signal-to-Quantizing Noise Ratios for Digital Encoding Systems," Proc. IEEE, *55,* No. 3 (March 1967), pp. 287–292.

On an Anomaly in the Mobility of Gaseous Ions

By GREGORY H. WANNIER*

(Manuscript received July 8, 1969)

Many mobility versus field curves for gaseous ions show a high mobility "bump" just above the ohmic range. The effect arises from the nature of the force between ions and molecules. It is effectively attractive for low speeds of encounter and repulsive for high speeds. A partial cancellation of deflections occurs in a range of intermediate speeds; the scattering cross section then appears to be abnormally low.

I. INTRODUCTION

The first semiquantitative understanding of the motion of gaseous ions in electric fields was achieved by Langevin.[1] He adopted as a model force between the ions and the gas molecules a superposition of the attractive polarization force and a hard core repulsion. He then applied kinetic theory to the mixture of ions and molecules and determined the response of the ions in such a mixture to a small field. A drift velocity proportional to the field was the result. The constant of proportionality is called the mobility. Langevin produced the first estimates for this number.

There has been no essential departure from Langevin's approach in subsequent years, but only refinements and extensions; they occurred generally in close correlation with experiment.[2-5] A useful extension was the one to high fields. One gets then a drift velocity versus field curve rather than a simple constant of proportionality. In favorable cases, the analysis of such data has been carried out in a quite satisfactory way.† The general rule is that if the results are expressed in terms of a mobility, then the mobility tends to decrease with increasing

* Mr. Wannier is Professor of Physics, University of Oregon, Eugene. He performed this work while on sabbatical leave at Bell Telephone Laboratories, Murray Hill, New Jersey.

† See Refs. 6 and 7. In Ref. 6, the abscissas on Figs. 3–7 are a power of ten too small.

field. The qualitative explanation of this trend is that high fields raise the mean random velocity of the ions above their thermal value. The mean speed of encounter of ions and molecules is thereby also increased. Under those conditions the mean free time between collisions would remain a constant only for inverse fifth power forces (Maxwellian molecules), but decrease for stiffer forces. This is normally the case in practice, except in the limit of very slow encounters when the polarization force prevails.

It is the purpose of this paper to focus attention on the "mobility bump" which is observed occasionally in a mobility versus field plot. While the mobility generally behaves as described in the preceding paragraph, there is sometimes found a short range of fields for which the mobility rises before the drop sets in.[4,8-10] The explanation proposed for this effect is the following. The drift velocity of the ions is controlled by their encounters with the molecules; these depend in turn on the mutual force. This force is attractive at long range and repulsive at short range. If one studies the momentum transfer cross section for such a force as function of speed one finds that it has a "dip" at intermediate speeds as compared to the limiting laws for high or low speed. The reason for this dip is a partial compensation of attraction and repulsion. The latter is responsible for the high speed behavior, and the former for low speed behavior. However, attraction and repulsion bend the path in the opposite sense, or give phase shifts of opposite sign. Hence a compensation with anomalous transparency must be expected for a small range of speeds. If these speeds are just slightly larger than thermal under the experimental conditions employed, a "bump" type anomaly will appear in the data. Furthermore the bump will be larger if the repulsive force is soft. The reason for this is that a soft repulsion can compensate the polarization attraction over a wider range of speeds than a hard force.

II. A STUDY OF CROSS SECTIONS

Computations of drift velocities and comparisons with experiment are presented in Section III. In this section, we show the effect of the compensation phenomenon on the behavior of the momentum transfer cross section, employing two simple models.

Within the range of validity of classical mechanics the momentum transfer cross section $\sigma(v)$ is defined as

$$\sigma(v) = 2\pi \int_0^\infty (1 - \cos \chi) b \, db. \tag{1}$$

Here b is the impact parameter for a collision, and χ is the angle of deflection in the center of mass frame. For the discussion of this section, we may think of the mobility as being inversely proportional to the quantity (1).

The first model to be discussed is the so-called Langevin force, consisting of the polarization force as the attractive force and a hard sphere radius as the repulsion. The deflection equals

$$\chi = 2 \int_0 \frac{b\, du}{\left\{1 - b^2u^2 + \dfrac{e^2P}{mv^2} u^4\right\}^{\frac{1}{2}}} - \pi. \tag{2}$$

Here P is the polarizability of the molecules and u an integration variable which equals the reciprocal radius. The upper limit of the integral is the smaller positive root of the denominator or $1/a$ whichever is less; a is the radius of the hard core. Formula (2) gives rise to a variety of elliptic integrals which one can teach a computer to distinguish and to look up in its library. After this is done, the integration (1) has to follow; this was carried out numerically. Results are shown in Fig. 1 in a log-log plot. On abscissa is V which equals

$$V = \left(\frac{m}{P}\right)^{\frac{1}{2}} \frac{a^2}{e} v. \tag{3}$$

It is a scaled dimensionless speed whose adjustable parameter is the hard sphere radius a. On ordinate is the cross section Σ in units πa^2. The curve is entirely determined by its two asymptotes. The equations for the asymptotes are

$$\Sigma = 2.210/V \tag{4a}$$

for the polarization force, and

$$\Sigma = 1 \tag{4b}$$

for the hard sphere repulsion.

Observation of Fig. 1 shows that a simple interpolation between the two straight lines, say, by adding the two cross sections, does not reproduce the actual behavior of the cross section even qualitatively. As the speed increases from very low values the cross section departs from the polarization value by being lower, not higher. The effect is admittedly small; the cross section falls to 85 percent of the polarization value and behaves normally as regards the hard sphere value: it approaches it from above.

To show up the effect more clearly, the calculation was repeated for

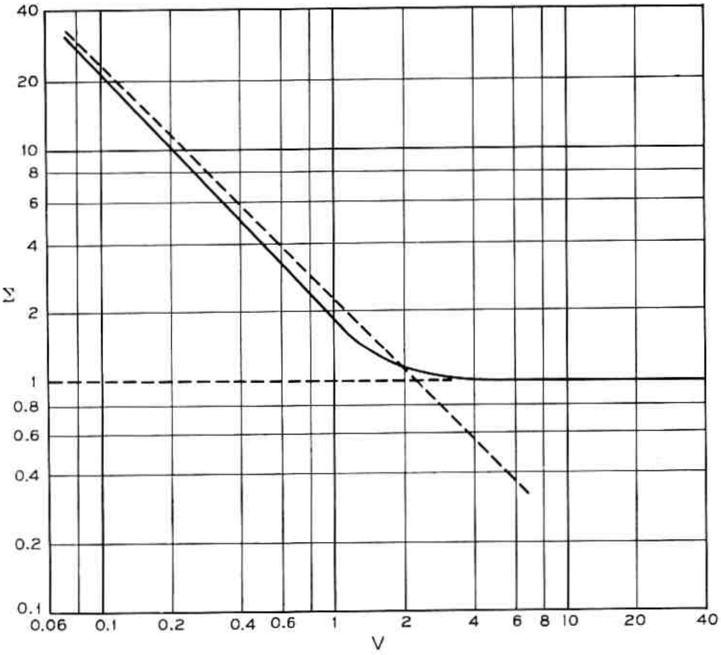

Fig. 1 — Momentum transfer cross section versus speed for the Langevin force. Cross section is relative to hard sphere, speed is rendered dimensionless through (4). Cross section falls below one limiting law (asymptote) albeit by a small amount.

another force model: the same polarization force plus a $1/r^6$ repulsive potential. The reason for this choice was that it is an extremely soft repulsion; so the two models should bracket the truth. An incidental advantage is that no orbit calculations are required because the angle of deflection is again an elliptic integral. In detail, the potential U was taken in the form

$$U = \tfrac{1}{2}e^2P\left\{-\frac{1}{r^4} + \frac{a^2}{r^6}\right\}. \tag{5}$$

a is the distance at which the potential vanishes, and thus resembles vaguely the hard core radius of the first example. With the help of standard mechanics, one finds for the angle of scattering

$$\chi = \int_{s_1}^{\infty} \frac{b\,ds}{\left(s^3 - b^2s^2 + \dfrac{e^2P}{mv^2}\,s - \dfrac{e^2P}{mv^2}\,a^2\right)^{\frac{1}{2}}} - \pi. \tag{6}$$

s is an integration variable which equals the square of the radius; the other letters have the same meaning as previously. s_1 is the largest positive root of the denominator. The integral thus is a complete elliptic integral. It takes two different forms, depending on whether the denominator has three real roots or one real root. One can teach a computer to find the roots from Cardano's formula and to look up the elliptic integral in its library. Once χ is found, the momentum transfer cross section (1) is computed in the same way as in the first example.

Results are shown in Fig. 2 on a log-log plot similar to Fig. 1. The parameter V defined in equation (3) is again used as abscissa, and the ordinate is again the cross section in units πa^2. On a log-log plot the curve has again two asymptotes representing high speed and low speed behavior. The equations for the asymptotes are

$$\Sigma = 2.210/V \tag{7a}$$

and

$$\Sigma = 1.112/V^{\frac{2}{3}}. \tag{7b}$$

They represent respectively the cross section which would prevail if the polarization force or the repulsive force were present alone.

This time the effect under discussion is very large. The curve for the cross section approaches either asymptote from below; in the central region it is substantially smaller than it would be according to either limiting law. The reduction is to 75 percent of the repulsive cross section and 36 percent of the polarization cross section.

Before comparing these results with experiment, we shall look at the theory internally and compare the two model cases with each other. The effect under discussion arises because there are strongly bent orbits which finally result in a small deflection; the reason is that bending toward and away from the center cancel. Langevin was aware of this effect.[1] In his Fig. 4, the fourth from the axis of the eleven orbits shown is of that nature. His results also contain the bump in the mobility curve. His Fig. 7 is essentially a plot of mobility versus speed. However, the effect is small. Our second example shows that if the repulsive force is made soft the effect can become very large. So it is rather the smallness of the effect for the Langevin force which needs some extra attention here. The effect is small because the model is discontinuous. Orbits which approach the hard core ever so closely do not experience any repulsion, and hence no cancellation leading to anomalously small angles. On the other hand, orbits colliding with the core do experience the attraction. But, once present, the repulsion predominates very quickly,

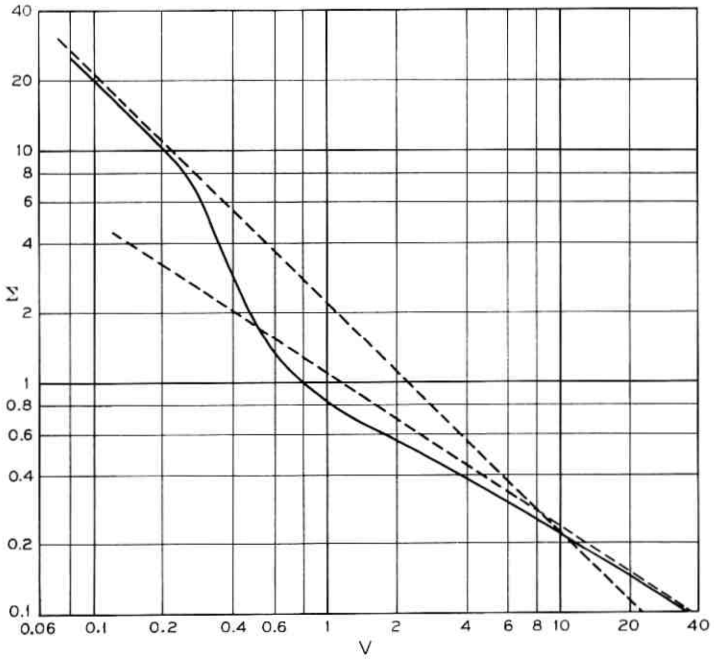

Fig. 2 — Momentum transfer cross section versus speed for the 4–6 potential. Abscissa and ordinate are essentially as in Fig. 1. Cross section falls very far below both limiting laws (asymptotes). Strong effect arises from softness of repulsion.

and the opportunity for small angle deflections is limited. In Fig. 3 a typical plot of deflection versus impact parameter is shown. The angle itself is a continuous function, but its derivative is infinite for orbits just barely touching the hard sphere. The effect under discussion arises from the small angles in the neighborhood of the point where the angle passes through zero. This is very close to the point having infinite derivative; hence, the relevant angular range is very small. If the repulsive force is softer, Fig. 3 will become smooth, and look somewhat like Fig. 4. Clearly, the range of initial conditions for which the angle of scattering is anomalously small will be much larger for such a situation.

III. COMPARISON OF THEORY AND EXPERIMENT

It is an essential feature of the "mobility bump" that it is observed outside the ohmic range. A simple mobility calculation is thus not quite right, but one should carry out an "intermediate field" type of

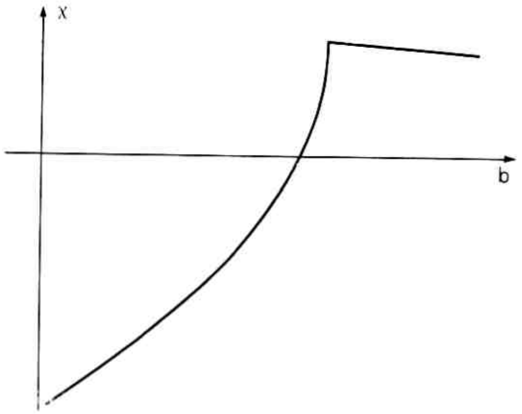

Fig. 3 — Detail on the dependence of the scattering angle on the impact parameter for the Langevin force. Infinite slope discontinuity inherent in the model makes also the passage through zero very rapid.

calculation.[6] However, as the bump appears at the very edge of the ohmic range, a mobility calculation should be indicative of precise results. What drives the drift velocity out of the ohmic range is the increase of the random speed of the ions above the thermal value. This speed can be very reliably estimated using experimental information only. As the first step, we "unreduce" a plot giving the reduced mobility as function of E/p_0 , in order to determine the observed drift velocity v_d . This is accomplished with the help of the formula

$$v_d = 760 \frac{E}{p_0} \mu_0 \; . \tag{8}$$

Thereupon we determine the mean square velocity by a formula which

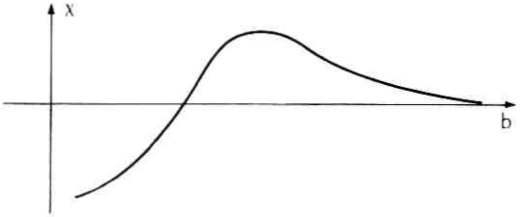

Fig. 4 — Detail on the dependence of the scattering angle on the impact parameter for the 4–6 potential. Curve is continuous and the passage through zero slower than in Fig. 3.

is discussed extensively elsewhere*

$$\langle v^2 \rangle = \frac{3kT}{m(\text{ion})} + \left(1 + \frac{m(\text{gas})}{m(\text{ion})}\right) v_d^2 \, . \tag{9}$$

What we need further on is the root mean square *relative* speed which we shall simply call "the speed" and denote by v:

$$v = \left\{\frac{3kT}{m(\text{gas})} + \langle v^2 \rangle\right\}^{\frac{1}{2}} \tag{10}$$

which comes out to be

$$v = \left\{\frac{3kT + m(\text{gas})v_d^2}{m}\right\}^{\frac{1}{2}} . \tag{11}$$

Here m is the reduced ion-molecule mass as used previously.

We may use equation (11) to convert the experimental data into a plot giving reduced mobility versus speed. Such a plot is shown in semilog form in Fig. 5 for H_3^+ in H_2 as published in Ref. 9. The plot shows the conventionally reduced mobility as function of the logarithm of the speed v.

We are in a position to find theoretical data which can be compared with a curve such as Fig. 5. The results on cross sections obtained in Section II can be exploited to yield a mobility μ with the help of the formula

$$\mu = \frac{1}{3} \frac{e}{mN} \left\langle \frac{1}{v^2} \frac{d}{dv} \left(\frac{v^2}{\sigma(v)}\right) \right\rangle^{\dagger} . \tag{12}$$

Here e is the charge of the ion, m the reduced mass of the ion-molecule system and N the number density of the gas. $\sigma(v)$ is the momentum transfer cross section as defined in equation (1). In addition, a calculation of Hershey can also be brought in for comparison.[5] Hershey carried out calculations of mobilities for a ninth power repulsive force in combination with the polarization force. His result as shown in curve II, Fig. 7, Ref. 5, is of the desired form. His abscissa, labelled $1/\mu$, is the random velocity (10) of this paper, apart from a scale factor. His ordinates

* See equation (21.20) of Ref. 7 or equation (122) of Ref. 6. Equation (97) of Ref. 6 also shows an instance in which the formula is not rigorously valid. Yet it still holds to within 5 percent.

† The formula is a modification of (20.10) of Ref. 7 for an isotropic situation. It also appears as (21.17), or results from (21.35). All three derivations fall short of being general. Indications are that the formula is close but not exact. Compare the comments to (168) of Ref. 6 where the same formula appears with a slightly different numerical factor.

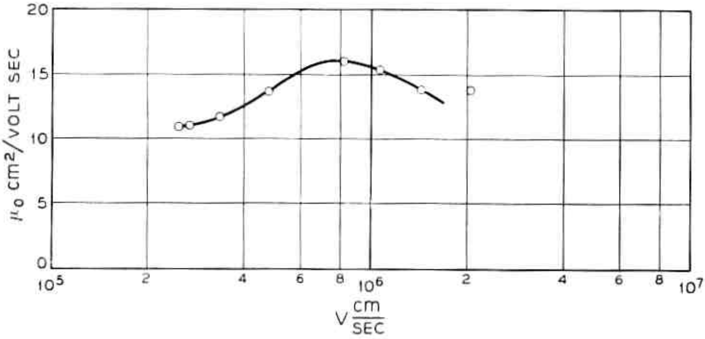

Fig. 5 — Adaptation to theoretical analysis of the data of Miller and others, on the motion of H_3^+ in H_2. Ordinate is the same as in the original paper, but on abscissa is plotted the root mean square speed of encounter.

must be multiplied with a factor to yield the polarization mobility at zero speed for the system under consideration.

In Figs. 6, 7 and 8 are shown the reduced mobilities predicted for H_3^+ in H_2 , using hard sphere repulsion, seventh power repulsion and ninth power repulsion, respectively, combined with the polarization attraction. On abscissa is the mean speed of encounter; the speed is plotted logarithmically, so that scale factors have no influence on the shape of the curves. Their only adjustability consists in a possible horizontal rigid displacement.

Comparison of the three theoretical curves among themselves bears out the point made at the end of the introduction. The bump is largest in Fig. 7 for which the repulsion is softest, and smallest in Fig. 6, for the

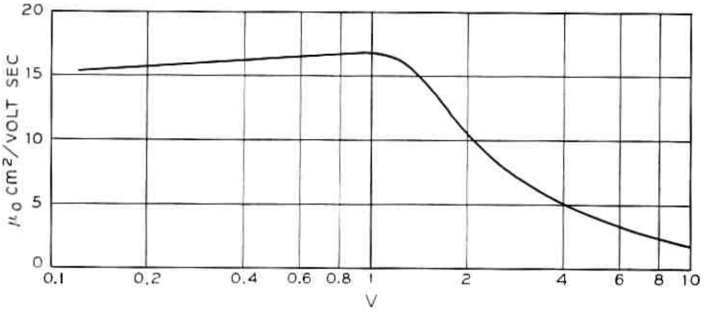

Fig. 6 — Theoretical mobility versus speed curve for H_3^+ in H_2, adopting the Langevin model. The speed has an adjustable scale factor which allows a horizontal shift without distortion of the curve shown.

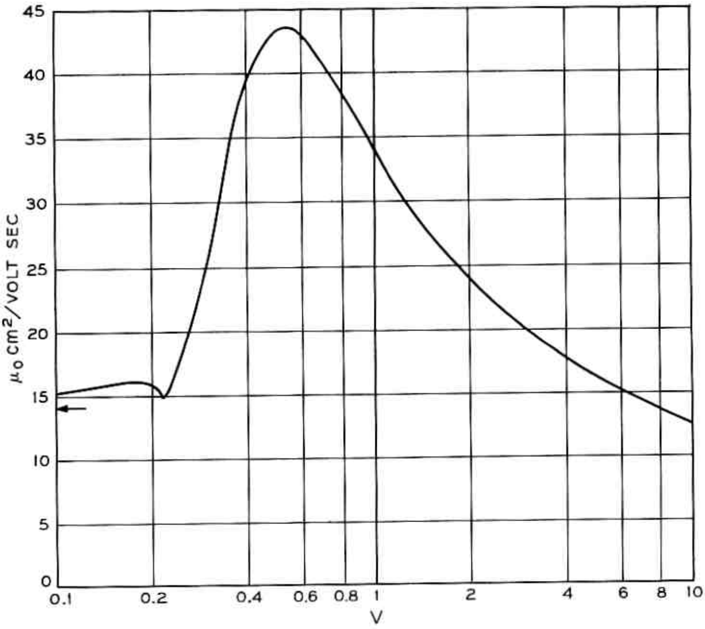

Fig. 7 — Theoretical mobility versus speed curve for H_3^+ in H_2, adopting a 4–6 potential model. The speed has an adjustable scale factor which allows a horizontal shift without distortion of the curve shown.

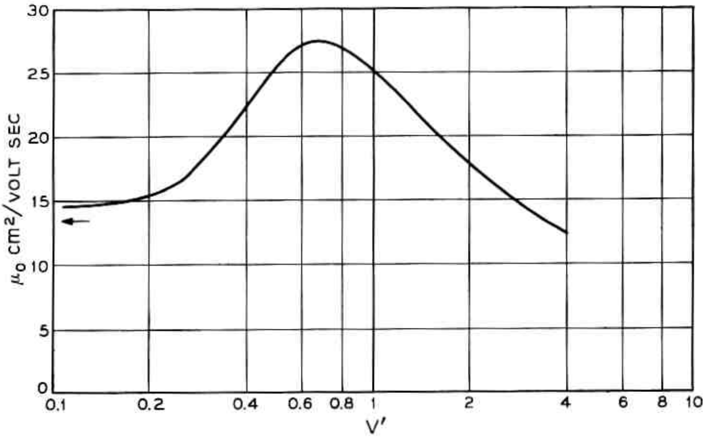

Fig. 8 — Theoretical mobility versus speed curve for H_3^+ in H_2, adopting a 4–8 potential model. Adaptation of results of Hershey.[5] V' is also a scaled speed. A horizontal shift without distortion of the curve shown is thus allowed.

hard sphere. Intermediate hardness yields an intermediate size bump. When we go on comparing these curves with the experimental curve shown in Fig. 5, we find the experimental bump in between the ninth power repulsion and the hard sphere. A Lennard-Jones type thirteenth power repulsion is thus quite a plausible candidate for a good fit.

We can make a further comparison between theory and experiment by identifying the velocities for which the maxima occur. The maxima in Fig. 6 and Fig. 7 occur roughly at $V = 1$, the experimental one at $v = 10^6$ cm/sec. We can thus use equation (3) to get an empirical value for the hard sphere radius a. We find

$$a = 1.7 \times 10^{-8} \text{ cm.}$$

Actually, the theoretical bump should be higher than the experimental one because the averaging process over velocities was omitted. Since the bump arises only for a restricted set of speeds it will be reduced by an averaging procedure.

The theoretical curves are not adjustable in a vertical direction and there is thus an unexplained discrepancy between theory and experiment in the low speed mobility value. The theoretical value results from the formula

$$\mu_0 = \frac{0.5105}{[\rho(\epsilon - 1)]^{\frac{1}{2}}} \left[1 + \frac{m(\text{gas})}{m(\text{ion})}\right]^{\frac{1}{2}} \frac{1}{300} \quad \text{cm}^2/\text{volt sec} \tag{13}$$

where ρ is the density and ϵ the dielectric constant of hydrogen. Taking for ρ the value 0.899×10^{-4} and for $\epsilon - 1$ 2.73×10^{-4}, we find a value of 14.03 cm^2/volt sec for μ_0 while the measured one is 11.2. The cause for this discrepancy is not known at this time. It is possible that formula (13) is not quite correct for a molecular gas. The molecular polarizability is a tensor function which depends on orientation. The dielectric constant represents the polarizability response to a uniform field. If the ion is capable of orienting the molecules or comes so close as to experience details of molecular structure then the effective polarizability will be larger and the mobility smaller.

Before leaving the subject of comparison with experiment, I wish to call attention to the data of Ref. 9 taken at the very highest fields. A second rise of the mobility is indicated. The theory proposed cannot explain such a rise. If the explanation is right, this rise must be an experimental error or arise from a quite extraneous feature.

IV. CONCLUSIONS

It is the conclusion of this paper that the mobility bump which shows up in recent experiments is a normal feature of the classical theory of ionic mobility. It may actually be found in the classical papers on the subject[1,2,5] but the effect happens to be quite small for the Langevin model. The second model discussed here and the work of Hershey[5] show that it can be quite large, with the mobility rising to 300 percent of its polarization value. The size of the bump is critically dependent on the softness of the repulsive part of the potential. It is thus plausible to expect that a $1/r^{12}$ repulsion such as occurs in the Lennard-Jones potential will give rise to curves resembling the experimental ones. With calculations of this type one might set up a correspondence between "bump size" and "softness". However, a glance at the experimental data indicates that such an identification is not easy to make because the bump occurs primarily when either the ion or the molecule or both are extended systems. "Softness" may thus be an indirect attribute arising because the force is different for different orientations.

V. ACKNOWLEDGMENT

I wish to express my thanks to Dr. L. G. McKnight and to Professor E. W. McDaniel for acquainting me with this interesting subject.

REFERENCES

1. Langevin, M. P., "Une formule fondamentale de théorie cinétique," Ann. Chim. et Phys., *5* (1905), 245–288.
2. Hasse, H. R. "Langevin's Theory of Ionic Mobility," Phil. Mag., *1,* No. 1 (January 1926), pp. 139–160.
3. Hasse, H. R. and Cook, W. R., "The Calculation of the Mobility of Monomolecular Ions," Phil. Mag., *12,* No. 77 (August 1931), pp. 554–566.
4. Hershey, A. V., "Measurements of the Mobility of Potassium Ions at High Field Intensity and Low Pressure," Phys. Rev., *56,* No. 9 (November 1939), pp. 908–915.
5. Hershey, A. V., "A Theory for the Mobility of Ions of High Velocity," Phys. Rev., *56,* No. 9 (November 1939), pp. 916–922.
6. Wannier, G. H., "Motion of Gaseous Ions in Strong Electric Fields," B.S.T.J., *32,* No. 1 (January 1953), pp. 170–254.
7. Wannier, G. H., *Statistical Physics,* New York: John Wiley and Sons, 1966, Chapter 21, pp. 455–471.
8. Albritton, D. L., Miller, T. M., Martin, D. W., and McDaniel, E. W., "Mobilities of Mass-Identified H_3^+ and H^+ Ions in Hydrogen," Phys. Rev., *171,* No. 1 (July 1968), pp. 94–102.
9. Miller, T. M., Moseley, J. T., Martin, D. M., and McDaniel, E. W., "Reactions of H^+ in H_2 and D^+ in D_2; Mobilities of Hydrogen and Alkali Ions in H_2 and D_2 Gases," Phys. Rev., *173,* No. 1 (September 1968), pp. 115–123.
10. Eiber, H., and Kandel, W., "Ueber die Beweglishkeit von Ionon in Sauerstoff-Wasserdampfgemischen," Zeits. f. Angewandte Phys., *25,* No. 1 (June 1968), pp. 18–23.

The Enumeration of Neighbors on Cubic and Hexagonal-Based Lattices

By J. D. WILEY and J. A. SEMAN

(Manuscript received October 2, 1969)

Radii and occupation numbers have been calculated for the first 50 shells of neighbors on each atomic sublattice for the CsCl, NaCl, *zincblende, wurtzite, and* CaF_2 *binary lattices. We present the results in tabular form along with rules for extending the tables to higher shell numbers. A sublattice approach is used and tables are given for key cubic and hexagonal-based sublattices. The generality of the sublattice approach is such as to allow easy application of the tables to more complex lattice structures or to such problems as enumeration of preferred interstitial sites. A number-theoretic explanation is offered for previously observed difficulties in obtaining a simple expression for the radius of the m-th shell in cubic-based structures.*

I. INTRODUCTION

In discussing phenomena involving the interaction of ions in a crystalline lattice it is often necessary to know the radii and occupation numbers of near-lying shells of lattice sites. Such information is extremely important, for example, in the interpretation of donor-acceptor pair recombination spectra[1,2] and in calculations of ion pairing[3,4] and other defect clustering phenomena. The present work was motivated by the apparent lack of any generally available tables or formulae for calculating these m-th neighbor shell parameters for common lattices. Shell radius formulae and partial tables have been published[1,2] for the interpretation of pair spectra in materials with zincblende lattices but these tables are inadequate for other applications. Wood[5] and Ferris-Prabhu[6] have given slightly more complete treatments but do not present sufficiently general rules to allow indefinite extension of their tables.* The methods which will be described here differ from those

* In fact, if the diamond lattice radius rules given by Ferris-Prabhu[6] were used to extend his table beyond the 25 shells which he lists, one would err in predicting the radius of the 28th shell and would have all higher shells improperly numbered. Further errors would be made for much higher shell numbers.

reported previously[1,2,5,6] in that greater attention is given to the formulation of general rules which allow extension of the tables to higher shell numbers. It is hoped, however, that the tables presented will be sufficiently large for most applications and will not require extension. The general approach will be discussed in Section II and final tables of shell parameters will be presented together in Section III.

II. DISCUSSION

The notation to be used throughout is as follows: A convenient lattice point will be chosen as the origin and will be taken as the center of a spherical shell which is allowed to expand. At certain radii, ρ_m , the shell will coincide with other points of the lattice. The number, Z_m , of lattice points on the shell of radius ρ_m will be referred to as the occupation number of the m-th shell or the number of m-th neighbors. To find the radius and occupation number of the m-th shell, the following general approach[5,6] will be used: For each lattice a rectangular set of basis vectors $(\mathbf{a}_1 , \mathbf{a}_2 , \mathbf{a}_3)$ will be chosen in such a way as to allow the coordinates of any lattice point to be written as $(\ell_1 , \ell_2 , \ell_3)$ where the ℓ_i are integers. All lattice points will therefore fall on the corners of rectangular parallelepiped (usually cubic) cells of the basis lattice but since the basis lattice is smaller than the actual lattice, there will be sets of integers which do not correspond to actual lattice points. Rules must therefore be formulated to allow these fictitious points to be rejected in the enumeration process. Points of the real lattice can then be enumerated by systematically counting all allowed combinations $(\ell_1 , \ell_2 , \ell_3)$. Since each point $(\ell_1 , \ell_2 , \ell_3)$ is located on a sphere of radius

$$\rho^2 = \ell_1^2 a_1^2 + \ell_2^2 a_2^2 + \ell_3^2 a_3^2 , \tag{1}$$

one can obtain Z_m by counting all lattice points with equal ρ^2 values and arranging the shells in order of ascending ρ^2. This process is simplified by making use of reflection and permutation symmetries but is best done by computer in any case.

It frequently turns out that one can write the radius of the m-th shell as a simple function of m: $\rho_m = f(m)$. Although there is no *a priori* reason to expect that such a formula will exist for any given lattice, it is very convenient if one can be found. (Radius formulae are useful, for example, in estimating the number of ℓ_i values which must be considered in order to count all lattice points of the m-th shell.) Since the subject of radius formulae has been a source of some confusion in the literature,[6] it will be given special attention in the discussions of specific lattices which follow.

2.1 *Cubic-Based Lattices*

Four monatomic cubic-based lattices will be considered first: Simple Cubic (sc), Body Centered Cubic (bcc), Edge Centered Cubic (ecc), and Face Centered Cubic (fcc). In each case the origin will be chosen to be at a cube corner and the basis vectors will be $(a/2)\mathbf{i}$, $(a/2)\mathbf{j}$, and $(a/2)\mathbf{k}$, where a is the length of a full cube edge and $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are unit vectors in the x, y, and z directions respectively which are taken to be cube edges. In the case of sc, one could choose vectors of length a but the $a/2$ choice turns out to be more convenient.

The bcc, fcc, and ecc lattices will each be decomposed into two sublattices: Sublattice 1 will consist of cube corners (as defined by the position of the origin) and sublattice 2 will consist of body centers (bc), face centers (fc), or edge centers (ec) as the case may be. This is illustrated in Fig. 1.

Since the basis lattice is a sc lattice with edge length $a/2$, all lattice points of the larger sc, fcc, ecc, and bcc structures can be written with integer coordinates (ℓ_1, ℓ_2, ℓ_3). Furthermore, it is seen by inspection that the following rules apply: (*i*) Points on sublattice 1 are obtained if and only if ℓ_1, ℓ_2, and ℓ_3 are all even. (*ii*) Points on the bc sublattice

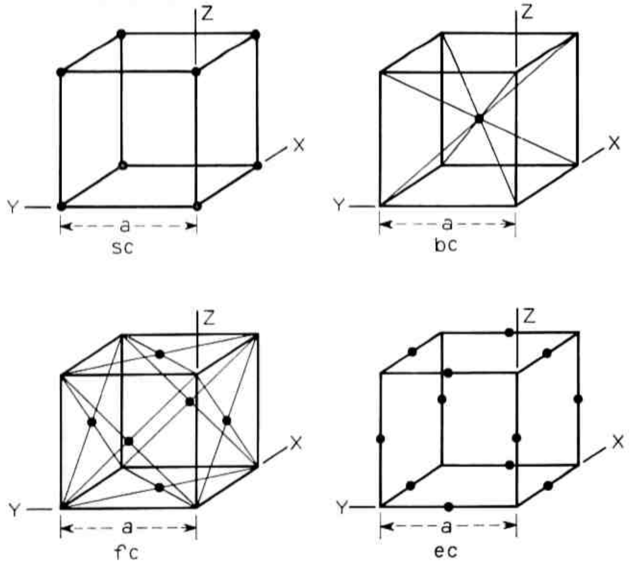

Fig. 1—The fundamental cubic-based lattices sc, bcc, fcc and ecc are shown decomposed into convenient sublattices: sc = sc, bcc = (bc + sc), fcc = (fc + sc), ecc = (ec + sc).

are reached if and only if the coordinates are all odd. (*iii*) Points on the fc sublattice are reached if and only if two of the coordinates are odd. (*iv*) Points on the ec sublattice are reached if and only if two of the coordinates are even.

By systematically enumerating all combinations of ℓ_1, ℓ_2, and ℓ_3 that satisfy these criteria, one obtains all points of the respective sublattices. For the simple lattices considered here one can make use of reflection symmetry and permutation symmetry by considering only points for which $\ell_1 \geqq \ell_2 \geqq \ell_3 \geqq 0$. Any point (ℓ_1, ℓ_2, ℓ_3) is then one-, three-, or sixfold degenerate under permutation of coordinates (p) and two-, four-, or eightfold degenerate under reflection in the coordinate planes (d). The total number of points equivalent to (ℓ_1, ℓ_2, ℓ_3) is then $Z_i = dp$ where d and p are given by Table I where A, B, and C are distinct integers and order is immaterial. In many cases there will be two or more nonequivalent sets of lattice points on the same shell. In such cases $Z_m = \sum_i d_i p_i$ where i ranges over the various distinguishable sets of lattice points. For example, the 22nd shell of the sc lattice has $\rho^2 = 100(a/2)^2$. This shell contains points of the type (8, 6, 0) and (10, 0, 0) (in units of $a/2$). There are $6 \times 4 = 24$ of the former and $3 \times 2 = 6$ of the latter for a total of 30 points on this shell.

Having chosen cartesian basis vectors of equal lengths we can write the distance from the origin to any lattice point in the form

$$r^2 = \ell_1^2 + \ell_2^2 + \ell_3^2, \tag{2}$$

where r is the shell radius in units of the basis vector length. Thus the square of the radius vector to any lattice point must be expressible as the sum of three perfect squares. This is highly relevant to previously observed difficulties in obtaining simple expressions for the radius r_m of the m-th shell in cubic based lattices. The usual difficulty is that one is able to find a formula which works only for a limited number of

TABLE I—NUMBER OF POINTS EQUIVALENT BY SYMMETRY

Coordinates of the Form	p
(A, A, A)	1
(A, A, C)	3
(A, B, C)	6
Number of Zero Coordinates	d
2	2
1	4
0	8

TABLE II—VALUES OF r_m^2 WHICH ARE FORBIDDEN IN CUBIC-BASED LATTICES*

	r				
s	0	1	2	3	4
0	7	28	112	448	1792
1	15	60	240	960	·
2	23	92	368	·	·
3	31	124	·	·	·
4	39	156	·	·	·
5	47	188	·	·	·
6	55	220	·	·	
7	63	252	·		
8	71	·			
9	79	·			
10	87	·			
11	95	·			
12	103	·			
·	·				
·	·				

* Numbers of the form $4^r\,(8s + 7)$ where r and s are integers $\geqq 0$.

shells and then fails by predicting a shell of lattice points at some radius r_m where, in fact, no actual lattice points exist. Thomas, and others,[1,2] call these "empty shells" and count them as m-th neighbors with $Z_m = 0$. Through this device a formula can be made to work for all r_m. Thomas, and others, also give a formula which predicts the shells which will require $Z_m = 0$ for the zincblende lattice but do not discuss the origin of this formula. In every case investigated in the present work, a failure of shell radius formulae occurred because these formulae predicted values for r_i^2 which were not expressible as the sum of three perfect squares.* It is known from the theory of numbers[7] that an integer can be expressed as the sum of three squares† if and only if it is not of the form $4^r(8s + 7)$ where r and s are integers $\geqq 0$. Thus whenever a radius formula predicts a value of r_m^2 of the form $4^r(8s + 7)$, no shell of lattice points will exist since r_m^2 will fail to satisfy the physical constraint given by equation (2). A few of these forbidden r_m^2 values are listed in Table II.

The sc, bc, fc, and ec sublattices form a basic set from which one can construct more complex lattices. They all have reflection and per-

* It is possible, in more complex lattices, for a radius rule to fail for other reasons.

† It can be shown that *any* integer can be expressed as the sum of not more than four squares, nine cubes, or nineteen fourth powers.[7] The important point, however, is that three squares are sufficient.

mutation symmetry, however, and will not be directly applicable to any lattices or sublattices that do not share these symmetries. This is illustrated in the final monatomic lattice to be considered in this section: diamond.

The diamond lattice is composed of two interpenetrating fcc lattices which are shifted along a common diagonal by an amount $(a/4, a/4, a/4)$ as indicated in Fig. 2. A corner of one of these fcc lattices will be chosen as the origin and this sublattice will be denoted "I." The shifted sublattice is then "II." The basis vectors are taken to be $a/4$ in length and span a sc basis lattice with cube edges $a/4$. The restrictions on ℓ_1 , ℓ_2 , ℓ_3 are found by inspection (this process is aided by consideration of projections in the coordinate planes), and are given in Table III along with a summary of similar results for the sc, bc, fc, and ec sublattices. The radius formulae given in column 5 of Table III are obtained by inserting general integers of the forms given in column 4

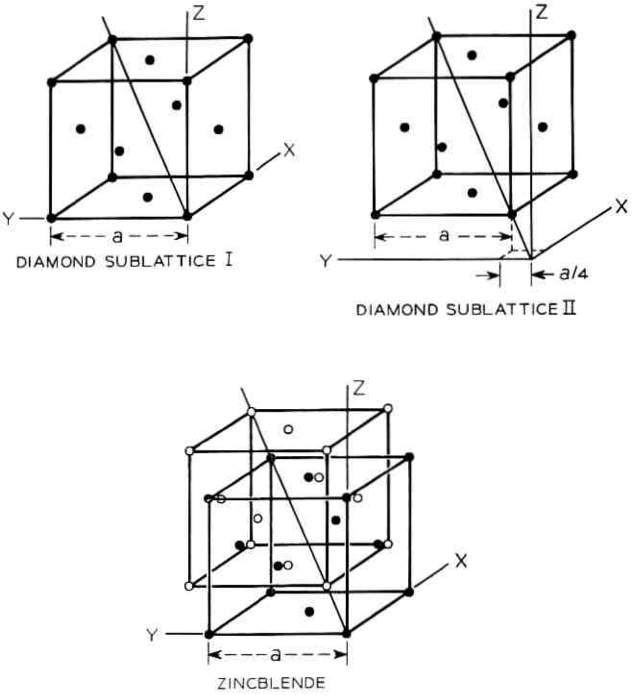

Fig. 2—The diamond and zincblende lattices are shown decomposed into two fcc sublattices. In diamond, atoms on sublattices I and II are identical. In zincblende, atoms on I are of one type and atoms on II are the other type.

TABLE III—SUMMARY OF THE FUNDAMENTAL CUBIC-BASED MONATOMIC SUBLATTICES

Lattice	Sublattice	Basis vectors chosen	Conditions on Coordinate Integers (ℓ_1, ℓ_2, ℓ_3)	r_m^2	Units of r_m^2	Notes
sc	sc (cube corners)	$(a/2)\mathbf{i}, (a/2)\mathbf{j}, (a/2)\mathbf{k}$	ℓ_1, ℓ_2, ℓ_3 all even	$r_m^2 = 4(m + n)$	$(a/2)^2$	a, b
bcc	bc (body centers)	$(a/2)\mathbf{i}, (a/2)\mathbf{j}, (a/2)\mathbf{k}$	ℓ_1, ℓ_2, ℓ_3 all odd	$r_m^2 = (8m - 5)$	$(a/2)^2$	a, c
fcc	fc (face centers)	$(a/2)\mathbf{i}, (a/2)\mathbf{j}, (a/2)\mathbf{k}$	exactly two of ℓ_1, ℓ_2, ℓ_3 must be odd	$r_m^2 = 2(2m - 1)$	$(a/2)^2$	a, c
ecc	ec (edge centers)	$(a/2)\mathbf{i}, (a/2)\mathbf{j}, (a/2)\mathbf{k}$	exactly two of ℓ_1, ℓ_2, ℓ_3 must be even	$r_m^2 = (4m - 3)$	$(a/2)^2$	a, c
diamond	I (unshifted fcc lattice)	$(a/4)\mathbf{i}, (a/4)\mathbf{j}, (a/4)\mathbf{k}$	ℓ_1, ℓ_2, ℓ_3 all even $\ell_1 + \ell_2 + \ell_3$ of the form $4s$	$r_m^2 = 8(m + n)$	$(a/4)^2$	a, b, d
diamond	II (fcc lattice shifted by $(a/4, a/4, a/4)$)	$(a/4)\mathbf{i}, (a/4)\mathbf{i}, (a/4)\mathbf{k}$	ℓ_1, ℓ_2, ℓ_3 all odd $\ell_1 + \ell_2 + \ell_3$ of the form $4s - 1$	$r_m^2 = (8m - 5)$	$(a/4)^2$	a, c

a. In the r_m^2 column m is an index numbering shells of neighbors *within each sublattice.*

b. n is an integer which starts at 0 and is incremented by 1 every time the formula (including previous increments) predicts an r_m^2 value of the form $4^r(8s + 7)$ where r and s are integers $\geqq 0$. (See also Table II and discussion in text.)

c. The radius formula in this case never predicts any r_m^2 values of the forbidden form $4^r(8s + 7)$.

d. Alternatively the sc and fc sublattices can be combined and used directly by multiplying the r_m^2 values of the fcc sublattices by 4 to convert to the $a/4$ basis chosen for diamond.

into equation (2) and identifying "m" in the resulting expression.

The diamond sublattice I can be worked out directly using the rules given in Table III or can be formed from the results of sc and fc by converting the r^2 values from the $a/2$ basis to the $a/4$ basis, interleaving the sc and fc sublattices to form a full fcc lattice, and renumbering the shells. The renumbering of shells invalidates the previous r_m^2 formulae but a new formula is found for the full fcc lattice. The diamond sublattice II enumeration proceeds in a straightforward manner but requires explicit counting of a greater number of points since there is less conveniently useable symmetry. In this case none of the simpler sublattices can be used directly since diamond II lacks the full cubic symmetries.

The simple monatomic lattices and sublattices can now be combined to describe physically interesting binary crystals. We adopt the following notation for sublattices of binary crystals: One X atom of the compound X_mY_n is chosen to lie at the origin and all other X atoms are said to occupy sublattice I_X. Sublattice II_X consists of all Y atoms when X is at the origin. Similarly, if Y is at the origin then Y atoms occupy sublattice I_Y and X atoms occupy II_Y. The distinction between I_X and I_Y or II_X and II_Y disappears for compounds of the type X_mY_m where all atoms of X and Y could be interchanged without any physically observable effect. The zincblende lattice, shown in the lower portion of Fig. 2, differs from diamond only in that sublattices I and II are occupied by different atomic species. Figure 3 shows three more commonly observed binary lattices: NaCl, CsCl, and CaF_2. In Table IV we show how these lattices can be formed from the basic cubic sublattices. Thus, for example, a table of r_m^2 and Z_m values for NaCl sublattice I (Na neighbors if Na is at the origin or Cl neighbors if Cl is at the origin) is composed of values from the sc and fc tables arranged in order of increasing r_m^2. The NaCl II sublattice is obtained by combining the bc and ec sublattices. All final tables of r_m^2 and Z_m will be given in Section III.

2.2 *Hexagonal-Based Lattices*

The hexagonal based lattices to be considered here are: (*i*) monatomic hexagonal close-packed (hcp) and (*ii*) wurtzite. These lattices are pictured in Fig. 4 along with a diagram of the basal plane showing how the basis vectors are chosen. The origin is placed at a corner of the hexagonal prism and the z-axis is taken along the c-axis of the crystal. The x and y axes are chosen as shown in Fig. 4 and the basis vectors are $(a/2)\mathbf{i}$, $[a/2(3)^{\frac{1}{2}}]\mathbf{j}$, $[2a/(6)^{\frac{1}{2}}]\mathbf{k}$. We assume that the hcp structure is the

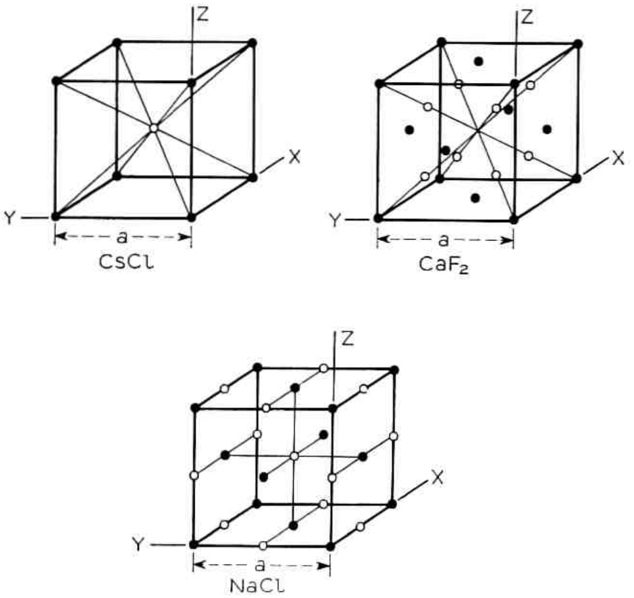

Fig. 3—The cubic-based binary lattices CsCl, NaCl, and CaF_2. CaF_2 is shown with Ca at the origin.

"ideal" one obtained by closest packing of spheres. In this case $c = (8/3)^{\frac{1}{2}}a$. The wurtzite structure is composed of two such hcp sublattices displaced along their common c-axis by an amount $u = (3/8)c$ and having atoms of different types occupying the two hcp sublattices. The unshifted sublattice will be called I and the shifted sublattice will be II. The enumeration of neighbors proceeds as in the cases already discussed and need not be detailed again. By inspection of the planes of lattice points one can obtain[5] the following conditions for ℓ_1, ℓ_2, and ℓ_3 for the hcp lattice:

$$\frac{3\ell_1 - \ell_2 + (-1)^{\ell_3} - 1}{6} = \text{integer}. \tag{3}$$

To identify points of the wurtzite II sublattice one can first locate points of the hcp lattice (wurtzite I) using equation (3) and then add $(\frac{3}{8})c$ to their z coordinates. The r_m^2 values are not integers for hcp and wurtzite because the separations between x-, y-, and z-planes of atoms are not related by rational numbers. Thus no $r_m^2 = f(m)$ formulae are expected. The radii are related to the coordinate integers by the follow-

TABLE IV—DECOMPOSITION OF CUBIC-BASED BINARY LATTICES INTO FUNDAMENTAL SUBLATTICES

Lattice	Sublattice	Basis Vectors	Equivalent Monatomic Lattice
NaCl	I Same as ion at origin	(a/2)i, (a/2)j, (a/2)**k**	fcc (sc + fc)
	II Opposite from ion at origin	(a/2)i, (a/2)j, (a/2)**k**	fcc shifted by (a/2, 0, 0) or (ec + bc)
CsCl	I Same	(a/2)i, (a/2)j, (a/2)**k**	sc
	II Opposite	(a/2)i, (a/2)j, (a/2)**k**	bc
Zincblende	I Same	(a/4)i, (a/4)j, (a/4)**k**	diamond I
	II Opposite	(a/4)i, (a/4)j, (a/4)**k**	diamond II
CaF_2	Ca Sublattice I_A Ca at orgin	(a/4)i, (a/4)j, (a/4)**k**	diamond I
	F Sublattice II_A Ca at origin	(a/4)i, (a/4)j, (a/4)k	bc with cube edge $a/2$
	F Sublattice I_B F at origin	(a/4)i, (a/4)j, (a/4)**k**	sc
	Ca Sublattice II_B F at origin	(a/4)i, (a/4)j, (a/4)**k**	diamond II

ing formulae

$$r^2 = \frac{3\ell_1^2 + \ell_2^2 + 8\ell_3^2}{12} \text{ (hcp)} \tag{4}$$

and

$$r^2 = \frac{6\ell_1^2 + 2\ell_2^2 + (4\ell_3 + 3)^2}{24} \text{ (wurtzite II)}. \tag{5}$$

III. TABLES

Tables V–VIII contain shell parameters for the basic monatomic lattices and sublattices. Table VIII can be used for both diamond and zincblende. In each case the shell number m refers to the m-th shell of neighbors *on that sublattice.* In order to combine two or more sublattices one must convert the r_m^2 values to the same basis (a, $a/2$, $a/4$, and so on) and interweave the appropriate columns in order of increasing r^2. The sc table includes r_m^2 columns for three choices of basis vector lengths:

a, $a/2$ and $a/4$. Tables IX–XIII give the corresponding results for CsCl, NaCl, CaF_2 and wurtzite.

The cubic based tables are easily extended to higher shell number by using the methods described earlier. In this regard the shell radius formulae are particularly useful if the search for lattice points is done by hand. For the diamond II and hexagonal based lattices, however, one must resort to computer enumeration and summation of allowed combinations of coordinates.

It should be pointed out here that there is a possible complication which can arise in utilizing these tables in physical applications involving more complex lattices. As was indicated in Section II, many of the shells in Tables V–XIII are degenerate. Although points on the same shell always belong to the same sublattice (for all sublattices which have been defined here), it is possible in some cases that different points

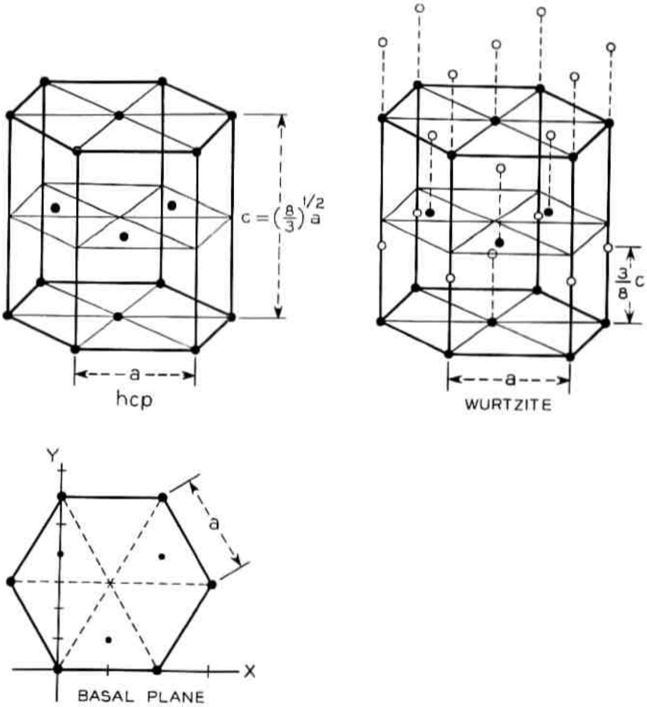

Fig. 4—Geometry of the hcp and wurtzite lattices. Wurtzite consists of two hcp lattices shifted $\frac{3}{8}c$ along their common c axis. The two hcp sublattices of wurtzite are occupied by different atomic species. $\mathbf{x}$ = (a/2)i; $\mathbf{y}$ = [a/2(3)$^{1/2}$]j; $\mathbf{z}$ = [2a/(6)$^{1/2}$$\mathbf{k}$ = (c/2)$\mathbf{k}$.

TABLE V—SHELL PARAMETERS FOR THE SC LATTICES*

Shell	$r^2(a^2)$	$r^2(a^2/4)$	$r^2(a^2/16)$	Z
1 (1)†	1	4	16	6
2 (2)	2	8	32	12
3 (3)	3	12	48	8
4 (4)	4	16	64	6
5 (5)	5	20	80	24
6 (6)	6	24	96	24
7 (8)	8	32	128	12
8 (9)	9	36	144	30
9 (10)	10	40	160	24
10 (11)	11	44	176	24
11 (12)	12	48	192	8
12 (13)	13	52	208	24
13 (14)	14	56	224	48
14 (16)	16	64	256	6
15 (17)	17	68	272	48
16 (18)	18	72	288	36
17 (19)	19	76	304	24
18 (20)	20	80	320	24
19 (21)	21	84	336	48
20 (22)	22	88	352	24
21 (24)	24	96	384	24
22 (25)	25	100	400	30
23 (26)	26	104	416	72
24 (27)	27	108	432	32
25 (29)	29	116	464	72
26 (30)	30	120	480	48
27 (32)	32	128	512	12
28 (33)	33	132	528	48
29 (34)	34	136	544	48
30 (35)	35	140	560	48
31 (36)	36	144	576	30
32 (37)	37	148	592	24
33 (38)	38	152	608	72
34 (40)	40	160	640	24
35 (41)	41	164	656	96
36 (42)	42	168	672	48
37 (43)	43	172	688	24
38 (44)	44	176	704	24
39 (45)	45	180	720	72
40 (46)	46	184	736	48
41 (48)	48	192	768	8
42 (49)	49	196	784	54
43 (50)	50	200	800	84
44 (51)	51	204	816	48
45 (52)	52	208	832	24
46 (53)	53	212	848	72
47 (54)	54	216	864	96
48 (56)	56	224	896	48
49 (57)	57	228	912	48
50 (58)	58	232	928	24

* For convenience, r^2 has been given for three choices of basis-vector lengths [a^2, $(a/2)^2$, and $(a/4)^2$].

† Numbers in parentheses conform to the notation of Refs. 1 and 2 in which "missing shells" are included in the sequential numbering as discussed in the text. If these shell numbers are used one must set $n = 0$ in the radius formula.

TABLE VI—SHELL PARAMETERS FOR THE BC AND EC SUBLATTICES*

Shell	bc Sublattice		ec Sublattice	
	r^2	Z	r^2	Z
1	3	8	1	6
2	11	24	5	24
3	19	24	9	30
4	27	32	13	24
5	35	48	17	48
6	43	24	21	48
7	51	48	25	30
8	59	72	29	72
9	67	24	33	48
10	75	56	37	24
11	83	72	41	96
12	91	48	45	72
13	99	72	49	54
14	107	72	53	72
15	115	48	57	48
16	123	48	61	72
17	131	120	65	96
18	139	72	69	96
19	147	56	73	48
20	155	96	77	96
21	163	24	81	102
22	171	120	85	48
23	179	120	89	144
24	187	48	93	48
25	195	96	97	48
26	203	96	101	168
27	211	72	105	96
28	219	96	109	72
29	227	120	113	96
30	235	48	117	120
31	243	104	121	78
32	251	168	125	144
33	259	96	129	144
34	267	48	133	48
35	275	120	137	96
36	283	72	141	96
37	291	96	145	96
38	299	192	149	168
39	307	72	153	144
40	315	144	157	72
41	323	96	161	192
42	331	72	165	96
43	339	144	169	78
44	347	120	173	168
45	355	96	177	48
46	363	104	181	120
47	371	192	185	192
48	379	72	189	192
49	387	120	193	48
50	395	192	197	120

[*] r^2 is in units of $(a/2)^2$.

Table VII—Shell Parameters for the fc and hcp Sublattices*

Shell	fc Sublattice		hcp Lattice	
	r^2	Z	r^2	Z
1	2	12	1.00	12
2	6	24	2.00	6
3	10	24	2.67	2
4	14	48	3.00	18
5	18	36	3.67	12
6	22	24	4.00	6
7	26	72	5.00	12
8	30	48	5.67	12
9	34	48	6.00	6
10	38	72	6.33	6
11	42	48	6.67	12
12	46	48	7.00	24
13	50	84	7.33	6
14	54	96	8.33	12
15	58	24	9.00	12
16	62	96	9.67	24
17	66	96	10.00	12
18	70	48	10.33	12
19	74	120	10.67	2
20	78	48	11.00	12
21	82	48	11.33	6
22	86	120	11.67	24
23	90	120	12.00	6
24	94	96	12.33	12
25	98	108	13.00	24
26	102	48	13.67	12
27	106	72	14.33	6
28	110	144	14.67	24
29	114	96	15.00	12
30	118	72	15.33	12
31	122	120	15.67	24
32	126	144	16.00	6
33	130	48	16.33	12
34	134	168	17.00	24
35	138	96	17.67	24
36	142	48	18.00	18
37	146	192	18.33	12
38	150	120	18.67	12
39	154	96	19.00	24
40	158	96	19.67	12
41	162	120	20.33	12
42	166	120	21.00	36
43	170	144	21.67	24
44	174	144	22.00	12
45	178	96	22.33	18
46	182	144	22.67	12
47	186	144	23.00	24
48	190	48	23.33	12
49	194	240	23.67	48
50	198	120	24.00	2

* For the fc sublattices r^2 is in units of $(a/2)^2$. For hcp r^2 is expressed in units of a^2.

TABLE VIII—SHELL PARAMETERS FOR SUBLATTICES OF THE DIAMOND OR ZINCBLENDE LATTICES[*]

Shell	Sublattice I		Sublattice II	
	r^2	Z	r^2	Z
1 (1)[†]	8	12	3	4
2 (2)	16	6	11	12
3 (3)	24	24	19	12
4 (4)	32	12	27	16
5 (5)	40	24	35	24
6 (6)	48	8	43	12
7 (7)	56	48	51	24
8 (8)	64	6	59	36
9 (9)	72	36	67	12
10 (10)	80	24	75	28
11 (11)	88	24	83	36
12 (12)	96	24	91	24
13 (13)	104	72	99	36
14 (15)	120	48	107	36
15 (16)	128	12	115	24
16 (17)	136	48	123	24
17 (18)	144	30	131	60
18 (19)	152	72	139	36
19 (20)	160	24	147	28
20 (21)	168	48	155	48
21 (22)	176	24	163	12
22 (23)	184	48	171	60
23 (24)	192	8	179	60
24 (25)	200	84	187	24
25 (26)	208	24	195	48
26 (27)	216	96	203	48
27 (28)	224	48	211	36
28 (29)	232	24	219	48
29 (31)	248	96	227	60
30 (32)	256	6	235	24
31 (33)	264	96	243	52
32 (34)	272	48	251	84
33 (35)	280	48	259	48
34 (36)	288	36	267	24
35 (37)	296	120	275	60
36 (38)	304	24	283	36
37 (39)	312	48	291	48
38 (40)	320	24	299	96
39 (41)	328	48	307	36
40 (42)	336	48	315	72
41 (43)	344	120	323	48
42 (44)	352	24	331	36
43 (45)	360	120	339	72
44 (47)	376	96	347	60
45 (48)	384	24	355	48
46 (49)	392	108	363	52
47 (50)	400	30	371	96
48 (51)	408	48	379	36
49 (52)	416	72	387	60
50 (53)	424	72	395	96

[*] r^2 is in units of $(a/4)^2$.

[†] Numbers in parentheses apply *only* to sublattice I and conform to the notation of Refs. 1 and 2 in which "missing shells" are included in the sequential numbering as discussed in the text. If these shell numbers are used one must set $n = 0$ in the radius formula.

Table IX—Shell Parameters for Sublattices of the CsCl Lattice*

Shell	Sublattice I		Sublattice II	
	r^2	Z	r^2	Z
1	4	6	3	8
2	8	12	11	24
3	12	8	19	24
4	16	6	27	32
5	20	24	35	48
6	24	24	43	24
7	32	12	51	48
8	36	30	59	72
9	40	24	67	24
10	44	24	75	56
11	48	8	83	72
12	52	24	91	48
13	56	48	99	72
14	64	6	107	72
15	68	48	115	48
16	72	36	123	48
17	76	24	131	120
18	80	24	139	72
19	84	48	147	56
20	88	24	155	96
21	96	24	163	24
22	100	30	171	120
23	104	72	179	120
24	108	32	187	48
25	116	72	195	96
26	120	48	203	96
27	128	12	211	72
28	132	48	219	96
29	136	48	227	120
30	140	48	235	48
31	144	30	243	104
32	148	24	251	168
33	152	72	259	96
34	160	24	267	48
35	164	96	275	120
36	168	48	283	72
37	172	24	291	96
38	176	24	299	192
39	180	72	307	72
40	184	48	315	144
41	192	8	323	96
42	196	54	331	72
43	200	84	339	144
44	204	48	347	120
45	208	24	355	96
46	212	72	363	104
47	216	96	371	192
48	224	48	379	72
49	228	48	387	120
50	232	24	395	192

* r^2 is in units of $(a/2)^2$.

TABLE X—SHELL PARAMETERS FOR SUBLATTICES OF THE NaCl LATTICE*

Shell	Sublattice I		Sublattice II	
	r^2	Z	r^2	Z
1	2	12	1	6
2	4	6	3	8
3	6	24	5	24
4	8	12	9	30
5	10	24	11	24
6	12	8	13	24
7	14	48	17	48
8	16	6	19	24
9	18	36	21	48
10	20	24	25	30
11	22	24	27	32
12	24	24	29	72
13	26	72	33	48
14	30	48	35	48
15	32	12	37	24
16	34	48	41	96
17	36	30	43	24
18	38	72	45	72
19	40	24	49	54
20	42	48	51	48
21	44	24	53	72
22	46	48	57	48
23	48	8	59	72
24	50	84	61	72
25	52	24	65	96
26	54	96	67	24
27	56	48	69	96
28	58	24	73	48
29	62	96	75	56
30	64	6	77	96
31	66	96	81	102
32	68	48	83	72
33	70	48	85	48
34	72	36	89	144
35	74	120	91	48
36	76	24	93	48
37	78	48	97	48
38	80	24	99	72
39	82	48	101	168
40	84	48	105	96
41	86	120	107	72
42	88	24	109	72
43	90	120	113	96
44	94	96	115	48
45	96	24	117	120
46	98	108	121	78
47	100	30	123	48
48	102	48	125	144
49	104	72	129	144
50	106	72	131	120

* r^2 is in units of $(a/2)^2$.

Table XI—Shell Parameters for Sublattices of the CaF_2 Lattice When Ca is Taken To Be at the Origin*

Shell	Sublattice I_{Ca}		Sublattice II_{Ca}	
	r^2	Z	r^2	Z
1	8	12	3	8
2	16	6	11	24
3	24	24	19	24
4	32	12	27	32
5	40	24	35	48
6	48	8	43	24
7	56	48	51	48
8	64	6	59	72
9	72	36	67	24
10	80	24	75	56
11	88	24	83	72
12	96	24	91	48
13	104	72	99	72
14	120	48	107	72
15	128	12	115	48
16	136	48	123	48
17	144	30	131	120
18	152	72	139	72
19	160	24	147	56
20	168	48	155	96
21	176	24	163	24
22	184	48	171	120
23	192	8	179	120
24	200	84	187	48
25	208	24	195	96
26	216	96	203	96
27	224	48	211	72
28	232	24	219	96
29	248	96	227	120
30	256	6	235	48
31	264	96	243	104
32	272	48	251	168
33	280	48	259	96
34	288	36	267	48
35	296	120	275	120
36	304	24	283	72
37	312	48	291	96
38	320	24	299	192
39	328	48	307	72
40	336	48	315	144
41	344	120	323	96
42	352	24	331	72
43	360	120	339	144
44	376	96	347	120
45	384	24	355	96
46	392	108	363	104
47	400	30	371	192
48	408	48	379	72
49	416	72	387	120
50	424	72	395	192

* r^2 is in units of $(a/4)^2$.

Table XII—Shell Parameters for Sublattices of the CaF_2 Lattice When F Is Taken To Be at the Origin*

Shell	Sublattice I_F		Sublattice II_F	
	r^2	Z	r^2	Z
1	4	6	3	4
2	8	12	11	12
3	12	8	19	12
4	16	6	27	16
5	20	24	35	24
6	24	24	43	12
7	32	12	51	24
8	36	30	59	36
9	40	24	67	12
10	44	24	75	28
11	48	8	83	36
12	52	24	91	24
13	56	48	99	36
14	64	6	107	36
15	68	48	115	24
16	72	36	123	24
17	76	24	131	60
18	80	24	139	36
19	84	48	147	28
20	88	24	155	48
21	96	24	163	12
22	100	30	171	60
23	104	72	179	60
24	108	32	187	24
25	116	72	195	48
26	120	48	203	48
27	128	12	211	36
28	132	48	219	48
29	136	48	227	60
30	140	48	235	24
31	144	30	243	52
32	148	24	251	84
33	152	72	259	48
34	160	24	267	24
35	164	96	275	60
36	168	48	283	36
37	172	24	291	48
38	176	24	299	96
39	180	72	307	36
40	184	48	315	72
41	192	8	323	48
42	196	54	331	36
43	200	84	339	72
44	204	48	347	60
45	208	24	355	48
46	212	72	363	52
47	216	96	371	96
48	224	48	379	36
49	228	48	387	60
50	232	48	395	96

* r^2 is in units of $(a/4)^2$.

Table XIII—Shell Parameters for Sublattices of the Wurtzite Lattice*

Shell	Sublattice I		Sublattice II	
	r^2	Z	r^2	Z
1	1.000	12	.375	4
2	2.000	6	1.042	1
3	2.667	2	1.375	9
4	3.000	18	2.042	6
5	3.667	12	2.375	9
6	4.000	6	3.375	9
7	5.000	12	3.708	3
8	5.667	12	4.042	6
9	6.000	6	4.375	18
10	6.333	6	4.708	3
11	6.667	12	5.042	7
12	7.000	24	5.375	3
13	7.333	6	5.708	6
14	8.333	12	6.042	6
15	9.000	12	6.375	12
16	9.667	24	7.042	1
17	10.000	12	7.375	15
18	10.333	12	7.708	6
19	10.667	2	8.042	24
20	11.000	12	8.375	9
21	11.333	6	8.708	3
22	11.667	24	9.042	6
23	12.000	6	9.375	12
24	12.333	12	9.708	9
25	13.000	24	10.042	12
26	13.667	12	10.375	9
27	14.333	6	10.708	3
28	14.667	24	11.042	6
29	15.000	12	11.375	6
30	15.333	12	11.708	9
31	15.667	24	12.042	12
32	16.000	6	12.375	21
33	16.333	12	12.708	6
34	17.000	24	13.042	6
35	17.667	24	13.375	15
36	18.000	18	13.708	12
37	18.333	12	14.042	30
38	18.667	12	14.375	18
39	19.000	24	14.708	3
40	19.667	12	15.042	1
41	20.333	12	15.708	12
42	21.000	38	16.042	12
43	21.667	24	16.375	27
44	22.000	12	17.042	12
45	22.333	18	17.375	9
46	22.667	12	17.708	9
47	23.000	24	18.042	18
48	23.333	12	18.375	16
49	23.667	48	18.708	6
50	24.000	2	19.042	12

* r^2 is in units of a^2.

on the same shell can occupy physically distinguishable sites in the lattice (that is, sites which, even though they belong to the same sublattice, have differing local environments). In addition to this, interactions between electrons and holes bound to donors and acceptors may depend upon the relative orientations of the electron and hole wavefunctions, the vector separation between ions, and the crystallographic axes. Thus in the interpretation of pair spectra, for example, one may expect energy splittings in such cases to cause deviations from the spacings and magnitudes predicted on the basis of neighbor tables.[1,2,8,9] Wurtzite is particularly complicated in this respect, providing a variety of local symmetries for donor-acceptor pairs involving substitutional and/or interstitial ions.[9,10] In the absence of externally imposed asymmetries, however, lattice sites on the same sublattice will usually be physically indistinguishable. Among the structures considered here, wurtzite is the sole exception. For this reason it was felt that the tables would be unnecessarily complicated by the inclusion of any information regarding degeneracies or coordinate types. In most cases, however, the computer programs were written in such a way as to preserve this information and it is available from the authors upon request.

IV. FURTHER APPLICATIONS

The crystal structures which have been explicitly discussed account for the vast majority of binary compounds $X_M Y_N$ and virtually all of the important XY compounds. This is indicated in Table XIV where we have shown the crystal structures of the common binary compounds formed by combining elements from groupings IA, IIA, and IB–IVB* with elements from groups IVB–VIIB.[11] In Table XIV the important elemental and compound semiconductors, the oxides and chalcogenides of group IIA and IIB metals, the alkali halides, and the noble metal salts have been enclosed in heavy lines. It is seen that nearly all of these compounds crystallize in one of the structures which has been treated here. Furthermore, it is noted that many of the structures which were not treated explicitly are cubic-based or hexagonal-based so that the one might be able to utilize one or more of the basic sublattices calculated here.

In certain kinds of defect interaction calculations it may also be convenient to know the distribution of available interstitial sites as a function of distance from a given ion. Inspection of crystalline lattice

* The A and B notation used here for subgroups of the periodic table was chosen to agree with Frederikse[11] but is not uniform throughout the literature.

structures shows the most of them have regular arrays of preferred interstitial sites. These sites form a sublattice which can be treated by the methods already discussed. As an example we list in Table XV the first 25 shells of interstitial sites for the fcc lattice. These are at ec and bc sites. This table is the same as the table for Type III pair spectra given by Ryan and Miller.[2] Similar tables can be constructed for interstitials in other lattices.

V. SUMMARY

We have described a general method for obtaining radii and occupation numbers of m-th order shells of neighboring lattice points for cubic and hexagonal based lattices. The method described here will, in

Table XIV—A Summary of Inorganic Binary Compounds $X_M Y_N$ and Their Crystal Structures*

	Y			
X	IVB C, Si, Ge, Sn, Pb	VB N, P, As, Sb	VIB O, S, Se, Te	VIIB F, Cl, Br, I
IVB C, Si, Ge, Sn, Pb	X, SiC 3,7	XY, X_3Y_2 1,8	XY, XY_2 1,4,6,7,8	XY_4 6,8
IIIB B, Al, Ga, In	—	XY 3,4	X_2Y_3 3,4,6,7	—
IIB Zn, Cd, Hg	—	X_3Y_2 6,7,8	XY 1,3,4	XY_2 5,6,7,8
IB Cu, Ag, Au	—	X_3Y, XY_2, X_2Y 6,7,8	XY, X_2Y, XY_2 1,2,5,6,7,8	XY 1,3,4
IA Li, Na, K, Rb, Cs	—	—	X_2Y 5	XY 1,2
IIA Be, Mg, Ca, Sr, Ba	X_2Y 5,7,8	X_3Y_2 6,7	XY 1,3,4	XY_2 5,6,7,8

1. NaCl Structure
2. CsCl Structure
3. Zincblende Structure (Diamond in case of element)
4. Wurtzite Structure
5. CaF_2 Structure
6. Other Cubic-based Structure
7. Other Hexagonal-based Structure
8. Complex

* Roman numerals and A or B refer to groups and subgroups of the periodic table in the notation of Ref. 11. Only representative compound-types are indicated and not all of the elements of any one group form in all of the combinations shown. See Ref. 11 for an extensive list of specific compounds.

Table XV—The First 25 Shells of Preferred Interstitial Sites for the fcc Lattice*

Shell	Interstitials	
	r^2	Z
1 (0)[†]	1	6
2 (1)	3	8
3 (2)	5	24
4 (4)	9	30
5 (5)	11	24
6 (6)	13	24
7 (8)	17	48
8 (9)	19	24
9 (10)	21	48
10 (12)	25	30
11 (13)	27	32
12 (14)	29	72
13 (16)	33	48
14 (17)	35	48
15 (18)	37	24
16 (20)	41	96
17 (21)	43	24
18 (22)	45	72
19 (24)	49	54
20 (25)	51	48
21 (26)	53	72
22 (28)	57	48
23 (29)	59	72
24 (30)	61	72
25 (32)	65	96

* This table was obtained by combining the first few shells of the bc and ec tables and is easily extended further. All r^2 values are in units of $(a/2)^2$.

[†] Numbers in parentheses conform to the notation of Refs. 1 and 2 in which "missing shells" are included in the sequential numbering as discussed in the text. Note that Ryan and Miller began the numbering with $m = 0$.

principle, work for *any* specific lattice if the basis vectors are properly chosen but is practical only in cases where the lattice contains no arbitrary angles or spacings. A general hexagonal lattice, for example, has an arbitrary c/a ratio which must be fixed before the basis vectors can be chosen. The principal results are contained in Tables V–XIII which contain shell parameters for simple building-block sublattices as well as physically interesting binary lattices. Extension of the tables to higher shell numbers and application of the tables to physical problems were discussed.

VI. ACKNOWLEDGMENTS

The authors would like to thank M. DiDomenico, Jr., for offering several helpful observations. We have also benefited from discussions with M. L. Howe and L. C. Luther.

REFERENCES

1. Thomas, D. G., Gershenzon, M., and Trumbore, F. A., "Pair Spectra and Edge Emission in Gallium Phosphide," Phys. Rev., *133,* No. 1A (January 1964), pp. A269–A279.
2. Ryan, F. M., and Miller, R. C., "Photoluminescence and Pair Spectrum in Boron Phosphide," Phys. Rev., *148,* No. 2 (August 1966), pp. 858–862.
3. Prener, J. S., "Ion Pairing in Solids," J. Chem. Phys., *25,* No. 6 (December 1956), pp. 1294.
4. Wiley, J. D., unpublished work.
5. Wood, V. E., "Enumerating Near Neighbors," Am. J. Phys., *33,* No. 8 (August 1965), pp. 632–634.
6. Ferris-Prabhu, A. V. M., "Enumerating Neighbors in Diamond-Like Lattices," Am. J. Phys., *34,* No. 8 (August 1966), pp. 645–646.
7. Griffin, H., *Elementary Theory of Numbers,* New York: McGraw-Hill, 1954, Chapter 10.
8. Patrick, L., "Donor and Acceptor Multipole Fields and their Effects in GaP," Phys. Rev., *180,* No. 3 (April 1969), pp. 794–796.
9. Henry, C. H., Faulkner, R. A., and Nassau, K., "Donor-Acceptor Pair Lines in Cadmium Sulfide," Phys. Rev., *183,* No. 3 (July 1969), pp. 798–806.
10. Boyn, R., "Polarization of Optical Absorption and Emission Bands due to Donor-Acceptor Associates in Wurtzite Type Crystals," Phys. Stat. Sol., *30* (1968), pp. 305–310.
11. Frederikse, H. P. R., *American Institute of Physics Handbook,* Edited by Dwight Gray, New York: McGraw-Hill, 1963, 2nd Ed., Chapter 9b.

An Asymmetric Encoding Scheme for Word Stuffing

By M. M. BUCHNER, JR.

(Manuscript received November 6, 1969)

The effectiveness of word stuffing for synchronization depends upon our ability to distinguish the stuff words from the data words at the destination and, thus, delete correctly the stuff words. If all input sequences are permitted, the data words must be encoded before stuffing occurs so that the stuff word can be distinct from the data words. In this paper, we give the code that, for a given redundancy, maximizes the minimum distance between the stuff word and any data word. This helps to prevent the loss of character synchronization because of transitions between data and stuff words due to transmission errors. In contrast to the perfect distance symmetries between the code words of the group codes normally encountered in error-control work, the primary virtue of the present code is its highly asymmetric distance structure.

I. INTRODUCTION

When a digital communication network is used for data transmission, it may be necessary to adjust transmission rates within the network to achieve synchronization. Word stuffing[1,2] is a technique that can be used for this purpose. The basic idea is to group the transmitted bits into words which we call data words. The data words are formed for transmission and are not related to any word structure that may exist in the customer's data stream. Stuff words, which are distinguishable from the data words, are inserted into the stream of data words at the transmitter. Thus, transmission rates can be adjusted within the network by inserting or deleting stuff words. At the destination, the stuff words are deleted whereas the data words are delivered to the customer. The effectiveness of word stuffing depends upon our ability to distinguish the stuff words from the data words at the destination and, thus, delete correctly the stuff words. If the stuff words are incorrectly deleted, bits will be inserted into or deleted from the customer's data stream. As a result, the received bits will be incorrectly formatted and incorrectly interpreted. When this occurs, we say that character synchronization is lost. Character synchronization is important in data transmission because once it is lost, subsequent bits are erroneously interpreted even if transmitted correctly.

Two problems arise. First, we require that all input sequences are allowed. Thus, redundancy must be added to the data words by an encoder so that it is possible to choose a stuff word that is distinct from the data words. Second, when transmission errors occur, it is possible for:

(*i*) a stuff word to be transformed into a data word,
(*ii*) a data word to be transformed into a stuff word, or
(*iii*) a data word to be transformed into another data word.

In most cases, (*i*) and (*ii*) are more serious than (*iii*) because of the resulting loss of character synchronization. The prevention of type (*iii*) errors is generally performed by the customer's terminal, when required, and is not considered in this paper.

Previously, Mattesich and Richters[1] proposed a format for the data words and the stuff word. The format results in the stuff word being distance one* from a data word. Therefore, a single transmission error can change a data word into the stuff word or vice versa with a corresponding loss of character synchronization.

We give an alternative encoding scheme that, for a given redundancy, maximizes the minimum distance between the stuff word and any data word. This helps to prevent the loss of character synchronization because of transitions between data and stuff words due to transmission errors. An implementation is given for arbitrary word size and redundancy. For a redundancy of one bit, a particularly simple encoding-decoding technique is described.

Some other methods of achieving synchronization by means of stuffing have been proposed. A word stuffing technique, proposed by Butman[2], achieves a distance d between the stuff word and any data word by inserting deliberate errors in certain data words at the transmitter to keep the data words at least distance d from the stuff word. Butman's technique requires some knowledge of the statistics of the transmitted data words for selection of the stuff word and a relatively large word size so that the deliberate insertion of errors is infrequent. It results in deliberate errors in the customer's data and prohibits the reception of

* The distance, frequently called the Hamming distance, between two binary words X and Y is the number of positions in which X and Y differ.

certain data words. The latter point is troublesome if the data words used in transmission are identical to the data words of the customer.

Pulse stuffing, rather than word stuffing, can be used to achieve synchronization. Individual pulses are inserted at the transmitter and a separate data link is used to signal the locations of the stuff pulses. References 3 through 6 are representative of the work in pulse stuffing.

II. PRELIMINARIES

A model of the processing before transmission is shown in Fig. 1. The input binary data stream is segmented into k-bit data words denoted by A, where

$$A = (a_k , a_{k-1} , \cdots , a_1).$$

We assume that all of the 2^k possible sequences for A are allowed. In order for the stuff word to be distinguishable from the data words, the alphabet is enlarged by adding redundancy to A. Thus, the encoder generates from A an n-tuple B where $n > k$ and

$$B = (b_n , b_{n-1} , \cdots , b_1).$$

The sequence B is transmitted. The dimensions of such a code are denoted by (n, k).

There are 2^n possible sequences for B of which 2^k are used to transmit data. Thus, there are $2^n - 2^k$ values of B that are not used for data but that can be used for other purposes including the stuff word. For $n > k$,

$$2^n - 2^k \geqq 2^k. \tag{1}$$

The reader should appreciate the tremendous flexibility available in the design of the coding scheme because, by (1), never more than half of the possible B sequences are used as data words.

The processing after transmission is indicated in Fig. 2. As shown later, it is possible to construct codes that have a minimum distance greater than one between the stuff word and any data word for practical values of n and k. For these codes, the first step at the destination is to delete the stuff word and all other received words "sufficiently close"

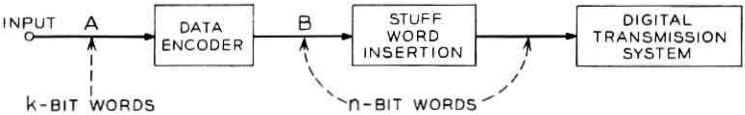

Fig. 1 — Processing before transmission.

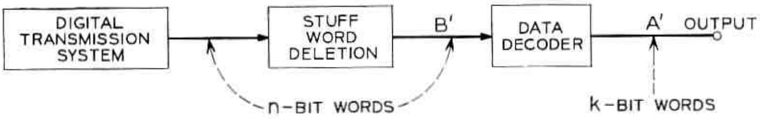

Fig. 2—Processing after transmission.

to the stuff word. In saying "sufficiently close," we mean that the received word is more likely the result of a stuff word corrupted by errors than a data word corrupted by errors. Let B' denote the n-tuples that are not deleted where

$$B' = (b'_n, b'_{n-1}, \cdots, b'_1).$$

The decoder operates on B' to form the k-bit word A' where

$$A' = (a'_k, a'_{k-1}, \cdots, a'_1).$$

The system functions properly if

$$A' = A.$$

At this point, we describe for reference the encoding and decoding schemes presented in Ref. 1. An (8, 7) code is used, that is, $n = 8$, $k = 7$. The encoder generates B from A by the relation

$$b_8 = 1$$

$$b_i = a_i \quad \text{for} \quad 1 \leqq i \leqq 7.$$

Thus,

$$B = (1, a_7, a_6, \cdots, a_1).$$

The stuff word is (00000001). At the receiver, only the word (00000001) is deleted as the stuff word; all other received words are decoded as data words. Thus, the stuff word is interpreted as a data word if one or more transmission errors occur. Alternatively, because the data word (10000001) is distance one from the stuff word, a single error can convert this data word into the stuff word. Any other data word requires at least two errors for conversion into the stuff word.

III. DESCRIPTION OF THE CODE

We construct a code that has one stuff word and 2^k data words such that the minimum distance between the stuff word and any data word is maximized. The problem is to divide the 2^n n-tuples into three sets; namely,

(*i*) the stuff word denoted by S,
(*ii*) the set D consisting of the 2^k data words, and
(*iii*) the set U consisting of the $2^n - 2^k - 1$ unused words,

such that, for given values of n and k, the minimum distance between S and any element of D is maximized.

Choose one of the n-tuples to be the stuff word S. For given n, k and S, let d_m denote the maximum possible minimum distance between S and any element of D. We determine d_m by observing that if all elements of D are to be at least distance d_m from S, then U must contain all n-tuples that are distance $d_m - 1$ or less from S. Thus, the set U contains

$$\begin{array}{lll} \text{the } \binom{n}{1} & n\text{-tuples distance } 1 & \text{from } S, \\ \text{the } \binom{n}{2} & n\text{-tuples distance } 2 & \text{from } S, \\ \quad\vdots & & \\ \text{the } \binom{n}{d_m - 1} & n\text{-tuples distance } d_m - 1 & \text{from } S, \end{array} \tag{2a}$$

and δ n-tuples distance greater than $d_m - 1$ from S such that

$$\sum_{l=1}^{d_m-1} \binom{n}{l} \leq 2^n - 2^k - 1 < \sum_{l=1}^{d_m} \binom{n}{l} \tag{2b}$$

and

$$\delta = 2^n - 2^k - 1 - \sum_{l=1}^{d_m-1} \binom{n}{l}. \tag{2c}$$

The data words are the remaining 2^k words.

The value of d_m is determined by (2b). From (2b), it follows that d_m is independent of S and is determined entirely by n and k. Also, it is easy to show that changing a code by adding* a constant n-tuple to all words simply rotates the code with no change in the distance properties, including d_m.

The code is specified by (2) up to the choice of the δ unused words that are at distance greater than $d_m - 1$ from S. While the choice of these δ words does not alter d_m, the probability that a data word is transposed into the stuff word is minimized if the δ words are all chosen

* The addition is component-by-component modulo two addition and is denoted by $\oplus$.

to be distance d_m from S. In practice, the choice of the δ words does not change this probability substantially and it appears preferable to assign the δ words to simplify encoder-decoder design.

The case $n = k + 1$ is of interest because of the low redundancy. In Appendix A, it is shown that for $n = k + 1$,

$$d_m = \left[\frac{k}{2}\right] + 1$$

where $[x]$ denotes the largest integer less than or equal to x. Also, for $n = k + 1$ and k even, it is shown in Appendix A that $\delta = 0$. A plot of d_m versus n for $n = k + 1$ and $n = k + 2$ is given in Fig. 3.

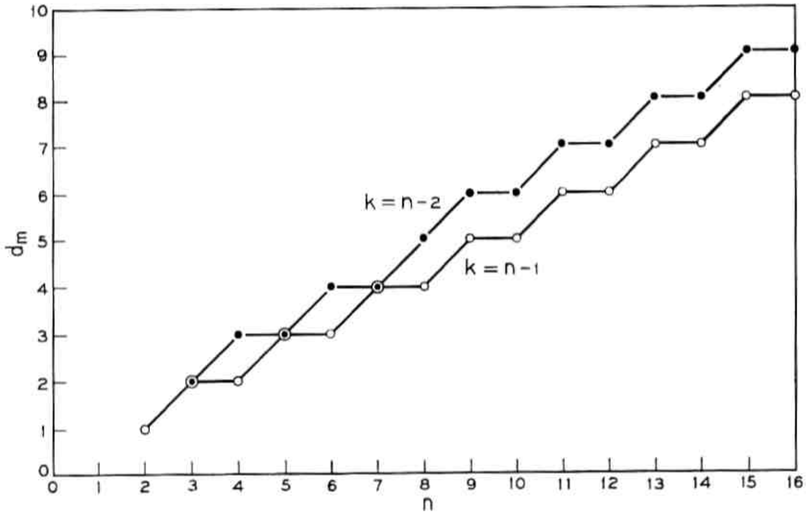

Fig. 3—Maximum minimum distance d_m for various n.

Transmission of the all-zero word can be avoided by choosing S so that the all-zero word is one of the unused words. However, it may be convenient, particularly when detecting stuff words, for S to be the all-zero word because then distance from S is equivalent to weight* and can be computed by counting ones. These two objectives can be simultaneously satisfied as follows. Design the encoder and stuff word using a code C in which S is the all-zero word. Let S' be one of the unused words in C. Add S' to each word immediately before transmission and

* The weight of a binary word X is the number of nonzero components in X and is denoted by $w(X)$.

add S' to each word immediately after transmission. The double addition of S' permits the suppression of the all-zero word for transmission but is transparent for the encoding, stuff word detection, and decoding operations.

Because the data words and the stuff word form a subset of the set of all possible n-tuples, we may, on the average, transmit an unequal number of zeros and ones. By varying $w(S)$ from 0 to n, the relative number of zeros and ones in the data words can be varied from mostly ones to mostly zeros.

IV. ENCODING, DECODING AND STUFF WORD DETECTION: GENERAL CASE

In this section, we specify an encoder that achieves d_m for arbitrary n and k. Let the stuff word be the all-one n-tuple. A necessary and sufficient condition for the encoder to achieve d_m is that each A must be encoded into a unique B such that the maximum weight of any B is $n - d_m$.

We begin by regarding each A as the k-bit natural binary representation of some integer α, $0 \leqq \alpha \leqq 2^k - 1$. Thus,

$$A = B_k(\alpha)$$

where

$$\alpha = \sum_{i=1}^{k} 2^{i-1} a_i \ .$$

We can construct a table that, for each α, gives the corresponding A and B sequences. The entries are in the order of increasing α. For illustration, consider the (9, 7) code where $d_m = 6$. The first 24 entries for the (9, 7) code are shown in Table 1. The A column corresponds to counting in binary from the all-zero k-tuple to the all-one k-tuple. The B column is also formed by counting in binary except that all n-tuples with weight greater than $n - d_m$ are omitted from the count. It follows that the maximum weight of any B is $n - d_m$. The arrows in Table 1 indicate where 9-tuples with weight greater than three have been omitted in the B column.

Let us examine the counting in the B column of Table 1 in greater detail. Consider counting to the B sequence for $\alpha = 22$, that is,

$$(0000\underset{\text{position 5}\,\uparrow}{1}\underset{\uparrow\,\text{position 4}}{1}000). \tag{3}$$

First, count to (000010000). At this point, all 4-tuples of weight not

Table I—Relationship Between A and B for (9, 7) code; Only the First 24 Values of A are Shown.

α	$A = B_7(\alpha)$	B
0	0000000	000000000
1	0000001	000000001
2	0000010	000000010
3	0000011	000000011
4	0000100	000000100
5	0000101	000000101
6	0000110	000000110
7	0000111	000000111
8	0001000	000001000
9	0001001	000001001
10	0001010	000001010
11	0001011	000001011
12	0001100	000001100
13	0001101	000001101
14	0001110	000001110 ←
15	0001111	000010000
16	0010000	000010001
17	0010001	000010010
18	0010010	000010011
19	0010011	000010100
20	0010100	000010101
21	0010101	000010110 ←
22	0010110	000011000
23	0010111	000011001

greater than three have been used in positions one through four. Then, count to the sequence in (3) by using all 3-tuples of weight not greater than two in positions one through three while keeping a one in position five. As shown in Fig. 4, the value of α associated with the sequence in (3) has two components. The first component, denoted by $\alpha_1(5)$, is the number of B sequences used in counting to obtain the one in position five. Similarly, the second component, denoted by $\alpha_2(4)$, is the number of B sequences used in counting to obtain the one in position four. Accordingly,

$$\alpha = \alpha_1(5) + \alpha_2(4)$$

where

$$\alpha_1(5) = \sum_{i=0}^{3} \binom{4}{i} = 15,$$

$$\alpha_2(4) = \sum_{i=0}^{2} \binom{3}{i} = 7.$$

The above ideas can be formalized so that, for an arbitrary B (de-

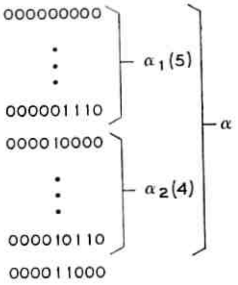

Fig. 4—Counting to determine α.

noted by B_0), it is possible to determine the corresponding value of α (denoted by α_0). Let $\omega = w(B_0)$, $\omega \leqq n - d_m$. Let $\beta_1, \beta_2, \cdots, \beta_\omega$ denote the position numbers of the ω nonzero components of B_0 where

$$\beta_1 > \beta_2 > \cdots > \beta_\omega .$$

For example, if $B_0 = (000011000)$, $\omega = 2$ and $\beta_1 = 5$, $\beta_2 = 4$. Let $\alpha_i(\beta_i)$ denote the contribution of the one in position β_i to α_0, that is,

$$\alpha_0 = \sum_{i=1}^{\omega} \alpha_i(\beta_i).$$

Observe that $\alpha_i(\beta_i)$ is the number of sequences in the B column from the sequence whose nonzero components are in positions $\beta_1, \beta_2, \cdots, \beta_{i-1}$ to the sequence whose nonzero components are in positions $\beta_1, \beta_2, \cdots, \beta_i$. Thus, $\alpha_i(\beta_i)$ is the number of $(\beta_i - 1)$-tuples of weight not greater than $n - d_m - i + 1$ and is given by

$$\alpha_i(\beta_i) = \sum_{m=0}^{\mathrm{Min}(n-d_m-i+1,\beta_i-1)} \binom{\beta_i - 1}{m}. \tag{4}$$

It follows that

$$\alpha_0 = \sum_{i=1}^{\omega} \sum_{m=0}^{\mathrm{Min}(n-d_m-i+1,\beta_i-1)} \binom{\beta_i - 1}{m}.$$

Equation (4) can be used to find B_0 when α_0 is given, that is, to design an encoder.* For each value of i, $1 \leqq i \leqq n - d_m$, construct an array. In the ith array, list j and $\alpha_i(j)$ as j runs from one to $n - i + 1$. To encode α_0, find in the first array the largest j (denoted by j_1) such that

$$\alpha_1(j_1) \leqq \alpha_0 .$$

* The ideas in this paragraph are illustrated by a numerical example in Appendix B.

Next, find in the second array the largest j (denoted by j_2) such that

$$\alpha_2(j_2) \leqq \alpha_0 - \alpha_1(j_1).$$

In the ith array, find the largest j such that

$$\alpha_i(j_i) \leqq \alpha_0 - \sum_{m=1}^{i-1} \alpha_m(j_m).$$

The process continues until

$$\sum_{m=1}^{\omega'} \alpha_m(j_m) = \alpha_0$$

for some $\omega' \leqq n - d_m$. It follows that B_0 has ones in positions $j_1, j_2, \cdots, j_{\omega'}$ and zeros in all remaining positions. Also, $\omega' = \omega$, the weight of B_0 .

It is possible to obtain a recurrence relation for the elements of the arrays. It is shown in Appendix C that

$$\alpha_i(j) + \alpha_{i+1}(j) = \alpha_i(j + 1) \tag{5}$$

for $1 \leqq j \leqq n - i, 1 \leqq i \leqq n - d_m - 1$. By knowing that $\alpha_i(1) = 1$ for $1 \leqq i \leqq n - d_m$ and that $\alpha_{n-d_m}(j) = j$ for $1 \leqq j \leqq d_m + 1$, equation (5) can be used to generate elements of the arrays.

Also, equation (5) can be used to construct an encoder directly. First, we specify the subtraction and storage device shown in Fig. 5. Let R denote the integer stored in the device. The output is equal to the integer stored in the device, that is, R. Let the input be an integer R', $0 \leqq R' \leqq R$. When the device is activated, the number stored in the device becomes $R - R'$ and, thus, the output also takes the value $R - R'$.

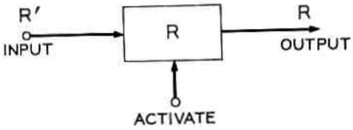

Fig. 5—Subtraction and storage device.

The encoder operates as shown in Fig. 6. The storage devices are preset so that $R_i = \alpha_i\ (n - i + 1), 1 \leqq i \leqq n - d_m$, and $T = \alpha$, the integer representation of the A sequence to be encoded.* Position the

* For the (9, 7) code, the preset values are $R_1 = 93$, $R_2 = 29$ and $R_3 = 7$.

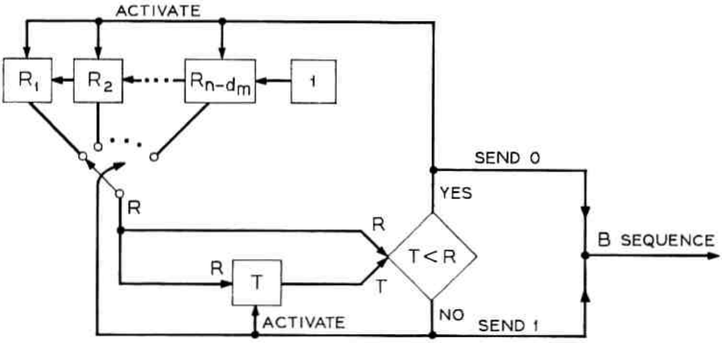

Fig. 6—Encoder for arbitrary n and k.
Preset Values: $T = \alpha$
$R_1 = \alpha_1(n)$
$R_2 = \alpha_2(n - 1)$
$\cdot$
$\cdot$
$\cdot$
$R_{n-d_m} = \alpha_{n-d_m}(d_m + 1)$

switch so $R = R_1$. If $T < R_1$, the $n - d_m$ storage devices are activated and their contents reduced. Also, a 0 is transmitted, that is, $b_n = 0$. However, if $T \geqq R_1$, a 1 is sent, R_1 is subtracted from T, and the switch is shifted so $R = R_2$. The process continues until the entire B sequence is generated. Notice that the B sequence is generated and, thus, transmitted with b_n first and b_1 last.

The decoder is shown in Fig. 7. The storage devices are preset so that $R_i = \alpha_i\ (n - i + 1)$, $1 \leqq i \leqq n - d_m$, and the accumulator is set equal to zero. Position the switch so $R = R_1$. Prior to the decoder, if the weight of the received n-tuple is greater than $n - d_m$ because errors have occurred, the weight is reduced to $n - d_m$ by arbitrarily converting sufficient ones to zeros. Thus, we assume that $w(B') \leqq n - d_m$ and that B' arrives with b_n' first and b_1' last. The decoder first considers b_n'. If $b_n' = 0$, the $n - d_m$ storage devices are activated and their contents reduced. If $b_n' = 1$, the accumulator is increased by R_1 and the switch is shifted so $R = R_2$. The process continues until b_1' has been used. After b_1' has been used, the accumulator contains the integer representation of A'.

For the stuff word detector, S and all other received words less than a specified distance from S are deleted. Thus, the detector counts the zeros in each word and, if the count is less than the specified distance, deletes the word.

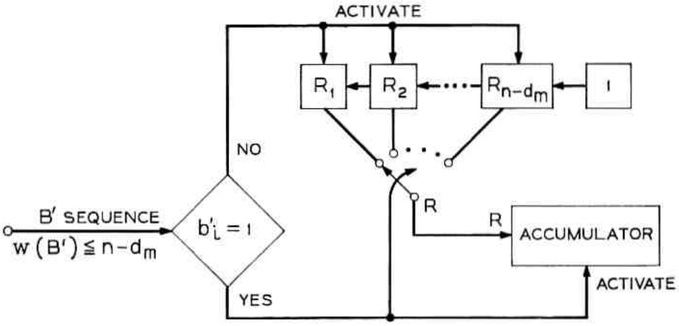

Fig. 7—Decoder for arbitrary n and k.
Preset Values: $R_1 = \alpha_1(n)$
$R_2 = \alpha_2(n - 1)$
$\cdot$
$\cdot$
$\cdot$
$R_{n-d_m} = \alpha_{n-d_m}(d_m + 1)$
Accumulator $= 0$

V. ENCODING, DECODING, AND STUFF WORD DETECTION: $n = k + 1$

In this section, we give a possible implementation for the encoder, decoder, and stuff word detector for $n = k + 1$.

5.1 *Case 1*

Let the stuff word S be the all-zero $(k + 1)$-tuple. The encoder is specified in (6) and shown in Fig. 8. In Fig. 8, the A sequence is assumed to arrive with a_k first and a_1 last. Similarly, b_{k+1} is transmitted first with b_1 last.

$$\begin{aligned} &\text{If} \quad w(A) \geqq \left[\frac{k}{2}\right] + 1, \qquad b_{k+1} = 0. \\ &\text{If} \quad w(A) \leqq \left[\frac{k}{2}\right], \qquad\quad b_{k+1} = 1. \\ &\text{Then} \quad b_i = b_{k+1} \oplus a_i \quad \text{for} \quad 1 \leqq i \leqq k. \end{aligned} \tag{6}$$

Decoding is also straightforward. The operation necessary for decoding is

$$a'_i = b'_{k+1} \oplus b'_i \quad \text{for} \quad 1 \leqq i \leqq k. \tag{7}$$

Figure 9 shows the decoder in equation (7). In Fig. 9, it is assumed that b'_{k+1} is received first and b'_1 is received last. The stuff word detector is the same as in Section IV except ones are counted instead of zeros.

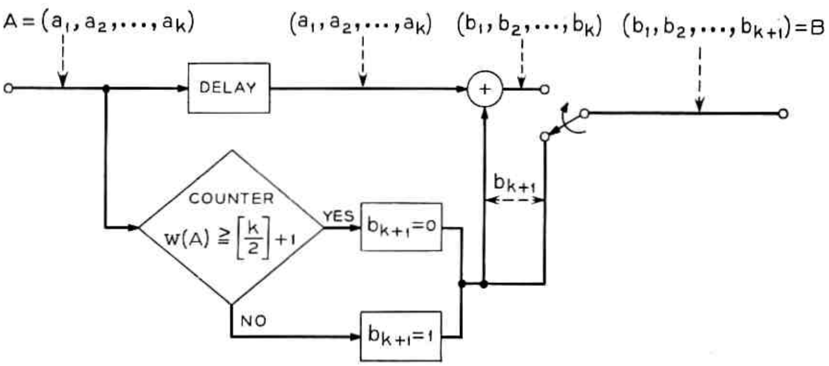

Fig. 8 — Encoder for Case 1.

The encoding-decoding technique in (6) results in some error multiplication. Suppose that position i, $1 \leqq i \leqq k$, is in error and that position $k + 1$ is correct. Then, from equations (6) and (7),

$$a'_i = b_{k+1} \oplus 1 \oplus b_i = 1 \oplus a_i \ .$$

The error in position i is delivered to the customer. However, if position $k + 1$ is in error, from equations (6) and (7),

$$a'_i = 1 \oplus b_{k+1} \oplus b_i = 1 \oplus a_i \quad \text{for} \quad 1 \leqq i \leqq k.$$

The point is that now all k data positions are in error. However, character synchronization is maintained because the correct number of bits are delivered to the destination.

As noted in Section III, it is possible to design a code for which the stuff word is the all-zero word and then, for transmission, suppress the all-zero word by adding an unused word to each word before trans-

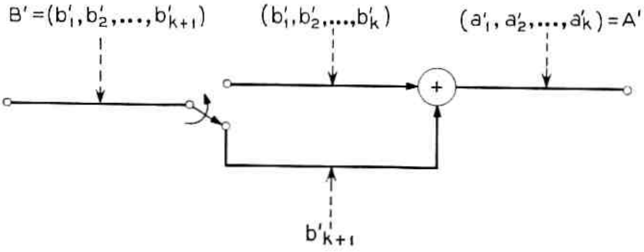

Fig. 9 — Decoder for Case 1.

mission. However, it is easy to combine the encoding and addition operations. We give two examples.

5.2 *Case 2*

Let the stuff word be

$$S = (\underset{\substack{\uparrow \\ \text{position } k+1}}{1}\ \underbrace{0 \cdots 0}_{k \text{ positions}}).$$

The encoder is given in (8).

$$\begin{aligned} &\text{If} \quad w(A) \geqq \left[\frac{k}{2}\right] + 1, \qquad b_{k+1} = 1. \\ &\text{If} \quad w(A) \leqq \left[\frac{k}{2}\right], \qquad\quad\ b_{k+1} = 0. \end{aligned} \tag{8}$$

$$\text{Then } b_i = 1 \oplus b_{k+1} \oplus a_i \quad \text{for} \quad 1 \leqq i \leqq k.$$

The decoder performs the operations in (9).

$$a_i' = 1 \oplus b_{k+1}' \oplus b_i' \quad \text{for} \quad 1 \leqq i \leqq k. \tag{9}$$

5.3 *Case 3*

The stuff word is

$$S = (\underset{\substack{\uparrow \\ \text{position } k+1}}{0}\ 0 \cdots 0\ \underbrace{1 \cdots 1}_{k_1 \text{ positions}}).$$

The encoder is specified in (10).

$$\begin{aligned} &\text{If} \quad w(A) \geqq \left[\frac{k}{2}\right] + 1, \qquad b_{k+1} = 0. \\ &\text{If} \quad w(A) \leqq \left[\frac{k}{2}\right], \qquad\quad\ b_{k+1} = 1. \end{aligned} \tag{10}$$

$$\begin{aligned} \text{Then } b_i &= 1 \oplus b_{k+1} \oplus a_i \quad \text{for} \qquad\quad 1 \leqq i \leqq k_1\,, \\ b_i &= b_{k+1} \oplus a_i \qquad\quad\ \text{for} \quad k_1 + 1 \leqq i \leqq k. \end{aligned}$$

The decoder performs the operations in (11).

$$a'_i = 1 \oplus b'_{k+1} \oplus b'_i \quad \text{for} \quad 1 \leqq i \leqq k_1 ,$$
$$a'_i = b'_{k+1} \oplus b'_i \quad \text{for} \quad k_1 + 1 \leqq i \leqq k. \tag{11}$$

Notice that if $k_1 = k$, the decoder for Case 3 is identical to the decoder for Case 2.

VI. THE (8, 7) CODE

Because Bell System PCM channels use a basic 8-bit word, the (8, 7) code is of interest. From Fig. 3, it is possible to design an (8, 7) code with $d_m = 4$. Let R_s and R_d denote the stuff rate and the rate of occurrence of the data words that are distance four from S, respectively. When the stuff rate is low ($R_s < R_d$), the decision rule at the receiver is biased in favor of the data words by deleting as stuff words all received words distance one from S and decoding as data words all remaining received words. Conversely, for high stuff rates ($R_s > R_d$), the decision rule is biased in favor of the stuff words by deleting as stuff words all received words distance two or less from S and decoding as data words all remaining received words. When the rates are approximately equal ($R_s \simeq R_d$), the two decision rules give comparable performance.

It is possible to modify the (8, 7) code by relaxing the minimum distance requirement so that all data words are merely required to be at least distance three from S. Received words distance one from S are deleted as stuff words; the remaining received words are decoded as data words. This balanced code gives roughly the same performance as either of the biased codes when $R_s \simeq R_d$. However, the reduction in minimum distance provides for less error multiplication (see Section V).

Let $\bar{T}_{d|s}$ denote the mean time between erroneous conversions of a stuff word into a data word and let $\bar{T}_{s|d}$ denote the mean time between erroneous conversions of a data word into a stuff word. We use the following assumptions:

(*i*) Transmission errors are independent of the transmitted bits, independent of each other, and occur with probability $p = 10^{-7}$.

(*ii*) The transmission rate is 64×10^3 bits per second or, because $n = 8$, 8000 words per second.

(*iii*) All 7-bit input words are equally likely.

(*iv*) The stuff rate is R_s stuff words per second.*

* The value of R_s will vary depending upon the application. For fine adjustment of clock rates, R_s typically would be less than 20 words per second. If word stuffing is used for speed padding as well as adjusting clock rates (for example, to send 50 kilobit service over a 64-kilobit line), R_s could be 850 words per second or larger.

The balanced (8, 7) code is applicable except for extremely high or low stuff rates. Let N_3 denote the number of data words distance three from S (the exact value of N_3 depends upon the encoder). Then

$$\bar{T}_{d|s} \cong \frac{1}{3600R_s\left[\binom{8}{2}p^2\right]} = \frac{9.93 \times 10^8}{R_s} \text{ hours}$$

and

$$\bar{T}_{s|d} \cong \frac{1}{3600(8000 - R_s)\left\{\frac{\binom{3}{2}N_3}{2^7}p^2\right\}} = \frac{1.18 \times 10^{12}}{N_3(8000 - R_s)} \text{ hours.}$$

By a modification of the encoder-decoder in Case 1 of Section V, it is possible to reduce the error multiplication in the decoder. The stuff word is

$$S = (00000000).$$

The encoder is given in (12).

$$\begin{aligned} &\text{If} \quad w(A) \geqq 3, \qquad b_8 = 0. \\ &\text{If} \quad w(A) \leqq 2, \qquad b_8 = 1. \\ \text{Then } b_i &= b_8 \oplus a_i \quad \text{for} \quad 1 \leqq i \leqq 4, \\ b_i &= a_i \qquad\quad \text{for} \quad 5 \leqq i \leqq 7. \end{aligned} \tag{12}$$

The decoder performs the operations in (13).

$$\begin{aligned} a_i' &= b_8' \oplus b_i' \quad \text{for} \quad 1 \leqq i \leqq 4, \\ a_i' &= b_i' \qquad\quad \text{for} \quad 5 \leqq i \leqq 7. \end{aligned} \tag{13}$$

Notice that an error in position eight now results in four rather than seven errors for the customer.

VII. COMPARISON WITH GROUP CODES

In the binary group codes normally encountered in error-control work, the code words are a set of 2^k n-tuples selected so that the code words form a group under component-by-component modulo two addition. Because of the resulting perfect distance symmetries between the code words,* the probability that a transmitted word is decoded in

* Let X and Y be code words. For any $\epsilon(1 \leqq \epsilon \leqq n)$, the number of code words distance ϵ from X is equal to the number of code words distance ϵ from Y for all X and Y.

error does not depend upon the transmitted word (provided the transmission errors are independent of the transmitted bits).

The code presented herein has $2^k + 1$ code words (the 2^k elements of D plus the stuff word S). The code words do not form a group and exhibit highly asymmetric distance properties. It is the asymmetric distance properties that enable us to use the available redundancy to protect against the loss of character synchronization due to transmission errors.

We note that it is possible to design other asymmetric codes that are, in a sense, generalizations of the code in (2). Instead of a single stuff word, there are now several special words with unique distance properties with respect to each other and the set D. Such codes might be used in a data transmission system where, for example, one wishes to provide more protection for control characters than for data words. A wide range of capabilities is possible and future work in the design of these codes should prove profitable.

VIII. CONCLUSIONS

For a given redundancy, we give the code that maximizes the minimum distance between the stuff word and any data word. An encoder and decoder are given for arbitrary n and k. For $n = k + 1$, a particularly simple encoding-decoding technique is described. Certain properties of the (8, 7) code are considered in detail.

IX. ACKNOWLEDGMENT

The author wishes to thank E. J. Hronik for pointing out the importance of this problem and for a number of useful discussions during the course of the work.

APPENDIX A

Derivation of d_m for $n = k + 1$

Let $n = k + 1$. From (2b), d_m is chosen so that

$$\sum_{l=1}^{d_m-1} \binom{k+1}{l} \leq 2^k - 1 < \sum_{l=1}^{d_m} \binom{k+1}{l}. \tag{14}$$

However,

$$2^{k+1} - 2 = \sum_{l=1}^{k} \binom{k+1}{l}.$$

Therefore, for k even,

$$2^k - 1 = \sum_{l=1}^{k/2} \binom{k+1}{l}$$

which, from (14), implies that

$$d_m = \frac{k}{2} + 1$$

and, from (2c), that $\delta = 0$. For k odd,

$$2^k - 1 = \sum_{l=1}^{[k/2]} \binom{k+1}{l} + \frac{1}{2} \begin{pmatrix} k+1 \\ \left[\frac{k}{2}\right] + 1 \end{pmatrix} \tag{15}$$

where $[k/2]$ denotes the largest integer $\leqq k/2$. Thus, from (14) and (15),

$$d_m = \left[\frac{k}{2}\right] + 1.$$

APPENDIX B

Encoder for (9, 7) *Code*

For the (9, 7) code, $d_m = 6$. Thus, construct the three arrays shown in Table II. Consider encoding $\alpha_0 = 22$. In the first array, $\alpha_1(5) = 15$ is the largest $\alpha_1(j)$ not greater than 22. Therefore, $j_1 = 5$. Next, in the second array, we find that the largest $\alpha_2(j)$ not greater than

$$\alpha_0 - \alpha_1(5) = 22 - 15 = 7$$

is $\alpha_2(4) = 7$. Thus, $j_2 = 7$. However,

$$\sum_{m=1}^{2} \alpha_m(j_m) = \alpha_1(5) + \alpha_2(4) = 22 = \alpha_0$$

Table II—Arrays for Encoding for the (9, 7) Code.

Array 1		Array 2		Array 3	
j	$\alpha_1(j)$	j	$\alpha_2(j)$	j	$\alpha_3(j)$
1	1	1	1	1	1
2	2	2	2	2	2
3	4	3	4	3	3
4	8	4	7	4	4
5	15	5	11	5	5
6	26	6	16	6	6
7	42	7	22	7	7
8	64	8	29		
9	93				

so the process terminates. Therefore, B_0 has ones in positions $j_1 = 5$ and $j_2 = 4$, that is,

$$B_0 = (000011000).$$

position 5 —↑↑— position 4

APPENDIX C

Proof of Equation (5)

After substituting (4) into (5), we must show that

$$\sum_{m=0}^{\mathrm{Min}(n-d_m-i+1,\,j-1)} \binom{j-1}{m} + \sum_{m=0}^{\mathrm{Min}(n-d_m-i,\,j-1)} \binom{j-1}{m} = \sum_{m=0}^{\mathrm{Min}(n-d_m-i+1,\,j)} \binom{j}{m} \qquad (16)$$

for $1 \leqq j \leqq n - i$, $1 \leqq i \leqq n - d_m - 1$. Choose an i and consider j as j increases from 1 to $n - i$. Suppose that $1 \leqq j \leqq n - d_m - i + 1$. Then (16) reduces to

$$\sum_{m=0}^{j-1} \binom{j-1}{m} + \sum_{m=0}^{j-1} \binom{j-1}{m} = \sum_{m=0}^{j} \binom{j}{m}$$

or

$$2^{j-1} + 2^{j-1} = 2^j.$$

Now, suppose that $n - d_m - i + 1 < j \leqq n - i$. Then (16) becomes

$$\sum_{m=0}^{n-d_m-i+1} \binom{j-1}{m} + \sum_{m=0}^{n-d_m-i} \binom{j-1}{m} = \sum_{m=0}^{n-d_m-i+1} \binom{j}{m}.$$

However,

$$\begin{aligned} 1 + \sum_{m=1}^{n-d_m-i+1} \binom{j-1}{m} + \sum_{m=0}^{n-d_m-i} \binom{j-1}{m} &= 1 + \sum_{m=0}^{n-d_m-i} \left[\binom{j-1}{m+1} + \binom{j-1}{m}\right] \\ &= 1 + \sum_{m=0}^{n-d_m-i} \binom{j}{m+1} \\ &= \sum_{m=0}^{n-d_m-i+1} \binom{j}{m}. \end{aligned}$$

The argument is valid for each i, $1 \leqq i \leqq n - d_m - 1$.

REFERENCES

1. Mattesich, R. R., and Richters, J. S., unpublished work, Bell Telephone Laboratories, December 13, 1968.
2. Butman, S., "Synchronization of PCM Channels by the Method of Word Stuffing," IEEE Trans. Comm. Tech., *COM-16,* No. 2 (April 1968), pp. 252–254.
3. Johannes, V. I., and McCullough, R. H., "Multiplexing of Asynchronous Digital Signals Using Pulse Stuffing with Added-Bit Signaling," IEEE Trans. Comm. Tech., *COM-14,* No. 5 (October 1966), pp. 562–568.
4. Mayo, J. S., "An Approach to Digital System Networks," IEEE Trans. Comm. Tech., *COM-15,* No. 2 (April 1967), pp. 307–310.
5. Mayo, J. S., "Experimental 224 Mb/s PCM Terminals," B.S.T.J., *44,* No. 9 (November 1965), pp. 1813–1841.
6. Witt, F. J., "An Experimental 224 Mb/s Digital Multiplexer-Demultiplexer Using Pulse Stuffing Synchronization," B.S.T.J., *44,* No. 9 (November, 1965), pp. 1843–1885.

Queues Served in Cyclic Order: Waiting Times

By R. B. COOPER

(Manuscript received December 2, 1969)

This paper extends the results of a previous paper in which two models of a system of queues served in cyclic order were studied. One model is an exhaustive service model, in which the server waits on all customers in a queue before proceeding to the next queue in cyclic order. The other is a gating model, in which a gate closes behind the waiting units when the server arrives, and the server waits on only those customers in front of the gate, deferring service of later arrivals until the next cycle.

In the present paper, the Laplace–Stieltjes transforms of the order-of-arrival waiting time distribution functions and, for the exhaustive service model, the mean waiting time for a unit arriving at a queue, are obtained.

I. INTRODUCTION

In a recent paper[1] we studied two models of a system of queues served in cyclic order:

In each model, the ith queue is characterized by general service time distribution function $H_i(\cdot)$ and Poisson input with parameter λ_i . In the exhaustive service model, the server continues to serve a particular queue until for the first time there are no units in service or waiting in that queue; at this time the server advances to and immediately starts service on the next nonempty queue in the cyclic order. The gating model differs from the exhaustive service model in that when the server advances to a nonempty queue, a gate closes behind the waiting units. Only those units waiting in front of the gate are served during this cycle, with the service of subsequent arrivals deferred to the next cycle.

In Ref. 1 we found, for the exhaustive service model, expressions for the mean number of units in a queue at the instant it starts service, the mean cycle time, and the Laplace–Stieltjes transform of the cycle time distribution function.

In the present paper, we extend the analysis to obtain, for each model, the Laplace–Stieltjes transform of the order-of-arrival waiting time distribution function and, for the exhaustive service model, the mean waiting time for a unit arriving at the ith queue.

In Ref. 1 we defined a switch point as a time epoch at which the server finishes serving a queue; and we defined $P_i(n_1, \cdots, n_N)$ as the joint probability that at a switch point the server has just completed a visit at queue i ($i = 0, 1, \cdots, N$) and n_1 units are waiting in queue $i + 1$, n_2 units in queue $i + 2$, $\cdots$, and n_N units in queue $i + N$.

The central results of Ref. 1 were an iterative algorithm for the calculation of the probability generating functions

$$g_i(x_1, \cdots, x_N) = \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} P_i(n_1, \cdots, n_N) x_1^{n_1} \cdots x_N^{n_N} \qquad (i = 0, 1, \cdots, N) \qquad (1)$$

and, for the exhaustive service model, an expression for the mean number $\bar{n}_i$ of units waiting in queue $i + 1$ when the server completes a visit at queue i. In particular, it was shown for the exhaustive service model that these generating functions satisfy the functional equations

$$\begin{aligned} g_i(x_1, \cdots, x_N) = {} & g_{i-1}\left(\beta_i\left(\sum_{m=1}^{N} \lambda_{i+m}(1 - x_m)\right), x_1, \cdots, x_{N-1}\right) \\ & + \frac{\lambda_i}{\lambda} \beta_i\left(\sum_{m=1}^{N} \lambda_{i+m}(1 - x_m)\right) P(0) - P_{i-1}(0, \cdots, 0) \\ & \qquad (i = 0, 1, \cdots, N) \qquad (2) \end{aligned}$$

and that $\bar{n}_i$ is given by $\bar{n}_i = \bar{m}_i / g_i(1, \cdots, 1)$, where

$$\bar{m}_i = \frac{\lambda_{i+1}}{\lambda} P(0) \frac{\rho - \rho_{i+1}}{1 - \rho} \qquad (i = 0, 1, \cdots, N) \qquad (3)$$

and where λ_i is the rate of arrivals of units at queue i, ρ_i is the traffic intensity at queue i,

$$\lambda = \sum_{i=0}^{N} \lambda_i,$$

$$\rho = \sum_{i=0}^{N} \rho_i,$$

$$P(0) = \sum_{i=0}^{N} P_i(0, \cdots, 0),$$

and $\beta_i(\cdot)$ is the Laplace–Stieltjes transform of the distribution function of the length of the busy period at queue i. Equations (1), (2), and (3) appear in Ref. 1 as equations (3), (5), and (34), respectively.

The distribution generated by $g_i(x_1, \cdots, x_N)$ is defined with respect to a Markov chain imbedded at the switch points. An analysis by Takács[2] of the exhaustive service model for the special case of two queues is based on a Markov chain imbedded at the set of service completion points. Clearly, the set of switch points is a proper subset of the set of service completion points. Our use of switch points instead of service completion points enabled us in Ref. 1 to analyze the multiqueue model with about the same degree of mathematical complexity as Takács required for the analysis of the 2-queue model. On the other hand, one would expect that our use of a chain imbedded in a "smaller" set of points would result in a corresponding loss of useful information.

Takács' analysis yielded waiting time results for the 2-queue model. At the time Ref. 1 was written, it was not apparent to us that our method of analysis provided enough information to enable us to obtain corresponding waiting time results for the multiqueue model. Accordingly, we concentrated on the cycle time, a quantity that seemingly gives the same kind of information as the waiting time. Unfortunately, we did not have complete freedom in choosing a precise definition of the quantity we would call cycle time. The mathematical formulation of the model dictated that the cycle time for queue i be defined, roughly speaking, as the length of time between two successive instants at which the server completes service at queue i, without regard to whether or not the server is continuously busy throughout this time interval. This definition introduces, among others, the following difficulty in the interpretation of realized values of the cycle time: A long cycle time could have resulted either from heavy congestion or from no congestion.

No such ambiguities exist with respect to interpretation of the waiting time, which is simply the elapsed time from the arrival instant of a unit to the instant at which service on this unit begins. Therefore, we would like to obtain waiting time results. Furthermore, we would like to obtain these results, if possible, without directly extending the previous analysis to include the entire set of service completion points.

In the present paper, we obtain the desired waiting time results without recourse to a complicated reformulation of the original analysis based on the complete set of service completion points. Rather, to obtain the waiting times at queue i, we use the generating function $g_{i-1}(x_1, \cdots, x_N)$, calculated in Ref. 1, to append to the original set of switch points only those service completion points that correspond to departures from queue i; and this is sufficient for our waiting time calculations.

The essence of the method is to calculate the probability generating function of the number of units left in queue i by an arbitrary departure from queue i, using only the (known) probability generating function of the number of units waiting in queue i when the server arrives. The Laplace–Stieltjes transform of the order-of-arrival waiting time distribution function for units at queue i is then easily obtained by a standard argument.

The preceding discussion refers mainly to the exhaustive service model, which was discussed in detail in Ref. 1. The gating model was shown to be characterized by equations that are essentially the same as those of the exhaustive service model, and was therefore not developed in detail. In the present paper, the waiting times for the gating model will also be discussed.

II. THE M/G/1 QUEUE WITH SERVER VACATION TIMES

In preparation for calculation of the waiting times in the exhaustive service model, we first consider the following generalization of the $M/G/1$ queue:

As usual, the server serves the queue continuously as long as there is at least one unit in the system (waiting or in service). When the server finishes serving a unit and finds the system empty, however, it goes away for a length of time called a vacation. At the end of the vacation the server returns to the queue, and begins to serve those units, if any, that have arrived during the vacation. If the server finds the system empty at the end of a vacation, it immediately takes another vacation, and continues in this manner until it finds at least one waiting unit upon return from a vacation.

Let X_k ($k = 1, 2, \cdots$) be the number of units left behind by the kth departing unit. Then

$$P\{X_{k+1} = n\} = \sum_{\nu=0}^{n+1} P\{X_k = \nu\} P\{X_{k+1} = n \mid X_k = \nu\}$$
$$(k = 1, 2, \cdots ; n = 0, 1, \cdots). \qquad (4)$$

Let $P(j)$ be the probability that at the end of a vacation the server finds $j \geqq 0$ units waiting for service. If the arrival rate and the service time distribution function are denoted by λ and $H(\cdot)$, respectively, then

$$P\{X_{k+1} = n \mid X_k = \nu > 0\} = \int_0^\infty \frac{(\lambda\xi)^{n+1-\nu}}{(n+1-\nu)!} \exp(-\lambda\xi)\, dH(\xi)$$
$$(n \geqq \nu - 1) \qquad (5)$$

and

$$P\{X_{k+1} = n \mid X_k = 0\} = \sum_{j=1}^{n+1} \frac{P(j)}{1 - P(0)} \int_0^\infty \frac{(\lambda\xi)^{n+1-j}}{(n + 1 - j)!} \exp(-\lambda\xi) \, dH(\xi). \tag{6}$$

The expression $P(j)/[1 - P(0)]$ in equation (6) is the conditional probability that when the server starts serving the queue there are j units waiting, given that at least one unit is waiting.

When the traffic intensity ρ is less than unity ($\rho = \lambda h$, where h is the mean service time), there exists a unique distribution

$$\pi_n = \lim_{k\to\infty} P\{X_k = n\} \qquad (n = 0, 1, \cdots) \tag{7}$$

that satisfies both the normalization equation

$$\sum_{n=0}^{\infty} \pi_n = 1 \tag{8}$$

and the limiting set of equations obtained from equation (4),

$$\pi_n = \pi_0 \sum_{j=1}^{n+1} \frac{P(j)}{1 - P(0)} \int_0^\infty \frac{(\lambda\xi)^{n+1-j}}{(n + 1 - j)!} \exp(-\lambda\xi) \, dH(\xi)$$
$$+ \sum_{\nu=1}^{n+1} \pi_\nu \int_0^\infty \frac{(\lambda\xi)^{n+1-\nu}}{(n + 1 - \nu)!} \exp(-\lambda\xi) \, dH(\xi) \qquad (n = 0, 1, \cdots). \tag{9}$$

Define the probability generating functions

$$f(x) = \sum_{n=0}^{\infty} \pi_n x^n \tag{10}$$

and

$$\psi(x) = \sum_{j=0}^{\infty} P(j) x^j. \tag{11}$$

Substitution of equation (9) into equation (10) yields, after some manipulation,

$$f(x) = \frac{\left\{\dfrac{\psi(x) - P(0)}{1 - P(0)} - 1\right\}\eta(\lambda - \lambda x)}{x - \eta(\lambda - \lambda x)} \, \pi_0 \tag{12}$$

where $\eta(\cdot)$ is the Laplace–Stieltjes transform of the service time distribution function $H(\cdot)$. Observe that the expression $[\psi(x) - P(0)]/[1 - P(0)]$ is the probability generating function of the number of units waiting when service commences.

The unknown probability π_0 is determined from equation (12) by the normalization condition $f(1) = 1$. Application of l'Hospital's rule to equation (12) yields

$$\pi_0 = \frac{1 - P(0)}{\psi'(1)} (1 - \rho). \tag{13}$$

Thus, the probability generating function $f(\cdot)$ of the number of units left behind by an arbitrary departure, and the probability generating function $\psi(\cdot)$ of the number of units waiting at the end of a vacation are related as follows:

$$f(x) = \frac{[\psi(x) - 1]\eta(\lambda - \lambda x)}{x - \eta(\lambda - \lambda x)} \frac{(1 - \rho)}{\psi'(1)}. \tag{14}$$

We now apply a standard argument to obtain the Laplace–Stieltjes transform $\omega(\cdot)$ of the order-of-arrival waiting time distribution function from the generating function (14).

Let $F(\cdot)$ be the distribution function of an arbitrary unit's sojourn time, defined as the elapsed time between the unit's arrival and departure epochs, and denote by $\phi(\cdot)$ the Laplace–Stieltjes transform of $F(\cdot)$. Since the sojourn time is the sum of the waiting time and the service time, and since these latter times are independent, therefore

$$\phi(s) = \omega(s)\eta(s). \tag{15}$$

When units are served in their arrival order, each departing unit leaves behind it precisely those units that arrived during the departure's sojourn time. Further, these remaining units arrived according to a Poisson process, independent of the sojourn time. Therefore, the probability π_n that a departing unit leaves behind n other units is

$$\pi_n = \int_0^\infty \frac{(\lambda\xi)^n}{n!} \exp(-\lambda\xi)\, dF(\xi). \tag{16}$$

Substitution of equation (16) into (10) gives the well known and fundamental relation

$$f(x) = \phi(\lambda - \lambda x). \tag{17}$$

Equations (14), (15), and (17) together give the Laplace–Stieltjes transform $\omega(\cdot)$ of the waiting time distribution function in terms of the probability generating function $\psi(\cdot)$ of the number of units waiting at the end of a vacation:

$$\omega(s) = \frac{\lambda}{\psi'(1)} \left[1 - \psi\left(\frac{\lambda - s}{\lambda}\right)\right] \frac{1 - \rho}{s - \lambda + \lambda\eta(s)}. \tag{18}$$

Note that for the ordinary $M/G/1$ queue, in which the vacation ends immediately whenever a unit arrives and finds the server idle, $\psi(x) = x$ and equation (18) reduces to the well known Pollaczek–Khinchin formula,

$$\omega(s) = \frac{s(1 - \rho)}{s - \lambda + \lambda\eta(s)}. \tag{19}$$

Finally, the mean wait for service $\bar{W} = -\omega'(0)$, obtained from equation (18), is given by

$$\bar{W} = \frac{\lambda\eta''(0)}{2(1 - \rho)} + \frac{\psi''(1)}{2\lambda\psi'(1)}. \tag{20}$$

The first term on the right side of equation (20) is identical with the mean waiting time in the ordinary $M/G/1$ queue, as would be obtained directly from the Pollaczek–Khinchin formula (19). The second term in equation (20) represents the component of the mean wait that arises because of the variability in the number of units waiting when service begins. Although service in arrival order was assumed in its derivation, equation (20) is valid for any order of service that is independent of the service times.

Note that the result (14) is true regardless of any relationship between the vacation lengths and the arrival process, whereas equations (18) and (20) are valid only when the vacation lengths are determined without regard to the arrival process. For example, if one were to consider a mechanism such that service begins as soon as a fixed number j ($j \geqq 2$) units are waiting, then equation (14) with $\psi(x) = x^j$ would correctly give the probability generating function of the number of units in the system just after a service completion epoch. On the other hand, equations (18) and (20) would not apply, because (16), and therefore (17), would no longer be true. (For if $\psi(x) = x^j$, then the first departing unit would always leave behind at least those $j - 1$ other units that were present when service commenced. Thus, for the first departing unit, $\pi_n = 0$ for $n < j - 1$, and this contradicts the assumption (16) if $j \geqq 2$.)

III. LAPLACE–STIELTJES TRANSFORM OF WAITING TIME DISTRIBUTION FUNCTION FOR EXHAUSTIVE SERVICE MODEL

We now proceed to apply the results of Section II to the analysis of waiting times in the exhaustive service model. In essence, the "vacation time" of Section II is the length of time that the server spends idle or working on other queues before registering a switch point at queue $i-1$ and beginning service on queue i.

For want of a better word, let us define as a supercycle the elapsed time between the arrival epoch of a unit at any queue when the system is completely empty, and the first instant at which the whole system again becomes empty. Then units that arrive at queue i can be classified into two exclusive and exhaustive categories:

(1) arrivals at queue i that either initiate a supercycle or occur during the 1-busy period generated by an arrival at queue i that initiated a supercycle; or

(2) all other arrivals at queue i.

Equivalently, units in category (1) are those arrivals at queue i whose service begins prior to the occurrence of the first switch point of a supercycle, whereas units in category (2) are arrivals at queue i whose service times begin after the first switch point of a supercycle.

Consider now the waiting times of units that arrive at queue i. Those units in category (1) are served during a busy period originated by one unit. Therefore, the Laplace–Stieltjes transform $\omega_i^{(1)}(\cdot)$ of the order-of-arrival waiting time distribution function for units at queue i that belong to category (1) is given by the Pollaczek–Khinchin formula (19):

$$\omega_i^{(1)}(s) = \frac{s(1-\rho_i)}{s-\lambda_i+\lambda_i\eta_i(s)} \tag{21}$$

where ρ_i, λ_i, and $\eta_i(\cdot)$ are the corresponding quantities in equation (19) defined now with respect to queue i.

Units in category (2) are served during a busy period originated by those units waiting in queue i when the server leaves queue $i-1$. Let $\psi_i(\cdot)$ be the probability generating function of the number of units waiting in queue i when the server leaves queue $i-1$; $\psi_i(\cdot)$ is the probability generating function of the number of units waiting for service in queue i when the server finishes a vacation, and is given by

$$\psi_i(x) = \frac{g_{i-1}(x, 1, \cdots, 1)}{g_{i-1}(1, 1, \cdots, 1)}. \tag{22}$$

[Note that $\psi_i'(1) = \bar{n}_{i-1} = \bar{m}_{i-1}/g_{i-1}(1, 1, \cdots, 1)$.] Thus, the Laplace–Stieltjes transform $\omega_i^{(2)}(\cdot)$ of the waiting time distribution function for units in category (2) is given by equation (18):

$$\omega_i^{(2)}(s) = \frac{\lambda_i}{\psi_i'(1)}\left[1-\psi_i\left(\frac{\lambda_i-s}{\lambda_i}\right)\right]\frac{1-\rho_i}{s-\lambda_i+\lambda_i\eta_i(s)}. \tag{23}$$

Let $p_i^{(1)}$ be the proportion of all arrivals at queue i that are in category (1). The mean number of units that arrive at queue i during an interval of length t is $\lambda_i t$. The probability that an arbitrary arrival at queue i finds the whole system empty is $1 - \rho$ $(\rho = \rho_0 + \cdots + \rho_N)$, so that $\lambda_i t(1 - \rho)$ is the mean number of arrivals at queue i that initiate a supercycle during any elapsed time t. The mean number of units served at queue i during the 1-busy period generated by each such arrival is $(1 - \rho_i)^{-1}$, and hence the mean number of units in category (1) served at queue i during an elapsed time t is $\lambda_i t(1 - \rho)/(1 - \rho_i)$. Therefore, the probability is $[\lambda_i t(1 - \rho)/(1 - \rho_i)]/\lambda_i t$ that an arbitrary arrival at queue i is in category (1); that is,

$$p_i^{(1)} = \frac{1 - \rho}{1 - \rho_i}, \tag{24}$$

and the probability $p_i^{(2)} = 1 - p_i^{(1)}$ that an arbitrary arrival at queue i is in category (2) is

$$p_i^{(2)} = \frac{\rho - \rho_i}{1 - \rho_i}. \tag{25}$$

The Laplace–Stieltjes transform $\omega_i(\cdot)$ of the waiting time distribution function for an arbitrary unit at queue i is the weighted sum of the transforms for each category:

$$\omega_i(s) = p_i^{(1)}\omega_i^{(1)}(s) + p_i^{(2)}\omega_i^{(2)}(s). \tag{26}$$

Finally, equation (26) becomes, with the help of equations (21) through (25) and equation (3),

$$\omega_i(s) = \frac{1 - \rho}{s - \lambda_i + \lambda_i \eta_i(s)} \cdot \left\{ \frac{\lambda}{P(\mathbf{0})} \left[g_{i-1}(1, 1, \cdots, 1) - g_{i-1}\left(\frac{\lambda_i - s}{\lambda_i}, 1, \cdots, 1 \right) \right] + s \right\} \qquad (i = 0, 1, \cdots, N). \tag{27}$$

Inherent in equation (27) is the assumption that units in queue i are served in their arrival order, but no assumption is made regarding the order of service of units in other queues. If at each queue units are served in their arrival order, then the waiting time distribution function for an arbitrary unit, without regard to the identity of the queue in which it is served, has Laplace–Stieltjes transform $\omega(\cdot)$ given by

$$\omega(s) = \lambda^{-1} \sum_{i=0}^{N} \lambda_i \omega_i(s). \tag{28}$$

IV. MEAN WAITING TIMES FOR EXHAUSTIVE SERVICE MODEL

Denote by $\bar{W}_i$ the mean wait for service suffered by units arriving at queue i. The mean wait for service for units in category (1) is $[\lambda_i \eta_i''(0)/2(1 - \rho_i)]$; the mean wait for service for units in category (2) is, in analogy with equation (20), $[\lambda_i \eta_i''(0)/2(1 - \rho_i)] + [\psi_i''(1)/2\lambda_i \psi_i'(1)]$. Weighting these values according to equations (24) and (25), respectively, we have

$$\bar{W}_i = \frac{\lambda_i \eta_i''(0)}{2(1 - \rho_i)} + \frac{\psi_i''(1)}{\psi_i'(1)} \frac{\rho - \rho_i}{2\lambda_i(1 - \rho_i)} \qquad (i = 0, 1, \cdots, N). \tag{29}$$

In equation (26) of Ref. 1 we defined

$$\bar{m}_i(k) = \frac{\partial}{\partial x_k} g_i(x_1, \cdots, x_N) \bigg|_{x_1 = \cdots = x_N = 1} \qquad (i = 0, 1, \cdots, N; \quad k = 1, \cdots, N) \tag{30}$$

and $\bar{m}_i = \bar{m}_i(1)$. Let us also define

$$\bar{m}_i(j, k) = \frac{\lambda(1 - \rho)}{P(\mathbf{0})} \frac{\partial^2}{\partial x_j \, \partial x_k} g_i(x_1, \cdots, x_N) \bigg|_{x_1 = \cdots = x_N = 1} \qquad (i = 0, 1, \cdots, N; \quad j = 1, \cdots, N; \quad k = 1, \cdots, N). \tag{31}$$

Then it follows from equation (22) and these definitions that

$$\frac{\psi_i''(1)}{\psi_i'(1)} = \frac{P(\mathbf{0})}{\lambda(1 - \rho)} \frac{\bar{m}_{i-1}(1, 1)}{\bar{m}_{i-1}(1)}. \tag{32}$$

Using equations (3) and (32), we can rewrite (29):

$$\bar{W}_i = \frac{\lambda_i \eta_i''(0)}{2(1 - \rho_i)} + \frac{\bar{m}_{i-1}(1, 1)}{2\lambda_i^2(1 - \rho_i)} \qquad (i = 0, 1, \cdots, N). \tag{33}$$

It remains to calculate the quantity $\bar{m}_{i-1}$ (1, 1) in (33). To this end, we define

$$\bar{\beta}_i(k) = \frac{\partial}{\partial x_k} \beta_i\left(\sum_{m=1}^{N} \lambda_{i+m}(1 - x_m)\right) \bigg|_{x_1 = \cdots = x_N = 1} \qquad (i = 0, 1, \cdots, N; \quad k = 1, \cdots, N) \tag{34}$$

and

$$\bar{\beta}_i(j, k) = \frac{\partial^2}{\partial x_j \, \partial x_k} \beta_i\left(\sum_{m=1}^{N} \lambda_{i+m}(1 - x_m)\right) \bigg|_{x_1 = \cdots = x_N = 1} \qquad (i = 0, 1, \cdots, N; \quad j = 1, \cdots, N; \quad k = 1, \cdots, N). \tag{35}$$

Note that in terms of the given parameters,

$$\bar{\beta}_i(k) = \lambda_{i+k} \frac{h_i}{1 - \rho_i} \tag{36}$$

and

$$\bar{\beta}_i(j, k) = \lambda_{i+j}\lambda_{i+k} \frac{\eta_i''(0)}{(1 - \rho_i)^3} \tag{37}$$

where h_i is the mean and $\eta_i(\cdot)$ the Laplace–Stieltjes transform of the service time distribution function for a unit at queue i.

We proceed to calculate $\bar{m}_i(1, 1)$ in the same way we calculated $\bar{m}_i(1)$ in Section VII of Ref. 1. Differentiating twice through equation (2) and setting $x_1 = \cdots = x_N = 1$, we obtain the three-dimensional set of linear equations

$$\begin{aligned} \bar{m}_i(j, k) = {} & \frac{\lambda(1 - \rho)}{P(0)} \bar{m}_{i-1}(1)\bar{\beta}_i(j, k) + (1 - \rho)\lambda_i\bar{\beta}_i(j, k) \\ & + \bar{\beta}_i(j)\bar{\beta}_i(k)\bar{m}_{i-1}(1, 1) + (1 - \delta(N - j))\bar{\beta}_i(k)\bar{m}_{i-1}(1, j + 1) \\ & + (1 - \delta(N - k))\bar{\beta}_i(j)\bar{m}_{i-1}(1, k + 1) \\ & + (1 - \delta(N - j))(1 - \delta(N - k))\bar{m}_{i-1}(j + 1, k + 1) \\ & (i = 0, 1, \cdots, N; \quad j = 1, \cdots, N; \quad k = 1, \cdots, N) \end{aligned} \tag{38}$$

where $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ if $x \neq 0$. Using equation (3) and combining the first two terms on the right side of equation (38), we can write

$$\begin{aligned} \bar{m}_i(j, k) = {} & \lambda_i(1 - \rho_i)\bar{\beta}_i(j, k) + \bar{\beta}_i(j)\bar{\beta}_i(k)\bar{m}_{i-1}(1, 1) \\ & + \bar{\beta}_i(k)\bar{m}_{i-1}(1, j + 1) + \bar{\beta}_i(j)\bar{m}_{i-1}(1, k + 1) + \bar{m}_{i-1}(j + 1, k + 1) \\ & (i = 0, 1, \cdots, N; \quad j = 1, \cdots, N; \quad k = 1, \cdots, N) \end{aligned} \tag{39}$$

where all undefined terms are taken to be zero. (The functions $\bar{m}_i(j, k)$ are defined only for $j, k = 1, \cdots, N$.) It is required to solve this set of $\frac{1}{2}N(N + 1)^2$ independent linear equations for the $\bar{m}_i(1, 1)$. [Note that $\bar{m}_i(j, k) = \bar{m}_i(k, j)$.]

Successive substitution into the last term on the right side of equation (39) gives

$$\begin{aligned} \bar{m}_i(1, k) = {} & \sum_{\nu=0} \lambda_{i-\nu}(1 - \rho_{i-\nu})\bar{\beta}_{i-\nu}(1 + \nu, k + \nu) \\ & + \sum_{\nu=0} \bar{\beta}_{i-\nu}(1 + \nu)\bar{\beta}_{i-\nu}(k + \nu)\bar{m}_{i-1-\nu}(1, 1) \end{aligned}$$

$$+ \sum_{\nu=0} \bar{\beta}_{i-\nu}(k + \nu)\bar{m}_{i-1-\nu}(1, 2 + \nu)$$

$$+ \sum_{\nu=0} \bar{\beta}_{i-\nu}(1 + \nu)\bar{m}_{i-1-\nu}(1, k + 1 + \nu)$$

$$(i = 0, 1, \cdots, N; \quad k = 1, \cdots, N) \qquad (40)$$

where each sum is continued as long as the terms are defined.

The set (40) consists of $N(N + 1)$ independent linear equations. Unfortunately, it does not appear that further algebraic simplification is likely. However, for particular values of the parameters and reasonable values of N, numerical solution should not be difficult.

Therefore, to calculate the mean wait $\bar{W}_i$ for service at queue i ($i = 0, 1, \cdots, N$) for any particular values of the basic parameters, simply solve the set (40) of $N(N + 1)$ linear equations numerically, and use the resulting value of $\bar{m}_{i-1}$ (1, 1) in equation (33). Note that these calculations for the mean waiting times require no iteration, since neither generating functions nor state probabilities appear. This last observation is remarkable in light of the complicated iteration process (discussed in Ref. 1) underlying the derivation of these results. Thus, despite the complicated derivation, calculations do not seem impractical.

In the particular case of two queues ($N = 1$), only $N(N + 1) = 2$ simultaneous equations must be solved to find $\bar{m}_i(1, 1)$, and an algebraic solution is easily obtained. For $N = 1$ equation (40) gives

$$\bar{m}_{i-1}(1, 1) = \lambda_i^2 \frac{\lambda_{i-1}\eta_{i-1}''(0)(1 - \rho_i)^2 + \lambda_i\eta_i''(0)\rho_{i-1}^2}{(1 - \rho_{i-1})^2(1 - \rho_i)^2 - \rho_{i-1}^2\rho_i^2}$$

and hence, for two queues,

$$\bar{W}_i = \frac{\lambda_i\eta_i''(0)}{2(1 - \rho_i)} + \frac{\lambda_{i-1}\eta_{i-1}''(0)(1 - \rho_i)^2 + \lambda_i\eta_i''(0)\rho_{i-1}^2}{2(1 - \rho_i)[(1 - \rho_{i-1})^2(1 - \rho_i)^2 - \rho_{i-1}^2\rho_i^2]}$$

$$(i = 0, 1). \qquad (41)$$

Our result (41) is in agreement with previous results of Takács,[2] Avi-Itzhak, Maxwell and Miller,[3] and Eisenberg.[4]

Although service in order of arrival was assumed throughout, the results for the mean waiting time are valid for any order of service that is independent of the service times.

V. LAPLACE–STIELTJES TRANSFORM OF WAITING TIME DISTRIBUTION FUNCTION FOR GATING MODEL

Turning to the gating model, we now study briefly the distribution of waiting times for units served in order of arrival at the ith queue.

As with the exhaustive service model, we first calculate the probability generating function of the number of units left in queue i by an arbitrary unit departing from queue i; and using the same arguments, we obtain from this generating function the Laplace–Stieltjes transform of the waiting time distribution function. As in Ref. 1, the notation for the gating model is the same as for the exhaustive service model; $g_{i-1}(x_1, \cdots, x_N)$ and related probabilities are defined and calculated as described in Section IX of Ref. 1.

Let $\pi_i(j)$ be the conditional probability that an arbitrary departure from queue i leaves behind it j units in queue i, given that this departure did not arrive when the system was completely empty. Then

$$\pi_i(j) = \sum_{n=1}^{\infty} \frac{P_{i-1}(n)}{1 - P_{i-1}(0)} \frac{1}{n} \sum_{k=1}^{n} \int_0^{\infty} \frac{(\lambda_i \xi)^{j-n+k}}{(j - n + k)!} \exp(-\lambda_i \xi) \, dH_i^{*k}(\xi) \qquad (j = 0, 1, \cdots) \tag{42}$$

where $P_{i-1}(n)/[1 - P_{i-1}(0)]$ is the conditional probability that $n \geqq 1$ units are waiting in queue i when the gate closes, given that at least one unit is waiting; and $1/n$ is the probability that a departing unit is kth in line for service ($k = 1, 2, \cdots, n$) given that n units are present at the closing of the gate. The integrand in equation (42) is taken to be zero when $n - k > j$.

Following the argument for the exhaustive service model we see that the (conditional) probability generating function of the number of arrivals at queue i that occur during the waiting time of a departing unit (given that the departing unit did not find the system empty on arrival) is

$$\sum_{j=0}^{\infty} \pi_i(j) x^j / \eta_i(\lambda_i - \lambda_i x).$$

A simple calculation gives

$$\frac{\sum_{j=0}^{\infty} \pi_i(j) x^j}{\eta_i(\lambda_i - \lambda_i x)} = \sum_{n=1}^{\infty} \frac{P_{i-1}(n)}{1 - P_{i-1}(0)} \frac{1}{n} \sum_{k=1}^{n} x^{n-k} \eta_i^{k-1}(\lambda_i - \lambda_i x). \tag{43}$$

The probability that an arbitrary departing unit did not find the system completely empty on arrival is $\rho = \sum_{i=0}^{N-1} \lambda_i h_i$. Thus, after summing the geometric series in equation (43), the unconditional order-of-arrival waiting time distribution function for units served at queue i has Laplace–Stieltjes transform $\omega_i(\cdot)$ given by

$$\omega_i(s) = (1 - \rho) + \rho \sum_{n=1}^{\infty} \frac{P_{i-1}(n)}{1 - P_{i-1}(0)} \frac{1}{n\lambda_i^{n-1}} \frac{[\lambda_i \eta_i(s)]^n - (\lambda_i - s)^n}{s - \lambda_i + \lambda_i \eta_i(s)} \qquad (i = 0, 1, \cdots, N - 1). \qquad (44)$$

Equation (44) allows numerical calculation (and hence numerical inversion) of the transform $\omega_i(\cdot)$. Unfortunately, this procedure requires knowledge of the distribution $\{P_{i-1}(n)\}$, which is specified only through its generating function $g_{i-1}(x, 1, \cdots, 1)$. Thus, to obtain numerical results for the gating model one must solve two distinct problems in numerical analysis, numerical calculation of the $\{P_{i-1}(n)\}$ and then numerical inversion of the transform. Note that the first of these numerical calculations is not required for the exhaustive service model. The subject of numerical inversion of Laplace–Stieltjes transforms and probability generating functions (the latter being, in fact, a special case of the former) is important for the reduction to practice of these cyclic queuing models. However, it is a subject best treated separately, without regard to the particular applications at hand, and will not be discussed further here.

VI. SUMMARY AND PROPOSALS FOR FUTURE WORK

We have extended our previous study of cyclic queues to obtain waiting time results. In particular we have obtained, for both the exhaustive service model and the gating model, the Laplace–Stieltjes transform of the waiting time distribution function for units arriving at the ith queue, when units at that queue are served in order of arrival. These transforms are given by equation (27) for the exhaustive service model and equation (44) for the gating model. Also, we have obtained for the exhaustive service model a formula (33) for the mean waiting time for units arriving at the ith queue. Use of equation (33) requires calculation of the value $\bar{m}_{i-1}(1, 1)$, which can be obtained in any particular case by numerical solution of the $N(N + 1)$ linear equations (40). It is noteworthy that the calculation of the mean waiting time requires no iteration.

The techniques used in this and our previous study might be useful in the analyses of priority queuing models and other cyclic queuing models that have important practical applications. Examples of the latter are: extensions of the present models to include arbitrary switching times and/or set up times; systems of queues served in arbitrary periodic order (of which cyclic order is a special case); and within-queue disciplines other than service in order of arrival, such as service in random order.

REFERENCES

1. Cooper, R. B., and Murray, G., "Queues Served in Cyclic Order," BSTJ, *48,* No. 3 (March 1969), pp. 675–689.
2. Takács, L., "Two Queues Attended by a Single Server," Operations Research, *16,* No. 3 (May-June 1968), pp. 639–650.
3. Avi-Itzhak, B., Maxwell, W. L., and Miller, L. W., "Queuing with Alternating Priorities," Operations Research, *13,* No. 2 (March-April 1965), pp. 306–318.
4. Eisenberg, M., "Multiqueues with Changeover Times," MIT Doctoral Dissertation, September 1967.

On the Capacity of an Ensemble of Channels with Differing Parameters

By E. A. WALVICK

(Manuscript received November 17, 1969)

To provide a mathematical tool for the evaluation of cable pairs, this paper suggests a quality measure which is based on information theory. While a group of cable pairs (paired wires) with a given gauge and construction are nominally equivalent, manufacturing tolerances and differences in installation and environment lead to variation from the nominal parameters.

To measure the quality of a group of channels (for example, cable pairs leaving a central office), this paper recommends the following procedure. Choose a fraction p of the original group and evaluate the mutual information between input and output of each channel in the subgroup, subject to the input to each channel being chosen from the same process. Then, by choosing the proper process, maximize the smallest mutual information in the subgroup. This largest possible minimum mutual information is a quality measure for the subgroup. Next, apply this measure to all subgroups of fractional size p; the subgroup with the highest measure provides the numerical value of the quality measure for the original group relative to fraction p. Repeat this procedure for all p ($0 \leqq p \leqq 1$). The resulting function is the suggested quality measure of the group.

To illustrate the above measure of quality, we derive the capacity of an ensemble of channels with stationary Gaussian inputs, additive noise, and crosstalk. In the Appendix we derive the capacity of a single such channel.

I. INTRODUCTION

Cables are usually analyzed as if all of the components had a particular set of parameters (for example, nominal, worst case, and so on). Because of manufacturing tolerances, installation differences and various environmental effects, however, transmission parameters actually vary from pair to pair. To account for these variations, this paper takes an approach based on information theory. We consider the cable network to be a statistical population of channels which have parameters that vary from channel to channel. We propose a quality measure for the network based on this model.

After defining channel capacity, we present the suggested capacity definition for a group of channels (as outlined in the Abstract). Using this definition, we provide an example in Section IV in which the capacity of a group of channels with stationary Gaussian inputs, additive noise and crosstalk is derived. The capacity for a single such channel is found in the Appendix. In Section 4.1 the crosstalk is assumed to differ from channel to channel; in Section 4.2 the channel attenuation is assumed variable; and in Section 4.3 both the crosstalk and attenuation vary from channel to channel.

These results indicate the trade-off between design rate for a transmission system and the expected fraction of channels which will be capable of error-free transmission at the design rate.

This technique can also be used to evaluate different parameter distributions as may result from tighter production controls.

II. CHANNEL CAPACITY

A channel is defined as a probabilistic mapping of one stochastic process onto another (for our problem we consider the processes to be time functions). (See Fig. 1.)

Let

$s(t)$ be the input stochastic process,
$r(t)$ be the output stochastic process,
$s_T = \{s(t) : t \, \varepsilon \, [-T/2, T/2]\}$,
s be s_∞ .

Then the operation of the channel can be written in terms of a probabilistic mapping F as

$$F\{s\} = r. \tag{1}$$

The capacity is defined as the maximum (over input processes) mutual information between input and output, that is,

s(t)
CHANNEL
r(t)

Fig. 1 — Channel model.

$$C \equiv \sup_{s} \limsup_{T\to\infty} \frac{1}{T} I(r_T , s_T) = \sup_{s} \langle I(r, s)\rangle \tag{2}$$

where

$$\langle I(r, s)\rangle \equiv \limsup_{T\to\infty} \frac{1}{T} I(r_T , s_T). \tag{3}$$

$I(r_T , s_T)$ is the mutual information* between s_T and r_T
the supremum is taken over all possible distributions of input signals subject to some constraint (for example, fixed power),
and for a large class of channels including memoryless channels and colored Gaussian channels C is the maximum information rate (that is, the maximum error free transmittable rate).

The maximization of equation (2) will yield not only C, but more importantly perhaps, the properties of $s(t)$ which will achieve C.

III. CAPACITY DEFINITIONS FOR A GROUP OF CHANNELS

Now, consider the extension of the capacity definition to a group of channels. Capacity is dependent not only on the nature of the channel, but also the nature of the constraints placed on the input. For different sets of input constraints, different capacities will be obtained. This section contains two possible alternate capacity definitions [equations (5) and (6)] followed by the recommended definition [equation (10)].

A natural extension of equation (2) to a class of channels (formally the set $\{\omega : \omega \varepsilon \Omega\}$)

$$F^{(\omega)}\{s\} = r^{(\omega)}, \qquad \omega \varepsilon \Omega$$

(that is, $F^{(\omega)}$ is the mapping corresponding to channel ω) would be to define the capacity as the sum of the individual capacities, or the average capacity for an infinite set Ω. That is, the capacity of each channel is:

$$C^{(\omega)} = \sup_{s^{(\omega)}} \langle I(r^{(\omega)}, s^{(\omega)})\rangle,$$

where the supremum is performed for each channel separately, constrained as before. Then one measure of the capacity of the ensemble could be the total capacity (for a finite set):

$$C_T = \sum_{\omega \varepsilon \Omega} C^{(\omega)}. \tag{4}$$

Another measure could be the average per-channel capacity

[*] See for example Gallager.[1]

$$\langle C \rangle_T = E\{C^{(\omega)}\}, \tag{5}$$

where E denotes the expectation.

As for the single channel, equation (4) [or (5)] yields both C_T (or $\langle C \rangle_T$) and the properties of the set $S = \{s^{(\omega)} : \omega \in \Omega\}$ which will achieve C_T (or $\langle C \rangle_T$). This number defines the maximum transmittable rate when the input processes are chosen for each channel individually. In many instances this may not be a practical measure in that it may be desirable to use a single signaling set on all members of Ω. A measure using a single signaling set has been suggested in the literature:[2]

$$C_B = \sup_s \inf_{\omega \in \Omega} \langle I(r^{(\omega)}, s) \rangle. \tag{6}$$

The desirable property of C_B is that it will result in a signal distribution which, when applied to any member, $\omega \in \Omega$, will permit transmission at rates arbitrarily close to C_B with arbitrarily small probability of error. That is, C_B is the maximum rate which will work on all members of the group when one process is sent over all channels. However, this seems to be an overly pessimistic measure in that if Ω should have even one member with poor transmission properties, C_B will reflect this single poor member in exactly the same way as if all of Ω were equally bad.

To overcome this difficulty a new capacity definition is introduced. This definition is actually a function rather than a single number for the group of channels. This definition is essentially C_B , restricted to the best subset, of size p, of the original group of channels, as a function of p. To formalize this notion:

Let $\Omega_\lambda(p)$ be a subset of Ω (indexed by λ) of fractional size p. That is, for $\omega \in \Omega$

$$\Pr\{\omega : \omega \in \Omega_\lambda(p)\} = p. \tag{7}$$

Let $\Omega(p)$ be the set of all such subsets:

$$\Omega(p) = \{\Omega_\lambda(p)\}. \tag{8}$$

Find C_B for each subset $\Omega_\lambda(p)$:

$$C_B[\Omega_\lambda(p)] = \sup_s \inf_{\omega \in \Omega_\lambda(p)} \langle I(r^{(\omega)}, s) \rangle. \tag{9}$$

Finally consider the supremum over all such subsets, that is

$$\begin{aligned} C(p) &\equiv \sup_\lambda \sup_s \inf_{\omega \in \Omega_\lambda(p)} \langle I(r^{(\omega)}, s) \rangle \\ &= \sup_\lambda C_{B\Omega_\lambda}. \end{aligned} \tag{10}$$

Note that equation (10) can also be written as:

$$C(p) = \sup_{s} \sup_{\lambda} \inf_{\omega \varepsilon \Omega_\lambda(p)} \langle I(r^{(\omega)}, s)\rangle .$$

The following coding theorem can be proved almost by inspection: $C(p)$ is the supremum of rates which can be transmitted error free over at least the fraction p of the original ensemble of channels when the inputs to all channels are from the same signal distribution. (Clearly, for larger p, more channels are required to be capable of error free transmission at rates arbitrarily close to $C(p)$ than for smaller p. Therefore, $C(p)$ decreases as p increases, more of the set of poor channels included.)

$C(p)$ would then be plotted as in Fig. 2 which intentionally represents an ensemble for which most of the channels have near nominal parameters, a small percentage have worse parameters, and a small percentage better parameters. From the figure, if all channels must have error-free transmission, then the design rate for any system can be no greater than C_B. However, if design criterion only requires p_1 of the channels to be error free then rate C_1 can be used. Similarly, if only p_2 of the channels need operate without errors, rate C_2 can be used.

$C(p)$ is a useful measure for the following reasons:

(*i*) If it is desired that a given fraction of the channels have satisfactory transmission, then $C(p)$ indicates the maximum permissible rate.

(*ii*) If the objective is to provide a given transmission rate $C(p)$, then the value of p indicates what percentage of the channels will be capable of operating without errors.

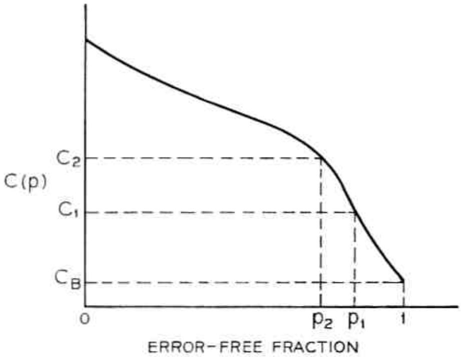

Fig. 2 — $C(p)$ vs. p.

(*iii*) When deciding between two alternative groups of channels, $C(p)$ can indicate those transmission rates for which one type is better than the other.

IV. EXAMPLE

To provide an example for the use of the $C(p)$ function, consider an ensemble of channels, each of which can be modeled as shown in Fig. 3.*

where

$H(\omega)$ is the channel transfer function,
$X(\omega)$ is the crosstalk transfer function,
$s(t)$ is the signal with one-sided power spectral density $S_S(\omega)$,
$\hat{s}(t)$ is the signal on an adjacent channel with the same power spectral density, and
$n(t)$ is the noise with power spectral density $S_N(\omega)$.

If only stationary Gaussian inputs are considered the average mutual information between the input and output of any channel in the ensemble is:[3,4]

$$\langle I(r, s)\rangle = -\frac{1}{2\pi}\int_0^\infty \log\left(1 - \frac{|\,S_{SR}(\omega)\,|^2}{S_R(\omega)S_S(\omega)}\right) d\omega, \tag{11}$$

where

$S_S(\omega)$ is the input signal power spectral density,

$$S_R(\omega) = S_S(\omega)\,|\,H(\omega)\,|^2\,(1 + |\,X(\omega)\,|^2) + S_N(\omega), \tag{12}$$

is the output power spectral density, and

$$S_{SR}(\omega) = S_S(\omega)H^*(\omega), \tag{13}$$

is the input-output cross power spectral density.
Then,

$$\langle I(r, s)\rangle = -\frac{1}{2\pi} \cdot \int_0^\infty \log\left[1 - \frac{S_S^2(\omega)\,|\,H(\omega)\,|^2}{S_S^2(\omega)\,|\,H(\omega)\,|^2\,(1 + |\,X(\omega)\,|^2) + S_S(\omega)S_N(\omega)}\right] d\omega. \tag{14}$$

* ω is used here to represent frequency and not probability spaces as in the first part of this paper.

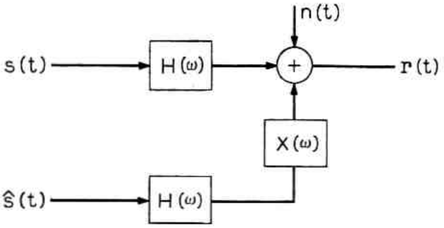

Fig. 3 — Channel model.

Define

$$\mathcal{S}_N(\omega) = \frac{S_N(\omega)}{\mid H(\omega) \mid^2}. \tag{15}$$

Then

$$\langle I(r, s)\rangle = \frac{1}{2\pi}\int_0^\infty \log\left[1 + \frac{S_S(\omega)}{\mid X(\omega) \mid^2 S_S(\omega) + \mathcal{S}_N(\omega)}\right] d\omega. \tag{16}$$

4.1 *Fixed Noise and Channel, Distributed Crosstalk*

Assume that the crosstalk parameters of all of the cables are the same except for a multiplier, that is:

$$\mid X_\xi(\omega) \mid = \epsilon_\xi \mid X(\omega) \mid. \tag{17}$$

Equation (16) indicates that for any choice of $S_S(\omega)$, $\langle I(r, s)\rangle$ decreases as ϵ_ξ increases. Thus, while equation (9) requires a minimization followed by a maximization, it is clear in the case that the "inf" for a given p occurs for the largest ϵ_ξ in the subset $\Omega_\lambda(p)$. Further, the "sup" is achieved using the spectrum that achieves capacity on the channel with the largest ϵ_ξ. Finally, the "sup" in equation (10) is achieved by choosing the $\Omega_\lambda(p)$ such that $0 \leqq \epsilon \leqq \epsilon_p$ where ϵ_p is such that

$$p = \int_0^{\epsilon_p} p_{\epsilon_\xi}(\epsilon)\, d\epsilon, \tag{18}$$

where $p_{\epsilon_\xi}(\epsilon)$ is the probability density of ϵ_ξ. Thus using the capacity result obtained in the Appendix

$$S_0(\omega) = [\mathcal{S}_N(\omega_{\max}) - \mathcal{S}_N(\omega)][1 - \tfrac{1}{4}\epsilon_p^2 \mid X(\omega) \mid^2(\epsilon_p^2 \mid X(\omega) \mid^2 + 1) + \cdots]. \tag{19}$$

The capacity function can now be found:

$$C(p) = \frac{1}{2\pi}\left\{\int_0^{\omega_{\max}} \left[\log\left[1 + \frac{1}{\epsilon_p^2 \mid X(\omega) \mid^2}\right] + \tfrac{1}{2}\log\left\{1 + 4\epsilon_p^2 \mid X(\omega) \mid [\epsilon_p^2 \mid X(\omega) \mid^2 + 1] \frac{S_N(\omega_{\max})}{S_N(\omega)}\right\}\right] d\omega\right\}. \tag{20}$$

Consider the following example to illustrate the above:

Let $S_N(\omega)$ be zero. Then from equation (16) the mutual information is independent of the signal spectrum and

$$C(p) = \frac{1}{2\pi}\int_0^\infty \log\left[1 + \frac{1}{\epsilon_p^2 \mid X(\omega) \mid^2}\right] d\omega. \tag{21}$$

(This is achieved for any spectrum with finite power which is strictly greater than zero for all frequencies.) Let

$$\mid X(\omega) \mid^2 = \omega^2 \tag{22}$$

and let $\log \epsilon_p$ be normally distributed with mean -8.2 and $\sigma = 0.17$. (The crosstalk figures are idealizations of typical figures for 22 gauge PIC). Then

$$p = \operatorname{erf}\left\{\frac{\log \epsilon_p + 8.2}{0.17}\right\} + 0.5 \tag{23}$$

where

$$\operatorname{erf}(x) = \frac{1}{(2\pi)^{\frac{1}{2}}}\int_0^x e^{-y^2/2}\, dy, \tag{24}$$

and

$$C(p) = \frac{1}{2\epsilon_p \ln 2} \quad \text{(in bits)}$$

$$= \frac{1}{2\epsilon_p} \quad \text{(in nats)}. \tag{25}$$

These results are indicated in Fig. 4. For $p = 0.8$, $C(p)$ is 6.3×10^7. For $p = 0.1$, $C(p)$ is 1.3×10^8. Thus a factor of two is realized in capacity by reducing the required usable fraction of channels by a factor of eight.

4.2 *Fixed Crosstalk and Noise, Distributed Channel*

Assume

$$S_N(\omega) = \epsilon_N , \tag{26}$$

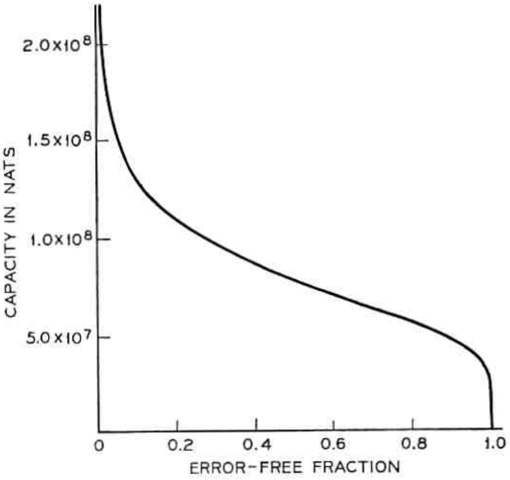

Fig. 4 — Capacity vs. fraction of satisfactory channels.

$$| H_\xi(\omega) | = e^{-\alpha\xi(\omega)}, \tag{27}$$

and that $\alpha_\xi(\omega)$ is such that

$$\alpha_{\xi_1}(\omega_i) > \alpha_{\xi_2}(\omega_i) \Rightarrow \alpha_{\xi_1}(\omega_j) \geqq \alpha_{\xi_2}(\omega_j). \tag{28}$$

That is, if the attenuation of channel ξ_1 is greater than the attenuation of ξ_2 at one frequency, it is greater at all frequencies. Then, again design for the worst case, that is,

$$p = \int_0^{\alpha_p} p_{\alpha_\xi}(\alpha) \, d\alpha. \tag{29}$$

(Again the smaller α, the "better" the channel.)

$$S_0(\omega) = \epsilon_N[e^{2\alpha_p(\omega_{\max})} - e^{2\alpha_p(\omega)}][1 - \tfrac{1}{4} | X(\omega) |^2 (| X(\omega) |^2 + 1)], \tag{30}$$

and

$$C(p) = \frac{1}{2\pi}\left\{\int_0^{\omega_{\max}} \left\{\log\left[1 + \frac{1}{| X(\omega) |^2}\right]\right.\right.$$
$$\left.\left. + \tfrac{1}{2}\log\left[1 + 4 | X(\omega) |^2 (| X(\omega) |^2 + 1) \frac{e^{2\alpha_p(\omega_{\max})}}{e^{2\alpha_p(\omega)}}\right]\right\} d\omega\right\}. \tag{31}$$

4.3 *Fixed Noise, Distributed Channel and Crosstalk*

Proceeding as in the previous two sections, choose the signal spectrum to achieve capacity on a channel with a particular ϵ_{p_s} and α_{p_s}, that is:

$$
\begin{aligned}
S_{0_x}(\omega) &= \epsilon_N[\exp[2\alpha_{p_x}(\omega_{\max})] - \exp[2\alpha_{p_x}(\omega)]] \\
&\quad \cdot [1 - \tfrac{1}{4}\epsilon_{p_x}^2 \mid X(\omega) \mid^2 (\epsilon_{p_x}^2 \mid X(\omega) \mid^2 + 1)], \qquad (32)
\end{aligned}
$$

$$
\begin{aligned}
C_x(p) &= \frac{1}{2\pi}\left\{\int_0^{\omega_{\max}} \left\{\log\left(1 + \frac{1}{\epsilon_{p_x}^2 \mid X(\omega) \mid^2}\right)\right.\right. \\
&\quad + \tfrac{1}{2}\log\left[1 + 4\epsilon_{p_x}^2 \mid X(\omega) \mid^2 (\epsilon_{p_x}^2 \mid X(\omega) \mid^2 + 1)\right. \\
&\quad \left.\left.\left.\cdot \frac{\exp[2\alpha_{p_x}(\omega_{\max})]}{\exp[2\alpha_{p_x}(\omega)]}\right]\right\} d\omega\right\}. \qquad (33)
\end{aligned}
$$

Here, p is found by integrating the joint density of α_ξ and ϵ_ξ over the region where the mutual information is greater than $C_x(p)$. That is,

$$
p = \iint_R p_{\epsilon_\xi,\alpha_\xi}(\epsilon, \alpha)\, d\epsilon\, d\alpha \qquad (34)
$$

where R is the region in the (ϵ, α) space, where

$$
\frac{1}{2\pi}\int_0^{\omega_{\max}} \log\left[1 + \frac{1}{\epsilon_\xi \mid X(\omega) \mid^2 + \epsilon_N \exp 2\alpha_\xi(\omega)}\right] d\omega \geqq C_x(p). \qquad (35)
$$

The difficulty with this problem at this point is that there is no unique ϵ_{p_x} and α_{p_x} which yield a given p in equation (34). Thus it is necessary to search all such sets $(\epsilon_{p_x}, \alpha_{p_x})$ which yield the same p, and choose the one which yields the largest $C_x(p)$. While this may seem difficult, it can be implemented with a computer search.

V. CONCLUDING REMARKS

The quality measure suggested herein should be considered part of the development of a technique for comparison of groups of cables.

It should be pointed out that, while the definition of capacity for a single channel is straightforward, the definition of capacity for an ensemble of channels depends on certain assumptions concerning the use of the channels.

$\langle C \rangle_T$, the average per channel capacity, [equation (5)] assumes each channel will be used individually, and optimized individually. C_B, Blackwell's definition, [equation (6)] assumes the same input distributions will be used for all channels in the group and that the channels are used individually. Further, no more than the minimum rate transmittable over the worst channel will be sent over each channel.

The recommended definition, $C(p)$, equation (12), is similar to C_B except that the fraction $1 - p$ of the worst channels will not work satisfactorily and could be discarded.

The example carried out in Section IV uses a new result on the capacity of a single channel with crosstalk and additive Gaussian noise which is derived in the Appendix. The discussion on the calculus of variations, contained before equation (40), can be used to rigorously prove some old results for optimum spectra in the presence of additive Gaussian noise [see footnote, following equation (52)].

In applying $C(p)$, it is noted that when only one parameter is unknown (for example, the magnitude of the crosstalk), the capacity of the group of channels as a function of p (the fraction of usable channels) is a simple calculation. However, when more than one parameter is unknown (for example, the magnitude of the crosstalk as well as the channel attenuation), search techniques are indicated.

APPENDIX

Capacity Calculation for a Single Channel

Consider the channel model of Fig. 3. The mutual information expression is given in equation (16) and is repeated below:

$$\langle I(r, s)\rangle = \frac{1}{2\pi}\int_0^\infty \log\left[1 + \frac{S_S(\omega)}{|\,X(\omega)\,|^2\, S_S(\omega) + \mathcal{S}_N(\omega)}\right] d\omega. \tag{36}$$

Now capacity for this channel is defined as

$$C = \sup_{S_S(\omega)} \langle I(r, s)\rangle, \tag{37}$$

subject to

$$-\frac{1}{2\pi}\int_0^\infty S_S(\omega)\, d\omega \leqq P = \text{power},$$

and

$$S_S(\omega) \geqq 0 \qquad \text{for all } \omega. \tag{38}$$

The method of variational calculus can be applied to determine the optimum solution.

Let

$S_0(\omega)$ be the assumed optimum solution, and
$\delta\cdot\epsilon(\omega)$ be a small perturbation, then
$S_S(\omega) = S_0(\omega) + \delta\cdot\epsilon(\omega)$.

In order to account for $S_S(\omega) \geqq 0$, $\delta \cdot \epsilon(\omega)$ must be nonnegative whenever $S_0(\omega) = 0$. That is, in order that $S_S(\omega)$ be a power spectral density, whenever $S_0(\omega) = 0$, the perturbation at that frequency must be such that the resulting density at that frequency remain nonnegative. Note that since $S_0(\omega)$ is to be the optimum solution, $\langle I(r, s) \rangle$ must be a maximum at $\delta = 0$ *for all permissible* $\delta \cdot \epsilon(\omega)$. This implies that $\partial \langle I \rangle / \partial \delta \mid_{\delta=0}$ will be negative for all permissible $\delta \cdot \epsilon(\omega)$ which approach zero from the positive side $(\delta \cdot \epsilon(\omega) \rightarrow 0^+)$, and $\partial \langle I \rangle / \partial \delta \mid_{\delta=0}$ will be positive for $\delta \cdot \epsilon(\omega) \rightarrow 0^-$. Or finally, $\partial \langle I \rangle / \partial \delta \mid_{\delta=0} = 0$ whenever $\delta \cdot \epsilon(\omega)$ can approach zero from either the positive or negative side and the derivative is defined. Now,

$$2\pi \frac{\partial \langle I \rangle}{\partial \delta} = \frac{\partial}{\partial \delta} \left[\int_0^{+\infty} \log \left\{ 1 + \frac{1}{\mid X(\omega) \mid^2 + \dfrac{\mathcal{S}_N(\omega)}{S_0(\omega) + \delta \cdot \epsilon(\omega)}} \right\} d\omega \right] + \frac{\partial}{\partial \delta} \left[-\mu \left\{ \int_0^{+\infty} [S_0(\omega) + \delta \cdot \epsilon(\omega)] \, d\omega - 2\pi P \right\} \right], \tag{39}$$

where μ is a Lagrangian multiplier used to introduce the power constraint.

Or

$$\left. \frac{\partial \langle I \rangle}{\partial \delta} \right|_{\delta=0} = \int_0^\infty \epsilon(\omega) \left[\frac{\mathcal{S}_N(\omega)}{A(\omega) S_0^2(\omega) + B(\omega) S_0(\omega) + \mathcal{S}_N^2(\omega)} - \mu \right] d\omega, \tag{40}$$

where

$$A(\omega) = \mid X(\omega) \mid^2 (\mid X(\omega) \mid^2 + 1),$$
$$B(\omega) = [2 \mid X(\omega) \mid^2 + 1] \, \mathcal{S}_N(\omega).$$

In equation (20) one now considers all permissible $\epsilon(\omega)$. Whenever $S_0(\omega)$ is nonzero (that is, $\delta \cdot \epsilon(\omega)$ is unrestricted), the integrand is zero, since for these frequencies, $\epsilon(\omega)$ is arbitrary. However, when $S_0(\omega)$ is zero $(\delta \cdot \epsilon(\omega) \geqq 0)$ all that can be said is that the integrand is *negative*. That is, as one approaches the boundary of permissible $\delta \cdot \epsilon(\omega)$, $\langle I(r, s) \rangle$ must be monotonically increasing.

This yields:

$$\frac{\mathcal{S}_N(\omega)}{A(\omega) S_0^2(\omega) + B(\omega) S_0(\omega) + \mathcal{S}_N^2(\omega)} = \mu \qquad \text{for all} \quad \omega \quad \text{such that} \quad S_0(\omega) \neq 0, \tag{41}$$

and

$$\frac{1}{\mathcal{S}_N(\omega)} - \mu \leqq 0 \quad \text{for all} \quad \omega \quad \text{such that} \quad S_0(\omega) = 0. \tag{42}$$

(This equation is simply the integrand for $S_0(\omega) = 0$.)

Or from the above for $S_0(\omega) \neq 0$,

$$S_0(\omega) = \frac{-B(\omega) \pm \{B^2(\omega) - 4A(\omega)C(\omega)\}^{\frac{1}{2}}}{2A(\omega)}$$

where

$$C(\omega) = \mathcal{S}_N^2(\omega) - \mathcal{S}_N(\omega)/\mu. \tag{43}$$

Now, $S_0(\omega)$ must be nonnegative. In order that equation (43) yield a nonnegative result, the positive root must be taken, and further the following inequality must be satisfied:

$$\{B^2(\omega) - 4A(\omega)C(\omega)\}^{\frac{1}{2}} \geqq B(\omega). \tag{44}$$

Relation (44) implies

$$C(\omega) \leqq 0 \quad \text{for all } \omega \text{ such that } S_0(\omega) > 0. \tag{45}$$

Relation (42) (which implies $1/\mu \leqq 0$) can be rewritten as

$$C(\omega) \geqq 0 \quad \text{for all } \omega \text{ such that } S_0(\omega) = 0. \tag{46}$$

If we assume that $\mathcal{S}_N(\omega)$ is a monotonically increasing function* of ω, relations (45) and (46) imply

$$\frac{1}{\mu} = \mathcal{S}_N(\omega_{\max}), \tag{47}$$

and

$$S_0(\omega) = 0 \qquad \omega \geqq \omega_{\max}. \tag{48}$$

Hence† rewriting equation (43)

$$S_0(\omega) = \mathcal{S}_N(\omega) \frac{\{1 + 4A(\omega)\,\mathcal{S}_N(\omega_{\max})/\,\mathcal{S}_N(\omega)\}^{\frac{1}{2}} - \{1 + 4A(\omega)\}^{\frac{1}{2}}}{2A(\omega)} \tag{49}$$

where $\omega_{\max}$ can be found from

$$\frac{1}{2\pi} \int_0^{\omega_{\max}} S_0(\omega)\, d\omega = P. \tag{50}$$

This can be used to obtain a relation between $\omega_{\max}$ and P.

* This is not at all necessary. It just simplifies the form of the following equations.

† In what follows, it is understood that the expression given for $S_0(\omega)$ holds for $\omega \leqq \omega_{\max}$, and that $S_0(\omega) = 0$ otherwise.

Now if the crosstalk is small [$|X(\omega)| \ll 1$], each of the radicals in the equation for $S_0(\omega)$ can be approximated by the first few terms in the binomial expansion. Then

$$S_0(\omega) = [\mathcal{S}_N(\omega_{\max}) - \mathcal{S}_N(\omega)]\Big[1 - \tfrac{1}{4} \,|\, X(\omega) \,|^2 \,(|\, X(\omega) \,|^2 + 1) + \tfrac{1}{8} \,|\, X(\omega) \,|^2 \,(|\, X(\omega) \,|^2 + 1)\Big(\frac{\mathcal{S}_N(\omega)_{\max}}{\mathcal{S}_N(\omega)} - 1\Big) + \cdots\Big], \qquad (51)$$

$$S_0(\omega) \cong [\mathcal{S}_N(\omega_{\max}) - \mathcal{S}_N(\omega)][1 - \tfrac{1}{4} \,|\, X(\omega) \,|^2 \,(|\, X(\omega) \,|^2 + 1)],$$
$$\cong [\mathcal{S}_N(\omega_{\max}) - \mathcal{S}_N(\omega)]. \qquad (52)$$

This result indicates that for small crosstalk, the signal spectrum ought to be designed independent of the crosstalk spectrum. Equation (52) is the familiar "spectrum filling" result for additive Gaussian noise.* This is shown in Fig. 5. The exact solution [equation (51)] (superimposed with the broken lines) simply implies some shaping. Figure 6 illustrates the solution when $S_N(\omega)$ is not monotonically increasing.

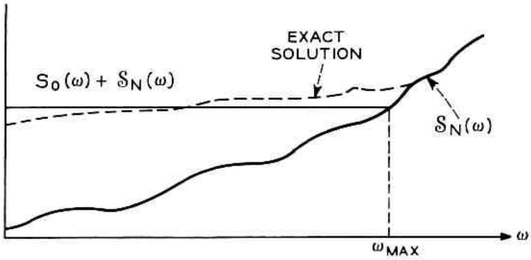

Fig. 5 — Typical optimum spectrum 1.

With the solution obtained for $S_0(\omega)$

$$\text{Capacity} = \frac{1}{2\pi}\Big\{\int_0^{\omega_{\max}} \log\Big[1 + \frac{1}{|\, X(\omega)\,|^2}\Big]\, d\omega + \frac{1}{2}\int_0^{\omega_{\max}} \log\Big\{1 + 4 \,|\, X(\omega) \,|^2 \,[|\, X(\omega) \,|^2 + 1]\, \frac{\mathcal{S}_N(\omega_{\max})}{\mathcal{S}_N(\omega)}\Big\}\, d\omega\Big\}. \qquad (53)$$

Using the small crosstalk approximation for $S_0(\omega)$:

* This result is contained in Fano[5], pp. 173ff. That proof is not as rigorous as the one presented here as Fano does not prove that the optimum spectrum is zero whenever the noise is greater than a threshold. Fano's result can be proved directly by noting the discussion preceding equation (21) herein.

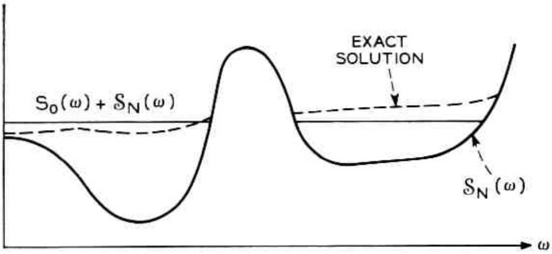

Fig. 6 — Typical optimum spectrum 2.

Capacity

$$
\begin{aligned}
&\cong \frac{1}{2\pi}\int_0^{\omega_{\max}} \log\left[1 + \frac{S_N(\omega_{\max}) - S_N(\omega)}{|X^2(\omega)|\,[S_N(\omega_{\max}) - S_N(\omega)] + S_N(\omega)}\right] d\omega, \\
&\cong \frac{1}{2\pi}\int_0^{\omega_{\max}} \log\left[1 + \frac{S_N(\omega_{\max}) - S_N(\omega)}{|X(\omega)|^2\, S_N(\omega_{\max}) + S_N(\omega)}\right] d\omega, \\
&\cong -\frac{1}{2\pi}\int_0^{\omega_{\max}} \log\left[|X(\omega)|^2 + \frac{S_N(\omega)}{S_N(\omega_{\max})}\right] d\omega. \qquad (54)
\end{aligned}
$$

(Note that even for small crosstalk, capacity is still a function of the crosstalk spectrum.)

REFERENCES

1. Gallager, R. G., *Information Theory and Reliable Communication,* New York, N. Y.: John Wiley & Sons, Inc., 1968, pp. 369–370.
2. Blackwell, D., Breiman, L., and Thomasian, A. J., "The Capacity of a Class of Channels," *Ann. of Math. Stat., 30,* No. 4, (December 1959), pp. 1229–1241.
3. Pinsker, M. S., "A Quantity of Information of a Gaussian Random Stationary Process, Contained in a Second Process Connected with it in a Stationary Manner," *Doklady Akad. Nauk S.S.S.R., 99,* No. 2, (February 1954), pp. 213–216. (Translation available as A.C.S.I.L. Trans. No. 1266, October 1960).
4. Kolmogorov, A. N., "On the Shannon Theory of Information Transmission in the Case of Continuous Signals," *IRE Trans., IT-2,* No. 4 (December 1956), pp. 102–108.
5. Fano, R. M., *Transmission of Information,* Cambridge, Mass.: The MIT Press, 1961, pp. 173–174.

A Stationary Phase Method for the Computation of the Far Field of Open Cassegrain Antennas

By W. H. IERLEY and H. ZUCKER

(Manuscript received October 17, 1969)

A method is presented for the computation of the far field radiation patterns of paraboloid reflector antennas by using a modified stationary phase approximation to eliminate one integration. This method is applicable to open cassegrain, offset paraboloid and horn reflector antennas. For symmetrical paraboloid antennas the modified approximation reduces to the exact expression obtained by direct integration.

The errors introduced by the stationary phase and modified stationary phase approximations are investigated. Specifically the far field of an open cassegrain with a 128 *wavelength aperture diameter is computed by the approximate method up to* 20 *degrees off-axis. The difference between these radiation patterns and those computed by double integration, is less than a few tenths of a* dB *up to* 1.0 *degree, and less than a few hundredths of a* dB *at larger angles off-axis.*

In order to estimate the computational advantage of this approximation, the number of points required for integration of an oscillatory function by Simpson's rule is also examined and it is determined that at least 6 *points per cycle are necessary to obtain* 4 *decimal accuracy. For fewer points the error is appreciable.*

I. INTRODUCTION

The computation of the far field radiation patterns of large reflector antennas is of importance in predicting the performance of satellite ground stations. For example, the antenna sidelobes contribute to the system noise temperature and may cause interference with other communication systems. The open cassegrain antenna[1] is a particularly suitable configuration for obtaining low sidelobe levels, since blocking by the subreflector and its supports are eliminated. A further advantage, resulting from this feature, is that the radiation pattern can be accurately predicted and it has been shown to be in good agreement with experimental results.[1]

The previous computations of the radiation patterns of open cassegrain antennas have been performed by precise computation of the appropriate diffraction integrals, generally requiring a double numerical integration. For large angles off-axis, such computations require considerable computation time. It is, however, for large angles that the integrals which are used for the computation of the far field radiation patterns are of a form which is suitable for approximation by the method of stationary phase. This method was initially applied to eliminate the azimuthal ϕ integration, but it was subsequently recognized that certain terms in the approximation are related to the asymptotic expansions of Bessel functions. The stationary phase approximation could therefore be modified, with the observed result that the far field radiation pattern can be computed with good accuracy also in the immediate vicinity of the main beam.

In the following sections we derive the stationary phase approximation and present a geometrical interpretation of the location of the stationary points. For the far field on axis, the ϕ integration is performed in closed form and it is shown that the antenna gain is the same for both perpendicular polarizations. Numerical computations are performed to estimate the error introduced by the stationary phase and modified methods. The extended range of applicability of the latter method is evident from the computations. The number of points per cycle needed to obtain an accurate value for an integral of an oscillatory function is also examined, and it is shown that for the functions considered at least 6 points/cycle are needed.

The far field radiation patterns of an open cassegrain with a 128 wavelength aperture diameter are computed with this method up to 20 degrees off-axis. In the vicinity of 20 degrees, the relative sidelobe levels are less than -65 dB or about 15 dB below isotropic.

1.1 *The Far Field*

The far electric field $\mathbf{E}_f$ of a paraboloid reflector antenna in an angular region about the axis, can be, based on the projected aperture field method, related to the reflected field at the aperture, $\mathbf{E}_r$, by the following expression:[2]

$$\mathbf{E}_f = \frac{j \exp(-jkR_a)}{\lambda R_a} \iint_A \mathbf{E}_r(x_p, y_p) \exp(jk\boldsymbol{\varrho}_p \cdot 1_{R_a})\, ds \qquad (1)$$

where

λ = wavelength,
$k = 2\pi/\lambda$ (propagation constant),
R_a = the distance to the far field observation point,
$\boldsymbol{\varrho}_p$ is a vector in the aperture plane,
$\mathbf{1}_{R_a}$ is a unit vector which specifies the direction of the observation point,
A is the aperture area.

Specifically the direction of the observation point $\mathbf{1}_{R_a}$ expressed in terms of the unit vectors of the aperture $(x_p\ ,\ y_p\ ,\ z_p)$ coordinate system is:

$$\mathbf{1}_{R_a} = \mathbf{1}_{x_p} \sin\theta_a \cos\phi_a + \mathbf{1}_{y_p} \sin\theta_a \sin\phi_a + \mathbf{1}_{z_p} \cos\theta_a \tag{2}$$

where θ_a and ϕ_a are the far field observation angles, and

$$\boldsymbol{\varrho}_p = \mathbf{1}_{x_p} x_p + \mathbf{1}_{y_p} y_p\ . \tag{3}$$

For an open cassegrain the incident fields at the main reflector can be more readily computed in a spherical coordinate system with the axis aligned with the horn subreflector axis as shown in Fig. 1. Therefore, the integrations in equation (1) are also performed in this coordinate system. The relations between aperture coordinates and the fields in the two coordinate systems were derived previously[1] and are

$$x_p = r[\cos\theta_0 \sin\theta \cos\phi + \sin\theta_0 \cos\theta], \tag{4}$$

$$y_p = r \sin\theta \sin\phi, \tag{5}$$

$$\begin{aligned}\frac{2f}{r}\mathbf{E}_r = \mathbf{1}_{x_p}\{&[\sin\theta_0 \sin\theta - \cos\phi(1 + \cos\theta\cos\theta_0)]E_\theta \\ &+ \sin\phi(\cos\theta + \cos\theta_0)E_\phi\} \\ &- \mathbf{1}_{y_p}\{\sin\phi(\cos\theta_0 + \cos\theta)E_\theta \\ &- [\sin\theta \sin\theta_0 - \cos\phi(1 + \cos\theta\cos\theta_0)]E_\phi\},\end{aligned} \tag{6}$$

where E_θ and E_ϕ are the θ and ϕ components of the incident electric field, f is the focal length of the paraboloid and r is the equation of the paraboloid surface in the θ, ϕ coordinate system

$$r = \frac{2f}{1 + \cos\theta_0 \cos\theta - \sin\theta \sin\theta_0 \cos\phi} \tag{7}$$

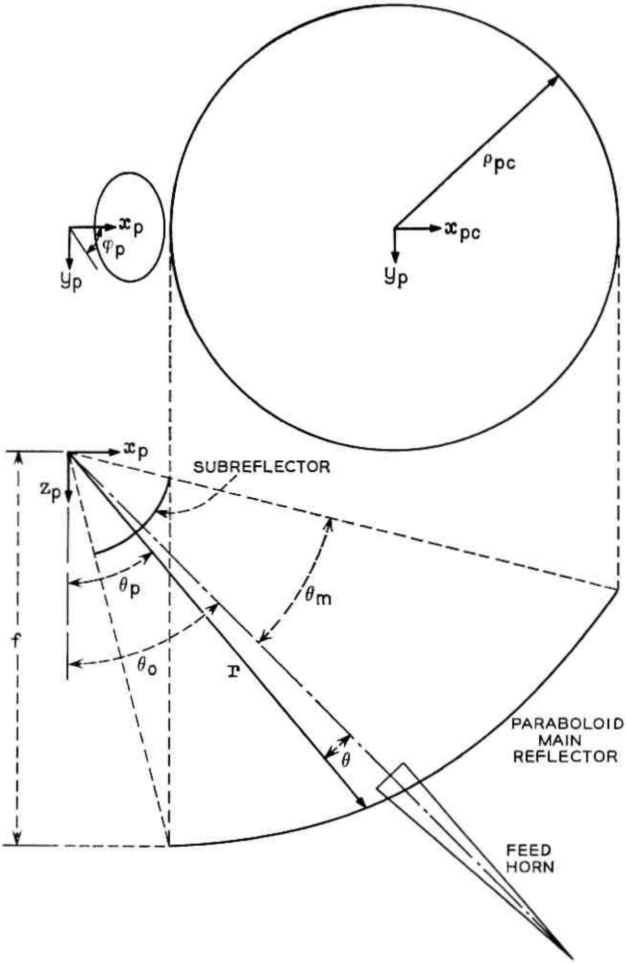

Fig. 1 — Open cassegrain antenna.

θ_0 is the offset angle. The surface element

$$ds = r^2 \sin \theta \, d\theta \, d\phi. \tag{8}$$

Heretofore the above integral has been evaluated by double integration using Simpson's rule. Although rather accurate results can be obtained in this fashion, computation time for a given angle, θ_a increases roughly proportional to $(\sin \theta_a)^2$ (see Appendix C). As a result, except for the mainlobe and first few sidelobes of the far field, this approach becomes extremely time consuming. To reduce the computation time for large values of the off-axis angle θ_a, the stationary phase approximation to the ϕ integration is investigated.

1.2 *The Method of Stationary Phase*

Consider an integral of the form

$$I(k \sin \theta_a) = \int_a^b g(\phi) \exp \{j(k \sin \theta_a)\psi(\phi)\} \, d\phi \tag{9}$$

where $(k \sin \theta_a)$ is large, $\psi(\phi)$ is a real function and $g(\phi)$ is a slowly varying function. The method of stationary phase[3] approximates the above integral to $O(1/k \sin \theta_a)$ by considering only contributions in the vicinity of the stationary points ϕ_i where $\psi'(\phi_i) = 0$.

Under these conditions

$$I(k \sin \theta_a) \approx \sum_i g(\phi_i)\left(\frac{2j\pi}{k \sin \theta_a \psi''(\phi_i)}\right)^{\frac{1}{2}} \exp \{jk \sin \theta_a \psi(\phi_i)\}. \tag{10}$$

This method has been applied to the ϕ integration in the expression for the far field (1).

From equations (1), (2) and (3) $\psi(\phi)$ can be written as

$$\psi(\phi) = [1_{x_p} \cos \phi_a + 1_{y_p} \sin \phi_a] \cdot \boldsymbol{\rho}_p \, . \tag{11}$$

For specified observation angles θ_a and ϕ_a, equation (11) can be considered as the projection of the vector $\boldsymbol{\rho}_p$ in the direction of the unit vector $1_{\rho_a} = 1_{x_p} \cos \phi_a + 1_{y_p} \sin \phi_a$. For the problem under consideration $\boldsymbol{\rho}_p$ is a function of θ and ϕ. It has been shown previously[1] that for constant θ, the vector $\boldsymbol{\rho}_p$ describes a circle as ϕ varies from 0 to 2π. The equation of the circle is

$$\left(x_p - \frac{2f \sin \theta_0}{\cos \theta_0 + \cos \theta}\right)^2 + y_p^2 = \left(\frac{2f \sin \theta}{\cos \theta + \cos \theta_0}\right)^2 . \tag{12}$$

A family of such circles for an offset angle θ_0 of 47.5° is shown in Fig. 2.

Therefore the condition $\psi'(\phi) = 0$ corresponds to determining the extreme values of the projections of the vector $\boldsymbol{\rho}_p$ in the direction of the unit vector 1_{ρ_a}. It is evident from Fig. 2 that as $\boldsymbol{\rho}_p$ describes a constant θ circle two extreme values for the projections exist, namely at those two points on the circle such that tangents to the circle passing through the points intersect normally a line in the direction of the unit vector 1_{ρ_a}. Furthermore the difference between the two extreme projections is the circle diameter.

The expressions for the stationary points and the other values which enter in the evaluation of the ϕ integration are derived in Appendix A.

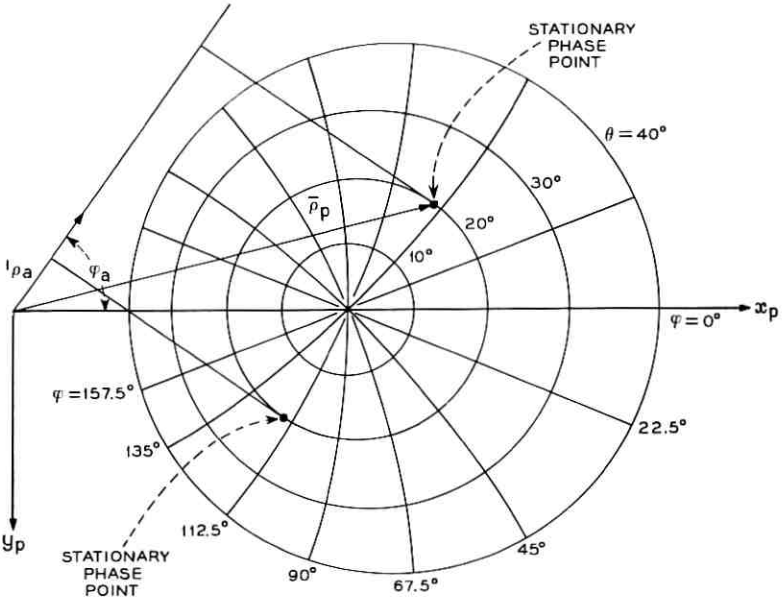

Fig. 2 — Projection circles of the paraboloid reflector for $\theta_0 = 47.5°$.

1.3 *Approximate Values*

In order to approximate the ϕ integration in equation (1), namely

$$\mathbf{E} = \int_0^{2\pi} r^2 \mathbf{E}_r \exp\{jk \sin \theta_a (x_p \cos \phi_a + y_p \sin \phi_a)\}\, d\phi \tag{13}$$

by the stationary phase method, the reflected field $\mathbf{E}_r$ has to be determined at the stationary points. This field has an explicit ϕ dependence for a TE_{11} mode or combined $TE_{11} - TM_{11}$ excitation. For these modes it has been shown[1] that for x and y polarization the field components E_θ and E_ϕ are

$$x \text{ polarization } \begin{cases} E_\theta = E(\pi/2) \cos \phi & (14) \\ E_\phi = -E(0) \sin \phi & (15) \end{cases}$$

$$y \text{ polarization } \begin{cases} E_\theta = E(\pi/2) \sin \phi & (16) \\ E_\phi = E(0) \cos \phi & (17) \end{cases}$$

where $E(\pi/2)$ and $E(0)$ denote the θ dependence of the fields when the feed horn is excited for y polarization in the planes $\phi = \pi/2$ and $\phi = 0$, respectively.

As shown in Appendix A, it is sufficient to evaluate the fields at the stationary points for one polarization only. For the other polarization the fields are then readily obtained. Therefore only y polarization will be considered. Furthermore it will be assumed that the radiated field from the subreflector has been computed at a constant radius from the focal point of the hyperboloid subreflector.

By assuming a $1/r$ dependence for subreflector fields, the relation between the fields is

$$\mathbf{E}_r(\theta, \phi) = \frac{2f}{1 + \cos \theta_0} \frac{\mathbf{E}_c(\theta, \phi)}{r} \tag{18}$$

where $\mathbf{E}_c$ is the field at the distance $2f/(1 + \cos \theta_0)$ and r is the equation of the paraboloid (7). The spherical phase dependence of the field in equation (18) is suppressed.

With the approximation (18) and the stationary phase approximation to the ϕ integration, the far field using equation (1) is obtained from the integral (see Appendix A)

$$(\mathbf{E}_{fy}) = j \frac{\exp(-jkR_a)}{\lambda R_a} \int_0^{\theta_m} (\mathbf{E}_y) \sin \theta \, d\theta \tag{19}$$

where the subscript y designates that the far field is for y polarization, and θ_m is the illumination angle. The y and x (cross polarization) components of $(\mathbf{E}_y)$ namely $(E_y)_y$ and $(E_y)_x$ are given by

$$\begin{aligned}(E_y)_y = {} & \frac{-\pi(2f)^2 e^{j\beta}}{c(a^2 - b^2 \cos^2 \phi_a)(1 + \cos \theta_0)} \left[\frac{2}{\pi\left(\alpha + \dfrac{\pi}{4}\right)}\right]^{\frac{1}{2}} \\ & \cdot [c(e^{j\alpha} + e^{-j\alpha})\{a \sin^2 \phi_a \, E_c(\pi/2) + c \cos^2 \phi_a \, E_c(0)\} \\ & - b \sin^2 \phi_a \cos \phi_a (e^{j\alpha} - e^{-j\alpha})\{cE_c(\pi/2) - aE_c(0)\}]\end{aligned} \tag{20}$$

and the cross polarized component $(E_y)_x$ is

$$\begin{aligned}(E_y)_x = {} & \frac{\pi(2f)^2 e^{j\beta} \sin \phi_a}{c(a^2 - b^2 \cos^2 \phi_a)(1 + \cos \theta_0)} \left[\frac{2}{\pi\left(\alpha + \dfrac{\pi}{4}\right)}\right]^{\frac{1}{2}} \\ & \cdot [b(e^{j\alpha} - e^{-j\alpha})\{a \sin^2 \phi_a \, E_c(0) + c \cos^2 \phi_a \, E_c(\pi/2)\} \\ & + c \cos \phi_a (e^{j\alpha} + e^{-j\alpha})\{cE_c(0) - aE_c(\pi/2)\}]\end{aligned} \tag{21}$$

where

$$a = 1 + \cos \theta \cos \theta_0, \tag{22a}$$

$$b = \sin\theta \sin\theta_0, \tag{22b}$$

$$c = \cos\theta + \cos\theta_0. \tag{22c}$$

$$\alpha = \frac{2kf \sin\theta_a \sin\theta}{\cos\theta + \cos\theta_0} - \pi/4, \tag{23}$$

$$\beta = 2kf \sin\theta_a \cos\phi_a \left[\frac{\sin\theta_0}{\cos\theta + \cos\theta_0} - \frac{\sin\theta_m}{\cos\theta_0 + \cos\theta_m}\right]. \tag{24}$$

The second term in equation (24) reduces the phase by a constant.

The expansions (20) and (21) are valid for ($k \sin\theta_a$) very large. It should be noted that (20) and (21) display singularities at $\alpha = -\pi/4$, that is, at $\theta_a = 0$, or $\theta = 0$. However, upon examination it can be seen that (20) and (21) contain the first terms of the asymptotic expansions for Bessel functions, that is,

$$\begin{aligned} J_n(x) &\sim \left(\frac{2}{\pi x}\right)^{\frac{1}{2}} \cos\left(x - \frac{n\pi}{2} - \frac{\pi}{4}\right) \\ &\sim \frac{1}{2}(j)^n \left(\frac{2}{\pi x}\right)^{\frac{1}{2}} \left\{(-1)^n \exp\left[j\left(x - \frac{\pi}{4}\right)\right] + \exp\left[-j\left(x - \frac{\pi}{4}\right)\right]\right\}. \end{aligned} \tag{25}$$

By identifying and replacing the asymptotic terms by the actual Bessel functions, the singularities are removed and it might be expected that the approximation for small θ and θ_a would improve and furthermore for large values of θ and θ_a the approximations would be equivalent. There is however, no unique method to introduce such a replacement. The method chosen was dictated by the requirement that for the symmetric case $\theta_0 = 0$, the expressions reduce to the exact expressions previously determined.[4] This necessitates associating a Bessel function of order n, $J_n(x)$ with terms $\cos n\phi_a$ or $\sin n\phi_a$.

On this basis the approximations to the ϕ integration are:

$$\begin{aligned} (E_y)_y = {} & \frac{-\pi(2f)^2 e^{i\beta}}{c(a^2 - b^2\cos^2\phi_a)(1 + \cos\theta_0)} \\ & \cdot [c(J_0(x)\{aE_c(\pi/2) + cE_c(0)\} + J_2(x)\cos 2\phi_a\{aE_c(\pi/2) - cE_c(0)\}) \\ & - j/2b\{cE_c(\pi/2) - aE_c(0)\}\{J_1(x)\cos\phi_a + J_3(x)\cos 3\phi_a\}] \end{aligned} \tag{26}$$

and the cross polarized component

$$\begin{aligned} (E_y)_x = {} & \frac{\pi(2f)^2 e^{i\beta}}{c(a^2 - b^2\cos^2\phi_a)(1 + \cos\theta_0)} \\ & \cdot [jb\{3aE_c(0) + cE_c(\pi/2)\}J_1(x)\sin\phi_a \\ & \qquad + 2c\{aE_c(\pi/2) - cE_c(0)\}J_2(x)\sin 2\phi_a \\ & \qquad + jb\{aE_c(0) - cE_c(\pi/2)\}J_3(x)\sin 3\phi_a] \end{aligned} \tag{27}$$

where

$$x = \frac{4\pi f}{\lambda}\left(\frac{\sin\theta_a \sin\theta}{\cos\theta + \cos\theta_0}\right). \tag{28}$$

It is noted the $(E_y)_y$ is symmetrical with respect to ϕ_a since by interchanging ϕ_a by $-\phi_a$ the expression remains the same. This would be expected since the plane $\phi = 0$ is a plane of antenna symmetry. It is also evident from equation (27) that the cross polarized component is zero in the plane of symmetry and is antisymmetrical with respect to ϕ_a.

The above expressions reduce to those obtained by the method of stationary phase for large values of x. As shown subsequently by numerical integration the latter approximations extend considerably the range of θ and θ_a beyond which the stationary phase approximations are applicable.

For x polarization as outlined in Appendix A the expressions are similar. In particular $(E_x)_x$ is of the same form as $(E_y)_y$ with $E_c(o)$ and $E_c(\pi/2)$ interchanged. The cross polarized component $(E_x)_y$ is of the same form as $-(E_y)_x$ with $E_c(o)$ and $E_c(\pi/2)$ interchanged.

The expressions (26) and (27) are of course approximate. This is evident by considering the special case $\theta_a = 0$, where the values for the fields must be the same independent of ϕ_a. For this special case the ϕ integration is performed in closed form in Appendix B.

For the antenna shown in Fig. 1, with $\theta_0 = 55°$ and $\theta_m = 34.0°$, by assuming the radiation fields of the subreflector at a constant distance are the same in the E and H planes, it is shown that the differences between the exact and the approximate values at $\theta_a = 0$ are 0.049 in the E-plane and -0.053 in the H-plane, both in comparison to one. The subsequent numerical computations indicate that these are the largest errors introduced by the approximation.

1.4 *Numerical Results*

In order to determine the validity of the above approximations, computations of the far field radiation patterns for the open cassegrain antenna have been performed using the subreflector radiation pattern $\mathbf{E}_c$ shown in Fig. 3.

The integration with respect to ϕ, as indicated in equation (13), has been performed, employing Simpsons rule, as a function of θ in the E and H planes, and for observation angles $\theta_a = 0$, 2.5, 5°, 10° and 20°. Estimates for the number of points required for the ϕ integration and the computation time are presented in Appendix C.

The normalized amplitudes obtained from the integration are shown in the upper portions of Figs. 4–8. The normalization was based on the

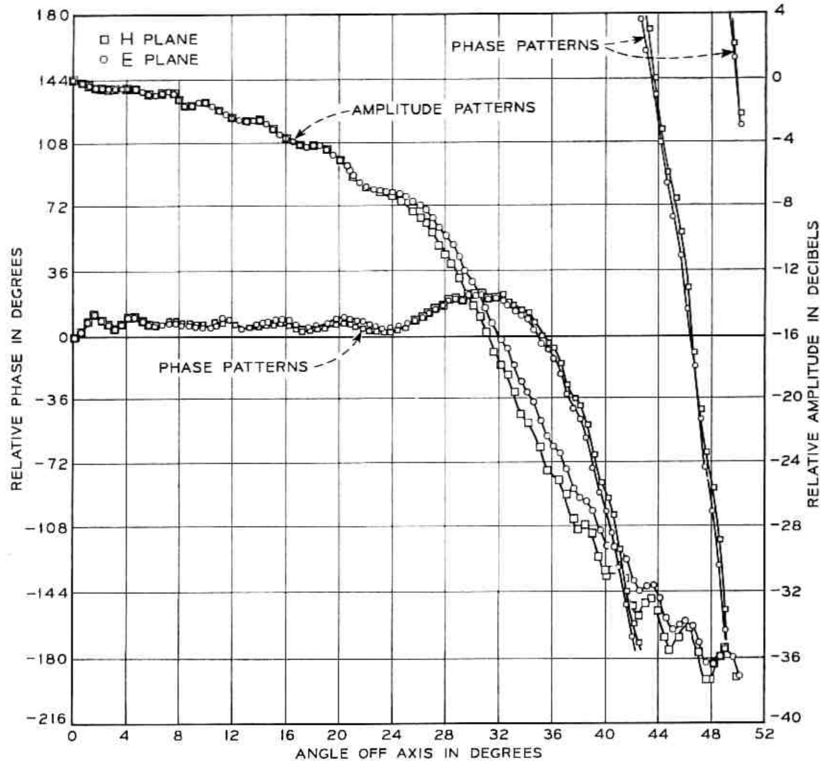

Fig. 3 — Amplitude and phase of subreflector radiation pattern.

stationary phase approximation (20) which shows that in the planes $\phi_a = 0$ or $\pi/2$ the integral (13) is proportional to $E_c(o)$ or $E_c(\pi/2)$ respectively. The normalized values for the integrals E_N shown are

$$E_N = \frac{(E_y)_y/E_c}{[(E_y)_y/E_c]_0} \tag{29}$$

where the subscript zero indicates the value at $\theta = 0$.

Immediately beneath E_N in Figs. 5–8 is shown a plot of the absolute value of the difference between the normalized values obtained by integration and the stationary phase approximation given by equation (20). The third plot in each figure show the corresponding difference using the modified stationary phase approximation (26). As predicted by the method of stationary phase and as shown in Figs. 5–8, the approximations improve as $k \sin \theta_a$ increases. However, where as for small values of θ the stationary phase approximation (20) introduces significant error, the modified approximation (26) becomes increasingly accurate. It should also be noted that with both approximations relative maximum differences occur near zeros, therefore resulting in less significance in the second integration.

Figures 9 and 10 show the amplitudes of the far field radiation as computed by single integration and the modified stationary phase approximation (26). Shown for y polarization are the far field in the plane of antenna symmetry $\phi_a = 0$ and π, and the fields in the plane of asymmetry $\phi_a = \pi/2$ up to 20° off axis. Figure 11 shows the difference in the plane $\phi_a = 0$ between the far field pattern computed by the approximate method and the same pattern computed using double integration. Excluding the vicinity of relative minima, errors were less than 0.2 dB up to off-axis angles of 1°, and on the order of a few hundredths of a dB for larger angles. It should be noted that on axis the difference is zero, since the exact expression for the ϕ integration as given in Appendix B is incorporated in the single integration program.

II. CONCLUSIONS

A method has been developed for the numerical computation of the far field radiation patterns of open cassegrain antennas and related

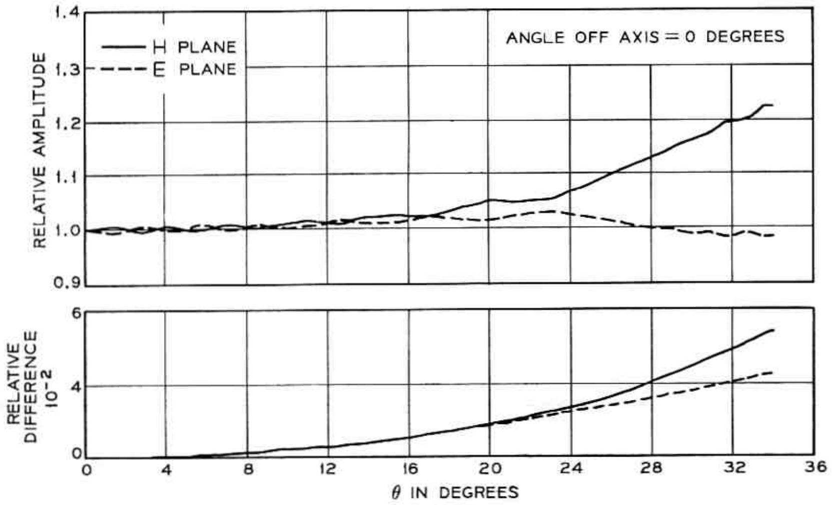

Fig. 4 — Error introduced by modified approximation [equation (26)] on axis.

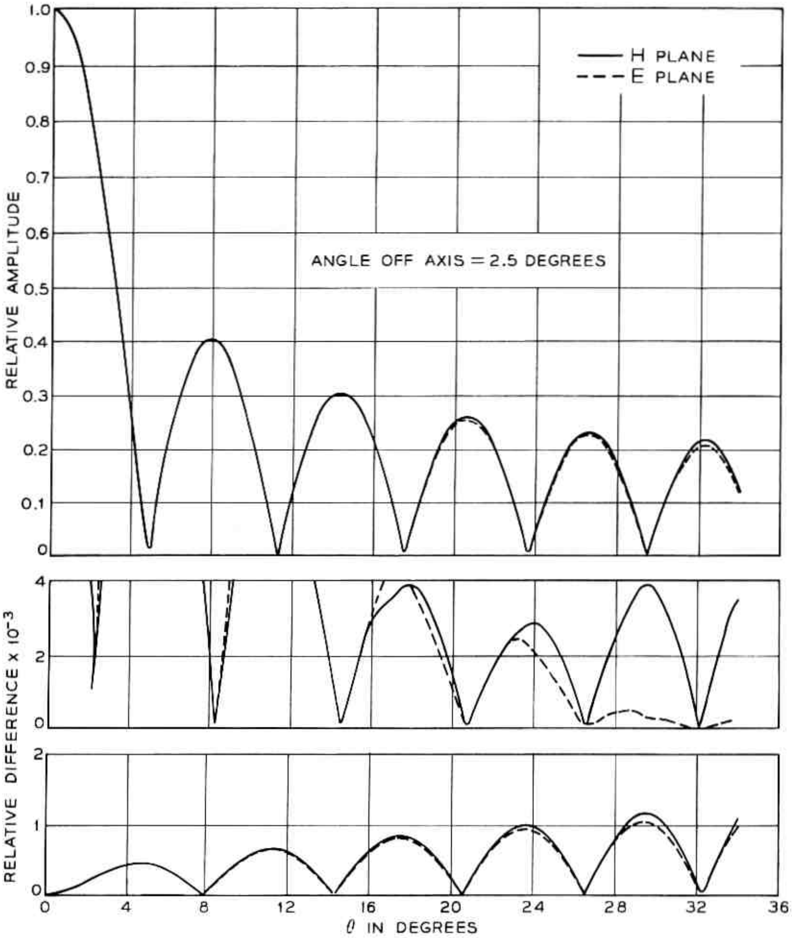

Fig. 5 — Error introduced by stationary phase approximation [equation (20)] and modified approximation [equation (26)].

antenna configurations using a single numerical integration, which reduces considerably the computation time. The method is based on the stationary phase approximation but modified such that for symmetrical paraboloid antennas the approximation reduces to the exact expression which is obtained by direct integration. The errors introduced by stationary phase and modified approximations are examined. It is shown by numerical computations that the error in the far field radiation pattern introduced by the modified stationary phase approximation at angles beyond the main beam is on the order of a few hundreths of a dB.

The number of points needed for numerical integration of oscillatory functions by Simpson's rule is also examined by using specific oscillatory functions. It was determined that at least 6 points per cycle are necessary to obtain four decimal-point accuracy.

The radiation patterns of an open cassegrain antenna with a 128 wavelength aperture diameter are computed up to 20° off axis. In this

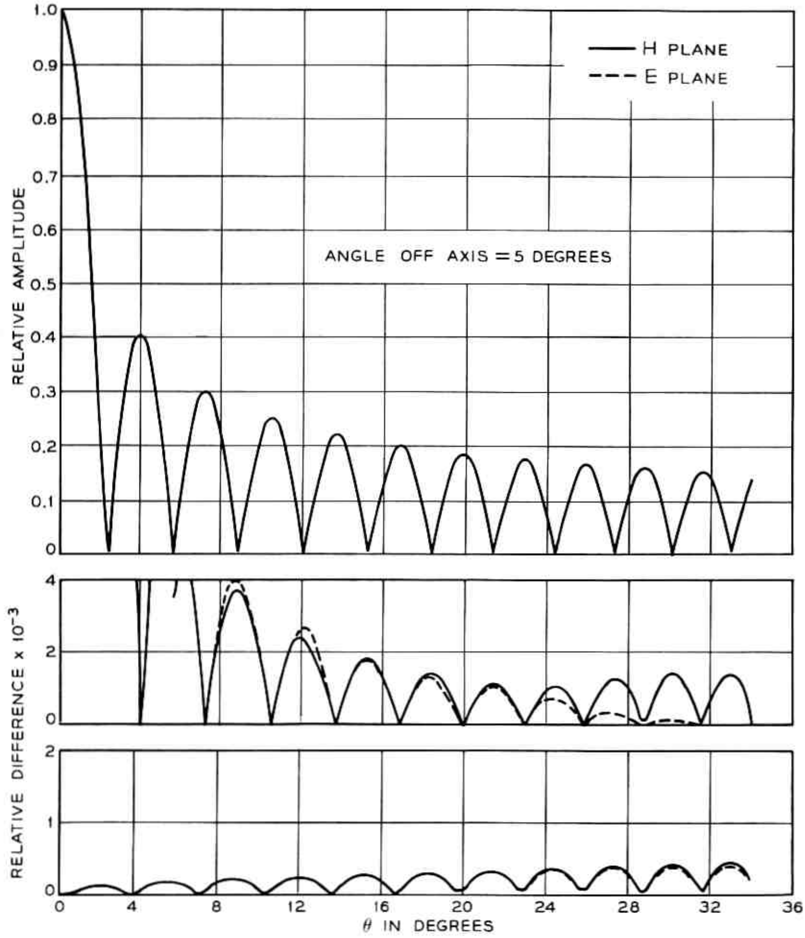

Fig. 6 — Error introduced by stationary phase approximation [equation (20)] and modified approximation [equation (26)]

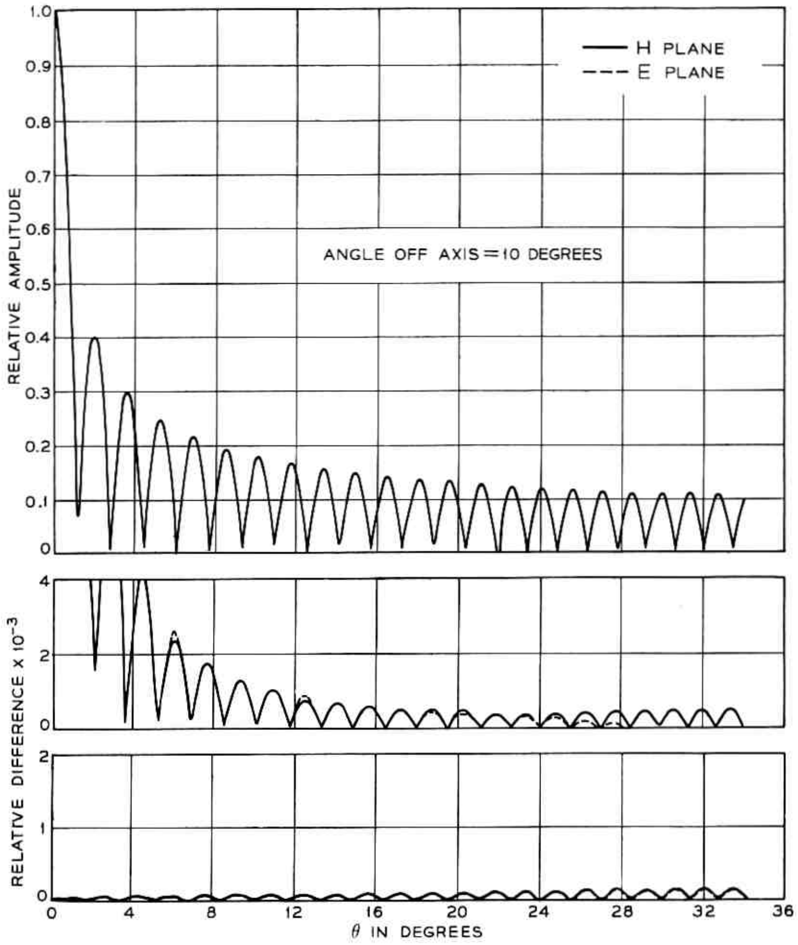

Fig. 7 — Error introduced by stationary phase approximation [equation (20)] and modified approximation [equation (26)].

region the sidelobe levels are −65 dB or −15 dB below the isotropic levels.

Although the stationary phase method has been applied here to the computation of the far field based on the projected aperture field method, the same approach may be used for the computations based on the current distribution method. In particular in the plane of antenna symmetry, the locations of the stationary phase stationary points are the same and therefore the presented approximations are readily extended to this plane.

APPENDIX A

Derivation of the Stationary Phase Approximation

Consider the ϕ integration of equation (1)

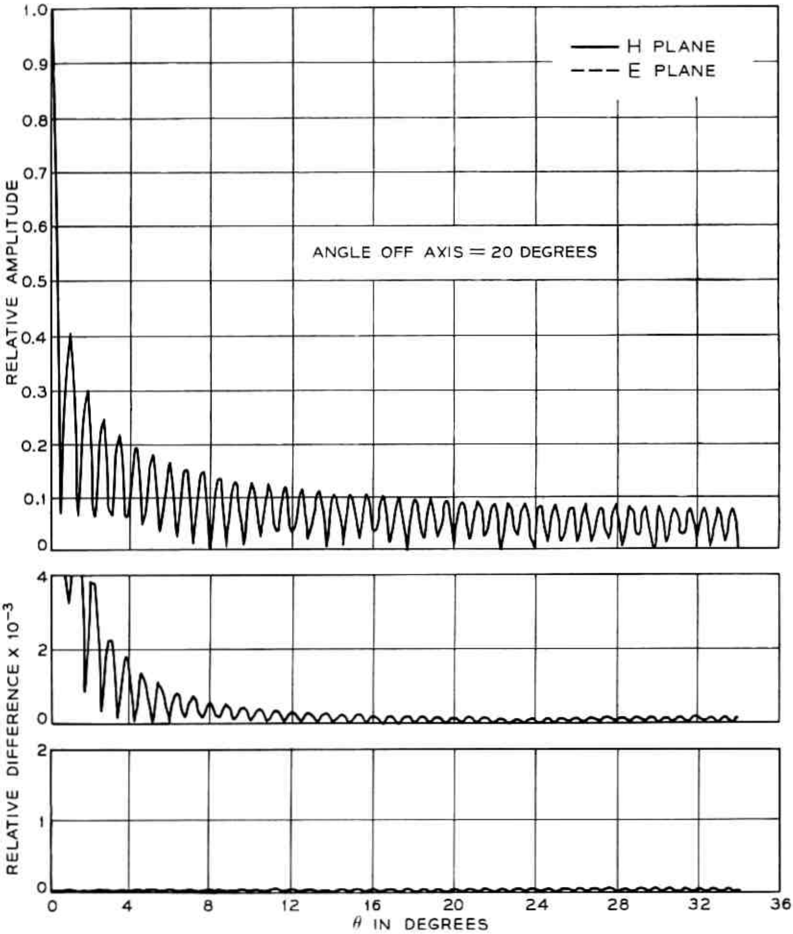

Fig. 8 — Error introduced by stationary phase approximation [equation (20)] and modified approximation [equation (26)].

$$\mathbf{E}(k,\ \theta) = \int_0^{2\pi} \mathbf{E}_r(\theta,\ \phi) \exp\ [jk \sin\ \theta_a(x_p\ \cos \phi_a\ +\ y_p \sin \phi_a)]r^2\ d\phi \qquad (30)$$

with x_p, y_p, $\mathbf{E}_r$ and r given by equations (4) through (7). The stationary points are determined by

$$\frac{\partial x_p}{\partial \phi} \cos \phi_a + \frac{\partial y_p}{\partial \phi} \sin \phi_a = \frac{d\psi}{d\phi} = 0. \qquad (31)$$

This leads to the relation

$$[(1 + \cos\theta\ \cos\ \theta_0)\ \cos\phi - \sin\theta \sin\ \theta_0] \sin \phi_a$$
$$- \sin\phi[\cos\ \theta + \cos\ \theta_0]\ \cos \phi_a = 0. \qquad (32)$$

Equation (32) is a quadratic equation in cos ϕ. Solving this equation

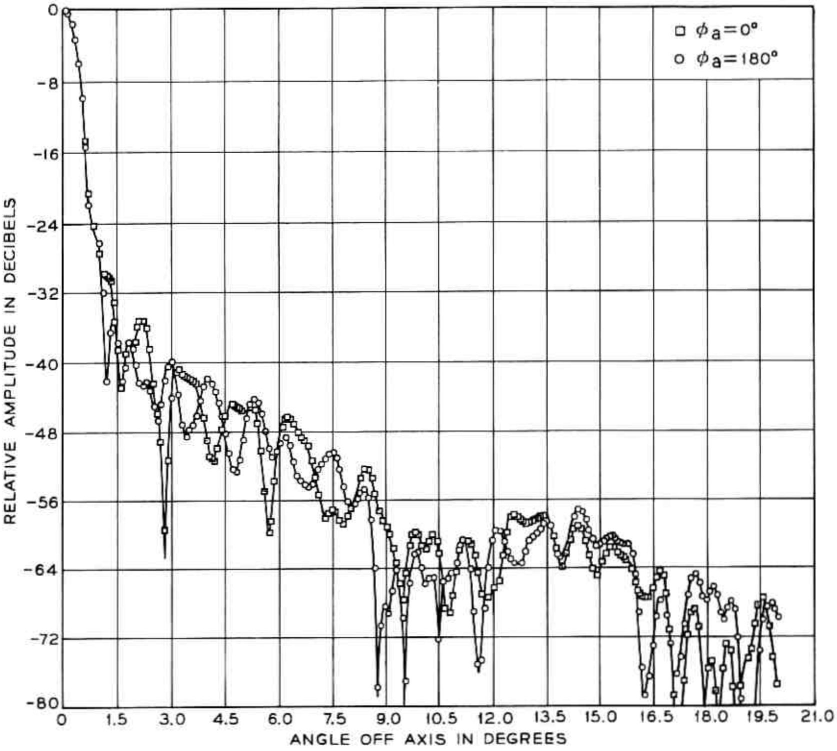

Fig. 9 — Far field radiation pattern in the plane of symmetry.

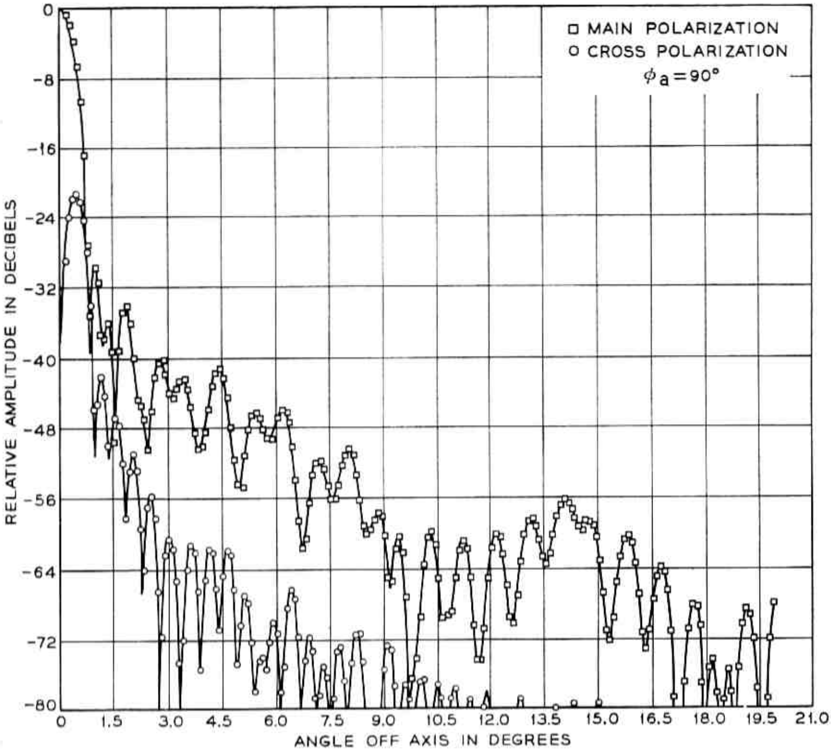

Fig. 10 — Far field radiation pattern in the plane of asymmetry.

yields the following values for the stationary points

$$\cos \phi_{1,2} = \frac{\sin \theta \sin \theta_0 \pm \cos \phi_a(1 + \cos \theta \cos \theta_0)}{1 + \cos \theta \cos \theta_0 \pm \cos \phi_a \sin \theta \sin \theta_0} \tag{33}$$

and from (32)

$$\sin \phi_{1,2} = \pm \frac{\sin \phi_a(\cos \theta + \cos \theta_0)}{1 + \cos \theta \cos \theta_0 \pm \cos \phi_a \sin \theta \sin \theta_0}. \tag{34}$$

Evaluating the phase factor ψ gives

$$\psi(\phi_1) = 2f \frac{(\sin \theta + \cos \phi_a \sin \theta_0)}{\cos \theta + \cos \theta_0} ; \tag{35}$$

$$\psi(\phi_2) = -2f \frac{(\sin \theta - \cos \phi_a \sin \theta_0)}{\cos \theta + \cos \theta_0}. \tag{36}$$

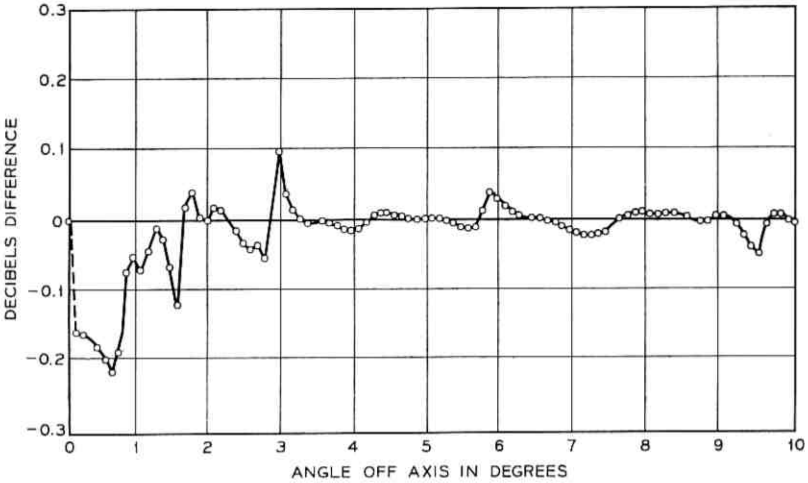

Fig. 11 — Difference in far field pattern as computed by double integration and approximate method (in the plane $\phi_a = 0$).

Note that

$$\psi(\phi_1) - \psi(\phi_2) = 2\left[\frac{2f \sin \theta}{\cos \theta + \cos \theta_0}\right] \tag{37}$$

which is the diameter of the circle given by equation (12). That is the stationary points are antipodes on the projection of the plane of the intersection of the paraboloid surface with the cone θ = constant.

Evaluation of the second derivative leads to

$$\left.\frac{d^2\psi}{d\phi^2}\right|_{1,2} = \mp 2f \sin \theta \frac{[1 + \cos \theta \cos \theta_0 \pm \cos \phi_a \sin \theta \sin \theta_0]^2}{[\cos \theta + \cos \theta_0]^3}. \tag{38}$$

Evaluation of r gives

$$r_{1,2} = 2f \frac{[1 + \cos \theta \cos \theta_0 \pm \sin \theta \sin \theta_0 \cos \phi_a]}{[\cos \theta + \cos \theta_0]^2}. \tag{39}$$

It remains to evaluate $\mathbf{E}_r$ at the stationary points. From equation (18)

$$\mathbf{E}_r = \frac{\mathbf{E}_c}{r} \frac{2f}{(1 + \cos \theta_0)}. \tag{40}$$

From equations (6) and (31)

$$(\mathbf{E}_r)_{1,2} = -\frac{\sin \phi_{1,2}}{\sin \phi_a} \frac{(\cos \theta + \cos \theta_0)}{1 + \cos \theta_0} [1_{xp}(E_{c\theta} \cos \phi_a - E_{c\phi} \sin \phi_a)$$
$$+ 1_{yp}(E_{c\theta} \sin \phi_a + E_{c\phi} \cos \phi_a)]. \tag{41}$$

Substituting the values for x and y polarization in equations (14) through (17) it is evident that the expressions are similar. Therefore the x component for x polarization can be obtained from the y component for y polarization by interchanging $E_c(o)$ with $E(\pi/2)$. Similarly the y component for x polarization can be obtained from the x component for y polarization also by interchanging $E_c(o)$ with $E(\pi/2)$ and changing the sign in front of the resulting expression.

Substituting the appropriate expressions for y polarization into the stationary phase approximation leads to equations (20) and (21).

APPENDIX B

The ϕ Integration on Axis ($\theta_a = 0$)

From equations (6), and (13) for $\theta_a = 0$, the integral for the y component of y polarization with respect to ϕ can be written

$$(E_y)_y = -\frac{(2f)^2}{(1 + \cos\theta_0)} \cdot \int_0^{2\pi} \left\{ \frac{[E_c(\pi/2)c - aE_c(0)]\sin^2\phi}{[a - b\cos\phi]^2} + \frac{E_c(0)}{a - b\cos\phi} \right\} d\phi \tag{42}$$

where a, b, and c are defined by equations (22a), (22b), and (22c). Integrating the first term by parts, reduces the evaluation of equation (42) to a tabulated integral,[5] that is,

$$\int_0^{2\pi} \frac{d\phi}{a - b\cos\phi} = \frac{2\pi}{(a^2 - b^2)^{\frac{1}{2}}} \tag{43}$$

hence

$$(E_{ry})_y = -2(2f)^2 \frac{[E_c(\pi/2) + E_c(0)]}{(1 + \cos\theta_0)^2(1 + \cos\theta)}. \tag{44}$$

The integration for x polarization $(E_{rx})_x$ gives the same result. As a consequence the on-axis gain for an open cassegrain antenna is the same for x and y polarization if the excitation is the same. Based on the approximation (26)

$$(E_y)_y = -\frac{(2f)^2[E_c(\pi/2)(1 + \cos\theta\cos\theta_0) + E_c(0)(\cos\theta + \cos\theta_0)]}{(1 + \cos\theta_0)[(1 + \cos\theta\cos\theta_0)^2 - (\sin\theta\sin\theta_0\cos\phi_a)^2]}. \tag{45}$$

To estimate the relative error it is assumed that $E_c(o) = E_c(\pi/2)$. This gives in the plane $\phi_a = 0$ or π

$$(\Delta E_y)_y = 1 - \left[\frac{(1 + \cos\theta)(1 + \cos\theta_0)}{2[\cos\theta + \cos\theta_0]}\right]^2.$$

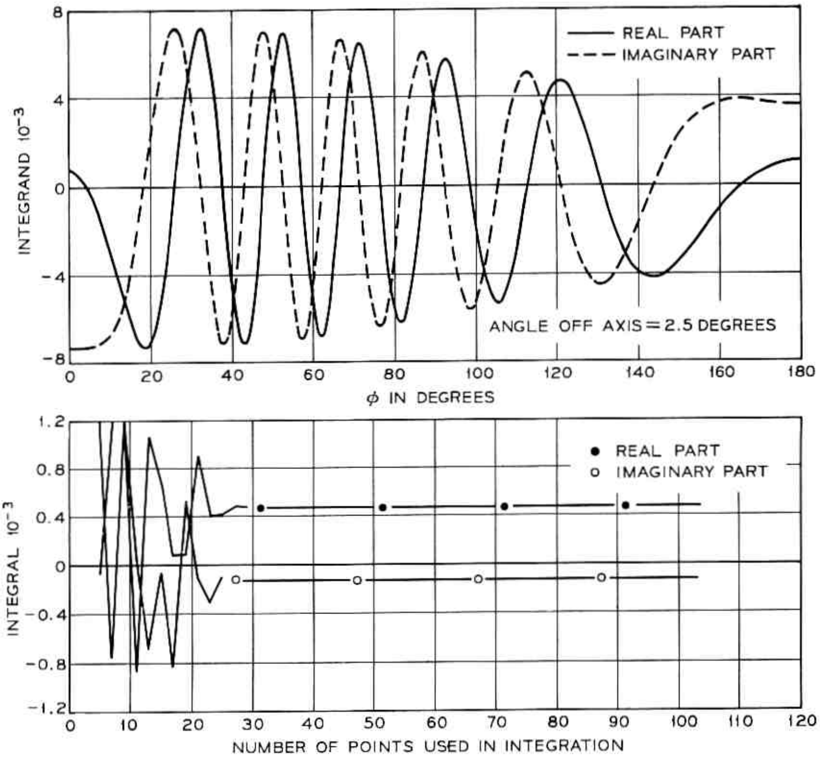

Fig. 12 — Numerical integration of oscillatory function.

Similarly in the plane $\phi_a = \pi/2$

$$(\Delta E_y)_y = 1 - \left[\frac{(1 + \cos\theta)(1 + \cos\theta_0)}{2(1 + \cos\theta\cos\theta_0)}\right]^2. \tag{46}$$

Equations (48) and (49) can be combined yielding

$$(\Delta E_y)_y = 1 - \frac{1}{[1 \pm (\tan\theta/2 \tan\theta_0/2)^2]^2} \tag{47}$$

where the minus sign corresponds to $\phi = 0$ or π and the plus sign is for $\phi_a = \pi/2$.

APPENDIX C

Computation Time Estimates

The attempt to find a suitable approximation to be used for the evaluation of far field radiation patterns was motivated, in fact necessitated, by the excessive computation time required for a double integration procedure.

Both the ϕ integral [equation (13)] and the θ integral [equation (19)] have oscillatory integrands, with the ϕ integrand, for the maximum value of θ, having approximately double the number of oscillations as the θ integrand. The maximum number of full cycles in the θ integrand is equal approximately to

$$\frac{D}{2\lambda} \sin \theta_a$$

where D is the antenna aperture diameter.

Figures 12 through 15 show the ϕ integrand for $\theta = 34°$, and for various values of observation angle θ_a . Below these figures is shown

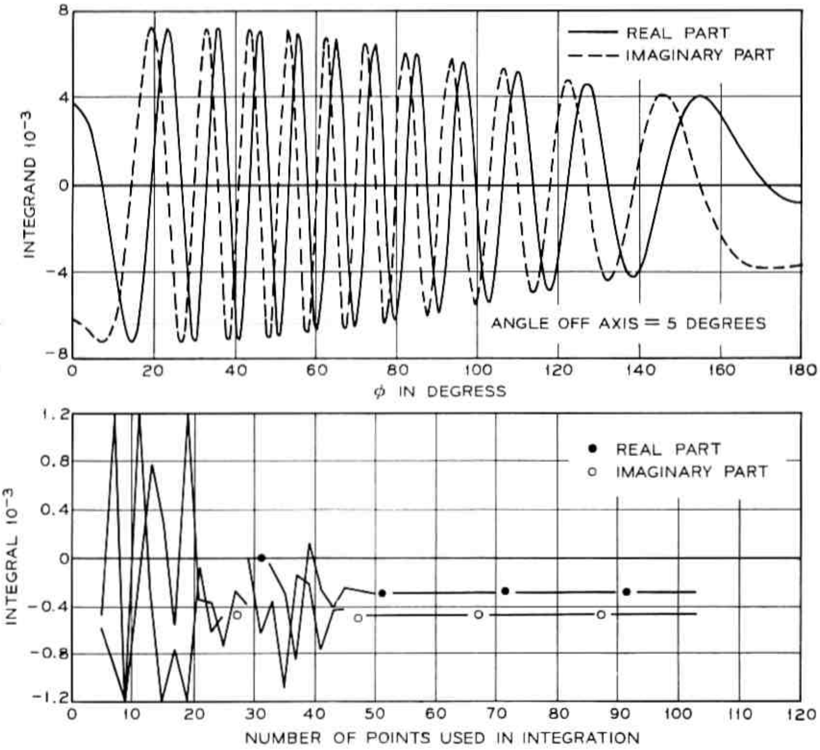

Fig. 13 — Numerical integration of oscillatory function.

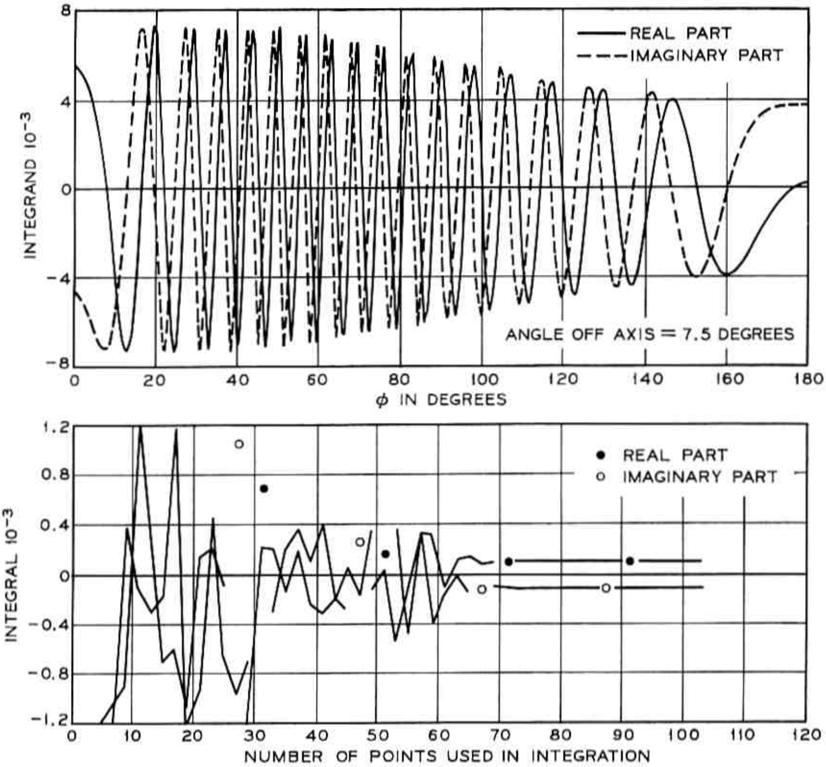

Fig. 14 — Numerical integration of oscillatory function.

the result of the ϕ integration using Simpson's rule with increasing number of points. We have found by numerical integration that for the integrands discussed above, a minimum number of 6 points per cycle is necessary to provide reasonable accuracy.

For a discussion of numerical integration of oscillatory integrands the reader is referred to Ref. 6.

An estimate for the processor time required to calculate both components of the far field pattern in both the E and H planes (4 patterns) out to an observation angle of θ_a degrees by the method of double integration is given by:

$$k_2 \frac{\theta_a}{\Delta \theta_a} N^2 \frac{\theta_a^2}{3}$$

where

k_2 is the fundamental program loop execution time (1.13×10^{-4} min., on the G. E. 635).

N is the number of integrand evaluations per cycle required by the integration procedure.

$\Delta\theta_a$ is the observation angular increment at which results are to be calculated.

An estimate for the same calculation by the approximate method derived herein is:

$$k_1 \frac{\theta_a}{\Delta\theta_a} N \frac{\theta_a}{2}$$

where

$$k_1 = 0.75 \times 10^{-3} \text{ min.}$$

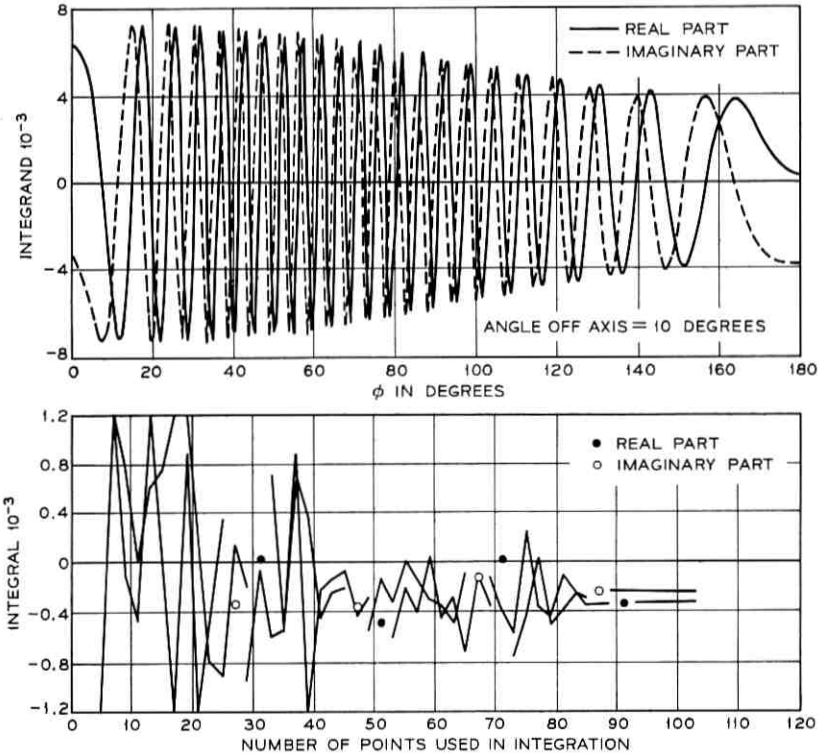

Fig. 15 — Numerical integration of oscillatory function.

TABLE I—EXECUTION TIMES

θ_a	5°	10°	20°
Double Int.	4.7 min.	38 min.	300 min.
Appr. Method	0.9 min.	3.7 min.	15 min.

Table I shows comparable execution times for various observation angle extremes, assuming $\Delta\theta_a = 0.1°$, $N = 10$.

REFERENCES

1. Cook, J. S., Elam, E. M., and Zucker, H., "The Open Cassegrain Antenna," B.S.T.J., *44*, No. 7 (September 1965), pp. 1255–1300.
2. Silver, S., *Microwave Antenna Theory and Design*, New York: McGraw-Hill, 1949.
3. Erdelyi, A., *Asymptotic Expansions*, New York: Dover Publ., 1950, pp. 51–57.
4. Zucker, H., and Ierley, W. H., "Computer Aided Analysis of Cassegrain Antennas," B.S.T.J., *47*, No. 6 (July-August 1968), pp. 897–932.
5. Dwight, H. B., *Tables of Integrals and Other Matehmatical Data*, New York: The MacMillan Company, 1957, p. 199.
6. Allen, C. C., "Numerical Integration Methods for Antenna Pattern Calculations," IRE Transactions on Antennas and Propagation, Special Supplement, Symposium on Electromagnetic Theory, December 1959, pp. 387–401.

A Decomposition of a Transfer Function Minimizing Distortion and Inband Losses

By ERNST LUEDER

(Manuscript received July 3, 1969)

A rational transfer function to be realized by an RC*-active network is usually decomposed into functions of at most second degree. We present a method for achieving this—which maximizes the dynamic range of the whole network while minimizing inband losses. The method is based on the "bottleneck problem."*

I. INTRODUCTION

A given rational transfer-function $T(s)$ of a passive network, which is real for real s, is to be realized by an inductorless two-port. This is usually done by breaking down $T(s)$ into functions $T_i(s)$ of the first or second degree in s. All functions of the first and those of the second degree with poles on the negative real axis are realized by passive RC-networks with buffer amplifiers between the different stages. Those of second degree but with poles not on the negative real axis are realized by RC-active networks containing amplifiers. We deal at first with the second group of functions. The extension to the general case follows easily. The voltage swing at the input of the different stages with functions $T_i(s)$ is often tightly limited by the threshold above which overdriving of the amplifiers (that is, distortion) occurs. A further result in many cases is high inband loss of the overall filter which cannot be overcome by amplification because of both distortion and a too low signal/noise ratio.

Our task is to find a method of factoring $T(s)$ into the different functions $T_i(s)$ such that the allowable voltage swing at the input is as high as possible without creating distortion and the inband losses as low as possible. We confine ourselves first to transfer functions

$$T(s) = \frac{V_2}{V_1} = K\,\frac{s^n + a_{n-1}s^{n-1} + \cdots a_1 s + a_0}{s^m + b_{m-1}s^{m-1} + \cdots b_1 s + b_0} \quad \text{with} \quad n \leqq m \qquad (1)$$

which, as mentioned, have no poles on the negative real axis. Thus m is even. If we count the zeros of $T(s)$ including those at infinity, then $T(s)$ has also m zeros. Let the number of zeros including the origin and infinity on the real axis be r_z. Then r_z is even, since m and the number of zeros not on the real axis are even. We are mainly dealing with transfer-functions, which belong to the class of networks having a passband. In the case of two ports without a passband a slightly different approach will be necessary. The functions $T_j(s)$ of second order have the general form.

$$T_j(s) = K_j \frac{s^2 + \frac{\omega_z}{q_z} s + \omega_z^2}{s^2 + \frac{\omega_p}{q_p} s + \omega_p^2}. \tag{2}$$

After normalization by

$$\frac{s}{\omega_p} = p \tag{3a}$$

with

$$p = \sigma + j\Omega \tag{3b}$$

we get from equation (2)

$$T_j(\omega_p \cdot p) = \overline{T_j}(p) = K_j \frac{p^2 + \frac{c}{q_z} p + c^2}{p^2 + \frac{1}{q_p} p + 1} \quad \text{with} \quad c = \frac{\omega_z}{\omega_p}. \tag{3c}$$

To meet the aforementioned requirements as to distortion and inband losses, we have these possibilities:

(*i*) There are in general a large number of ways of finding the different functions $T_j(s)$, because there are many methods of choosing pairs of poles and zeros in forming $T_j(s)$. In Sections II and III we discuss the best choice for our task.

(*ii*) In RC-active two-ports there is some freedom in evaluating the constant K_j in equation (2).

(*iii*) The functions $T_j(s)$ and their realizations once found, there are many possibilities for the sequence in cascading the different stages. In Section IV, we discuss some guidelines for this point as well as for *ii*.

II. A CRITERION FOR THE GOODNESS OF AN ASSIGNMENT OF POLES AND ZEROS

We need a criterion which tells us when a chosen assignment meets the requirement of voltage swing and inband losses. For this reason we are looking for the shape of the function $| T_j(j\Omega) |^2$, which from equation (3c) has the normalized form,

$$\frac{| T_j(j\Omega) |^2}{K_j^2} = F_j(x) = \frac{(c^2 - x)^2 + \frac{c^2}{q_z^2} x}{(1 - x)^2 + \frac{x}{q_p^2}} \quad \text{with} \quad x = \Omega^2. \tag{4}$$

We are interested in the shape of the function in equation (4) for real values of Ω, that is, for real nonnegative values of x. The extrema of $F_j(x)$ occur, as can be easily calculated, at the values

$$x_{a,b} = \frac{c^4 - 1 \pm \left[(c^4-1)^2 - c^2 \left\{ 2(c^2-1) + \frac{1}{q_z^2} - \frac{c^2}{q_p^2} \right\} \left\{ 2(c^2-1) + \frac{1}{q_p^2} - \frac{c^2}{q_z^2} \right\} \right]^{\frac{1}{2}}}{2(c^2-1) + \frac{1}{q_p^2} - \frac{c^2}{q_z^2}} \tag{5a}$$

and

$$x = \infty \tag{5b}$$

with

$$F_j(\infty) = 1. \tag{5c}$$

If $x_{a,b}$ are real and nonnegative, extrema occur with the ordinate values $F_j(x_{a,b})$ from equations (4) and (5a) as follows:

$$F_j(x_{a,b}) = \frac{(c^2 - x_{a,b})^2 + \frac{c^2}{q_z^2} x_{a,b}}{(1 - x_{a,b})^2 + \frac{x_{a,b}}{q_p^2}}. \tag{6}$$

For the moment we need not know if $F_j(x_a)$ or $F_j(x_b)$ is a maximum or a minimum. It is sufficient to note that besides the extremum at $x = \infty$, at most two other extrema of $F_j(x)$ can occur.* Let the maximum be at x_m and the minimum at x_0. We note from equation (4):

$$F_j(0) = c^4 \tag{7}$$

with $F_j(0) \lesseqgtr F_j(\infty) = 1$ depending on c. An example for a function

* For Ω as abscissa, a further extremum can occur at $\Omega = 0$. For $c = 0$, one of the extrema in equation (5a) lies at $x = 0$.

$F_j(x)$ with $F_j(x_m) > c^4$ and $F_j(0) > F_j(\infty)$ is shown in Fig. 1a, while Fig. 1b represents a function $F_j(x)$ with $F_j(x_m) < 1$ and $F_j(0) < F_j(\infty)$.

The passband of the whole filter may be in $x \in [x_1, x_2]$ with $x_1 \geqq 0$ and $x_2 > x_1$, as shown in Fig. 1. Let us first assume that a peak $F_{\max}$ of $F_j(x)$ may occur at x_m with $x_m \in [x_1, x_2]$. Considering only frequencies in the passband, we are stating that overdrive of the amplifiers will first occur at the maximal value $F_{\max}$ of $F_j(x)$, when the spectrum of the input signal is assumed constant at least in $x \in [x_1, x_2]$. To prevent overdrive of the amplifiers $F_{\max}$ should exceed the "mean" values in the passband as little as possible. On the other hand, we have to regard the minimum value $F_{\min}$ of $F_j(x)$ in $x \in [x_1, x_2]$. $F_{\min}$ gives us the strongest attenuation of the signal, which we have to overcome by amplification. When this is not possible because of overdrive or a too low signal/noise ratio, then a low $F_{\min}$ yields high inband losses. For this reason, $F_{\min}$ should be as close to the "mean" values in the passband as possible. It would seem at first sight that both requirements can be met, namely that $F_{\max}$ and $F_{\min}$ be as close to the mean values in the passband as possible, if we look for a transfer function $T_j(s)$ such that $d_j = F_{\max} - F_{\min}$ be minimized. But this criterion does not always cover our requirements as can be seen in Fig. 2. Both functions $F_j(x)$ have the same value $d_j = F_{\max} - F_{\min}$. Their practical behavior however is very different. The two port with the transfer function of Fig. 2a almost entirely cuts off the frequencies in the passband in the neighborhood of x_1 and x_2 and we would need a very high gain to bring them up, which is not true in the case of Fig. 2b. To avoid this error, we redefine the d_j-value by the ratio

$$d_j = \frac{F_{\max}}{F_{\min}} \tag{8}$$

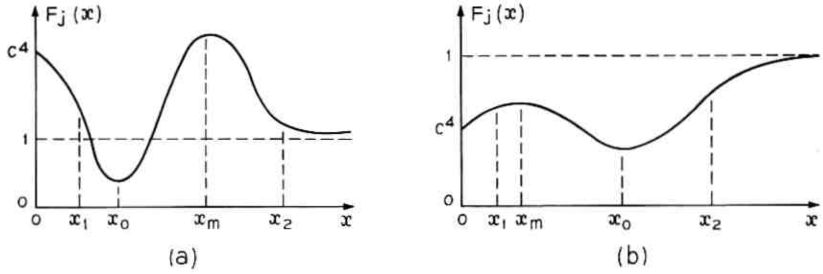

Fig. 1 — Functions $F_j(x)$ with (a) $F_j(x_m) > c^4$ and $F_j(0) > F_j(\infty)$; and (b) $F_j(x_m) < 1$ and $F_j(0) < F_j(\infty)$.

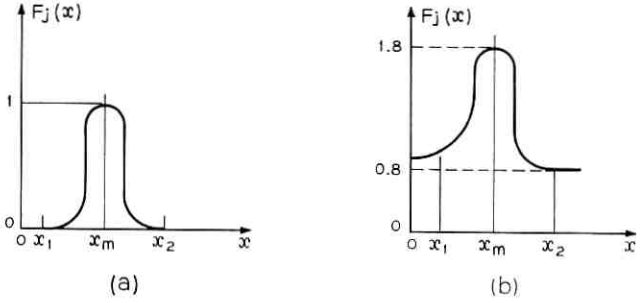

Fig. 2 — Two-port almost entirely cuts off the frequencies in passband in neighborhood of x_1 and x_2 in (a) but not in (b).

with $F_{\max}$ and $F_{\min}$ for $x \in [x_1, x_2]$ and require it to be as close to unity as possible. In order to compress the range of values we can let

$$d_i = \log \frac{F_{\max}}{F_{\min}} \tag{9}$$

where d_i is obviously a positive number. The values d_i in equation (9) should be as close as possible to zero, that is, as small as possible.

Until now, we have assumed $x_m \in [x_1, x_2]$. If x_m is outside the passband we have to argue in a slightly different manner. Now we could have two points of a peak value, one at $x = x_m$ and one at the boundary $x = \infty$. We look for that point among these two with the highest $F_{\max}$ value and we denote this point by x_m'. Let us assume that the amplitude of a signal from a neighboring channel occurring at x_m' is so high that the amplifiers are overdriven. This will change the operating points of the amplifiers resulting in an impaired transmission of signals in the passband. This can sometimes be avoided by inserting the stage under consideration at such a place in the cascade, that the amplitude of the input signal at x_m' is not too high. This, however restricts the freedom of choosing the cascade sequence. We therefore request that the maximum $F_{\max}$ at x_m' be as close to the "mean" values in the passband as possible, even if $x_m' \notin [x_1, x_2]$. Thus we look for $F_{\max}$ for $x \in [0, \infty]$. Minimum values of $F_i(x)$ for $x \notin [x_1, x_2]$ are of no importance since we don't have to amplify those values outside the passband. For these reasons the d_i-values in equations (8) and (9) are replaced by

$$d_i = \frac{F_{\max}}{F_{\min}} \tag{10}$$

or

$$d_j = \log \frac{F_{\max}}{F_{\min}} \tag{11}$$

with

$$F_{\max} \quad \text{for} \quad x \in [0, \infty] \tag{12a}$$

and

$$F_{\min} \quad \text{for} \quad x \in [x_1, x_2]. \tag{12b}$$

Thus each assignment of a pair of poles to a pair of zeros, that is, each function $T_j(s)$, is described by a number d_j, defined by one of the equations (10) or (11), which in the case of (11) should be as small as possible for all $j = 1, 2, \cdots, m/2$. In other words

$$\max \{d_j\}_{j=1,2,\cdots m/2} \rightarrow \min \tag{13}$$

for the d_j in equation (11).

M. Segal pointed out that this problem is an assignment problem of the bottleneck type.[1] O. Gross has given a solution which is convenient also for large numbers of poles and zeros.[2] This algorithm was adopted by S. Halfin to find an optimal pairing and an optimal nested solution.[3] He also presented a method of listing all equivalent solutions. A further solution suitable for smaller numbers of poles and zeros (for example, $\leqq 20$) has been described in Ref. 4, where also all equivalent solutions may be found. The next paragraph shows how the various types of transfer functions should be treated as to this assignment problem.

III. THE PAIRING OF POLES AND ZEROS

We have to check all possibilities of assigning a pair of poles to a pair of zeros. For simplicity, we first assume all zeros including the origin and infinity to lie off the real axis. We consider the case of zeros on the real axis later. Remember that all poles of $\bar{T}(p)$* lie in the interior of the left half plane of p but not on the negative real axis. If we assign a pole at $p = p_\gamma$ to a zero at $p = z_\mu$ as shown in Fig. 3, we have to assign, as is well known, the conjugate complex pole $p = p_\nu^*$ to the conjugate complex zero $p = z_\mu^*$, thus forming the second order function $T_j(p)$ in equation (3c). Therefore we need only regard in Fig. 3 the assignments of the poles p_1, $p_2 \cdots p_{m/2}$ to the zeros z_1,

* $\bar{T}(p)$ is the function $T(s)$ of equation (1) normalized by equation (3a).

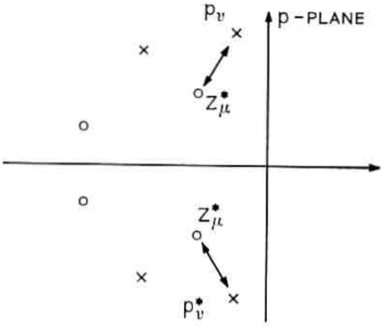

Fig. 3 — An example for a pole-zero assignment.

$z_2 \cdots z_{m/2}$ where all values have positive imaginary part. To each assignment (p_ν, z_μ), $\nu, \mu = 1, 2 \cdots m/2$, belongs a number defined in equations (10) or (11), which will here be denoted by $d_{\nu,\mu}$. All possible assignments (p_ν, z_μ), with their associated $d_{\nu,\mu}$ are listed in Table I. Obviously their total number is $(m/2)^2$. The solution of the assignment problem[3,4] starts with Table I.

Now we have to regard the case of zeros lying on the real axis including the origin and infinity, while as before, all poles are assumed to be complex. The problem now is to assign a pair of zeros to each conjugate complex pair of poles. The zeros on the real axis with number r_z, where r_z is even, can be arranged pairwise in many ways. For example if we have the four distinct zeros 1, 2, 3, 4 in Fig. 4, where one of them may be at infinity, then we have these three possibilities to arrange them in pairs: (1, 2)(3, 4); (1, 3)(2, 4); (1, 4)(2, 3). In general, if we have r_z different zeros on the real axis with r_z even, then we have

$$a_0 = (r_z - 1)(r_z - 3)(r_z - 5) \cdots 5 \cdot 3 \cdot 1 \tag{14}$$

TABLE I

	z_1	z_2	-------	z_h
p_1	d_{11}	d_{12}	-------	d_{1h}
p_2	d_{21}	d_{22}	-------	d_{2h}
⋮	⋮	⋮		⋮
p_h	d_{h1}	d_{h2}	-------	d_{hh}

$h = \frac{m}{2}$

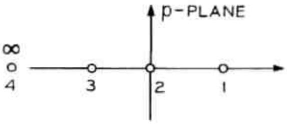

Fig. 4 — We have these possibilities to arrange zeros in pairs: (1, 2) (3, 4); (1, 3) (2, 4); (1, 4) (2, 3).

different possibilities of a pairwise arrangement. In the case of multiple zeros on the real axis we have less than a_0 possibilities of pairwise assignments. For the zero arrangement shown in Fig. 5, we have only the following two possibilities: (1, 2)(3, 4) and (1, 3)(2, 4).

Now we have to complete Table I by regarding also zeros on the real axis. We pick out one pairwise combination of the zeros, for example, the arrangement (1, 2)(3, 4) of Fig. 5, and add each pair of these zeros to the pairs of conjugate complex zeros in Table I, where we treat them like the other zeros. From that, one solution of the assignment problem will be found. However, we did not yet regard all possible assignments. We have to replace the pairwise combinations of the real zeros by another possible combination, for example, by (1, 3)(2, 4) in Fig. 5. This provides a second table like Table I from which a further solution can be found, and so on. The solution with the least maximum value of the $d_{\nu,\mu}$ is the solution to the whole problem.

Finally we have to deal with the case in which poles are also located on the real axis. We consider first the simplest and most important case of only one pole and r_z different zeros on the real axis. The pole on the real axis can be assigned to one of the zeros on the real axis. There are r_z ways of doing this.

The $r_z - 1$ zeros left can be pairwise arranged according to equation (14) in $a_1 = (r_z - 2)(r_z - 4) \cdots 5 \cdot 3 \cdot 1$ ways where each pair of these a_1 sets is handled like a conjugate complex pair of zeros. Thus we get $r_z a_1$ sets of zeros to be assigned to the poles, which means $r_z a_1$ different tables of the kind of Table I. The solution with the least maximum value of the $d_{\nu,\mu}$ is the solution of the whole problem. The case where the r_z zeros on the real axis are not different is handled in a

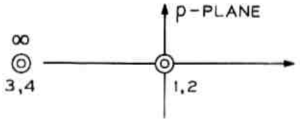

Fig. 5 — We have here two possibilities: (1, 2) (3, 4) and (1, 3) (2, 4).

similar way. The more general case where $r_p > 1$ poles, with r_p odd or even, are on the real axis is usually not so important and is therefore only described briefly.

If r_p is odd, we assign one pole to one zero on the real axis, which can be done in several ways. The zeros and poles left on the real axis are arranged in pairs, which once more can be done in several ways. Then we form for each pairwise arranged set of poles and zeros a table like Table I from which we get the solution. If r_p is even, we start with the pairwise arrangement of poles and zeros and proceed as above.

Incidentally more first order functions of the kind described could be formed with negative real poles and zeros. However this is undesirable because it would require more buffer amplifiers.

In some cases the assignment of one particular pole to one particular zero is prescribed by the realization procedure. Then we simply assign them and eliminate this pole-zero pair from consideration. On the other hand, if a particular assignment of one pole to one zero is forbidden, we provide it a high $d_{\nu,\mu}$-value.

IV. THE CHOICE OF THE FACTORS K_i AND OF THE SEQUENCE OF CASCADING

In most realization procedures the factor K_i in equation (2) can be chosen within certain limits by evaluating the gain of the amplifiers. We describe here one way to do that. The choice of K_i should be made in such a way that the "gain" in the passband of all stages is as close together as possible. This prevents one stage from having a much lower gain than the others which results in a lower signal/noise level in that stage. This is only a short hint because work on this point is continuing. Within the building block concept of G. S. Moschytz[5] there is enough freedom to choose the appropriate K_i. In the case of the low pass, bandpass and the high pass, this has been shown in Ref. 6.

Some guidelines for the sequence in cascading the different stages follow where we use observations by Moschytz. The first stage should be a low pass or a bandpass, thus keeping higher frequencies from the amplifiers and avoiding slew rate problems. Also the last stage should be a low or a bandpass for the purpose of suppressing noise created by the amplifiers themselves. Where a peak in the frequency response of a stage cannot be avoided, this stage should be preceded by stages delivering attenuation at the peak point. In special cases different considerations for cascading could be necessary.

If the assignment problem gives us several solutions, then the one best meeting these guidelines for cascading should be chosen. If the two port has no passband, then x_1 and x_2 should define the frequency range in which the network has to operate. The choice of the values K_i should also be made in such a way that there is no great difference in gain in the frequency range $x \in [x_1, x_2]$.

V. AN EXAMPLE

Given a transfer function $T(s)$ for a Single-Side-Band (SSB) filter with the following poles $p_1 \cdots p_5$ and the zeros $z_1 \cdots z_5$:

$$\begin{aligned}
z_1 &= \pm j0.32233523 \cdot 10^6, \\
z_2 &= \pm j0.36742346 \cdot 10^6, \\
z_3 &= -0.31480 \cdot 10^5 \pm j0.3132295 \cdot 10^6, \\
z_4 &= 0 \text{ (twice)}, \\
z_5 &= \infty \text{ (twice)}, \\
p_1 &= -0.276100 \cdot 10^5 \pm j0.2961048 \cdot 10^6, \\
p_2 &= -0.31480 \cdot 10^5 \pm j0.3132295 \cdot 10^6, \\
p_3 &= -0.8706 \cdot 10^4 \pm j0.314697 \cdot 10^6, \\
p_4 &= -0.93340 \cdot 10^5 \pm j0.18670202 \cdot 10^6, \\
p_5 &= -0.25280 \cdot 10^5 \pm j0.62888167 \cdot 10^5.
\end{aligned}$$

The zero z_3 and the pole p_2 are a phantom pair which have been introduced for the realization procedure.

The four zeros on the real axis z_4 and z_5 can be pairwise arranged in two ways: (z_4, z_4), (z_5, z_5) or (z_4, z_5), (z_4, z_5). Let the pairs of the first arrangement be denoted by $z_{4,1} = (z_4, z_4)$ and $z_{5,1} = (z_5, z_5)$ and the pairs of the second assignment by $z_{4,2} = (z_4, z_5)$ and $z_{5,2} = (z_4, z_5)$.

Now we calculate the d_j-values of equation (11) from which we obtain Table II, corresponding to Table I. Then we obtain with the help of Ref. 4, the following pairings*

$$(z_5, p_4); \quad (z_{41}, p_5); \quad (z_{51}, p_4); \quad (z_2, p_2); \quad (z_1, p_1) \cdot \tag{15}$$

In the realization procedures there is usually a constraint such that one particular assignment of a pole to a zero is prescribed. In the realization by building blocks,[5] pole p_3 has to be assigned to zero z_3. The rest of the assignments are free. In this case the solution is

$$(z_3, p_3); \quad (z_{41}, p_5); \quad (z_{51}, p_4); \quad (z_2, p_2); \quad (z_1, p_1). \tag{16}$$

* The pairings in this example have been calculated by a procedure described in Ref. 4, which is suitable for small numbers of poles and zeros.

TABLE II

	z_1	z_2	z_3	z_{41}	z_{51}	z_{42}	z_{52}
p_1	0.15	0.57	0.25	3.7	1.36	2.47	2.47
p_2	0.18	0.3	0	3.8	1.49	2.69	2.69
p_3	0.18	1.25	1.1	4.9	2.5	3.77	3.77
p_4	1.87	1.4	1.6	1.9	0.56	0.87	0.87
p_5	4.1	3.3	3.7	0.4	2.64	1.5	1.5

Solution (16) has the five transfer functions T_j , $j = 1, 2, \cdots 5$ listed and drawn with full lines in Figs. 6a through 6e. There the factors K_j , $j = 1, 2, 3, 4, 5$ in equation (3b) have been chosen in accordance with Section IV such that the whole filter has an attenuation of 0 dB at 30 kHz. The sequence in cascading the different stages is as described in Section IV, using the denotations for the different stages in Figs. 6a through 6e

$$T_4(s)\,T_1(s)\,T_2(s)\,T_3(s)\,T_5(s).$$

T_4 , T_1 and T_2 deliver the attenuation for the peak of T_3 . Since it is not possible to have a low pass as both the first and the last sections, we chose the low pass to be the first because in this case, noise coming in at the input terminals was stronger than noise created by the amplifiers. The magnitude of the transfer function of the whole filter can be seen in Fig. 7.

We wish to compare the solution described above with an earlier solution realizing the same transfer function in a different way. In the earlier version the phantom pair $(z_3$, $p_2)$ had the following location

$$z_3 = -0.12\cdot 10^6 \pm j0.28\cdot 10^6; \quad p_2 = -0.12\cdot 10^6 \pm j0.28\cdot 10^6.$$

In the new realization we shifted this phantom pair closer to the pole p_3 and thus were able to decrease the peak in the FEN* section as shown in Fig. 6c.

The earlier realization had the following assignment of poles and zeros; $(z_1p_2)(z_2p_1)(z_3p_3)(z_4p_4)(z_5p_5)$, which leads to the following transfer functions of Ref. 7. T'_j , $j = 1, 2. \cdots 5$, are listed in Figs. 6a through 6e. The magnitude in dB of these functions is plotted as dotted lines in Figs. 6a through 6e. The passband lies between 12 and 46 kHz. The functions T_1 , T_5 and especially T_4 of the new version have obviously less attenuation in the passband, while the functions T_2 and especially T_3 have a lower peak than in the earlier version. The whole filter has

* Frequency Emphasizing Network.

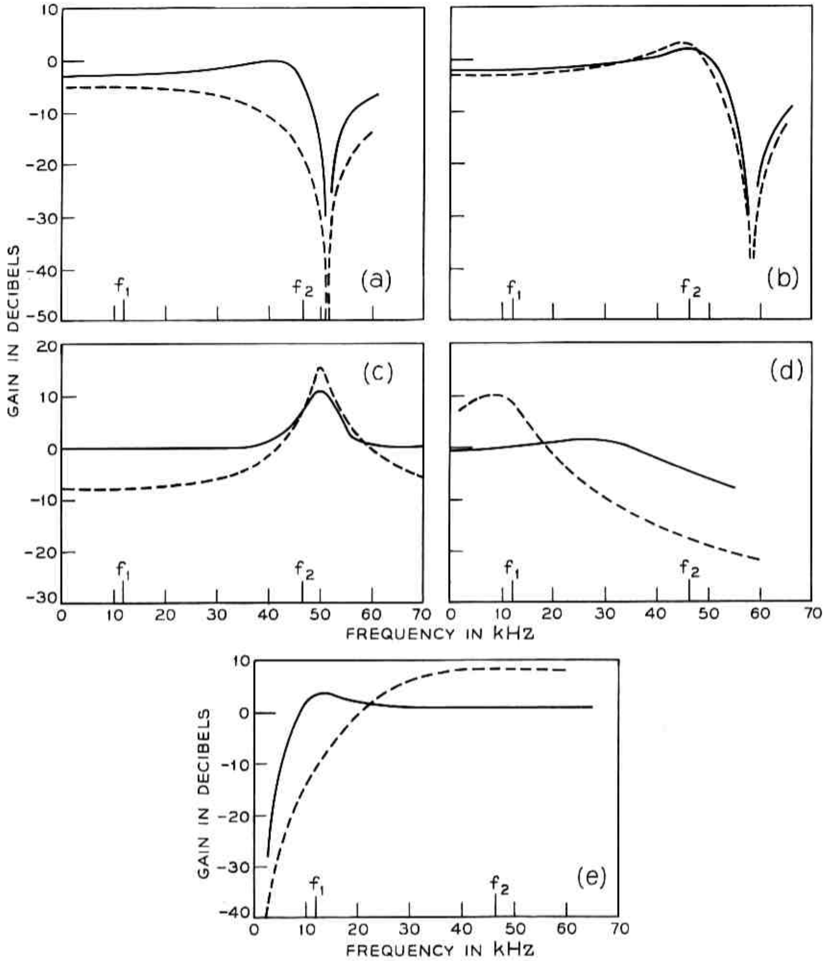

Fig. 6a — Transfer function $T_j, j = 1$.

$$T_1(s) = 0.85 \frac{s^2 + 0.1039 \cdot 10^{12}}{s^2 + (0.5522 \cdot 10^5)s + 0.8844 \cdot 10^{11}} \text{ (solid line);}$$

$$T_1'(s) = \frac{.575s^2 + 59.742 \cdot 10^9}{s^2 + (25.909 \cdot 10^4)s + 9.911 \cdot 10^{10}} \text{ (dotted line).}$$

Fig. 6b — Transfer function $T_j, j = 2$.

$$T_2(s) = 0.75 \frac{s^2 + 0.135 \cdot 10^{12}}{s^2 + (0.06296 \cdot 10^6)s + 9.911 \cdot 10^{10}} \text{ (solid line);}$$

$$T_2'(s) = .4586 \left[\frac{s^2 + .135 \cdot 10^{12}}{s^2 + (.5522 \cdot 10^5)s + .8844 \cdot 10^{11}} \right] \text{ (dotted line).}$$

the frequency response of Fig. 7, where the full line represents the new, the dotted line the earlier version. The new version has a minimum attenuation of 0 dB in the passband, instead of −17 dB in the earlier case. If needed, the new version is able to deliver some amplification in the passband. In the earlier version a maximum voltage swing of $0.3V_{pp}$ at the input was admissible because of overdriving, while in the new version the maximum voltage swing is limited by the amplifiers and not by the peaks of the transfer-functions. Using the op. amp. RCA 3015A, the maximum voltage swing in the new version is $1.8V_{pp}$, since the amplifiers alone have a voltage swing of $1.8V_{pp}$. A way to improve the dynamic range of the amplifiers by minimizing the current drain can be found in Ref. 8. This method can also be used in connection with the transfer-functions $T_j(s)$ found by the method presented in this paper.

VI. CONCLUSIONS

The given transfer function of a filter, which is to be realized by an RC-active two-port, is generally factored into second order functions. A method has been presented to achieve this so that the whole filter has minimum inband losses and maximum dynamic range in which no overdrive of the amplifiers (that is, no distortion) occurs. The problem led to an assignment problem of the bottleneck type. The efficiency of the method has been shown in the case of an SSB-filter, where the in-

Fig. 6c — Transfer function T_j, $j = 3$.

$$T_3(s) = \frac{s^2 + 0.06296 \cdot 10^6 s + 9.911 \cdot 10^{10}}{s^2 + (1.7412 \cdot 10^4)s + 9.911 \cdot 10^{10}} \text{ (solid line);}$$

$$T_3'(s) = \frac{4015^2 + 10.412 \cdot 10^4 s + 39.84 \cdot 10^9}{s^2 + (1.7412 \cdot 10^4)s + 9.911 \cdot 10^{10}} \text{ (dotted line).}$$

Fig. 6d — Transfer function T_j, $j = 4$.

$$T_4(s) = 3.277 \cdot 10^{10} \frac{1}{s^2 + (1.8668 \cdot 10^5)s + 4.357 \cdot 10^{10}} \text{ (solid line);}$$

$$T_4'(s) = \frac{15.287 \cdot 10^7}{s^2 + (5.056 \cdot 10^4)s + 4.594 \cdot 10^9} \text{ (dotted line).}$$

Fig. 6e — Transfer function T_j, $j = 5$.

$$T_5(s) = 0.834 \frac{s^2}{s^2 + (5.056 - 10^4)s + 4.594 \cdot 10^9} \text{ (solid line);}$$

$$T_5'(s) = \frac{2.106 s^2}{s^2 + (1.8668 \cdot 10^5)s + 4.357 \cdot 10^{10}} \text{ (dotted line).}$$

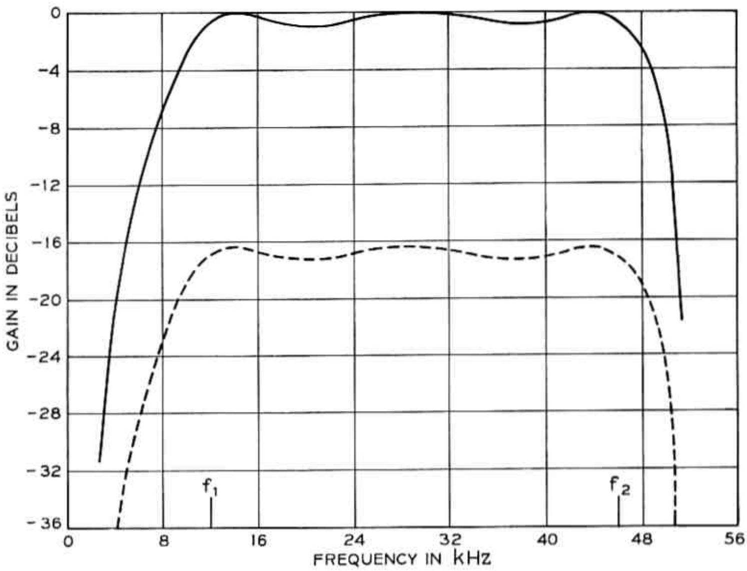

Fig. 7 — Gain of the bandpass. Earlier version with dotted line; new version with solid line.

band loss could be reduced to 0 dB instead of −17 dB in an earlier version and where the dynamic range of the input signal could be increased to $1.8V_{pp}$ instead of $0.3V_{pp}$ as before.

VII. ACKNOWLEDGMENTS

The author acknowledges the stimulating and helpful discussions with G. S. Moschytz, M. Segal and C. J. Steffen. He thanks D. G. Medill, who calculated the component values for three sections and W. D'Zio, who built up the whole SSB-filter and did all the measurements.

REFERENCES

1. Segal, M., private communication.
2. Gross, O., "The Bottleneck Assignment Problem, An Algorithm," Rand Corporation Paper, March 6, 1959, p. 1630.
3. Halfin, S., "Optimization Method for Cascading Filters," B.S.T.J., *49,* No. 2 (February 1970), pp. 185–190.
4. Lueder, E., "Cascading of RC-Active Two-Ports in Order to Minimize Inband Losses and to Avoid Distortion," Proc. Int. Conf. on Comm., Boulder, Colorado, June 9–11, 1969.
5. Moschytz, G. S., "Active RC Filter Building Blocks Using Frequency Em-

phasizing Networks," IEEE Journal of Solid State Circuits, *SC-2,* No. 2 (June 1967), pp. 59–62.

6. Lueder, E., "Some Filters with Variable Gain," unpublished work.
7. Malek, G., "Design of the Modified Raised Cosine and the SSB Transmit and Receive Filters for the 304 Data Set Using Active RC Filter Building Blocks," unpublished work.
8. Paris, D. J., "A Strategy for Improved Dynamic Range of Active Filter Sections," unpublished work.

Contributors to This Issue

MORGAN M. BUCHNER, JR., B.E.S., 1961, Ph.D., 1965, Johns Hopkins University; U. S. Army active duty, 1966–1968; Bell Telephone Laboratories, 1965–66 and 1968—. Mr. Buchner has been interested in the design and performance of data transmission systems. He is presently a supervisor in the Traffic Research Department. Member, IEEE, Tau Beta Pi, Sigma Xi, Eta Kappa Nu.

ROBERT B. COOPER, B.S., 1961, Stevens Institute of Technology; M.S., 1962, and Ph.D., 1968, University of Pennsylvania; Bell Telephone Laboratories, 1961–1969; Georgia Institute of Technology, 1969—. At Bell Labs, Mr. Cooper worked on a variety of problems concerned with applications of probability theory to the analysis of telephone systems and taught the GSP course, Probability Applied to Traffic Engineering. As Associate Professor of Industrial and Systems Engineering at Georgia Tech, he teaches courses in probability, statistics, and queuing theory. He is writing a textbook in queuing theory.

WILLIAM H. IERLEY, B.A., 1959, Drew University; M.S., 1967, N.Y.U.; Bell Telephone Laboratories, 1966—. As a resident visitor at Bell Laboratories from 1964 to 1966, Mr. Ierley worked on various phases of the Nike-X program. He has since been engaged in applications of computer technology to electromagnetic research, emphasizing computer graphics.

NUGGEHALLY S. JAYANT, B.Sc., 1962, University of Mysore (India); B.E. (Dist.), (1965), Ph.D. (1970), Indian Institute of Science, Bangalore; Research Associate (1967–68), Stanford Electronics Laboratories; Bell Telephone Laboratories, October 1968—. Mr. Jayant has worked on digital communication in the presence of burst-noise, and on the detection of fading signals. He is currently conducting research in the encoding of speech and video signals. Member, IEEE, Sigma Xi.

ERNST LUEDER, Dipl.-Ing., 1958; Dr.-Ing., 1962; Habilitation, 1967, University of Stuttgart (Germany); Bell Telephone Laboratories, 1968—. At Stuttgart Mr. Lueder was engaged in network synthesis, theory of nonlinear and electromechanical circuits and system theory.

Since joining Bell Laboratories he has specialized in design of RC-active filters. Member, German associations of Engineers VDE and NTG.

J. A. SEMAN, Graduate, 1968, RCA Institutes; Bell Telephone Laboratories, 1968—. Mr. Seman is engaged in measurements of the bulk properties of optoelectronic materials. His current work involves infrared absorption studies of Gallium Phosphide.

EDWARD A. WALVICK, B.E., 1964, Cooper Union; M.S. (E.E.), 1966, Ph.D. (E.E.), 1969, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1968—. Mr. Walvick has been concerned with problems relating to the exchange cable networks, including methods of determining and specifying their quality. Member, IEEE, Eta Kappa Nu, Tau Beta Pi.

GREGORY H. WANNIER, Louvain University, 1930–31; University of Cambridge, 1933–34; Ph.D., University of Basel, 1935. Assistant, University of Geneva, 1935–36; Swiss-American Exchange Fellow, Princeton University, 1936–37; instructor, University of Pittsburgh, 1937–38; assistant lecturer, Bristol University, 1938–39; instructor, University of Texas, 1939–41; University of Iowa, 1941–46; Socony-Vacuum Laboratories, 1946–49; Bell Telephone Laboratories, 1949–60; Professor of Physics, University of Oregon, 1961—. In the Physical Research Department at Bell Labs, Mr. Wannier worked on electron motion and related solid state phenomena, and the motion of ions in gases. Member, American Physical Society, Swiss Physical Society.

J. D. WILEY, B.S., 1964, Indiana University; M.S., 1965, University of Wisconsin; Ph.D., 1968, University of Wisconsin; Bell Telephone Laboratories, 1968—. Mr. Wiley is engaged in studies of the optical and transport properties of semiconductors. Current work involves infrared absorption studies of defects in Gallium Phosphide. Member, APS.

H. ZUCKER, Dipl. Ing., 1950, Technische Hochschule, Munich, Germany; M.S.E.E., 1954, Ph.D., 1959, Illinois Institute of Technology; Bell Telephone Laboratories, 1964—. Mr. Zucker has been concerned with satellite communication antennas, optical resonators and problems related to physical and geometrical optics. Member, IEEE, Eta Kappa Nu, Sigma Xi.

B.S.T.J. BRIEF

Electrochemically Controlled Thinning of Silicon

By H. A. WAGGENER

(Manuscript received December 31, 1969)

A method for precision thinning silicon integrated circuit slices has been developed whereby either n or p type regions may be selectively removed from material of opposite conductivity. The existence of a simple and economical means to attain precise thickness control permits more complete advantage to be taken of many silicon IC structures. For example, precise thickness control, together with anisotropic[1] etching of isolation/separation slots, is expected to permit economical fabrication of high component density, air-isolated monolithic[2] integrated circuits.

This method differs from previous electrochemical techniques[3] in that unwanted silicon is removed chemically, while the regions to be retained are passivated electrochemically. Accordingly, etchants are used for which silicon to be retained is passive when biased above some critical voltage, V_{pass}, while regions to be removed are at a potential below V_{pass}.

Hot aqueous alkaline solutions form a useful class of etchants for this application, for orientations other than (111). These etchants are characterized by a relatively sharp active/passive transition ($V_{\text{cell}} = V_{\text{pass}} \approx 0.5$ volt) and by a large ratio of silicon etch rates between the active and passive states. Ratios of greater than 200 : 1 are readily obtained. The ratio of active etch rate/passive etch rate is very important, because this quantity in part determines the thickness uncertainty.

Application of the technique to the formation of thin, uniformly thick n type silicon slices is illustrated in Fig. 1. If $V_{\text{cell}} > V_{\text{pass}}$, then $V_n = V_{\text{cell}} > V_{\text{pass}}$ and the n region will be retained. If V_{cell} is restricted to voltages such that the leakage of the reverse biased junction is

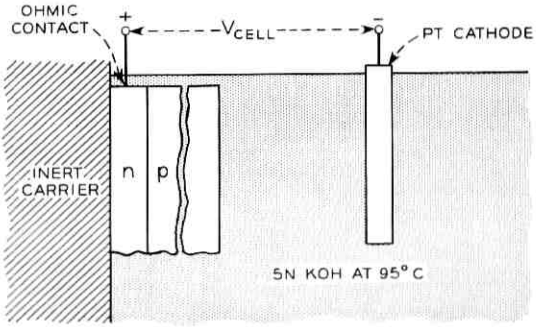

Fig. 1 — Schematic illustration of the thinning technique as used to form n type slices with accurately controlled thickness. An n region is formed by epitaxy or by diffusion. The starting material is high-resistivity p material. When the cell is biased as described in the text, the p region can be removed while the n region is retained. Neither the exact composition nor temperature of the etching solution is critical.

sufficiently small, then $V_p < V_{pass}$, and the p region will be removed. It is assumed that the contacts are arranged so that the lateral ohmic drop in the n type silicon is small enough to be negligible.

For the cell arrangement shown, V_{pass} is about 0.5 volt for either n or p material. Thus thin p-type slices can be formed by reversing the location of the n and p layers as shown in Fig. 2. The maximum allowable cell voltage is reduced because the controlling junction is now forward biased.

Structures of this type have been made where the n layer was formed by diffusion into a background approximately 2.5×10^{-2} cm thick.

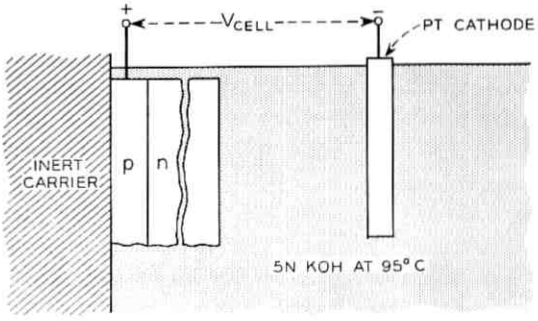

Fig. 2 — Schematic illustration of the thinning technique as used to form p type slices with accurately controlled thickness. Cell polarity is unchanged.

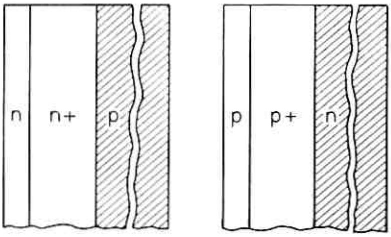

Fig. 3 — Multilayered structures which can be fabricated using the thinning technique. These structures are particularly suitable for making npn and pnp integrated circuits.

After thinning, the difference between the thickness of the remaining n layer and the depth of the diffusion was about 3×10^{-5} cm.

Multilayered structures can also be fabricated as illustrated in Fig. 3. These structures are particularly suitable for fabrication of npn and pnp air-isolated or dielectric-isolated integrated circuits. The thickness control is determined by the combined thickness of the diffused and epitaxial layers, and is expected to be easily controllable to within 10 percent.

Experimental beam leaded, air-isolated monolithic integrated circuits have been made on n/n+/p starting material and have been thinned by the technique described. A total of approximately 2.5×10^{-2} cm of p material was removed in one step, without prior mechanical operations.

REFERENCES

1. Waggener, H. A., Kragness, R. C., and Tyler, A. L., "Anisotropic Etching for Forming Isolation Slots in Silicon Beam Leaded Integrated Circuits," IEEE International Electron Devices Meeting, Washington, D. C., October 19, 1967 (talk).
2. Lepselter, M. P., "Beam Lead Technology," B.S.T.J., *45,* No. 2 (February 1966), pp. 233–253.
3. Dutch patent No 67030B, Aug 26, 1968.